И.П. Гайдышев

Моделирование стохастических и детерминированных систем: Руководство пользователя программы AtteStat

Версия от 16.02.15

2015

Гайдышев И.П. Моделирование стохастических и детерминированных систем: Руководство пользователя программы AtteStat. – Курган, 2015.

Оглавление

Глава 1. Введение в практический анализ

Настоящая монография посвящена теоретическому обоснованию и описанию приемов работы с программным обеспечением AtteStat. С другой стороны, упоминание конкретного программного обеспечения может рассматриваться лишь как повод для изложения результатов научных литературных изысканий, попытки систематизации известных математических и статистических алгоритмов и представления некоторых оригинальных теоретических исследований, выполненных автором самостоятельно. Данное упоминание любых программных реализаций при необходимости (например, стандартом программного обеспечения в учреждении выбрано иное программное обеспечение) легко может быть опущено, либо изложение может быть привязано к другому программному обеспечению, в котором могут быть представлены реализации аналогичных алгоритмов или которое читатель сочтет более пригодным для решения его практических задач[1].
Напомним, что начиная с версии 13, программное обеспечение AtteStat включает в себя все методы программного обеспечения «Математические и инженерные компоненты ME.com», которое как самостоятельный программный продукт более не предлагается.
В настоящее время программа AtteStat поставляется как единое целое. Однако увеличение числа методов не увеличило объема программы, не сделало более длительным процесс загрузки, что стало доступным благодаря оптимизации архитектуры программы. Кроме того, данное структурированное руководство пользователя заменило Справочную систему.

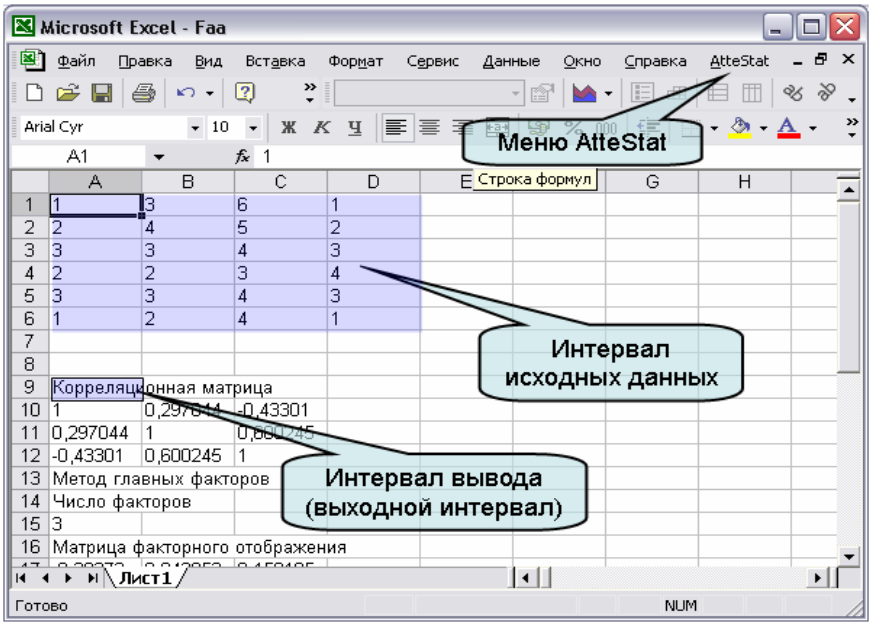
1.1. Как начать работу

Общие принципы работы с программным обеспечением описаны в главе «Введение в практичекий анализ». Раздел «Особенности представления результатов» поможет получить отображение результатов расчета в наиболее удобной для пользователя форме. От неверных результатов вычислений предостережет одноименный раздел. Об ошибках при работе с программным обеспечением и способах их локализации рассказано в соответствующих главах.

1.2. Работа с программным обеспечением

Программное обеспечение AtteStat использует интерфейс 32-разрядных версий электронных таблиц Excel®, функционирующих под управлением 32- или 64-разрядных версий операционных систем Windows®.
Для работы с программным обеспечением AtteStat запустите электронные таблицы или воспользуйтесь специальным скриптом для запуска программного обеспечения AtteStat, находящимся в меню Пуск. При установленном программном обеспечении AtteStat меню станет выглядеть примерно так, как на показанном рисунке (внешний вид окна и расположение меню зависит от типа операционной системы и версии электронных таблиц).

1 Работа над монографией продолжалась параллельно работе над программой AtteStat, поэтому подробность описания различных методов существенно различна – от конспективного описательного и до теории и подробного вывода расчетных формул (в составленных последними по времени главах). Поэтому предполагается, что если у читателя возникнут вопросы по реализации того или иного алгоритма, ему следует обратиться к доступным исходным текстам программы.

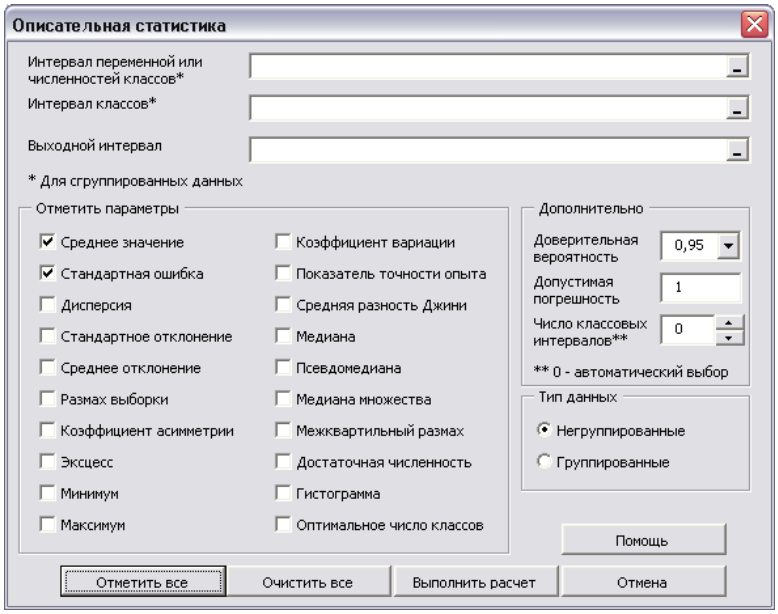


1.2.1. Основные определения

- Под интервалом исходных данных (см. рисунок выше) в программе AtteStat понимается диапазон ячеек листа электронных таблиц, содержащих числовые или иные данные для расчета. Требования к формату данных предъявляются соответствующим методом расчета. Интервалов исходных данных может быть несколько, в зависимости от метода. Они могут содержать как отдельные выборки, так и матрицы исходных данных, а также содержать или не содержать пропущенные данные. Более подробная информация дана при описании конкретных методов расчета.
- Под интервалом вывода (выходным интервалом) понимается диапазон ячеек, в которые будет производиться вывод результатов расчета. Обратим внимание, что вывод производится, начиная с левого верхнего угла выбранного диапазона. Таким образом, в качестве интервала вывода можно указать только одну ячейку, начиная с которой будут выводиться результаты расчета. Методы программного обеспечения AtteStat не оценивают предполагаемый объем выдачи, поэтому следует быть осторожным, указывая интервал вывода, в том смысле, чтобы не затереть нужные данные, содержащиеся в расположенных ниже ячейках листа.

1.2.2. Основные действия

Выберите из меню программы добавленный программным обеспечением пункт AtteStat, затем нужный для расчета раздел. На экране появится диалоговое окно, подобное изображенному на рисунке (может отличаться от реального образа):



Дальнейшие действия пользователя зависят от требований соответствующих методов. От пользователя обычно требуется указать интервал исходных данных и интервал вывода, как описано выше. Также необходимо выбрать, ввести, отметить или оставить по умолчанию метод расчета и вспомогательные параметры, номенклатура и количество которых зависят от применяемого метода. Некоторые параметры могут относиться ко всем представленным методам. Другие параметры – только к некоторым. В любом случае перед началом расчета новым для пользователя методом рекомендуется ознакомиться с предпосылками и порядком его применения. В противном случае велика вероятность неуспеха, особенно ошибочной трактовки полученных результатов.

Пользователям, не имеющим твердых навыков работы с электронными таблицами или имеющим навыки работы с другими программами анализа данных, полезно посетить раздел «Ввод данных». В разделе дан иллюстрированный обзор элементов управления и порядок работы с ними.

Ко всем методам программного обеспечения AtteStat неприменимы стандартные операции отмены типа выбора из меню Правка | Отменить …, поэтому следует быть внимательным, а перед производством расчета настоятельно рекомендуется сохранить свои файлы выбором из меню Файл | Сохранить.

1.2.3. Ввод данных

В программном обеспечении AtteStat для управления исходными данными применяются стандартные средства. Кратко рассмотрим их возможности. Вот некоторые из них, доступные пользователям:

- CheckBox – флажок (кнопка независимого выбора),
- ComboBox – элемент управления, отображающий список величин и позволяющий выбрать одну их них,
- CommandButton – командная кнопка,
- MultiPage – многостраничный элемент управления,
- OptionButton – переключатель (кнопка зависимого выбора, иначе радиокнопка, название заимствовано от переключателя частотных диапазонов старинного радиоприемника),
- RefEdit – поле ссылки на ячейки листа электронной таблицы,

- SpinButton – кнопка инкремента/декремента числового значения,
- TextBox – поле ввода (окно редактирования, «текстовое» поле).

Данные наименования хорошо знакомы программистам на языке Visual Basic. Здесь мы их приводим для пользователей только с тем, чтобы обозначить и различить обсуждаемые сущности.

Работа с представленными элементами управления эффективна с помощью манипулятора «мышь» или его аналога. Возможна работа и с клавиатуры с использованием клавиш управления курсором, клавиш Tab («табуляция») и Space («пробел»). Работа «мышью» и с клавиатуры стандартна.

Флажок CheckBox позволяет пользователю выбрать ту или иную опцию, привязанную к данному флажку. В программе флажок имеет два состояния: «не выбран» и «выбран». Значение флажка по умолчанию зависит от контекста.

Элемент управления ComboBox отображает список величин (например, стандартных доверительных уровней) и позволяет пользователю выбрать только одну из этих величин.

Командная кнопка CommandButton служит для запуска на выполнение той или иной процедуры, например, расчета («Выполнить расчет») или выхода из формы без производства вычислений (Отмена). К командной кнопке иногда привязаны и другие события, например, установка всех флажков в значение «выбран» или «не выбран».

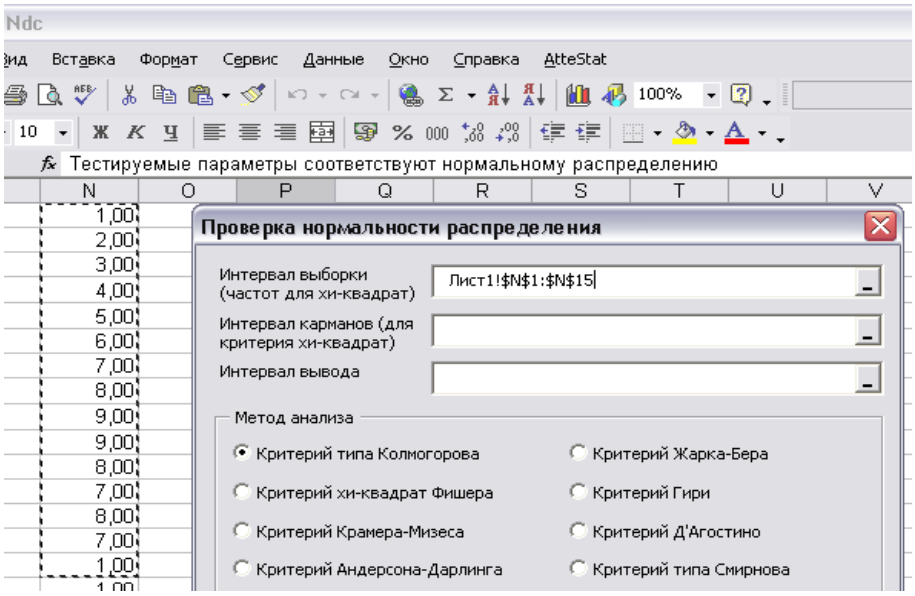
Элемент управления MultiPage применяется в главе «Распознавание образов с обучением». Данный элемент позволяет отобразить в компактной форме большое число элементов управления. Например, элемент позволяет отобразить как обучающие, так и распознающие алгоритмы. Для перехода на ту или иную страницу описываемого элемента достаточно щелкнуть левой кнопкой «мыши» по ярлычку нужной страницы.

Переключатель OptionButton служит для обеспечения зависимого выбора, иначе одного выбора из нескольких возможных.

Основным элементом управления, обеспечивающим удобную передачу данных с листа электронных таблиц в расчетные модули, является поле RefEdit. Для начала работы необходимо установить курсор на данный элемент управления. Затем следует воспользоваться одним из следующих стандартных (!) приемов работы:

- Ввести с клавиатуры интервал ячеек, содержащих данные для расчета.
- Указать методом протаскивания курсора интервал ячеек, содержащих данные для расчета. При использовании «мыши» протаскивание производится проведением курсора «мыши» по ячейкам при нажатой левой кнопке (для стандартной настройки «мыши»).
- Воспользоваться комбинированным методом, особенно удобным для больших объемов данных (например, содержащих несколько сотен строк): протащить курсор по небольшой части данных, а затем отредактировать интервал «вручную». Например, пусть нам нужно указать интервал ячеек, содержащий 768 строк (от 1 до 768) на 8 столбцов (от A до I). Пролистывание всей таблицы было бы утомительным. Поэтому, протащив курсор по части таблицы, мы получаем интервал, скажем, Лист2!$A$1:$I$7. Отредактировав данный интервал «вручную», получаем требуемый интервал Лист2! $A$1:$I$768.

На рисунке показан типичный результат протаскивания курсора по ячейкам листа электронной таблицы при заполнении поля.



Поле RefEdit имеет еще одно удобство. Нажатие значка, расположенного справа внутри поля приводит к сворачиванию формы, после чего пользователь имеет возможность свободно перемещаться по рабочему листу (если нажатие на значок не было произведено, форма сворачивается автоматически на момент протаскивания курсора, затем автоматически восстанавливается). Повторное нажатие на данный значок восстанавливает форму для продолжения ввода или выбора остальных параметров расчета.
Кнопка SpinButton служит для увеличения на единицу (инкремента) или уменьшения на единицу (декремента) некоторого числового значения. Применяется для плавного изменения пользователем целочисленного параметра расчета (например, числа классов в главе «Кластерный анализ»). Слева от кнопки в текстовом поле, программно привязанном к данной кнопке, отображается результат ее работы.
В поле ввода TextBox с клавиатуры вводятся данные различных типов. Верификация введенных данных производится программой.

1.2.4. Примеры

«Долог путь поучений, короток и успешен путь примеров» (Сенека Младший). Для удобства пользователей и с целью пояснений по технологии работы с программой приводятся примеры применения, позволяющие пользователю на конкретных данных, почерпнутых из известных источников, изучить порядок работы с программой и оценить ее работоспособность.

1.2.5. Ссылки

Данный литературный труд находится в постоянном развитии, поэтому стиль ссылок, обычно принятый в книгах (монографиях) – т. е. по номеру источника в списке литературы, нам показался неприемлемым из-за необходимости постоянного перестроения данных списков (к сожалению, безупречных программных систем, позволяющих гибко и надежно делать это автоматически, в настоящее время не существует). Поэтому принят стиль ссылок на источники, похожий на научные статьи (а именно – по фамилиям авторов). По этой же причине ссылки на главы данной же монографии выполнены аналогично. Кроме того, в списки источников включены работы, прямых ссылок на которые в данной монографии не делается, но на которые автор хотел бы обратить внимание читателей.

Возможно, такой стиль ссылок покажется неудачным, если относиться к данному произведению, как к классическому научному труду, которым оно не является. При необходимости, определенном усердии и доступе к оригиналам работ цитируемых авторов местонахождение ссылок (с точностью до страницы источника) может быть легко установлено.

1.3. Теоретическое обоснование

Программное обеспечение, представленное в монографии, является реализацией определенных научных концепций, которые в итоге получили воплощение в некоторых математических формулах. В силу этого могут возникнуть вопросы, правильно ли данные формулы запрограммированы, а также в каких диапазонах обрабатываемых исходных данных или параметров гарантируется получение правильных результатов вычислений. Ответам на данные вопросы посвящается настоящий параграф.

1.3.1. Правильность вычислений

Механистическое применение вычислительных методов, без ознакомления с порядком и предпосылками их применения, редко приводит к успеху. Однако и при соблюдении всех предпосылок возможно получение неверных, с точки зрения пользователей, результатов вычислений. Если осталось убеждение, что проблема в программе, автор будет чрезвычайно рад получить информацию об ошибке, включающую исходные данные (в любом формате) и указание на ошибочный метод. В общих интересах сделать программу максимально полезной. Мы не исключаем ни наличия алгоритмических ошибок, ни таких наборов данных, которые способны привести программу к краху.
Программное обеспечение как продукт научного творчества является благоприятным объектом, в отличие от статьи или монографии, в плане постоянного контроля над ним. Оно всегда может быть скорректировано автором или его коллегами и, таким образом, находится в состоянии постоянного развития. Прежние версии в любое удобное время могут быть заменены более совершенными вариантами, а современные средства коммуникации позволяют оперативно произвести обновление копий программного обеспечения на компьютерах пользователей.

1.3.1.1. Методики проверки

Предлагаем ознакомиться с методиками проверки, использованными при разработке программного обеспечения, а также с руководством пользователя к действию при возможном обнаружении неверных, с его точки зрения, результатов.
Для любой вычислительной программы существенной является проблема уверенности пользователя в правильности вычислений. В процессе кодирования алгоритмов программы использовались следующие методы проверки правильности программных реализаций:

1. «Ручной» счет, чтобы убедиться в правильности расчета программой как в особых контрольных точках, обычно соответствующих основным этапам алгоритма, так и в правильности результата в целом. Данный метод осложняется тем обстоятельством, что трудоемкость «ручного» счета растет с ростом численности выборок настолько, что при определенных пределах объема исходных данных он становится нереализуемым в приемлемые сроки.
2. Разновидностью «ручного» счета является выполненный вручную графический метод, применяемый для проверки тестов, которые основаны на представлениях, могущих иметь интуитивно понятное графическое отображение. Например, расчет критериев, основанных на тех или иных функциях распределения, типа критериев Колмогорова

или Смирнова (см. главу «Проверка нормальности распределения»), можно представить графически и статистику критерия получить измерением линейкой.

3. Сопоставление результатов расчета с опубликованными результатами. Данный метод проверки прост, очевиден и быстр, если программист или исследователь настолько наивен, чтобы безоговорочно доверять любому опубликованному тексту. Метод осложняется тем обстоятельством, что слишком часто опубликованные результаты, независимо от авторитета издательств и авторов работ, содержат неизбежные опечатки и ошибки и в формулах, и в математических вычислениях. В практике встречались случаи, когда поиск гипотетических ошибок в собственных программах приводил к нахождению ошибок в источниках, и было жаль потраченного времени.
4. Сопоставление с результатами расчета аналогами. Данный метод проверки – первое, что пытается сделать любознательный пользователь. Он осложняется тремя обстоятельствами: высокой стоимостью тестируемых программных продуктов, возможным отсутствием в их составе требуемых алгоритмов, возможно неправильной работой программного продукта, выбранного на роль эталона. Кроме всего, в разных программах методы могут быть запрограммированы правильно, но разными способами, использовать различные критические значения статистик (односторонние и двухсторонние, для простых и сложных гипотез, более или менее точные результаты компьютерного моделирования), иметь различные поправки, условности и ограничения своего применения. Часто одинаковые названия носят не только разные модификации одно и того же теста, но и разные методы вообще.
5. Сопоставление результатов расчета с функциональными аналогами собственной разработки. Одним из основных тезисов политики разработки AtteStat является уникальная номенклатура методов с целью удовлетворения вычислительных потребностей исследователей. Поэтому на определенном этапе количество наработок позволило использовать для тестирования новых алгоритмов свои собственные разработки, как внедренные в пакет, так и оставшиеся в опытных версиях. Подробные рассуждения об эквивалентных алгоритмах см. в главе «Непараметрическая статистика».
6. Вычисления на основе специально сгенерированных или вымышленных выборок. Для контроля устойчивости программы обязательно производится расчет на особых экстремальных выборках.

1.3.1.2. Действия пользователя

Если пользователь предположил возможность получения неверных результатов вычислений вследствие ошибок в программе и его действительно интересует благополучное разрешение ситуации, рекомендуется придерживаться следующего порядка действий:

1. Убедиться, что исходные данные введены в той форме, которая требуется, а выводы интерпретированы так, как это указано в описании используемого алгоритма. Требования к представлению исходных данных излагаются в соответствующих разделах.
2. Убедиться, что исходные данные соответствуют требованиям алгоритма. Прежде всего, адекватными должны быть: шкала измерения, размерность выборки, тип данных (исходная выборка, вариационный ряд, корреляционная матрица, таблица сопряженности и другие). Убедиться, что исходные данные соответствуют теоретическим допущениям алгоритма. Для некоторых методов – это нормальность распределения, для других – отсутствие линейной зависимости в матрице исходных данных, сложность гипотезы и т. д.
3. При сравнительной проверке убедиться, что в аналогичных программах используется

тот же метод расчета, а также проверяются те же параметры выборок.

1.3.2. Типы и размеры данных

Алгоритмы AtteStat выполнены на стандартном языке программирования C++. Интерфейс пользователя выполнен на языке Visual Basic for Application. Приведем сводку типов и размеров (максимальных и минимальных) исходных данных, которыми может корректно оперировать программа.

Язык	Тип данных	Биты	Минимум	Максимум
Visual Basic for Application	Integer	16	–32768	32767
	Long	32	–2147483648	2147483647
	Double	64	4,94065645841247E–324	1,79769313486231E+308
C++	short	16	–32768	32767
	unsigned short	16	0	65535
	long	32	–2147483648	2147483647
	double	64	1,7E–308	1,7E+308
	long double	80	3,4E–4932	3,4E+4932

Комментарии к таблице:

1. Целые типы данных, как правило, относятся к адресации ячеек таблицы или перечислению элементов массивов (вариант выборок), а также к данным, по своей природе имеющим целый тип. Типы данных с плавающей десятичной точкой относятся к самим вариантам, имеющим количественный тип. Соответственно, все соглашения и ограничения по типам данных относятся к упомянутым характеристикам исходных данных.
2. В таблице приведены некоторые применяемые в программе AtteStat типы, для которых установлено адекватное соответствие между языками программирования, а именно: Long – long и Double – double для Visual Basic for Application и C++, соответственно. Отметим, что хотя типы данных C++ и стандартизованы, их размер определяется компилятором.
3. Максимально допустимое число строк составляет 65536 (при нумерации с 1). Поэтому программное обеспечение AtteStat может оперировать только данным количеством строк.
4. В программных модулях, составленных на языке C++, иногда применяется тип данных long double, указанный в таблице и не имеющий аналога в Visual Basic for Application. Это сделано для повышения точности (а иногда и самой возможности выполнения) некоторых промежуточных процедур вычислений. При этом окончательные результаты всегда конвертируются в тип double для совместимости с типом Double языка Visual Basic for Application.

На допустимые размеры исходных данных может накладывать ограничения и применяемый алгоритм. Так, часто в промежуточных вычислениях различных параметров применяются квадраты исходных вариант. В этом случае необходимо учитывать естественное изменение допустимых значений исходных вариант.

1.3.3. Шкала измерения

Перед применением метода необходимо убедиться, что он соответствует шкале измерения исходных данных (признаков). Распределение признаков по шкалам измерения обычно основано на анализе допустимых логических и арифметических операций, которые могут

быть проведены над признаками, как это показано в нижеприведенной таблице.

Шкала измерения	Допустимые действия
Номинальная	Различение
Порядковая	Различение, сравнение
Количественная	Различение, сравнение, сложение, умножение

Классификация включает признаки:

1. Номинальные признаки (nominal) – качественные признаки с неупорядоченными состояниями, классификационные признаки, категоризированные признаки. Например, переменная «тип транспортного средства» принимает значения: «велосипед», «мотоцикл», «автомобиль». Номинальные признаки могут быть оцифрованы, однако смысла эти цифры, за исключением возможности различать признаки между собой, не имеют. Частным случаем номинальных признаков являются бинарные (качественные, дихотомические) признаки, представляющие собой номинальные признаки с двумя градациями, например: «нет» – 0, «да» – 1. Подробнее о представлении бинарных выборок см. в разделе «Таблицы 2 x 2». Отметим, что некоторыми (особенно зарубежными) авторами вводятся так называемые «естественным образом упорядоченные» (ordered) номинальные признаки. Несомненно, что под данным определением на самом деле имеются в виду не номинальные, а порядковые признаки (ordinal), ибо номинальные признаки не могут быть никаким образом, в том числе естественным, упорядоченными по определению.
2. Порядковые признаки (ordinal) – качественные признаки с упорядоченными состояниями, ординальные признаки (от английского order – порядок, последовательность). Например: отлично, хорошо, удовлетворительно, плохо. Порядок состояний имеет смысл, признаки могут быть осмысленно оцифрованы (в данном примере: 5, 4, 3, 2) и сравниваться между собой, однако расстояния между ними не определены. Особым типом порядковой шкалы является шкала ранжировок, о которой подробно рассказано в главе «Обработка экспертных оценок».
3. Количественные (численные, вариационные) признаки, иногда подразделяемые на интервальные и относительные. Они различаются положением нулевой отметки на шкале измерения. Например, год рождения – относительный количественный признак, а срок службы в рядах вооруженных сил – интервальный количественный признак. Если в первом примере определены только операции различения, сравнения и вычитания, то во втором к ним добавляются операции сложения и отношения. Численные признаки определяют измеряемые количества (величины) и являются истинными количественными, причем могут измеряться как непрерывные, так и целочисленные признаки.
4. Фиктивные (индикаторные) переменные (dummy variables) – это вспомогательные бинарные переменные, принимающие значения только 1 либо 0, которые применяются для введения в регрессионные модели качественных переменных.

1.3.3.1. Фиктивные переменные

Рассмотрим подробнее фиктивные (dummy) переменные и принципы возможного кодирования (не путать с «оцифровкой» – кодирование не меняет шкалу измерения, см. следующий раздел) качественных переменных, что необходимо для обеспечения участия качественных переменных в количественных расчетах наряду с истинно количественными переменными.

Если качественная переменная принимает S фиксированных значений, то теоретически она может быть закодирована N фиктивными переменными, где минимальное значение N, очевидно, определяется из целочисленного неравенства $S \leq 2^N$.
Для пояснения рассмотрим пример. Пусть имеется номинальная переменная «тип двухколесного транспортного средства», принимающая три значения: «мотоцикл», «мотороллер», «велосипед». Согласно показанной формуле, переменная может быть минимально закодирована двумя ($3 \leq 2^2$) фиктивными переменными. При этом возможны случаи:

- первая фиктивная переменная равна 1 (при этом вторая фиктивная переменная равна 0), если транспортное средство – мотоцикл;
- вторая фиктивная переменная равна 1 (при этом первая фиктивная переменная равна 0), если транспортное средство – мотороллер;
- первая фиктивная переменная равна 0 и вторая фиктивная переменная равна 0, если транспортное средство – велосипед.

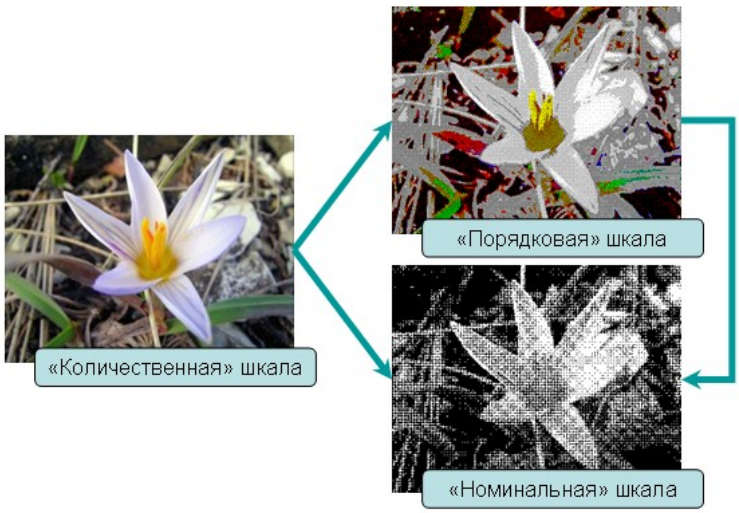
Если стоит задача описать качественную переменную минимальным количеством фиктивных переменных, то следует поступить так, как описано выше. Однако удобством интерпретации такая кодировка не отличается.
Поэтому возможны и иные варианты кодировки. Например, может оказаться очевиднее качественную переменную с S фиксированными значениями закодировать S фиктивными переменными. В аналогичном примере переменная может быть закодирована тремя фиктивными переменными. При этом возможны случаи:

- первая фиктивная переменная равна 1, если транспортное средство – мотоцикл, или 0 – если не мотоцикл;
- вторая фиктивная переменная равна 1, если транспортное средство – мотороллер, или 0 – если не мотороллер;
- третья фиктивная переменная равна 1, если транспортное средство – велосипед, или 0 – если не велосипед.

Для полноты описания отметим еще одну бытующую точку зрения. Некоторые авторы полагают, что всегда $N = S - 1$. Никаких логических препятствий для применения данного подхода нет.

1.3.3.2. Преобразование шкалы

Шкалы могут приводиться одна к другой, как показано стрелками на рисунке: количественная шкала – к порядковой шкале или номинальной, порядковая шкала – к номинальной шкале. Обратные операции считаются некорректными, хотя, к примеру, проблеме т.н. «оцифровки» неколичественных данных посвящено немало источников. Очевидно, что во фрагменте под названием «Количественная шкала» (условно – полноцветное изображение) содержится гораздо больше полезной информации, чем в двух других: «Порядковая шкала» (условно – 256–цветная стандартная палитра) и «Номинальная шкала» (условно – черно–белое изображение). Исследователю обычно не приходится выбирать между шкалами измерения – данные получают и исследуют в той шкале, которая отражает физическую природу изучаемого явления. Важно лишь применять адекватные методы анализа.



Частой ошибкой является попытка применения методов, развитых для признаков, измеренных в количественной шкале, для признаков, измеренных в других шкалах. Пусть некоторый параметр эксперт измеряет в баллах (например, преподаватель «измеряет» успеваемость студента). Иногда пытаются определить некоторый средний балл, забывая, что баллы относятся к порядковой шкале, для которой операции суммирования и деления не определены. Более того, между величинами в порядковой шкале не определены также и расстояния. Например, для пятибалльной шкалы успеваемости нельзя утверждать, что оценка «5» отличается от «4» настолько же, насколько «3» отличается от «2». Можно лишь утверждать, что «5» в определенном смысле лучше, чем «4», а «3» лучше, чем «2».
Некоторые методы программного обеспечения AtteStat, когда это возможно и необходимо (например, см. главу «Непараметрическая статистика»), проверяют адекватность типа исходных данных, например, не позволяя ввести в качестве номинальных данных выборки, содержащие величины, отличные от нуля и единицы. Данная проверка сделана для повышения устойчивости программы к ошибкам ввода и предостережения от получения бессмысленных результатов расчета.

Более подробно о шкалах измерения см. в учебном пособии Борцова.

1.3.4. Табличные данные

1.3.4.1. Таблицы 2 x 2

Двухвходовые таблицы сопряженности типа 2 x 2 возникают в результаты сопоставления двух бинарных (дихотомических) выборок, т. е. выборок, состоящих из значений 1 и 0, причем под значением 1 обычно понимают наличие признака, под значением 0 понимают отсутствие признака.
Выборки рассматриваемого могут быть представлены в виде таблиц типа 2 x 2 различными способами, в зависимости от того, являются ли выборки независимыми или парными. Ниже представлены способы получения таблиц 2 x 2 и указаны их существенные особенности.

1.3.4.1.1. Независимые выборки

Порядок построения таблицы из вариант независимых выборок иллюстрируется следующей таблицей:

	Наличие эффекта *A*	
	Да	Нет
Выборка (группа) 1	*a*	*b*
Выборка (группа) 2	*c*	*d*

При этом в ячейки заносятся:
a – число значений с эффектом *A* первой выборки,
b – число значений без эффекта *A* первой выборки,
c – число значений с эффектом *A* второй выборки,
d – число значений без эффекта *A* второй выборки.
Таблицы данного типа могут применяться при анализе данных типа «опыт – контроль» или сравнении двух независимых методов воздействия типа «группа 1 – группа 2». Численности выборок могут как совпадать, так и различаться.
В настоящем программном обеспечении указанные выборки вводятся стандартно.

1.3.4.1.2. Парные выборки

Для парных (сопряженных) выборок порядок построения таблицы иллюстрируется следующей таблицей:

		Эффект *B*	
		Да	Нет
Эффект *A*	Да	*a*	*b*
	Нет	*c*	*d*

В данном случае анализу подвергается фактически одна двумерная выборка – выборка пар значений, первое значение пары – наличие или отсутствие эффекта *A*, второе – наличие или отсутствие эффекта *B*. Поэтому в ячейки таблицы заносятся:
a – число пар значений с эффектом *A* и с эффектом *B*,
b – число пар значений с эффектом *A* и без эффекта *B*,
c – пар число значений без эффекта *A* и с эффектом *B*,
d – пар число значений без эффекта *A* и без эффекта *B*.
Таблицы данного типа могут эффективно применяться при анализе данных типа «до — после».
В описаниях методов иногда применяется формальная система обозначений, отличная от показанной выше системы. Система более громоздка, но более удобна с математической точки зрения: *a* – это n_{11}, *b* – это n_{12}, *c* – это n_{21}, *d* – это n_{22}. При этом первая цифра индекса указывает номер строки таблицы, вторая – номер столбца.

1.3.4.2. Двухвходовые таблицы типа r x c

Пусть обозначено:
r – число градаций первого признака,
c – число градаций второго признака,
n_{ij}, $i = 1,2,...,r$; $j = 1,2,...,c$ – численность вариант, обладающих одновременно *i*–й градацией первого признака и *j*–й градацией второго признака.
Тогда таблица сопряженности будет иметь вид:

n_{11} n_{12} … n_{1c}
n_{21} n_{22} … n_{2c}
… … … …
n_{r1} n_{r2} … n_{rc}

Порядок признаков (столбцы или строки) значения не имеет. При анализе таблицы сопряженности условились количество строк таблицы обозначать символом *r* (от английского слова rows), а количество столбцов – символом *c* (от английского слова columns), хотя могут встречаться и любые другие обозначения. Например, в некоторых источниках, количество строк и столбцов может быть обозначено, соответственно, *I* и *J*. Таким образом, в общем случае двумерная таблица сопряженности будет именоваться *r* x *c* или *R* x *C* таблицей. Каждая клетка таблицы сопряженности с индексами *i*, *i* = 1,2,...,*r*, (номер строки) и *j*, *j* = 1,2,...,*c*, (номер столбца) представляет собой количество индивидуумов, обладающих одновременно градацией *i* первого признака и градацией *j* второго признака. Данное количество называется наблюдаемой (наблюденной) частотой встречаемости признаков. Таким образом, в общем случае методами кросстабуляции исследуется зависимость первого номинального признака с числом градаций *r* от второго номинального признака с числом градаций *c*. Если таблица сопряженности квадратная (числа градаций для первого и второго признаков одинаковы), то часто используется обозначение: таблица типа *k* x *k*, где *k* – число градаций каждого признака.

В программе электронных таблиц двухвходовые таблицы сопряженности органично задаются прямоугольным фрагментом рабочего листа размером *r* x *c*. Для анализа таблица сопряженности должна быть полностью заполненной.

Некоторые практические вопросы

1. Иногда пользователи задают вопрос, как строить таблицу сопряженности для массива данных из двух зависимых или независимых выборок равной или неравной численности. Так, например, могут быть представлены для анализа выборки, одна из которых является опытной, а другая – контрольной. Порядок действий тут прост и формально повторяет изложенную выше процедуру. Первый признак при этом является тем физическим признаком, влияние которого исследуется. Число его градаций равно *r*. Вторым признаком является принадлежность к выборке. Число градаций второго признака, очевидно, равно 2, а сами эти градации: «принадлежит к первой выборке» и «принадлежит ко второй выборке». Дальнейший анализ ничем не отличается от стандартного подхода к анализу таблицы сопряженности. Таким образом, в данном частном случае методами кросстабуляции исследуется зависимость первого номинального признака с числом градаций *r* от второго номинального (дихотомического) признака с числом градаций, равным 2.
2. Для некоторых расчетных методов имеет значение, получены таблицы сопряженности из порядковых (в терминологии некоторых авторов – естественным образом упорядоченных) или номинальных (неупорядоченных) признаков. Более того, можно привести примеры методов анализа [двухвходовых] таблиц сопряженности, имеющих различные расчетные формулы для случаев:
    - оба признака неупорядочены (таблица построена на основе двух номинальных признаков);
    - один из признаков упорядочен (таблица построена на основе одного номинального и одного порядкового признака);
    - оба признака упорядочены (таблица построена на основе двух порядковых признаков).

1.3.4.3. Многовходовые таблицы

Многовходовые таблицы сопряженности возникают, когда число признаков превышает 2. Сначала для пояснения принципа обозначений рассмотрим трехвходовую таблицу, а затем обобщим результаты на таблицы сопряженности произвольной размерности.
Введем новые обозначения. Пусть
k_i, i = 1,2,3 – число градаций i–го признака,
(.) – обозначение фиксированного уровня 3–го признака.
Тогда таблица сопряженности для признаков 1 и 2 при фиксированном $k_3 = 1$ имеет вид обычной двухвходовой таблицы:

$$\begin{matrix} n_{11}^{(1)} & n_{12}^{(1)} & \dots & n_{1k_2}^{(1)} \\ n_{21}^{(1)} & n_{22}^{(1)} & \dots & n_{2k_2}^{(1)} \\ \dots & \dots & \dots & \dots \\ n_{k_1 1}^{(1)} & n_{k_1 2}^{(1)} & \dots & n_{k_1 k_2}^{(1)} \end{matrix}$$

Действуя аналогично, получаем и все остальные таблицы:

$$\begin{matrix} n_{11}^{(2)} & n_{12}^{(2)} & \dots & n_{1k_2}^{(2)} & n_{11}^{(\dots)} & n_{12}^{(\dots)} & \dots & n_{1k_2}^{(\dots)} & n_{11}^{(k_3)} & n_{12}^{(k_3)} & \dots & n_{1k_2}^{(k_3)} \\ n_{21}^{(1)} & n_{22}^{(1)} & \dots & n_{2k_2}^{(1)} & n_{21}^{(\dots)} & n_{22}^{(\dots)} & \dots & n_{2k_2}^{(\dots)} & n_{21}^{(k_3)} & n_{22}^{(k_3)} & \dots & n_{2k_2}^{(k_3)} \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ n_{k_1 1}^{(2)} & n_{k_1 2}^{(2)} & \dots & n_{k_1 k_2}^{(2)} & n_{k_1 1}^{(\dots)} & n_{k_1 2}^{(\dots)} & \dots & n_{k_1 k_2}^{(\dots)} & n_{k_1 1}^{(k_3)} & n_{k_1 2}^{(k_3)} & \dots & n_{k_1 k_2}^{(k_3)} \end{matrix}.$$

Таким образом, видим, что трехвходовая таблица сопряженности представляет собой своеобразный «куб» со сторонами k_1 x k_2 x k_3. Хорошо заметно, насколько громоздко и неудобно такое представление данных для многовходовых таблиц, поэтому было принято более удобное представление, фактически отражающее ту же самую сущность, т. е. представления взаимозаменяемы. Данное эквивалентное табличное представление многовходовых таблиц (для рассмотренного примера) представлено ниже. Благодаря тем же самым обозначениям понятен порядок построения табличной формы многовходовой таблицы:

- Первый столбец таблицы – градации признака 1.
- Второй столбец таблицы – градации признака 2.
- Третий столбец таблицы – градации признака 3.
- Последний столбец любой таблицы – это количества индивидуумов (частоты), обладающих одновременно градациями признаков, перечисленных в строке, соответствующей данной частоте.

Для уменьшения объема примера ограничимся «размерами»: $k_1 = 3$, $k_2 = 3$, $k_3 = 2$. Тогда искомое представление таблицы будет иметь вид

1 1 1 $n_{11}^{(1)}$
1 2 1 $n_{12}^{(1)}$
1 3 1 $n_{13}^{(1)}$
2 1 1 $n_{21}^{(1)}$
2 2 1 $n_{22}^{(1)}$
2 3 1 $n_{23}^{(1)}$
3 1 1 $n_{31}^{(1)}$
3 2 1 $n_{32}^{(1)}$
3 3 1 $n_{33}^{(1)}$
1 1 2 $n_{11}^{(2)}$
1 2 2 $n_{12}^{(2)}$
1 3 2 $n_{13}^{(2)}$
2 1 2 $n_{21}^{(2)}$
2 2 2 $n_{22}^{(2)}$
2 3 2 $n_{23}^{(2)}$
3 1 2 $n_{31}^{(2)}$
3 2 2 $n_{32}^{(2)}$
3 3 2 $n_{33}^{(2)}$

В общем случае таблица представляет собой все возможные сочетания градаций признаков k_i, i = 1,2,..., и соответствующие им частоты. Поэтому размер таблицы будет равен $\prod_{i=1}^{n} k_i$ строк на n + 1 столбцов, где n – количество изучаемых признаков. Если пользователем введено число строк меньшее, чем вычислено по указанной формуле, либо перебор сочетаний признаков не полный, программа должна отслеживать такие ошибочные ситуации.

Рассмотренное представление позволяет изобразить на «плоскости» таблицы сопряженности произвольной размерности. Данное плоское представление многовходовых таблиц сопряженности иногда используется и в литературе. Рассмотренное представление, в силу своей универсальности, применяется также и для представления таблиц r x c в ряде стандартных программ анализа данных. Поэтому информация дана как для полноты, так и для указания пользователю пути сравнения возможностей различных программ.

Для анализа многовходовых таблиц сопряженности применяются специальные модификации стандартных критериев, указанные в главе «Кросстабуляция».

1.3.5. Проблема пропущенных данных

Понятие пропусков и анализ причин их появления приводятся в главе «Обработка пропущенных данных». Напомним, что такое цензурирование. В ходе контролируемого процесса (научного исследования, производственного процесса, хода лечения и т. д.) часть контролируемых объектов может не отказать за период наблюдения. Другая часть может отказать, причем моменты отказов точно неизвестны. Это явление носит наименование

цензурирования, а получаемые выборки – цензурированных.
Функции AtteStat, кроме особо оговоренных случаев (см. главу «Анализ выживаемости»), не работают с пропущенными данными и с цензурированными выборками, поэтому пользователь обязан позаботиться о получении пригодного для анализа применяемыми методами диапазона ячеек исходных данных без пропусков до производства расчетов.
Имеется несколько путей решения проблемы:

1. Ячейки, содержащие значения, пропущенные по условиям эксперимента (объект исследования выбыл до окончания эксперимента), могут быть просто исключены.
2. Если данные утрачены по причинам, связанным или не связанным с условиями эксперимента (лаборант забыл сделать отсчет), они могут быть восстановлены с помощью специальных компьютерных программ.

См. главу «Обработка пропущенных данных».

1.3.6. Проблемы малых и больших выборок

Проблемы малых и больших выборок относятся к основным проблемам, возникающим при практическом применении методов анализа данных, причем некоторые авторы обоснованно полагают, что понятие «малости» выборки тесно связано с решаемой задачей.
Можно предложить такую классификацию выборок по численности, исходя из требований представленных в программе критериев:

- очень малые выборки – от 5 до 12,
- малые выборки – от 13 до 40,
- выборки средней численности – от 41 до 100,
- большие выборки – от 101 и выше.

Минимальную численность выборки лимитирует не столько алгоритм вычисления критерия, сколько распределение его статистики. Так, для ряда алгоритмов при слишком малых численностях нормальная аппроксимация распределения статистики критерия будет под вопросом.
Максимальная численность выборки лимитируется повышенной трудоемкостью вычисления статистики критерия, особенно, если в схеме его вычисления применяются комбинаторные алгоритмы. При больших численностях выборок становится оправданным применение менее трудоемких в вычислении тестов, в том числе параметрических.
При стремящейся к бесконечности численности выборки независимых одинаково распределенных случайных величин, согласно центральной предельной теореме, распределение их суммы приближается к нормальному, а среднее арифметическое случайных величин (теорема Маркова) сходится по вероятности к среднему арифметическому их математических ожиданий. Данные и другие параметры как раз являются основой схем вычисления различных параметрических тестов.
Итак, с большими выборками хорошо справляются параметрические методы, например, из серии методов, представленных в главк «Параметрическая статистика».
Отметим, что большая численность выборки вовсе не означает абсолютной гарантии верного применения параметрических тестов, как ошибочно полагают некоторые исследователи. Проверку нормальности распределения с помощью методов главы «Проверка нормальности распределения» провести рекомендуется в любом случае.
Непараметрические методы могут анализировать любые, в том числе большие и малые выборки, однако предел «малости» конкретного метода обычно ограничен численностью выборки, указанной в описании теста. Меньшие выборки представленными методами анализировать не рекомендуется.

1.3.7. Общая методология

Согласно энциклопедии «Вероятность и математическая статистика», математической статистикой называют «раздел математики, посвященный математическим методам сбора, систематизации, обработки и интерпретации статистических данных, а также использование их для научных или практических выводов». Под математической статистикой также обычно понимают прикладное, практическое приложение достижений теории вероятностей. Официальное определение термина «прикладная статистика» отсутствует в словарях и энциклопедиях. Термин, по сути, означает «прикладную математическую статистику». Так как математическая статистика – это уже «прикладная теория вероятности», сам термин «прикладная статистика» научного смысла не имеет, и может использоваться сугубо в «бытовом» смысле.
Теория вероятности носит всеобщий характер безотносительно к физической природе явления. Поэтому методы математической статистики одинаковы для изучения любой объективной реальности, живой и неживой природы, научных и технических объектов. Однако исторически сложились некоторые специфические области конкретных приложений методов математической статистики:

1. Биометрия (biometry), биометрика (biometrics) – раздел биологии, основная задача которого – планирование количественных биологических экспериментов и обработка результатов методами математической статистики. Данное определение показывает, что биометрия – это просто приложение методов математической статистики к биологии. Иначе, биометрия – это совокупность приемов планирования и обработки данных биологического исследования методами математической статистики. Термин «биометрия» обычно применяют к биологическим и агрокультурным приложениям. Для медицинских приложений применяют термин «биостатистика» (biostatistics), поэтому использование терминов «биометрия» и «биометрика» для медицинских приложений является нонсенсом. Термин «биометрика» считается синонимом «биометрии», однако в последнее время под биометрикой в зарубежных публикациях понимают персональную идентификацию людей по биометрическим показателям, что не мешает некоторым авторам применять для той же цели термин «биометрия».
2. Эконометрия (econometry), эконометрика (econometrics) – наука, изучающая конкретные количественные взаимосвязи экономических объектов и процессов с помощью математических и статистических методов и моделей. Данное определение показывает, что эконометрия – понятие более широкое, чем просто приложение методов математической статистики к экономике. Иначе, эконометрика – экономические измерения, наука о применении статистических и математических методов в экономическом анализе для проверки правильности экономических теоретических моделей и способов решения экономических проблем.
3. Логично было бы ввести термины «технометрия» и/или «технометрика», относящиеся к приложениям методов математической статистики к физико–химическим и инженерным наукам. Наименование Technometrics, однако, зарезервировано за одноименным иностранным журналом, оперирующим в названных областях.
4. Логично было бы ввести термины «психометрия» и/или «психометрика», относящиеся к приложениям методов математической статистики к психиатрии, психологии, психофизиологии. Наименование Psychometrica, однако, зарезервировано за одноименным иностранным журналом, оперирующим в названных областях.
5. Геостатистика – это математическая теория разведки месторождений и оценки их характеристик.

Приведенные примеры доказывают, что в толкованиях, взятых из энциклопедических словарей, явно просматривается неумение (или сознательная позиция) авторов данных

статей отделить содержательную специфическую часть научной проблемы от всеобщей расчетной части: биометрию от биологии, эконометрию от экономики и т.п. «Расчлените каждую изучаемую вами задачу на столько частей …, сколько потребуется, чтобы их было легко решить» (Р. Декарт).
Некоторые пользователи программы анализа данных (обычно, это – ученые) сообщают, что у них на руках имеется некоторое количество экспериментальных данных, которые нужно обработать. Они полагают, что этапом научных изысканий, следующим за сбором данных, должна быть статистическая обработка этих самых данных, а проблема заключается в выборе метода математико–статистической обработки. На самом деле первой, главной и единственной проблемой таких ученых является неудачное планирование научной работы (включая планирование финансовых затрат на информационное и лицензионное программное обеспечение). Хотя «наука, как и добродетель, сама себе награда» (Ч. Кингсли), подобное отношение к планированию научных исследований неприемлемо.
Если пользователей математико–статистических алгоритмов и их программных воплощений интересует качество исследований, следует до производства какого–либо исследования проделать следующие шаги:

1. Изучить философские основания методологии научного исследования, начиная с трудов Поппера и Лакатоса. Как обзорное введение можно использовать популярную статью Баюка.
2. Сформировать четкое понимание о шкалах измерения. Именно через шкалы измерения исходные данные диктуют, какие методы могут быть использованы для их обработки. Перед применением каждого метода следует ознакомиться с его предпосылками и ограничениями и спланировать потребный объем исследований исходя из мощности критериев.
3. Приступить к сбору данных. Здесь уже предполагаемый метод обработки укажет, в какой форме должны быть представлены экспериментальные данные, пригодные для адекватного применения предполагаемого метода.
4. Математико–статистическая обработка – это предпоследний, технический, этап, содержание которого должно быть полностью понятно после реализации 2–го этапа, когда еще не было больших затрат на экспериментальные исследования. Данный этап не имеет никакого отношения к предметной области. Математическая статистика, как уже было сказано в начале предыдущего раздела, не интересуется природой исходных данных (природой данных и физикой явлений интересуется математическое моделирование, см. главу «Обыкновенные дифференциальные уравнения»).
5. Последний этап – предметные научно обоснованные выводы по результатам исследования, рекомендации и прогноз.

1.3.7.1. Статистическая популяция

Областью исследований прикладного статистического анализа является статистическая популяция (генеральная совокупность), о параметрах которой делается предположение на основании репрезентативной эмпирической выборки (выборочной совокупности) из популяции.
Статистической популяцией называется совокупность всех объектов одного класса, различия между которыми определяются только случайными факторами. Рассматривая популяцию с точки зрения различий между объектами, мы неизбежно вынуждены решать, по каким именно параметрам различаются объекты, составляющие популяцию (т. е., что именно мы исследуем).
Статистическая популяция определяется по тому параметру, который нас интересует. Соответственно, в одном случае (например, исследование распределения по росту и массе

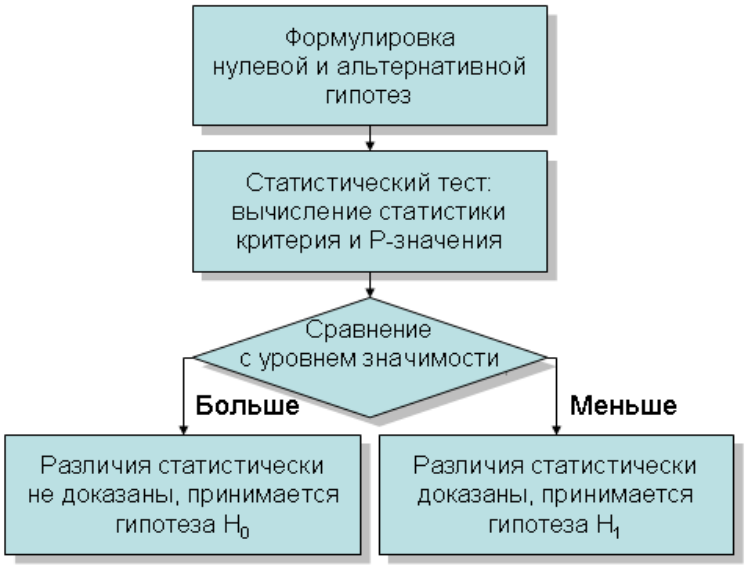
тела) у нас будет популяция людей, а в другом (исследование распределения половых признаков) мужчин и женщин.
Чтобы показать, что статистическая популяция не тождественна популяции в биологическом, социальном или ином предметном смысле, рассмотрим курьезный пример. К примеру, есть две группы военнослужащих, которые мы решили сравнить по росту. В каждой группе рост (в сантиметрах) – нормально распределенная величина. И, действительно, можно предположить (нулевая гипотеза), что обе группы взяты из одной популяции, если не учитывать, что первая группа – это подразделение народно–освободительной армии, а вторая группа – терракотовая армия императора Цинь Ши Хуанди.

1.3.7.2. Статистическая гипотеза

Статистической гипотезой H_0 называется утверждение, в котором предполагается, что истинное распределение вероятностей, описывающее изменчивость, принадлежит подмножеству семейства возможных вероятностных распределений. Проверяемая гипотеза H_0 называется нулевой гипотезой. Альтернативной (конкурирующей) гипотезой H_1 обычно называется отрицание нулевой гипотезы, однако могут быть иные варианты.
Пусть, например, статистический критерий проверяет нулевую гипотезу H_0 о равенстве («нет статистически значимого различия») функций распределения двух выборочных совокупностей $F(x) = G(x)$ Альтернативная гипотеза H_1 в данном случае может быть сформулирована одним из трех способов:

1. $F(x) \neq G(x)$– «нулевая гипотеза неверна» – это двусторонняя (two–tailed, two–sided) гипотеза;
2. $F(x) < G(x)$ – это односторонняя (upper–tailed) гипотеза;
3. $F(x) > G(x)$ – это односторонняя (low–tailed) гипотеза.

Критерий T проверки статистической гипотезы H_0 есть процедура выработки решения о том, принять или отклонить данную нулевую гипотезу. Критической областью (областью непринятия нулевой гипотезы) U является та часть выборочного пространства, которая приводит к отклонению гипотезы H_0.
Уровнем значимости критерия является вероятность α того, что этот критерий приведет к отклонению нулевой гипотезы в случае ее истинности: $P(T \in U) = \alpha$. Если результаты проверки находятся в критической области $P(T > T_\alpha) < \alpha$, нулевая гипотеза отклоняется и принимается альтернативная гипотеза. Здесь критическому значению критерия соответствует уровень значимости α.
Отклонение нулевой гипотезы в случае ее истинности называется ошибкой I (первого) рода. Принятие нулевой гипотезы, когда она не верна, называется ошибкой II (второго) рода. Вероятность ошибки второго рода обозначается β. Величину 1 – β называют мощностью статистического критерия.
С целью унификации статистических таблиц и стандартизации выводов уровень значимости выбирается из стандартной линейки типа 0,001; 0,005; 0,01; 0,05 …, либо то же в процентах. Величина уровня значимости зависит от важности предметной области (см. раздел о доверительной вероятности). Чем проводятся исследования более важные (в биомедицине и смежных дисциплинах – более социально значимые), тем меньшим уровнем значимости следует оперировать.
На схеме показан алгоритм действий при практическом решении задачи проверки гипотезы. Пусть нулевая гипотеза H_0 сформулирована как «нет статистически значимого различия», а альтернативная гипотеза H_1 сформулирована как «нулевая гипотеза неверна».



Результатом статистической проверки является вывод о том, в скольких случаях, например, на каждые 100 проведенных испытаний отклонения можно считать случайными. Таким образом, на заданном стандартном уровне значимости исследователь может остановиться на одной из двух гипотез.

1.3.7.2.1. Односторонние и двусторонние гипотезы

Рассмотрим понятия односторонней (upper–tailed и low–tailed) и двусторонней (two–tailed) гипотез, которым соответствуют односторонний (one–sided) и двусторонний (two–sided) критерии значимости.
Считается, что когда исследователь имеет достаточное количество данных, позволяющих предсказать в альтернативной гипотезе направление различий (например, доля желательных эффектов в опытной группе не просто отличается от доли в контрольной группе, а превышает ее), используется односторонний критерий. В противном случае (доля эффектов в опытной группе просто отличается от доли в контрольной группе) используется двухсторонний критерий. Даже если интересующее различие должно быть в одностороннем направлении, исследователю рекомендуется подстраховаться от неожиданных результатов, выполнив двусторонний тест.
Порядок действий при решении о принятии гипотезы такой.

1. Нулевая гипотеза H_0 (двусторонняя альтернатива) отклоняется, если $p_2 < \alpha$.
2. Нулевая гипотеза H_0 (односторонняя upper–tailed альтернатива) отклоняется, если $(1 - p_U) < \alpha$.
3. Нулевая гипотеза H_0 (односторонняя low–tailed альтернатива) отклоняется, если $p_L < \alpha$.

Здесь обозначено:
p_2 – достигнутый уровень значимости двусторонней статистической гипотезы,
p_U и p_L – достигнутый уровень значимости соответствующей односторонней статистической гипотезы.
При выполнении перечисленных условий соответствующая альтернативная гипотеза H_1 может быть принята.
Если оперировать значением статистики критерия, нулевая гипотеза может быть принята при нахождении вычисленного значения статистики критерия T в области:

- $T_{1-\alpha} < T \le T_{\alpha}$ для двусторонней альтернативы,
- $T_{1-\alpha/2} < T \le T_{\alpha/2}$ для односторонней альтернативы.

Обсуждение см. в монографиях Тюрина с соавт., Селезнева с соавт., Теннанта–Смита, Брандта, Клюшина с соавт., Мостеллера (Mosteller) с соавт., учебном пособии Тутубалина, книге Глотова с соавт., в статьях Гудмана.

1.3.7.3. P–значение

При подстановке статистики в ее функцию распределения получается величина, имеющая смысл вероятности и интерпретацию, зависящую от решаемой проблемы. Эта вероятность называется фактически достигнутым уровнем значимости, иначе *P*–значением.
Различные виды *P*–значений:

1. *P*–значение статистики критерия, полученное в результате подстановки статистики критерия в его функцию распределения. Данное *P*–значение не дает возможности сделать вывод о значимости статистической гипотезы в силу того, что оно ни к какой статистической гипотезе не относится.
2. *P*–значение статистической гипотезы. Данное *P*–значение дает возможность сделать вывод о значимости альтернативной статистической гипотезы. Поэтому рассматриваемый показатель может быть рассчитан только после формулировки альтернативной гипотезы.

P–значение дает возможность принимать или отклонять данную гипотезу при любом заранее заданном уровне значимости α путем простого сравнения вычисленного *P*–значения с принятым стандартным уровнем значимости. Поэтому возможен иной подход к проверке статистической гипотезы. А именно, сначала вычисляется по выборке статистика *T*. Затем вычисляется вероятность *P* попадания *T* в критическую область.
Рассмотрим, как нужно делать выводы относительно *P*–значения статистической гипотезы на основе вычисленного *P*–значения статистики критерия в стандартных случаях статистической гипотезы. Итак, пусть вычислено *P*–значение статистики критерия *p* путем подстановки статистики критерия в его функцию распределения. Тогда:

1. В случае двусторонней статистической гипотезы ее *P*–значение (говорят проще – двустороннее *P*–значение) вычисляется как $p_2 = 2 \cdot \min(p, 1 - p)$.
2. Если схема вычисления статистического критерия позволяет сразу вычислить два *P*–значения стандартных односторонних статистических гипотез (говорят проще – одностороннее *P*–значение): p_U (верхний хвост, upper–tailed) и p_L (нижний хвост, low–tailed), то двустороннее *P*–значение равно $p_2 = p_U + p_L$. Если распределение статистки критерия несимметрично, то $p_U \neq p_L$. При этом обычно приводится одностороннее *P*–значение, вычисляемое как $p_1 = \min(p_U, p_L)$.
3. Если распределение статистки критерия симметрично, то $p_U = p_L$ и $p_2 = 2 \cdot p_U = 2 \cdot p_L$. Поэтому, если вычислено двустороннее *P*–значение, а распределение статистики критерия симметричное, одностороннее *P*–значение можно получить из двустороннего *P*–значения по формуле $p_1 = p_2 / 2$.

Рассмотрим пример. Проверяется нулевая гипотеза о равенстве средних значений двух выборок, а также сформулирована двусторонняя альтернатива о том, что средние значения не равны. Зададимся уровнем значимости $\alpha = 0{,}05$. Пусть на основе статистики критерия вычислен достигнутый уровень значимости $p = 0{,}988095$. Тогда двустороннее *P*–значение равно $p_2 = 2 \cdot \min(p, 1 - p) = 2 \cdot \min(0{,}988095; 0{,}011905) = 0{,}023810$. Очевидно, что $p_2 < \alpha$, поэтому нулевая гипотеза отклоняется и принимается альтернативная гипотеза о статистически значимом различии средних значений на уровне значимости $\alpha = 0{,}05$. Данный факт записывают как $p < 0{,}05$.

Обсуждение см. в монографиях Петровича с соавт., Боровкова, Браунли. О калибровке *P*–

значений см. работы Селлке (Sellke) с соавт., Байарри (Bayarri) с соавт.

1.3.7.4. Доверительная вероятность

Доверительная вероятность (доверительный уровень, коэффициент доверия) определяется формулой
$P = 1 – \alpha$,
где α – уровень значимости.
Доверительная вероятность требуется для вычисления ряда выборочных статистических показателей, и в отличие от ряда других параметров является не вычисляемой по выборке, а задаваемой пользователем программы величиной. Она выбирается из следующей стандартной линейки (в основном, следуя классификации Плохинского):

- Нулевой порог 0,90 применяется для работы с пониженной ответственностью, при первом ознакомлении с явлением.
- Первый порог 0,95 применяется в большинстве исследований (например, биологические исследования).
- Второй порог 0,99 применяется для работ с повышенной ответственностью (например, медицинские исследования).
- Третий порог 0,999 применяется для работ с высокой ответственностью (например, исследования эффективности лекарств).

Доверительный уровень может быть выражен в долях, например, 0,95, либо в процентах, то же самое, 95%.

1.3.7.5. Мощность критерия

Мощностью называют величину $1 – \beta$, где β – вероятность ошибки второго рода статистической гипотезы. Мощность характеризует качество статистического критерия. Мощность – это не число, а функция. Чем эффективнее данная функция стремится к 1, тем более эффективен статистический критерий. От чего зависит мощность критерия?

1. Для критериев согласия (см. главу «Проверка нормальности распределения») функция мощности зависит от выбора конкретного альтернативного распределения. Знание вида функций мощности различных критериев в зависимости от свойств конкретного параметрического семейства важно на том основании, что авторами рекомендуется выбирать более мощный статистический критерий для анализа. Сравнение различных критериев согласия по мощности считается задачей типичной, но контрпродуктивной, т.к. всегда можно указать альтернативу, при которой мощность именно данного теста является наибольшей.
2. В любом случае мощность критерия тем выше (функция зависимости мощности от численности ближе к единице), чем выше численность анализируемой выборки. Данная зависимость позволяет определить необходимую численность выборки, чтобы при исследовании гарантировать заданную мощность (для медицинских приложений достаточной считается мощность не менее 0,80 или 80%). Подробнее о зависимости мощности от численности (и наоборот) см. главу «Описательная статистика».
3. Иногда от других параметров схемы вычисления критерия.

При численном исследовании мощности основным моментом часто является даже не количество повторений численного эксперимента, хотя этот параметр очень важен и должен быть максимально большим (порядка нескольких тысяч или десятков тысяч), а способ получения качественной последовательности псевдослучайных чисел с заданным законом распределения.

Численное исследование мощности методом Монте–Карло представлено в работах

Золотухиной с соавт., Селезнева с соавт., Хассана (Hassan), серии работ Лемешко с соавт., монографии Хана с соавт. О разработке и тестировании генераторов псевдослучайных чисел см. классические статьи Лекюйе (L'Ecuyer), Марсалья (Marsaglia), а также программное обеспечение и библиографию, указанную данными авторами. Подробный обзор методов и список источников представлены в главе «Рандомизация и генерация случайных последовательностей».

1.3.7.6. Сопряженность выборок

Данные, полученные в реальных экспериментах, могут быть представлены независимыми либо сопряженными (связанными) выборками. Соответственно, к этим выборкам применимы критерии значимости для независимых выборок либо для сопряженных выборок.

1.3.7.6.1. Независимые выборки

Независимыми будут выборки, отобранные из причинно независимых совокупностей. При этом обычно не имеет значения, равны между собой или не равны численности совокупностей.
Критерии для независимых выборок применяются, чтобы выявить статистическую значимость различий двух различных групп индивидуумов. Примерами независимых выборок могут служить:

- параметры двух групп пациентов, к которым применялись различные методики лечения с целью изучения значимости различий между методиками;
- частный случай предыдущей схемы: параметры двух групп пациентов, к одной из которых (опытная группа) применялось воздействие методики, а к другой (контрольной) не применялось, с целью изучения значимости влияния данной методики на результат лечения; данная схема называется «опыт – контроль»;
- частный случай предыдущей схемы: параметры группы пациентов, к которой применяется некоторое лекарственное средство, и контрольной группы пациентов, к которой применятся плацебо, а исследование производится с целью проверки эффективности препарата.

Напомним, что случайное распределение всей совокупности пациентов на группы называется рандомизацией и может быть выполнено с помощью методов главы «Рандомизация и генерация случайных последовательностей».

1.3.7.6.2. Сопряженные выборки

Критерии, применяемые к выборкам с попарно сопряженными вариантами, называются парными критериями либо критериями для связанных или сопряженных выборок. При анализе сопряженных выборок численности сравниваемых совокупностей всегда равны между собой. Примеры сопряженных выборок:

- параметры одной и той же испытуемой группы до и после воздействия какого–либо фактора, например, методики лечения; данная схема называется «до и после»;
- параметры одной и той же группы индивидуумов (например, список политических партий, участвующих в парламентских выборах) при воздействии на нее различных факторов (предпочтения электората в различных избирательных округах);
- параметры одного и того же объекта экспериментального исследования, но относящиеся к различным его частям, например состояния двух конечностей в процессе лечения, одна из которых подвергается лечебному воздействию, а другая нет.

1.3.8. Статистические распределения

В разделе представлены некоторые применяемые стандартные статистические распределения и комментарии относительно их вычислений.
В программе применяются следующие стандартные распределения:

1. биномиальное распределение,
2. гипергеометрическое распределение,
3. нормальное распределение и обратное к нему,
4. многомерное нормальное распределение,
5. *t*–распределение Стьюдента и обратное к нему,
6. *F*–распределение и обратное к нему,
7. бета–распределение,
8. распределение χ^2 и обратное к нему,
9. нецентральное распределение χ^2 и обратное к нему,
10. обобщенное гамма–распределение,
11. логнормальное распределение,
12. распределение S_U Джонсона,
13. распределение выборочного размаха,
14. распределение стьюдентизированного размаха,
15. распределение стьюдентизированного максимума модулей,
16. распределение статистики критерия Колмогорова,
17. распределение статистики критерия Койпера,
18. распределения статистик критериев Вилкоксона,
19. распределение статистики критерия Манна–Уитни,
20. распределение статистики a_1 критериев типа омега–квадрат.

Для удобства восприятия (и программной реализации) мы не будем придерживаться традиции обозначать дополнительные параметры распределений, там, где это необходимо, греческими символами, формально заменив их латиницей. Кроме того, мы часто для экономии записи используем обозначение типа «Литера(.)», в котором точка означает подстановку любого допустимого выражения для введенного математического объекта «Литера». Для обозначения некоторых функций применяются сходные литеры, поэтому необходимо быть внимательным к контексту и расшифровывать все обозначения в формулах, хотя бы даже стандартные или общепринятые.

Основные источники: Большев с соавт., Брандт, Гайдышев (2001), Де Гроот, Попов с соавт., Родионов, Родионов с соавт., Хан с соавт., Хастингс с соавт., Шор с соавт., Бьюри (Bury), Эванс (Evans) с соавт. Сводку распределений и аппроксимаций дал Кобзарь.

1.3.8.1. Биномиальное распределение

Функция биномиального распределения вычисляется по формуле

$$P(k < K) = \sum_{k=0}^{K-1} W_k^n,$$

где n – число степеней свободы, $n > 0$,

W_k^n – вероятности биномиального распределения, вычисляемые по формуле

$$W_k^n = C_n^k p^k (1-p)^{n-k},$$

где C_n^k – число сочетаний из n по k.

Для обеспечения численной устойчивости алгоритма число сочетаний может вычисляться

как (Попов с соавт., Брандт)

$$C_n^k = \frac{\Gamma(n+1)}{\Gamma(k+1)\Gamma(n-k+1)},$$

где Γ(.) – гамма–функция.

1.3.8.2. Гипергеометрическое распределение

Функция гипергеометрического распределения вычисляется по формуле

$$P(k < k') = \sum_{k=0}^{k'-1} W_k,$$

где W_k – вероятности гипергеометрического распределения, вычисляемые по формуле

$$W_k = \frac{C_K^k C_{N-K}^{n-k}}{C_N^n}, n \le N, k \le K,$$

где C_K^k – число сочетаний из K по k,

C_{N-K}^{n-k} – число сочетаний из $N - K$ по $n - k$,

C_N^n – число сочетаний из N по n,

K, N, n – параметры распределения.

1.3.8.3. Нормальное распределение

Нормальным называется одно из важнейших распределений вероятностей случайной величины. Теоретическое обоснование роли нормального распределения дается центральными предельными теоремами, рассматриваемыми в курсе «Теории вероятностей». Функция плотности нормального распределения имеет вид:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-a)^2}{2\sigma^2}},$$

$-\infty < a < \infty$, $\sigma > 0$, $-\infty < x < \infty$.

Путем введения нормированной величины

$$t = \frac{x-a}{\sigma},$$

где a – математическое ожидание (обычно его оценка – среднее значение, но могут применяться и другие параметры положения),

σ^2 – дисперсия (параметр разброса),

показанной выше формуле придан несколько иной вид. Этой формулой удобно пользоваться при расчете теоретических частот эмпирического распределения. К тому же таблицы обычно даются для функции, называемой также плотностью вероятности стандартизованной (стандартной) нормальной случайной величины,

$$f(t) = \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}.$$

Функция стандартного нормального распределения равна

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-y^2/2} dy$$

и называется функцией Лапласа (вторым законом распределения Лапласа) либо интегралом вероятности Гаусса (законом Гаусса, гауссовым распределением) в честь применения данного закона распределения для изучения ошибок наблюдений.

Практически вычисление функции стандартного нормального распределения производится по формуле

$$\Phi(x) = \frac{1}{2}\left[1 + sign(x) P_{x^2/2}(1/2)\right],$$

где $P_{\cdot}(.)$ – неполная гамма–функция.
Находит применение интеграл вероятностей

$$I(x) = \frac{1}{\sqrt{2\pi}} \int_{-x}^{x} e^{-y^2/2} dy.$$

С использованием свойства симметрии подынтегральной функции стандартного нормального распределения $\Phi(.)$ расчетная формула интеграла вероятностей $I(.)$ сводится к простому выражению
$I(x) = 2\Phi(|x|) - 1)$.
В литературе нормальное распределение кратко обозначают как $N(a,\sigma^2)$. Стандартное нормальное распределение обозначается как $N(0,1)$. Обратная к $\Phi(.)$ функция иногда называется пробитом и может обозначаться как $\Psi(.)$, $\Phi^{-1}(.)$, *probit*(.).

1.3.8.4. Многомерное нормальное распределение

В случае многомерного нормального распределения плотность распределения совокупности определяется формулой

$$P(X) = \frac{1}{(2\pi)^{d/2}|S|^{1/2}} e^{-\frac{1}{2}(X-\overline{X})' S^{-1} (X-\overline{X})},$$

где S – дисперсионно–ковариационная матрица,
$\overline{X}$ – вектор математического ожидания,
d – «число измерений» – порядок матрицы S и длина вектора $\overline{X}$,
$'$ – операция транспонирования.
Дисперсионно–ковариационная матрица в случае многомерного распределения является параметром, аналогичным дисперсии в одномерном случае. На диагонали данной матрицы располагаются дисперсии компонент случайного вектора. Внедиагональные члены матрицы являются ковариациями.
Иногда нормальное многомерное распределение ошибочно понимается в том смысле, что каждая переменная, составляющая многомерную совокупность (реализацию случайного многомерного вектора), имеет нормальное распределение. Это неверно: исследуя такое распределение «одномерных составляющих», анализируют только маргинальные распределения компонент случайного многомерного вектора, составляющих многомерное распределение, но не само многомерное распределение. Для исследования нормальности многомерного распределения разработаны специальные методы.

Многомерное нормальное распределение в литературе кратко обозначают как $N(\overline{X}, S)$.

О многомерном нормальном распределении см. статью Мартынова.

1.3.8.5. t–распределение

Функция t–распределения Стьюдента выражается формулой

$$F_n(x) = \frac{1}{\sqrt{n} B(1/2, n/2)} \int_{-\infty}^{x} \left(1 + \frac{y^2}{n}\right)^{-(n+1)/2} dy,$$

где n – число степеней свободы, $n > 0$,
$B(.)$ – бета–функция.
Практически вычисление функции производится по формуле

$$F_n(x) = \frac{1}{2}\left\{1 + sign(x)\left[1 - I_{n/(n+x^2)}(n/2, 1/2)\right]\right\},$$

где $I_{.}(.,.)$ – регуляризованная неполная бета–функция.

1.3.8.6. F–распределение

Функция F–распределения выражается формулой

$$F_x(n_1, n_2) = \frac{\Gamma((n_1+n_2)/2)}{\Gamma(n_1/2)\Gamma(n_2/2)}\left(\frac{n_1}{n_2}\right)^{n_1/2}\int_0^x y^{n_1-1}\left(1+\frac{n_1}{n_2}y\right)^{-(n_1+n_2)/2} dy,$$

где n_1 – число степеней свободы, $n_1 > 0$,
n_2 – число степеней свободы, $n_2 > 0$,
$\Gamma(.)$ – гамма–функция.
Практически вычисление функции производится по формуле

$$F_x(n_1, n_2) = 1 - I_{n_2/(n_2+n_1 x)}(n_2/2, n_1/2),$$

где $I_{.}(.,.)$ – регуляризованная неполная бета–функция.

1.3.8.7. Бета–распределение

Функция бета–распределения – эквивалентное наименование регуляризованной неполной бета–функции. Подробнее см. в разделе «Специальные функции».

1.3.8.8. Хи–квадрат распределение

Функция распределения χ^2 выражается формулой

$$F_n(x) = \frac{1}{2^{n/2}\Gamma(n/2)}\int_0^x y^{n/2-1}e^{-y/2}dy,$$

где n – число степеней свободы, $n > 0$,
$\Gamma(.)$ – гамма–функция.
Практически вычисление функции производится по формуле
$F_n(x) = 1 - P_{x/2}(n/2)$,
где $P_{.}(.)$ – неполная гамма–функция.

1.3.8.9. Нецентральное хи–квадрат распределение

Функция нецентрального распределения χ^2 выражается формулой

$$F_n'(x, \lambda) = e^{-\lambda/2}\sum_{k=1}^{\infty}\frac{\lambda^k}{k!2^{n/2+2k}\Gamma(n/2+k)}\int_u^{\infty} y^{n/2+k-1}e^{-y/2}dy,$$

где n – число степеней свободы, $n > 0$,
λ – параметр нецентральности, $\lambda \geq 0$,
u – обратная функция распределения χ^2,
$\Gamma(.)$ – гамма–функция.
При $\lambda = 0$ нецентральное распределение χ^2 совпадает с распределением χ^2.
Практически вычисление функции производится посредством аппроксимации, предложенной Пирсоном,

$$F_n'(x,a) = F_{n'}(x)\frac{n+3\lambda}{n+2\lambda} - \frac{\lambda^2}{n+3\lambda},$$

где $F_{n'}(x)$ – функция распределения χ^2 с числом степеней свободы, равным

$$n' = \frac{(n+2\lambda)^3}{(n+3\lambda)^2}.$$

Свойства, аппроксимации и приложения распределения изучены Большевым с соавт., Оуэном, Кобзарем, Кульбаком. Один из частных случаев рассмотрен Фишером (Fisher).

1.3.8.10. Обобщенное гамма–распределение

Функция гамма–распределения может иметь один, два или три параметра. Гамма–функция с тремя параметрами, называемая обобщенной гамма–функцией, вычисляется по формуле

$$F_x(a,b,c) = \frac{1}{b^a\Gamma(a)}\int_0^x (t-c)^{a-1}e^{-(t-c)/b}dt.$$

Практически вычисление функции производится по формуле
$F_x(a,b,c) = P_{(x-c)/b}(a)$,
где $P_.(.)$ – неполная гамма–функция.

1.3.8.11. Логнормальное распределение

Функция логнормального (логарифмически нормального) распределения с двумя параметрами вычисляется по формуле

$$P_x(a,b) = \frac{1}{b\sqrt{2\pi}}\int_0^x y^{-1}e^{-(\ln y-a)^2/2b^2}dy.$$

Заменой переменной $\ln y = t$, $y^{-1}dy = dt$ и, соответственно, меняя пределы интегрирования $y \in [0;x]$ на $t \in]-\infty;\ln x]$, получаем формулу, которая пригодится в дальнейших выкладках,

$$P_x(a,b) = \frac{1}{b\sqrt{2\pi}}\int_{-\infty}^{\ln x} e^{-(t-a)^2/2b^2}dt.$$

Функция нормального распределения от нестандартизованной случайной величины равна
$F_x(a,b) = \Phi((x-a)/b)$,
где $\Phi(.)$ – функция стандартного нормального распределения.
Таким образом, расчетная формула путем преобразований примет вид
$P_x(a,b) = F_{\ln x}(a,b) = \Phi((\ln x - a)/b)$.
Рассмотренное распределение является частным случаем логнормального распределения с тремя параметрами, называемого также распределением S_L Джонсона.
Плотность логнормального распределения с двумя параметрами вычисляется по формуле

$$f(x,a,b) = \frac{1}{xb\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{\ln(x)-a}{b}\right)^2}, x>0, b>0, -\infty<a<\infty.$$

1.3.8.12. Распределение S_U Джонсона

Функция распределения S_U Джонсона вычисляется по формуле

$$P_x(a,b,c,d) = \frac{b}{d\sqrt{2\pi}}\int_{-\infty}^{x}\frac{1}{\sqrt{\left(((y-c)/d)^2+1\right)}}e^{-\frac{1}{2}\left\{a+b\ln\left[(y-c)/d+\sqrt{((y-c)/d)^2+1}\right]\right\}^2}dy$$

.

Действуя аналогично предыдущему случаю и руководствуясь материалами монографии Хана с соавт. (с. 233) по распределениям Джонсона, устанавливаем, что расчетная формула будет иметь вид
$P_x(a,b,c,d) = 1 - \Phi(a + b \cdot Arsh(x - c) / d))$,
где $\Phi(.)$ – функция стандартного нормального распределения,
$Arsh(.)$– функция гиперболического арксинуса.

1.3.8.13. Распределение выборочного размаха

Функция распределения $P_n(W \le w)$ выборочного размаха (range) W для выборки численности n, иначе вероятность того, что он не превысит w, определяется формулой

$$P_n(W \le w) = n \int_{-\infty}^{\infty} [F(x+w) - F(x)]^{n-1} dF(x).$$

Если совокупность распределена нормально, то выражение $F(.)$, входящее в формулу, представляет собой функцию распределения нормально распределенной стандартизованной случайной величины

$$P(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2} dt.$$

Однако для практического вычисления функции распределения размаха дополнительно необходимо выразить величину $dF(x)$, входящую в формулу ее вычисления, через dx. Можно записать

$$dF(x) = \frac{dF(x)}{dx} dx,$$

где $\frac{dF(x)}{dx}$ – производная $F(x)$ по x – плотность распределения вероятности – для стандартизованной нормальной случайной величины вычисляется по формуле

$$\frac{dP(x)}{dx} = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}.$$

Сделав все необходимые подстановки, получаем пригодную для практических вычислений формулу

$$P_n(W \le w) = \frac{n}{\sqrt{2\pi}} \int_{-\infty}^{\infty} [P(x+w) - P(x)]^{n-1} e^{-x^2/2} dx.$$

Вычисление рассматриваемого распределения производится численным интегрированием методом Симпсона. Метод см. у Лоренсель (Laurencelle) с соавт. См. источники: Мюллер с соавт., Хальд, Оуэн, Барндорф–Нильсен с соавт., Математический энциклопедический словарь, энциклопедия «Вероятность и математическая статистика», Гайдышев (2001).

1.3.8.14. Распределение стьюдентизированного размаха

Пусть из нормальной совокупности извлекается выборка численностью n и по данной выборке вычисляется выборочный размах W. Затем из той же нормальной совокупности или из другой нормальной совокупности с тем же стандартным отклонением извлекается выборка численностью f и по данной выборке вычисляется выборочное стандартное отклонение s.
Тогда отношение W / s называется стьюдентизированным размахом (размахом Стьюдента, studentized range) и его распределение зависит только от величин n и f. Функция

распределения $P_{n,f}(W / s \le q)$ выборочного стьюдентизированного размаха, иначе вероятность того, что он не превысит q, определяется формулой

$$P_{n,f}(W/s \le q) = \frac{f^{f/2}}{\Gamma(f/2)\cdot 2^{f/2-1}} \int_0^\infty x^{f-1} e^{-fx^2/2} P_n(qx)dx,$$

где $P_n(.)$ – функция распределения выборочного размаха.

Вычисление рассматриваемого распределения производится численным интегрированием методом Симпсона. Метод см. у Лоренсель (Laurencelle) с соавт. См. также источники: Оуэн, Ликеш с соавт., Мюллер с соавт., Хальд, Гайдышев (2001), Дэйвид, Дэйвид (David) с соавт., Хартер (Harter) с соавт., Шеффе. Аппроксимации рассмотрены Копенгауэр (Copenhaver) с соавт., Глизон (Gleason), Рамсей (Ramsey) с соавт., Карри (Currie), Пирсон (Pearson), Типпетт (Tippett). Методику вычислений см. также в статье Копенховер (Copenhaver) с соавт., Баум (Baum) с соавт.

1.3.8.15. Распределение стьюдентизированного максимума модулей

Функция распределения $P(Q_{k,n} \le q)$ стьюдентизированного максимума модулей (studentized maximum modulus) $Q_{k,n}$ с параметром k и числом степеней свободы n, иначе вероятность того, что он не превысит q, определяется формулой

$$P(Q_{k,n} \le q) = \int_0^\infty [2\Phi(qx) - 1]^k \, d\mu_n(x),$$

где $\Phi(.)$ – функция стандартного нормального распределения,

$\mu_.(.)$ – деленная на $\sqrt{n}$ плотность функции χ-распределения.
Плотность χ-распределения имеет вид

$$f(x,n) = \frac{x^{n-1}e^{-x^2/2}}{2^{n/2-1}\Gamma(n/2)}.$$

Выполнив необходимые преобразования, по смыслу аналогичные тем, что произведены при вычислении выборочного размаха, получим, что дифференциал $d\mu_n(x)$ определяется формулой

$$d\mu_n(x) = \frac{n^{n/2}x^{n-1}e^{-nx^2/2}}{2^{n/2-1}\Gamma(n/2)}dx.$$

Тогда искомый вид функции распределения

$$P(Q_{k,n} \le q) = \frac{n^{n/2}}{2^{n/2-1}\Gamma(n/2)} \int_0^\infty [2\Phi(qx) - 1]^k \, x^{n-1}e^{-nx^2/2}dx.$$

Вычисление рассматриваемого распределения производится численным интегрированием методом Симпсона. Метод см. у Лоренсель (Laurencelle) с соавт. Способ интегрирования методом разложения в ряд см. в статьях Пиллаи (Pillai) и Пиллаи с соавт. Таблицы и аппроксимацию см. в статье Юри (Ury) с соавт. Как указывают Сахаи (Sahai) с соавт., функция распределения стьюдентизированного максимума модулей может быть получена также как корень квадратный из функции распределения стьюдентизированного максимума хи–квадрат (studentized maximum chi–square) – см. Армитэйдж (Armitage) с соавт. О стьюдентизированном максимуме и минимуме хи–квадрат (studentized minimum chi–square) см. монографию Гупта (Gupta) с соавт., о стьюдентизированном минимуме хи–квадрат см. статью Алан (Alam). См. также Бечхофер (Bechhofer) с соавт., Столайн (Stoline) с соавт.

1.3.8.16. Распределение статистики критерия Колмогорова

Распределение статистики критерия Колмогорова (λ–распределение) вычисляется по точной формуле

$$K(x) = \begin{cases} \sum_{i=-\infty}^{\infty} (-1)^i e^{-2i^2x^2}, x > 0, \\ 0, x \le 0. \end{cases}$$

Бесконечная последовательность быстро сходится, и для получения приемлемой для практических вычислений точности критического значения достаточно небольшого числа ее членов (в программе используется 31 член последовательности, т. е. $-15 \le i \le 15$).

1.3.8.17. Распределение статистики критерия Койпера

Распределение статистики критерия Койпера вычисляется по точной формуле

$$Q(x) = \sum_{i=1}^{\infty} (4i^2x^2 - 1)e^{-2i^2x^2}.$$

Бесконечная последовательность быстро сходится, и для получения приемлемой для практических вычислений точности критического значения достаточно небольшого числа ее членов (в программе используется 15 членов последовательности, т. е. $1 \le i \le 15$).

1.3.8.18. Распределения статистик критериев Вилкоксона

Различают распределения статистик критерия Вилкоксона для независимых выборок и критерия Вилкоксона для связанных выборок.
Для независимых выборок рекуррентные формулы вычисления критических значений критерия Вилкоксона суть
$f(n_1,n_2,W) = f(n_1,n_2 - 1,W - n_1) + f(n_1 - 1,n_2,W)$,
$f(n_1,n_2,-x) = 0$, $x > 0$,
$f(n_1,n_2,0) = 1$,
$f(n_1,0,W) = 0$,
где n_1 – численность одной выборки,
n_2 – численность другой выборки.
P–значение вычисляется как

$$P = \frac{n_2}{n_1 + n_2} f(n_1, n_2 - 1, W - n_1) + \frac{n_1}{n_1 + n_2} f(n_1 - 1, n_2, W).$$

Для связанных выборок рекуррентные формулы вычисления критических значений критерия Вилкоксона суть
$f(N,W^+) = f(N - 1,W^+) + f(N - 1,W^+ - N)$,
$f(N,0) = 1$,
$f(N,-x) = 0$, $x > 0$,
$f(N,W^+) = f(N,N(N + 1) / 2)$, $W^+ \ge N(N + 1) / 2$,
где N – численность каждой выборки.
P–значение вычисляется как

$$P = \frac{f(N, W^+)}{2^N}.$$

См. таблицы Оуэна.

1.3.8.19. Распределение статистики критерия Манна–Уитни

Рекуррентные формулы вычисления критических значений статистики критерия Манна–Уитни суть
$f(n_1,n_2,U) = f(n_1 - 1,n_2,U - n_1) + f(n_1,n_2 - 1,U)$,
$f(n_1,n_2,-x) = 0$, $x > 0$,
$f(n_1,n_2,0) = 1$,
$f(n_1,0,U) = 1$,
$f(n_1,n_2,U) = f(n_2,n_1,U)$,
где n_1 – численность одной выборки,
n_2 – численность другой выборки.
P–значение вычисляется как

$$P = \frac{n_1!n_2!}{(n_1+n_2)!} f(n_1,n_2,U).$$

См. работы Ван де Виля (Van de Wiel) с соавт., Ди Буччианико (Di Bucchianico) с соавт.

1.3.8.20. Распределение статистики критериев типа омега–квадрат

Предельная функция распределения a_1 критериев типа омега–квадрат вычисляется как

$$a_1(x) = \frac{1}{\sqrt{2x}} \sum_{j=0}^{\infty} \frac{\Gamma(j+1/2)}{\Gamma(1/2)\Gamma(j+1)} \sqrt{4j+1} \cdot \exp\left(-\frac{(4j+1)^2}{16x}\right) \cdot \left\{ I_{-1/4}\left(\frac{(4j+1)^2}{16x}\right) - I_{1/4}\left(\frac{(4j+1)^2}{16x}\right) \right\},$$

где $I_{\cdot}(.)$ – модифицированная функция Бесселя.

См. Большева с соавт.

1.3.8.21. Маргинальные распределения

Маргинальным (частным) распределением называют проекцию многомерного распределения на подпространство, порожденное некоторым набором координатных векторов. Пусть $F(x_1,x_2,...,x_n)$ – функция распределения случайного n–мерного вектора $(X_1,X_2,...,X_n)$. Функция распределения $(X_{i_1}, X_{i_2},..., X_{i_m}), 1 \le i_1 < i_2 < ... < i_m \le n, m < n,$ называется маргинальной функцией распределения по отношению к $F(.)$, а соответствующее распределение – маргинальным.

См. энциклопедию «Вероятность и математическая статистика» (с. 299).

1.3.8.22. Специальные функции

Гамма–функция Эйлера определяется формулой

$$\Gamma(a) = \int_0^{\infty} y^{a-1}e^{-y}dy.$$

Неполная гамма–функция (с одним параметром) определяется формулой

$$P_x(a) = \frac{1}{\Gamma(a)} \int_0^{x} t^{a-1}e^{-t}dt.$$

Бета–функция определяется формулой

$$B(a,b)=\int_0^1 y^{a-1}(1-y)^{b-1}dy=\frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}.$$

Неполная бета–функция определяется формулой

$$B_x(a,b)=\int_0^x y^{a-1}(1-y)^{b-1}dy.$$

Регуляризованная неполная бета–функция (иногда для краткости именуемая просто регуляризованной бета–функцией) определяется формулой

$$I_x(a,b)=\frac{B_x(a,b)}{B(a,b)},$$

причем для целых значений аргументов имеет место простая формула

$$I_x(a,b)=\sum_{j=a}^{a+b-1}\frac{(a+b-1)!}{j!(a+b-1-j)!}x^j(1-x)^{a+b-1-j}.$$

Модифицированная функция Бесселя 1 рода вычисляется по формуле

$$I_\nu(z)=\sum_{k=0}^{\infty}\frac{(z/2)^{2k+\nu}}{k!\Gamma(k+\nu+1)}.$$

Для вычисления специальных функций разработаны компьютерные программы, которые производят вычисления с гарантированной точностью при помощи разложения в ряд и взятием конечного числа его членов, либо при помощи непрерывных усеченных, при достижении заданной точности, цепных дробей, либо при помощи аппроксимаций. Некоторые специальные случаи точного рекурсивного вычисления рассмотрены Де Гроотом. Кроме представленных функций, в программе используется также некоторое количество программных реализаций функций, различным элементарным образом преобразованных от данных функций. Это сделано с целью сохранения гарантированной точности и экономичности вычислений.

См. Абрамовиц с соавт., Брандт, Де Гроот, Краковский.

1.3.8.23. Методы вычисления

В вычислительном аспекте наименование «функция распределения» в настоящем программном обеспечении употребляется стандартно для математического объекта, в который подставляются статистика, а также некоторый набор параметров (в том числе, возможно, так называемые степени свободы), а в результате получается значение, имеющее смысл вероятности и, следовательно, заключенное в интервале [0;1].
И наоборот, «обратная функция распределения» в настоящем программном обеспечении – это такая функция, в которую подставляется параметр, имеющий смысл вероятности, а также, возможно, некоторый набор дополнительных параметров, аналогичный рассмотренной выше функции распределения, а в результате получается значение статистики. Вычисляется обратная функция распределения на основе численной реализации функции распределения методом деления отрезка пополам либо иным методом локальной оптимизации.
Для вычисления, по возможности, используется точное приведение формул к распределениям, вычисление которых проще (по крайней мере, хорошо отработано), и к известным специальным функциям. Если такое приведение неизвестно, используется прямое интегрирование с заданной, достаточной для практических применений, точностью.

1.3.8.23.1. Пример практического вычисления

Вычисление по теоретическим формулам часто не удается выполнить, если формулы запрограммировать в точности так, как указано в математической записи формулы. Подобное прямое вычисление возможно лишь для некоторых обычных значений параметров. Но данное вычисление не универсально. Оно совершенно не работает на малых или, наоборот, больших значениях параметров. Проблема тут заключается даже не в программировании, а в некоем промежуточном этапе между теорией (математикой) и практикой (программированием), который мы называем организацией вычислений. Для примера рассмотрим организацию вычислений регуляризированной неполной бета–функции. Стандартно данная функция вычисляется с помощью цепной дроби:

$$I_x(a,b) = \frac{x^a(1-x)^b}{aB(a,b)} \cdot \cfrac{1}{1+\cfrac{r_1}{1+\cfrac{r_2}{1+\cfrac{r_3}{\dots}}}},$$

где r_i, i = 1,2,3,... – коэффициенты, вычисляемые, в зависимости от их четности, по формулам:

$r_{2k+1} = -\frac{(a+k)(a+b+k)x}{(a+2k)(a+2k+1)}$ для нечетных или

$r_{2k} = \frac{k(b-k)x}{(a+2k-1)(a+2k)}$ для четных номеров.

Вычислительная проблема показанной формулы вычисления регуляризованной неполной бета–функции кроется в коэффициенте цепной дроби. Дело в том, что при больших значениях параметров a и b значения числителя и знаменателя быстро стремятся к нулю, что ведет к неопределенности. Стандартным приемом исключения такой неопределенности является преобразование формулы вычисления таким образом, чтобы слишком малые (или, наоборот, слишком большие) величины компенсировались до возникновения вычислительных проблем переполнения или потери значимости.

Пользуясь случаем, напомним, что другим стандартным приемом является логарифмирование выражения (с целью замены умножения величин сложением их логарифмов) с последующим выполнением операции взятия экспоненты. Третьим приемом является установление такой последовательности вычисления выражения, чтобы промежуточные результаты ни на одном из этапов вычислений не были слишком большими или слишком маленькими.

В преобразованиях нам понадобится аппроксимация бета–функции формулой Стирлинга, справедливая для больших значений параметров:

$$B(a,b) \approx \frac{\sqrt{2\pi}\,x^{x-0,5}\,y^{y-0,5}}{(x+y)^{x+y-0,5}}.$$

Подставив формулу Стирлинга в выражение коэффициента перед цепной дробью, получим следующее выражение для данного коэффициента

$$\frac{x^a(1-x)^b}{aB(a,b)} = \exp\left[\left(a+b-\frac{1}{2}\right)\ln(a+b) - \left(a-\frac{1}{2}\right)\ln a - \left(b-\frac{1}{2}\right)\ln b + a\ln x + b\ln(1-x) - \ln a - \frac{\ln(2\pi)}{2}\right].$$

Последний вопрос, который осталось решить – это установить, при каких значениях параметров a и b допустимо заменить точную теоретическую формулу показанной аппроксимацией. Оказалось удобным сделать это, анализируя абсолютное значение бета–

функции. При увеличении параметров данное значение весьма быстро стремится к нулю, что приводит вычисление по точной формуле к неизбежному краху. Поэтому данное значение можно взять порядка 10^{-36} для 32–разрядной вычислительной системы.
Приведенная схема вычисления наглядно показывает, насколько практическое вычисление может отличаться от теоретической формулы.

Список использованной и рекомендуемой литературы

1. Alam K. A Monotonicity property of the distribution of the studentized smallest chi–square // The Annals of Mathematical Statistics, 1970, vol. 41, no. 1, pp. 318–320.
2. Anderson S.A. Statistical methods for comparative studies: Techniques for bias reduction / S.A. Anderson, A. Auquier, W.W. Hauck et al. – New York, NY: John Wiley & Sons, 1980.
3. Armitage J.V., Krishnaiah P.R. Tables for the studentized largest chi–square distribution and their applications. – Columbus, OH: Wright–Patterson Air Force Base, 1964.
4. Armitage P. Encyclopedia of biostatistics / Ed. by P. Armitage, T. Colton. – New York, NY: John Wiley & Sons, 2005.
5. Balakrishnan N., Nevzorov V.B. A primer on statistical distributions. – New York, NY: John Wiley & Sons, 2003.
6. Baum J.–J., Chen H.J., Xiong M. Percentage points of the studentized range test for dispersion of normal means // Journal of Statistical Computation and Simulation, 1993, vol. 44, no. 3, pp. 149–163.
7. Bayarri M.J., Berger J.O. Quantifying surprise in the data and model verification // Bayesian Statistics 6 / Ed. by J.M. Bernardo et al. – Oxford: Oxford University Press, 1998, pp. 53–82.
8. Bayarri M.J., Berger, J.O. P–values for composite null models // Journal of the American Statistical Association, 2000, vol. 95, pp. 1127–1142.
9. Bechhofer R.E., Dunnett C.W. Comparisons for orthogonal contrasts: Examples and tables // Technometrics, 1982, vol. 24, pp. 213–222.
10. Breslow N.E., Day N.E. Statistical methods in cancer research. Volume I – The analysis of case–control studies (IARC Scientific Publications No. 32). – Lyon, France: International Agency for Research of Cancer, 1980.
11. Breslow N.E., Day N.E. Statistical methods in cancer research. Volume II – The design and analysis of cohort studies (IARC Scientific Publications No. 82). – Lyon, France: International Agency for Research of Cancer, 1987.
12. Bury K. Statistical distributions in engineering. – Cambridge, UK: Cambridge University Press, 1999.
13. Copenhaver M.D., Holland B.S. Computation of the distribution of the maximum studentized range statistic with application to multiple significance testing of simple effects // Journal of Statistical Computation and Simulation, 1988, vol. 30, no. 1, pp. 1–15.
14. Curran–Everett D., Benos D.J. Guidelines for reporting statistics in journals published by the American Physiological Society // American Journal of Physiology – Renal Physiology, 2004, vol. 287, pp. F169–F171.
15. Curran–Everett D., Benos D.J. Guidelines for reporting statistics in journals published by the American Physiological Society // Physiological Genomics, 2004, vol. 18, pp. 249–251.
16. Fisher R.A. The general sampling distribution of the multiple correlation coefficient // Proceedings of the Royal Society, Series A, 1928, vol. 121, pp. 654–673.
17. Currie I.D. On the distribution of the studentized range in a single normal sample // Scandinavian Journal of Statistics, 1980, no. 7, pp. 150–154.
18. David H.A., Nagaraja H.N. Order statistics. – Hoboken, NJ: John Wiley & Sons, 2003.
19. Dey D.K. Handbook of statistics. Vol. 25. Bayesian thinking: Modeling and computation /

Ed. by D.K. Dey, C.R. Rao. – New York, NY: Elsevier, 2005.
20. Di Bucchianico A. Combinatorics, computer algebra and Wilcoxon–Mann–Whitney test // Memorandum COSOR, Eindhoven University of Technology, 1996.
21. Esteve J., Benhamou E., Raymond L. Statistical methods in cancer research. Volume IV – Descriptive epidemiology (IARC Scientific Publications No. 128). – Lyon, France: International Agency for Research of Cancer, 1994.
22. Evans M., Hastings N., Peacock B. Statistical distributions. – New York, NY: John Wiley & Sons, 2000.
23. Fisher R.A. Statistical tables for biological, agricultural and medical research / Ed. by R.A. Fisher, F. Yates. – Edinburgh: Oliver and Boyd, 1963.
24. Ghosh S. Handbook of statistics. Vol. 13. Design and analysis of experiments / Ed. by S. Ghosh, C.R. Rao. – New York, NY: Elsevier, 1996.
25. Gleason J.R. An accurate, non–iterative approximation for studentized range quantiles // Computational statistics & data analysis, August 1999, vol. 31, no. 2, pp. 147–158.
26. Greenhalgh T. How to read a paper: Statistics for the non–statistician. I: Different types of data need different statistical tests // BMJ, 9 August 1997, vol. 315, pp. 364–366.
27. Greenhalgh T. How to read a paper: Statistics for the non–statistician. II: «Significant» relations and their pitfalls // BMJ, 16 August 1997, vol. 315, pp. 422–425.
28. Gupta S.S., Panchapakesan S. Multiple decision procedures; theory and methodology of selecting and ranking populations. – Philadelpia, PA: The Society for Industrial and Applied Mathematics, 2002.
29. Harter H.L., Balakrishnan N. Tables for the use of range and studentized range in tests of hypotheses. – Boca Raton, FL: CRC Press LLC, 1998.
30. Hassan A.S. Goodness–of–fit for the generalized exponential distribution // InterStat (Statistics on the Internet), July 2005, No. 1.
31. Higham N.J. Accuracy and stability of numerical algorithms. – Philadelphia, PA: Society for Industrial and Applied Mathematics, 1996.
32. Jaiswal A.K., Khandelwal A. A textbook of computer based numerical and statistical techniques. – New Delhi: New Age International, 2009.
33. Khattree R. Handbook of statistics. Vol. 22. Statistics in industry / Ed. by R. Khattree, C.R. Rao. – New York, NY: Elsevier, 2003.
34. Krishnaiah P.R. Handbook of statistics. Vol. 2. Classification, pattern recognition and reduction of dimensionality / Ed. by P.R. Krishnaiah, L.N. Kanal. – New York, NY: Elsevier, 1982.
35. Krishnaiah P.R. Handbook of statistics. Vol. 4. Nonparametric methods / Ed. by P.R. Krishnaiah, P.K. Sen. – New York, NY: Elsevier, 1984.
36. Krishnaiah P.R. Handbook of statistics. Vol. 6. Sampling / Ed. by P.R. Krishnaiah, C.R. Rao. – New York, NY: Elsevier, 1988.
37. Krzanowski W.J. Statistical principles and techniques in scientific and social investigations. – Oxford, NY: Oxford University Press, 2007.
38. L'Ecuyer P. Random number generation // Elsevier Handbooks in Operations Research and Management Science: Simulation / Ed. by S.G. Henderson, B.L. Nelson. – Elsevier Science, 2005.
39. L'Ecuyer P. Random number generation // The Handbook of Computational Statistics / Ed. by J.E. Gentle, W. Haerdle, Y. Mori. – Heidelberg: Springer–Verlag, 2004, pp. 35–70.
40. L'Ecuyer P., Hellekalek P. Random number generators: Selection criteria and testing // Random and Quasi–Random Point Sets (Lecture Notes in Statistics, vol. 138) / Ed. by P. Hellekalek, G. Larcher. – New York: Springer, 1998, pp. 223–265.
41. Laurencelle L., Dupuis F. Statistical tables, explained and applied. – Singapore: World

Scientific Publishing 2000.
42. Lentner C. Geigy scientific tables. Vol. 2. Introduction to statistics, statistical tables, mathematical formulae / Ed. by C. Lentner. – Basle, Switzerland: Ciba–Geigy, 1982.
43. Lester D. Exact statistics and continued fractions // Journal of Universal Computer Science, 1995, vol. 1, no. 7, pp. 504–513.
44. Maddala G.S. Handbook of statistics. Vol. 11. Econometrics / Ed. by G.S. Maddala, C.R. Rao, H.D. Vinod. – New York, NY: Elsevier, 1993.
45. Maddala G.S. Handbook of statistics. Vol. 14. Statistical methods in finance / Ed. by G.S. Maddala, C.R. Rao. – New York, NY: Elsevier, 1996.
46. Maddala G.S. Handbook of statistics. Vol. 15. Robust inference / Ed. by G.S. Maddala, C.R. Rao. – New York, NY: Elsevier, 1997.
47. Marsaglia G. Random number generators // Journal of Modern Statistical Methods, May 2003, vol. 2, no. 1, pp. 2–13.
48. Marsaglia G., Tsang W.W., Wang J. Fast generation of discrete random variables // Journal of Statistical Software, July 2004, vol. 11, no. 3.
49. McCullough B.D. Assessing the reliability of statistical software: Part I // The American Statistician, November 1998, vol. 52, no. 4, pp. 358–366.
50. McCullough B.D. Assessing the reliability of statistical software: Part II // The American Statistician, May 1999, vol. 53, no. 2, pp. 149–159.
51. Mosteller F., Bailar J.C. Medical uses of statistics. – Boston, MA: NEJM Books, 1992.
52. Nash J.C., Nash M.M. Scientific computing with PCs. – Ottawa: Nash Information Services, 1993.
53. Patil G.P. Handbook of statistics. Vol. 12. Environmental Statistics / Ed. by G.P. Patil, C.R. Rao. – New York, NY: Elsevier, 1994.
54. Pearson E.S. Further note on the distribution of range in samples taken from a normal population // Biometrika, 1926, vol. 18, no. 1–2, pp. 173–194.
55. Pillai K.C.S. On the distributions of midrange and semi–range in samples from a normal population // The Annals of Mathematical Statistics, 1950, vol. 21, no. 1, pp. 100–105.
56. Pillai K.C.S., Ramachandran K.V. On the distribution of the ratio of the ith observation in an ordered sample from a normal population to an independent estimate of the standard deviation // The Annals of Mathematical Statistics, 1954, vol. 25, no. 3, pp. 565–572.
57. Ramsey P.H., Ramsey P.P. Critical values for two multiple comparison procedures based on the studentized range distribution // Journal of Educational and Behavioral Statistics, 1990, vol. 15, no. 4, pp. 341–352.
58. Rao C.R. Handbook of statistics. Vol. 24. Data mining and data visualization / Ed. by C.R. Rao, E.J. Wegman, J.L. Solka. – New York, NY: Elsevier, 2005.
59. Rao C.R. Handbook of statistics. Vol. 26. Psychometrics / Ed. by C.R. Rao, S. Sinharay. – New York, NY: Elsevier, 2007.
60. Rao C.R. Handbook of statistics. Vol. 8. Statistical methods in biological and medical sciences / Ed. by C.R. Rao, R. Chakraborty. – New York, NY: Elsevier, 1991.
61. Rao C.R. Handbook of statistics. Vol. 9. Computational statistics / Ed. by C.R. Rao. – New York, NY: Elsevier, 1993.
62. Sahai H., Ageel M.I. The analysis of variance: fixed, random, and mixed models. – Boston, MA: Birkhauser, 2000.
63. Sellke T., Bayarri M.J., Berger J.O. Calibration of P–values for testing precise null hypotheses // The American Statistician, 2001, vol. 55, pp. 62–71.
64. Sen P.K. Handbook of statistics. Vol. 18. Bioenvironmental and public health statistics / Ed. by P.K. Sen, C.R. Rao. – New York, NY: Elsevier, 2000.
65. Simon J.L. Resampling: The new statistics. – Arlington, VA: Resampling Stats Inc., 1997.

66. Sterne J.A.C., Smith G.D. Sifting the evidence – what's wrong with significance tests? // BMJ, 27 January 2001, vol. 322, pp. 226–231.
67. Stoline M.R., Ury H.K. Tables of the studentized maximum modulus distribution and an application to multiple comparisons among means // Technometrics, February 1979, vol. 21, no. 1, pp. 87–93.
68. Taylor J.K. Cihon C. Statistical techniques for data analysis. – Boca Raton, FL: CRC Press LLC, 2004.
69. Tiku M.L., Akkaya A.D. Robust estimation and hypothesis testing. – New Delhi: New Age International, 2004.
70. Tippett L.H.C. On the extreme individuals and the range of samples taken from a normal population // Biometrika, 1925, vol. 17, no. 3–4, pp. 364–387.
71. Ury H.K., Stoline M.R., Mitchell B.T. Further tables of the studentized maximum modulus distribution // Communications in Statistics – Simulation and Computation, 1980, vol. 9, no. 2, pp. 167–178.
72. Van de Wiel M.A., Di Bucchianico A., Van der Laan P. Exact distributions of nonparametric test statistics using computer algebra // Memorandum COSOR, 1997, Eindhoven University of Technology.
73. Young D.H. Recurrence relations between the P.D.F.'s of order statistics of dependent variables, and some applications // Biometrika, 1967, vol. 54, no. 1–2, pp. 283–292.
74. Абрамовиц М. Справочник по специальным функциям с формулами, графиками и математическими таблицами / Под ред. М.Абрамовица, И. Стиган. – М.: Наука, 1979.
75. Барндорф–Нильсен О., Кокс Д. Асимптотические методы в математической статистике. – М.: Мир, 1999.
76. Бащинский С.Е. Качество Российских научных публикаций, посвящённых лечебным и профилактическим вмешательствам // Международный журнал медицинской практики, 2005, №1, с. 32–36.
77. Баюк Д.А. Почему мы доверяем науке? // Вокруг света, Март 2008, № 3 (2810).
78. Биглхол Р., Бонита Р., Кьельстрем Т. Основы эпидемиологии. – М.: Медицина, 1994.
79. Боровин Г.К., Комаров М.М., Ярошевский В.С. Ошибки–ловушки при программировании на фортране. – М.: Наука, 1987.
80. Боровков А.А. Математическая статистика. – М.: Наука, 1984.
81. Борцов Ю.С. Социология. Учебное пособие. – Ростов–на–Дону: Издательство «Феникс», 2002.
82. Брандт З. Анализ данных. Статистические и вычислительные методы для научных работников и инженеров. – М.: Мир, ООО «Издательство АСТ», 2003.
83. Браунли К.А. Статистическая теория и методология в науке и технике. – М.: Наука, 1977.
84. Вадзинский Р.Н. Справочник по вероятностным распределениям. – СПб.: Наука, 2001.
85. Власов В.В. Эффективность диагностических исследований. – М.: Медицина, 1988.
86. Гайдышев И. Анализ и обработка данных: специальный справочник. – СПб: Питер, 2001.
87. Гайдышев И.П. Статистика в публикациях // Гений ортопедии, 2005, № 4, с. 155–161.
88. Гланц С. Медико–биологическая статистика. – М.: Практика, 1998.
89. Глотов Н.В. Биометрия / Н.В. Глотов, Л.А. Животовский, Н.В. Хованов и др. – Л.: Издательство Ленинградского государственного университета, 1982.
90. Гудман С.Н. На пути к доказательной биостатистике. Часть 1: Обманчивость величины p // Международный журнал медицинской практики, 2002, № 1, с. 8–17.
91. Гудман С.Н. На пути к доказательной биостатистике. Часть 2: Байесовский критерий // Международный журнал медицинской практики, 2002, № 2, с. 5–14.

92. Де Гроот М. Оптимальные статистические решения. – М.: Мир, 1974.
93. Дэйвид Г. Порядковые статистики. – М.: Наука, 1979.
94. Иванов Е.Г. Исследования типа случай–контроль и когортные исследования // Акушерство и Гинекология в World Wide Web, 2002, № 7.
95. Камень Ю.Э., Камень Я.Э., Орлов А.И. Реальные и номинальные уровни значимости в задачах проверки статистических гипотез // Заводская лаборатория. Диагностика материалов, 1986, т. 52, № 12, с. 55–57.
96. Клюшин Д.А., Петунин Ю.И. Доказательная медицина. Применение статистических методов. – М.: ООО «И.Д. Вильямс», 2008.
97. Кобзарь А.И. Прикладная математическая статистика. Для инженеров и научных работников. – М.: ФИЗМАТЛИТ, 2006.
98. Козлов М.В. Мнимые повторности (pseudo replications) в экологических исследованиях: проблема, не замеченная российскими учеными // Журнал общей биологии, 2003, т. 64, № 4, с. 292–307.
99. Козлов М.В., Хелберт С.Х. Мнимые повторности, бесплодные дискуссии, и интернациональная сущность науки: Ответ Д.В. Татарникову // Журнал общей биологии, 2006, т. 67, № 2, с. 145–152.
100. Кокс Д., Снелл Э. Прикладная статистика. Принципы и примеры. – М.: Мир, 1984.
101. Колкот Э. Проверка значимости. – М.: Статистика, 1968.
102. Краковский Ю.М. Имитационное моделирование: Методические указания. – Иркутск: Издательство ИГЭА, 2002.
103. Кульбак С. Теория информации и статистика. – М.: Наука, 1967.
104. Лакатос И. Методология исследовательских программ. – М.: ООО «Издательство АСТ»: ЗАО НПП «Ермак», 2003.
105. Ланг Т. Двадцать ошибок статистического анализа, которые вы сами можете обнаружить в биомедицинских статьях // Международный журнал медицинской практики, 2005, № 1, с. 21–31.
106. Леонов В.П. Наукометрика статистической парадигмы экспериментальной биомедицины // Вестник Томского государственного университета, серия «Математика. Кибернетика. Информатика», апрель 2002, № 275, с. 17–24.
107. Леонов В.П. Применение статистики в статьях и диссертациях по медицине и биологии. Часть II. История биометрии и ее применения в России // Международный журнал медицинской практики, 1999, № 4, с. 7–19.
108. Леонов В.П. Применение статистики в статьях и диссертациях по медицине и биологии. Часть IV. Наукометрия статистической парадигмы экспериментальной биомедицины // Международный журнал медицинской практики, 2002, № 3, с. 6–10.
109. Леонов В.П., Ижевский П.В. Об использовании прикладной статистики при подготовке диссертационных работ по медицинским и биологическим специальностям // Бюллетень ВАК РФ, 1997, № 5, с. 56–61.
110. Леонов В.П., Ижевский П.В. Применение статистики в статьях и диссертациях по медицине и биологии. Часть I. Описание методов статистического анализа в статьях и диссертациях // Международный журнал медицинской практики, 1998, № 4, с. 7–12.
111. Ликеш И., Ляга Й. Основные таблицы математической статистики. – М.: Финансы и статистика, 1985.
112. Мак–Кракен Д., Дорн У. Численные методы и программирование на ФОРТРАНе. – М.: Мир, 1977.
113. Мардиа К., Земроч П. Таблицы F–распределений и распределений, связанных с

ними. – М.: Наука, 1984.
114. Мартынов Г.В. Вычисление функции нормального распределения // Итоги науки и техники. Серия «Теория вероятностей. Математическая статистика. Теоретическая кибернетика», 1979, т. 17, с. 57–84.
115. Оуэн Д.Б. Сборник статистических таблиц. – М.: ВЦ АН СССР, 1973.
116. Петри А., Сэбин К. Наглядная статистика в медицине. – М.: ГЭОТАР–МЕД, 2003.
117. Петрович М.Л., Давидович М.И. Статистическое оценивание и проверка гипотез на ЭВМ. – М.: Финансы и статистика, 1989.
118. Плохинский Н.А. Достаточная численность выборки // В сб. Биометрический анализ в биологии. – М.: Издательство МГУ, 1982, с. 152–157.
119. Попов Б.А., Теслер Г.С. Вычисление функций на ЭВМ. Справочник. – Киев: Наукова Думка, 1984.
120. Поппер К.Р. Логика и рост научного знания. Избранные работы. – М.: Прогресс, 1983.
121. Поппер К.Р. Объективное знание. Эволюционный подход. – М.: Эдиториал УРСС, 2002.
122. Прохоров А.М. Большой энциклопедический словарь: В 2–х тт. / Гл. ред. А.М. Прохоров. – М.: Советская энциклопедия, 1991.
123. Прохоров Ю.В. Вероятность и математическая статистика. Энциклопедия / Гл. ред. Ю.В. Прохоров. – М.: Научное издательство «Большая Российская энциклопедия», 1999.
124. Прохоров Ю.В. Математический энциклопедический словарь / Гл. ред. Ю.В. Прохоров. – М.: Научное издательство «Большая Российская энциклопедия», 1995.
125. Рекомендации. Прикладная статистика. Методы обработки данных. Основные требования и характеристики. – М.: ВНИИС,1987.
126. Родионов Д.А. Справочник по математическим методам в геологии / Д.А. Родионов, Р.И.Коган, В.А. Голубева и др. – М.: Недра, 1987.
127. Родионов Д.А. Статистические решения в геологии. – М.: Недра, 1981.
128. Теннант–Смит Дж. Бейсик для статистиков. – М.: Мир, 1988.
129. Тутубалин В.Н. Теория вероятностей и случайных процессов: Учебное пособие. – М.: Издательство МГУ, 1992.
130. Тюрин Ю.Н., Макаров А.А. Статистический анализ данных на компьютере. – М.: ИНФРА–М, 1999.
131. Урбах В.Ю. Биометрические методы. Статистическая обработка опытных данных в биологии, сельском хозяйстве и медицине. – М.: Наука, 1964.
132. Флетчер Р., Флетчер С., Вагнер Э. Клиническая эпидемиология: Основы доказательной медицины. – М.: Медиа Сфера, 2004.
133. Хан Г., Шапиро С. Статистические модели в инженерных задачах. – М.: Мир, 1969.
134. Хастингс Н., Пикок Дж. Справочник по статистическим распределениям. – М.: Статистика, 1980.
135. Хромов–Борисов Н.Н. Биометрические аспекты популяционной генетики / В кн. Кайданов Л.З. Генетика популяций. – М.: Высшая школа, 1996, с. 251–308.
136. Шенк Х. Теория инженерного эксперимента. – М.: Мир, 1972.
137. Шеффе Г. Дисперсионный анализ. – М.: Наука, 1980.
138. Шмерлинг Д.С. О проверке согласованности экспертных оценок // В сб. Статистические методы анализа экспертных оценок. Ученые записки по статистике, т. 29 / Под ред. Ю.Н. Тюрина, А.А. Френкель. – М.: Наука, 1977, с. 77–83.

139. Шмерлинг Д.С. Экспертные оценки. Методы и применение (обзор) / Д.С. Шмерлинг, С.А. Дубровский, Т.Д. Аржанова и др. // В сб. Статистические методы анализа экспертных оценок. Ученые записки по статистике, т. 29 / Под ред. Ю.Н. Тюрина, А.А. Френкель. – М.: Наука, 1977, с. 290–382.
140. Шор Я.Б., Кузьмин Ф.И. Таблицы для анализа и контроля надежности. – М.: Советское радио, 1968.
141. Эренштайн В. Исследования типа случай – контроль // Международный журнал медицинской практики, 2007, № 1, с. 39–50.
142. Эренштайн В. Обсервационные исследования // Международный журнал медицинской практики, 2006, № 3, с. 18–30.

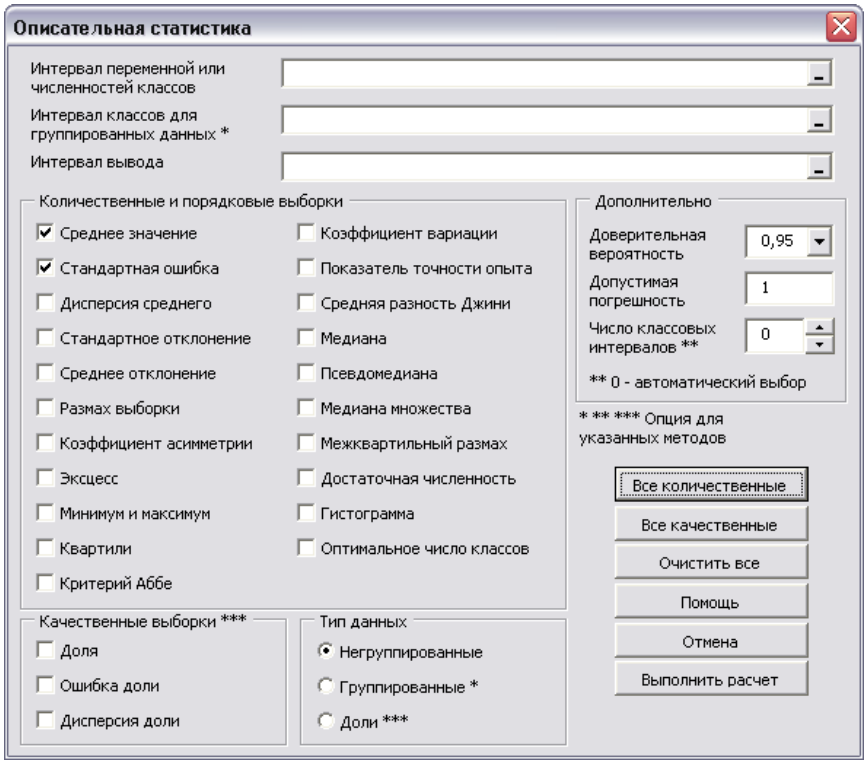
Глава 2. Описательная статистика

2.1. Введение

Программное обеспечение описательной статистики обеспечивает вычисление основных показателей описательной статистики количественных и качественных показателей. При этом исходные данные могут быть представлены в качестве эмпирической выборки или в сгруппированном виде. Подробнее представление исходных данных рассмотрено в одноименном разделе.

2.2. Работа с программным обеспечением

Выберите из меню программы пункт AtteStat | Описательная статистика. На экране появится диалоговое окно, изображенное на рисунке:



Затем проделайте следующие шаги:

- Выберите или введите интервал исходной выборочной совокупности. Если исходные данные представлены в сгруппированном виде, в данном поле выбирается или

вводится интервал численностей классов.
- Если исходные данные представлены в сгруппированном виде, в данном поле выбирается или вводится интервал классов. Содержимое интервала классов определяется шкалой измерения исходных данных (см. главу «Введение»). При выборе данного представления для типа данных укажите опцию «Группированные».
- В случае качественных (бинарных) данных возможен ввод в виде долей. Подробности см. в разделе «Представление исходных данных». При выборе данного представления для типа данных укажите опцию «Доли».
- Выберите или введите выходной интервал для выдачи результатов расчета статистических показателей. Начиная с первой ячейки выходного интервала (следовательно, можно указать только одну ячейку, т.к. остальные ячейки интервала игнорируются), будут выведены вычисленные отмеченные Вами выборочные показатели описательной статистики.
- Отметьте необходимые параметры расчета статистических показателей, пользуясь соответствующими кнопками.
- Нажмите кнопку «Выполнить расчет».

Для вычисления ряда показателей требуется выбрать доверительную вероятность (доверительный уровень) или ввести допустимую погрешность. Для построения гистограммы потребуется ввести число классов либо оставить нулевое значение. В последнем случае программа сама вычислит число классов. Отметим, что программа предоставляет возможность с помощью различных методов вычислить оптимальное число классов. Данный показатель, предварительно вычислив, можно использовать для построения гистограммы.
При ошибках, вызванных неверными действиями пользователя при вводе исходных данных для расчета, выдаются сообщения об ошибках.

2.2.1. Представление исходных данных

Настоящее программное обеспечение может обрабатывать исходные данных, представленные в следующих шкалах измерения:
- количественная,
- порядковая,
- качественная (номинальные признаки),
- качественная (бинарные признаки).

Исходные данные различных типов определяют разнообразные способы их представления для расчета. Поэтому в данное программное обеспечение введена опция «Тип данных», которая может иметь следующие значения:
- негруппированные,
- группированные,
- доли.

Опция «Тип данных» имеет значение по умолчанию «Негруппированные». В данном случае необходимо указать только интервал исходных данных. Применимо для количественных, порядковых и качественных (как номинальных, так и бинарных) данных.
Опция «Тип данных» со значением «Группированные» требует ввода интервала численностей классов и интервала классов. Применимо для количественных, порядковых и качественных (как номинальных, так и бинарных) данных.
Опция «Тип данных» со значением «Доли» требует ввода интервала данных, содержащего всего два числа. Эти числа представляют собой численности двух классов в случае бинарных данных. Интервал классов в данном случае вводить необязательно – он уже подразумевается программой. Применимо только для качественных (бинарных) данных.

В качестве примера рассмотрим все возможные случаи представления в данном программном обеспечении бинарной выборки численностью 7 с количеством случаев 3. В программе предполагается, что бинарная выборка может состоять только из нулей и единиц. При этом наличие признака кодируется единицей (например – наличие симптома заболевания). Отсутствие признака кодируется нулем. Как указано выше, для данной выборки допустимо вводить данные для расчета в негруппированном, группированном виде, в виде долей. Ввод стандартный.
В негруппированном виде выборка может быть введена как (транспонировано)
1 0 0 1 1 0 0
В группированном виде та же выборка будет на рабочем листе электронных таблиц выглядеть как

3	1
4	0

В виде долей та же выборка будет представлена как (транспонировано)
3 4
О группировке данных см. также раздел «Формулы для сгруппированных выборок».

2.2.2. Сообщения об ошибках

При ошибках ввода исходных данных для расчета могут выдаваться диагностические сообщения следующих типов:

Ошибка	Комментарий
Пустая ячейка в области данных.	Проверьте исходные данные и заполните все ячейки, отмеченные Вами как входной интервал. Для избежания ошибок расчета, вызванных разногласиями, трактовать ли пустую ячейку как нуль, программное обеспечение требует заполнения всех ячеек. Если в ячейке не должно быть данных по физической природе исследуемого процесса, введите в данную ячейку нуль.
Нечисловой тип данных.	Проверьте типы ячеек входного интервала. Тип может быть только числовым. Проще всего выделить интервал ячеек и явно определить их тип как числовой стандартными средствами.
Мало данных.	Для расчета необходимо выбрать интервал, содержащий хотя бы четыре ячейки с числовыми значениями. Данная минимальная численность выборки лимитируется формулами вычисления статистических показателей описательной статистики.
Не определен интервал переменной.	Вы не выбрали или неверно ввели входной интервал. Лучшим способом избежать ошибки является не ввод, а выделение интервала стандартным образом, т. е. протаскиванием курсора.
Не определен интервал вывода.	Вы не выбрали или неверно ввели выходной интервал. Лучшим способом избежать ошибки является не ввод, а выделение интервала стандартным образом, т. е. протаскиванием курсора.
Не определен интервал классов.	Вы не выбрали или неверно ввели интервал классов. Лучшим способом избежать ошибки является не ввод, а выделение интервала стандартным образом, т. е. протаскиванием курсора.
Разное число классов и	При расчете по сгруппированным данным количество классовых интервалов и численностей классов должны совпадать.

численностей классов.	
Не задана допустимая погрешность.	Для вычисления достаточной численности выборки следует корректно задать величину допустимой погрешности, как это указано в разделе «Достаточная численность выборки», посвященном анализу репрезентативности. Данная величина должна быть числом в тех же единицах измерения, что и исследуемая эмпирическая выборка.
Нулевая допустимая погрешность.	Для вычисления достаточной численности выборки следует корректно задать величину допустимой погрешности, как это указано в разделе «Достаточная численность выборки», посвященном анализу репрезентативности. Данная величина должна быть числом в тех же единицах измерения, что и исследуемая эмпирическая выборка, и не может равняться нулю.

2.3. Теоретическое обоснование

Эмпирические (опытные, экспериментальные) выборки (совокупности) состоят из отдельных вариант (элементов), которые объединены общностью некоторых свойств (признаков, переменных). Выборки могут быть получены в результате медико–биологического или технического эксперимента, научного опыта, социологического опроса и т. п. Источник появления выборок для статистического анализа значения не имеет. Единственное требование к анализируемым данным программным обеспечением выборкам определяется представленными методами расчета. Они применимы только к таким выборкам, варианты которых измерены в соответствующей шкале. Для большинства методов, представленных в данной программе, предполагается количественная шкала измерения исходных данных. Если варианты выборок измерены в порядковой или номинальной шкале, следует применять иные методы расчета описательной статистики.
В представленном программном обеспечении рассчитываются следующие выборочные статистические показатели описательной статистики:

- численность выборки,
- показатели положения: среднее значение и его стандартная ошибка, медиана, псевдомедиана,
- показатели разброса (рассеяния, масштаба): дисперсия, стандартное отклонение, среднее отклонение, размах, коэффициент вариации, средняя разность Джини, квартили, межквартильный размах,
- показатели формы распределения: коэффициент асимметрии, эксцесс.

Кроме перечисленных показателей, по выборке рассчитываются:

- достаточная численность выборки, из анализа заданных и рассчитанных выборочных показателей,
- оптимальное число классов.
- минимум и максимум.

Для качественных (бинарных) выборок может быть рассчитана доля, ошибка доли и дисперсия доли.
Для всех показателей рассчитываются как точечные, так и интервальные оценки. При этом на распечатке для краткости доверительные интервалы обозначаются аббревиатурой ДИ.
Напомним, что параметры положения и разброса количественной выборки могут оцениваться двумя методами: методом моментов и методом квантилей.

- Использование метода моментов дает в качестве параметрической точечной оценки положения среднее значение, в качестве параметра разброса – дисперсию.

- Использование метода квантилей в качестве непараметрической точечной оценки параметра положения приводит к медиане, в качестве параметра разброса – к межквартильному размаху.

В программе рассчитываются как точечные оценки параметров эмпирической выборки, так и параметрические и непараметрические интервальные оценки всех параметров, для которых данное понятие применимо. Таким образом, пользователь получает возможности гибкого представления описательной статистики, доступные только в данной программе.
Доверительная вероятность (доверительный уровень) требуется для вычисления ряда выборочных статистических показателей, и, в отличие от ряда других параметров, является не вычисляемой по выборке, а задаваемой пользователем программы величиной. Она выбирается из следующей стандартной линейки (в основном, следуя классификации Плохинского):

- Нулевой порог 0,90 применяется для работы с пониженной ответственностью, при первом ознакомлении с явлением.
- Первый порог 0,95 применяется в большинстве исследований (например, биологические исследования).
- Второй порог 0,99 для работ с повышенной ответственностью (например, медицинские исследования).
- Третий порог 0,999 применяется для работ с высокой ответственностью (например, исследования эффективности лекарств).

Доверительный уровень может быть выражен как долях, например, 0,95, так и в процентах, что то же самое, 95%.
С конкретными примерами использования доверительной вероятности можно ознакомиться по описаниям соответствующих статистических выборочных показателей.

2.3.1. Численность выборки

Количество вариант совокупности в источниках называют по–разному. Так, если речь идет об эмпирической выборке, количество ее элементов может называться численностью, величиной или размером. Термин «размерность» употреблять в значении «численность» не следует, т.к. он зарезервирован для описания так называемых многомерных совокупностей. Традиционными в отечественной статистической литературе являются термины «выборка», «варианта» и «численность», поэтому по возможности следует придерживаться их.

2.3.2. Среднее значение

Выборочное среднее значение – традиционно наиболее часто применяемый статистический показатель, характеризующий середину эмпирической совокупности. Иначе, выборочное среднее значение – популярная оценка параметра положения.

2.3.2.1. Общая методика

Среднее значение – это параметрическая оценка параметра положения статистического распределения. Следовательно, для вычисления оценки среднего значения мы должны задаться (или установить на основе эмпирических данных) типом распределения статистической совокупности. Затем следует воспользоваться одним из методов оптимизации (обычно используется метод максимального правдоподобия) с целью вычисления данной оценки. Процесс включает следующие этапы:

- Составление функционала.
- Определение производных функционала по искомым параметрам.
- Приравнивание производных нулю с целью получения системы линейных или

нелинейных алгебраических уравнений для вычисления оптимальных (доставляющих экстремум функционалу: минимум – для метода наименьших квадратов, максимум – для метода максимального правдоподобия) значений параметров.

- Решение уравнений. Для некоторых моделей бывает достаточно одного уравнения для данного параметра, в результате преобразования которого получается простая алгебраическая формула. Для других моделей приходится аналитически или численно решать систему уравнений.

2.3.2.2. Оценка среднего на основе теории распределений

Пусть имеется количественная выборка, имеющая нормальное распределение с плотностью

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}},$$

где μ – параметр положения статистического распределения,
σ – параметр масштаба.
Вычислим оценку максимального правдоподобия для параметра положения. В рассматриваемом случае функция максимального правдоподобия (ФМП) запишется как

$$L(\mu,\sigma) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}},$$

где n – численность выборки,
x_i, i = 1,2,...,n,– значения вариант выборки.
Оптимальные значения параметров доставляют максимум ФМП. Вычисления упрощаются, если исследовать не саму ФМП, а ее логарифм, т.к. ФМП и логарифм ФМП достигают максимума при одних и тех же значениях параметров. Логарифмическая ФМП имеет вид

$$\ln L(\mu,\sigma) = -\frac{1}{\sqrt{2\pi}}\left[n\ln\sigma + \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2 \right].$$

Максимум логарифмической ФМП достигается при равенстве нулю частных производных по параметрам. Частная производная логарифмической ФМП по интересующему нас сейчас параметру μ будет

$$\frac{\partial L(\mu,\sigma)}{\partial\mu} = -\frac{1}{\sigma^2\sqrt{2\pi}}\sum_{i=1}^{n}(x_i-\mu) = 0,$$

откуда очевидно получается

$$\sum_{i=1}^{n}(x_i-\mu) = \sum_{i=1}^{n}x_i - n\mu = 0$$

и, окончательно,

$$\mu = \frac{1}{n}\sum_{i=1}^{n}x_i.$$

Обращаем внимание, что классическая формула выборочной оценки среднего значения получена в предположении нормального распределения количественной случайной величины. Следовательно, вычисленную по данной формуле оценку допустимо применять только для нормально распределенной количественной величины, но не для величин в других шкалах измерения и с другими функциями распределения.
При описании результатов экспериментального исследования в медико–биологических науках выборочную оценку среднего значения принято обозначать символами $\bar{x}$, M или E, причем последние символы стандартно принято использовать в смысле оператора над случайной величиной. В некоторых источниках среднее [значение] часто эквивалентно

называют средней [величиной].
Доверительный интервал оцениваемого среднего значения вычисляется на заданном доверительном уровне, выражаемом в долях или процентах. Доверительный интервал, вычисленный на доверительном уровне, например, 95% (или, то же самое в долях, 0,95), означает, что 95% вариант выборочной совокупности попадают в данный интервал. Иначе, истинное значение среднего значения генеральной совокупности (математического ожидания) находится между нижней и верхней значениями доверительного интервала с вероятностью, равной доверительной.
Для вычисления доверительного интервала оцениваемого среднего значения в случае, если эмпирическая выборка распределена нормально, используется формула:

$$I_m = \left(\bar{x} - t_{(1+\beta)/2} \frac{\sigma}{\sqrt{n}}; \bar{x} + t_{(1+\beta)/2} \frac{\sigma}{\sqrt{n}} \right),$$

где σ – стандартное отклонение,
$t_{(1+\beta)/2}$ – значение обратной функции t–распределения Стьюдента с параметрами n – 1 и $(1 + \beta) / 2$,
β – доверительный уровень, выраженный в долях.
Для вычисления доверительного интервала оцениваемого среднего значения, когда выборка не является нормальной, применяется формула:

$$I_m = \left(\bar{x} - \Psi((1+\beta)/2) \frac{\sigma}{\sqrt{n}}; \bar{x} + \Psi((1+\beta)/2) \frac{\sigma}{\sqrt{n}} \right),$$

где Ψ(.) – обратная функция стандартного нормального распределения.

Метод максимума правдоподобия представлен Тику (Tiku) с соавт. О вычислении доверительных интервалов оцениваемого среднего значения см. статью Орлова, книгу Мюллера с соавт., статью Жоу (Zhou) с соавт.

2.3.2.3. Оценка среднего на основе теории множеств

Попытаемся дать универсальное понятие среднего значения, не зависящее ни от шкалы измерения эмпирических данных, ни от их размерности. Пусть имеется эмпирическая случайная выборка $\{X_1, X_2, \ldots, X_n\}$ численностью n, где X_i, $i = 1,2,...,n$ – варианты, скалярные или векторные, иначе эмпирические реализации случайной величины. Обозначим через $d(X, X_i)$ расстояние между произвольной реализацией случайной величины X и величиной X_i, $i = 1,2,...,n$. Основное требование к данному расстоянию состоит в его допустимости в используемой шкале измерения эмпирической выборки.
Средним значением выборки будет случайная величина X, удовлетворяющая условию

$$M\{X_1, X_2, ..., X_n\} = \arg\min_{X \in R} \sum_{i=1}^{n} d(X, X_i),$$

где R – пространство всех допустимых, с точки зрения шкалы измерений, реализаций случайной величины X.
Иначе, среднее ищется среди всех возможных, а не только среди полученных в опыте, реализаций X. Поэтому в общем случае среднее значение не является никакой из вариант X_i, $i = 1,2,...,n$. Это свойство можно считать слабостью в понятии точечной оценки среднего значения, компенсировать которую призваны представленные выше интервальные оценки.
В настоящем разделе реализовано вычисление среднего количественной выборки. Оно несложно благодаря использованию в качестве расстояния Евклидовой метрики. Сложности вычисления среднего значения возникают в шкалах категорий. Так, в шкале ранжировок без связей (см. главу «Обработка экспертных оценок») для поиска среднего значения,

удовлетворяющего представленному выше условию, необходимо перебрать $n!$ всевозможных реализаций случайной величины X, что при больших численностях представляет собой трудную задачу при современном уровне развития компьютерной техники.

2.3.2.4. Стандартная ошибка

Стандартная ошибка среднего значения определяется по формуле:

$$\mu = \frac{\sigma}{\sqrt{n}},$$

где σ – стандартное отклонение,
n – численность выборки.
При описании результатов экспериментального исследования в медико–биологических науках стандартную ошибку принято обозначать символом m. Обычно используется понятная большинству исследователей традиционная запись, характеризующая среднее значение и его стандартную ошибку, в виде $M \pm m$. Почему это именно так, поясняется на с. 24 и далее классической монографии Тейлора. Тем не менее, в работах следует помещать расшифровку обозначений всех используемых показателей во избежание разночтений, не полагаясь на общеупотребительность тех или иных обозначений.
Некоторые авторы через символ ± при описании параметра положения (среднего значения или медианы) или иного статистического параметра пытаются записывать не ошибку, а доверительные (толерантные) интервалы. Так поступать следует с осторожностью, ибо доверительные интервалы не для всех статистических параметров бывают симметричными.

2.3.2.5. Дисперсия

Основным статистическим показателем, характеризующим разброс выборки, является выборочная дисперсия. Общая методика оценки и вид функционала для нормальной количественной выборки представлены в разделе «Среднее значение».
Пусть имеется количественная выборка, имеющая нормальное распределение с плотностью

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}},$$

где μ – параметр положения статистического распределения,
σ – параметр масштаба.
Вычислим оценку максимального правдоподобия для параметра масштаба. В рассматриваемом случае функция максимального правдоподобия (ФМП) запишется как

$$L(\mu,\sigma) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}},$$

где n – численность выборки,
x_i, $i = 1,2,...,n$ – значения вариант выборки.
Оптимальные значения параметров доставляют максимум ФМП. Вычисления упрощаются, если исследовать не саму ФМП, а ее логарифм, т.к. ФМП и логарифм ФМП достигают максимума при одних и тех же значениях параметров. Логарифмическая ФМП имеет вид

$$\ln L(\mu,\sigma) = -\frac{1}{\sqrt{2\pi}}\left[n\ln\sigma + \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2 \right].$$

Максимум логарифмической ФМП достигается при равенстве нулю частных производных по параметрам. Частная производная логарифмической ФМП по интересующему нас сейчас параметру σ будет

$$\frac{\partial L(\mu,\sigma)}{\partial \sigma} = -\frac{1}{\sigma\sqrt{2\pi}}\left[n - \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right] = 0,$$

откуда очевидно получается

$$n\sigma^2 - \sum_{i=1}^{n}(x_i-\mu)^2 = 0$$

и, окончательно,

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2.$$

Величину σ² называют выборочной дисперсией и часто обозначают как *S* или *D*, причем последний символ стандартно принято использовать в смысле оператора над случайной величиной.

Обращаем внимание, что классическая формула выборочной оценки дисперсии получена в предположении нормального распределения количественной случайной величины. Следовательно, вычисленную по данной формуле оценку допустимо применять только для нормально распределенной количественной величины, но не для величин в других шкалах измерения и с другими функциями распределения.

Хотя для больших выборок это несущественно, считается, что будет неверным пользоваться полученной выше формулой для дисперсии, если оценка среднего значения совокупности производится также по выборке. Обозначим:

ξ– случайная величина,

*M*ξ – математическое ожидание,

*D*ξ – выборочная дисперсия,

$\bar{x}$ – выборочное среднее значение.

Согласно определению и учитывая, что

$$M(\bar{x}-a)^2 = D\bar{x} = \frac{1}{n}D\xi,$$

где *a* – известное среднее значение совокупности,

можно записать

$$nD\xi = \sum_{i=1}^{n}(x_i-a)^2 = \sum_{i=1}^{n}\left[(x_i-\bar{x})-(a-\bar{x})\right]^2 =$$

$$= \sum_{i=1}^{n}(x_i-\bar{x})^2 - 2(a-\bar{x})\sum_{i=1}^{n}(x_i-\bar{x}) + n(a-\bar{x})^2 .$$

В последнем выражении сумма во втором члене, очевидно, дает нуль, поэтому, перенеся первый член этого выражения в левую часть и сменив знак, получаем

$$\sum_{i=1}^{n}(x_i-\bar{x})^2 = nD\xi - D\xi = (n-1)D\xi,$$

откуда непосредственно следует, что в случае оценки среднего значения по выборке в качестве оценки дисперсии выборочной совокупности берется величина, определяемая по формуле:

$$D = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2.$$

Представленная формула, используемая в программе, вычисляет так называемую несмещенную выборочную оценку дисперсии генеральной совокупности (эмпирическую дисперсию).

Получим эквивалентную формулу для выборочной дисперсии, не содержащую значения

выборочного среднего.

$$D = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i^2 - 2x_i\bar{x} + \bar{x}^2) = \frac{1}{n-1}\left[\sum_{i=1}^{n}x_i^2 - \sum_{i=1}^{n}(2x_i\bar{x} - \bar{x}^2)\right] =$$

$$= \frac{1}{n-1}\left[\sum_{i=1}^{n}x_i^2 - \bar{x}\sum_{i=1}^{n}(2x_i - \bar{x})\right] = \frac{1}{n-1}\left[\sum_{i=1}^{n}x_i^2 - \bar{x}\left(2\sum_{i=1}^{n}x_i - \sum_{i=1}^{n}\bar{x}\right)\right].$$

Обратим внимание, что в круглых скобках получилась разность удвоенной суммы вариант выборки и просто суммы вариант, ибо

$$\sum_{i=1}^{n}\bar{x} = n\bar{x} = n\frac{1}{n}\sum_{i=1}^{n}x_i = \sum_{i=1}^{n}x_i.$$

Поэтому продолжим, подставив выражение для среднего арифметического значения,

$$D = \frac{1}{n-1}\left[\sum_{i=1}^{n}x_i^2 - \frac{1}{n}\sum_{i=1}^{n}x_i\sum_{i=1}^{n}x_i\right] = \frac{1}{n-1}\left[\sum_{i=1}^{n}x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n}x_i\right)^2\right].$$

Для вычисления доверительного интервала оцениваемой дисперсии в случае, если эмпирическая выборка распределена нормально, применяется формула:
$I_D = (D - t_{(1+\beta)/2}d;\ D + t_{(1+\beta)/2}d)$,
где $t_{(1+\beta)/2}$ – значение обратной функции t–распределения Стьюдента с параметрами $n - 1$ и $(1 + \beta) / 2$,
β – доверительный уровень, выраженный в долях,
d – величина, рассчитанная по формуле

$$d = \sqrt{\frac{1}{n}\left(m_4 - \left(\frac{n-1}{n}\right)^4 D^2\right)},$$

где m_4 – четвертый центральный выборочный момент, вычисляемый по формуле

$$m_4 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4.$$

Для вычисления доверительного интервала оцениваемой дисперсии, когда выборка не является нормальной, применяется формула:
$I_D = (D - \Psi((1 + \beta) / 2)d;\ D + \Psi((1 + \beta) / 2)d)$,
где $\Psi(.)$ – обратная функция стандартного нормального распределения.

Метод максимума правдоподобия представлен Тику (Tiku) с соавт. О вычислении доверительных интервалов оцениваемой дисперсии см. статью Орлова, книгу Мюллера с соавт.

2.3.2.6. Стандартное отклонение

Стандартным отклонением (средним квадратическим отклонением, средним квадратичным отклонением, стандартом, сигмой) называют корень квадратный из дисперсии. Вычисление стандартного отклонения производится по формуле:

$$\sigma = \sqrt{D},$$

где D – выборочная дисперсия.
Для вычисления доверительного интервала оцениваемого стандартного отклонения количественной выборки в случае, если эмпирическая выборка распределена нормально, применяется формула:

$$I_\sigma = \left(\sigma \cdot \sqrt{\frac{n-1}{\chi^2_{(1-\beta)/2}}}; \sigma \cdot \sqrt{\frac{n-1}{\chi^2_{(1+\beta)/2}}} \right),$$

где n – численность выборки,

$\chi^2_{(1-\beta)/2}$ – значение обратной функции χ^2–распределения с параметрами $n - 1$ и $(1 - \beta) / 2$,

$\chi^2_{(1+\beta)/2}$ – значение обратной функции χ^2–распределения с параметрами $n - 1$ и $(1 + \beta) / 2$,

β – доверительный уровень, выраженный в долях.

Для вычисления доверительного оцениваемого интервала стандартного отклонения, когда выборка не является нормальной, применяется формула:

$I_\sigma = (\sigma - \Psi((1 + \beta) / 2)d / (2\sigma); \sigma + \Psi((1 + \beta) / 2)d / (2\sigma))$,

где $\Psi(.)$ – обратная функция стандартного нормального распределения,

d – величина, рассчитанная по формуле

$$d = \sqrt{\frac{1}{n}\left(m_4 - \left(\frac{n-1}{n} \right)^4 D^2 \right)},$$

где m_4 – четвертый центральный выборочный момент, вычисляемый по формуле

$$m_4 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^4,$$

где x_i, $i = 1,2,...,n$ – значения вариант выборки,

$\bar{x}$ – среднее значение.

О вычислении доверительных интервалов оцениваемого стандартного отклонения см. статью Орлова, книгу Мюллера с соавт.

2.3.2.7. Среднее отклонение

Выборочное среднее отклонение (выборочная оценка среднего отклонения), подобно стандартному отклонению, характеризует разброс эмпирической выборки относительно среднего значения и вычисляется по формуле

$$\hat{x} = \frac{1}{n}\sum_{i=1}^{n}\left|x_i - \bar{x}\right|,$$

где n – численность выборки,

x_i, $i = 1,2,...,n$ – значения вариант выборки,

$\bar{x}$ – выборочное среднее значение.

Среднее отклонение отражает так называемый модульный подход к вычислению меры отклонения между величинами в противоположность тому, что стандартное отклонение отражает квадратический подход. Подобный выбор возникает перед исследователем не только в описательном статистическом анализе, а и во многих других областях математики. Квадратический подход находит применение из–за удобства дифференцирования квадратического функционала (см. главу «Многомерное шкалирование»). Кроме того, квадратический функционал имеет еще ряд преимуществ перед модульным функционалом, анализ которых выходит за рамки настоящего повествования.

2.3.2.8. Средняя разность Джини

Средняя разность Джини характеризует разброс значений вариант эмпирической выборки друг относительно друга и не зависит от какого–либо центрального значения, например, от

среднего значения или медианы. Вычисление выборочной средней разности Джини производится по формуле

$$g = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{\substack{j=1 \\ j \neq i}}^{n} \left| x_i - x_j \right|,$$

где n – численность выборки,
x_i, $i = 1,2,...,n$ – значения вариант выборки.

2.3.3. Асимметрия

Асимметрия характеризует форму статистического распределения. Если коэффициент асимметрии больше нуля, асимметрия правосторонняя (положительная), форма кривой распределения скошена вправо относительно кривой плотности нормального распределения. Если коэффициент асимметрии меньше нуля, то асимметрия левосторонняя (отрицательная), форма кривой распределения скошена влево относительно кривой плотности нормального распределения. Коэффициент асимметрии выборочной совокупности вычисляется по уточненной формуле:

$$A = \frac{n}{(n-1)(n-2)} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{\sigma} \right)^3,$$

где n – численность выборки,
x_i, $i = 1,2,...,n$ – значения вариант выборки,
$\bar{x}$ – выборочное среднее значение,
σ – выборочное стандартное отклонение.
Вычисление доверительного интервала оцениваемого коэффициента асимметрии производится по формуле:

$$I_A = \left(A - \sqrt{\frac{D_A}{\beta}}; A + \sqrt{\frac{D_A}{\beta}} \right),$$

где D_A – дисперсия коэффициента асимметрии,
β – доверительный уровень, выраженный в долях.
Дисперсия коэффициента асимметрии вычисляется по формуле

$$D_A = \frac{6(n-2)}{(n+1)(n+3)}.$$

Асимметрия находит применение, в частности, при исследовании формы распределения выборки. Подробнее см. главу «Проверка нормальности распределения». Доверительные интервалы оцениваемого коэффициента асимметрии вычислены в книге Иглина.

2.3.4. Эксцесс

Эксцесс характеризует форму статистического распределения. Если эксцесс больше нуля, то форма кривой распределения островершинная по сравнению с кривой плотности нормального распределения. Если эксцесс меньше нуля, то форма кривой распределения плосковершинная по сравнению с кривой плотности нормального распределения. Эксцесс выборочной совокупности вычисляется по уточненной формуле:

$$E = \frac{n(n+1)}{(n-1)(n-2)(n-3)} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{\sigma} \right)^4 - \frac{3(n-1)^2}{(n-2)(n-3)},$$

где n – численность выборки,
x_i, $i = 1,2,...,n$ – значения вариант выборки,

$\bar{x}$ – выборочное среднее значение,
σ – стандартное отклонение.
Вычисление доверительного интервала оцениваемого эксцесса производится по формуле:

$$I_E = \left(E - \sqrt{\frac{D_E}{\beta}}; E + \sqrt{\frac{D_E}{\beta}} \right),$$

где D_E – дисперсия эксцесса,
β – доверительный уровень, выраженный в долях.
Дисперсия эксцесса вычисляется по формуле

$$D_E = \frac{24n(n-2)(n-3)}{(n+1)^2(n+3)(n+5)}.$$

Эксцесс находит применение, в частности, при исследовании формы распределения выборки. Подробнее см. главу «Проверка нормальности распределения». Доверительные интервалы оцениваемого эксцесса вычислены в книге Иглина.

2.3.5. Коэффициент вариации

Коэффициент вариации представляет собой характеристику рассеяния случайной величины. Он показывает, какой процент составляет стандартное отклонение от среднего значения. Коэффициент вариации используется для установления степени выравненности совокупности по тому или иному признаку. Коэффициент вариации вычисляется по формуле:

$V = \frac{\sigma}{\bar{x}}$ в долях (выдается программой) или

$\vartheta = \frac{\sigma}{\bar{x}} \cdot 100\%$ в процентах,

где σ – стандартное отклонение,
$\bar{x}$ – выборочное среднее значение.
Для вычисления доверительного интервала оцениваемого коэффициента вариации применяется формула:
$I_V = (V - \Psi((1 + \beta) / 2)d; V + \Psi((1 + \beta) / 2)d)$,
где Ψ(.) – обратная функция стандартного нормального распределения,
β – доверительный уровень, выраженный в долях,
d – величина, рассчитанная по формуле

$$d = \sqrt{\frac{1}{n}\left(V^4 - \frac{V^2}{4} + \frac{m_4}{4D\bar{x}^2} - \frac{m_3}{\bar{x}^3} \right)},$$

где n – численность выборки,
D – выборочная дисперсия,
m_3 – третий центральный выборочный момент, вычисляемый по формуле

$$m_3 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^3,$$

m_4 – четвертый центральный выборочный момент, вычисляемый по формуле

$$m_4 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^4,$$

где x_i, $i = 1,2,...,n$ – значения вариант выборки.

О вычислении доверительных интервалов оцениваемого коэффициента вариации см. статью Орлова, книгу Мюллера с соавт. Некоторыми авторами коэффициент вариации применяется при проверке репрезентативности (оценке достаточной численности) выборки.

2.3.6. Минимум и максимум

Программой выводятся значения минимальной и максимальной вариант выборки:
x_{max} – значение максимальной варианты выборки,
x_{min} – значение минимальной варианты выборки.

2.3.6.1. Размах выборки

Размах выборки (размах вариации, амплитуда ряда) характеризует степень разброса данных в абсолютных числах. Выборочный размах – это разность между максимумом и минимумом вариант выборки. Вычисление размаха количественной выборки производится по формуле:
$R = x_{max} - x_{min}$,
где x_{max} – значение максимальной варианты выборки,
x_{min} – значение минимальной варианты выборки.

2.3.7. Медиана

Существует два типичных определения медианы. Энциклопедия «Вероятность и математическая статистика» определяет медиану случайной величины X как любое число m такое, что $P\{X \geq m\} \geq ½$ и $P(X \leq m\} \leq ½$. Математический энциклопедический словарь определяет медиану m непрерывно распределенной случайной величины X со строго монотонной функцией распределения $F(x)$ как единственный корень уравнения $F(m) = ½$. Алгоритм определения выборочной медианы количественной выборки, реализованный в настоящей программе, все источники определяют следующим образом. Для вычисления медианы эмпирической количественной выборки x_i, $i = 1,2,...,n$, численностью n сначала строится интервальный вариационный ряд y_i, $i = 1,2,...,n$, т. е. упорядоченная по возрастанию исходная выборка. Для нечетного $n = 2k + 1$ медианой будет варианта с номером k. Для четного $n = 2k$ медианой будет полусумма вариант с номерами k и $k + 1$.
Приведенный алгоритм может применяться также и для порядковой выборки нечетной численности. Для порядковой выборки четной численности некоторые авторы рассматривают левую медиану – варианту вариационного ряда с номером k – и правую медиану – варианту вариационного ряда с номером $k + 1$ – ввиду того, что для порядковой шкалы измерения операция деления не определена. Данные вычисления производятся и выводятся в разделе «Медиана множества». О шкалах измерения см. главу «Введение».
Доверительный интервал оцениваемой медианы задается формулой
$I_m = (y_{c+1}; y_{n-c})$,
где c – параметр, вычисляемый по формуле
$c = [n / 2 - \Psi((1 + \beta) / 2)n^{1/2} / 2]$,
где [.] – целая часть числа,
$\Psi(.)$ – обратная функция стандартного нормального распределения,
β – доверительный уровень, выраженный в долях.
Некоторые исследователи предпочитают медиану среднему значению (для шкалы измерения, в котором данный показатель имеет смысл), считая ее более точной оценкой меры положения выборки.

2.3.7.1. Оценка медианы на основе теории множеств

Рассмотрим выборочный показатель, представляющий собой варианту выборки,

равноудаленную от всех остальных вариант этой же эмпирической выборки. Данный показатель называется медианой множества (далее – медианой). При этом смысловое наполнение термина «равноудаленная» определяется шкалой измерения выборки. Попытаемся дать универсальное определение медианы, не зависящее ни от шкалы измерения эмпирических данных, ни от их размерности. Пусть имеется множество реализаций некоторой случайной величины, представляющее собой случайную эмпирическую выборку $\{X_1,X_2,\ldots,X_n\}$, где X_i, $i = 1,2,...,n$ – варианты, скалярные или векторные. Обозначим через $d(X, X_i)$ расстояние между произвольной реализацией случайной величины X и величиной X_i, $i = 1,2,...,n$. Основное требование к данному расстоянию состоит в его допустимости в используемой шкале измерения эмпирической выборки. Определим медиану как решение оптимизационной задачи. Медианой будет случайная величина X, удовлетворяющая условию

$$\mu\{X_1, X_2,..., X_n\} = \arg\min_{X \in D} \sum_{i=1}^{n} d(X, X_i),$$

где D – выборочное пространство реализаций случайной величины X.
Иначе, медианой множества является одна из вариант X_i, $i = 1,2,...,n$, удовлетворяющая данному условию.
Поиск медианы множества не вызывает вычислительной сложности в любой шкале измерения и может производиться на основе формального применения представленного определения. Для количественной выборки медиану множества можно найти, построив эмпирическую функцию распределения, подобно тому, как это сделано в главе «Непараметрическая статистика».
Показатель, вычисленный в настоящем разделе, может применяться как для количественных, так и для порядковых выборок. В случае количественной выборки нечетной численности показатель совпадает с обычной медианой.
Для порядковой выборки четной численности некоторые авторы рассматривают левую медиану – варианту вариационного ряда с номером k – и правую медиану – варианту вариационного ряда с номером $k + 1$ – ввиду того, что для порядковой шкалы измерения операция деления не определена. Данные показатели выводятся программой.

О вычислении точечной и интервальной оценки медианы см. статью Орлова, книгу Холлендера с соавт., монографию Кормена с соавт. (с. 240). Вычисление медианы ранжировок (медианы Кемени) производится в главе «Обработка экспертных оценок».

2.3.7.2. Псевдомедиана

Пусть вычислено $m = n(n + 1) / 2$ значений $w_1 \le w_2 \le ... \le w_m$ величин $(x_i + x_j) / 2$, $i \le j$; $i = 1,2,...,n$; $j = 1,2,...,n$, где x_i, x_j, $i = 1,2,...,n$; $j = 1,2,...,n$ – значения вариант исходной количественной выборки. Тогда медиана μ полученной выборки w_i, $i = 1,2,...,m$, называется псевдомедианой (оценкой Ходжеса–Лемана).
Итак, для вычисления медианы полученной выше количественной выборки w_i, $i = 1,2,...,m$, численностью m сначала строится интервальный вариационный ряд y_i, $i = 1,2,...,m$, т. е. упорядоченная по возрастанию выборка. Для нечетного m медианой является варианта полученного интервального вариационного ряда, имеющая порядковый номер $(m + 1) / 2$. Для четного m медиана равна среднему значению двух средних вариант. Утверждается, что если распределение симметрично, выборочные оценки медианы и псевдомедианы совпадают.
Доверительный интервал оцениваемой псевдомедианы задается формулой
$I_\mu = (y_{c+1}; y_{m-c})$

где c – параметр, вычисляемый по формуле

$$c = \left[\frac{n(n+1)}{4} - \Psi((1+\beta)/2) \left(\frac{n(n+1)(2n+1)}{24} \right)^{1/2} \right],$$

где [.] – целая часть числа,
n – численность выборки,
Ψ(.) – обратная функция стандартного нормального распределения,
β – доверительный уровень, выраженный в долях.

2.3.8. Квартили

Квартили, а также медиана (50% процентиль), обеспечивают разбиение упорядоченной количественной выборки (в виде вариационного ряда) на 4 подмножества равной численности. Вычисление данных показателей производится по правилам, принятым для вычисления медианы. Верхняя квартиль представляет собой 75% процентиль выборки. Нижняя квартиль представляет собой 25% процентиль выборки.
Укажем на одно полезное употребление квартилей. Тьюки предложил так называемый график «ящик с усами», представляющий собой совокупность следующих элементов:

- точки, обозначающей медиану,
- прямоугольника с верхней и нижней границами (если график изображается вертикально), соответствующими квартилям,
- отрезками, соответствующими максимуму и минимуму выборки.

Иногда изображается график «ящик с усами» для выборки, из которой уже исключены выбросы. В этом случае выбросы накладываются на график в виде точек. О методах исключения выбросов см. главу «Обработка выбросов».

2.3.8.1. Межквартильный размах

Как известно, квартили, а также медиана (50% процентиль), обеспечивают разбиение упорядоченной количественной выборки (в виде вариационного ряда) на 4 подмножества равной численности. Вычисление данных показателей производится по правилам, принятым для вычисления медианы.
Межквартильный (интерквартильный) размах выборки характеризует степень разброса данных в абсолютных числах. Выборочный межквартильный размах – это разность между верхней и нижней квартилями выборки, иначе 75% и 25% процентилями выборки.
Вычисление межквартильного размаха упорядоченной по возрастанию количественной выборки производится по формуле:
$f = f_{3/4} - f_{1/4}$,
где $f_{3/4}$ – значение верхней квартили выборки,
$f_{1/4}$ – значение нижней квартили выборки.
Утверждается, что межквартильный размах является более репрезентативной оценкой разброса значений выборки по сравнению с точечной оценкой стандартного отклонения. Точечная оценка стандартного отклонения для нормально распределенной совокупности может быть получена из межквартильного размаха как

$$\sigma = \frac{f}{2\Psi(0{,}75)} \approx 0{,}741301 f,$$

где Ψ(.) – функция, обратная функции стандартного нормального распределения,
Межквартильный размах находит применение в качестве основы одного из методов выявления аномальных наблюдений (выбросов), применяемых в главе «Обработка выбросов». Величина $f / 2$ используется как характеристика рассеяния и называется

семиинтерквартильной широтой.

2.3.9. Гистограмма

Гистограмма представляет собой дискретный или интервальный вариационный ряд (ряд распределения), полученный в результате группировки исходной эмпирической выборки, измеренной в порядковой или количественной шкале, по особым образом подобранным классовым интервалам. Данный вариационный ряд служит основой для многих статистических алгоритмов, таких, как глазомерный метод проверки нормальности распределения, установление типа распределения (как для дискретных, так и для непрерывных распределений), критерии типа хи–квадрат и других.
Имеется два пути практической группировки: задавшись границами классовых интервалов (классов) или задавшись их количеством, а затем вычислить границы. Во втором случае для вариационного ряда число классов равно числу градаций переменной, выбранному пользователем или вычисленному программой. При этом число классов дискретного вариационного ряда обычно равно числу градаций вариант выборки, измеренной в порядковой шкале. Для интервального вариационного ряда число классов задается пользователем на основе одного из применяемых правил, рассмотренных ниже.
Критерием правильности выбора количества классов считается верная передача типа распределения эмпирических частот данной выборочной совокупности. Если выбрано слишком мало классов, можно потерять характерную картину эмпирического распределения. При слишком подробном делении можно затушевать реальную картину распределения частот случайными отклонениями.
Инструмент «Гистограмма» позволяет пользователю, при его желании, задать число классов либо делает это автоматически. Выделяются несколько общеупотребительных способов вычисления числа классов для выборок умеренной численности. Применяемое правило Стержесса (Стургеса, Старджеса, Sturgess) основано на формуле
$k = 1{,}44 \ln n + 1$,
где k – число классов,
n – численность выборочной совокупности.
После решения вопроса о числе классов производится вычисление границ классовых интервалов и разнесение вариант исходной количественной выборки по классовым интервалам. Программа выводит число классов, размер классового интервала, середины классовых интервалов и количества вариант, попавших в данный класс, а также моду.
Недопустимо заменять гистограмму ломаной линией. Такая замена предполагает, что между ординатами существуют или могут существовать какие–то значения, чего на самом деле не имеет места.

Подробный обзор элементарных способов выбора числа классов см. в книге Новицкого с соавт. См. также статью Скотта (Scott).

2.3.9.1. Мода

Мода представляет собой значение переменной, при котором функция плотности распределения достигает максимального значения. Визуальным отображением эмпирической функции плотности эмпирического распределения является гистограмма (деленная на численность выборки), поэтому моду удобно рассчитать и вывести в разделе «Гистограмма». Имеет особенность расчет моды для группированных исходных данных. Помимо моды, вычисленной как обычно, программой дополнительно выдается значение моды, полученной непосредственно из группированных исходных данных, полагая, что в частном случае данная группировка и гистограмма – это практически одно и то же.

2.3.9.2. Оптимальное число классов

Очевидным критерием правильности выбора количества классов считается верная передача типа распределения эмпирических частот данной выборочной совокупности. Если выбрано слишком мало классов, можно потерять характерную картину эмпирического распределения. При слишком подробном делении можно затушевать реальную картину распределения частот случайными отклонениями. Большинство источников ограничиваются данными рекомендациями, предлагая различные эвристические формулы вычисления числа классовых интервалов. Выбор некоторого оптимального количества классов позволит не только верно визуально передать тип распределения, но и минимизировать существенные потери информации, содержащейся в исходных данных, которая происходит при фактическом понижении исходной количественной шкалы до шкалы номинальной. О шкалах измерения и их преобразовании см. главу «Введение».

2.3.9.2.1. Метод оптимизации числа классов

Предлагается алгоритм, дающий математическое обоснование критерия, с формулировки которого начат данный раздел. Под оптимальным числом классов мы понимаем минимально допустимое, но верно передающее распределение исходной случайной величины. Алгоритм состоит из следующих шагов.

1. Пусть дана количественная эмпирическая выборка x_i, $i = 1,2,...,n$.
2. Берется минимальное имеющее смысл число классов $k = 2$.
3. Производится классификация, в результате которой получается вариационный ряд y_j, $j = 1,2,...,k$.
4. По вариационному ряду восстанавливается выборка z_i, $i = 1,2,...,n$, фактически представляющая собой огрубленную до номинальной шкалы с числом градаций, равным k, исходную выборку.
5. Сравниваются функции распределения исходной выборки x_i, $i = 1,2,...,n$, и выборки z_i, $i = 1,2,...,n$. Может использоваться один из тестов, предназначенных для сравнения двух эмпирических функций распределения. В программе применяется критерий Койпера, аналогичный представленному в главе «Непараметрическая статистика».
6. Контролируется P–значение статистики критерия, вычисленного на шаге 5. Первое же значение k, при котором различия окажутся незначимы (в программе $p \geq 0,05$), будет оптимальным числом классов – на этом процесс завершается (процесс завершается также при достижении $k = n$). Иначе, при установлении значимости $p < 0,05$, значение k увеличивается на 1 и осуществляется переход к шагу 3.

Значение k, полученное в результате работы алгоритма, дает необходимую объективную нижнюю оценку числа классовых интервалов равной ширины, при котором тип распределения исходной случайной величины передается верно. В дальнейших расчетах можно уверенно брать любое число классов, равное или немного превышающее данную величину.

Преимуществом предложенного алгоритма является возможность использования для сравнения распределений: исходного и гистограммы – различных метрик, которые зависят от применяемого критерия, и различных уровней значимости. О критериях сравнения функций распределения см. главу «Непараметрическая статистика».

2.3.9.2.2. Метод Шимазаки–Шиномото

Метод предложен Шимазаки (Shimazaki) и Шиномото (Shinomoto). Оригинальный метод оптимизирует ширину классового интервала, поэтому мы немного видоизменили схему метода с целью оптимизации числа классов (данные параметры в случае классовых интервалов равной ширины являются однозначно взаимозависимыми).

- Пусть дана количественная эмпирическая выборка x_i, $i = 1,2,...,n$.
- Берется минимальное имеющее смысл число классов $k = 2$.
- Вычисляется соответствующая ширина классового интервала $\Delta(k)$.
- Производится классификация, в результате которой получается вариационный ряд y_j, $j = 1,2,...,k$. По вариационному ряду вычисляются параметры: среднее значение $\bar{y} = \frac{1}{k}\sum_{j=1}^{k} y_j$ и дисперсия $D = \frac{1}{k}\sum_{j=1}^{k}\left(y_j - \bar{y}\right)^2$.
- Вычисляется функционал («функция стоимости», в терминологии авторов) $C(\Delta) = \frac{2\bar{y} - D}{\Delta^2}$.
- Значение k увеличивается на 1 и осуществляется переход к шагу 3. Процесс повторяется до достижения $k = n$.
- Оптимальным числом классов будет то число, которое обеспечивает минимум функционалу $C(\Delta)$.

Программа выводит все упомянутые параметры: оптимальное число классов, а также зависимость функции стоимости от числа классов и ширины классового интервала. Упомянутыми авторами сконструированы и другие функционалы.

2.3.10. Доля

Для бинарной выборки оценка доли (распространенности, binomial proportion), т. е. количества вариант – «случаев», отнесенное к численности выборки, может быть рассчитана по формуле максимального правдоподобия:

$$\hat{p} = \frac{m}{n},$$

где m – число случаев,
n – численность выборки.
Доверительные интервалы для оцениваемой доли могут вычисляться различными методами. Методы, реализованные в программе, представлены ниже.
Стандартно доверительный интервал оцениваемой доли в источниках рассчитывается по классической формуле Вальда (Wald interval)

$$I_{\hat{p}} = \left(\hat{p} - \Psi((1+\beta)/2)\sqrt{D_{\hat{p}}}; \hat{p} + \Psi((1+\beta)/2)\sqrt{D_{\hat{p}}}\right),$$

где $\Psi(.)$ – обратная функция стандартного нормального распределения,
β – доверительный уровень, выраженный в долях,
$D_{\hat{p}}$ – дисперсия доли.
Доверительные интервалы оцениваемой доли в программе могут рассчитываться по «точным» формулам Клоппера–Пирсона (Clopper–Pearson interval). При этом нижняя граница доверительного интервала оцениваемой доли считается как

$$L_P = \left[1 + \frac{n - m + 1}{m \cdot F^{-1}_{2m,2(n-m+1)}(1-(1-\beta)/2)}\right]^{-1},$$

где $F^{-1}_{\dots}(.)$ – обратная функция *F*–распределения.
Верхняя граница доверительного интервала оцениваемой доли считается как

$$H_P = \left[1 + \frac{n-m}{(m+1)\cdot F^{-1}_{2(m+1),2(n-m)}((1-\beta)/2)}\right]^{-1}.$$

Доверительный интервал оцениваемой доли в программе рассчитывается также по формуле Агрести–Коула (Agresti–Coull interval, иначе называемый уточненным методом Вальда)

$$I_{\tilde{p}} = \left(\tilde{p} - \Psi((1+\beta)/2)\sqrt{D_{\tilde{p}}}; \tilde{p} + \Psi((1+\beta)/2)\sqrt{D_{\tilde{p}}}\right),$$

где $\tilde{p}$ – скорректированное значение доли,
$D_{\tilde{p}}$ – значение дисперсии скорректированной доли.
Скорректированное значение доли рассчитывается по формуле

$$\tilde{p} = \frac{m+2}{n+4}.$$

Дисперсия скорректированной доли вычисляется по формуле

$$D_{\tilde{p}} = \frac{\tilde{p}\cdot(1-\tilde{p})}{n}.$$

Доверительный интервал оцениваемой доли в программе рассчитывается также по формуле Вилсона (Wilson interval)

$$I_{\tilde{p}} = \left(\tilde{p} - \Psi((1+\beta)/2)\sqrt{D_{\tilde{p}}}; \tilde{p} + \Psi((1+\beta)/2)\sqrt{D_{\tilde{p}}}\right),$$

где $\tilde{p}$ – скорректированное значение доли,
$D_{\tilde{p}}$ – значение дисперсии скорректированной доли.
В дальнейшей записи для простоты обозначим $k = \Psi((1 + \beta) / 2)$.
Тогда, с учетом введенного обозначения, скорректированное значение доли рассчитывается по формуле

$$\tilde{p} = \frac{m + k^2/2}{n + k^2}.$$

Дисперсия скорректированной доли вычисляется по формуле

$$D_{\tilde{p}} = \frac{n}{(n+k^2)^2}\cdot\left(\tilde{p}\cdot(1-\tilde{p}) + \frac{k^2}{4n}\right)$$

Обзоры методов (их несколько десятков) оценки доли см. в статьях Льюис (Lewis), Льюис с соавт., Брауна (Brown) с соавт. См. оригинальные статьи Агрести (Agresti) с соавт., Клоппера (Clopper) с соавт., а также работы Пирес (Pires) с соавт., Болбоака (Bolboaca) с соавт., Друган (Drugan) с соавт., доклады Пирес, Сауро (Sauro) с соавт., приложение Хромова–Борисова к книге Кайданова, монографии Флейс, Флейс (Fleiss), Флейс с соавт.

2.3.10.1. Ошибка доли

Ошибка доли вычисляется по формуле

$$m_{\hat{p}} = \sqrt{D_{\hat{p}}},$$

где $D_{\hat{p}}$ – дисперсия доли.
Исследователи иногда задаются вопросом, как рассчитать процент и ошибку процента случаев от численности выборки. Идея заключается в том, что вычисления в данном случае

производятся по стандартным формулам для доли. Результат же переводится в проценты следующим образом. Процент вычисляется как $100 \cdot \hat{p}$, где $\hat{p}$ – оценка доли. Ошибка процента вычисляется как $100 \cdot m_{\hat{p}}$.

2.3.10.2. Дисперсия доли

Дисперсия доли может быть вычислена по формуле

$$D_{\hat{p}} = \frac{\hat{p} \cdot (1 - \hat{p})}{n},$$

где $\hat{p}$ – выборочная оценка доли,
n – численность выборки.

2.3.11. Показатель точности опыта

Показатель точности опыта, иначе – показатель точности определения среднего значения, выражает величину ошибки среднего значения в процентах от самого среднего. Точность считается удовлетворительной, если величина показателя не превышает 5%, а при значениях, больших 5%, рекомендуется увеличить число наблюдений или повторений. Иногда величину показателя точности можно уменьшить, если повысить точность измерений параметров объектов опыта. Показатель точности опыта вычисляется по формуле:

$p = \frac{m}{\bar{x}}$ в долях или

$P = \frac{m}{\bar{x}} \cdot 100\%$ в процентах (выдается программой),

где m – стандартная ошибка,
$\bar{x}$ – выборочное среднее значение.

Очевидно, показатель точности определения среднего значения – это именно то, что имеют в виду исследователи в медико–биологических науках, указывая в публикациях после $M \pm m$ через запятую, к примеру, выражение $p < 0{,}05$, называя его достоверностью. Хотя это определение в данном случае не совсем верно, но оно используется традиционно. Для того чтобы читатель понял, что именно имеет в виду исследователь, в работе следует расшифровывать абсолютно все используемые математические обозначения и аббревиатуры, не полагаясь на то, что данные показатели общеупотребительны. Сказанное относится и к остальным применяемым в исследовании статистическим показателям.

2.3.12. Достаточная численность выборки

Анализ репрезентативности выборки (иначе – способности выборки адекватно представить всю генеральную совокупность, популяцию) особенно важен на начальном этапе исследований, когда численность генеральной совокупности неизвестна в принципе, но уже известны некоторые параметры опыта, позволяющие оценить репрезентативность.
Достаточная численность выборки может быть рассчитана как для количественных, так и для качественных выборок.
В программе представлен метод вычисления достаточной численности количественной выборки, основанной на формуле

$$n = \frac{t^2_{(1+\beta)/2}\sigma^2}{\Delta^2},$$

где $t_{(1+\beta)/2}$ – значение обратной функции t–распределения Стьюдента с числом степеней

свободы ∞ и параметром (1 + β) / 2,
β – доверительный уровень, выраженный в долях, к примеру 0,95 (что соответствует 95%),
σ – выборочная оценка стандартного отклонения, к примеру, 50 рублей,
Δ – абсолютная погрешность определения среднего арифметического значения, к примеру, 5 рублей.
Абсолютная погрешность вводится в именованных числах, т. е. в тех же единицах измерения, что и варианты выборки. Например, при подсчете количества неделимых объектов исследования (например, избирательных бюллетеней) абсолютная погрешность может быть установлена равной 1.
В литературе представлена также формула, аналогичная приведенной выше, за исключением того, что используется значение не обратной функции распределения Стьюдента, а обратной функции нормального распределения

$$n = \frac{\Psi^2((1+\beta)/2)\sigma^2}{\Delta^2},$$

где Ψ(.) – обратная функция стандартного нормального распределения.
Метод (не представленный в программе) вычисления достаточной численности качественной выборки основан на формуле

$$n = \frac{\Psi^2((1+\beta)/2)\cdot\hat{p}\cdot(1-\hat{p})}{\Delta^2},$$

где $\hat{p}$ – выборочная оценка доли, к примеру, 0,35,
Δ – абсолютная погрешность определения доли, к примеру, 0,05.
Если известна численность популяции *N*, а вычисленная достаточная численность оказывается 10% и более от численности популяции, то достаточная численность выборки должна быть скорректирована в соответствии с формулой

$$n' = \frac{nN}{N+n-1}.$$

Большинство данных формул не реализованы в программе по причине сложности учета многообразных форм представления исходных данных, однако при необходимости вычислить достаточную численность выборки не представит никакой сложности.

О вычислении достаточной численности см. монографии Зайцева, Малхотра, Девятко, Голубкова, Лванга (Lwanga) с соавт., Чау (Chow) с соавт., статьи Делл (Dell) с соавт., Кук (Cook) с соавт., Инг (Eng). Вычисление численности для различных статистических методов и для исходных данных в различных шкалах см. в статьях Кэмпбелл (Campbell) с соавт., Чен (Chan), Бонетт (Bonett) с соавт., Вальтер (Walter) с соавт., отчете Калвани (Kalwani) с соавт.

2.3.13. Критерий Аббе

Для проверки, извлечена ли выборка случайно из нормальной генеральной совокупности либо, с другой точки зрения, независимы ли одинаково нормально распределенные случайные величины, можно воспользоваться критерием Аббе. Статистика критерия (отношение фон Ноймана, von Neuman Ratio) может быть подсчитана по формуле:

$$\gamma = \frac{\sum_{i=1}^{n-1}\left(x_{i+1}-x_i\right)^2}{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2},$$

где *n* – численность выборки,

x_i, i = 1,2,...,n – значения вариант выборки,
$\bar{x}$ – выборочное среднее значение.
В литературе (и в программе) под названием статистики Аббе фигурирует величина $q = \gamma / 2$.
При этом P–значение может быть вычислено с помощью модифицированной статистики

$$T = (q-1)\sqrt{\frac{2n+1}{2-(q-1)^2}},$$

которая для больших выборок распределена приближенно нормально по закону $N(0,1)$.

Распределение статистики γ изучил фон Нойманн (von Neumann). Аппроксимацию P–значений предложили Бингхэм (Bingham) с соавт. В отчете Кемпбелла (Campbell) с соавт. указаны аппроксимации для больших выборок. Программу вычисления P–значения опубликовал Нельсон (Nelson). См. монографии Браунли, Петровича с соавт., Айвазяна с соавт., справочник Большева с соавт., статьи Хэрта (Hart), Лемешко.

2.3.14. Формулы для сгруппированных выборок

Группировка выборок может быть как следствием их естественного исходного представления (номинальная либо порядковая шкала измерения), так и результатом понижения количественной шкалы измерения до порядковой или номинальной шкалы. Более подробная информация о шкалах измерения и их преобразовании приводится в «Введение».
Исходные данные в группированном виде могут, к примеру, иметь следующий вид (пусть верхняя строка – оценка за курсовую работу, а нижняя – число студентов, получивших данную оценку):

b_i, i = 1,2,...,5	1	2	3	4	5
ν_i, i = 1,2,...,5	0	1	10	19	25

Здесь обозначено:
b_i, i = 1,2,...,k – середины классовых интервалов (для количественных выборок) либо значения для порядковых и номинальных выборок,
ν_i, i = 1,2,...,k – частоты наблюдаемых случаев в классах, иначе – численности классов,
k – число классов (групп).
Для вычислений выборочных показателей используются формулы для среднего значения, среднего отклонения и дисперсии (несмещенная оценка), соответственно, в следующей форме:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{k} b_i \nu_i,$$

$$\hat{x} = \frac{1}{n}\sum_{i=1}^{k} |b_i - \bar{x}| \nu_i,$$

$$D = \frac{1}{n-1}\sum_{i=1}^{k} (b_i - \bar{x})^2 \nu_i$$

либо в эквивалентной форме

$$D = \frac{1}{n-1}\left[\sum_{i=1}^{k} b_i^2 \nu_i - \frac{1}{n}\left(\sum_{i=1}^{k} b_i \nu_i\right)^2\right],$$

где n – общее число наблюдений, вычисляемое по формуле

$$n = \sum_{i=1}^{k} \nu_i.$$

Логика вычислений заключается в суммировании по числу классов и домножении каждого

выражения под знаком суммы на соответствующую данному классу частоту. На основании данной информации записать эквивалентные формулы для вычисления других статистических показателей не составит труда.
Статистические показатели, в формулы вычислений которых не входят значения вариант выборки, вычисляются по тем самым формулам и для негруппированных, и для сгруппированных данных.
В программе реализован полный комплект вычислений описательной статистики для сгруппированных данных. Однако показанные в разделе формулы приводятся только для полноты, т.к. в программе они не применяются. В программе реализован более удобный в вычислительном отношении прием – сначала из сгруппированных данных «восстанавливается» исходная выборка, а затем все расчеты проводятся в обычном режиме.

Более подробная информация о наименованиях, использованных выше, приводится в справочнике Гайдышева.

Список использованной и рекомендуемой литературы

1. Agresti A., Coull B. Approximate is better than «exact» for interval estimation of binomial proportions // The American Statistician, 1998, vol. 52, pp. 119–126.
2. Armitage P., Berry G., Matthews J.N.S. Statistical methods in medical research. – Oxford, UK: Blackwell Science, 2001.
3. Bingham C., Nelson L.S. An approximation for the distribution of the von Neuman Ratio // Technometrics, 1981, vol. 23, pp. 285–288.
4. Bolboaca S.-D., Achimas Cadariu A.B. Binomial distribution sample confidence intervals estimation 2. Proportion–like medical key parameters // Leonardo Electronic Journal of Practices and Technologies, July–December, 2003, no. 3, pp. 75–110.
5. Bonett D.G., Wright T.A. Sample size requirements for estimating Pearson, Kendall and Spearman correlations // Psychometrika, March 2000, vol. 65, no. 1, pp. 23–28.
6. Brown L., Cai T., DasGupta A. Confidence intervals for a binomial proportion and asymptotic expansions // The Annals of Statistics, 2002, vol. 30, pp. 160–201.
7. Brown L., Cai T., DasGupta A. Interval estimation for a binomial proportion // Statistical Science, 2001, vol. 16, pp. 101–133.
8. Campbell K. Fundamental data analyses for measurement control / K. Kempbell, G.L. Barlich, B. Fazal et al. // Technical Report LA–10811–MS. – Los Asamos, NM: Los Alamos National Laboratory, 1987.
9. Campbell M.J., Julious S.A., Altman D.G. Estimating sample sizes for binary, ordered categorical, and continuous outcomes in two group comparisons // BMJ, 28 October 1995, vol. 311, pp. 1145–1148.
10. Chan Y.H. Randomised controlled trials (RCTs) – sample size: The magic number? Singapore Medical Journal, April 2003, vol. 44, no. 4, pp. 172–174.
11. Chow S.-C., Shao J., Wang H. Sample size calculations in clinical research. – Boca Raton, FL: Chapman & Hall / CRC, 2008.
12. Clopper C.J., Pearson E.S. The use of confidence or fiducial limits illustrated in the case of binomial // Biometrika, December 1934, vol. 26, no. 4, pp. 404–413.
13. Cook R.J., Sackett D.L. The number needed to treat: a clinically useful measure of treatment effect // BMJ, 18 February 1995, vol. 310, pp. 452–454.
14. Dell R.B., Holleran S., Ramakrishnan R. Sample size determination // ILAR Journal, 2002, vol. 43, no. 4.
15. Diekhoff G. Statistics for the social and behavioral sciences: Univariate, bivariate, multivariate. – Dubuque, IA: WM. C. Brown Company Publishers Dubuque, 1992.

16. Drugan T. Binomial distribution sample confidence intervals estimation. 1. Sampling and medical key parameters calculation / T. Drugan, S.–D. Bolboaca, L. Jantschi et al. // Leonardo Electronic Journal of Practices and Technologies, July–December 2003, no. 3, pp. 45–74.
17. Eng J. Sample size estimation: A glimpse beyond simple formulas // Radiology, 2004, vol. 230, no. 3, pp. 606–612.
18. Eng J. Sample size estimation: How many individuals should be studied? // Radiology, 2003, vol. 227, no. 2, pp. 309–313.
19. Fielding A. Determining adequate sample size: A statistical consultant's advice in a legal brief // Teaching Statistics, 1996, vol. 18, no. 1, pp. 6–9.
20. Fisher R.A. Statistical tables for biological, agricultural and medical research / Ed. by R.A. Fisher, F. Yates. – Edinburgh: Oliver and Boyd, 1963.
21. Fleiss J.L. Statistical methods for rates and proportions. – New York, NY: John Wiley & Sons, 1981.
22. Fleiss J.L., Levin B., Paik M.C. Statistical methods for rates and proportions. – New York, NY: John Wiley & Sons, 2003.
23. Galassi M. GNU Scientific Library Reference Manual / M. Galassi, J. Davies, J. Theiler et al. – Network Theory, 2005.
24. Gonick L., Smith W. The cartoon guide to statistics. – New York, NY: Harper Perennial, 1993.
25. Good P.I., Hardin J.W. Common errors in statistics, and how to avoid them. – New York, NY: John Wiley & Sons, 2003.
26. Goodman S.N., Berlin J.A. The use of predicted confidence intervals when planning experiments and the misuse of power when interpreting results // Annals of Internal Medicine, 1 August 1994, vol. 121, no. 3, pp. 200–206.
27. Greenhalgh T. How to read a paper: Statistics for the non–statistician. I: Different types of data need different statistical tests // BMJ (British Medical Journal), 1997, vol. 315, pp. 364–366.
28. Greenhalgh T. How to read a paper: Statistics for the non–statistician. II: «Significant» relations and their pitfalls // BMJ (British Medical Journal), 1997, vol. 315, pp. 422–425.
29. Grimm L.G., Yarnold P.R. Reading and understanding more multivariate statistics. – American Psychological Association, 2000.
30. Guyatt G. Basic statistics for clinicians: 2. Interpreting study results: confidence intervals. / G. Guyatt, R. Jaeschke, N. Heddle et al. // Canadian Medical Association Journal, January 1995, vol. 152, no. 2, pp. 169–173.
31. Hahn G.J., Meeker W.Q. Statistical intervals: A guide for practitioners. – New York, NY: John Wiley & Sons, 1991.
32. Hart B.I. Significance levels for the ratio of the mean square successive difference to the variance // The Annals of Mathematical Statistics, 1942, vol. 13, no. 4, pp. 445–447.
33. Kalwani M.U., Morrison D.G. Estimating the proportion of «always buy» and «never buy» consumers: A likelihood ratio test with sample size implications. – Cambridge, MA: M.I.T. Alfred P. Sloan School of Management, 1976.
34. Kerlinger F.N. Foundation of behavioral research. – New York, NY: Holt, Rinehart & Winston, 1986.
35. Le C.T. Introductory biostatistics. – New York, NY: John Wiley & Sons, 2003.
36. Lewis J.R. Evaluation of procedures for adjusting problem–discovery rates estimated from small samples // The International Journal of Human–Computer Interaction, 2001, vol. 13, no. 4, pp. 445–479.
37. Lewis J.R., Sauro J. When 100% really isn't 100%: Improving the accuracy of small–sample

estimates of completion rates // Journal of Usability Studies, May 2006, vol. 1, no. 3, pp. 136–150.
38. Lucy D. Introduction to statistics for forensic scientists. – Chichester, UK: John Wiley & Sons, 2005.
39. Lwanga S.K., Lemeshow S. Sample size determination in health studies. A practical manual. – Geneva: World Health Organization, 1991.
40. Mosteller F., Bailar J.C. Medical uses of statistics. – Boston, MA: NEJM Books, 1992.
41. Nelson L.S. The mean square successive difference test automated // Journal of Quality Technology, October 1998, vol. 30, no. 4, pp. 401–402.
42. Pires A.M. Confidence intervals for a binomial proportion: comparison of methods and software evaluation // Proceedings of the Conference CompStat 2002 – Short Communications and Posters / Ed. by S. Klinke, P. Ahrend, L. Richter, 2002.
43. Pires A.M., Amado C. Interval estimators for a binomial proportion: Comparison of twenty methods // REVSTAT – Statistical Journal, June 2008, vol. 6, no. 2, pp. 165–197.
44. Salvatore D., Reagle D. Statistics and econometrics. – New York, NY: McGraw–Hill, 2003.
45. Santiago Medina L., Zurakowski L. Measurement variability and confidence intervals in nedicine: Why should radiologists care? // Radiology, 2003, vol. 226, no. 2, pp. 297–301.
46. Sauro J., Lewis J.R. Estimating completion rates from small samples using binomial confidence intervals: Comparisons and recommendations // Proceedings of the Human Factors and Ergonomics Society Annual Meeting (HFES 2005) Orlando, FL, 2005.
47. Scott D.W. Optimal and data–based histograms // Biometrika, 1979, vol. 66, no. 3, pp. 605–610.
48. Shimazaki H., Shinomoto S. A method for selecting the bin size of a time histogram // Neural Computation, 2007, vol. 19, no. 6, pp. 1503–1527.
49. Shimazaki H., Shinomoto S. A recipe for optimizing a time–histogram // Neural Information Processing Systems, 2007, vol. 19, pp. 1289–1296.
50. Sonnad S.S. Describing data: Statistical and graphical methods // Radiology 2002, vol. 225, no. 3, pp. 622–628.
51. Tiku M.L., Akkaya A.D. Robust estimation and hypothesis testing. – New Delhi: New Age International, 2004.
52. Von Neumann J. Distribution of the ratio of the mean square successive difference to the variance // The Annals of Mathematical Statistics, 1941, vol. 12, no. 4, pp. 367–395.
53. Walter S.D., Yao X. Effect sizes can be calculated for studies reporting ranges for outcome variables in systematic reviews // Journal of Clinical Epidemiology, August 2007, vol. 60, no. 8, pp. 849–852.
54. Wand M.P. Data–based choice of histogram bin width // The American Statistician, February 1997, vol. 51, no. 1, pp. 59–64.
55. Wilcox R.R. Fundamentals of modern statistical methods: Substantially improving power and accuracy. – New York, NY: Springer, 2001.
56. Zhou X.–H., Dinh P. Nonparametric confidence intervals for the one– and two–sample problems // UW Biostatistics Working Paper Series. Working Paper 233. September 14, 2004.
57. Айвазян С.А. Прикладная статистика: Основы моделирования и первичная обработка данных. Справочное издание / С.А. Айвазян, И.С. Енюков, Л.Д. Мешалкин. – М.: Финансы и статистика, 1983.
58. Александров В.В., Шнейдеров В.С. Обработка медико–биологических данных на ЭВМ. – Л.: Медицина, 1982.
59. Белова Е.Б. Компьютеризованный статистический анализ для историков. Учебное пособие / Е.Б. Белова, Л.И. Бородкин, И.М. Гарскова и др. – М.: МГУ, 1999.

60. Большев Л.Н., Смирнов Н.В. Таблицы математической статистики. – М.: Наука, 1983.
61. Боровков А.А. Математическая статистика. Оценка параметров. Проверка гипотез. – М.: Наука, 1984.
62. Браунли К.А. Статистическая теория и методология в науке и технике. – М.: Наука, 1977.
63. Вентцель Е.С. Теория вероятностей. – М.: Высшая школа, 1999.
64. Вероятность и математическая статистика. Энциклопедия. – М.: Научное издательство «Большая Российская энциклопедия», 1999.
65. Власов В.В. Эпидемиология: Учебное пособие для вузов. – М.: Издательский дом «ГЭОТАР–МЕД», 2004.
66. Власов В.В. Эффективность диагностических исследований. – М.: Медицина, 1988.
67. Гайдышев И. Анализ и обработка данных: Специальный справочник. – СПб: Питер, 2001.
68. Гайдышев И.П. Статистика в публикациях // Гений ортопедии, 2005, № 4, с. 155–161.
69. Голубков Е.П. Маркетинговые исследования: теория, методология и практика. – М.: Издательство «Финпресс», 1998.
70. Гринхальх Т. Основы доказательной медицины. – М.: Издательский дом «ГЭОТАР–МЕД», 2004.
71. Гудман С.Н. На пути к доказательной биостатистике. Часть 1: обманчивость величины p // Международный журнал медицинской практики, 2002, № 1, с. 8–17.
72. Гудман С.Н. На пути к доказательной биостатистике. Часть 2: байесовский критерий // Международный журнал медицинской практики, 2002, № 2, с. 5–14.
73. Девятко И.Ф. Методы социологического исследования. – Екатеринбург: Издательство Уральского университета, 1998.
74. Дерффель К. Статистика в аналитической химии. – М.: Мир, 1994.
75. Джини К. Средние величины. – М.: Статистика, 1970.
76. Длин А.М. Математическая статистика в технике. – М.: Советская наука, 1958.
77. Доспехов Б.А. Методика полевого опыта (с основами статистической обработки результатов исследований). – М.: Агропромиздат, 1985.
78. Зайцев Г.Н. Математическая статистика в экспериментальной ботанике. – М.: Наука, 1984.
79. Зуева Л.П., Яфаев Р.Х. Эпидемиология: Учебник. – СПб: ООО «Издательство ФОЛИАНТ», 2005.
80. Иванов Ю.И., Погорелюк О.Н. Статистическая обработка результатов медико–биологических исследований на микрокалькуляторах по программам. – М.: Медицина, 1990.
81. Иглин С.П. Математические расчеты на базе MATLAB. – СПб: БХВ–Петербург, 2007.
82. Кайданов Л.З. Генетика популяций. – М.: Высшая школа, 1996.
83. Кобзарь А.И. Прикладная математическая статистика. Для инженеров и научных работников. – М.: ФИЗМАТЛИТ, 2006.
84. Кокрен У. Методы выборочного исследования. – М.: Статистика, 1976.
85. Конушин А. Устойчивые алгоритмы оценки параметров модели на основе случайных выборок // On–line журнал «Графика и мультимедиа», 2003, выпуск 3.
86. Кормен Е.Ч. Алгоритмы: построение и анализ / Е.Ч. Кормен, Ч.И. Лейзерсон, Р.Л. Ривест и др. – М.: Издательский дом «Вильямс», 2005.
87. Крянев А.В., Лукин Г.В. Математические методы обработки неопределенных данных. – М.: ФИЗМАТЛИТ, 2006.
88. Кудлаев Э.М., Орлов А.И. Вероятностно–статистические методы исследования в работах А.Н. Колмогорова // Заводская лаборатория. Диагностика материалов, 2003, т.

69, № 5, с. 55–61.
89. Кюн Ю. Описательная и индуктивная статистика. – М.: Финансы и статистика, 1981.
90. Лакин Г.Ф. Биометрия. – М.: Высшая школа, 1990.
91. Ланг Т. Двадцать ошибок статистического анализа, которые Вы сами можете обнаружить в биомедицинских статьях // Международный журнал медицинской практики, 2005, № 1, с. 21–31.
92. Леман Э. Теория точечного оценивания. – М.: Наука, 1991.
93. Лемешко С.Б. Критерий независимости Аббе при нарушении предположений нормальности // Измерительная техника, 2006, № 10, с. 9–14.
94. Леонов В.П., Ижевский П.В. Об использовании прикладной статистики при подготовке диссертационных работ по медицинским и биологическим специальностям // Бюллетень ВАК РФ, 1997, № 5, с. 56–61.
95. Ллойд Э. Справочник по прикладной статистике. В 2–х т. Т. 2. / Под ред Э. Ллойда, У. Ледермана, С.А. Айвазяна и др. – М.: Финансы и статистика, 1990.
96. Малхотра Н.К. Маркетинговые исследования и эффективный анализ статистических данных. – М.: Издательство «ДиаСофт», 2002.
97. Малхотра Н.К. Маркетинговые исследования. Практическое руководство. – М.: Издательский дом «Вильямс», 2002.
98. Новицкий П.В., Зограф И.А. Оценка погрешностей результатов измерений. – Л.: Энергоатомиздат, 1991.
99. Орлов А.И. Непараметрическое точечное и интервальное оценивание характеристик распределения // Заводская лаборатория. Диагностика материалов, 2004, т. 70, № 5, с. 65–70.
100. Петрович М.Л., Давидович М.И. Статистическое оценивание и проверка гипотез на ЭВМ. – М.: Финансы и статистика, 1989.
101. Плохинский Н.А. Достаточная численность выборки / В сб. Биометрический анализ в биологии. – М.: Издательство Московского университета, 1982, с. 152–157.
102. Прохоров Ю.В. Математический энциклопедический словарь / Гл. ред. Ю.В. Прохоров. – М.: Научное издательство «Большая Российская энциклопедия», 1995.
103. Рокицкий П.Ф. Биологическая статистика. – Мн.: Вышэйшая школа, 1973.
104. Сборник научных программ на Фортране. Выпуск 1. Статистика. – М.: Статистика, 1974.
105. Сидоренко Е.В. Методы математической обработки в психологии. – СПб.: ООО «Речь», 2001.
106. Солонин И.С. Математическая статистика в технологии машиностроения. – М.: Машиностроение, 1972.
107. Тейлор Дж. Введение в теорию ошибок. – М.: Мир, 1985.
108. Технический отчет ISO/TR 10017:2003. Руководство по статистическим методам применительно к ISO 9001:2000. – М.: ВНИИКИ, 2004.
109. Тутубалин В.Н. Математическое моделирование в экологии: Историко–методологический анализ / В.Н. Тутубалин, Ю.М. Барабашева, А.А. Григорян и др. – М.: Языки русской культуры, 1999.
110. Фишер Р.А. Статистические методы для исследователей. – М.: Госстатиздат, 1958.
111. Флейс Дж. Статистические методы для изучения таблиц долей и пропорций. – М.: Финансы и статистика, 1989.
112. Холлендер М., Вулф Д. Непараметрические методы статистики. – М.: Финансы и статистика, 1983.
113. Чипулис В.П. Оценка достоверности результатов измерений в

теплоэнергетике // Измерительная техника, 2005, № 5, с. 53–58.
114. Яншин В.В., Калинин Г.А. Обработка изображений на языке Си для IBM PC: Алгоритмы и программы. – М.: Мир, 1994.

Глава 3. Параметрическая статистика

3.1. Введение

Все представленные методы применимы только для анализа выборок признаков, измеренных в количественной шкале.
Серьезной проблемой, которая касается представленных методов проверки гипотез, является применимость методов в случае малой численности выборок, что может иметь следствием низкую мощность. Дополнительно о влиянии численности на мощность критериев см. в главе «Введение в практичекий анализ».
Число наблюдений (численность выборки) для использования параметрических критериев должно быть по возможности большим. Минимальные численности выборок можно установить по таблицам, данным в книге Джонсона с соавт.
Считается, что параметрические методы могут применяться, только если эмпирическое распределение анализируемых выборок не противоречит статистической гипотезе о нормальности распределения. В этой связи необходимо отметить два обстоятельства:

- Данную проверку можно выполнить с помощью статистических тестов главы «Проверка нормальности распределения» (в данной главе содержатся рекомендации, какие именно параметры выборок подлежат проверке). Перед нами – яркий пример того, когда проверка предпосылок применения метода гораздо сложнее самого метода.
- Перед использованием параметрических методов, если данные не показывают нормальности распределения, возможна их нормализация. Методы нормализации представлены в главе «Преобразования данных».

Исследования показывают, что острота проблемы отклонения от нормальности и утверждение, что выборка тем нормальнее, чем многочисленнее, преувеличена. Ряд авторов посвятил свои исследования данной теме.

См. работы Виккерса (Vickers), Бриджа (Bridge) с соавт., Мюллера с соавт., Блэйр (Blair) с соавт.

3.2. Работа с программным обеспечением

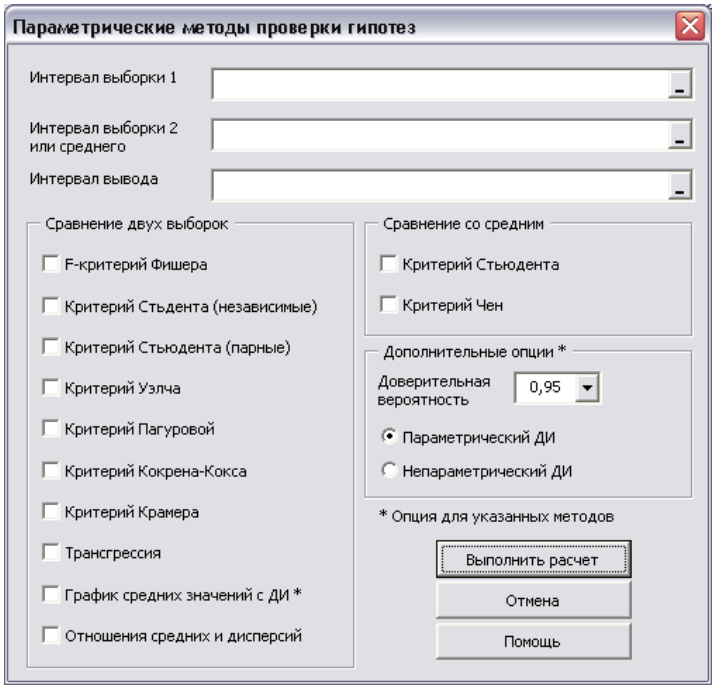
Выберите из меню программы пункт AtteStat | Параметрическая статистика. На экране появится диалоговое окно, изображенное на рисунке.
Затем проделайте следующие шаги:

- Выберите или введите интервалы сравниваемых выборок. При использовании критерия Стьюдента и критерия Чен в качестве второй выборки должна быть введена одна ячейка, в которую следует поместить тестируемое математическое ожидание (при выборе интервала в его качестве будет взято содержимое первой ячейки выделенного интервала). Для парного критерия Стьюдента численности сравниваемых выборок должны быть равны между собой.
- Выберите или введите выходной интервал. Начиная с первой ячейки выходного интервала (следовательно, можно указать только одну ячейку, т.к. остальные ячейки интервала игнорируются), будут выведены результаты вычислений.
- Отметьте критерии для проведения статистического расчета. Возможен выбор

нескольких тестов одновременно. Естественно, не имеет смысла выбирать одновременно критерии из двух групп: и сравнение двух выборок, и сравнение со средним. При выборе нескольких критериев следует выбирать сходные по назначению критерии только из одной группы.

- Выберите дополнительные опции методов, для которых предназначены данные опции.
- Нажмите кнопку «Выполнить расчет».

После выполнения вычислений будет, начиная с первой ячейки выходного интервала, выведено название статистического критерия, значение статистики критерия, вычисленное *P*–значение и предлагаемый программой вывод о результате проверки статистической гипотезы.



Программное обеспечение берет на себя верификацию исходных данных, выдавая подробную диагностику. При ошибках, вызванных неверными действиями пользователя, или ошибках периода выполнения выдаются сообщения об ошибках.

3.2.1. Сообщения об ошибках

При ошибках ввода и во время выполнения программы могут выдаваться диагностические сообщения следующих типов:

Ошибка	Комментарий
Не определен интервал переменной.	Вы не выбрали или неверно ввели интервал эмпирической выборки. Лучшим способом избежать ошибки является не ввод, а выделение интервала стандартным образом, т. е. протаскиванием курсора.
Пустая ячейка в области данных.	Проверьте исходные данные и заполните все ячейки, отмеченные Вами как входной интервал. Для избежания ошибок расчета, вызванных разногласиями, трактовать ли пустую ячейку как нуль, программное обеспечение требует заполнения всех ячеек. Если в ячейке не должно быть данных по физической природе исследуемого процесса, введите в данную

Ошибка	Комментарий
	ячейку нуль.
Нечисловой тип данных.	Проверьте типы ячеек входного интервала. Тип может быть только числовым. Проще всего выделить интервал ячеек и явно определить их тип как числовой стандартными средствами.
Не определена область вывода.	Не выбран или неверно введен выходной интервал. Лучшим способом избежать ошибки является не ввод, а выделение интервала стандартным образом, т. е. протаскиванием курсора.

3.3. Теоретическое обоснование

Критерии (тесты), при помощи которых могут быть сравнены статистические совокупности, разделяются на две группы: параметрические и непараметрические. Особенностью параметрических критериев является ряд требований:

- Распределение признака в генеральной (!) совокупности подчиняется некоторому известному, в данном случае нормальному, закону. Нормальность распределения генеральной совокупности может быть статистически установлена на основе проверки эмпирического распределения выборки из данной совокупности до применения любого параметрического теста с помощью одного из методов, представленных в главе «Проверка нормальности распределения». Задача проверки нормальности в целом сложнее задачи проверки гипотезы о математических ожиданиях. Она может быть уверенно решена лишь при больших объемах выборок.
- Для адекватного применения ряда критериев требуется равенство дисперсий сравниваемых выборок. Поэтому многие авторы рекомендуют проверить нулевую гипотезу о равенстве дисперсий сравниваемых совокупностей с помощью критерия Фишера.

Пусть обе выборки извлечены из генеральных совокупностей, имеющих нормальные распределения с равными или неравными между собой неизвестными дисперсиями. Нулевая гипотеза состоит в том, что средние значения совокупностей равны. При анализе выборок из нормальных генеральных совокупностей с неизвестными дисперсиями, равенство которых не предполагается, либо если отношение дисперсий неизвестно, возникает так называемая проблема Беренса–Фишера (Behrens–Fisher problem), решаемая с помощью параметрических методов: критерия Уэлча, критерия Пагуровой или критерия Кокрена–Кокса.

В практических исследованиях решение данной проблемы актуально, т. к. при анализе реальных экспериментальных данных, особенно в сложных социально–экономических, научно–технических и медико–биологических исследованиях, все параметры распределения чаще всего действительно оцениваются по эмпирическим выборкам. Многие исследователи совершают методическую ошибку, применяя для анализа таких выборок варианты тестов, предназначенных для выборок с известными средними или дисперсиями, или тем и другим одновременно.

Можно предположить, что параметры распределений бывают известными лишь при анализе простых и часто повторяющихся производственных процессов. Показано также, что при больших и примерно равных объемах выборок учет представленных требований не является необходимым.

Параметрические критерии в большинстве случаев являются более мощными, чем их непараметрические аналоги. Если существуют предпосылки использования параметрических критериев, но используются непараметрические, увеличивается вероятность ошибки II рода. Об исследовании ошибки II рода и мощности критерия, а также о влиянии отклонений от

некоторых исходных предположений см. главу «Введение».

См. работы Пинто (Pinto), Рейнеке (Reineke), Райел (Rhiel).

3.3.1. Критерий Стьюдента

Критерий Стьюдента предназначен для проверки нулевой гипотезы о равенстве среднего значения выборочной совокупности заданному математическому ожиданию. Вычисление производится по формуле

$$t = \frac{|\bar{x} - \lambda_0|\sqrt{n}}{s},$$

где $\bar{x}$ – среднее значение совокупности,
λ_0 – заданное математическое ожидание,
n – численность совокупности,
s^2 – оценка выборочной дисперсии.
Статистика критерия Стьюдента подчиняется t–распределению с числом степеней свободы $n - 1$. В отличие от некоторых других представленных в программе тестов, в качестве второй выборки вводится ячейка электронной таблицы, содержащая заданное математическое ожидание. Другие ячейки второй выборки, кроме первой, будут проигнорированы.

Согласно Мюллеру с соавт. (с. 127, см. также ссылку в источнике), «критерий t относительно нечувствителен к небольшим отклонениям от распределения генеральной совокупности от нормального (т. е. практически является робастным)».

3.3.2. Критерий Чен

Критерий Чен (Chen’s test) в качестве обобщения критерия Стьюдента предназначен для проверки нулевой гипотезы о том, что среднее значение выборочной совокупности не превышает заданного математического ожидания

$$\bar{x} \le \lambda_0,$$

где $\bar{x}$ – среднее значение совокупности,
λ_0 – заданное математическое ожидание.
Метод может применяться только при положительном коэффициенте асимметрии.
Вычисление статистики критерия производится по формуле

$$T = t + a(1 + 2t^2) + 4a^2(t + 2t^3),$$

где

$$a = \frac{b}{6\sqrt{n}},$$

b – коэффициент асимметрии,
n – численность совокупности,

$$t = \frac{|\bar{x} - \lambda_0|\sqrt{n}}{s},$$

– статистика критерия Стьюдента,
s^2 – оценка выборочной дисперсии.
Статистика критерия подчиняется стандартному нормальному распределению.
В отличие от некоторых других представленных в программе тестов, в качестве второй выборки вводится ячейка электронной таблицы, содержащая заданное математическое ожидание. Другие ячейки второй выборки, кроме первой, будут проигнорированы.

3.3.3. Критерий Стьюдента для независимых выборок

Критерий Стьюдента для независимых выборок (two–group unpaired *t*–test) предназначен для проверки нулевой гипотезы о равенстве средних значений двух нормальных выборочных совокупностей в случае равных неизвестных дисперсий.
Распределение нормальной случайной величины полностью определяется двумя параметрами: математическим ожиданием (его выборочная оценка – среднее значение) и дисперсией. Поэтому в данном случае нулевая гипотеза может быть сформулирована как гипотеза о том, что выборки извлечены из одной статистической популяции.
Вычисление статистики критерия производится по формуле

$$t = \frac{\left|\bar{x}_1 - \bar{x}_2\right|}{s\sqrt{1/n_1 + 1/n_2}},$$

где $\bar{x}_1$ и $\bar{x}_2$ – средние значения совокупностей,
n_1 и n_2 – численности совокупностей,
s^2 – оценка выборочной дисперсии.
Оценка выборочной дисперсии рассчитывается как

$$s^2 = \frac{s_1^2(n_1 - 1) + s_2^2(n_2 - 1)}{n_1 + n_2 - 2},$$

где s_1^2 и s_2^2 – оценки дисперсий, которые считаются по соответствующим выборкам.
Статистика критерия подчиняется *t*–распределению с числом степеней свободы $n_1 + n_2 - 1$.

Доверительные интервалы для оцениваемой разности средних значений вычислены в статье Сим (Sim) с соавт. Хотя оригинальный критерий изначально предназначен для нормальных количественных выборок, имеется исследование Хирен (Heeren) с соавт. о применении рассмотренного теста к порядковым выборкам.

3.3.4. Парный критерий Стьюдента

Критерий Стьюдента для связанных выборок (парный критерий Стьюдента, two–group paired *t*–test) предназначен для проверки нулевой гипотезы о равенстве средних значений двух выборочных совокупностей в случае неравных неизвестных дисперсий. В источниках критерий может называться одновыборочным критерием Стьюдента. Это название вызвано тем обстоятельством, что на самом деле, исходя из представленной схемы расчета, анализируется действительно одна совокупность, составленная из попарных разностей вариант исходных связанных выборок. Понятно, что в данном случае проверяется нулевая гипотеза о равенстве среднего значения полученной выборки известному значению, а именно – нулю.
Вычисления производятся по формуле

$$t = \frac{\sum_{i=1}^{n} \delta_i}{\sqrt{\frac{n\sum_{i=1}^{n} \delta_i^2 - \left(\sum_{i=1}^{n} \delta_i\right)^2}{n-1}}},$$

где n – численность каждой выборки,
$\delta_i = x_i - y_i$, $i = 1,2,...,n$ – попарные разности вариант совокупностей, где
x_i, $i = 1,2,...,n$ – варианты первой совокупности,

y_i, $i = 1,2,...,n$ – варианты второй совокупности.
Статистика имеет распределение Стьюдента с числом степеней свободы $n - 1$.

Модификацию критерия, с учетом корреляции между выборками, и рассуждения о влиянии типа распределения исходных выборок на мощность критерия приводит Циммерман (Zimmerman).

3.3.5. Критерий Лорда

Критерий Лорда (Lord's range test) разработан для проверки нулевой гипотезы о равенстве средних двух совокупностей. Статистика критерия вычисляется по формуле

$$L = \frac{\left|\bar{x}_1 - \bar{x}_2\right|}{r_1 + r_2},$$

где $\bar{x}_1$ и $\bar{x}_2$ – средние значения совокупностей,
r_1 и r_2 – значения размахов. Подробнее о размахе см. главу «Описательная статистика».
Статистику применяют для очень малых выборок. В таблице представлены уровни значимости. Значение L, равное или большее табличного значения, значимо.

n_1	n_2	5%	1%
2	2	1,71	3,96
3	3	0,64	1,05
4	4	0,41	0,62

Описание критерия и ссылки даны для полноты информации.

Метод представлен в книге Закса, монографии Лэнгли (Langley). См. также работу Пэтнэйка (Patnaik).

3.3.6. Критерий Уэлча

Критерий Уэлча (критерий Велча, критерий Вэлча, критерий Крамера–Уэлча, критерий Саттерзвайта, Satterthwaite's test) предназначен для решения проблемы Беренса–Фишера. Вычисления производятся по формуле

$$d = \frac{\left|\bar{x}_1 - \bar{x}_2\right|}{\sqrt{s_1^2/n_1 + s_2^2/n_2}},$$

где $\bar{x}_1$ и $\bar{x}_2$ – средние значения совокупностей,
n_1 и n_2 – численности совокупностей,
s_1^2 и s_2^2 – оценки дисперсий, которые считаются по соответствующим выборкам.
Распределение статистики критерия близко к t–распределению Стьюдента при числе степеней свободы, равном

$$\nu = \frac{\left(s_1^2/n_1 + s_2^2/n_2\right)^2}{\frac{\left(s_1^2/n_1\right)^2}{n_1 - 1} + \frac{\left(s_2^2/n_2\right)^2}{n_2 - 1}}.$$

Описание критерия см. в книге Закса, Когана с соавт. См. также описание критерия Юена–Уэлча (Yuen–Welch test) в книге Вилкокса (Wilcox).

Уместно указать еще одну модификацию критерия Стьюдента, предложенную Хатчесоном (Hutcheson) и предназначенную для сравнения индексов Шеннона двух совокупностей (см. главу «Информационный анализ»):

$$t = \frac{\left|H_1 - H_2\right|}{\sqrt{D_{H_1}^2 / n_1 + D_{H_2}^2 / n_2}},$$

где H_1 и H_2 – индексы Шеннона (энтропии) совокупностей,

D_{H_1} и D_{H_2} – соответствующие оценки дисперсий индексов Шеннона.

Распределение статистики критерия Хатчесона близко к *t*–распределению Стьюдента при числе степеней свободы, равном

$$\nu = \frac{\left(D_1 + D_2\right)^2}{D_{H_1}^2 / n_1 + D_{H_2}^2 / n_2}.$$

См. статью Хатчесона, работу Шитикова с соавт.

3.3.7. Критерий Пагуровой

Приближенное решение проблемы Беренса–Фишера дано Пагуровой, которая предположила, что распределение статистики критерия существенно зависит от отношения неизвестных дисперсий. Вычисление критерия Пагуровой производится по формуле, аналогичной формуле Уэлча,

$$\upsilon = \frac{\left|\bar{x}_1 - \bar{x}_2\right|}{\sqrt{s_1^2 / n_1 + s_2^2 / n_2}},$$

где $\bar{x}_1$ и $\bar{x}_2$ – средние значения совокупностей,

n_1 и n_2 – численности совокупностей,

s_1^2 и s_2^2 – оценки дисперсий, которые считаются по соответствующим выборкам.

Двустороннее *P*–значение вычисляется как решение нелинейного уравнения

$$\upsilon = t_{n_2,1-p/2}\frac{(\theta-\eta)^2(1-\eta)}{\theta^2} +$$

$$+ t_{n_1+n_2,1-p/2}\frac{[\theta(1-\theta)+(2\theta-1)(\eta-\theta)]\eta(1-\eta)}{\theta^2(1-\theta)^2} + t_{n_1,1-p/2}\frac{(\theta-\eta)^2\eta}{(1-\theta)^2},$$

где *t*.,. – значение обратной функции *t*–распределения,

$$\eta = c - 2c(1-c)\left(\frac{1-c}{n_2} - \frac{c}{n_1}\right),$$

$$\theta = \frac{n_1}{n_1 + n_2},$$

$$c = \frac{s_1^2 / n_1}{s_1^2 / n_1 + s_2^2 / n_2}.$$

Уравнение может быть решено одним из методов локальной оптимизации. В простейшем случае используется метод деления отрезка пополам.

Описание критерия приводится в работе Пагуровой.

3.3.8. Критерий Кокрена–Кокса

Критерий Кокрена–Кокса (Cochran and Cox test) предназначен для решения проблемы Беренса–Фишера. Вычисления производятся по формуле

$$d = \frac{\left|\bar{x}_1 - \bar{x}_2\right|}{\sqrt{\frac{s_1^2}{n_1 - 1} + \frac{s_2^2}{n_2 - 1}}},$$

где $\bar{x}_1$ и $\bar{x}_2$ – средние значения совокупностей,
n_1 и n_2 – численности совокупностей,
s_1^2 и s_2^2 – оценки дисперсий, которые считаются по соответствующим выборкам.
Распределение статистики критерия близко к *t*–распределению Стьюдента при числе степеней свободы, равном

$$\nu = \frac{\left(s_1^2/(n_1 - 1) + s_2^2/(n_2 - 1)\right)^2}{\frac{\left(s_1^2/(n_1 - 1)\right)^2}{n_1 + 1} + \frac{\left(s_2^2/(n_2 - 1)\right)^2}{n_2 + 1}} - 2.$$

3.3.9. Критерий Крамера

Критерий Крамера предназначен для проверки нулевой гипотезы о равенстве средних значений двух выборочных совокупностей в случае равных неизвестных дисперсий.
Вычисление статистики критерия производится по формуле

$$t = \frac{\sqrt{n_1 n_2}\left|\bar{x}_1 - \bar{x}_2\right|}{\sqrt{n_2 s_1^2 + n_1 s_2^2}},$$

где $\bar{x}_1$ и $\bar{x}_2$ – средние значения совокупностей,
n_1 и n_2 – численности совокупностей,
s_1^2 и s_2^2 – оценки дисперсий, которые считаются по соответствующим выборкам.
Статистика критерия подчиняется стандартному нормальному распределению.

См. монографию Крамера.

3.3.10. Критерий Фишера

F–критерий Фишера (критерий Фишера–Снедекора) применяют для сравнения дисперсий двух нормальных выборочных совокупностей. Критерий часто называют дисперсионным отношением или просто статистикой Фишера. Вычисление ведется по формуле, предложенной Снедекором:

$$F = \frac{s_1^2}{s_2^2},$$

где в числителе – оценка дисперсии одной выборки, в знаменателе – оценка дисперсии другой выборки. Принято (см. Лакина) брать отношение большего значения дисперсии к меньшему значению, хотя принципиальной разницы нет.
Числа степеней свободы для поиска критического значения по таблице *F*–распределения (данная таблица – двухвходовая) следует взять $n_1 - 1$ и $n_2 - 1$, где n_1 и n_2 – соответствующие численности совокупностей.

См. книгу Когана с соавт.

3.3.11. Трансгрессия

У независимых выборок из различных генеральных совокупностей часть вариант может оказаться в одних и тех же классах вариационного ряда. Такие ряды называются трансгрессирующими, а их неполное разобщение – трансгрессией. При статистически доказанном различии в средних значениях большая величина трансгрессии (которая может выражаться в долях, как в настоящей программе, или в процентах) заставляет предположить, что разделение рядов по анализируемому фактору не является единственным.
В случае нормальных генеральных совокупностей трансгрессия вычисляется по формуле:

$$Tr = \frac{n_1 P_1 + n_2 P_2}{n_1 + n_2},$$

где n_1 и n_2 – численности совокупностей.
Остальные величины вычисляются по формулам, соответственно,

$$P_1 = 0{,}5 + 0{,}5 \cdot I\left(\frac{\min_2 - \bar{x}_1}{s_1}\right) \text{ и } P_2 = 0{,}5 + 0{,}5 \cdot I\left(\frac{\max_1 - \bar{x}_2}{s_2}\right),$$

где $I(.)$ – интеграл вероятностей,
$\bar{x}_1$ и $\bar{x}_2$ – средние значения совокупностей,
s_1^2 и s_2^2 – оценки дисперсий, которые считаются по соответствующим выборкам, а остальные величины вычисляются по формулам, соответственно,
$\min_2 = \bar{x}_2 - 3s_2$ и $\max_1 = \bar{x}_1 - 3s_1$,
Если окажется, что $\min_2 > \bar{x}_1$ или $\max_1 < \bar{x}_2$, то значения величин P_1 и P_2 рассчитываются по формулам, соответственно,

$$P_1 = 0{,}5 - 0{,}5 \cdot I\left(\frac{\min_2 - \bar{x}_1}{s_1}\right) \text{ и } P_2 = 0{,}5 - 0{,}5 \cdot I\left(\frac{\max_1 - \bar{x}_2}{s_2}\right).$$

См. монографию Лакина.

3.3.12. График средних значений с ДИ

Представленное программное обеспечение дает возможность табличного и графического вывода средних значений сравниваемых выборок, включая доверительные интервалы. При этом на график накладываются, по выбору пользователя, параметрические либо непараметрические доверительные интервалы, вычисленные для доверительного уровня, заданного из стандартной линейки.
Доверительные интервалы оцениваемых средних значений нормальных выборок вычисляются по формуле

$$I_m = \left(\bar{x} - t_{(1+\beta)/2}\frac{s}{\sqrt{n}}; \bar{x} + t_{(1+\beta)/2}\frac{s}{\sqrt{n}}\right),$$

где s – выборочная оценка стандартного отклонения,
$t_{(1+\beta)/2}$ – значение обратной функции t–распределения Стьюдента с параметрами $n - 1$ и $(1 + \beta) / 2$,
β – доверительный уровень, выраженный в долях.
Для вычисления двустороннего доверительных интервалов оцениваемых средних значений,

когда выборки не является нормальными, применяется формула:

$$I_m = \left(\bar{x} - \Psi((1+\beta)/2)\frac{s}{\sqrt{n}}; \bar{x} + \Psi((1+\beta)/2)\frac{s}{\sqrt{n}} \right),$$

где Ψ(.) – обратная функция стандартного нормального распределения.
Дополнительно в таблице выводится разность средних анализируемых выборок $(\bar{x}_1 - \bar{x}_2)$,

где $\bar{x}_1$ и $\bar{x}_2$ – средние значения совокупностей.
Также в таблице выводятся заданные доверительные интервалы. Доверительный интервал оцениваемой разности средних значений (выборки нормальные) вычисляется по формуле

$$I_d = \left((\bar{x}_1 - \bar{x}_2) - t_{(1+\beta)/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}; (\bar{x}_1 - \bar{x}_2) + t_{(1+\beta)/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \right),$$

где s_1^2 и s_2^2 – оценки дисперсий, которые считаются по соответствующим выборкам.
n_1 и n_2 – численности совокупностей,
$t_{(1+\beta)/2}$ – значение обратной функции *t*–распределения Стьюдента с параметрами ν (число степеней свободы) и (1 + β) / 2. При этом число степеней свободы считается как

$$\nu = \frac{\left(s_1^2/n_1 + s_2^2/n_2\right)^2}{\frac{\left(s_1^2/n_1\right)^2}{n_1 - 1} + \frac{\left(s_2^2/n_2\right)^2}{n_2 - 1}}.$$

Доверительный интервал оцениваемой разности средних значений (выборки не являются нормальными) вычисляется по формуле

$$I_d = \left((\bar{x}_1 - \bar{x}_2) - \Psi((1+\beta)/2)\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}; (\bar{x}_1 - \bar{x}_2) + \Psi((1+\beta)/2)\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \right).$$

Результаты представленного графического анализа интерпретируются следующим образом. Если 100β% доверительные интервалы оцениваемых средних значений сравниваемых выборок пересекаются, конкурирующая гипотеза (средние не равны) может быть принята на уровне значимости $p \le \beta$. Если 100β% доверительные интервалы оцениваемых средних значений сравниваемых выборок не пересекаются, нулевая гипотеза (средние равны) не отвергается на уровне значимости $p > \beta$. Т. к. доверительные интервалы тем шире, чем больше значение β, выбирая различные стандартные значения β, можно получить значение уровня значимости, более точно соответствующее представленным данным.

См. статьи Массон (Masson) с соавт., Вольфе (Wolfe) с соавт., Пэйтон (Payton) с соавт., Остин (Austin) с соавт., Маршалл (Marshall).

3.3.13. Отношения средних и дисперсий

Представленное программное обеспечение дает возможность вычисления точечных и интервальных оценок отношения средних и отношения дисперсий двух нормальных выборок. В дальнейших выкладках подразумевается, что первая выборка – та, соответствующее значение которой стоит в числителе, вторая выборка – в знаменателе.
Для вычисления доверительного интервала оцениваемого отношения средних значений двух выборок

$$q = \frac{m_1}{m_2},$$

m_1 – среднее значение первой выборки,
m_2 – среднее значение второй выборки,
сначала вычисляется промежуточная переменная

$$g = \left[t_{(1+\beta)/2} \frac{\mu_2}{m_2} \right]^2,$$

где $t_{(1+\beta)/2}$ – значение обратной функции t–распределения Стьюдента с параметрами $n_1 + n_2 - 2$ и $(1 + \beta) / 2$,
β – доверительный уровень, выраженный в долях,
μ_2 – стандартная ошибка среднего значения второй выборки.
Дальнейшие вычисления зависят от значения промежуточной переменной, которая является численной характеристикой отношения стандартной ошибки среднего значения второй выборки к самому ее среднему значению.
1. При $g \geq 1$ искомой интервальной оценки не существует.
2. При малом значении g стандартная ошибка отношения средних значений вычисляется по формуле (в программе – метод 1)

$$SE_q = q \sqrt{\frac{\mu_1}{m_1} + \frac{\mu_2}{m_2}},$$

где μ_1 – стандартная ошибка среднего значения первой выборки.
При этом доверительный интервал оцениваемого отношения средних значений
$I_{\frac{m_1}{m_2}} = \left(q - t_{(1+\beta)/2} SE_q ; q + t_{(1+\beta)/2} SE_q \right)$.
3. При большом значении g стандартная ошибка отношения средних значений вычисляется по уточненной формуле (в программе – метод 2)

$$SE_q = \frac{q}{1-g} \sqrt{(1-g)\frac{\mu_1}{m_1} + \frac{\mu_2}{m_2}},$$

При этом доверительный интервал оцениваемого отношения средних значений

$$I_{\frac{m_1}{m_2}} = \left(\frac{q}{1-g} - t_{(1+\beta)/2} SE_q ; \frac{q}{1-g} + t_{(1+\beta)/2} SE_q \right)$$

Доверительный интервал оцениваемого отношения дисперсий вычисляется по формуле

$$I_{\frac{\sigma_1^2}{\sigma_2^2}} = \left(\frac{s_1^2}{s_2^2} F^{-1}_{(1+\beta)/2}(n_1 - 1, n_2 - 1); \frac{s_1^2}{s_2^2} F^{-1}_{1-(1+\beta)/2}(n_1 - 1, n_2 - 1) \right),$$

где s_1^2 – выборочное значение дисперсии 1–й выборки,
s_2^2 – выборочное значение дисперсии 2–й выборки,
n_1 и n_2 – численности совокупностей,
$F_{\cdot}^{-1}(.,.)$ – обратная функция F–распределения.

Алгоритмы вычислений и поясняющие примеры см. в монографиях Мотульски (Motulsky), Бетеа (Bethea) с соавт. См. также статью Ли (Lee A.F.S.) с соавт.

Список использованной и рекомендуемой литературы

1. Austin P., Hux J. A brief note on overlapping confidence intervals // Journal of Vascular Surgery, July 2002, vol. 36, issue 1, pp. 194–195.
2. Best D.I., Rayner C.W. Welch's approximate solution for the Behrens–Fisher problem //

Technometrics, 1987, vol. 29, pp. 205 –210.
3. Bethea R.M., Duran B.S., Boullion T.L. Statistical methods for engineers and scientists. – New York, NY: Marcel Dekker, 1995.
4. Blair R.C., Higgins J.J. Comparison of the power of the paired samples t test to that of Wilcoxon's signed–ranks test under various population shapes // Psychological Bulletin, January 1985, vol. 97, no. 1, pp. 119–128.
5. Bridge P.D., Sawilowsky S.S. Increasing physicians' awareness of the impact of statistics on research outcomes: comparative power of the t–test and and Wilcoxon rank–sum test in small samples applied research // Journal of Clinical Epidemiology, 1999, vol. 52, no.3, pp. 229–235.
6. Chen L. Testing the mean of skewed distribution // Journal of the American Statistical Association, 1995, vol. 90, pp. 767–772.
7. Chernick M.R. Friis R.H. Introductory biostatistics for the health sciences. Modern application including bootstrap. – New York, NY: John Wiley & Sons, 2003.
8. Cochran W.G., Cox G.M. Experimental designs. – New York, NY: John Wiley & Sons, 1950.
9. Cohen J. Statistical power analysis for the behavioral sciences. – Hillsdale, NJ: Lawrence Erlbaum Associates, 1988.
10. Dunlop W.P. Meta–analysis of experiments with matched groups or repeated measures designs / W.P. Dunlop, J.M. Corina, J.B. Vaslow et al. // Psychological Methods, 1996, vol. 1, pp. 170–177.
11. Fisher R.A. Statistical tables for biological, agricultural and medical research / Ed. by R.A. Fisher, F. Yates. – Edinburgh: Oliver and Boyd, 1963.
12. Guyatt G. Basic statistics for clinicians: 1. Hypothesis testing / G. Guyatt, R. Jaeschke, N. Heddle et al. // Canadian Medical Association Journal, January 1995, vol. 152, issue 1, pp. 27–32.
13. Heeren T., D'Agostino R. Robustness of the two independent samples *t*–test when applied to ordinal scaled data // Statistics in Medicine, 1987, vol. 6, pp. 79–90.
14. Hinkle D.E., Wiersma W., Jurs S.G. Applied statistics for the behavioral sciences. – Boston, MA: Houghton Mifflin, 1994.
15. Huntsberger D.V., Billingsley P.P. Elements of statistical inference. – Dubuque, IA: WM. C. Brown Publishers, 1989.
16. Hussien A., Carriere K.C. Robustness of procedures for the Behrens–Fisher problems: extension to bivariate normal mixtures // Communications in Statistics – Simulation and Computation, 2001, vol. 30, no. 4, pp. 831–846.
17. Hutcheson K. A test for comparing diversities based on the Shannon formula // Journal of Theoretical Biology, October 1970, vol. 29, issue 1, pp. 151–154.
18. Jaworski S., Zielinski W. A procedure for ε–comparison of means of two normal distributions // Applicationes Mathematicae, 2004, vol. 31, no. 2, pp. 155–160.
19. Langley R. Practical statistics for non–mathematical people. – Newton Abbot, UK: David and Charles, 1971.
20. Lee A.F.S., Gurland J. Size and power of tests for equality of means of two normal populations with unequal variances // Journal of the American Statistical Association, 1975, vol. 70, pp. 933–941.
21. Lee J.C., Lin S.–H. Generalized confidence intervals for the ratio of means of two normal populations // Journal of Statistical Planning and Inference, 2004, vol. 123, no. 1, pp. 49–60.
22. Lehmann E.L. Testing statistical hypotheses. – New York, NY: John Wiley & Sons, 1986.
23. Marshall S.W. Testing with confidence: The use (and misuse) of confidence intervals in biomedical research // Journal of Science and Medicine in Sport, June 2004, vol. 7, issue 2, pp. 135–137.

24. Masson M.E.J., Loftus G.R. Using confidence intervals for graphically based data interpretation // Canadian Journal of Experimental Psychology, 2003, vol. 57, no. 3, pp. 203–220.
25. McDonald L.L. Evaluation and comparison of hypothesis testing techniques for bond release applications / L.L. McDonald, S. Howlin, J. Polyakova et al. // Final Report, May 28, 2003, Western EcoSystems Technology.
26. Misanan J.R., Hinderliter C.F. Fundamentals of statistics for psychology students. – New York, NY: Harper Collins, 1991.
27. Moore D.S. The basic practice of statistics. – New York, NY: W.H. Freeman and Company, 1995.
28. Motulsky H.J. Intuitive biostatistics. – New York, NY: Oxford University Press, 1995.
29. Myers J.L., Well A.D. Research design and statistical analysis. – New York, NY: Harper Collins, 1991.
30. Patnaik P.B. The use of mean range as an estimator of variance in statistical tests // Biometrika, June 1950, Vol. 37, no. 1/2, pp. 78–87.
31. Payton M.E., Greenstone M.H., Schenker N. Overlapping confidence intervals or standard error intervals: What do they mean in terms of statistical significance? // Journal of Insect Science, October 2003, vol. 3, issue 34, pp. 1–6.
32. Pinto J.V., Ng P., Allen D.S. Logical extremes, beta, and the power of the test // Journal of Statistics Education, 2003, vol. 11, no. 1.
33. Posten H.O., Yeh Y.Y., Owen D.B. Robustness of the two–sample t test under violations of the homogeneity of variance assumption // Communications in Statistics, 1982, vol. 11, pp. 109 –126.
34. Ramsey P.H. Exact type I error rates for robustness of Student's t test with unequal variances // Journal of Educational Statistics, 1980, vol. 5, pp. 337–349.
35. Reineke D.M., Baggett J., Elfessi A. A note on the effect of skewness, kurtosis, and shifting on one–sample t and sign tests // Journal of Statistics Education, 2003, vol. 11, no. 3.
36. Rhiel S.G., Chaffin W.W. An investigation of the large–sample/small–sample approach to the one–sample test for a mean (sigma unknown) // Journal of Statistics Education, 1996, vol. 4, no. 3.
37. Robinson G.K. Properties of Student's t and of the Behrens–Fisher solution to the two mean problem // Annals of Statistics, 1976, vol. 4, pp. 963–971.
38. Rossi J.A. An application of Welch's approximate t–solution of the Behrens–Fisher problem to confidence intervals // Technometrics, February 1975, vol. 17, no. 1, pp. 57–60.
39. Salvatore D., Reagle D. Statistics and econometrics. – London, UK: McGraw–Hill, 2003.
40. Satterthwaite F.W. An approximate distribution of estimates of variance components // Biometrics Bulletin, 1946, vol. 2, pp. 110–114.
41. Scheffe H. Practical solutions of the Behrens–Fisher problem // Journal of the American Statistical Association, 1970, vol. 65, pp. 1501–1508.
42. Sheskin D.J. Handbook of parametric and nonparametric statistical procedures. – Boca Raton, FL: Chapman & Hall / CRC, 2000.
43. Sim J., Reid N. Statistical inference by confidence intervals: Issues of interpretation and utilization // Physical Therapy, February 1999, vol. 79, no. 2, pp. 186–195.
44. Snedecor G.W., Cochran W.G. Statistical methods. – Ames, IA: Iowa State University Press, 1980.
45. Steel R.G.D., Torrie J.H. Principles and procedures of statistics. – New York, NY: McGraw–Hill, 1980.
46. Vickers A.J. Parametric versus non–parametric statistics in the analysis of randomized trials with non–normally distributed data // BMC Medical Research Methodology, November

2005, vol. 5, pp. 35–47.
47. Wang Y.Y. Probabilities of the type I error of the Welch tests for the Behrens–Fisher Problem // Journal of the American Statistical Association, 1971, vol. 66, pp. 605–608.
48. Wilcox R.R. Introduction to robust estimation and hypothesis testing. – New York, NY: Elsevier Academic Press, 2005.
49. Wolfe R., Cumming G. Communicating the uncertainty in research findings: Confidence intervals // Journal of Science and Medicine in Sport, 2004, vol. 7, pp. 138–143.
50. Wolfe R., Hanley J. If we're so different, why do we keep overlapping? When 1 plus 1 doesn't make 2 // Canadian Medical Association Journal, 8 January 2002, vol. 166, no. 1, pp. 65–66.
51. Yuen K.K. The two–sample trimmed t for unequal population variances // Biometrika, 1974, vol. 61, pp. 165 –170.
52. Zimmerman D.W. Increasing power in paired–samples designs by correcting the Student t statistic for correlation // InterStat (Statistics on the Internet), September 2005, No. 2.
53. Афифи А., Эйзен С. Статистический анализ: Подход с использованием ЭВМ. – М.: Мир, 1982.
54. Белова Е.Б. Компьютеризованный статистический анализ для историков. Учебное пособие / Е.Б. Белова, Л.И. Бородкин, И.М. Гарскова и др. – М.: МГУ, 1999.
55. Бендат Дж., Пирсол А. Прикладной анализ случайных данных. – М.: Мир, 1989.
56. Большев Л.Н., Смирнов Н.В. Таблицы математической статистики. – М.: Наука, 1983.
57. Боровков А.А. Математическая статистика. Оценка параметров. Проверка гипотез. – М.: Наука, 1984.
58. Брандт З. Анализ данных. Статистические и вычислительные методы для научных работников и инженеров. – М.: Мир, ООО «Издательство АСТ», 2003.
59. Браунли К.А. Статистическая теория и методология в науке и технике. – М.: Наука, 1977.
60. Гайдышев И. Анализ и обработка данных: специальный справочник. – СПб: Питер, 2001.
61. Гудман С.Н. На пути к доказательной биостатистике. Часть 1: обманчивость величины *p* // Международный журнал медицинской практики, 2002, № 1, с. 8–17.
62. Гудман С.Н. На пути к доказательной биостатистике. Часть 2: байесовский критерий // Международный журнал медицинской практики, 2002, № 2, с. 5–14.
63. Джонсон Н., Лион Ф. Статистика и планирование эксперимента в технике и науке. Методы обработки данных. – М.: Мир, 1980.
64. Зайцев Г.Н. Математическая статистика в экспериментальной ботанике. – М.: Наука, 1984.
65. Закс Л. Статистическое оценивание. – М.: Статистика, 1976.
66. Иванов Ю.И., Погорелюк О.Н. Статистическая обработка результатов медико–биологических исследований на микрокалькуляторах по программам. – М.: Медицина, 1990.
67. Коган Р.И., Белов Ю.П., Родионов Д.А. Статистические ранговые критерии в геологии. – М.: Недра, 1983.
68. Корнилов С.Г. Оптимальные объемы групп при сравнении средних / Биометрический анализ в биологии. – М.: Издательство Московского университета, 1982, с. 71–90.
69. Крамер Г. Математические методы статистики. – М.: Мир, 1975.
70. Лакин Г.Ф. Биометрия. – М.: Высшая школа, 1990.
71. Леман Э. Проверка статистических гипотез. – М.: Наука, 1979.
72. Леонов В.П., Ижевский П.В. Об использовании прикладной статистики при подготовке диссертационных работ по медицинским и биологическим

специальностям // Бюллетень ВАК РФ, 1997, № 5, с. 56–61.
73. Медик В.А., Токмачев М.С., Фишман Б.Б. Статистика в медицине и биологии: Руководство. В 2–х томах. / Под ред. Ю.М. Комарова. Т.1. Теоретическая статистика. – М.: Медицина, 2000.
74. Мюллер П., Нойман П., Шторм Р. Таблицы по математической статистике. – М.: Финансы и статистика, 1982.
75. Налимов В.В. Применение математической статистики при анализе вещества. – М.: Государственное издательство физико–математической литературы, 1960.
76. Новиков Д.А., Новочадов В.В. Статистические методы в медико–биологическом эксперименте (типовые случаи). – Волгоград: Издательство ВолГМУ, 2005.
77. Пагурова В.И. Критерий сравнения средних значений по двум нормальным выборкам. – М.: ВЦ АН СССР, 1968.
78. Петрович М.Л., Давидович М.И. Статистическое оценивание и проверка гипотез на ЭВМ. – М.: Финансы и статистика, 1989.
79. Поллард Дж. Справочник по вычислительным методам статистики. – М.: Финансы и статистика, 1982.
80. Прохоров Ю.В. Вероятность и математическая статистика. Энциклопедия / Гл. ред. Ю.В. Прохоров. – М.: Научное издательство «Большая Российская энциклопедия», 1999.
81. Романовский В.И. Математическая статистика. Кн.2. Оперативные методы математической статистики. – Ташкент: Издательство Академии наук УзССР, 1963.
82. Сергиенко В.И., Бондарева И.Б. Математическая статистика в клинических исследованиях – М.: ГЭОТАР–МЕД, 2001.
83. Сидоренко Е.В. Методы математической обработки в психологии. – СПб.: ООО «Речь», 2001.
84. Тюрин Ю.Н., Макаров А.А. Статистический анализ данных на компьютере. – М.: ИНФРА–М, 1998.
85. Фишер Р.А. Статистические методы для исследователей. – М.: Госстатиздат, 1958.
86. Чубенко А.В. Применение современных методов математической статистики при проведении клинических научных исследований и их анализе. Сравнение двух пропорций / А.В. Чубенко, П.Н. Бабич, С.Н. Лапач и др. // Аптека, 8 сентября 2003, № 34 (405).
87. Шитиков В.К., Розенберг Г.С. Оценка биоразнообразия: попытка формального обобщения // Количественные методы экологии и гидробиологии / Под ред. Г.С. Розенберга. – Тольятти: ИЭВБ РАН, 2005, с. 91–129.

Глава 4. Непараметрическая статистика

4.1. Введение

Программное обеспечение реализует непараметрические методы проверки статистических гипотез и методы анализа качественных (бинарных) данных.
Бытует несколько основных соображений относительно полезности непараметрических методов (по данным литературы):

- Параметрические методы могут применяться, только если доказана нормальность распределения анализируемых выборок, но эмпирические выборки, полученные в реальных экспериментах, очень часто не являются нормально распределенными.
- Параметрические методы могут применяться для больших выборок. Реальные выборки часто содержат небольшое число вариант, что тем более делает полезным

непараметрические методы.

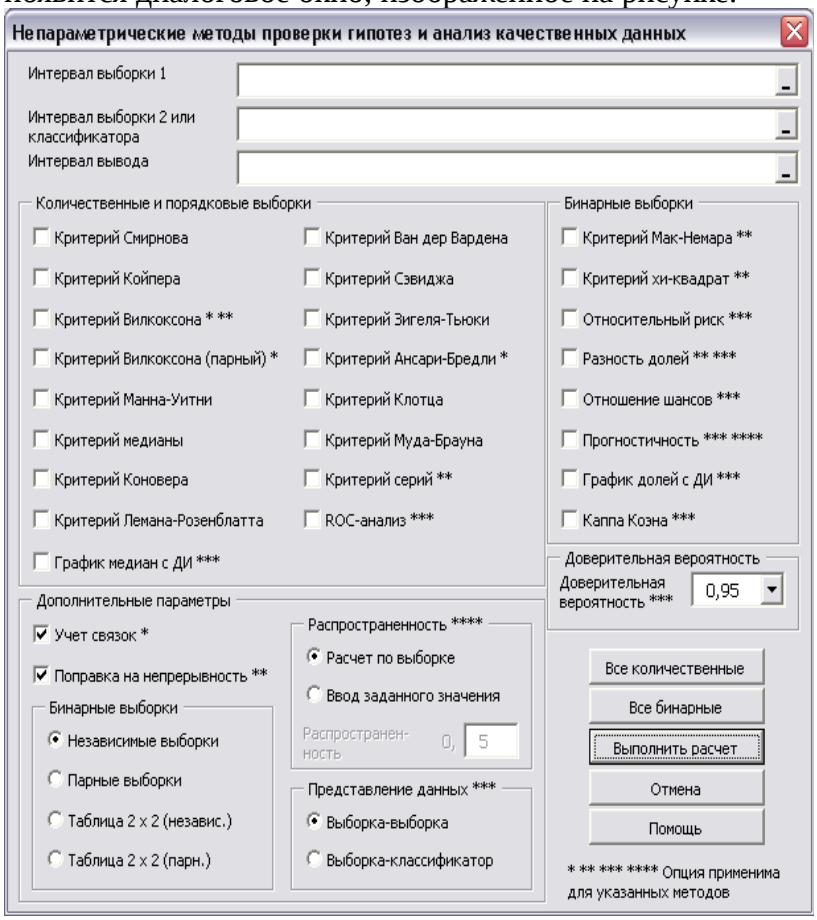
Исследования показывают, что острота проблемы отклонения от нормальности преувеличена, а утверждение, что выборка тем нормальнее, чем многочисленнее, не имеет основания. Ряд авторов посвятил свои исследования данной теме. См. работы Виккерса (Vickers), Бриджа (Bridge) с соавт., Мюллера с соавт., Блэйр (Blair) с соавт.

Серьезной проблемой, которая касается представленных методов проверки гипотез так же, как и параметрических, является применимость методов в случае малой численности выборок, что может иметь следствием низкую мощность критерия (напомним, что мощность – это не число, а монотонная функция численности – чем больше численности выборок, тем выше мощность критерия, к тому же зависящая от альтернативы). Дополнительно о влиянии численности на мощность критериев см. в главе «Введение в практичекий анализ».

Перед применением любого статистического метода необходимо убедиться, что проверяется статистическая значимость различий именно тех параметров выборок, которые интересуют исследователя, а также в том, что метод соответствует шкале измерения исходных данных (признаков). О шкалах измерения см. главу «Введение».

4.2. Работа с программным обеспечением

Выберите из меню программы пункт AtteStat | Непараметрическая статистика. На экране появится диалоговое окно, изображенное на рисунке:



Затем проделайте следующие шаги:

- Выберите или введите интервалы сравниваемых выборок. Если анализируется заранее составленная пользователем таблица 2 x 2 (это следует выбрать опционно в разделе

«Выбор параметров» рассматриваемой формы), в качестве ее первого столбца укажите «Интервал выборки 1», в качестве второго столбца «Интервал выборки 2».

- Выберите или введите выходной интервал. Начиная с первой ячейки выходного интервала (следовательно, можно указать только одну ячейку, т.к. остальные ячейки интервала игнорируются), будут выведены результаты вычислений.
- Выберите критерий или группу критериев для проведения статистического расчета. Для выбора группы критериев можно воспользоваться кнопками Все количественные (для выбора всех критериев для количественных или порядковых выборок) или Все бинарные (для выбора всех критериев для дихотомических выборок).Для отмеченных критериев оставьте по умолчанию или отмените учет поправок. О влиянии и необходимости тех или иных поправок см. описания соответствующих тестов.
- Для отмеченных методов выберите дополнительные опции. Подробнее см. описания соответствующих методов.
- Нажмите кнопку «Выполнить расчет».

После выполнения вычислений будет, начиная с первой ячейки выходного интервала, выведено название статистического критерия, значение статистики критерия, *P*–значение и предлагаемый программой вывод о результате проверки статистической гипотезы. Для ряда методов может быть выдан также доверительный интервал.
Программное обеспечение берет на себя верификацию исходных данных, выдавая подробную диагностику. При ошибках, вызванных неверными действиями пользователя, или ошибках периода выполнения выдаются сообщения об ошибках.

4.2.1. Сообщения об ошибках

При ошибках ввода и во время выполнения программы могут выдаваться диагностические сообщения следующих типов:

Ошибка	Комментарий
Не определен интервал переменной.	Вы не выбрали или неверно ввели входной интервал. Лучшим способом избежать ошибки является не ввод, а выделение интервала стандартным образом, т. е. протаскиванием курсора.
Пустая ячейка.	Проверьте исходные данные и заполните все ячейки, отмеченные Вами как входной интервал. Для избежания ошибок расчета, вызванных разногласиями, трактовать ли пустую ячейку как нуль, данное программное обеспечение требует заполнения всех ячеек. Если в ячейке не должно быть данных по физической природе исследуемого процесса, введите в данную ячейку нуль.
Нечисловой тип данных.	Проверьте типы ячеек входного интервала. Тип может быть только числовым. Проще всего выделить интервал ячеек и явно определить их тип как числовой стандартными средствами.
Не определена область вывода.	Не выбран или неверно введен выходной интервал. Лучшим способом избежать ошибки является не ввод, а выделение интервала стандартным образом, т. е. протаскиванием курсора.
Мало данных.	Указан интервал исходных данных слишком малой численности (менее 2). Укажите верные интервалы данных.

4.3. Теоретическое обоснование

Существует большое количество опытных данных, которые не показывают нормальности распределения, поэтому применение параметрических критериев не может быть обоснованным для данных рассматриваемого класса.
Практически ценными явились робастные методы, которые применимы в широком диапазоне условий. Робастные, непараметрические и свободные от распределения процедуры традиционно относят к одному классу, хотя в литературе есть и альтернативные мнения. Сам термин «непараметрическая статистика» был введен в 1942 году Вольфовицем.

4.3.1. Робастность

Подробное обсуждение этой темы приводится Хьюбером. Под робастностью мы понимаем слабую чувствительность к отклонениям от стандартных условий (например, эмпирическое распределение может отличаться от теоретического нормального), а методы, применимые в широком диапазоне реальных условий, называем робастными. В этом качестве понятие робастности статистического метода практически совпадает со смыслом данного понятия, которое вкладывается в него в механике и смежных прикладных дисциплинах.
Понятие робастности не тождественно устойчивости статистической процедуры (не путать с численной устойчивостью алгоритма). Как указывает Хьюбер, статистическую процедуру называют устойчивой, если на значение оценки не оказывают влияния малые изменения в выборке (малые изменения всех или большие изменения нескольких значений – см. «Обработка выбросов»). Понятия устойчивости и робастности различны, но иногда их применяют в качестве синонимов.
Непараметрические критерии не требуют предварительных предположений относительно вида исходного распределения и являются более робастными, чем их параметрические аналоги. Их называют также критериями значимости, независимыми от типа распределения. Естественно, непараметрические критерии применимы и для случая нормального распределения. Однако непараметрические критерии в большинстве случаев являются менее мощными, чем их параметрические аналоги. Если существуют предпосылки использования параметрических критериев, но используются непараметрические, увеличивается вероятность ошибки II рода.

4.3.2. Тестируемые параметры

Многие пользователи задают вопрос, почему, к примеру, одним методом между выборками выявляются статистически значимые различия, другим – нет. Дело в том, что все методы предназначены для проверки отсутствия статистических различий в различных параметрах (иногда – в совокупности параметров) выборок. Так, можно себе представить такие выборки, которые имеют одинаковые параметры положения (медианы), но разные параметры рассеяния (дисперсии). В таком гипотетическом случае критерий Ансари–Бредли покажет наличие различий, критерий Вилкоксона – нет. Становится понятным, почему исследователи часто не ограничиваются одним тестом, а пытаются выполнить их совокупность для статистического сравнения всевозможных параметров выборок: средних, медиан, дисперсий, функций распределения.
При формулировании нулевой гипотезы обязательно следует указывать, какие конкретные параметры эмпирических выборок сравниваются с помощью используемого критерия. Данная информация приводится в описании каждого критерия. Нужно указывать это в научной публикации, чтобы читатель имел возможность проверить правильность рассуждений автора. В таблице указаны тестируемые параметры выборок для различных критериев.

Тестируемые параметры	Статистический критерий
Положение (location tests)	Вилкоксона, Манна–Уитни, Ван дер Вардена, Уайта, Фишера–Йейтса–Терри–Гефтинга, Розенбаума, медианы, медианный Муда–Брауна, Гехана, Блома, Тьюки, Мак–Немара, серий Вальда–Вольфовица
Рассеяние/масштаб (scale tests)	Ансари–Бредли, Клотца, Сэвиджа, Коновера, Муда, Дэвида, Зигеля–Тьюки, Мозеса
Функция распределения	Смирнова, Крамера–фон Мизеса, Койпера, Лемана–Розенблатта

В показанной таблице не конкретизировано, какие именно параметры являются параметрами положения, а какие параметрами рассеяния. Уточнение приводится в таблице.

Параметр	Параметрика	Непараметрика
Положение	Среднее значение	Медиана или псевдомедиана (оценка Ходжеса–Лемана)
Рассеяние	Стандартное отклонение[2]	Межквартильный размах или семиинтерквартильная широта

Подробнее обо всех перечисленных параметрах см. главу «Описательная статистика».

4.3.3. Типы критериев

Все непараметрические критерии проверки гипотез, в зависимости от их конструкции, могут принадлежать к одному из следующих типов:

- ранговые критерии (рангом называют номер варианты в ряду упорядоченных по возрастанию или убыванию вариант),
- критерии, основанные на сравнении функций распределения,
- точные критерии.

Представленное разделение критериев на типы очень условно и часто относится только к реализации. Лучше говорить о тестируемых параметрах, как это описано в предыдущем разделе. В описании некоторых критериев авторами устанавливаются параллели между ранговыми и перестановочными критериями, ранговыми критериями и критериями на основе функций распределения. Описаны комбинаторные алгоритмы вычисления ранговых критериев. К точным критериям относятся как перестановочные критерии для таблиц сопряженности, являющихся продуктом анализа номинальных признаков, так и критерии первых других типов, для которых известно (и практически применимо) точное распределение статистик.
Многие из представленных критериев имеют многомерные аналоги, представленные в главе «Дисперсионный анализ».

См. монографии Холлендера с соавт., Гаека с соавт., Хеттсманпергера, Коновера (Conover), Руниона, нормативный документ EPA QA/G–9.

4.3.3.1. Ранговые критерии

К ранговым критериям рассматриваемого класса, представленным в программе, относятся:

2 Обычно в качестве параметра рассеяния применяют дисперсию, однако в данном случае удобно взять стандартное отклонение для сопоставления с параметром рассеяния в непараметрическом случае.

- критерий Вилкоксона для независимых выборок,
- критерий Вилкоксона для связанных выборок,
- критерий Манна–Уитни,
- критерий Ван дер Вардена,
- критерий Сэвиджа,
- критерий Ансари–Бредли,
- критерий Клотца,
- критерий Зигеля–Тьюки,
- критерий Коновера,
- медианный критерий Муда–Брауна.

Некоторые из представленных тестов являются эквивалентными. Критерии называются эквивалентными, по определению Холлендера и Вулфа, если для любых возможных выборок решение, принятое с помощью одного из критериев, согласуется с решением, принятым с помощью другого критерия.

Нетрудно показать эквивалентность ряда критериев, например, критериев Дэвида (Бартона–Дэвида, Barton–David, метод не представлен в настоящей программе), Ансари–Бредли и Зигеля–Тьюки. Для упомянутых методов Клотц дает следующую формулу:

$(T + T' + 1) / 4 = W = (N / 2 + 1) / N / 2 – S$,

где T и T' – статистики Зигеля–Тьюки,

W – статистика Ансари–Бредли,

S – статистика Бартона–Дэвида,

N – численность объединенной выборки.

Также эквиваленты критерии Вилкоксона (для независимых выборок, без учета поправок) и Манна–Уитни. Простая формула их связи имеет вид

$$U = n_1 n_2 + \frac{n_1(n_1+1)}{2} - W,$$

где U – статистика Манна–Уитни,

W – статистика Вилкоксона,

n_1 – численность той выборки, для которой вычислялись статистики,

n_2 – численность другой выборки.

Сказанное означает, что вместо одного из названных критериев можно с успехом применить другой и, казалось бы, нет причин помещать в программу эквивалентные тесты. Однако при составлении программы для конечного пользователя обязательно следует учитывать такой субъективный фактор, как традиции в конкретных лабораториях, институтах или даже областях знаний. Понятно стремление исследователя использовать именно тот тест, которым пользуются его коллеги. Поэтому необходимо дать возможность пользователю использовать тест, к которому он привык и которому доверяет.

Данные критерии называются ранговыми (Ф. Вилкоксон, 1945 г.), так как они оперируют не численными значениями вариант, а их рангами. Сначала производят совместное ранжирование сравниваемых выборок. Данная процедура может быть организована различными способами, однако предпочтительным в смысле простоты понимания процесса и его реализации является объединение двух сравниваемых выборок, их сортировка, ранжирование по требуемой схеме и последующее разнесение рангов на места соответствующих им вариант в обеих выборках. Если имеются совпадающие значения, совпавшим наблюдениям условились назначать средний ранг.

Ранговые критерии могут применяться к признакам, измеренным в количественной или порядковой шкале. Применение ранговых критериев к количественным признакам фактически понижает исходную количественную шкалу до порядковой шкалы (напомним, что ранг – это номер варианты по порядку в ранжированном ряду). Это вызывает опасение

некоторых авторов, хотя в литературе показано, что точность выводов снижается гораздо меньше, чем можно было бы себе вообразить.
Схемы вычислений всех ранговых критериев могут быть описаны одними и теми же универсальными соотношениями, отличающимися только способом вычисления ранговых отметок (функций от рангов). Кроме того, перед ранжированием исходные выборки, в зависимости от схемы алгоритма, могут быть подвергнуты преобразованиям. Таким образом, в зависимости от требований алгоритма, могут получаться различные формулы вычисления ранговых критериев.
Обозначим:
$N = n_1 + n_2$ – общее число наблюдений в двух тестируемых выборках, которое может быть скорректировано при наличии совпадающих вариант,
n_1 – число наблюдений в одной выборке,
n_2 – число наблюдений в другой выборке.
Общая формула вычисления статистики рангового критерия, согласно Хеттсманпергеру, может быть представлена в виде

$$S = \sum_{i=1}^{n_1} a(R_i),$$

где $R_i, i = 1,2,...,n_1,$ – ранги наблюдений выборки,
$a(R_i), i = 1,2,...,n_1,$ – ранговые метки общего вида.
Для статистик ранговых критериев могут быть известны точные формулы вычисления критических значений, однако вычисления по точным формулам часто трудоемки уже при средних и всегда при больших численностях выборок. Подробнее о вычислениях распределений см. главу «Введение».
Они удобны для построения точных статистических таблиц, однако в практических вычислениях, как показали Гаек и Шидак, может применяться нормальная аппроксимация статистики рангового критерия

$$Z = \frac{S - ES}{\sqrt{DS}},$$

где ES – математическое ожидание,
DS – дисперсия, которая может быть скорректирована при наличии связок.
Параметры нормального распределения вычисляются по формулам, данным Хеттсманпергером,

$$ES = n_1\bar{a},$$

$$DS = \frac{n_1 n_2}{N(N-1)} \sum_{i=1}^{N} \left(a(R_i) - \bar{a}\right)^2,$$

где $$\bar{a} = \frac{1}{N} \sum_{i=1}^{N} a(R_i).$$

Обратите внимание, что суммирование в параметрах нормальной аппроксимации производится по обеим выборкам, тогда как статистика критерия вычисляется для одной (любой) из выборок.
Возможно и точное вычисление P–значений ранговых критериев. Например, методика точного вычисления критериев Вилкоксона полностью совпадает с соответствующими критериями рандомизации компонент, представленными в главе «Точные критерии», с той разницей, что все манипуляции производятся не с вариантами выборок, а с их рангами. По этой причине критерии Вилкоксона могут быть интерпретированы как критерии ранговой рандомизации.

В различных программных продуктах результаты вычисления того или иного критерия могут различаться. Это вызвано введением поправок. Причем в программах все указанные поправки могут учитываться одновременно, по отдельности или не учитываться вовсе, что ведет к возможному получению различных результатов расчета в разных программах.

4.3.3.1.1. Учет связок

Связкой (ties) называют совпадающие ранги. При наличии связок статистика критерия (точнее, дисперсия при нормальной аппроксимации статистики критерия) обычно корректируется с помощью особым образом вычисляемой поправки на объединение рангов.

4.3.3.1.2. Учет поправки на непрерывность

Поправка на непрерывность (continuity) фактически вводится в формулу вычисления нормальной аппроксимации статистики критерия, т.к. дискретное распределение ранговой статистики аппроксимируется непрерывным нормальным распределением.

См. результаты Пури (Puri), Раджарама (Rajaram).

4.3.3.1.3. Критерий Вилкоксона для независимых выборок

W–критерий Вилкоксона (критерий ранговых сумм Вилкоксона, двухвыборочный критерий Вилкоксона, статистика ранговой суммы Уилкоксона, Wilcoxon signed–rank test, Wilcoxon sum–of–ranks test for comparing two unmatched samples) применяется для проверки однородности двух независимых совокупностей одинаковой или разной численности. Выборки могут принадлежать порядковой или количественной шкале.
Вычисление статистики критерия производится по формуле:

$$W = \min\left(\sum_{i=1}^{n_1} R_i, \sum_{i=1}^{n_2} S_i \right),$$

где $R_i, i = 1,2,...,n_1$ – ранги выборки, имеющей наименьшую сумму рангов,
$S_i, i = 1,2,...,n_2$ – ранги выборки, имеющей наибольшую сумму рангов.
Другой прием – в качестве статистики критерия берется сумма рангов выборки наименьшей численности, хотя принципиальной разницы тут нет.
Вычисленное значение статистики критерия сравнивается с точным критическим значением, однако при большой численности выборок данными формулами пользуются неохотно из–за определенных вычислительных сложностей. Формулы пригодны для построения таблиц, но для практического вычисления значимости критерия применяется подход, учитывающий факт, что статистика
$\frac{W - EW}{\sqrt{DW}}$ распределена по стандартному нормальному закону.
Здесь обозначено:
$EW = n_1(N + 1) / 2$ – математическое ожидание,
дисперсия без связок $DW = n_1 n_2(N + 1) / 12$
или, при наличии связок,

$$DW = \frac{n_1 n_2}{12}\left[N + 1 - \frac{b}{N(N-1)} \right],$$

$N = n_1 + n_2$ – численность объединенной выборки.

$b = \sum_{j=1}^{g} t_j (t_j^2 - 1)$ – поправка на объединение рангов,
где t_j, $j = 1,2,...,g$ – численность связки,
g – число связок.
При расчете числа связок, при наличии хотя бы одной связки, учитываются также все группы с численностью 1, что, однако, исключает учет данных групп (подобно критерию Ансари–Бредли) из–за особенностей вычисления поправки на объединение рангов. В отчете Хельзель (Helsel) с соавт. со ссылкой на Коновер (Conover) приводится иной способ учета связок.
Если полученное значение статистики превышает 0,02, то в формулу вводится поправка на непрерывность: считается, что новое значение наименьшей суммы рангов равно $W + 0,5$.
В различных программных продуктах результаты вычисления критерия Вилкоксона могут незначительно различаться. Это вызвано введением поправок, рассмотренных выше, а именно:

- Учет связок (ties).
- Учет поправки на непрерывность (continuity).

В программах указанные поправки могут учитываться одновременно, по отдельности или не учитываться вовсе.
Критерий рекомендуется для выборок умеренной численности (численность каждой выборки от 12 до 40).
Имеется простая формула связи рассматриваемого критерия с критерием Манна–Уитни, поэтому представленный тест в некоторых источниках носит наименование критерия Вилкоксона–Манна–Уитни.

См. Когана с соавт., Черник (Chernick) с соавт., статью ЛаВанж (LaVange) с соавт. Точное вычисление распределения статистики Вилкоксона см. в работе Лемана (Lehman). Подробный анализ проблем, возникающих при применении критерия, см. в учебнике Орлова. Влияние различных поправок в критериях Вилкоксона–Манна–Уитни рассмотрено в работе Бергмана (Bergmann) с соавт. На связь статистики критерия Вилкоксона и площади, отсекаемой ROC кривой (AUC), указано в монографии Власова.

4.3.3.1.4. Критерий Вилкоксона для связанных выборок

T–критерий Вилкоксона (знаковый ранговый критерий Уилкоксона, критерий знаковых рангов Уилкоксона, Wilcoxon signed–ranks test for matched pairs), в отличие от *W*–критерия Вилкоксона, применяется для проверки однородности двух совокупностей с попарно сопряженными вариантами. Выборки могут принадлежать порядковой или количественной шкале.
Критерием проверяется статистическая значимость нулевой гипотезы о том, что распределение случайных величин симметрично относительно нуля. Эти случайные величины в рассматриваемом случае представляют собой разности случайных величин, соответствующих двум другим выборкам, поэтому часто критерий называют одновыборочным критерием Вилкоксона. Другое название критерия – критерий Вилкоксона для сопряженных пар, *T*–дельта–критерий, *W*–критерий Вилкоксона либо просто критерий Вилкоксона.
Методика приближенного вычисления похожа на процедуру вычисления *W*–критерия Вилкоксона, однако здесь мы оперируем абсолютными величинами разностей вариант. Массив разностей ранжируется. Если среди разностей есть нулевые, они отбрасываются (при этом численность сокращается на число отброшенных нулевых разностей). Затем рангам

добавляются знаки разностей, и вычисляется наименьшая из сумм положительных W^+ рангов, которая сравнивается с точным критическим значением, однако при большой численности выборок данными формулами пользуются неохотно из–за определенных вычислительных сложностей. Формулы пригодны для построения таблиц, но для практического вычисления значимости критерия применяется подход, учитывающий факт, что статистика

$\frac{W^+ - EW^+}{\sqrt{DW^+}}$ распределена по стандартному нормальному закону.

Здесь обозначено:

где $EW^+ = N(N + 1) / 4$ – математическое ожидание,

дисперсия без связок $DW^+ = N(N+1)(2N+1)/24$

$$DW^+ = \frac{1}{24}\left[N(N+1)(2N+1) - \frac{b}{2} \right],$$

или, при наличии связок,

N – численность каждого ряда (после отбрасывания нулевых значений),

$b = \sum_{j=1}^{g} t_j (t_j^2 - 1)$ – поправка на объединение рангов,

где t_j, $j = 1,2,...,g$ – численность связки,

g – число связок, причем, при наличии хотя бы одной связки, следовало бы учитывать также все группы с численностью 1; однако учет данных групп (подобно критерию Ансари–Бредли) из–за особенностей вычисления поправки на объединение рангов исключен из алгоритма (данные слагаемые – нулевые).

Критерий рекомендуется для выборок умеренной численности (численность каждой выборки от 12 до 40).

Критерий описан практически во всех источниках, посвященных проверке гипотез, непараметрической статистике и ранговым критериям, в частности. Критерий популярен среди биостатистиков. См. например, книгу Черник (Chernick) с соавт.

4.3.3.1.5. Критерий Манна–Уитни

U–критерий Манна–Уитни (Вилкоксона–Манна–Уитни) применяется для проверки однородности двух независимых совокупностей одинаковой или разной численности. Выборки могут принадлежать порядковой или количественной шкале. Наблюдения должны быть независимыми (непарными). Вычисления могут производиться по формулам (в источниках описаны различные схемы, приводящие к аналогичным результатам)

$U_1 = n_1 n_2 + n_1(n_1 + 1) / 2 - R_1$,

$U_2 = n_1 n_2 + n_2(n_2 + 1) / 2 - R_2$,

$U = \max(U_1, U_2)$,

где R_1 и R_2 – суммы рангов выборок,

n_1 и n_2 – численности соответствующих выборок.

Вычисленное значение статистики критерия сравнивается с точным критическим значением распределения Манна–Уитни, однако при большой численности выборок данными формулами пользуются неохотно из–за определенных вычислительных сложностей. Формулы пригодны для построения таблиц, но для практического вычисления значимости критерия применяется подход, учитывающий факт, что статистика

$\frac{U - EU}{\sqrt{DU}}$,

где $EU = n_1 n_2 / 2$ – математическое ожидание,
$DU = n_1 n_2 (N + 1) / 12$ – дисперсия, которая в случае наличия связок корректируется,
$N = n_1 + n_2$ – численность объединенной выборки.
распределена по стандартному нормальному закону.
Критерий эквивалентен критерию Вилкоксона. Статистические свойства *U*–критерия Манна–Уитни и *W*–критерия Вилкоксона совпадают. Отметим только, что в критерии Манна–Уитни не используются поправки, разработанные для критерия Вилкоксона, поэтому результаты расчета для одних и тех же данных могут различаться.

См. монографию Уилкса. Точное вычисление распределения статистики Манна–Уитни см. в работе Манна (Mann) с соавт.

4.3.3.1.6. Критерий Ван дер Вардена

Ранговый *X*–критерий Ван дер Вардена (Van der Waerden's *X*–test) применяется для проверки однородности двух независимых совокупностей одинаковой или разной численности. Выборки могут принадлежать порядковой или количественной шкале. Статистика критерия вычисляется по формуле

$$X = \sum_{i=1}^{n_1} \Psi\left(\frac{R_i}{N+1}\right),$$

где n_1 – численность одной выборки,
n_2 – численность другой выборки,
R_i, $i = 1,2,...,n_1$ – ранговые метки одной из выборок,
$\Psi(.)$ – функция, обратная функции стандартного нормального распределения,
$N = n_1 + n_2$ – численность объединенной выборки.
Практически значимость может вычисляться посредством нормальной аппроксимации критического значения критерия. При этом модифицированная статистика

$$\frac{X - EX}{\sqrt{DX}},$$

где $EX = 0$ – математическое ожидание,

$$DX = \frac{n_1 n_2}{N(N-1)} \sum_{i=1}^{N} \left[\Psi\left(\frac{i}{N+1}\right)\right]^2$$ – дисперсия.

распределена по стандартному нормальному закону.

См. также родственный представленному тесту критерий Флигнера–Киллина (Fligner–Killeen test of homogeneity of variances), описанный Гарретом (Garrett) с соавт.

4.3.3.1.7. Критерий Сэвиджа

Критерий Сэвиджа предназначен для проверки применяется для проверки однородности двух независимых совокупностей одинаковой или разной численности. Выборки могут принадлежать порядковой или количественной шкале. Предложено несколько эквивалентных форм записи формулы вычисления статистики критерия. В программе статистика критерия вычисляется по формуле, полагаемой наиболее удобной с практической точки зрения,

$$S = \sum_{i=1}^{n_1} \sum_{j=N+1-R_i}^{N} \frac{1}{j},$$

где R_i, $i = 1,2,...,n_1$ – ранги выборки с наибольшей численностью,
$N = n_1 + n_2$ – численность объединенной выборки,

n_1 – численность выборки с наибольшей численностью,
n_2 – численность выборки с наименьшей численностью.
Практически значимость может вычисляться посредством нормальной аппроксимации критического значения критерия. При этом модифицированная статистика
$\frac{S - ES}{\sqrt{DS}}$,
где $ES = n_1$ – математическое ожидание,
$DS = \frac{n_1 n_2}{N-1}\left(1 - \frac{1}{N}\sum_{j=1}^{N}\frac{1}{j}\right)$ – дисперсия.
распределена по стандартному нормальному закону.
Обобщением критерия Сэвиджа является широко известный критерий Кокса (логарифмический ранговый критерий), иногда называемый обобщенным критерием Сэвиджа, предназначенный для анализа цензурированных выборок и представленный в главе «Анализ выживаемости».

См. монографию Скрипника с соавт.

4.3.3.1.8. Критерий Ансари–Бредли

Критерий Ансари–Бредли (Фройнда и Ансари, Freund–Ansari test) применяется для проверки однородности двух независимых совокупностей одинаковой или разной численности. Выборки могут принадлежать порядковой или количественной шкале. Статистика критерия вычисляется по формуле

$$W = \sum_{i=1}^{n_1} R_i,$$

где R_i, $i = 1,2,...,n_1$ – ранги выборки с наибольшей численностью,
n_1 – численность одной выборки,
n_2 – численность другой выборки,
$N = n_1 + n_2$ – численность объединенной выборки.
Для построения критерия ранжирование производится особым образом. Если N четно, ранги присваиваются по схеме $1,2,3,...,N/2,N/2,...,3,2,1$. Если N нечетно, ранги присваиваются по схеме $1,2,3,...,(N-1)/2,(N+1)/2,...,3,2,1$. При наличии одинаковых наблюдений используются связанные (средние) ранги.
Практически значимость может вычисляться посредством нормальной аппроксимации критического значения критерия. При этом модифицированная статистика
$\frac{W - EW}{\sqrt{DW}}$ распределена по стандартному нормальному закону.
Здесь обозначено:
$EW = n_1(N+2)/4$ – математическое ожидание для четного N,
$EW = n_1(N+1)^2/4/N$ – математическое ожидание для нечетного N,
дисперсия для четного N без связок $DW = \frac{n_1 n_2 (N+2)(N-2)}{48(N-1)}$,
или, при наличии связок, $DW = \frac{n_1 n_2 \left[16b - N(N+2)^2\right]}{48N(N-1)}$,

дисперсия для нечетного N без связок $$DW = \frac{n_1 n_2 (N+1)(N^2+3)}{48N^2}$$

или, при наличии связок, $$DW = \frac{n_1 n_2 \left[16Nb - (N+1)^4\right]}{48N^2(N-1)},$$

$b = \sum_{j=1}^{g} t_j r_j^2$ – поправка на объединение рангов,

где t_j, $j = 1,2,...,g$ – численность связки,
r_j, $j = 1,2,...,g$ – средний ранг в связке,
g – число связок, причем, при наличии хотя бы одной связки, учитываются также и все группы с численностью 1.

Полное описание метода дано в монографии Шескин (Sheskin). См. также Джонсона с соавт., Петровича с соавт.

4.3.3.1.9. Критерий Клотца

Критерий Клотца (Klotz test) применяется для проверки однородности двух независимых совокупностей одинаковой или разной численности. Выборки могут принадлежать порядковой или количественной шкале. Статистика критерия вычисляется по формуле

$$K = \sum_{i=1}^{n_1} \left[\Psi\left(\frac{R_i}{N+1} \right) \right]^2,$$

где n_1 – численность одной выборки,
n_2 – численность другой выборки,
R_i, $i = 1,2,...,n_1$ – ранговые метки одной из выборок,
$\Psi(.)$ – функция, обратная функции стандартного нормального распределения,
$N = n_1 + n_2$ – численность объединенной выборки.
Практически значимость может вычисляться посредством нормальной аппроксимации критического значения критерия. При этом модифицированная статистика

$$\frac{K - EK}{\sqrt{DK}},$$

где $EK = \frac{n_1}{N} \sum_{i=1}^{N} \left[\Psi\left(\frac{i}{N+1} \right) \right]^2$ – математическое ожидание,

$DK = \frac{n_1 n_2}{N(N-1)} \sum_{i=1}^{N} \left[\Psi\left(\frac{i}{N+1} \right) \right]^4 - \frac{n_2}{n_1(N-1)} (EK)^2$ – дисперсия,

распределена по стандартному нормальному закону.

Критерий описан в монографии Гаека (Hajek) с совт., в книге Кулаичева, справочном издании Айвазяна с соавт. (1983).

4.3.3.1.10. Критерий Зигеля–Тьюки

Критерий Зигеля–Тьюки (Сиджела–Тьюки, Сайджела–Тьюки, Siegel–Tukey test) применяется для проверки однородности двух независимых совокупностей одинаковой или разной численности. Выборки могут принадлежать порядковой или количественной шкале. Статистика критерия вычисляется по формуле

$$T = \sum_{i=1}^{n_1} R_i,$$

где R_i, $i = 1,2,...,n_1$ – ранги выборки с наибольшей численностью,
n_1 – численность одной выборки,
n_2 – численность другой выборки.
Для построения критерия ранжирование производится особым образом. Ранги присваиваются по схеме $1,\underline{4,5},\underline{8,9},...,\underline{7,6},\underline{3,2}$ до исчерпания вариант объединенной выборки.
При наличии одинаковых наблюдений используются связанные (средние) ранги.
В названии рассмотренного критерия на самом деле объединены два теста – критерий Зигеля и критерий Тьюки. Эти тесты различаются только направлением ранжирования вариант. Присвоение рангов вариантам в схеме Тьюки начинается не слева направо, как в схеме Зигеля, а справа налево. Построенный таким способом критерий обозначается как T'.
Практически значимость может вычисляться посредством нормальной аппроксимации критического значения критерия. При этом модифицированная статистика

$$\frac{T - ET}{\sqrt{DT}},$$

где $ET = n_1(N + 1) / 2$ – математическое ожидание,
$DT = n_1 n_2(N + 1) / 12$ – дисперсия,
$N = n_1 + n_2$ – численность объединенной выборки,
распределена по стандартному нормальному закону.
Представленный критерий эквивалентен критерию Ансари–Бредли.

Критерий представлен во многих источниках. Описание см. в статье Клотца (Klotz), монографиях Благовещенского с соавт., Когана с соавт., Шескин (Sheskin).

4.3.3.1.11. Критерий Коновера

Критерий Коновера (Conover's two–sample squared ranks test for equality of variance) применяется для проверки однородности двух независимых совокупностей одинаковой или разной численности. Выборки могут принадлежать порядковой или количественной шкале.
Перед расчетом статистики критерия исходные выборки подвергаются преобразованиям по формулам

$$U_i = \left|x_i - m_x\right|, i = 1,2,...,n_1,$$

$$V_i = \left|y_i - m_y\right|, i = 1,2,...,n_2,$$

где x_i, $i = 1,2,...,n_1$ – одна из выборок,
y_i, $i = 1,2,...,n_2$ – другая из выборок,
n_1 – численность одной выборки,
n_2 – численность другой выборки,

$m_x = \frac{1}{n_1}\sum_{i=1}^{n_1} x_i$ – среднее значение одной выборки,

$m_y = \frac{1}{n_2}\sum_{i=1}^{n_2} y_i$ – среднее значение другой выборки.

Статистика критерия вычисляется по формуле

$$K = \sum_{i=1}^{n_1}\left[R(U_i)\right]^2,$$

R_i, i = 1,2,...,n_1 – ранговые метки одной из выборок.
Практически значимость может вычисляться посредством нормальной аппроксимации критического значения критерия. При этом модифицированная статистика

$$\frac{K - EK}{\sqrt{DK}},$$

где $EK = n_1\overline{R}^2$ – математическое ожидание,

$DK = \frac{n_1 n_2}{N(N-1)}\sum_{i=1}^{N}\left[R(U_i)\right]^4 - \frac{n_1 n_2}{N-1}\left(\overline{R}^2\right)^2$ – дисперсия,

$\overline{R}^2 = \frac{1}{N}\sum_{i=1}^{N}\left[R(U_i)\right]^2$ – среднее значение суммы квадратов рангов,

$N = n_1 + n_2$ – численность объединенной выборки,
распределена по стандартному нормальному закону.

Описание метода приводится в монографии Коновера, книге Спрента (Sprent) с соавт., статьях Коновера с соавт., работах Вилкокса (Wilcox), диссертации Бучана (Buchan).

4.3.3.1.12. Критерий Муда–Брауна

Медианный критерий Муда–Брауна (критерий Муда References) применяется для проверки однородности двух независимых совокупностей одинаковой или разной численности. Выборки могут принадлежать порядковой или количественной шкале. Статистика критерия вычисляется по формуле

$$V_+ = \sum_{i=1}^{n_1} sign\left[R_i - \frac{N+1}{2}\right],$$

где R_i, i = 1,2,...,n_1 – ранги выборки с наименьшей численностью,
n_1 – численность одной выборки,
n_2 – численность другой выборки,
$N = n_1 + n_2$ – численность объединенной выборки.
Практически значимость может вычисляться посредством нормальной аппроксимации критического значения критерия. При этом модифицированная статистика

$$\frac{V_+ - EV_+}{\sqrt{DV_+}},$$

где $EV_+ = \frac{n_1}{2}$ – математическое ожидание,

$DV_+ = \frac{n_1 n_2}{4(N-1)}$ – дисперсия.

распределена по стандартному нормальному закону.

4.3.3.2. Критерии на основе сравнения функций распределения

Идея сравнения функций распределения (А.Н. Колмогоров, 1933 г.) оказалась наиболее плодотворной при конструировании критериев согласия. Более подробная информация дана в главе «Проверка нормальности распределения».
Идея стала полезной и при сравнении эмпирических функций распределения эмпирических выборок. Из критериев данного класса нами представлены:

- критерий Смирнова,

- критерий Лемана–Розенблатта,
- критерий Койпера.

Существует группа критериев на основе распределения χ^2, предназначенная для анализа таблиц сопряженности, являющихся продуктом сопоставления эмпирических выборок.
Из критериев данного класса нами представлены:

- критерий Мак–Немара (для сопряженных бинарных выборок) в его асимптотическом варианте,
- критерий хи–квадрат (для независимых бинарных выборок),
- критерий медианы (для порядковых или количественных выборок).

Для применения критерия Мак–Немара и критерия хи–квадрат (в представленной форме) анализируемые выборки должны принадлежать дихотомической шкале измерения, т. е. (согласно принятому здесь соглашению) состоять только из нулей и единиц, причем нуль означает отсутствие признака, а единица означает наличие признака.

4.3.3.2.1. Критерий Смирнова

Критерий Смирнова (критерий Колмогорова–Смирнова, Kolmogorov–Smirnov test, Kolmogorov–Smirnoff test) применяется для проверки однородности двух независимых совокупностей одинаковой или разной численности. Проверяется нулевая гипотеза о том, являются ли одинаковыми непрерывные функции распределения генеральных совокупностей, из которых взяты выборки. Иначе, проверяется принадлежность двух выборок одной и той же генеральной совокупности при условии непрерывности ее функции распределения.
Статистика критерия имеет вид

$$D_{m,n} = \sup_{-\infty<x<\infty} \left|F_n(x') - G_m(x)\right|,$$

где $D_{m,n}$ – максимальная разность между частостями рядов x' и x,
m и n – численности вариационных рядов, построенных по эмпирическим выборкам,
$G_m(.)$ и $F_n(.)$ – соответствующие эмпирические функции распределения.
Практически вычисления производятся по формулам:

$$D = \max\left(D^+_{m,n}, D^-_{m,n}\right),$$

$$D^+_{m,n} = \max_{1\le r\le m}\left(\frac{r}{m} - F_n(x'_r)\right) = \max_{1\le s\le n}\left(G_m(x_s) - \frac{s-1}{n}\right),$$

$$D^-_{m,n} = \max_{1\le r\le m}\left(F_n(x'_r) - \frac{r-1}{m}\right) = \max_{1\le s\le n}\left(\frac{s}{n} - G_m(x_s)\right).$$

Функция распределения модифицированной статистики критерия $D\sqrt{N}$ (имеются и иные формулы) при $N = mn / (m + n) \to \infty$ сходится к функции распределения Колмогорова.
Критерий рекомендуется для выборок средней и большой численности (численность каждой выборки от 40 до 100 и выше). При большей численности выборок становится больше теоретических оснований для применения параметрических тестов.

См. учебник Айвазяна с соавт. (критерий однородности Смирнова), статью Лемешко с соавт.
Статистика рассматриваемого теста может быть записана как максимум линейных ранговых статистик – модифицированных статистик Муда. Поэтому некоторые авторы рассматривают метод в курсе ранговых критериев. Гудман (Goodman) предложил аппроксимировать статистику критерия распределением χ^2 (статистика хи–квадрат Гудмана, Goodman approximation of Kolmogorov–Smirnov test).

4.3.3.2.2. Критерий Лемана–Розенблатта

Критерий Лемана–Розенблатта (Lehmann–Rosenblatt test, Lehmann's two–sample test) применяется для проверки однородности двух независимых совокупностей одинаковой или разной численности. Проверяется нулевая гипотеза о том, являются ли одинаковыми непрерывные функции распределения генеральных совокупностей, из которых взяты выборки. Иначе, проверяется принадлежность двух выборок одной и той же генеральной совокупности при условии непрерывности ее функции распределения.
Статистика критерия вычисляется по формуле

$$T = \frac{1}{m+n}\left[\frac{1}{m}\sum_{i=1}^{n}(R_i - i)^2 + \frac{1}{n}\sum_{j=1}^{m}(S_j - j)^2 - \frac{4mn-1}{6}\right],$$

где R_i, $i = 1,2,...,n$ – ранги одной выборки,
S_j, $j = 1,2,...,m$ – ранги другой выборки.
n и m – численности выборок.
Функция распределения статистики критерия при $m, n \to \infty$ совпадает с функцией распределения a_1 критериев типа омега–квадрат.

См. таблицы Большева с соавт., книгу Мартынова, статьи Лемана (Lehmann), Розенблатта (Rosenblatt), Сандрама (Sundrum), Вегнера (Wegner), Лемешко (Lemeshko) с соавт., Лемешко с соавт., Фиша (Fisz).

4.3.3.2.3. Критерий Койпера

Критерий Койпера (Kuiper test) применяется для проверки однородности двух независимых совокупностей одинаковой или разной численности. Проверяется нулевая гипотеза о том, являются ли одинаковыми непрерывные функции распределения генеральных совокупностей, из которых взяты выборки. Иначе, проверяется принадлежность двух выборок одной и той же генеральной совокупности при условии непрерывности ее функции распределения.
Статистика критерия имеет вид

$$V = D^{+}_{m,n} + D^{-}_{m,n},$$

$$D^{+}_{m,n} = \sup_{-\infty<x<\infty}\left|F_n(x') - G_m(x)\right|,$$

$$D^{-}_{m,n} = \sup_{-\infty<x<\infty}\left|G_m(x) - F_n(x')\right|,$$

где $D^{+}_{m.n}$ – максимальная разность $F_n(x')$ «выше» $G_m(x)$,
$D^{-}_{m.n}$ – максимальная разность $F_n(x')$ «ниже» $G_m(x)$,
$F_n(x')$ и $G_m(x)$ – эмпирические функции распределения вариационных рядов x' и x, построенных по эмпирическим выборкам,
n и m – численности вариационных рядов x' и x.

Функция распределения модифицированной статистики критерия $V\sqrt{N}$ (имеются и иные формулы) при $N = mn / (m + n) \to \infty$ сходится к функции распределения Койпера.
Критерий рекомендуется для выборок средней и большой численности (численность каждой выборки от 40 до 100 и выше). При большей численности выборок становится больше теоретических оснований для применения параметрических тестов.

Ряд авторов полагает критерий Койпера предпочтительным относительно критерия

Смирнова. Имеются литературные данные о попытках применения критерия Койпера, подобно критериям типа Колмогорова, для проверки согласия распределений (подробнее о проверке нормальности см. в главе «Проверка нормальности распределения»). См. также статью Цирроне (Cirrone) с соавт.

4.3.3.2.4. Критерий Мак–Немара

Критерий Мак–Немара (McNemar's chi–square test) применяется для проверки нулевой гипотезы о том, отобраны ли две исследуемые попарно сопряженные бинарные выборки из генеральных совокупностей с одинаковой частотой встречаемости изучаемого эффекта. Рассматриваемый метод предназначен для обработки так называемых четырехпольных (четырехклеточных) таблиц, иначе таблиц типа 2 x 2. Анализируемые выборки должны принадлежать дихотомической шкале измерения, т. е. состоять только из нулей и единиц, причем нуль означает отсутствие признака, а единица означает наличие признака. Перед применением метода необходимо ознакомиться с разделом, посвященным описанию таблиц 2 x 2.
Вычисление статистики критерия производится по формуле:

$$X^2 = \frac{(|b-c|-Y)^2}{b+c},$$

где b – число пар наблюдений с эффектом A во первой выборке, но без эффекта B во второй выборке,
c – число наблюдений без эффекта A в первой выборке, но с эффектом B во второй выборке,
$Y = 0$ – поправка на непрерывность (поправка Йейтса), в случае ее неучета,
$Y = 1$ – в случае учета поправки (режим по умолчанию).
Считается, что при величине $b + c \geq 10$ статистика критерия (двусторонняя гипотеза) удовлетворительно аппроксимируется распределением χ^2 с числом степеней свободы, равным 1. При $b + c < 10$ можно использовать точные методы, представленные в главе «Точные критерии», в котором для полноты изложения представлен также критерий Мак–Немара, дополненный его точным вариантом.

О вычислении критерия и точном распределении его статистики см. заметки Беннетта (Bennett) с соавт., материалы компании Cytel. Существует вариант рассмотренного критерия (критерий Стюарта–Максвелла, Stuart–Maxwell test), предназначенный для анализа таблиц типа k x k, получающихся из номинальных выборок с числом градаций признаков, равным k. Аналогичное назначение имеют критерий симметрии Баукера (Bowker's test of symmetry) и критерий Бхапкара (Bhapkar's test). Данные методы представлены в главе «Кросстабуляция».

4.3.3.2.5. Критерий хи–квадрат

Критерий хи–квадрат применяется для проверки нулевой гипотезы о том, отобраны ли две исследуемые независимые бинарные выборки из генеральных совокупностей с одинаковой частотой встречаемости изучаемого эффекта. Рассматриваемый метод предназначен для обработки так называемых четырехпольных (четырехклеточных) таблиц, или таблиц 2 x 2. Анализируемые выборки должны принадлежать дихотомической шкале измерения, т. е. состоять только из нулей и единиц, причем нуль означает отсутствие признака, а единица означает наличие признака. Перед применением метода необходимо ознакомиться с разделом, посвященным описанию таблиц 2 x 2.
Вычисление статистики критерия для данного случая производится по формуле

$$X^2 = \sum_{i=1}^{2}\sum_{j=1}^{2} \frac{(|\ f_{ij} - \hat{f}_{ij}\ | - Y)^2}{\hat{f}_{ij}},$$

где f_{ij}, $i, j = 1,2$ – вычисленные частоты – значения в клетках a, b, c, d в дальнейшем для наглядности обозначим их этими же литерами,

$\hat{f}_{ij}, i, j = 1,2,$ – соответствующие ожидаемые частоты, вычисляемые по формулам:

$$\hat{f}_{11} = \frac{(a+b)(a+c)}{n},$$

$$\hat{f}_{12} = \frac{(a+b)(b+d)}{n},$$

$$\hat{f}_{21} = \frac{(c+d)(a+c)}{n},$$

$$\hat{f}_{22} = \frac{(c+d)(b+d)}{n},$$

где a – число наблюдений с эффектом A в первой выборке,
b – число наблюдений без эффекта A в первой выборке,
c – число наблюдений с эффектом A во второй выборке,
d – число наблюдений без эффекта A во второй выборке,
$n = a + b + c + d$ – общая численность всех наблюдений,
$Y = 0$ – поправка на непрерывность (поправка Йейтса), в случае ее неучета,
$Y = 0,5$ – в случае учета поправки (режим по умолчанию).
Статистика критерия удовлетворительно аппроксимируется распределением χ^2 с числом степеней свободы, равным 1.

Критерий стандартизован в отечественных и международных нормативных документах. См., например, методическую разработку Лванга (Lwanga) и Тыэ (Tye). Поправки обсуждаются в статье Лузен (Loosen). Существует вариант критерия для анализа таблиц типа k x k, получающихся из выборок с числом градаций признаков более 2, представленный в главе «Кросстабуляция».

4.3.3.2.6. Критерий медианы

Критерий медианы (медианный критерий) применяется для проверки однородности двух независимых совокупностей одинаковой или разной численности. Выборки могут принадлежать порядковой или количественной шкале. Этапы вычисления критерия для двух выборок численностями n_1 и n_2 включают:

- Объединение исходных выборок, вычисление медианы объединенной выборки.
- Формирование таблицы типа 2 x 2 по следующему правилу: в ячейку A заносится число отметок первой выборки, превышающих медиану; в ячейку B заносится число отметок второй выборки, превышающих медиану; в ячейки C и D заносится число отметок, соответственно, первой и второй выборок, не превышающих медиану.

В случае $n_1 > 15$ и/или $n_2 > 15$ к полученной таблице применяется критерий хи–квадрат с числом степеней свободы, равным 1.
Существует вариант критерия для анализа таблиц типа 2 x k, получающихся из k порядковых выборок с числом вариаций признаков, равным 2. Этот метод в настоящем программном обеспечении не представлен.

4.3.3.3. Прочие критерии

В программе реализованы также некоторые традиционно применяемые критерии, которые трудно отнести к перечисленным выше типам. В программе представлены:

- критерий серий Вальда–Вольфовица.

Точная версия критерия серий реализована в главе «Точные критерии».

4.3.3.3.1. Критерий серий Вальда–Вольфовица

Критерий серий Вальда–Вольфовица (Wald–Wolfowitz runs test) применяется для проверки однородности двух независимых совокупностей одинаковой или разной численности. Проверяется нулевая гипотеза о равенстве целого ряда параметров двух сравниваемых выборок, включая медианы и коэффициенты асимметрии. Критерий применяется в случае, если исследователя интересует, имеют ли место любые различия между совокупностями. Выборки могут принадлежать порядковой или количественной шкале. Суть расчета заключается в объединении выборок с численностями n_1 и n_2 в одну выборку общей численностью $N = n_1 + n_2$, ее сортировке по возрастанию или убыванию и подсчете числа серий элементов R, относящихся к первой и второй выборкам.
Значимость при численности выборок $n_1 > 20$ и $n_2 > 20$ может вычисляться посредством нормальной аппроксимации. При этом модифицированная статистика

$$\frac{|R - ER| - 0{,}5}{\sqrt{DR}},$$

где $ER = \frac{2n_1n_2}{N} + 1$ – математическое ожидание,

$DR = \frac{2n_1n_2(2n_1n_2 - N)}{N^2(N-1)}$ – дисперсия,

0,5 – поправка на непрерывность,
распределена по стандартному нормальному закону.
Точная версия критерия реализована в главе «Точные критерии».

Варианты критерия серий и аппроксимации представлены в монографии Браунли. Метод описан в справочнике Руниона, книге Зайцева, диссертации Хешл (Heschl). Замечания о применении см. в книге Гаека с соавт., статье Камень с соавт.

4.3.4. Таблицы 2 x 2

Рассчитываются следующие продукты анализа таблиц типа 2 x 2

- относительный риск,
- отношение шансов,
- разность долей,
- прогностичность.

Таблицы 2 x 2 возникают в результаты сопоставления двух бинарных (дихотомических) выборок, т. е. выборок, состоящих из значений 1 и 0, причем под значением 1 понимают наличие признака, под значением 0 понимают отсутствие признака.
Для расчета пользователь может указать одну из опций расчета таблицы:

- Для независимых выборок.
- Для связанных (парных) выборок.
- Расчет по готовой таблице для независимых выборок.
- Расчет по готовой таблице для связанных выборок.

Важно знать, что таблицы типа 2 x 2 могут быть получены из исходных выборок различными способами, в зависимости от того, являются ли выборки независимыми или связанными. Более подробная информация о типах и представлениях указанных данных см. в главе «Введение в практичекий анализ».
Для ввода готовой таблицы 2 x 2 в настоящем программном обеспечении в качестве первого столбца данной заранее составленной пользователем таблицы укажите «Интервал выборки 1», в качестве второго столбца «Интервал выборки 2».

См. монографию Ньюмен (Newman).

4.3.4.1. Относительный риск

Относительный риск (relative risk, RR), или отношение рисков – отношение заболеваемости среди лиц, подвергавшихся и не подвергавшихся воздействию факторов риска.
Относительный риск не несет информации о величине абсолютного риска (заболеваемости). Даже при высоких значениях относительного риска абсолютный риск может быть совсем небольшим, если заболевание редкое. Относительный риск показывает силу связи между воздействием и заболеванием.
Рассматриваемый метод предназначен для обработки так называемых четырехпольных (четырехклеточных) таблиц, или таблиц 2 x 2. Анализируемые выборки должны принадлежать дихотомической шкале измерения, т. е. состоять только из нулей и единиц, причем нуль означает отсутствие признака, а единица означает наличие признака. Перед применением метода необходимо ознакомиться с разделом, посвященным описанию таблиц 2 x 2.
Вычисление отношения рисков производится по формуле

$$RR = \frac{n_{11}(n_{21} + n_{22})}{n_{21}(n_{11} + n_{12})},$$

где n_{11}, n_{12}, n_{21}, n_{22} – ячейки таблицы.
Двусторонний доверительный интервал вычисляется по формуле
$I_{RR} = (RR - \Psi((1 + \beta) / 2)S_{RR}; RR + \Psi((1 + \beta) / 2)S_{RR})$,
где $\Psi(.)$ – обратная функция стандартного нормального распределения,
β – доверительный уровень, выраженный в долях,
S_{RR} – стандартная ошибка отношения рисков.
Стандартная ошибка логарифма отношения рисков вычисляется по формуле

$$S_{\ln(RR)} = \sqrt{\frac{1}{n_{11}} - \frac{1}{n_{11} + n_{12}} + \frac{1}{n_{21}} - \frac{1}{n_{21} + n_{22}}}.$$

Таким образом, окончательная формула двустороннего доверительного интервала оцениваемого отношения рисков будет:
$I_{RR} = (\exp(\ln(RR) - \Psi((1 + \beta) / 2)S_{\ln(RR)}); \exp(\ln(RR) + \Psi((1 + \beta) / 2)S_{\ln(RR)}))$.

См. монографии Агрести (Agresti), Хайнес (Haynes) с соавт., статьи Бертелла (Bertell), Гарта (Gart), Барратт (Barratt) с соавт., Подольной с соавт.

4.3.4.2. Отношение шансов

Отношение шансов (odds ratio, OR) определяется как отношение шансов события в одной группе к шансам события в другой группе, или как отношение шансов того, что событие произойдет, к шансам того, что событие не произойдет. В исследованиях случай–контроль отношение шансов используется для оценки относительного результата.

Рассматриваемый метод предназначен для обработки так называемых четырехпольных (четырехклеточных) таблиц, или таблиц 2 x 2. Анализируемые выборки должны принадлежать дихотомической шкале измерения, т. е. состоять только из нулей и единиц, причем нуль означает отсутствие признака, а единица означает наличие признака. Перед применением метода необходимо ознакомиться с разделом, посвященным описанию таблиц 2 x 2.
Вычисление отношения шансов производится по формуле

$$OR = \frac{n_{11}n_{22}}{n_{12}n_{21}},$$

где n_{11}, n_{12}, n_{21}, n_{22} – ячейки таблицы.
Двусторонний доверительный интервал вычисляется по формуле
$I_{OR} = (OR - \Psi((1 + \beta) / 2)S_{OR}; OR + \Psi((1 + \beta) / 2)S_{OR})$,
где $\Psi(.)$ – обратная функция стандартного нормального распределения,
β – доверительный уровень, выраженный в долях,
S_{OR} – стандартная ошибка отношения шансов.
Стандартная ошибка логарифма отношения шансов вычисляется по формуле

$$S_{\ln(OR)} = \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}.$$

Таким образом, окончательная формула двустороннего доверительного интервала оцениваемого отношения шансов будет
$I_{OR} = (\exp(\ln(OR) - \Psi((1 + \beta) / 2)S_{\ln(OR)}); \exp(\ln(OR) + \Psi((1 + \beta) / 2)S_{\ln(OR)}))$.

См. монографии Агрести (Agresti), Хайнес (Haynes) с соавт., статьи Бабич с соавт., Бленда (Bland) с соавт.

4.3.4.3. Разность долей

Рассматриваемый метод вычисления разности долей (difference of proportions) предназначен для анализа так называемых четырехпольных (четырехклеточных) таблиц, или таблиц 2 x 2, возникающих при обработке независимых либо связанных признаков. Анализируемые выборки должны принадлежать дихотомической шкале измерения, т. е. состоять только из нулей и единиц, причем нуль означает отсутствие признака, а единица означает наличие признака. Перед применением метода необходимо ознакомиться с разделом, посвященным описанию таблиц 2 x 2. Предоставляется возможность как ввода исходных массивов, так и готовых таблиц. В последнем случае обязательно необходимо указать, продуктом каких признаков является таблица, ибо формулы их обработки существенно различаются.

4.3.4.3.1. Разность долей в таблице независимых признаков

Вычисление разности долей производится по формуле
$d = |p_2 - p_1|$,
где $p_1 = n_{11} / (n_{11} + n_{12})$ – частота эффекта в первой выборке,
$p_2 = n_{21} / (n_{21} + n_{22})$ – частота эффекта во второй выборке,
n_{11}, n_{12}, n_{21}, n_{22} – ячейки таблицы.
Значимость разности долей тестируется с помощью *z*–критерия, вычисление статистики которого в данном случае производится по формуле

$$z = \frac{|p_2 - p_1| - Y}{\sqrt{\bar{p}(1-\bar{p})\left(\frac{1}{n_{11}+n_{12}} + \frac{1}{n_{21}+n_{22}}\right)}},$$

где $\bar{p} = (n_{11} + n_{21})/(n_{11} + n_{12} + n_{12} + n_{22})$,

$Y = 0$ – поправка на непрерывность (поправка Йейтса), в случае ее неучета,

$Y = 0{,}5 \cdot \left(\frac{1}{n_{11}+n_{12}} + \frac{1}{n_{21}+n_{22}}\right)$ – в случае учета поправки (режим по умолчанию).

Квадрат статистики критерия удовлетворительно аппроксимируется распределением χ^2 с числом степеней свободы, равным 1.

Программой также вычисляется двусторонний доверительный интервал оцениваемой разности долей по формуле Вальда:

$I_{p_2-p_1} = \left(d - \Psi((1+\beta)/2)S_{p_2-p_1} - Y; d + \Psi((1+\beta)/2)S_{p_2-p_1} + Y\right)$,

где $\Psi(.)$ – обратная функция стандартного нормального распределения,

β – доверительный уровень, выраженный в долях.

$S_{p_2-p_1}$ – стандартная ошибка разности долей, вычисляемая по формуле

$$S_{p_2-p_1} = \sqrt{\frac{p_1(1-p_1)}{n_{11}+n_{12}} + \frac{p_2(1-p_2)}{n_{21}+n_{22}}}.$$

4.3.4.3.2. Разность долей в таблице связанных признаков

Вычисление разности долей производится по формуле

$d = |p_2 - p_1|$,

где $p_1 = (n_{11} + n_{12}) / n$ – частота эффекта в первой выборке,

$p_2 = (n_{11} + n_{21}) / n$ – частота эффекта во второй выборке,

n – сумма таблицы, вычисляемая по формуле

$n = n_{11} + n_{12} + n_{21} + n_{22}$,

n_{11}, n_{12}, n_{21}, n_{22} – ячейки таблицы.

Значимость разности долей тестируется с помощью критерия хи-квадрат, вычисление статистики которого в данном случае производится по формуле

$$\chi^2 = \frac{(|n_{21} - n_{12}| - Y)^2}{n},$$

где $Y = 0$ – поправка на непрерывность (поправка Йейтса), в случае ее неучета,

$Y = 1 / n$ – в случае учета поправки (режим по умолчанию).

Квадрат статистики критерия удовлетворительно аппроксимируется распределением χ^2 с числом степеней свободы, равным 1.

Программой также вычисляется двусторонний доверительный интервал оцениваемой разности долей по формуле Вальда (Wald interval for difference of proportions):

$I_{p_2-p_1} = \left(d - \Psi((1+\beta)/2)S_{p_2-p_1} - Y; d + \Psi((1+\beta)/2)S_{p_2-p_1} + Y\right)$,

где стандартная ошибка разности долей вычисляется по формуле

$$S_{p_2-p_1} = \frac{1}{n}\sqrt{b + c - \frac{(b-c)^2}{n}}.$$

Дополнительно программой вычисляется двусторонний доверительный интервал оцениваемой разности долей по уточненной формуле Вальда (adjusted Wald interval for difference of proportions):

$$I_{p_2-p_1} = \left(\left| \hat{p}_2 - \hat{p}_1 \right| - \Psi((1+\beta)/2)\hat{S}_{p_2-p_1} - Y; \left| \hat{p}_2 - \hat{p}_1 \right| + \Psi((1+\beta)/2)\hat{S}_{p_2-p_1} + Y \right),$$

где $\left| \hat{p}_2 - \hat{p}_1 \right| = \left| n_{21} - n_{12} \right| / (n+2),$

$$\hat{S}_{p_2-p_1} = \frac{1}{n+2}\sqrt{b+c+1-\frac{(b-c)^2}{n+2}}.$$

См. монографию Флейс (Fleiss) с соавт., статьи Бурмана (Buhrman), Брауна (Brown) с соавт., Хаука (Hauck) с соавт., Биггерстаффа (Biggerstaff), Чубенко с соавт. Обзор методов вычисления доверительных интервалов оцениваемой разности долей в таблице независимых признаков см. в статье Сантнера (Santner) с соавт. Методы вычисления доверительных интервалов оцениваемой разности долей см. в монографиях Агрести (Agresti), Флейс с соавт., статьях Агрести с соавт., Бергер (Berger) с соавт., Хсие (Hsieh), Сюисса (Suissa) с соавт., Ньюскомб (Newcombe), Гарднер (Gardner) с соавт., Танг (Tang) с соавт.

4.3.4.4. Прогностичность

Рассматриваемая опция дает возможность вычислить общепринятые стандартные показатели прогностичности (прогностической ценности) диагностического теста (predictive values). Это следующие показатели:

- чувствительность (*Se*, sensitivity),
- специфичность (*Sp*, specificity),
- распространенность (*p*, преваленс, доля, prevalence),
- прогностичность положительного результата (*PPV*, positive predictive value),
- прогностичность отрицательного результата (*NPV*, negative predictive value).

Распространенность – это априорная (претестовая) вероятность наличия болезни до того, как стали известны результаты диагностического теста.
Прогностичность (собственно прогностическая ценность) – это апостериорная (посттестовая) вероятность наличия болезни при известном результате исследования. Различают прогностичность положительного результата и прогностичность отрицательного результата. Ниже представлены подробные описания данных показателей, включая формулы вычисления их точечных и интервальных оценок.
Рассматриваемые методы предназначены для обработки так называемых четырехпольных (четырехклеточных) таблиц, или таблиц 2 x 2. Анализируемые выборки должны принадлежать дихотомической шкале измерения, т. е. состоять только из нулей и единиц, причем нуль означает отсутствие признака, а единица означает наличие признака. Перед применением метода необходимо ознакомиться с разделом, посвященным описанию таблиц 2 x 2. Все рассматриваемые в настоящем разделе понятия основаны на следующей четырехпольной таблице:

		Наличие заболевания	
		Присутствует	Отсутствует
Результат диагностического теста	Положительный	n_{11}	n_{12}
	Отрицательный	n_{21}	n_{22}
		n_1	n_0

Положительным результатом теста считается такой результат, который показывает наличие заболевания. Отрицательным результатом теста считается такой результат, который показывает отсутствие заболевания. Обозначено:

n_{11} – численность индивидуумов с наличием заболевания, диагностированных тестом как больные,
n_{21} – численность индивидуумов с наличием заболевания, диагностированных тестом как здоровые,
n_{12} – численность индивидуумов без наличия заболевания, диагностированных тестом как больные,
n_{22} – численность индивидуумов без наличием заболевания, диагностированных тестом как здоровые,
$n_1 = n_{11} + n_{21}$ – численность больных,
$n_0 = n_{12} + n_{22}$ – численность здоровых.
Дополнительные пояснения см. в разделе, посвященном ROC–анализу.

4.3.4.4.1. Чувствительность

Чувствительностью называют долю положительных результатов диагностического теста в популяции. Чем чувствительнее тест, тем выше прогностическая ценность его отрицательного результата.
Вычисление оценки чувствительности производится по формуле

$$Se = \frac{n_{11}}{n_1}.$$

Двусторонний доверительный интервал оцениваемой чувствительности вычисляется по формуле

$$I_{Se} = \left(Se - \Psi((1+\beta)/2)S_{Se}; Se + \Psi((1+\beta)/2)S_{Se}\right),$$

где Ψ(.) – обратная функция стандартного нормального распределения,
β – доверительный уровень, выраженный в долях,
S_{Se} – стандартная ошибка чувствительности.
Стандартная ошибка чувствительности вычисляется по формуле

$$S_{Se} = \sqrt{\frac{Se \cdot (1 - Se)}{n_1}}.$$

4.3.4.4.2. Специфичность

Специфичностью называют долю отрицательных результатов диагностического теста в популяции. Чем специфичнее тест, тем выше прогностическая ценность его положительного результата.
Вычисление оценки специфичности производится по формуле

$$Sp = \frac{n_{22}}{n_0}.$$

Двусторонний доверительный интервал оцениваемой специфичности вычисляется по формуле

$$I_{Sp} = \left(Sp - \Psi((1+\beta)/2)S_{Sp}; Sp + \Psi((1+\beta)/2)S_{Sp}\right),$$

где S_{Sp} – стандартная ошибка специфичности.
Стандартная ошибка специфичности вычисляется по формуле

$$S_{Sp} = \sqrt{\frac{Sp \cdot (1 - Sp)}{n_0}}.$$

4.3.4.4.3. Распространенность

В литературе встречаются два различных мнения по поводу вычисления распространенности. Согласно источникам, распространенность может быть:

- отношением числа выявленных случаев [заболеваний] ко всем обследованным за определенный промежуток времени (например, за год),
- отношением числа выявленных случаев к численности популяции.

Когда распространенность стремится к нулю, прогностическая ценность положительного результата теста стремится к нулю. Когда распространенность стремится к 1, прогностическая ценность отрицательного результата теста стремится к нулю.
В программе предусмотрено два варианта: ввода либо вычисления распространенности. Ввод известной из предварительных исследований распространенности относится к Байесовской идеологии, когда те или иные выводы по результатам анализа представленных данных делаются с учетом некоторой априорной (известной до опыта) информации. Распространенность имеет область определения от нуля до 1, поэтому для удобства пользователей, с целью совместимости с различными версиями базовой программы и во избежание ошибок пользовательского ввода для известного значения распространенности в предлагаемом поле ввода программы вводить следует только десятичную часть числа. Обратите внимание, что целая часть уже показана на форме. Например, для ввода значения распространенности 0,124 следует ввести число 124. Другой пример. Пусть требуется ввести распространенность 23 случая на 1000 обследованных пациентов. В поле вводится значение 023.
В данном способе вычисление интервальной оценки не производится. О вычислении распространенности (доли) см. главу «Описательная статистика», где в статье, посвященной доле, показано вычисление ее интервальной оценки методом Клоппера–Пирсона.
В следующем способе (см. Флетчер с соавт.) вычисление точечной оценки распространенности на основе тех же самых представленных для анализа выборочных данных производится по формуле
$p = n_1 / n$,
где $n = n_{11} + n_{12} + n_{21} + n_{22}$ – общая численность.
Доверительный интервал оцениваемой распространенности рассчитываются стандартно по формуле Вальда
$I_p = \left(p - \Psi((1+\beta)/2)S_p ; p + \Psi((1+\beta)/2)S_p \right)$,
где S_p – стандартная ошибка распространенности.
Стандартная ошибка распространенности может быть вычислена по формуле

$$S_p = \sqrt{\frac{p \cdot (1-p)}{n}}.$$

Программа выводит график зависимости *PPV* и величины 1 – *NPV* от распространенности. Все величины на графике показаны в процентах.

4.3.4.4.4. Прогностичность положительного результата

Вычисление прогностичности положительного результата производится по формуле

$$PPV = \frac{Se \cdot p}{Se \cdot p + (1-Sp) \cdot (1-p)}.$$

Двусторонний доверительный интервал вычисляется по формуле
$I_{PPV} = (PPV - \Psi((1 + \beta) / 2)S_{PPV}; PPV + \Psi((1 + \beta) / 2)S_{PPV})$,
где S_{PPV} – стандартная ошибка прогностичности положительного результата.
Стандартная ошибка прогностичности положительного результата вычисляется по формуле

$$S_{PPV} = \sqrt{\frac{\left[p \cdot (1 - Sp) \cdot (1 - p)\right]^2 \frac{Se \cdot (1 - Se)}{n_1} + \left[p \cdot Se \cdot (1 - p)\right]^2 \frac{Sp \cdot (1 - Sp)}{n_0}}{\left[Se \cdot p + (1 - Sp) \cdot (1 - p)\right]^4}}.$$

4.3.4.4.5. Прогностичность отрицательного результата

Вычисление прогностичности отрицательного результата производится по формуле

$$NPV = \frac{Sp \cdot (1 - p)}{(1 - Se) \cdot p + Sp \cdot (1 - p)}.$$

Двусторонний доверительный интервал вычисляется по формуле
$I_{NPV} = (NPV - \Psi((1 + \beta) / 2)S_{NPV}; NPV + \Psi((1 + \beta) / 2)S_{NPV})$,
где S_{NPV} – стандартная ошибка прогностичности отрицательного результата.
Стандартная ошибка прогностичности отрицательного результата вычисляется по формуле

$$S_{NPV} = \sqrt{\frac{\left[p \cdot Sp \cdot (1 - p)\right]^2 \frac{Se \cdot (1 - Se)}{n_1} + \left[p \cdot (1 - Se) \cdot (1 - p)\right]^2 \frac{Sp \cdot (1 - Sp)}{n_0}}{\left[(1 - Se) \cdot p + Sp \cdot (1 - p)\right]^4}}.$$

См. монографии Власова, Флетчер с соавт., Флейс, Флейс (Fleiss), Флейс с соавт., Хайнес (Haynes) с соавт., Халли (Hulley) с соавт., статью и отчет Меркалдо (Mercaldo) с соавт., статьи Воробьева, Моссман (Mossman) с соавт., Линн (Linn), Зайкин (Zaykin) с соавт., Кроенке (Kroenke) с соавт., Альтман (Altman) с соавт., Сауро (Sauro) с соавт., Агрести (Agresti) с соавт.

4.3.5. График медиан с ДИ

Представленное программное обеспечение дает возможность табличного и графического вывода медиан сравниваемых выборок, включая доверительные интервалы. При этом на график накладываются доверительные интервалы, вычисленные для доверительного уровня, заданного из стандартной линейки.
Доверительный интервал оцениваемой медианы задается формулой
$I_m = (y_c; y_{n+1-c})$,
где c – параметр, вычисляемый по формуле
$c = [n / 2 - \Psi((1 + \beta) / 2)n^{1/2} / 2]$,
где [.] – целая часть числа,
$\Psi(.)$ – обратная функция стандартного нормального распределения,
β – доверительный уровень, выраженный в долях.
Дополнительно в таблице выводится разность медиан анализируемых выборок. Пусть вычислено $m = n_1 n_2$ разностей значений $w_1 \le w_2 \le ... \le w_m$ всех величин $(x_i - y_j)$, $i = 1,2,...,n_1$; $j = 1,2,...,n_2$, где x_i, $i = 1,2,...,n_1$ и y_j, $j = 1,2,...,n_2$ – значения вариант исходных количественных выборок. Тогда медиана μ полученной выборки w_i, $i = 1,2,...,m$, будет разностью медиан. Для нечетного m медианой является варианта полученного интервального вариационного ряда, имеющая порядковый номер $(m + 1) / 2$. Для четного m медиана равна среднему значению двух средних вариант.
Доверительный интервал оцениваемой разности медиан (интервал Мозеса) задается формулой
$I_\mu = (z_c; z_{m+1-c})$,
где z_i, $i = 1,2,...,m$ – интервальный вариационный ряд, представляющий собой упорядоченный по возрастанию ряд разностей w_i, $i = 1,2,...,m$,
c – параметр, вычисляемый по формуле

$$c = \left[\frac{m}{2} - \Psi((1+\beta)/2) \left(\frac{n_1 n_2 (m+1)}{12} \right)^{1/2} \right],$$

где [.] – целая часть числа,
Ψ(.) – обратная функция стандартного нормального распределения,
β – доверительный уровень, выраженный в долях.
Результаты представленного графического анализа интерпретируются следующим образом. Если 100β% доверительные интервалы оцениваемых медиан сравниваемых выборок пересекаются, конкурирующая гипотеза (медианы не равны) может быть принята на уровне значимости $p \le \beta$. Если 100β% доверительные интервалы оцениваемых средних значений сравниваемых выборок не пересекаются, нулевая гипотеза (медианы равны) не отвергается на уровне значимости $p > \beta$. Т.к. доверительные интервалы тем шире, чем больше значение β, выбирая различные стандартные значения β, можно получить значение уровня значимости, более точно соответствующее представленным данным.

О графическом изображении показателей и интерпретации результатов см. работу Голдстейн (Goldstein) с соавт. Дополнительную информацию о вычислении медианы, псевдомедианы и их доверительных интервалов см. в главе «Описательная статистика».

4.3.6. График долей с ДИ

Представленное программное обеспечение дает возможность вывода долей сравниваемых выборок. При этом на график накладываются доверительные интервалы, вычисленные для доверительного уровня, заданного из стандартной линейки. Границы доверительного интервала доли рассчитываются по «точным» формулам Клоппера–Пирсона (Clopper–Pearson interval). При этом нижняя граница доверительного интервала оцениваемой доли считается как

$$L_P = \left[1 + \frac{n-m+1}{m \cdot F^{-1}_{2m,2(n-m+1)}(1-(1-\beta)/2)} \right]^{-1},$$

где m – число случаев,
n – численность выборки,
$F^{-1}_{\dots}(.)$ – обратная функция F–распределения.
β – доверительный уровень, выраженный в долях.
Верхняя граница доверительного интервала оцениваемой доли считается как

$$H_P = \left[1 + \frac{n-m}{(m+1) \cdot F^{-1}_{2(m+1),2(n-m)}((1-\beta)/2)} \right]^{-1}.$$

Результаты представленного графического анализа интерпретируются следующим образом. Если 100β%, доверительные интервалы оцениваемых долей сравниваемых выборок пересекаются, конкурирующая гипотеза (доли не равны) может быть принята на уровне значимости $p \le \beta$. Если 100β% доверительные интервалы оцениваемых средних значений сравниваемых выборок не пересекаются, нулевая гипотеза (доли равны) не отвергается на уровне значимости $p > \beta$. Т.к. доверительные интервалы тем шире, чем больше значение β, выбирая различные стандартные значения β, можно получить значение уровня значимости, более точно соответствующее представленным данным.

О графическом изображении показателей и интерпретации результатов см. работу Голдстайн (Goldstein) с соавт. Дополнительную информацию о вычислении долей и доверительных

интервалов см. в главе «Описательная статистика».

4.3.7. ROC анализ

ROC (Receiver Operating Characteristic) анализ может иметь различные применения для анализа данных. Дальнейшие обозначения проще всего пояснить с помощью таблицы 2 x 2.

Исследуемый метод	Стандартный метод	
	Положительный исход	Отрицательный исход
Положительный исход	T_P	F_P
Отрицательный исход	F_N	T_N

Суть обозначений ясна из первых букв английских терминов:

- True – истинно,
- False – ложно,
- Positive – положительный,
- Negative – отрицательный.

Термины «положительный» и «отрицательный» здесь относятся не к объекту исследования, а, скажем, к способности диагностического теста установить диагноз. Так, при исследовании заболевания положительным исходом будет являться наличие заболевания, отрицательным исходом – отсутствие заболевания.

Термин ROC curve (ROC кривая) в адекватном переводе, заимствованном из радиотехники, означает кривую соотношений правильного и ложного обнаружения сигналов. ROC кривая представляет собой график параметрического типа. При этом абсцисса и ордината кривой являются функциями некоторого параметра, произвольно изменяемого или конкретно измеряемого в эксперименте. В исследовательской практике могут иметь место различные сочетания данных функций, что приводит к различным ROC кривым. Программа строит и анализирует наиболее употребительный тип ROC кривой, параметрически отображающий величину чувствительности *Se* и величину неспецифичности 1 – *Sp*, где *Sp* – специфичность. Порог чувствительности на графике не отображается, однако каждому [в данном случае] заданному значению порога соответствует пара «чувствительность–неспецифичность». На графике данные величины принято изображать в процентах. Показатели определяются следующими формулами.

Чувствительность показывает долю истинно положительных случаев, т. е.

$$Se = \frac{T_P}{T_P + F_N}.$$

Специфичность показывает долю истинно отрицательных случаев, т. е.

$$Sp = \frac{T_N}{T_N + F_P}.$$

Некоторые авторы величину *Sp* называют частотой истинно отрицательных результатов (true negative rate), а величину 1 – *Sp* называют ценой метода либо частотой ложно положительных результатов (false positive rate, FPR). По аналогии величину *Se* иногда называют частотой истинно положительных результатов (true positive rate, TPR). Некоторые авторы полагают, что в таких терминах ROC кривая более понятна для чтения. Также условились для построения ROC кривой использовать показатели в процентах.

Сочетание значений чувствительности и специфичности, рассчитываемое программой, в дальнейшем анализе может быть выбрано различным в зависимости от требований исследователя. При этом соответствующее значение диагностического параметра, выводимое программой, называют порогом отсечения. В программе используется критерий Юдена

(Йоден, Youden), максимизирующий сумму чувствительности и специфичности. О порогах отсечения дополнительно см. главу «Распознавание образов с обучением».
Рассмотрим алгоритм построения ROC кривой. Пусть даны исследуемая выборка численностью n и стандартная выборка численностью m.
Алгоритм ROC анализа предлагается сформулировать следующим образом:

1. Задаться интервалом изменения параметра. Удобнее всего данный интервал получить, объединив представленные выборки в массив диагностических параметров численностью $n + m$, а затем отсортировав данный массив по убыванию.
2. Используя варианты полученного в предыдущем пункте алгоритма массива диагностических параметров в качестве порогов отсечения, составить на основе исходных выборок для каждой варианты данного массива таблицу 2 x 2. При этом решающее правило имеет вид «параметр ≥ порога».
3. Подсчитать для каждой составленной в предыдущем пункте алгоритма таблицы чувствительность и неспецифичность. Массив чувствительностей численностью $n + m$ будет массивом абсцисс ROC кривой. Массив неспецифичностей численностью $n + m$ будет массивом ординат ROC кривой.
4. Построить график ROC кривой по парам точек «абсцисса–ордината», полученным в предыдущем пункте алгоритма.
5. Подсчитать площадь, отсекаемую ROC кривой.

Позиции 2, 3, 4 и 5 представленного алгоритма выгоднее выполнять в цикле по всем $n + m$ вариантам массива диагностических параметров.
Объективную оценку качества диагностического метода может показать площадь под ROC кривой, в литературе кратко называемая AUC (Area Under Curve). Оценка данной площади подсчитывается по формуле трапеций:

$$\hat{A} = \frac{1}{2}\sum_{j=1}^{n+m-1}(Se_i + Se_{i+1})(Sp_i - Sp_{i+1}).$$

При расчете оценки площади условились использовать показатели в долях. Чем выше AUC, тем большую прогностическую ценность имеют представленные данные (представленный метод). Максимальное значение AUC равно 1. При значении AUC, равном 0,5, прогностическая ценность отсутствует. Возможна такая конфигурация исходных данных, что кривая ROC окажется ниже диагонали, а AUC окажется, соответственно, в интервале от 0 до 0,5. В этом случае следует изменить решающее правило (позиция 2 алгоритма) на противоположное: «параметр ≤ порога» – и выполнить алгоритм заново.
Стандартная ошибка оценки AUC подсчитывается по формуле, представленной Хэнли (Hanley) с соавт. (1982),

$$SE(\hat{A}) = \sqrt{\frac{\hat{A}(1-\hat{A}) + (n-1)(Q_1 - \hat{A}^2) + (m-1)(Q_2 - \hat{A}^2)}{n \cdot m}},$$

где для краткости записи обозначено:

$$Q_1 = \hat{A}/(2 - \hat{A}),$$

$$Q_2 = 2\hat{A}^2/(1 + \hat{A}).$$

Хэнли с соавт. предложили метод сравнения двух ROC кривых по отсекаемым ими AUC. Для этого в простейшем случае используется статистика

$$Z = \frac{\left|\hat{A}_1 - \hat{A}_2\right|}{\sqrt{SE(\hat{A}_1)^2 + SE(\hat{A}_2)^2}},$$

распределенная асимптотически нормально.
Вычисление статистики Z на основе оценок AUC и ее дисперсий для двух ROC кривых не

представляет сложности и может быть выполнено пользователем самостоятельно. В программе же реализовано вычисление статистики *Z* при сравнении оценки AUC для данной ROC кривой с величиной AUC, равной 0,5 (случай «бесполезной» классификации). Статистика *Z*, вычисленная таким образом, позволяет объективно судить о статистической значимости полученной классификации. При этом *SE*(0,5) вычисляется по показанной выше формуле.

Для вычисления двустороннего доверительного интервала оцениваемой AUC применяется формула:

$$I_{AUC} = \left(\hat{A} - \Psi\left((1+\beta)/2\right) \cdot SE(\hat{A}); \hat{A} + \Psi\left((1+\beta)/2\right) \cdot SE(\hat{A})\right),$$

где Ψ(.) – обратная функция стандартного нормального распределения,

β – доверительный уровень, выраженный в долях.

Программное обеспечение предлагает дополнительную гибкость ввода исходных данных для ROC анализа. Доступны возможности:

- Ввод данных типа «выборка – выборка». При этом в качестве первой выборки указывается выборка с одним из значений классификатора (например, общепринятое значение 1, или наличие симптома болезни). В качестве второй выборки указывается выборка с другим значением классификатора (например, общепринятое значение 0, или отсутствие симптома болезни). При этом сами значения классификатора не вводятся.
- Ввод данных типа «выборка – классификатор». При этом в качестве первой выборки вводится весь массив исходных данных (для всех значений классификатора). В качестве второй выборки вводится [соответствующий массиву исходных данных] массив классификатора, состоящий из единиц и нулей. Кодировка классификатора аналогична предыдущему случаю. Если в массиве классификатора встречаются значения, отличные от общепринятых стандартных значений 1 или 0, программой выдается диагностическое сообщение и вычисление не производится.

Рассматривая возможность ROC анализа для исходных данных, представленных в виде таблицы 2 x 2, необходимо сделать вывод, что такой анализ сделать нельзя. Дело в том, что ROC кривая, как уже упоминалось выше – это не просто изображенная на графике зависимость *Se* от 1 – *Sp*. ROC кривая представляет собой особый математический объект, называемый параметрической кривой. Параметрическая кривая возникает, когда две величины, участвующие в построении графика, на самом деле зависят не одна от другой, а от третьего параметра (на графике не изображаемого) – в данном случае от порога отсечения. Аргументом является именно порог отсечения, а изображаются на графике произведенные от него чувствительность и неспецифичность. Вот порога–то отсечения как раз и нет в представленной таблице. Формально, конечно, можно для таблицы посчитать чувствительность и неспецифичность (в %), добавить еще две точки – (0;0) и (100;100) и нарисовать некий график, даже посчитать площадь под таким объектом, но это будет не ROC анализ. При необходимости данные формальные построения пользователь выполнит самостоятельно.

В некоторых публикациях бытует ошибочное изображение ROC в виде кривой гладкой. Это демонстрирует непонимание авторами публикаций самой сути ROC анализа как графического отображения результатов бинарной классификации. ROC кривая – не график зависимости одной непрерывной величины от другой непрерывной величины. ROC кривая может изображаться только в виде лесенки (в этом смысле название «кривая» – curve не является корректным). Она дискретна по своей природе, меняя значения абсциссы и ординаты скачками даже при непрерывном изменении порога отсечения, не может быть гладкой, поэтому ее нельзя аппроксимировать гладкой кривой.

Популярное введение в ROC см. в статье Сюэтс (Swets) с соавт. В дополнение к упомянутым источникам по ROC анализу см. монографии Флетчер с соавт., Жоу (Zhou) с соавт., Хайнес (Haynes) с соавт., статьи Метц (Metz), Обучовски (Obuchowski), Дэвис (Davis) с соавт., Фараджи (Faraggi) с соавт., Парк (Park) с соавт., Шистерман (Schisterman) с соавт., Цвайг (Zweig) с соавт., Альтман (Altman) с соавт., Ланглотц (Langlotz), Клотше (Klotsche) с соавт., статьи и отчет Фосетт (Fawcett). Тема упомянута в книгах Петри с соавт., ван Бель (van Belle) с соавт. О порогах отсечения см. также статью Флусс (Fluss) с соавт. О статистическом сравнении ROC кривых см. статьи Вергара (Vergara) с соавт., Хэнли с соавт. (1983), Метц с соавт., ДеЛонг (DeLong) с соавт. На связь AUC и статистики непараметрического критерия Вилкоксона указано в работе Фосетт (Fawcett), монографии Власова. Обзор компьютерных программ представили Стефан (Stephan) с соавт.

4.3.8. Каппа Коэна

Для оценки согласия двух классификаций применяется показатель – каппа Коэна (Cohen's Kappa). Интерпретация каппы поясняется в следующей таблице.

Значение каппы	Уровень согласия
< 0,00	Плохое согласие (poor)
0,00 – 0,20	Небольшое согласие (slight)
0,21 – 0,40	Удовлетворительное согласие (fair)
0,41 – 0,60	Среднее согласие (moderate)
0,61 – 0,80	Существенное согласие (substantial)
0,81 – 1,00	Почти прекрасное согласие (almost perfect)

Вычисление выборочной оценки каппы производится по формуле

$$\hat{\kappa} = \frac{p_0 - p_e}{1 - p_e},$$

где p_0 – доля случаев, относительно которых существует согласие,
p_e – доля случаев, относительно которых ожидается согласие.
Упомянутые доли вычисляются по формулам, соответственно,

$$p_0 = \frac{n_{11}}{n} \cdot \frac{n_{22}}{n}$$ и

$$p_e = \frac{r_1}{n} \cdot \frac{c_1}{n} + \frac{r_2}{n} \cdot \frac{c_2}{n},$$

где $r_1 = n_{11} + n_{12}$ – численность первой строки таблицы,
$c_1 = n_{11} + n_{21}$ – численность первого столбца таблицы,
$r_2 = n_{21} + n_{22}$ – численность второй строки таблицы,
$c_2 = n_{12} + n_{22}$ – численность второго столбца таблицы,
$n = n_{11} + n_{21} + n_{12} + n_{22}$ – численность таблицы,
n_{11}, n_{21}, n_{12}, n_{22} – ячейки таблицы.
Стандартная ошибка каппы вычисляется по формуле

$$SE(\hat{\kappa}) = \sqrt{\frac{p_0(1-p_0)}{n(1-p_e)^2}}.$$

Двусторонний доверительный интервал оцениваемой каппы вычисляется по формуле
$I_\kappa = \left(\hat{\kappa} - \Psi\left((1+\beta)/2\right) \cdot SE(\hat{\kappa}); \hat{\kappa} + \Psi\left((1+\beta)/2\right) \cdot SE(\hat{\kappa})\right),$
где $\Psi(.)$ – обратная функция стандартного нормального распределения,

β – доверительный уровень, выраженный в долях.

См. монографию Флейс (Fleiss) с соавт., статьи Коэн (Cohen), Крюсон (Crewson), Флейс, Брайнгтон (Bryington) с совт., Костина, Заславского с соавт., Виера (Viera) с соавт., Ли (Lee) с соавт., Кундел (Kundel) с соавт. Взвешенную каппу также рассмотрели Коэн, Чиччетти (Cicchetti), Флейс с соавт. Расчет доверительных интервалов оцениваемой каппы см. также в статьях Блэкман (Blackman) с соавт., Гарнер (Garner), Гарнер с соавт.

Список использованной и рекомендуемой литературы

1. Agresti A. Categorical data analysis. – Hoboken, NJ: John Wiley & Sons, 2002.
2. Agresti A., Coull B. Approximate is better than «exact» for interval estimation of binomial proportions // The American Statistician, 1998, vol. 52, pp. 119–126.
3. Agresti A., Min Y. Simple improved confidence intervals for comparing matched proportions // Statistics in Medicine, 2005, vol. 24, pp. 729–740.
4. Ahmad I.A. Modification of some goodness of fit statistics II: two–sample and symmetry testing // Sankhya: The Indian Journal of Statistics, 1996, vol. 58, ser. A, pt. 3, pp. 464–472.
5. Altman D.G., Bland J.M. Statistics notes: Diagnostic tests 1: sensitivity and specificity // British Medical Journal, 11 June, 1994, vol. 308 p. 1552.
6. Altman D.G., Bland J.M. Statistics notes: Diagnostic tests 3: receiver operating characteristic plots // British Medical Journal, 16 July 1994, vol. 309, p. 188.
7. Anderson T.W. On the distribution of the two–sample Cramer–von Mises criterion // Annals of Mathematical Statistics, 1962, vol. 33, no. 3, pp. 1148–1159.
8. Ansari A.R., Bradley R.A. Rank–sum tests for dispersions // Annals of Mathematical Statistics, 1960, vol. 31, no. 4, pp. 1174–1189.
9. Applegate K.E., Tello R., Ying J. Hypothesis testing III: Counts and medians // Radiology, 2003, vol. 228, no. 3, pp. 603–608.
10. Balakrishnan N. Handbook of statistics. Vol. 16. Order statistics – Theory and methods / Ed. by N. Balakrishnan, C.R. Rao. – New York, NY: Elsevier, 1997.
11. Balakrishnan N. Handbook of statistics. Vol. 17. Order Statistics: Applications / Ed. by N. Balakrishnan, C.R. Rao. – New York, NY: Elsevier, 1998.
12. Barratt A. Tips for learners of evidence–based medicine: 1. Relative risk reduction, absolute risk reduction and number needed to treat / A. Barratt, P.C. Wyer, R. Hatala et al. // Canada's learning medical journal, 17 August 2004, vol. 171, no. 4, pp. 353–358.
13. Bennett B.M., Underwood R.E. On McNemar's test for the 2 x 2 table and its power // Biometrics, June 1970, vol. 26, no. 2, pp. 339–343.
14. Berger R.L., Sidik K. Exact unconditional tests for a 2 x 2 matched–pairs design // Statistical Methods in Medical Research, 2003, vol. 12, pp. 91–108.
15. Bergmann R., Ludbrook J., Spooren W.P.J.M. Different outcomes of the Wilcoxon–Mann–Whitney from different statistics packages // The American Statistician, February 2000, vol. 54, no. 1, pp. 72–77.
16. Bertell H.R. Extensions of the relative risk concept // Cellular and Molecular Life Sciences, January 1975, vol. 31, no. 1, pp. 1–10.
17. Best D.J. Nonparametric comparison of two histograms // Biometrics, June 1994, vol. 50, no. 2, pp. 538–541.
18. Biggerstaff B.J. Confidence intervals for the difference of two proportions estimated from pooled samples // Journal of Agricultural, Biological, and Environmental Statistics, December 2008, vol. 13, no. 4, pp. 478–496.
19. Birnbaum Z.W. On a use of the Mann–Whitney statistic // Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, December 1954 and July–August

1955, vol. 1: Contributions to the Theory of Statistics / Ed. by J. Neyman. – Berkeley, CA: University of California Press, 1956, pp. 13–17.
20. Bishop Y.M.M., Fienberg S.E., Holland P.W. Discrete multivariate analysis: theory and practice. – Cambridge, MA: MIT Press, 1975.
21. Blackman N.J., Koval J.J. Interval estimation for Cohen's kappa as a measure of agreement // Statistics in medicine, 2000, vol. 19, no. 5, pp. 723–741.
22. Blair R.C., Higgins J.J. Comparison of the power of the paired samples t test to that of Wilcoxon's signed–ranks test under various population shapes // Psychological Bulletin, January 1985, vol. 97, no. 1, pp. 119–128.
23. Bland J.M., Altman D.G. The odds ratio // British Medical Journal, 27 May 2000, vol. 320, p. 1468.
24. Bland M., Peacock J. Interpreting statistics with confidence // The Obstetrician & Gynaecologist, 2002, vol. 4, no. 3, p. 176–180.
25. Box G.E.P., Hunter W.G., Hunter J.S. Statistics for experimenters: An introduction to design, data analysis, and model building. – New York, NY: John Wiley & Sons, 1978.
26. Bridge P.D., Sawilowsky S.S. Increasing physicians' awareness of the impact of statistics on research outcomes: comparative power of the t–test and Wilcoxon Rank–Sum test in small samples applied research // Journal of Clinical Epidemiology, 1999, vol. 52, no.3, pp. 229–235.
27. Brown L., Li X. Confidence intervals for two sample binomial distribution // Journal of Statistical Planning and Inference, 1 March 2005, vol. 130, issues 1–2, pp. 359–375.
28. Bryington A.A., Palmer D.J., Watkins M.W. The estimation of interobserver agreement in behavioral assessment // Journal of Early and Intensive behavior Intervention, 2004, vol. 1, no. 1, pp. 115–119.
29. Buchan I.E. The development of a statistical computer software resource for medical research. Thesis for the degree of Doctor of Medicine. – Liverpool: University of Liverpool, 2000.
30. Buhrman J.M. Tests and confidence intervals for the difference and ratio of two probabilities // Biometrika, 1977, vol. 64, no. 1, pp. 160–162.
31. Callaert H. Nonparametric hypotheses for the two–sample location problem // Journal of Statistics Education, 1999, vol. 7, no. 2.
32. Chernick M.R. Friis R.H. Introductory biostatistics for the health sciences. Modern application including bootstrap. – New York, NY: John Wiley & Sons, 2003.
33. Chernoff H., Savage I.R. Asymptotic normality and efficiency of certain nonparametric test statistics // Annals of Mathematical Statistics, 1958, vol. 29, no. 4, pp. 972–994.
34. Cicchetti D.V. A new measure of agreement between rank ordered variables // Proceedings of the American Psychological Association, 1972, vol. 7, pp. 17–18.
35. Cicchetti D.V. Comparison of the null distributions of weighted kappa and the C ordinal statistic // Applied Psychological Measurement, 1977, vol. 1, pp. 195–201.
36. Cirrone G.A.P. A goodness–of–fit statistical toolkit / G.A.P. Cirrone, S. Donadio, S. Guatelli et al. // IEEE Transactions on Nuclear Science, October 2004, vol. 51, no. 5, pp. 2056 – 2063.
37. Cohen J. A coefficient of agreement for nominal scales // Educational and Psychological Measurement, 1960, vol. 20, pp. 37–46.
38. Cohen J. Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit // Psychological Bulletin, 1968, vol. 70, pp. 213–220.
39. Conover W.J. Practical nonparametric statistics. – New York, NY: John Wiley & Sons, 1999.
40. Conover W.J., Iman R.L. Rank transformations as a bridge between parametric and nonparametric statistics // The American Statistician, 1981, vol. 35, pp. 124–129.

41. Conover W.J., Johnson M.E., Johnson M.M. A comparative study of tests for homogeneity of variance, with applications to the outer continental shelf bidding data // Technometrics, November 1981, vol. 23, no. 4, pp. 351–361.
42. Crewson P.E. Fundamentals of clinical research for radiologists. Reader agreement studies // American Journal of Roentgenology, 2005, vol. 184, pp. 1391–1397.
43. Dahiya R.C., Gurland J. Pearson chi–square test of fit with random intervals // Biometrika, 1972, vol. 59, no. 1, pp. 147-153.
44. Dahiya R.C., Gurland J. How many classes in the Pearson chi–square test? // Journal of the American Statistical Association, September 1973, vol. 68, no. 343, pp. 707–712.
45. Daniel W.W. Applied nonparametric statistics. – Florence, KY: Wadsworth Publishing, 1990.
46. Davis J., Goadrich M. The relationship between precision–recall and ROC curves // Proceedings of 23 International Conference on Machine Learning, Pittsburgh, PA, 2006.
47. DeLong E.R., DeLong D.M., Clarke–Pearson D.L. Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach // Biometrics, 1988, vol. 44, pp. 837–845.
48. Deshpande J.V., Gore A.P., Shanubhogue A. Statistical analysis of nonnormal data. – New York, NY: John Wiley & Sons, 1995.
49. Di Bucchianico A. Combinatorics, computer algebra and Wilcoxon–Mann–Whitney test // Memorandum COSOR 96–24, 1996, Eindhoven University of Technology.
50. Ederer F., Mantel N. Confidence limits for the ratio of two Poisson variables // American Journal of Epidemiology, September 1974, vol. 100, no. 3, pp. 165–167.
51. Everitt B.S. The analysis of contingency tables. – Boca Raton, FL: Chapman & Hall / CRC, 1977.
52. Faraggi D., Reiser B., Schisterman E. ROC curve analysis for biomarkers based on pooled assessments // Statistics in Medicine, 2003, vol. 22, pp. 2515–2527.
53. Fawcett T. An introduction to ROC analysis // Pattern Recognition Letters, 2006, vol. 27, no. 8, pp. 861–874.
54. Fawcett T. ROC graphs with instance varying costs // Pattern Recognition Letters, June 2006, vol. 27, no. 8, pp. 882–891.
55. Fawcett T. ROC graphs: Notes and practical considerations for researchers // Technical Report HPL–2003–4, HP Laboratories, 2003.
56. Fisher R.A. On the interpretation of χ^2 from contingency tables, and the calculation of P // Journal of the Royal Statistical Society, 1922, vol. 85, pp. 87–94.
57. Fisher R.A. Statistical tests of agreement between observation and hypothesis // Economica, 1923, vol. 3, pp. 139–147.
58. Fisz M. On a result by M. Rosenblatt concerning the Von Mises–Smirnov test // The Annals of Mathematical Statistics, 1960, vol. 31, no. 2, pp. 427–429.
59. Fleiss J.L, Cohen J. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability // Educational and Psychological Measurement, 1973, vol. 33, pp. 613–619.
60. Fleiss J.L. Measuring nominal scale agreement among many raters // Psychological Bulletin, 1971, vol. 76, pp. 378–81.
61. Fleiss J.L., Cohen J., Everitt, B.S. Large sample standard errors of kappa and weighted kappa // Psychological Bulletin, 1969, vol. 72, pp. 323–327.
62. Fleiss J.L., Levin B., Paik M.C. Statistical methods for rates and proportions. – New York, NY: John Wiley & Sons, 2003.
63. Fluss R., Faraggi D., Reiser B. Estimation of the Youden Index and its associated cutoff point // Biometrical Journal, 2005, vol. 47, pp. 458–472.
64. Gardner M.J., Altman D.G. Confidence intervals rather than P values: estimation rather than

hypothesis testing // British Medical Journal, 15 March 1986, vol. 292, pp. 746–750.
65. Gardner M.J., Altman D.G. Confidence intervals rather than P values: estimation rather than hypothesis testing // British Medical Journal, 1986, vol. 292, pp. 746–750.
66. Garner B., Hale C.A., Fleiss J.L. Interval estimation for kappa // Biometrics, 1994, vol. 50, no. 1, pp. 309–310.
67. Garner J.B. The standard error of Cohen's Kappa // Statistics in medicine, 1991, vol. 10, no. 5, pp. 767–75.
68. Garrett L., Nash J.C. Issues in teaching the comparison of variability to non–statistics students // Journal of Statistics Education, 2001, vol. 9, no. 2.
69. Gart J.J. Approximate confidence limits for the relative risk // Journal of the Royal Statistical Society, series B (Methodological), 1962, vol. 24, no. 2, pp. 454–463.
70. Geisser S. Significance testing for the 2 x 2 table // Bulletin of the International Statistical Institute, 52nd Session, Proceedings, Tome LVIII, Finland, 1999. Contributed Paper Meeting 7: Statistical tests.
71. Gibbons J.D., Chakraborti S. Nonparametric statistical inference. – New York, NY: Marcel Dekker, 1992.
72. Goldstein H., Healy M.J.R. The graphical presentation of a collection of means // Journal of the Royal Statistical Society: Series A (Statistics in Society), 1995, vol. 158, no. 1, pp. 175–177.
73. Goodman L.A. Kolmogorov–Smirnov tests for psychological research // Psychological Bulletin, 1954, vol. 51, pp. 160–168.
74. Graf R.G. A Visual Basic program for estimating missing cell frequencies in chi square tests for association / R.G. Graf, E.F. Alf, S. Williams et al. // InterStat (Statistics on the Internet), August 1997.
75. Graham P.L., MacEachern S.N., Wolfe D.A. The unconditional and conditional censored Wilcoxon rank sum null distributions: Tabulated values and P–value program // InterStat (Statistics on the Internet), August 2003, No. 1.
76. Greenwood P.E., Nikulin M.S. Guide to chi–squared testing. – New York, NY: John Wiley & Sons, 1996.
77. Guidance for data quality assessment. Practical methods for data analysis. EPA QA/G–9. – Washington, DC: United States Environmental Protection Agency, 2000.
78. Guyatt G. Basic statistics for clinicians: 1. Hypothesis testing / G. Guyatt, R. Jaeschke, N. Heddle et al. // Canadian Medical Association Journal, January 1995, vol. 152, no. 1, pp. 27–32.
79. Hajek J., Sidak Z., Sen P.K. Theory of rank tests. – New York, NY: Academic Press, 1999.
80. Hajian–Tilaki K.O. A comparison of parametric and nonparametric approaches to ROC analysis of quantitative diagnostic tests / K.O. Hajian–Tilaki, J.A. Hanley, L. Joseph et al. // Medical Decision Making, 1997, vol. 17, no. 1, pp. 94–102.
81. Hanley J.A., McNeil B.J. A method of comparing the areas under receiver operating characteristic curves derived from the same cases // Radiology, September 1983, vol. 148, no. 3, pp. 839–843.
82. Hanley J.A., McNeil B.J. The meaning and use of the area under a receiver operating characteristic (ROC) curve // Radiology, April 1982, vol. 143, no. 1, pp. 29–36.
83. Hauck W.W., Anderson S. A comparison of large–sample confidence interval methods for the difference of two binomial probabilities // The American Statistician, November 1986, vol. 40, no. 4, pp. 318–322.
84. Haynes R.B. Clinical epidemiology: how to do clinical practice research / R.B. Haynes, D.L. Sackett, G.H. Guyatt et al. – Philadelphia, PA: Lippincott Williams & Wilkins, 2006.
85. Helsel D.R., Hirsch R.M. Techniques of Water–Resources Investigations Reports. Book 4:

Hydrologic Analysis and Interpretation. Section A: Statistical analysis. Chapter A3: Statistical methods in water resources. – Denver, CO: U.S. Geological Survey, 2002.
86. Heschl W.C. An investigation of the power of the Wald–Wolfowitz, two sample, runs test. Master's thesis. – Monterey, CA: Naval Postgraduate School, 1972.
87. Hettmansperger T.P. Statistical inference based on ranks. – New York, NY: John Wiley & Sons, 1984.
88. Hodges J.L., Lehmann E.L. Comparison of the normal scores and Wilcoxon tests // Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability. vol. 1: Contributions to the Theory of Statistics, June 20–July 30, 1960 / Ed. by J. Neyman. – Berkeley, CA: University of California Press, 1961, pp. 307–317.
89. Hoeffding W. «Optimum» nonparametric tests // Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability, July 31–August 12, 1950 / Ed. by J. Neyman. – Berkeley, CA: University of California Press, 1951, pp. 83–92.
90. Hollander M., Wolfe D.A. Nonparametric statistical methods. – New York, NY: John Wiley & Sons, 1999.
91. Hoover D.R. Extending power and sample size approaches developed for McNemar's procedure to general sign tests // International Statistical Review, 2005, vol. 73, no. 1, pp. 103–110.
92. Hora S.C., Conover W.J. The F statistic in the two–way layout with rank–score transformed data // Journal of the American Statistical Association, 1984, vol. 79, pp. 668–673.
93. Hora S.C., Iman R.L. Asymptotic relative efficiencies of the rank–transformation procedure in randomized complete block designs // Journal of the American Statistical Association, 1988, vol. 83, pp. 462–470.
94. Hsieh C.C. Note on interval estimation of the difference between proportions from correlated series // Statistics in Medicine, January–March 1985, vol. 4, no. 1, pp. 23–27.
95. Hulley S.B. Designing clinical research: An epidemiologic approach / S.B. Hulley, S.R. Cummings, W.S. Browner et al. – Philadelphia, PA: Lippincott Williams & Wilkins: 2000.
96. Iman R.L., Conover W.J. The use of the rank transform in regression // Technometrics, 1979, vol. 21, pp. 499–509.
97. Iman R.L., Hora S.C., Conover W.J. Comparison of asymptotically distribution–free procedures for the analysis of complete blocks // Journal of the American Statistical Association, 1984, vol. 79, pp. 674–685.
98. Klotsche J. A novel nonparametric approach for estimating cut–offs in continuous risk indicators with application to diabetes epidemiology / J. Klotsche, D. Ferger, L. Pieper et al. // BMC Medical Research Methodology 2009, vol. 9, no. 63.
99. Klotz J.H. Nonparametric tests for scale // Annals of Mathematical Statistics, 1962, vol. 33, no. 2, pp. 498–512.
100. Kraft S. Nonparametric tests based on area–statistics // Bulletin of the International Statistical Institute, 52nd Session, Proceedings, Tome LVIII, Finland, 1999. Contributed Paper Meeting 82: Nonparametric statistics.
101. Kraft S., Schmid F. Nonparametric tests based on area–statistics // Discussion papers in statistics and econometrics, August 2000, no.2/00. Seminar of economic and social statistics, University of Cologne.
102. Kroenke K. Causes of persistent dizziness: a prospective study of 100 patients in ambulatory care / K. Kroenke, C.A. Lucas, M.L. Rosenberg // Annals of Internal Medicine, 1 December 1992, vol. 117, no. 11, pp. 898–904.
103. Kuiper N.H. Tests concerning random points on a circle // Proceedings of the Koninklijke Nederlandse Akademie van Wetenschappen, ser. A, 1962, vol. 63, pp. 38–47.
104. Kundel H.L., Polansky M. Measurement of observer agreement // Radiology, 2003,

vol. 228, no. 2, pp. 303–308.
105. Langlotz C.P. Fundamental measures of diagnostic examination performance: Usefulness for clinical decision making and research // Radiology, 2003, vol. 228, no. 1, pp. 3–9.
106. LaVange L.M., Koch G.G. Rank score tests // Circulation, 2006, vol. 114, pp. 2528–2533.
107. Le C.T. Introductory biostatistics. – New York, NY: John Wiley & Sons, 2003.
108. Lee P.S.C., Suen H.K. The estimation of kappa from percentage agreement interobserver reliability // Behavioral Assessment, 1984, vol. 6, pp. 375–378.
109. Lehman S.Y. Exact and approximate distribution for the Wilcoxon statistic with ties // Journal of the American Statistical Association, June 1961, vol. 56, no. 294, pp. 293–298.
110. Lehmann E.L. Consistency and unbiasedness of certain nonparametric tests // The Annals of Mathematical Statistics, 1951, vol. 22, no. 2, pp. 165–179.
111. Lehmann E.L. Nonparametrics: statistical methods based on ranks. – New York, NY: Prentice Hall, 1998.
112. Lehmann E.L. Testing statistical hypotheses. – New York, NY: John Wiley & Sons, 1986.
113. Lemeshko B., Lemeshko S. Statistical distribution convergence and homogeneity test power for Smirnov and Lehmann–Rosenblatt tests // Measurement Techniques, December 2005, vol. 48, no. 12, pp. 1159–1166.
114. Li G., Zhou K. A unified approach to nonparametric comparison of receiver operating characteristic curves for longitudinal and clustered data // Journal of the American Statistical Association, June 2008, vol. 103, no. 482, pp. 705–713.
115. Linn S. A new conceptual approach to teaching the interpretation of clinical tests // Journal of Statistics Education, 2004, vol. 12, no. 3.
116. Littell R.C. On the efficiency of a competitor of the two–sample Kolmogorov–Smirnov and Kuiper tests // The Annals of Mathematical Statistics, vol. 43, no. 6, pp. 1991–1992.
117. Loosen F. Note on the chi–square statistic of association in 2 x 2 contingency tables and the correction for continuity // Mathematiques et Sciences Humaines, 1978, vol. 61, pp. 29–37.
118. Macskassy S.A., Provost F. Confidence bands for ROC curves: Methods and an empirical study // European Conference on Artificial Intelligence. First Workshop on ROC Analysis in Artificial Intelligence, Valencia, Spain, 22 August, 2004.
119. Mann H.B., Whitney D.R. On a test of whether one of two random variables is stochastically larger than the other // The Annals of Mathematical Statistics, March 1947, vol. 18, no. 1, pp. 50–60.
120. Maxwell A.E. Comparing the classification of subjects by two independent judges // British Journal of Psychiatry, 1970, vol. 116, pp. 651–655.
121. McNemar Q. Psychological statistics. – New York, NY: John Wiley & Sons, 1966.
122. Mercaldo N.D., Lau K.F., Zhou X.–H. Confidence intervals for predictive values with an emphasis to case–control studies // Statistics in Medicine, May 2007, vol. 26, no. 10, pp. 2170–2183.
123. Mercaldo N.D., Zhou X.–H., Lau K.F. Confidence intervals for predictive values using data from a case control study // UW Biostatistics Working Paper Series, Working Paper 271, 7 December 2005.
124. Metz C.E. Basic principles of ROC analysis // Seminars in nuclear medicine, October 1978, vol. 8, no. 4, pp. 283–298.
125. Metz C.E., Herman B.A., Roe C.A. Statistical comparison of two ROC–curve

estimates obtained from partially–paired datasets // Medical Decision Making, January–March 1998, vol. 18, no. 1, pp. 110–121.
126. Montgomery D.C., Runger G.C. Applied statistics and probability for engineers. – New York, NY: John Wiley & Sons, 2003.
127. Mossman D., Berger J.O. Intervals for posttest probabilities: A comparison of 5 methods // Medical Decision Making, November–December 2001, vol. 21, no. 6, pp. 498–507.
128. Motulsky H.J. InStat guide to choosing and interpreting statistical tests. – San Diego, CA: GraphPad Software, 1998.
129. Motulsky H.J. Intuitive biostatistics. – New York, NY: Oxford University Press, 1995.
130. Myers J., Huang S.–F., Tsay J. Exact conditional inference for two–way randomized Bernoulli experiments // Journal of Statistical Software, September 2007, vol. 21, code snippet 1.
131. Newcombe R.G. Improved confidence intervals for the difference between binomial proportions based on paired data // Statistics in Medicine, 1998, vol. 17, pp. 2635–2650.
132. Newman S.C. Biostatistical methods in epidemiology. – New York, NY: John Wiley & Sons, 2001.
133. NIST/SEMATECH e–Handbook of statistical methods (NIST Handbook 151, ver. 1/27/2005). – Gaithersburg, MD: National Institute of Standards and Technology, 2005.
134. Obuchowski N.A. Fundamentals of clinical research for radiologists. ROC analysis // American Journal of Roentgenology, February 2005, vol. 184, no. 2, pp. 364–372.
135. Obuchowski N.A. Receiver operating characteristic curves and their use in radiology // Radiology, October 2003, vol. 229, no. 1, pp. 3–8.
136. Park S.H., Goo J.M., Jo C.–H. Receiver operating characteristic (ROC) curve: Practical review for radiologists // Korean Journal of Radiology, March 2004, vol. 5, no. 1, pp. 11–18.
137. Pinto J.V., Ng P., Allen D.S. Logical extremes, beta, and the power of the test // Journal of Statistics Education, 2003, vol. 11, no. 1.
138. Puri M.L., Rajaram N.S. Asymptotic normality and convergence rates of linear rank statistics under alternatives // Mathematical Statistics Banach Center Publications (PWN–Polish Scientific Publishers, Warsaw), 1980, vol. 6, pp. 267–277.
139. Randles R.H., Wolfe D.A. Introduction to the theory of nonparametric statistics. – New York, NY: John Wiley & Sons, 1979.
140. Rao C.R. Handbook of statistics. Vol. 27. Epidemiology and medical statistics / Ed. by C.R. Rao, J.P. Miller, D.C. Rao. – New York, NY: Elsevier, 2008.
141. Rayner J.C.W., Best D.J. A contingency table approach to nonparametric testing. – Boca Raton, FL: Chapman & Hall / CRC, 2000.
142. Reineke D.M., Baggett J., Elfessi A. A note on the effect of skewness, kurtosis, and shifting on one–sample t and sign tests // Journal of Statistics Education, 2003, vol. 11, no. 3.
143. Rhiel S.G., Chaffin W.W. An investigation of the large–sample/small–sample approach to the one–sample test for a mean (sigma unknown) // Journal of Statistics Education, 1996, vol. 4, no. 3.
144. Rosenblatt M. Limit theorems associated with variants of the Von Mises statistic // The Annals of Mathematical Statistics, 1952, vol. 23, no. 4, pp. 617–623.
145. Sahai H., Khurshid A. Confidence intervals for the mean of a Poisson distribution: A review // Biometrical Journal, 2007, vol. 35, no. 7, pp. 857–867.
146. Salvatore D., Reagle D. Statistics and econometrics. – London, UK: McGraw–Hill, 2003.
147. Santer T.J. Small–sample comparisons of confidence intervals for the difference of

two independent binomial proportions / T.J. Santer, V. Pradhan, P. Senchaudhuri et al. // Computational Statistics & Data Analysis, 2007, vol. 51, pp. 5791–5799.
148. Sauro J., Lewis J.R. Estimating completion rates from small samples using binomial confidence intervals: Comparisons and recommendations // Proceedings of the Human Factors and Ergonomics Society Annual Meeting (HFES 2005) Orlando, FL, 2005.
149. Schisterman E. Statistical inference for the area under the ROC curve in the presence of random measurement error / E. Schisterman, D. Faraggi, B. Reiser et al. // American Journal of Epidemiology, 2001, vol. 154, pp. 174–179.
150. Sheskin D.J. Handbook of parametric and nonparametric statistical procedures. – Boca Raton, FL: Chapman & Hall / CRC, 2000.
151. Siegel S., Castellan Jr. N.J. Nonparametric statistics for the behavioral sciences. – London, UK: McGraw–Hill, 1988.
152. Siegel S., Tukey J.W. A nonparametric sum of ranks procedure for relative spread in unpaired samples. // Journal of the American Statistical Association, 1960, vol. 55, pp. 429–445.
153. Sistrom C.L., Garvan C.W. Proportions, odds, and risk // Radiology, 2004, vol. 230, no. 1, pp. 12–19.
154. Snedecor G.W., Cochran W.G. Statistical methods. – Ames, IA: Iowa State University Press, 1980.
155. Solorzano E. Nonparametric multiple comparisons with more than one control using normal scores and Savage statistics // InterStat (Statistics on the Internet), November 2004.
156. Sprent P., Smeeton N.C. Applied nonparametric statistical methods. – London, UK: Chapman & Hall / CRC, 2005.
157. Stephan C. Comparison of eight computer programs for receiver–operating characteristic analysis / C. Stephan, S. Wesseling, T. Schink et al. // Clinical Chemistry, March 2003, vol. 49, no. 3, pp. 433–439.
158. Sterne J.A.C., Smith G.D., Cox D.R. Sifting the evidence – what's wrong with significance tests? Another comment on the role of statistical methods // British Medical Journal, 2001, vol. 322, pp. 226–231.
159. Stuart A.A. A test for homogeneity of the marginal distributions in a two–way classification // Biometrika, 1955, vol. 42, pp. 412–416.
160. Suissa S., Shuster J.J. The 2 x 2 matched pairs trial: exact unconditional design and analysis // Biometrics, 1991, vol. 47, pp. 361–372.
161. Sundrum R.M. On Lehmann's two–sample test // The Annals of Mathematical Statistics, 1954, vol. 25, no. 1, pp. 139–145.
162. Swets J.A., Dawes, R.M., Monahan, J. Better decisions through science // Scientific American, 2000, vol. 283, pp. 82–87.
163. Tang M.–L., Tang N.–S., Chan I. S. F. Confidence interval construction for proportion difference in small–sample paired studies // Statistics in Medicine, 2005, vol. 24, no. 23, pp. 3565–3579.
164. Tello R., Crewson P.E. Hypothesis testing II: Means // Radiology, 2003, vol. 227, no. 1, pp. 1–4.
165. Van Belle G. Biostatistics: A methodology for the health sciences // G. van Belle, L.D. Fisher, P.J. Heagerty et al. – New York, NY: John Wiley & Sons, 2003.
166. Van de Wiel M.A. Exact distributions of multiple comparisons rank statistics // Journal of the American Statistical Association, 2002, vol. 97, no. 460, pp. 1081–1089.
167. Van de Wiel M.A. Exact non–null distributions of rank statistics // Communications in Statistics – Simulation and Computation, 2001, vol. 30, no. 4, pp. 1011–1030.
168. Vergara I. StAR: a simple tool for the statistical comparison of ROC curves // I.

Vergara, T. Norambuena, E. Ferrada et al. // BMC Bioinformatics, 2008, vol. 9, no. 265.
169. Vickers A.J. Parametric versus non–parametric statistics in the analysis of randomized trials with non–normally distributed data // BMC Medical Research Methodology, November 2005, vol. 5, pp. 35–47.
170. Viera A.J., Garrett J.M. Understanding interobserver agreement: The Kappa statistic // Family Medicine, May 2005, vol. 37, no. 5, pp. 360–363.
171. Wegner L.H. Properties of some two–sample tests based on a particular measure of discrepancy // The Annals of Mathematical Statistics, December 1956, vol. 27, no. 4, pp. 1006–1016.
172. Wellek S. Testing statistical hypotheses of equivalence. – Boca Raton, FL: Chapman & Hall / CRC, 2003.
173. Wilcox R.R. Fundamentals of modern statistical methods. – New York, NY: Springer–Verlag, 2001.
174. Wilcox R.R. New designs in analysis of variance // Annual Review of Psychology, January 1987, vol. 38, pp. 29–60.
175. Wolfowitz J. Non–parametric statistical inference // Proceedings of the Berkeley symposium on mathematical statistics and probability, August 13–18, 1945 and January 27–29, 1946 / Ed. by J. Neyman. – Berkeley, CA: University of California Press, 1949, pp. 93–113.
176. Xiao Y., Gordon A., Yakovlev A. A C++ program for the Cramer–Von Mises two–sample test // Journal of Statistical Software, December 2006, vol. 17, no. 8.
177. Yates F., Irwin J.O. Contingency tables involving small numbers and the χ^2 test // Supplement to Journal of the Royal Statistical Society, 1934, vol. 1, 217–235.
178. Youden W.J. Index for rating diagnostic tests // Cancer, 1950, vol. 3, no. 1, pp. 32–35.
179. Zaykin D.V., Meng Z., Ghosh S.K. Interval estimation of genetic susceptibility for retrospective case–control studies // BMC Genetics, 11 May 2004, vol. 5, no. 9.
180. Zhou X.–H., McClish D.K., Obuchowski N.A. Statistical methods in diagnostic medicine. – New York, NY: John Wiley & Sons, 2002.
181. Zou K.H. Hypothesis testing I: Proportions / K.H. Zou, J.R. Fielding, S.G. Silverman et al. // Radiology, 2003, vol. 226, no. 3, pp. 609–613.
182. Zweig M.H., Campbell G. ROC plots: A fundamental evaluation tool in clinical medicine // Clinical Chemistry, vol. 39, no. 4, 1993.
183. Айвазян С.А., Енюков И.С., Мешалкин Л.Д. Прикладная статистика: Основы моделирования и первичная обработка данных. Справочное издание. – М.: Финансы и статистика, 1983.
184. Айвазян С.А., Мхитарян В.С. Прикладная статистика и основы эконометрики. – М.: Издательское объединение «ЮНИТИ», 1998.
185. Афифи А., Эйзен С. Статистический анализ: Подход с использованием ЭВМ. – М.: Мир, 1982.
186. Бабич П.Н., Чубенко А.В., Лапач С.Н. Применение современных статистических методов в практике клинических исследований. Сообщение третье. Отношение шансов: понятие, вычисление и интерпретация // Український Медичний Часопис, 2005, № 2 (46), с. 113–119.
187. Белова Е.Б. Компьютеризованный статистический анализ для историков. Учебное пособие / Е.Б. Белова, Л.И. Бородкин, И.М. Гарскова и др. – М.: Издательство Московского государственного университета, 1999.
188. Бендат Дж., Пирсол А. Прикладной анализ случайных данных. – М.: Мир, 1989.

189. Благовещенский Ю.Н., Самсонова В.П., Дмитриев Е.А. Непараметрические методы в почвенных исследованиях. – М.: Наука, 1987.
190. Большев Л.Н., Смирнов Н.В. Таблицы математической статистики. – М.: Наука, 1983.
191. Боровков А.А. Математическая статистика. Оценка параметров. Проверка гипотез. – М.: Наука, 1984.
192. Брандт З. Анализ данных. Статистические и вычислительные методы для научных работников и инженеров. – М.: Мир, ООО «Издательство АСТ», 2003.
193. Браунли К.А. Статистическая теория и методология в науке и технике. – М.: Наука, 1977.
194. Ван дер Варден Б.Л. Математическая статистика. – М.: Издательство иностранной литературы, 1960.
195. Власов В.В. Эпидемиология: Учебное пособие для вузов. – М.: Издательский дом «ГЭОТАР–МЕД», 2004.
196. Власов В.В. Эффективность диагностических исследований.– М.: Медицина, 1988.
197. Воинов В.Г. Об оптимальных свойствах критерия Рао–Робсон–Никулина // Заводская лаборатория. Диагностика материалов, 2006, № 3, с. 65–70.
198. Воробьев К.П. Формат современной журнальной публикации по результатам клинического исследования. Часть 3. Дизайн клинического исследования // Український медичний часопис, 2008, № 2, с. 150–160.
199. Гаек Я., Шидак З. Теория ранговых критериев. – М.: Наука, 1971.
200. Гайдышев И. Анализ и обработка данных: специальный справочник. – СПб: Питер, 2001.
201. Гайдышев И.П. Статистический анализ результатов выборов // Наука и образование Зауралья, 2004, № 1, с. 189–194.
202. Гланц С. Медико–биологическая статистика. – М.: Практика, 1998.
203. Глотов Н.В. Биометрия / Н.В. Глотов, Л.А. Животовский, Н.В. Хованов и др. – Л.: Издательство Ленинградского государственного университета, 1982.
204. Гублер Е.В. Вычислительные методы распознавания патологических процессов. – Л.: Медицина, 1970.
205. Гублер Е.В. Информатика в патологии, клинической медицине и педиатрии. – Л.: Медицина, 1990.
206. Гублер Е.В., Генкин А.А. Применение непараметрических критериев статистики в медико–биологических исследованиях. – Л.: Медицина, 1973.
207. Джонсон Н., Лион Ф. Статистика и планирование эксперимента в технике и науке. Методы обработки данных. – М.: Мир, 1980.
208. Ефимов А.Н. Порядковые статистики – их свойства и приложения. – М.: Знание, 1980.
209. Зайцев Г.Н. Математическая статистика в экспериментальной ботанике. – М.: Наука, 1984.
210. Закс Л. Статистическое оценивание. – М.: Статистика, 1976.
211. Заславский А.А., Пригарина Т.А. Оценка согласованности субъективных классификаций при заданных классах // Социология: методология, методы, математические модели, № 3–4, с. 84–109.
212. Коган Р.И., Белов Ю.П., Родионов Д.А. Статистические ранговые критерии в геологии. – М.: Недра, 1983.
213. Корнилов С.Г. Оптимальные объемы групп при сравнении средних / Биометрический анализ в биологии. – М.: Издательство Московского

государственного университета, 1982, с. 71–90.
214. Костин В.С. Статистика для сравнения классификаций // Информационные технологии в гуманитарных исследованиях: Сборник трудов. Выпуск 6. – Новосибирск, 2003, с. 57–65.
215. Крамер Г. Математические методы статистики. – М.: Мир, 1975.
216. Кулаичев А.П. Методы и средства анализа данных в среде Windows®. STADIA. – М.: Информатика и компьютеры, 1999.
217. Лакин Г.Ф. Биометрия. – М.: Высшая школа, 1990.
218. Лванга С.К. Обучение медицинской статистике: Двадцать конспектов лекций и семинаров / Под ред. С.К. Лванга, Ч.–Е. Тыэ. – М.: Медицина, 1989.
219. Леман Э. Проверка статистических гипотез. – М.: Наука, 1979.
220. Лемешко Б.Ю., Лемешко С.Б. О сходимости распределений статистик и мощности критериев однородности Смирнова и Лемана–Розенблатта // Измерительная техника, 2005, № 12, с .9–14.
221. Мартынов Г.В. Критерии омега–квадрат. – М.: Наука, 1978.
222. Медик В.А., Токмачев М.С., Фишман Б.Б. Статистика в медицине и биологии: Руководство. В 2–х томах / Под ред. Ю.М. Комарова. Т. 1. Теоретическая статистика. – М.: Медицина, 2000.
223. Мюллер П., Нойман П., Шторм Р. Таблицы по математической статистике. – М.: Финансы и статистика, 1982.
224. Налимов В.В. Применение математической статистики при анализе вещества. – М.: Государственное издательство физико–математической литературы, 1960.
225. Налимов В.В. Теория эксперимента. – М.: Наука, 1971.
226. Никитин Я.Ю. Асимптотическая эффективность непараметрических критериев. – М.: Наука, 1995.
227. Никулин М.С., Юсас Й. Об учете числа совпадений в двухвыборочном критерии Вилкоксона // Записки научного семинара ЛОМИ, т. 119, «Проблемы теории вероятностных распределений. VII». – Л.: Наука, 1982, с. 195–197.
228. Новиков Д.А. Статистические методы в педагогических исследованиях (типовые случаи). – М.: МЗ–Пресс, 2004.
229. Новиков Д.А., Новочадов В.В. Статистические методы в медико–биологическом эксперименте (типовые случаи). – Волгоград: Издательство ВолГМУ, 2005.
230. Новиков Ф.А. Дискретная математика для программистов. Учебник для вузов. – СПб.: Питер, 2005.
231. Новицкий П.В., Зограф И.А. Оценка погрешностей результатов измерений. – Л.: Энергоатомиздат, 1985.
232. Орлов А.И. Прикладная статистика. Учебник. – М.: Издательство «Экзамен», 2006.
233. Оуэн Д.Б. Сборник статистических таблиц. – М.: Вычислительный центр АН СССР, 1966.
234. Петри А., Сэбин К. Наглядная статистика в медицине. – М.: Издательский дом «ГЭОТАР–МЕД», 2003.
235. Петрович М.Л., Давидович М.И. Статистическое оценивание и проверка гипотез на ЭВМ. – М.: Финансы и статистика, 1989.
236. Подольная М.А., Кобринский Б.А. Показатели и методика расчета эпидемиологических характеристик риска // Российский вестник перинатологии и педиатрии, 2000, № 6, с. 52–54.
237. Поллард Дж. Справочник по вычислительным методам статистики. – М.:

Финансы и статистика, 1982.
238. Прохоров Ю.В. Вероятность и математическая статистика: Энциклопедия / Ред. Ю.В. Прохоров. – М.: Научное издательство «Большая Российская энциклопедия», 1999.
239. Родионов Д.А. Справочник по математическим методам в геологии / Д.А. Родионов, Р.И. Коган, В.А. Голубева и др. – М.: Недра, 1987.
240. Родионов Д.А. Статистические решения в геологии. – М.: Недра, 1981.
241. Романовский В.И. Математическая статистика. Кн.2. Оперативные методы математической статистики. – Ташкент: Издательство Академии наук УзССР, 1963.
242. Рунион Р. Справочник по непараметрической статистике. – М.: Финансы и статистика, 1982.
243. Сборник научных программ на Фортране. Выпуск 1. Статистика. – М.: Статистика, 1974.
244. Сергиенко В.И., Бондарева И.Б. Математическая статистика в клинических исследованиях – М.: Издательский дом «ГЭОТАР–МЕД», 2001.
245. Скрипник В.М. Анализ надежности технических систем по цензурированным выборкам / В.М. Скрипник, А.Е. Назин, Ю.Г. Приходько и др. – М.: Радио и связь, 1988.
246. Тюрин Ю.Н., Макаров А.А. Статистический анализ данных на компьютере. – М.: ИНФРА–М, 1999.
247. Уилкс С. Математическая статистика. – М.: Наука, 1967.
248. Фишер Р.А. Статистические методы для исследователей. – М.: Госстатиздат, 1958.
249. Флейс Дж. Статистические методы для изучения таблиц долей и пропорций. – М.: Финансы и статистика, 1989.
250. Флетчер Р., Флетчер С., Вагнер Э. Клиническая эпидемиология: Основы доказательной медицины. – М.: Медиа Сфера, 2004.
251. Хеттманспергер Т. Статистические выводы, основанные на рангах. – М.: Финансы и статистика, 1987.
252. Холлендер М., Вулф Д. Непараметрические методы статистики. – М.: Финансы и статистика, 1983.
253. Хромов–Борисов Н.Н. Биометрические аспекты популяционной генетики / В кн. Кайданов Л.З. Генетика популяций. – М.: Высшая школа, 1996, с. 251–308.
254. Хьюбер П. Робастность в статистике. – М.: Мир, 1984.
255. Чубенко А.В. Применение современных статистических методов в практике клинических исследований. Сообщение первое. Сравнение двух пропорций / А.В. Чубенко, П.Н. Бабич, С.Н. Лапач и др. // Український Медичний Часопис, 2003, № 4, с. 139–143.

Глава 5. Точные критерии

5.1. Введение

Программное обеспечение реализует точные (exact) методы проверки статистических гипотез, иначе известные как комбинаторные (перестановочные, permutational), а также еще ряд методов, допускающих точное решение задачи. Отметим, что точность здесь понимается в смысле решения задачи с установленными ограничениями и принятыми допущениями используемой статистической модели.
В настоящей главе собраны непараметрические методы проверки гипотез, отличительной

особенностью которых является точное вычисление *P*–значений статистик критериев. К данной группе критериев принято относить как методы, основанные на перестановках, так и методы, для которых известны точные распределения статистик критериев (в частности, некоторые ранговые критерии).
Имеется несколько соображений относительно полезности точных непараметрических методов:

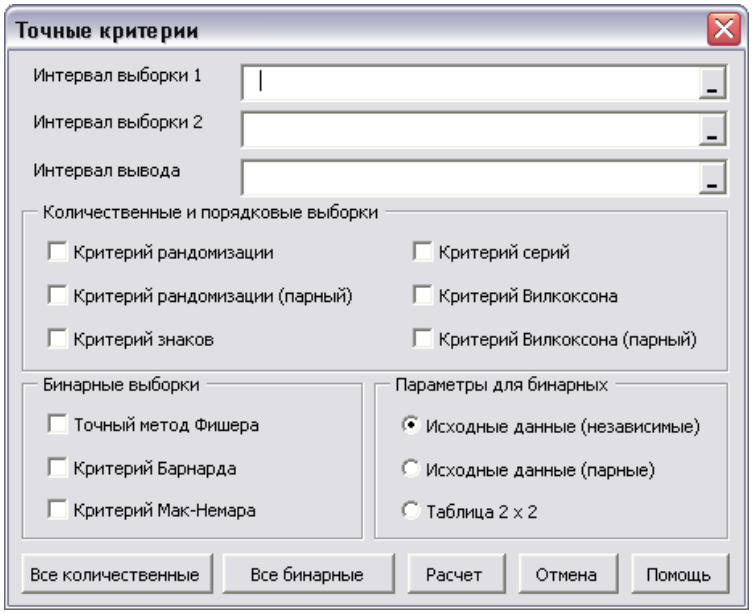
- По данным литературы, параметрические методы могут применяться, только если доказана нормальность распределения (см. главу «Проверка нормальности распределения») анализируемых выборок, но эмпирические выборки, полученные в реальных экспериментах, очень часто не являются нормально распределенными.
- Опять же по данным литературы, параметрические методы могут применяться для больших выборок. Реальные выборки часто содержат небольшое число вариант, что тем более делает полезным непараметрические методы.

Номенклатура методов, входящих в данное программное обеспечение, обеспечивает адекватный анализ выборок произвольного распределения, практически любой численности.
Перед применением любого статистического метода необходимо убедиться, что проверяется статистическая значимость различий именно тех параметров выборок, которые интересуют исследователя, а также в том, что метод соответствует шкале измерения исходных данных (признаков). При выборе метода, неадекватного шкале измерения представленных данных, полученный числовой результат расчета может оказаться лишен какого–либо смысла.

5.2. Работа с программным обеспечением

Выберите из меню программы пункт AtteStat | Точные критерии. На экране появится диалоговое окно, изображенное на рисунке.
Затем проделайте следующие шаги:

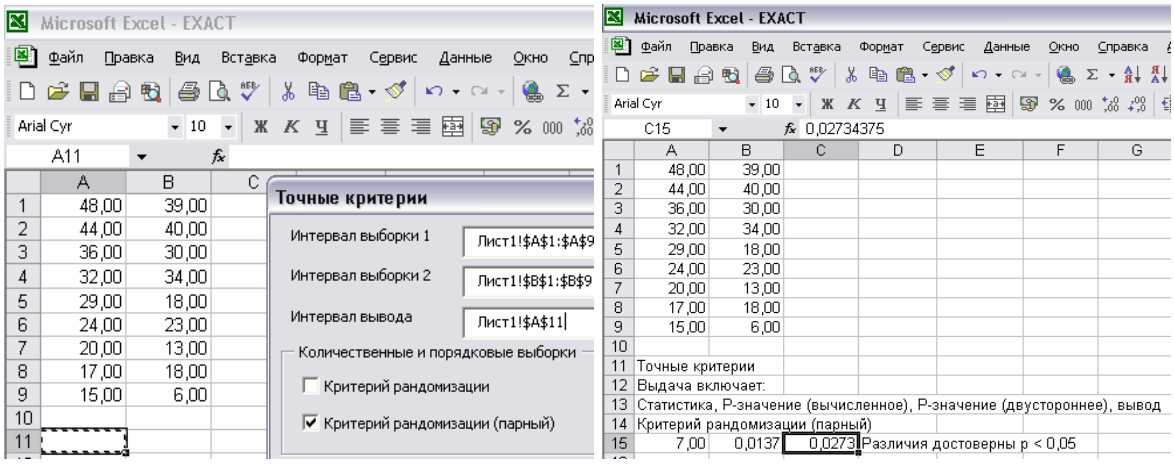
- Выберите или введите интервалы сравниваемых выборок.
- Выберите или введите выходной интервал. Начиная с первой ячейки выходного интервала (следовательно, можно указать только одну ячейку, т.к. остальные ячейки интервала игнорируются), будут выведены результаты вычислений.
- Выберите критерий или группу критериев для проведения статистического расчета. Для выбора группы критериев можно воспользоваться кнопками Все количественные (для выбора всех критериев для количественных или порядковых выборок) или Все бинарные (для выбора всех критериев для дихотомических выборок).
- Для бинарных критериев есть возможность указать программе, заданы исходные данные в виде выборок (по умолчанию) либо в виде таблицы сопряженности типа 2 x 2. Во втором случае в качестве первого столбца таблицы сопряженности укажите интервал выборки 1, в качестве второго столбца укажите интервал выборки 2. Данный метод выделения таблицы сопряженности отличается от принятого в главе «Кросстабуляция» (там выделяется таблица сопряженности целиком). Это сделано ради обеспечения совместимости с другими методами.
- Нажмите кнопку «Выполнить расчет».



После выполнения вычислений будет, начиная с первой ячейки выходного интервала, выведено название статистического критерия, значение статистики критерия, вычисленное *P*–значение, двустороннее *P*–значение и предлагаемый программой вывод о результате проверки статистической гипотезы.
Программное обеспечение берет на себя верификацию исходных данных, выдавая подробную диагностику. При ошибках, вызванных неверными действиями пользователя, или ошибках периода выполнения выдаются сообщения об ошибках.

5.2.1. Пример применения

В качестве примера исследуем исходные данные, приведенные на с. 121 монографии Руниона. Как и в источнике, воспользуемся критерием рандомизации для связанных выборок.



Введем исходные данные: первую выборку в интервал ячеек A1:A9, вторую выборку в интервал ячеек B1:B9. В качестве интервала вывода (начала интервала) укажем ячейку A11. Выберем указанный метод анализа.
После нажатия кнопки «Выполнить расчет» экран примет вид, показанный на фрагменте.

Нулевая гипотеза может быть принята. Результаты совпадают с источником. Подробную интерпретацию результатов см. в описании метода.

5.2.2. Сообщения об ошибках

При ошибках ввода и во время выполнения программы могут выдаваться диагностические сообщения следующих типов:

Ошибка	Комментарий
Не определен интервал переменной.	Вы не выбрали или неверно ввели интервал эмпирической выборки. Лучшим способом избежать ошибки является не ввод, а выделение интервала стандартным образом, т. е. протаскиванием курсора.
Пустая ячейка в области данных.	Проверьте исходные данные и заполните все ячейки, отмеченные Вами как входной интервал. Для избежания ошибок расчета, вызванных разногласиями, трактовать ли пустую ячейку как нуль, программное обеспечение требует заполнения всех ячеек. Если в ячейке не должно быть данных по физической природе исследуемого явления, введите в данную ячейку нуль.
Нечисловой тип данных.	Проверьте типы ячеек входного интервала. Тип может быть только числовым. Проще всего выделить интервал ячеек и явно определить их тип как числовой стандартными средствами.
Не определена область вывода.	Вы не выбрали или неверно ввели выходной интервал. Лучшим способом избежать ошибки является не ввод, а выделение интервала стандартным образом, т. е. протаскиванием курсора.

5.3. Теоретическое обоснование

Все точные критерии базируются на возможности точного вычисления *P*–значения.
Основной группой точных критериев являются методы, основанные на перестановках.
Алгоритмы методов позволяют вычислить точное *P*–значение, зная число благоприятных исходов и общее число исходов опыта, представляющее собой все мыслимые варианты исхода. Следовательно, при вычислении критериев не избежать применения комбинаторных алгоритмов и вычисления дискретных функций распределения, которые могут быть очень трудоемкими в реализации, особенно для больших выборок.
Областью применения большинства рассматриваемых критериев является анализ именно малых выборок, причем для критериев рандомизации ограничение лимитируется только приемлемым для диалоговой программы быстродействием.
В принципе представленными методами могут анализироваться и малые, и большие выборки.
К точным критериям (exact tests), представленным в программе, относятся:

- критерий рандомизации для независимых выборок,
- критерий рандомизации для связанных выборок,
- критерий Вилкоксона для независимых выборок,
- критерий Вилкоксона для связанных выборок,
- точный метод Фишера,
- критерий Барнарда,
- критерий Мак–Немара,
- критерий знаков,

- критерий серий Вальда–Вольфовица.

При формулировании нулевой гипотезы обязательно следует указывать, какие конкретные параметры эмпирических выборок сравниваются с помощью используемого критерия. Данная информация приводится в описании каждого критерия. Нужно указывать это в научной публикации, чтобы читатель имел возможность проверить правильность рассуждений автора. В таблице указаны тестируемые параметры выборок для различных критериев.

Тестируемые параметры	Точный статистический критерий
Положение: среднее и/или медиана (location tests)	Рандомизации, точный метод Фишера, критерий Барнарда, серий Вальда–Вольфовица, критерий Мак–Немара, Вилкоксона
Функция распределения	Знаков

Для пользователей рекомендуются простые и практические источники, например, переведенные на русский язык монография Кендалла с соавт. и книга Руниона. Сравнительный обзор критериев для проверки однородности таблиц сопряженности приводится в статье Мехротра (Mehrotra) с соавт. Обзор подходов Фишера и Барнарда к анализу таблиц сопряженности см. в работе Мартина Андреса (Martin Andres) с соавт.

5.3.1. Критерий рандомизации для независимых выборок

Критерий рандомизации компонент Фишера (критерий рандомизации Фишера–Питмана) для независимых выборок применяется для проверки нулевой гипотезы о том, отобраны ли две независимые выборки из совокупностей с одинаковыми средними значениями. Выборки должны принадлежать количественной шкале.
Критерий рандомизации называется также критерием перестановок, выборочное распределение которого при каждом вычислении должно быть получено заново перебором всех возможных исходов.
Методика теста базируется на идее перебора всех комбинаций наблюдаемых отметок. Пусть даны две выборки: x_i, $i = 1,2,...,n_x$, и y_i, $i = 1,2,...,n_y$, где n_x, n_y – численности выборок. Сумма, меньшая из двух наблюдаемых, будет

$$S = \min\left(\sum_{i=1}^{n_x} x_i, \sum_{i=1}^{n_y} y_i\right).$$

Число благоприятных исходов вычисляется по формуле

$$N = 2\sum_{i=1}^{C_n^m} n_i, n_i = \begin{cases} 0, s_i < S, \\ 1, s_i \geq S, \end{cases}$$

где n_i – оценка i–го исхода,
C_n^m – общее число исходов – число сочетаний из n по m,
$n = n_x + n_y$ – численность объединенной выборки,
m – численность выборки, соответствующей минимальной сумме

$$s_i = \sum_{j=1}^{m} z_j, i = 1,...,C_n^m,$$

где z_j, $j = 1,2,...,m$ – массив сочетаний из объединенной выборки.
Двустороннее P–значение вычисляется по формуле

$$p = \frac{N}{C_n^m}$$

и сравнивается с заданным уровнем значимости.
Критерий рекомендуется для малых выборок (численность каждой выборки от 5 до 12); при численности выборок, большей 12, время расчета может стать неприемлемо большим для диалоговой программной системы, поэтому при больших численностях выборок вместо описанного здесь критерия рекомендуется применять W–критерий Вилкоксона (см. главу «Непараметрическая статистика»), являющийся критерием ранговой рандомизации.

См. также описание и пример критерия Питмана–Уэлча в монографии Файнштайн (Feinstein). См. справочник Руниона, статьи Питмана (Pitman), Кайзера (Kaiser), монографии Фишера (Fisher), Зигеля (Siegel) с соавт.

5.3.2. Критерий рандомизации для связанных выборок

Критерий рандомизации компонент Фишера (критерий рандомизации Фишера–Питмана) для связанных выборок применяется для проверки нулевой гипотезы о равенстве средних значений двух связанных совокупностей. Выборки должны принадлежать количественной шкале.
Критерий рандомизации называется также критерием перестановок, выборочное распределение которого при каждом вычислении должно быть получено заново перебором всех возможных исходов.
Основным моментом в реализации критерия является перебор возможных исходов, построенных из разностных отметок. Пусть даны две выборки: x_i, y_i, i = 1,2,...,n , где n – число пар экспериментальных значений. Тогда сумма массива разностных отметок будет

$$S = \sum_{i=1}^{n} s_i .$$

Определим значения разностных отметок:

$$s_i = \sum_{j=1}^{n} a_{ij}(x_j - y_j), i = 1,...,2^n ,$$

где a_{ij}, i = 1,2,...,2n; j = 1,2,...,n – элементы матрицы возможных исходов.
Отметим, что в некоторых источниках разность вариант в показанной выше формуле берется по модулю. Однако анализ формулы показывает, что в процессе перебора операция взятия модуля в данном случае значения не имеет.
Систематизацию перебора всех возможных исходов удобно провести в соответствии с ортогональным планом эксперимента первого порядка. Размер полного ортогонального плана составляет 2n строк на n столбцов, причем j–й столбец размером 2n представляет собой чередующиеся с шагом $2j^{-1}$ величины +1 и –1, j = 1,2,...,n.
Число благоприятных исходов вычисляется по формуле:

$$N = \sum_{i=1}^{2^n} n_i, n_i = \begin{cases} 0, s_i < S, \\ 1, s_i \geq S. \end{cases}$$

Двустороннее P–значение, вычисляемое по формуле

$$p = \frac{N}{2^n},$$

сравнивается с заданным уровнем значимости.
Критерий рекомендуется для малых выборок (численность каждой выборки от 5 до 12); при численности выборок, большей 12, время расчета может стать неприемлемо большим для

диалоговой программной системы, поэтому при больших численностях выборок рекомендуется применять *T*–критерий Вилкоксона (см. главу «Непараметрическая статистика»), являющийся критерием ранговой рандомизации.

См. справочник Руниона, статьи Питмана (Pitman), Кайзера (Kaiser), монографии Фишера (Fisher), Зигеля (Siegel) с соавт. Об ортогональных планах см. монографии Шеффлера (Scheffler), Корикова, Монтгомери.

5.3.3. Критерий Вилкоксона для независимых выборок

Критерий Вилкоксона для независимых выборок является аналогом критерия рандомизации для независимых выборок с той разницей, что все операции производятся не над вариантами выборок, а над их рангами.
Метод имеет те же ограничения, что и критерий рандомизации, уступает ему в мощности, но не уступает в трудоемкости вычислений и поэтому находит ограниченное применение.
Для больших выборок следует использовать асимптотический *W*–критерий Вилкоксона (см. главу «Непараметрическая статистика»).

О точном вычислении критерия Вилкоксона для независимых выборок см. Браунли, Уилкса.

5.3.4. Критерий Вилкоксона для связанных выборок

Критерий Вилкоксона для связанных выборок является аналогом критерия рандомизации для связанных выборок с той разницей, что все операции производятся не над вариантами выборок, а над их рангами.
Метод имеет те же ограничения, что и критерий рандомизации, уступает ему в мощности, но не уступает в трудоемкости вычислений и поэтому находит ограниченное применение.
Для больших выборок следует использовать асимптотический *T*–критерий Вилкоксона, реализованный в главе «Непараметрическая статистика».

О точном вычислении критерия Вилкоксона для связанных выборок см. Браунли.

5.3.5. Точный метод Фишера

Точный метод Фишера (критерий Фишера, точный метод Фишера–Ирвина, критерий Фишера–Ирвина, Fisher's exact test, Fisher–Irwin test, Fisher–Yates–Irwin exact test) применяется для проверки нулевой гипотезы о том, отобраны ли две исследуемые бинарные выборки из генеральных совокупностей с одинаковой частотой встречаемости изучаемого эффекта. Рассматриваемый метод предназначен для обработки так называемых четырехпольных (четырехклеточных) таблиц, или таблиц типа 2 x 2. Для применения критерия анализируемые выборки должны принадлежать дихотомической шкале измерения. В настоящей программе принято, что исходные выборки должны состоять только из нулей и единиц, причем нуль означает отсутствие признака (эффекта), а единица означает наличие признака (эффекта).

	Наличие эффекта A		
	Да	Нет	
Выборка (группа) 1	a	b	$M_1 = a + b$
Выборка (группа) 2	c	d	$M_2 = c + d$
Сумма	$N_1 = a + c$	$N_2 = b + d$	$n = a + b + c + d$

Вычисление односторонних достигнутых уровней значимости критерия производится путем суммирования вероятностей всех вариантов $p(X)$ заполнения таблицы сопряженности:

$$P_U = \sum_{\substack{T(X)>T(X_0)\\ ad<bc}} p(X),$$

$$P_L = \sum_{\substack{T(X)>T(X_0)\\ ad\geq bc}} p(X),$$

где $T(X)$ – статистика Вальда для текущего варианта заполнения таблицы,
$T(X_0)$ – статистика Вальда исходной таблицы сопряженности.
Двусторонний достигнутый уровень значимости критерия равен
$P_F = P_U + P_L$.
Статистика Вальда в данном случае вычисляется по формуле.

$$T(X) = \frac{n\left|\frac{a}{a+c} - \frac{b}{b+d}\right|}{\sqrt{(a+b)(c+d)\left(\frac{1}{a+c} + \frac{1}{b+d}\right)}},$$

где a – число наблюдений с эффектом A в первой выборке,
b – число наблюдений без эффекта A в первой выборке,
c – число наблюдений с эффектом A во второй выборке,
d – число наблюдений без эффекта A во второй выборке,
$n = a + b + c + d$ – численность таблицы сопряженности.
Статистика Вальда выводится программой в качестве критериальной статистики.
Варианты заполнения таблицы сопряженности планируются при условии сохранения всех маргинальных сумм. Это означает, что для всех вариантов таблицы маргинальные суммы N_1, N_2, M_1, M_2 должны быть одинаковыми.
Некоторыми авторами приводится эквивалентная (и гораздо более быстрая в вычислении) формула вычисления критерия. Отличие заключается в замене статистики Вальда текущего варианта заполнения таблицы и статистики Вальда исходной таблицы на, соответственно, вероятность $p(X)$ и вероятность исходной таблицы $p(X_0)$. Именно данная формула используется в настоящем программном обеспечении.
Вместо указанной точной условной вероятности биномиального распределения Фишер предложил использовать вероятность гипергеометрического распределения

$$p(X) = \frac{C_{a+b}^{a} C_{c+d}^{c}}{C_{n}^{a+c}} = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{n!a!b!c!d!}.$$

По этой причине описываемый критерий и все методы, основанные на данной идее, называются условными критериями (conditional tests). Подробные соображения по данному вопросу изложены в т. 1 справочника под ред. Ллойда с соавт. В этом же источнике описаны методики получения таблиц сопряженности.
Настоящее программное обеспечение может производить вычисление точным методом Фишера как на основе исходных выборок, так и обрабатывая заранее полученную таблицу сопряженности. В первом случае программа сама вычисляет таблицу сопряженности, а вводить следует исходные выборки, как это предусмотрено для всех других функций данной главы. Во втором случае в качестве первого столбца таблицы сопряженности укажите интервал выборки 1, в качестве второго столбца укажите интервал выборки 2.
При работе нужно учитывать, что критерий трудоемок в вычислении, причем снять задачу с выполнения, не дожидаясь ее нормального окончания, можно только средствами операционной системы.

См. монографии ван Бель (van Belle) с соавт., Черник (Chernick) с соавт., Ле (Le), диссертацию Бучана (Buchan). Сравнительный обзор приводится в статье Мехротра (Mehrotra) с соавт. Описание критерия см. в книгах Лемана, Руниона, Флейса, Кендалла с соавт., Гайдышева, статьях Бауэра (Bower), Бергера (Berger) с соавт. На основе идеи Фишера для обработки таблиц сопряженности типа *r* x *c* Фриман (Freeman) и Холтон (Halton) разработали расширенный тест, представленный в главе «Кросстабуляция».

5.3.6. Критерий Барнарда

Критерий Барнарда (Barnard's test) применяется для проверки нулевой гипотезы о том, отобраны ли две исследуемые бинарные выборки из генеральных совокупностей с одинаковой частотой встречаемости изучаемого эффекта. Рассматриваемый метод предназначен для обработки так называемых четырехпольных (четырехклеточных) таблиц, или таблиц типа 2 x 2. Для применения критерия анализируемые выборки должны принадлежать дихотомической шкале измерения. В программе принято, что исходные выборки должны состоять только из нулей и единиц, причем нуль означает отсутствие признака (эффекта), а единица означает наличие признака (эффекта).

	Наличие эффекта A		
	Да	Нет	
Выборка (группа) 1	a	b	$a + b$
Выборка (группа) 2	c	d	$c + d$
Сумма	$N_1 = a + c$	$N_2 = b + d$	$n = a + b + c + d$

Точный двусторонний достигнутый уровень значимости критерия P_B определяется как

$$P_B = \sup_{0<\pi<1}\left\{\sum_{T(X)>T(X_0)} p(X,\pi)\right\},$$

где π – параметр распределения,
$p(X,\pi)$– вероятность варианта заполнения таблицы,
$T(X)$ – статистика Вальда варианта заполнения таблицы сопряженности,
$T(X_0)$ – статистика Вальда исходной таблицы сопряженности.
Максимум целевой функции может быть найден с помощью методов оптимизации.
Вероятность таблицы сопряженности вычисляется по формуле вероятности биномиального распределения

$$p(X,\pi) = C_{a+c}^{a} C_{b+d}^{b} \pi^{a+b} (1-\pi)^{c+d}.$$

Это первое отличие критерия Барнарда от критерия Фишера (см. точный метод Фишера). По этой причине и в противоположность условным критериям описываемый критерий и все методы, основанные на данной идее, называются безусловными (unconditional tests).
Статистика Вальда в рассматриваемом случае имеет вид

$$T(X) = \frac{n\left|\frac{a}{a+c} - \frac{b}{b+d}\right|}{\sqrt{(a+b)(c+d)\left(\frac{1}{a+c} + \frac{1}{b+d}\right)}},$$

где a – число наблюдений с эффектом A в первой выборке,
b – число наблюдений без эффекта A в первой выборке,
c – число наблюдений с эффектом A во второй выборке,

d – число наблюдений без эффекта A во второй выборке,
$n = a + b + c + d$ – численность таблицы сопряженности.
Вычисленное «оптимальное» значение параметра π выводится программой, после чего в качестве критериальной статистики программой выводится статистика Вальда и P–значение, как предусмотрено и во всех остальных методах, реализованных в программе.
Варианты заполнения таблицы сопряженности планируются при условии сохранения маргинальных сумм N_1 и N_2. Это означает, что для всех вариантов таблицы значения маргинальные суммы N_1 и N_2 должны быть одинаковыми. Это второе отличие критерия Барнарда от критерия Фишера. По сути, эти два отличия отражают все основные подходы к обработке таблиц сопряженности (как 2 x 2, так r x c) и определяют их два основных направления развития:

- Подход Фишера: гипергеометрическое распределение и все фиксированные маргинальные суммы.
- Подход Барнарда: биномиальное распределение (с вычислением оптимального параметра) и фиксированные суммы столбцов.

Настоящее программное обеспечение может производить вычисление критерия Барнарда как на основе исходных выборок, так и обрабатывая заранее полученную таблицу сопряженности. В первом случае программа сама вычисляет таблицу сопряженности, а вводить следует исходные выборки, как это предусмотрено для всех других функций настоящей главы. Во втором случае в качестве первого столбца таблицы сопряженности укажите интервал выборки 1, в качестве второго столбца укажите интервал выборки 2.
Критерий Барнарда более трудоемок в вычислении, чем точный метод Фишера. Это вызвано необходимостью решать задачу поиска оптимального значения параметра. Задача упрощается тем, что зависимость целевой функции от данного параметра, как показывают исследования, симметрична относительно π = 0,5 и имеет форму либо «шляпы», либо «сомбреро», в зависимости от соотношения частот таблицы сопряженности. Стратегия решения – стандартная. На начальном (глобальном) этапе простым методом перебора производится поиск интервала локализации параметра распределения, который затем уточняется до нужной точности с помощью быстродействующего локального метода.
При работе нужно учитывать, что критерий трудоемок в вычислении, причем снять задачу с выполнения можно, не дожидаясь ее нормального окончания, только средствами операционной системы.

Описание критерия см. в оригинальных статьях Барнарда (Barnard), работах Мехта (Mehta) с соавт., статье Мартин Андрес (Martin Andres) с соавт. В сравнительном обзоре Мехротра (Mehrotra) с соавт. дано современное состояние вопроса и представлены дальнейшие развития идеи Барнарда. Методы локальной оптимизации см. в пособии Вержбицкого.

5.3.7. Критерий Мак–Немара

Критерий Мак–Немара (McNemar’s test) применяется для проверки нулевой гипотезы о том, отобраны ли две исследуемые парные бинарные выборки из генеральных совокупностей с одинаковой частотой встречаемости изучаемого эффекта. Рассматриваемый метод предназначен для обработки так называемых четырехпольных (четырехклеточных) таблиц, или таблиц типа 2 x 2:

		Эффект *B*	
		Да	Нет
Эффект *A*	Да	*a*	*b*

	Нет	*c*	*d*

Метод идеально подходит для анализа данных типа «до и после».
Вычисление статистики критерия производится по формуле:

$$X^2 = \frac{(|b-c|-Y)^2}{b+c},$$

где b – число индивидуумов с наличием эффекта A и отсутствием эффекта B,
c – число индивидуумов с отсутствием эффекта A и наличием эффекта B,
$Y = 0$ – если не используется поправка на непрерывность (поправка Йэйтса),
$Y = 1$ – если используется поправка на непрерывность.
В программе представлены 3 варианта критерия:

1. Асимптотика хи–квадрат.
2. Асимптотика хи–квадрат с поправкой Йэйтса.
3. Точный вариант критерия.

По поводу вычисления P–значений в первых двух вариантах критерия см. главу «Непараметрическая статистика».
Точный двусторонний достигнутый уровень значимости критерия P_X определяется суммированием вероятностей всех вариантов заполнения таблицы, при условии сохранения суммы ячеек b и c исходной таблицы, как

$$P_X = \sum_{p(X)\le p(X_0)} p(X),$$

где $p(X)$ – вероятность варианта заполнения таблицы,
$p(X_0)$ – вероятность исходной таблицы.
Вероятность заполнения таблицы вычисляется по формуле биномиального распределения (адаптированной к данному случаю)

$$p(X) = \frac{0{,}5^{b+c}(b+c)!}{b!c!}.$$

Настоящее программное обеспечение может производить вычисление критерия Мак–Немара как на основе исходных выборок, так и обрабатывая заранее полученную таблицу 2 x 2. В первом случае программа сама вычисляет таблицу 2 x 2, а вводить следует исходные выборки, как это предусмотрено для всех других функций данной главы. Во втором случае в качестве первого столбца таблицы 2 x 2 укажите интервал выборки 1, в качестве второго столбца укажите интервал выборки 2.

Описание критерия см. в статьях Беннетта (Bennett) с соавт., Дуайера (Dwyer), Лиделла (Liddell).

5.3.8. Критерий знаков

Критерий знаков (критерий знаков Фишера) предназначен для проверки гипотезы об однородности распределения совокупности, что эквивалентно проверке гипотезы о равенстве функций распределения. Критерий часто используется при сравнении эффективности двух различных способов воздействия на n объектов и, таким образом, он может применяться и для связанных выборок. Выборки могут принадлежать порядковой или количественной шкале. Требованием является равная численность сравниваемых выборок, в том числе и независимых выборок. Статистика критерия вычисляется как число положительных разностей вариант выборок:

$$B = \sum_{i=1}^{n} s(x_i, y_i),$$

$$s(x_i, y_y) = \begin{cases} 1, x_i > y_i, \\ 0, x_i < y_i, \end{cases}$$

где

x_i, y_i, i = 1,2,...,n – варианты выборок,

n – численность каждой выборки.

Если среди значений вариант есть совпадающие, т. е. $x_i = y_i$, i = 1,2,...,n, то данные пары значений отбрасываются и, соответственно, на число отброшенных значений сокращается численность n.

В представленной программе точное критическое значение критерия знаков для любой численности вычисляется на основе функции биномиального распределения с параметрами (B; n; 0,5) .

Для больших выборок (на самом деле аппроксимация хорошо работает уже при численности, равной 25 вариант в каждой выборке) P–значение может также вычисляться посредством нормальной аппроксимации критического значения критерия. При этом модифицированная статистика

$$\frac{B - EB}{\sqrt{DB}},$$

где $EB = n / 2$ – математическое ожидание,

$DB = n / 4$ – дисперсия,

распределена по стандартному нормальному закону. Данная аппроксимация в программе не используется и дана для полноты информации.

См. книги Браунли, Лемана.

5.3.9. Критерий серий Вальда–Вольфовица

Критерий серий Вальда–Вольфовица (Wald–Wolfowitz runs test) предназначен для проверки нулевой гипотезы о равенстве целого ряда параметров двух сравниваемых выборок, включая медианы и коэффициенты асимметрии. Критерий применяется в случае, если исследователя интересует, имеют ли место любые различия между совокупностями. Выборки могут принадлежать порядковой или количественной шкале. Суть расчета заключается в объединении выборок с численностями n_1 и n_2 в одну выборку общей численностью $n_1 + n_2$, ее сортировке по возрастанию или убыванию и подсчете числа серий элементов R, относящихся к первой и второй выборкам.

Точное одностороннее P–значение статистики критерия вычисляется как

$$P(r \le R) = \frac{1}{C_{n_1+n_2}^{n_1}} \sum_{i=2}^{R} F_i,$$

где величины под знаком суммы вычисляются так:

для четных индексов $F_i = 2C_{n_1-1}^{k-1} C_{n_2-1}^{k-1}$, где $k = i / 2$,

для нечетных индексов $F_i = C_{n_1-1}^{k-2} C_{n_2-1}^{k-1} + C_{n_1-1}^{k-1} C_{n_2-1}^{k-2}$, где $k = (i + 1) / 2$.

Отметим, что в формуле для вероятности не имеет значения, стоит в знаменателе число сочетаний из $n_1 + n_2$ по n_1 или по n_2, т. к. показано (Браунли), что

$$C_{n_1+n_2}^{n_1} = C_{n_1+n_2}^{n_2} = \frac{(n_1 + n_2)!}{n_1! n_2!}.$$

В асимптотической версии критерия (см. главу «Непараметрическая статистика») для вычисления *P*–значения используется нормальная аппроксимация.

Метод описан в монографии Браунли, диссертации Хешл (Heschl). Замечания о применении см. в книге Гаека с соавт., статье Камень с соавт.

Список использованной и рекомендуемой литературы

1. Agresti A. A survey of exact inference for contingency tables (with discussion) // Statistical Science, 1992, vol. 7, pp. 131–172.
2. Agresti A. An introduction to categorical data analysis. – New York, NY: John Wiley & Sons, 1996.
3. Agresti A., Mehta C.R., Patel N.R. Exact inference for contingency tables with ordered categories // Journal of the American Statistical Association, June 1990, vol. 85, no. 410, pp. 453–458.
4. Agresti A., Wackerly D., Boyett J.M. Exact conditional tests for cross–classifications: approximations of attained significance levels // Psychometrika, 1979, vol. 44, pp. 75–83.
5. Ahmad I.A. Modification of some goodness of fit statistics II: two–sample and symmetry testing // Sankhya: The Indian Journal of Statistics, 1996, vol. 58, ser. A, pt. 3, pp. 464–472.
6. Albers W., Bickel P.J., Van Zwet W.R. Asymptotic expansions for the power of distribution–free tests in the one–sample problem // The Annals of Statistics, 1976, vol. 4, pp. 108–156.
7. Anderson M.J. Permutation tests for univariate or multivariate analysis of variance and regression // Canadian Journal of Fisheries and Aquatic Sciences, March 2001, vol. 58, no. 3, pp. 629–636.
8. Arnold H.J. Permutation support for multivariate techniques // Biometrika, 1964, vol. 51, pp. 65–70.
9. Baker F.B., Collier R.O. Analysis of experimental designs by means of randomization: A Univac 1103 program // Behavioral Science, 1961, vol. 6, p. 369.
10. Barnard G.A. A new test for 2 x 2 tables // Nature, 1945, vol. 156, no. 177, pp. 783–784.
11. Barnard G.A. Must clinical trials be large? The interpretation of p–values and the combination of test results // Statistics in Medicine, 1990, vol. 9, pp. 601–614.
12. Barnard G.A. On alleged gains in power from lower p–values // Statistics in Medicine, 1989, vol. 8, pp. 1469–1477.
13. Barnard G.A. Sequential tests in industrial statistics // Journal of the Royal Statistical Society Supplement, 1946, vol. 8, pp. 1–26.
14. Barnard G.A. Significatce tests for 2 x 2 tables // Biometrika, 1947, vol. 34, pp. 123–138.
15. Barnard G.A. Statistical inference // Journal of the Royal Statistical Society: Series B (Statistical Methodology), 1949, vol. 11, pp. 115–149.
16. Barton D.E., David F.N. Randomization basis for multivariate tests // Bulletin of the International Statistical Institute, 1961, vol. 39, pp. 455–467.
17. Basu D. Randomization analysis of experimental data: The Fisher randomization test // Journal of the American Statistical Association, 1980, vol. 75, pp. 575–582.
18. Bennett B.M., Underwood R.E. On McNemar’s test for the 2 x 2 table and its power // Biometrics, June 1970, vol. 26, no. 2, pp. 339–343.
19. Berger R.L. More powerful tests from confidence interval p values // American Statistician, 1996, vol. 50, pp. 314–318.
20. Berger R.L., Sidik K. Exact unconditional tests for a 2 x 2 matched–pairs design // Statistical Methods in Medical Research, 2003, vol. 12, pp. 91–108.
21. Bishop Y.M.M., Fienberg S.E., Holland P.W. Discrete multivariate analysis: theory and practice. – Cambridge, MA: MIT Press, 1975.

22. Boschloo R.D. Raised conditional level of significance for 2 x 2 table when testing equality of probability // Statistica Neerlandica, 1970, vol.24, pp. 1–35.
23. Bower K.M. The two–sample t–test and randomization test // Six Sigma Forum – American Society for Quality, June 2003.
24. Bower K.M. When to use Fisher's exact test // ASQ Six Sigma Forum Magazine, August 2003, vol. 2, no. 4.
25. Box G.E.P., Anderson S.L. Permutation theory in the development of robust criteria and the study of departures from assumptions (with discussion) // Journal of the Royal Statistical Society: Series B (Statistical Methodology), 1955, vol. 17, pp. 1–34.
26. Box G.E.P., Hunter W.G., Hunter J.S. Statistics for experimenters: An introduction to design, data analysis, and model building. – New York, NY: John Wiley & Sons, 1978.
27. Bradley J.V. Distribution–free statistical tests. – Englewood Cliffs, NJ: Prentice–Hall, 1968.
28. Bross I.D.J. How to use ridit analysis // Biometrics, 1958, vol. 14, pp. 18–38.
29. Bross I.D.J. Taking a covariable into account // Journal of the American Statistical Association, 1964, vol. 59, pp. 725–736.
30. Buchan I.E. The development of a statistical computer software resource for medical research. Thesis for the degree of Doctor of Medicine. – Liverpool: University of Liverpool, 2000.
31. Chernick M.R. Friis R.H. Introductory biostatistics for the health sciences. Modern application including bootstrap. – New York, NY: John Wiley & Sons, 2003.
32. Corcoran C.D., Mehta C.R. Exact level and power of permutation, bootstrap and asymptotic tests of trend // Journal of Modern Applied Statistical Methods, 2002, vol. 1, pp. 42–51.
33. Cox D.R. A note on weighted randomization // Annals of Mathematical Statistics, 1956, vol. 27, pp. 1144–1150.
34. Deshpande J.V., Gore A.P., Shanubhogue A. Statistical analysis of nonnormal data. – New York, NY: John Wiley & Sons, 1995.
35. Diaconis P., Efron B. Computer–intensive methods in statistics // Scientific American, 1983, vol. 247, no. 5, pp. 96–129.
36. Diks C.G.H., Panchenko V. Nonparametric tests for serial independence based on quadratic forms // CeNDEF Working Paper no. 05–13, University of Amsterdam.
37. Draper D. Exchangeability and data analysis (with discussion) / D. Draper, J.S. Hodges, C.L. Mallows et al. // Journal of the Royal Statistical Society: Series A (Statistics in Society), 1993, vol. 156, 9–37.
38. Dufour J.–M., Khalaf L. Exact tests for contemporaneous correlation of disturbances in seemingly unrelated regressions // CIRANO Working Paper no. 2000s–16, Montreal, May 2000.
39. Dwass M. Modified randomization tests for nonparametric hypotheses // The Annals of Mathematical Statistics, 1957, vol. 28, pp. 181–187.
40. Dwyer A.J. Matchmaking and McNemar in the comparison of diagnostic modalities // Radiology, February 1991, vol. 178, no. 2, pp. 328–330.
41. Edgington E.S. Randomization tests. – New York, NY: Marcel Dekker, 1995.
42. Edgington E.S. Statistical inference and nonrandom samples // Psychological Bulletin, 1966, vol. 66, pp. 485–487.
43. Edgington E.S. Statistical inference: The distribution–free approach. – New York, NY: McGraw–Hill, 1969.
44. Efron B. Bootstrap methods: Another look at the jacknife // Annals of Statistics, 1979, vol. 7, pp. 1–26.
45. Efron B. The jacknife, the bootstrap and other resampling plans. – Philadelphia, PA: SIAM, 1982.

46. Efron B., Tibshirani R.J. An introduction to the bootstrap. – New York, NY: Chapman & Hall, 1993.
47. Ernst M.D. Permutation methods: A basis for exact inference // Statistical Science, 2004, vol. 19, no. 4, pp. 676–685.
48. Everitt B.S. The analysis of contingency tables. – New York, NY: Chapman & Hall, 1977.
49. Feinstein A.R. Principles of medical statistics. – New York, NY: Chapman & Hall / CRC, 2002.
50. Fisher R.A. A new test for 2 x 2 tables // Nature, 1945, vol. 156, p. 388.
51. Fisher R.A. Coefficient of racial likeness and the future of craniometry // Journal of the Royal Anthropological Society, 1936, vol. 66, pp. 57–63.
52. Fisher R.A. Statistical tests of agreement between observation and hypothesis // Economica, 1923, vol. 3, pp. 139–147.
53. Fisher R.A. The design of experiments. – Edinburgh: Oliver & Boyd, 1966.
54. Fleiss J.L. Statistical methods for rates and proportions. – New York, NY: John Wiley & Sons, 1981.
55. Freeman G.H., Halton J.H. Note on an exact treatment of contingency, goodness–of–fit, and other problems of significance // Biometrika, 1951, vol. 38, pp. 141–149.
56. Gabriel K.R., Hsu C.F. Evaluation of the power of rerandomization tests, with application to weather modification experiments // Journal of the American Statistical Association, December 1983, vol. 78, no. 384, pp. 766–775.
57. Gart J. Point and interval estimation of the common odds ratio in the combination of 2 x 2 tables with fixed marginals // Biometrika, 1970, vol. 57, pp. 471–475.
58. Geisser S. Significance testing for the 2 x 2 table // Bulletin of the International Statistical Institute, 52nd Session, Proceedings, Tome LVIII, Finland, 1999. Contributed Paper Meeting 7: Statistical tests.
59. Good P. Permutation tests: A practical guide to resampling methods for testing hypotheses. – New York, NY: Springer–Verlag, 2000.
60. Good P. Resampling methods: A practical guide to data analysis. – Boston, MA: Birkhauser, 2006.
61. Goodman L.A. Simple models for the analysis of association in cross–classifications having ordered categories // Journal of the American Statistical Association, 1979, vol. 74, pp. 537–552.
62. Green B.F. A practical interactive program for randomization tests of location // American Statistician, February 1977, vol. 31, no. 1, pp. 37–39.
63. Gridgeman N.T. The lady tasting tea, and allied topics // Journal of the American Statistical Association, 1959, vol. 54, pp. 776–783.
64. Guidance for data quality assessment. Practical methods for data analysis. EPA QA/G–9. – Washington, DC: United States Environmental Protection Agency, 2000.
65. Hall P. The bootstrap and Edgeworth expansion. – New York, NY: Springer–Verlag, 1992.
66. Heschl W.C. An investigation of the power of the Wald–Wolfowitz, two sample, runs Test. Master's thesis. – Monterey, CA: Naval Postgraduate School, 1972.
67. Hinkelmann K., Kempthorne O. Design and analysis of experiments, Vol. 1: Introduction to experimental design. – New York, NY: Wiley & Sons, 1994.
68. Hoeffding W. Combinatorial central limit theorem // Annals of Mathematical Statistics, 1951, vol. 22, pp. 556–558.
69. Hoeffding W. The large sample power of tests based on permutations of observations // The Annals of Mathematical Statistics, 1952, vol. 23, pp. 169–192.
70. Hoover D.R. Extending power and sample size approaches developed for McNemar's procedure to general sign tests // International Statistical Review, 2005, vol. 73, no. 1, pp.

103–110.
71. Hora S.C., Iman R.L. Asymptotic relative efficiencies of the rank–transformation procedure in randomized complete block designs // Journal of the American Statistical Association, 1988, vol. 83, pp. 462–470.
72. Hubert L.J. Assignment methods in combinatorial data analysis. – New York, NY: Marcel Dekker, 1987.
73. Jockel K.–H. Finite sample properties and asymptotic efficiency of Monte Carlo tests // Annals of Statistics, March 1986, vol. 14, no. 1, pp. 336–347.
74. Kaiser J. An exact and a Monte Carlo proposal to the Fisher–Pitman permutation tests for paired replicates and for independent samples // The Stata Journal, 2007, vol. 7, no. 3, pp. 402–412.
75. Kempthorne O. Design and analysis of experiments. – New York, NY: Wiley & Sons, 1952.
76. Kempthorne O., Doerfler T.E. The behavior of some significance tests under experimental randomization // Biometrika, 1969, vol. 56, pp. 231–247.
77. Khan H.A. A Visual Basic software for computing Fisher's exact probability // Journal of Statistical Software, 2003, vol. 8, no. 21.
78. Kopit J.S., Berger R.L. A more powerful exact test for a practical difference between binomial proportions // Proceedings of the Biopharmaceutical Section of the ASA, 1998, pp. 251–256.
79. Krauth J. Distribution–free statistics: An application oriented approach. – Amsterdam: Elsevier, 1988.
80. Kuiper N.H. Tests concerning random points on a circle // Proceedings of the Koninklijke Nederlandse Akademie van Wetenschappen, ser. A, 1962, vol. 63, pp. 38–47.
81. Lancaster H.O. Significance tests in discrete distributions // Journal of the American Statistical Association, 1961, vol. 56, pp. 223–234.
82. Le C.T. Introductory biostatistics. – Hoboken, New Jersey: John Wiley & Sons, 2003.
83. LePage R. Exploring the limits of bootstrap / Ed. by R. LePage, L. Billard. – New York, NY: Wiley & Sons, 1992.
84. Liddell F.D. Simplified exact analysis of case–referent studies: matched pairs; dichotomous exposure // Journal of Epidemiology and Community Health, March 1983, vol. 37, no.1, pp. 82–84.
85. Little R.J.A. Testing the equality of two independent binomial proportions // The American Statistician, 1989, vol. 43, pp. 283–288.
86. Ludbrook J. Statistical techniques for comparing measurers and methods of measurement: A critical review // Clinical and Experimental Pharmacology and Physiology, 2002, vol. 29, pp. 527–536.
87. Manly B.F.J. Randomization and Monte Carlo methods in biology. – London: Chapman & Hall, 1991.
88. Mantel N. The detection of disease clustering and a generalized regression approach // Cancer Research, 1967, vol. 27, pp. 209–220.
89. Mantel N., Haenszel W. Statistical aspects of the analysis of data from retrospective studies of disease // Journal of the National Cancer Institute, 1959, vol. 22, pp. 719–748.
90. Maritz J.S. Distribution free statistical methods. – London: Chapman & Hall, 1995.
91. Martin Andres A. On the validity condition of the chi–squared test in 2 x 2 tables / A. Martin Andres, M.J. Sanchez Quevedo, J.M. Tapia Garcia et al. // Sociedad de Estadistica e Investigacion Operativa Test, 2005, vol. 14, no. 1, pp. 99–128.
92. Martin Andres A., Tapia Garcia J.M. Optimal unconditional test in 2 x 2 multinomial trials // Computational Statistics & Data Analysis, 1999, vol. 31, pp. 311–321.
93. Maxwell A.E. Comparing the classification of subjects by two independent judges // British

Journal of Psychiatry, 1970, vol. 116, pp. 651–655.
94. May R.B., Masson M.E.J., Hunter M.A. Application of statistics in behavioral research. – New York, NY: Harper & Row, 1990.
95. McDonald L.L., Davis B.M., Milliken G.A. A nonrandomized unconditional test for comparing two proportions in 2 x 2 contingency tables // Technometrics, 1977, vol. 19, pp. 145–157.
96. Mehrotra D.V., Chan I.S.F., Berger R.L. A cautionary note on exact unconditional inference for a difference between two independent binomial proportions // Biometrics, 2003, vol. 59, pp. 441–450.
97. Mehta C.R. , Patel N.R., Gray R. Computing an exact confidence interval for the common odds ratio in several 2 x 2 contingency tables // Journal of the American Statistical Association, 1985, vol. 80, no. 392, pp. 969–973.
98. Mehta C.R., Patel N.R. A network algorithm for performing Fisher's exact test in r x c contingency tables // Journal of the American Statistical Association, 1983, vol. 78, no. 382, pp. 427–434.
99. Mehta C.R., Patel N.R. A network algorithm for the exact treatment of the 2 x K contingency table // Communications in Statistics: Simulation and Computation, 1980, vol. 9, pp. 649–664.
100. Mehta C.R., Patel N.R., Gray R. Correction: Computing an exact confidence interval for the common odds ratio in several 2 x 2 contingency tables // Journal of the American Statistical Association, 1986, vol. 81, no. 396, p. 1132.
101. Mewhort D.J.K. A comparison of the randomization test with the F test when error is skewed // Behavior Research Methods, August 2005, vol. 37, no. 3, pp. 426–435.
102. Mielke P.W., Berry K.J., Johnson E.S. Multiresponse permutation procedures for a priori classifications // Communications in Statistics: Theory and Methods, 1976, vol. 5, pp; 1409–1424.
103. Neyman J. First course in probability and statistics. – New York, NY: Holt, 1950.
104. Nichols T.E., Holmes A.P. Nonparametric permutation tests for functional neuroimaging: A primer with examples // Human Brain Mapping, 2001, vol. 15, pp. 1–25.
105. Noreen E. Computer–intensive methods for testing hypotheses. – New York, NY: Wiley & Sons, New York.
106. Oden A., Wedel H. Arguments for Fisher's permutation test // The Annals of Statistics, 1975, vol. 3, pp. 518–520.
107. Ogawa J. Effect of randomization on the analysis of a randomized block design // Annals of the Institute of Statistical Mathematics (Tokyo), 1961, vol. 13, pp. 105–117.
108. Pearson E.S. Some aspects of the problem of randomization // Biometrika, 1937, vol. 29, pp. 53–64.
109. Pitman E.J.G. Significance tests which may be applied to samples from any population. Part I. // Royal Statistical Society Supplement, 1937, vol. 4, pp. 119–130.
110. Pitman E.J.G. Significance tests which may be applied to samples from any population. Part II. The correlation coefficient test // Royal Statistical Society Supplement, 1937, vol. 4, pp. 225–232.
111. Pitman E.J.G. Significance tests which may be applied to samples from any population. Part III. The analysis of variance test // Biometrika, 1938, vol. 29, pp. 322–335.
112. Plackett R.L. Random permutations // Journal of the Royal Statistical Society: Series B (Statistical Methodology), 1968, vol. 30, pp. 517–534.
113. Reineke D.M., Baggett J., Elfessi A. A note on the effect of skewness, kurtosis, and shifting on one–sample t and sign tests // Journal of Statistics Education, 2003, vol. 11, no. 3.
114. Robinson J. A converse to a combinatorial central limit theorem // Annals of

Mathematical Statistics, 1972, vol. 43, pp. 2055–2057.
115. Rosenbaum P.R. Conditional permutation tests and the propensity score in observational studies // Journal of the American Statistical Association, 1984, vol. 79, pp. 565–574.
116. Scheffe H. Statistical inference in the non–parametric case // Annals of Mathematical Statistics, 1943, vol. 14, pp. 305–332.
117. Scheffler E. Einfurung in die Praxis der statistischen Versuchsplannung. – Leipzig: VEB Deutscher Verlag fur Grundstoffindustrie, 1973.
118. Senchaudhuri P., Mehta C.R., Patel N.R. Estimating exact P values by the method of control variates or Monte Carlo rescue // Journal of the American Statistical Association, 1995, vol. 90, no. 430, pp. 640–648.
119. Siegel S., Castellan N.J. Nonparametric Statistics for the Behavioral Sciences. – New York, NY: McGraw Hill, 1988.
120. Simon J.L. Basic research methods in social science. – New York, NY: Random House, 1969.
121. Simon J.L. Resampling: The new statistics. – Arlington, VA: Resampling Stats Inc., 1997.
122. Simon J.L., Burstein P. Basic research methods in social science. – New York, NY: Random House, 1985.
123. Sokal R.R., Rohlf F.J. Biometry: the principles and practice of statistics in biological research. – New York, NY: W.H. Freeman, 1995.
124. Sprent P. Applied nonparametric statistical methods. – London: Chapman & Hall / CRC, 1993.
125. Sprent P., Smeeton N.C. Applied nonparametric statistical methods. – Boca Raton, FL: Chapman & Hall / CRC, 2001.
126. Streitberg B., Rohmel J. Exact nonparametrics in APL // International Conference on APL archive. Proceedings of the international conference on APL, Finland, 1984, pp. 313–325.
127. Suissa S., Shuster J. Exact unconditional sample sizes for the 2 x 2 binomial trial // Journal of the Royal Statistical Society: Series A (Statistics in Society), 1985, vol. 148, pp. 317–327.
128. Tukey J.W., Brillinger D.R., Jones L.V. Management of weather resources. Vol. II. The role of statistics in weather resources management. – Washington, DC: Department of Commerce, US Government Printing Office, 1978.
129. Van Belle G. Biostatistics: A methodology for the health sciences // G. van Belle, L.D. Fisher, P.J. Heagerty et al. – New York, NY: John Wiley & Sons, 2003.
130. Wald A., Wolfowitz J. Statistical tests based on permutations of the observations // Annals of Mathematical Statistics, 1944, vol. 15, pp. 358–372.
131. Welch W.J. Construction of permutation tests // Journal of American Statistical Association, 1990, vol. 85, pp. 693–698.
132. Westfall P.H., Young S.S. Resampling–based multiple testing: Examples and methods for p–value adjustment. – New York, NY: Wiley & Sons, 1993.
133. Yates F. Tests of significance for 2 x 2 contingency tables (with discussion) // Journal of the Royal Statistical Society: Series A (Statistics in Society), 1984, vol. 147, pp. 426–463.
134. Zelen M. The analysis of several 2 x 2 contingency tables // Biometrika, 1971, vol. 58, pp. 129–137.
135. Браунли К.А. Статистическая теория и методология в науке и технике. – М.: Наука, 1977.
136. Вержбицкий В.М. Численные методы (линейная алгебра и нелинейные

уравнения): Учебное пособие для вузов. – М.: Высшая школа, 2000.
137. Гаек Я., Шидак З. Теория ранговых критериев. – М.: Наука, 1971.
138. Гайдышев И. Анализ и обработка данных: специальный справочник. – СПб: Питер, 2001.
139. Камень Ю.Э., Камень Я.Э., Орлов А.И. Реальные и номинальные уровни значимости в задачах проверки статистических гипотез // Заводская лаборатория. Диагностика материалов, 1986, т. 52, № 12, с. 55–57.
140. Кендалл М., Стьюарт А. Статистические выводы и связи. – М.: Наука. 1973.
141. Колмогоров А.Н. Комбинаторные основания теории информации и исчисления вероятностей // Успехи математических наук, 1983, т. 38, вып. 4 (232), с. 27–36.
142. Колмогоров А.Н. Теория информации и теория алгоритмов. – М.: Наука, 1987.
143. Кориков А.М. Математические методы планирования эксперимента. – Томск: ТГУ, 1973.
144. Кулаичев А.П. Методы и средства анализа данных в среде Windows®. STADIA. – М.: Информатика и компьютеры, 1999.
145. Леман Э. Проверка статистических гипотез. – М.: Наука, 1979.
146. Ллойд Э. Справочник по прикладной статистике. В 2–х т. Т. 1 / Под ред. Э. Ллойда, У. Ледермана, Ю.Н. Тюрина. – М.: Финансы и статистика, 1989.
147. Монтгомери Д.К. Планирование эксперимента и анализ данных. – Л.: Судостроение, 1980.
148. Мюллер П., Нойман П., Шторм Р. Таблицы по математической статистике. – М.: Финансы и статистика, 1982.
149. Налимов В.В. Теория эксперимента. – М.: Наука, 1971.
150. Новиков Ф.А. Дискретная математика для программистов. Учебник для вузов. – СПб.: Питер, 2005.
151. Прохоров Ю.В. Вероятность и математическая статистика: Энциклопедия / Гл. ред. Ю.В. Прохоров. – М.: Научное издательство «Большая Российская энциклопедия», 1999.
152. Рождественский А.В., Чеботарев А.И. Статистические методы в гидрологии. – Л.: Гидрометеоиздат, 1974.
153. Рунион Р. Справочник по непараметрической статистике. – М.: Финансы и статистика, 1982.
154. Тюрин Ю.Н., Макаров А.А. Статистический анализ данных на компьютере. – М.: ИНФРА–М, 1999.
155. Уилкс С. Математическая статистика. – М.: Наука, 1967.
156. Фишер Р.А. Статистические методы для исследователей. – М.: Госстатиздат, 1958.
157. Флейс Дж. Статистические методы для изучения таблиц долей и пропорций. – М.: Финансы и статистика, 1989.
158. Христофоров А.В. Теория вероятностей и математическая статистика. – М.: Издательство Московского университета, 1988.

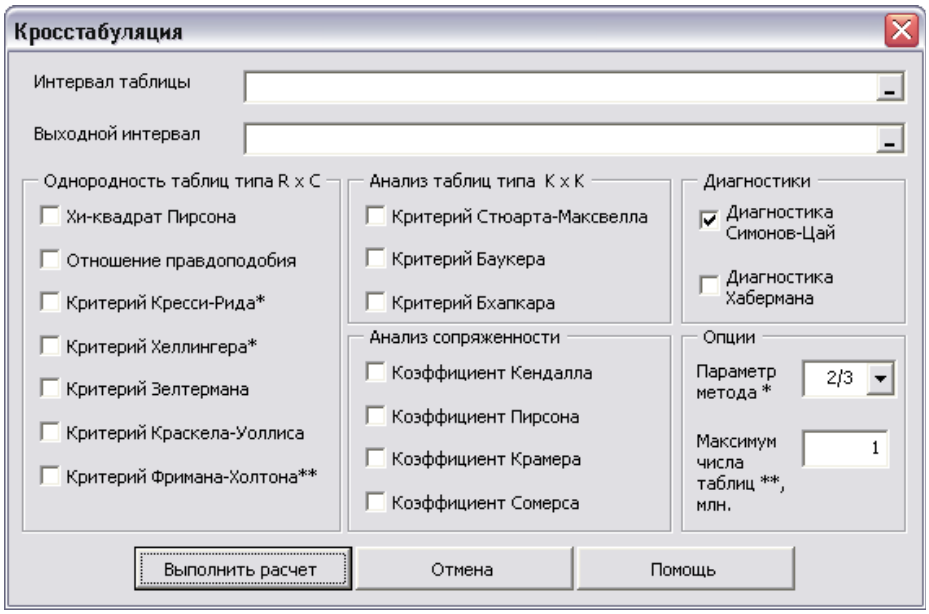
Глава 6. Кросстабуляция

6.1. Введение

Предлагаются методы анализа однородности и сопряженности (связи типа корреляции) в таблицах сопряженности, полученных на основе выборок, измеренных в номинальной шкале.

6.2. Работа с программным обеспечением

Выберите из меню программы пункт AtteStat | Кросстабуляция. На экране появится диалоговое окно, изображенное на рисунке:



Затем проделайте следующие шаги:

- Выберите или введите интервал таблицы сопряженности признаков. О порядке заполнения и требованиях к таблице сопряженности см. пояснения в теоретическом разделе.
- Выберите или введите выходной интервал для выдачи результатов расчета. Начиная с первой ячейки выходного интервала (следовательно, можно указать только одну ячейку, т.к. остальные ячейки интервала игнорируются), будут выведены вычисленные показатели.
- Выберите один или несколько методов анализа таблицы сопряженности.
- Для некоторых методов, отмеченных знаком *, возможно задание дополнительных параметров. Выберите значение дополнительного параметра из предлагаемого списка.
- Нажмите кнопку «Выполнить расчет».

После выполнения вычислений будет, начиная с первой ячейки выходного интервала, выведено название метода и результаты расчета: статистика критерия, одностороннее P – значение. Интерпретация полученных результатов статистических расчетов подробно рассмотрена в теоретическом разделе.
При ошибках, вызванных неверными действиями пользователя при вводе исходных данных для расчета, выдаются сообщения об ошибках.

6.2.1. Сообщения об ошибках

При ошибках ввода исходных данных для расчета могут выдаваться диагностические сообщения следующих типов:

Ошибка	Комментарий
Пустая ячейка в области данных.	Проверьте исходные данные и заполните все ячейки, отмеченные Вами как входной интервал. Для избежания ошибок расчета, вызванных

Ошибка	Комментарий
	разногласиями, трактовать ли пустую ячейку как нуль, программа требует заполнения всех ячеек. Если в ячейке не должно быть данных по физической природе исследуемого процесса, введите в данную ячейку нуль. Следует, однако, помнить, что ряд методов данного программного обеспечения, в частности, все методы, основанные на хи–квадрат, не работает с нулевыми значениями в исходных данных.
Нечисловой тип данных.	Проверьте типы ячеек входного интервала. Тип может быть только числовым. Проще всего выделить интервал ячеек и явно определить их тип как числовой стандартными средствами.
Мало данных.	Для расчета необходимо выбрать интервал таблицы сопряженности, содержащий хотя бы четыре ячейки таблицы сопряженности с заполненными числовыми значениями.
Не определена область данных.	Вы не выбрали или неверно ввели входной интервал. Лучшим способом избежать ошибки является не ввод, а выделение интервала стандартным образом, т. е. протаскиванием курсора.
Не определена область вывода.	Вы не выбрали или неверно ввели выходной интервал. Лучшим способом избежать ошибки является не ввод, а выделение интервала стандартным образом, т. е. протаскиванием курсора.
Ошибочные данные.	Таблица сопряженности содержит ошибочные данные. Таблица сопряженности должна содержать только неотрицательные целые числа, т.к. в клетки таблицы помещаются количества вариант, обладающих данными признаками.

6.3. Теоретическое обоснование

Кросстабуляцией (cross–tabulation, analysis of cross–tabulated data) называют анализ двухвходовых (двумерных) таблиц сопряженности (таблиц смежности, contingence tables) . Таблицы сопряженности возникают при анализе признаков, измеренных в номинальной шкале либо в более высоких шкалах, преобразованных к номинальной шкале.
Специфическими исходными данными для рассматриваемого программного обеспечения служат не первичные выборки, а уже построенные на их основе таблицы сопряженности. Представление таблиц сопряженности подробно описано главе «Введение». Методами кросстабуляции исследуется статистическая зависимость не выборок, как, например, в методах проверки гипотез, а признаков.
Двухвходовые таблицы могут быть проанализированы с помощью настоящего программного обеспечения. Одни методы предназначены для исследования однородности, это:
анализ однородности (согласия) в таблицах типа r x c:

- критерий Кресси–Рида,
- критерий Хеллингера,
- критерий хи–квадрат,
- критерий отношения правдоподобия,
- критерий Зелтермана,
- критерий Фримана–Холтона.

анализ однородности (согласия) и симметрии в таблицах типа k x k:

- критерий Стюарта–Максвелла,
- критерий Баукера,

- критерий Бхапкара.

Для исследования сопряженности признаков (связи типа корреляции, не путать с корреляцией, которая для номинальных признаков, отражением которых являются таблицы сопряженности, не определена), предназначены специальные методы, как–то:

- коэффициент Кендалла,
- коэффициент Крамера,
- коэффициент Сомерса,
- коэффициент сопряженности Пирсона.

Дисперсионный анализ выборок, представленных таблицами типа *r* x *c*, может быть выполнен методами:

- критерий Краскела–Уоллиса.

С другой стороны, данные методы отражают два подхода к решению проблемы:

- Критерии первой группы включают в себя, наряду с классическим критерием хи–квадрат, основанные на хи–квадрат методы: коэффициенты Крамера и сопряженности Пирсона.
- Непараметрические ранговые методы включают: коэффициенты Кендалла (его одноименный аналог для порядковых выборок см. в главе «Корреляционный анализ») и Сомерса, критерий Краскела–Уоллиса.

Отметим, что критерии, основанные на хи–квадрат, с одной стороны, и коэффициенты, на хи–квадрат не основанные (например, коэффициент Кендалла), могут при вычислении давать различные результаты. Это вызвано тем, что критерии, основанные на хи–квадрат, нечувствительны к упорядочению строк и столбцов таблицы сопряженности. Именно поэтому данные методы в других образцах программного обеспечения по анализу данных могут быть сгруппированы иначе, чем это сделано в настоящем программном обеспечении.
Для исследования сопряженности признаков также предназначены не рассмотренные здесь мера τ_c Стьюарта, коэффициент ранговой корреляции r_S Спирмэна (его одноименный аналог для порядковых выборок см. в главе «Корреляционный анализ»), ряд других методов. Все эти методы представлены в монографии Афифи с соавт., включая нормальные аппроксимации, позволяющие использовать данные методы для проверки значимости связи.
Коэффициент *G* Гудмана–Кендалла подробно описан в книгах Кулаичева. Реализация расчета данных коэффициентов на основе уже запрограммированных в программе методов не представляет совершенно никакой сложности и, по возможности, будет представлена в будущих версиях программы.
Другие типы тестов для таблиц сопряженности, например, информационный критерий Кульбака–Лейблера (Kullback–Leibler information criterion), логлинейный анализ, также будут представлены в будущих версиях программы.
Укажем на особенность тестов, основанных на распределении хи–квадрат. Распределения статистик данных критериев лишь приблизительно соответствуют хи–квадрат:

- Согласно Кокрену (см. Сергиенко с соавт., с. 79), если для таблицы 2 x 2 сумма таблицы < 20 или сумма таблицы от 20 до 40, но при этом в одной из ячеек ожидаемая частота < 5, то следует использовать не критерий хи–квадрат, а точный метод Фишера (см. главу «Точные критерии»).
- Согласно Аптону (глава 3), приближение работает достаточно хорошо, пока ожидаемые частоты в ячейках таблицы сопряженности не опустятся примерно до трех.

Объективными критериями допустимости аппроксимации хи–квадрат являются так называемые диагностики: Симонов–Цай, Хабермана, Мудхолкара–Хадсона и другие.
В практике иногда возникает необходимость проверки однородности данных, представленных в виде строк таблицы сопряженности, методами дисперсионного анализа.

Как известно, построение таблицы сопряженности из количественных данных понижает шкалу. При этом восстановление исходных данных по имеющейся таблице невозможно. Однако в случае, если исходные данные были порядковыми, понижения шкалы не происходит (хотя исходные данные также восстановить нельзя).
Для проведения дисперсионного анализа исходных порядковых данных и данных, восстановленных по таблице сопряженности, могут применяться одни и те же методы непараметрического дисперсионного анализа. Результат непараметрического дисперсионного анализа восстановленных с точностью до коэффициентов данных будет совпадать с результатами анализа исходных порядковых данных. Данная возможность обеспечивается процедурой ранжирования, применяемых в данных методах.
На следующем примере показана возможность восстановления порядковой выборки из строки таблицы сопряженности. Пусть имеется таблица результатов лечения для группы пациентов.

	Результат лечения			
	Плохо	Удовлетворительно	Хорошо	
Группа 1	2	5	10	17
Группа 2	5	4	4	13

Не имеет значения для непараметрического дисперсионного анализа, какие величины имели те или иные варианты до построения таблицы сопряженности, однако их соотношение должно соблюдаться. Поэтому для определенности можно выбрать кодировку: плохо – 1, удовлетворительно – 2, хорошо – 3. Таким образом, в показанном примере можно восстановить исходные порядковые выборки:

1. Группа 1 (численность 17): 1 1 2 2 2 2 2 3 3 3 3 3 3 3 3 3 3,
2. Группа 2 (численность 13): 1 1 1 1 1 2 2 2 2 3 3 3 3.

Данные выборки могут быть проанализированы любыми непараметрическими методами дисперсионного анализа. При восстановлении исходных порядковых данных из таблиц сопряженности неизбежно появление связок, поэтому применяемые методы, как вариант метода Краскела–Уоллиса, используются только с учетом связок.
Представленные в данном программном обеспечении параметры сопряженности (связи типа корреляции для номинальных признаков) могут применяться в качестве показателей статистической значимости связи между признаками. Данную связь допустимо интерпретировать как корреляционную, но нельзя называть корреляцией, т. к. для номинальных признаков корреляция не определена. Поэтому лучше использовать термины «сопряженность» или «связь типа корреляции». Подробнее о корреляции см. «Корреляционный анализ».
Афифи с соавт. приводят общую формулу расчета показателей, для которых неизвестно или затруднительно вычисление критических значений. Используется тот факт, что статистика

$$z = \frac{X}{\sqrt{DX}},$$

где X – статистика критерия,
DX – дисперсия,
асимптотически имеет стандартное нормальное распределение $N(0,1)$.

6.3.1. Критерий Кресси–Рида

Критерий Кресси–Рида (power–divergence family Cressie–Read) является наиболее общим методом анализа однородности таблиц сопряженности. Вычисление критерия производится

по формуле

$$CR(\lambda) = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{2}{\lambda(1+\lambda)} A_{ij}\left[\left(\frac{A_{ij}}{E_{ij}}\right)^{\lambda} - 1\right],$$

где A_{ij}, $i = 1,2,...,r$; $j = 1,2,...,c$ – заданные частоты таблицы сопряженности,
E_{ij}, $i = 1,2,...,r$; $j = 1,2,...,c$ – ожидаемые частоты таблицы сопряженности,
r – число строк таблицы сопряженности,
c – число столбцов таблицы сопряженности,
λ – параметр, обычно равный 2/3.
По условиям вычисления статистики критерия, при $A_{ij} = 0$, во избежание численных проблем, условились считать, что $A_{ij}[\ldots] = 0$.
Программа позволяет производить выбор параметра λ из нескольких предлагаемых вариантов.
Ожидаемые частоты вычисляются по формуле

$$E_{ij} = \frac{n_{i.} n_{.j}}{n}, i = 1,2,...,r; j = 1,2,...,c,$$

где $n_{i.} = \sum_{k=1}^{c} A_{ik}, i = 1,2,...,r,$ – суммы строк таблицы сопряженности,

$n_{.j} = \sum_{k=1}^{r} A_{kj}, j = 1,2,...,c,$ – суммы столбцов таблицы сопряженности,

$n = \sum_{k=1}^{r}\sum_{l=1}^{c} A_{kl}$ – общее число наблюдений.

Статистика критерия подчиняется распределению χ^2 с числом степеней свободы $(r - 1)(c - 1)$.
Представленный критерий является основой семейства тестов, которые получаются в результате того или иного выбора параметра λ. Например, можно получить при значениях параметра:

- $\lambda = 1$ – критерий хи–квадрат ,
- $\lambda = 0$ – критерий отношения правдоподобия.

Находят применение и другие значения параметра, в том числе отрицательные.

См. работы фон Давье (Von Davier), Браво (Bravo), Базу (Basu) с соавт.

6.3.2. Критерий Хеллингера

Критерий Хеллингера (blended weight Hellinger) является методом анализа однородности в таблицах сопряженности. Вычисление критерия производится по формуле

$$BWH(\alpha) = \sum_{i=1}^{r}\sum_{j=1}^{c}\left(\frac{A_{ij} - E_{ij}}{\alpha\sqrt{A_{ij}} + (1-\alpha)\sqrt{E_{ij}}}\right)^2,$$

где A_{ij}, $i = 1,2,...,r$; $j = 1,2,...,c$ – заданные частоты таблицы сопряженности,
E_{ij}, $i = 1,2,...,r$; $j = 1,2,...,c$ – ожидаемые частоты таблицы сопряженности,
r – число строк таблицы сопряженности,
c – число столбцов таблицы сопряженности,
α – параметр, обычно равный 1/2 или 1/9.
Программа позволяет производить выбор параметра α из нескольких предлагаемых вариантов.

Ожидаемые частоты вычисляются по формуле

$$E_{ij} = \frac{n_{i.} n_{.j}}{n}, i = 1,2,...,r; j = 1,2,...,c,$$

где $n_{i.} = \sum_{k=1}^{c} A_{ik}, i = 1,2,...,r,$ – суммы строк таблицы сопряженности,

$n_{.j} = \sum_{k=1}^{r} A_{kj}, j = 1,2,...,c,$ – суммы столбцов таблицы сопряженности,

$n = \sum_{k=1}^{r} \sum_{l=1}^{c} A_{kl}$ – общее число наблюдений.

Статистика критерия подчиняется распределению χ^2 с числом степеней свободы $(r - 1)(c - 1)$.

См. работы Бхаттачарья (Bhattacharya) с соавт., Парк (Park) с соавт., Базу (Basu) и Рэй (Ray) с соавт.

6.3.3. Критерий хи–квадрат

Классический критерий хи–квадрат (критерий хи–квадрат Пирсона, Pearson chi–square test, Pearson's X^2 test) является стандартным для анализа таблиц сопряженности. Вычисление критерия производится по формуле

$$X^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{\left(A_{ij} - E_{ij}\right)^2}{E_{ij}},$$

где A_{ij}, $i = 1,2,...,r$; $j = 1,2,...,c$ – заданные частоты таблицы сопряженности,
E_{ij}, $i = 1,2,...,r$; $j = 1,2,...,c$ – ожидаемые частоты таблицы сопряженности,
r – число строк таблицы сопряженности,
c – число столбцов таблицы сопряженности.
Ожидаемые частоты вычисляются по формуле

$$E_{ij} = \frac{n_{i.} n_{.j}}{n}, i = 1,2,...,r; j = 1,2,...,c,$$

где $n_{i.} = \sum_{k=1}^{c} A_{ik}, i = 1,2,...,r,$ – суммы строк таблицы сопряженности,

$n_{.j} = \sum_{k=1}^{r} A_{kj}, j = 1,2,...,c,$ – суммы столбцов таблицы сопряженности,

$n = \sum_{k=1}^{r} \sum_{l=1}^{c} A_{kl}$ – общее число наблюдений.

Для больших выборок статистика критерия подчиняется распределению χ^2 с числом степеней свободы $(r - 1)(c - 1)$.

См. работы Мехта (Mehta) с соавт., Стаффорда (Stafford). В работе Мудхолкара (Mudholkar) и Хатсона (Hutson) проведен анализ возможности аппроксимации распределения статистики критерия, введены т. н. диагностики, которые позволяют судить о правомерности данной процедуры.
Модификации критерия хи-квадрат для анализа многовходовых таблиц сопряженности см. в монографии Аптона, статьях Кастенбаума (Kastenbaum) с соавт., Гудмана (Goodman).
Модификация критерия хи-квадрат для анализа таблиц сопряженности типа 2 x k носит

наименование критерия тренда Кокрена-Армитейджа (Cochran–Armitage test for trend) и представлена Агрести (Agresti, 2002). Особенностью критерия является введение т. н. весовых функций, позволяющих формулировать различные нулевые гипотезы в рамках представленной таблицы.

6.3.4. Критерий отношения правдоподобия

Классический критерий отношения правдоподобия (likelihood ratio test, G^2 test) является стандартным методом исследования однородности таблиц сопряженности. Вычисление критерия производится по формуле

$$G^2 = 2\sum_{i=1}^{r}\sum_{j=1}^{c} A_{ij} \log \frac{A_{ij}}{E_{ij}},$$

где A_{ij}, $i = 1,2,...,r$; $j = 1,2,...,c$ – заданные частоты таблицы сопряженности,
E_{ij}, $i = 1,2,...,r$; $j = 1,2,...,c$ – ожидаемые частоты таблицы сопряженности,
r – число строк таблицы сопряженности,
c – число столбцов таблицы сопряженности.
По условиям вычисления статистики критерия, при $A_{ij} = 0$, во избежание численных проблем, условились считать, что $A_{ij} \log\ldots = 0$.
Ожидаемые частоты вычисляются по формуле

$$E_{ij} = \frac{n_{i.} n_{.j}}{n}, i = 1,2,...,r; j = 1,2,...,c,$$

где $n_{i.} = \sum_{k=1}^{c} A_{ik}, i = 1,2,...,r,$ – суммы строк таблицы сопряженности,

$n_{.j} = \sum_{k=1}^{r} A_{kj}, j = 1,2,...,c,$ – суммы столбцов таблицы сопряженности,

$n = \sum_{k=1}^{r}\sum_{l=1}^{c} A_{kl}$ – общее число наблюдений.

Статистика критерия подчиняется распределению χ^2 с числом степеней свободы $(r - 1)(c - 1)$.

Описание см. в работах Мехта (Mehta) с соавт.

6.3.5. Критерий Зелтермана

Статистика критерия Зелтермана (Zelterman's statistic) для исследования однородности таблиц сопряженности. Вычисление критерия производится по формуле

$$D_z^2 = X^2 - \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{A_{ij}}{E_{ij}} + rc,$$

где X^2 – статистика критерия хи–квадрат,
A_{ij}, $i = 1,2,...,r$; $j = 1,2,...,c$ – заданные частоты таблицы сопряженности,
E_{ij}, $i = 1,2,...,r$; $j = 1,2,...,c$ – ожидаемые частоты таблицы сопряженности,
r – число строк таблицы сопряженности,
c – число столбцов таблицы сопряженности.
Ожидаемые частоты вычисляются по формуле

$$E_{ij} = \frac{n_{i.} n_{.j}}{n}, i = 1,2,...,r; j = 1,2,...,c,$$

$n_{i.} = \sum_{k=1}^{c} A_{ik}, i = 1,2,...,r,$
где – суммы строк таблицы сопряженности,

$n_{.j} = \sum_{k=1}^{r} A_{kj}, j = 1,2,...,c,$ – суммы столбцов таблицы сопряженности,

$n = \sum_{k=1}^{r}\sum_{l=1}^{c} A_{kl}$ – общее число наблюдений.

Статистика критерия подчиняется распределению χ^2 с числом степеней свободы $(r-1)(c-1)$.

См. материалы Лаззаротто (Lazzarotto) с соавт.

6.3.6. Критерий Фримана–Холтона

Критерий Фримана–Холтона (Фишера–Фримана–Холтона, Fisher–Freeman–Halton test) предназначен для проверки однородности таблицы сопряженности. Критерий является расширением точного метода Фишера.

Пусть X_0 – заданная таблица сопряженности, а X – вариант заполнения таблицы сопряженности при условии сохранения сумм строк и сумм столбцов заданной таблицы (маргинальных сумм). Тогда достигнутый уровень значимости критерия Фримана–Холтона будет вычисляться как сумма вероятностей всех таблиц X, таких, что $P(X) < P(X_0)$, иначе

$$p = \sum_{P(X)<P(X_0)} P(X).$$

Вероятности таблиц сопряженности $P(X)$ вычисляются по формуле вероятности гипергеометрического распределения

$$P(X) = \frac{1}{N!}\prod_{i=1}^{r} R_i! \prod_{j=1}^{c}\left[C_j! \Big/ \prod_{i=1}^{r} x_{ij}! \right],$$

где N – численность заданной таблицы сопряженности,
R_i, i =1,2,...,r – суммы строк заданной таблицы,
C_j, j =1,2,...,c – суммы столбцов заданной таблицы,
x_{ij}, i = 1,2,...,r; j = 1,2,...,c – частоты, составляющие таблицу сопряженности,
r – число строк таблицы,
c – число столбцов таблицы.

Для информации программой выдается оценка требуемого числа генерируемых таблиц. Знание примерного объема вычислений позволяет прогнозировать время, затраченное на расчет. Для этого применяется алгоритм, разработанный Гейлом (Gail) и Мантелем (Mantel), согласно которому оценка числа генерируемых таблиц производится по приближенной формуле

$$n \approx \left(\frac{c-1}{2\pi\sigma^2 c} \right)^{(c-1)/2} \sqrt{c} e^{-Q/2} \prod_{i=1}^{r} C_{R_i+c-1}^{c-1},$$

$\sigma^2 = \frac{c-1}{(c+1)c^2}\sum_{i=1}^{r} R_i(R_i + c)$
где – дисперсия,

$Q = \frac{c-1}{\sigma^2 c}\left(\sum_{j=1}^{c} C_j^2 - \frac{N^2}{c} \right)$ – параметр.

По условиям алгоритма, если $c > r$, для вычислений оценки числа таблиц исходная таблица сопряженности [автоматически] транспонируется.

В программе реализован расчет критерия методом Монте–Карло (генерируется заданное число таблиц). Генерация таблиц осуществляется по алгоритму Пэйтфилда (Patefield). Результатом расчета является приближенное *P*–значение, получающееся как отношение числа таблиц, удовлетворяющего показанному выше условию, к общему числу сгенерированных таблиц. По умолчанию число генерируемых таблиц равно 1 миллиону. Этого достаточно для многих задач и не является трудоемким в предложенной реализации (рассчитывается практически мгновенно). Если это число окажется равным или меньшим числа, примерно оцениваемого по алгоритму Гейла–Мантеля, следует увеличить число генерируемых таблиц как минимум до оцениваемого по алгоритму Гейла–Мантеля, затем повторить расчет.
Если пользователем указано максимальное число генерируемых таблиц, не адекватное быстродействию компьютера, преждевременный выход из программы возможен только аварийным ее завершением средствами операционной системы.

Алгоритм Пэйтфилда был модифицирован с тем, чтобы использовать более качественные псевдослучайные числа. Использован алгоритм, представленный в работах Лекюйе (L'Ecuyer), Фокс (Fox), Брэтли (Bratley) с соавт. Описание критерия см. в работах Мехта (Mehta) с соавт. Методы решения предложены Халворсеном (Halvorsen), Борковым (Borkowf), Сандерсом (Saunders), Бойеттом (Boyett).

6.3.7. Критерий Стюарта–Максвелла

Критерий однородности Стюарта–Максвелла (Stuart–Maxwell test) является расширением критерия Мак–Немара (см. главу «Непараметрическая статистика») для анализа таблиц сопряженности типа *k* x *k*. Вычисление критерия производится по формуле
$X^2 = D'S^{-1}D$,
где D – вектор–столбец, составленный из величин $d_i = n_{i.} - n_{.i}$, $i = 1,2,...,k-1$,
S – квадратная матрица порядка $k-1$, составленная из величин

$$s_{ij} = \begin{cases} -(A_{ij} + A_{ji}), i \neq j, \\ n_{i.} + n_{.i} - 2A_{ii}, i = j, \end{cases}$$

где k – число строк и столбцов таблицы сопряженности,
A_{ij}, $i = 1,2,...,k$; $j = 1,2,...,k$ – заданные частоты таблицы сопряженности,

$$n_{i.} = \sum_{j=1}^{k} A_{ij}, i = 1,2,...,k,$$

– суммы строк таблицы сопряженности,

$$n_{.i} = \sum_{j=1}^{k} A_{ji}, i = 1,2,...,k,$$

– суммы столбцов таблицы сопряженности.
Статистика критерия подчиняется распределению χ^2 с числом степеней свободы $k-1$.

См. статьи Максвелла (Maxwell), Стюарта (Stuart).

6.3.8. Критерий Баукера

Критерий симметрии Баукера (Bowker test) является расширением критерия Мак–Немара (см. главу «Непараметрическая статистика») для анализа таблиц сопряженности типа *k* x *k*. Вычисление критерия производится по формуле

$$X^2 = \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} \frac{(A_{ij} - A_{ji})^2}{A_{ij} + A_{ji}},$$

где k – число строк и столбцов таблицы сопряженности,
A_{ij}, $i = 1,2,...,k$; $j = 1,2,...,k$ – заданные частоты таблицы сопряженности.
Статистика критерия подчиняется распределению χ^2 с числом степеней свободы $k(k-1)/2$.

См. статью Баукера (Bowker), отчет Льюиса (Lewis) с соавт., отчет Крампе (Krampe) с соавт.

6.3.9. Критерий Бхапкара

Критерий однородности Бхапкара (Bhapkar's test) предназначен для анализа таблиц сопряженности типа k x k. Вычисление критерия производится по формуле
$W = nD'S^{-1}D$,

$n = \sum_{i=1}^{k}\sum_{j=1}^{k} A_{ij}$
где – сумма таблицы сопряженности,
A_{ij}, $i = 1,2,...,k$; $j = 1,2,...,k$ – заданные частоты таблицы сопряженности,
k – число строк и столбцов таблицы сопряженности,
D – вектор–столбец, составленный из величин $d_i = n_{i.} - n_{.i}$, $i = 1,2,...,k-1$,
S – квадратная матрица порядка $k-1$, составленная из величин

$$s_{ij} = \begin{cases} -(n_{ij} + n_{ji}) - (n_{i.} - n_{.i})(n_{j.} - n_{.j}), i \neq j, \\ n_{i.} + n_{.i} - 2n_{ii} - (n_{i.} - n_{.i})^2, i = j, \end{cases}$$

где $n_{ij} = A_{ij} / n$, $i = 1,2,...,k$; $j = 1,2,...,k$ – частости,

$n_{i.} = \sum_{j=1}^{k} n_{ij}, i = 1,2,...,k,$ – суммы строк таблицы частостей,

$n_{.i} = \sum_{j=1}^{k} n_{ji}, i = 1,2,...,k,$ – суммы столбцов таблицы частостей.

Статистика критерия подчиняется распределению χ^2 с числом степеней свободы $k-1$.

См. статьи Бхапкара (Bhapkar), Бхапкара с соавт., отчет Льюиса (Lewis) с соавт.

6.3.10. Коэффициент Кендалла

Коэффициент τ_b Кендалла (коэффициент Кендэла, Kendall's τ_b) вычисляется по формуле, подробно рассмотренной Афифи с соавт. и удобной для численных расчетов:

$$\tau_b = S \Big/ \sqrt{\left[\frac{1}{2}n(n-1) - T_1\right]\left[\frac{1}{2}n(n-1) - T_2\right]},$$

где $S = P - Q$,

$P = \sum_{i=1}^{r}\sum_{j=1}^{c} A_{ij}\left(\sum_{k>i}\sum_{l>j} A_{kl}\right)$ – число пар объектов с взаимно возрастающими переменными,

$Q = \sum_{i=1}^{r}\sum_{j=1}^{c} A_{ij}\left(\sum_{k>i}\sum_{l<j} A_{kl}\right)$ – число пар объектов с взаимно убывающими переменными,

A_{ij}, $i = 1,2,...,r$; $j = 1,2,...,c$ – заданные частоты таблицы сопряженности,
r – число строк таблицы сопряженности,
c – число столбцов таблицы сопряженности,

$n = \sum_{k=1}^{r}\sum_{l=1}^{c} A_{kl}$ – общее число наблюдений,

$T_1 = \frac{1}{2}\sum_{i=1}^{r} n_{i.}(n_{i.} - 1)$ – число пар объектов с взаимно равными значениями по одной переменной,

$n_{i.} = \sum_{k=1}^{c} A_{ik}, i = 1,2,...,r,$ – суммы строк таблицы сопряженности,

$T_2 = \frac{1}{2}\sum_{j=1}^{c} n_{.j}(n_{.j} - 1)$ – число пар объектов с взаимно равными значениями по другой переменной,

$n_{.j} = \sum_{k=1}^{r} A_{kj}, j = 1,2,...,c,$ – суммы столбцов таблицы сопряженности.

Вычисление значимости связи основано на том факте, что статистика

$$\frac{\tau_b}{\sqrt{D\tau_b}},$$

где $D\tau_b = (4n + 10) / (9(n^2 – n))$ – дисперсия,
асимптотически имеет стандартное нормальное распределение $N(0,1)$.
Вариантами рассмотренного коэффициента являются коэффициенты τ_a и τ_c Кендалла, которые подробно описаны в монографии Кендалла (Кендэла), посвященной ранговым корреляциям. В данной монографии приведены также точные рекуррентные формулы распределений для малых выборок.

6.3.11. Коэффициент Крамера

Коэффициент Крамера (мера связанности Крамера) рассчитывается по формуле

$$V = \sqrt{\frac{\chi^2}{n\min(r-1, c-1)}},$$

где χ^2 – статистика критерия хи–квадрат ,

$n = \sum_{k=1}^{r}\sum_{l=1}^{c} A_{kl}$ – общее число наблюдений,

A_{ij}, $i = 1,2,...,r$; $j = 1,2,...,c$ – заданные частоты таблицы сопряженности.
r – число строк таблицы сопряженности,
c – число столбцов таблицы сопряженности.
Вычисление значимости связи основано на том факте, что статистика

$$\frac{V}{\sqrt{DV}},$$

где $DV = \frac{1}{n(q-1)}$ – дисперсия,
асимптотически имеет стандартное нормальное распределение $N(0,1)$. Таким образом, для больших выборок данная статистика подчиняется распределению $N(0,1)$.

См. источники: Крамер, Аптон, Кендалл с соавт.

6.3.12. Коэффициент Сомерса

Мера связанности Сомерса (коэффициент Сомерса, дельта Сомерса, Somers' D) является одной из разновидностей семейства мер Гудмана–Краскела. Он аналогичен коэффициенту Кендалла с той разницей, что при его вычислении производится дифференциальный учет пар с равными значениями переменных, учитывающих равенство первой и второй переменной. Коэффициент вычисляется по формулам

$D_x = S \Big/ \left[\frac{1}{2} n(n-1) - T_1 \right]$ – статистика «для строк»,

$D_y = S \Big/ \left[\frac{1}{2} n(n-1) - T_2 \right]$ – статистика «для столбцов»,

где $S = P - Q$,

$P = \sum_{i=1}^{r} \sum_{j=1}^{c} A_{ij} \left(\sum_{k>i} \sum_{l>j} A_{kl} \right)$ – число пар объектов с взаимно возрастающими переменными,

$Q = \sum_{i=1}^{r} \sum_{j=1}^{c} A_{ij} \left(\sum_{k>i} \sum_{l<j} A_{kl} \right)$ – число пар объектов с взаимно убывающими переменными,

A_{ij}, $i = 1,2,...,r$; $j = 1,2,...,c$ – заданные частоты таблицы сопряженности,

r – число строк таблицы сопряженности,

c – число столбцов таблицы сопряженности,

$n = \sum_{k=1}^{r} \sum_{l=1}^{c} A_{kl}$ – общее число наблюдений,

$T_1 = \frac{1}{2} \sum_{i=1}^{r} n_{i.}(n_{i.} - 1)$ – число пар объектов с взаимно равными значениями по одной переменной,

$n_{i.} = \sum_{k=1}^{c} A_{ik}, i = 1,2,...,r,$ – суммы строк таблицы сопряженности,

$T_2 = \frac{1}{2} \sum_{j=1}^{c} n_{.j}(n_{.j} - 1)$ – число пар объектов с взаимно равными значениями по другой переменной,

$n_{.j} = \sum_{k=1}^{r} A_{kj}, j = 1,2,...,c,$ – суммы столбцов таблицы сопряженности.

Асимптотические распределения статистик D_x и D_y вычисляются наподобие асимптотического распределения меры τ_c Стьюарта, приводятся в ряде источников и, возможно, будут реализованы в будущих версиях программного распределения.

6.3.13. Коэффициент сопряженности Пирсона

Коэффициент сопряженности Пирсона (Pearson's contingency coefficient) рассчитывается по формуле

$C = \sqrt{\frac{\chi^2}{\chi^2 + n}},$

где χ^2 – статистика критерия хи–квадрат ,

$n = \sum_{i=1}^{r}\sum_{j=1}^{c} A_{ij}$ – общее число наблюдений,
где A_{ij}, $i = 1,2,...,r$; $j = 1,2,...,c$ – заданные частоты таблицы сопряженности,
r – число строк таблицы сопряженности,
c – число столбцов таблицы сопряженности.
Значимость статистики критерия может быть оценена, ориентируясь на значимость статистики хи–квадрат, которая подчиняется распределению χ^2 с числом степеней свободы $(r-1)(c-1)$.

6.3.14. Критерий Краскела–Уоллиса

Критерий Краскела–Уоллиса (ранговый однофакторный анализ Краскела–Уоллиса) является непараметрическим аналогом однофакторного дисперсионного анализа и предназначен для проверки нулевой гипотезы о равенстве эффектов обработки (воздействия) на выборки с неизвестными, но равными средними. Нулевая гипотеза заключается в том, что все совокупности одинаково распределены. Вычисление критерия производится по формуле

$$H = \frac{12}{N(N+1)}\sum_{i=1}^{k}\frac{R_i^2}{n_i} - 3(N+1),$$

где R_i, $i = 1,2,...,k$ – сумма рангов наблюдений i–ой выборки,

$N = \sum_{i=1}^{k} n_i$ – общая численность,
n_i, $i = 1,2,...,k$ – численность i–й выборки,
k – количество столбцов (групп).
В программе введена поправка на объединение рангов

$$b = 1 - \frac{1}{N(N^2-1)}\sum_{j=1}^{g} t_j(t_j^2 - 1),$$

где t_j, $j = 1,2,...,g$ – численность связки,
g – число связок.
Тогда модифицированная статистика, выводимая программой, будет записана как $H' = H / b$.
Статистика критерия (равно и модифицированная статистика) имеет χ^2–распределение с параметром $k - 1$.

См. работы Бикела с соавт., Петровича с соавт., Холлендера с соавт. Точное вычисление критерия Краскела–Уоллиса см. в работе Клотца (Klotz) с соавт.

6.3.15. Диагностика Симонов–Цай

Диагностика Симонов–Цай (Simonoff–Tsai diagnostic) применяется для решения вопроса, допустима ли аппроксимация хи–квадрат в решении задачи кросстабуляции для конкретной таблицы сопряженности. Вычисление диагностики производится по формуле

$$S = \frac{(\chi^2(\nu,\alpha))^{1/2}}{3(X^2)^{3/2}}\sum_{i=1}^{r}\sum_{j=1}^{c}\frac{\left|(A_{ij} - E_{ij})\right|^3}{E_{ij}^2},$$

где $\chi^2(\nu,\alpha)$ – значение обратной функции распределения χ^2 для $\nu = (r-1)(c-1)$ степеней свободы и доверительного уровня α (обычно берется 0,95),
X^2 – статистика критерия хи–квадрат ,

A_{ij}, $i = 1,2,...,r$; $j = 1,2,...,c$ – заданные частоты таблицы сопряженности,
E_{ij}, $i = 1,2,...,r$; $j = 1,2,...,c$ – ожидаемые частоты таблицы сопряженности,
r – число строк таблицы сопряженности,
c – число столбцов таблицы сопряженности.
Ожидаемые частоты вычисляются по формуле

$$E_{ij} = \frac{n_{i.} n_{.j}}{n}, i = 1,2,...,r; j = 1,2,...,c,$$

где $n_{i.} = \sum_{k=1}^{c} A_{ik}, i = 1,2,...,r,$ – суммы строк таблицы сопряженности,

$n_{.j} = \sum_{k=1}^{r} A_{kj}, j = 1,2,...,c,$ – суммы столбцов таблицы сопряженности,

$n = \sum_{k=1}^{r} \sum_{l=1}^{c} A_{kl}$ – общее число наблюдений.

Если значение диагностики превышает значение 0,25, то это указывает на потенциальные проблемы с аппроксимацией χ^2.

См. материалы Хромова (Khromov) с соавт., Лаззаротто (Lazzarotto) с соавт.

6.3.16. Диагностика Хабермана

Диагностика Хабермана (Haberman diagnostic) применяется для решения вопроса, допустима ли аппроксимация хи–квадрат в решении задачи кросстабуляции для конкретной таблицы сопряженности. Вычисление диагностики производится по формуле

$$S = \frac{1}{\sqrt{32(rc-1)}} \sum_{i=1}^{r} \sum_{j=1}^{c} \left(\frac{1}{E_{ij}} - \frac{rc}{n} \right),$$

где r – число строк таблицы сопряженности,
c – число столбцов таблицы сопряженности,

$E_{ij} = \frac{n_{i.} n_{.j}}{n}, i = 1,2,...,r; j = 1,2,...,c,$ – ожидаемые частоты таблицы сопряженности,

$n_{i.} = \sum_{k=1}^{c} A_{ik}, i = 1,2,...,r,$ – суммы строк таблицы сопряженности,

$n_{.j} = \sum_{k=1}^{r} A_{kj}, j = 1,2,...,c,$ – суммы столбцов таблицы сопряженности,

$n = \sum_{k=1}^{r} \sum_{l=1}^{c} A_{kl}$ – общее число наблюдений,

A_{ij}, $i = 1,2,...,r$; $j = 1,2,...,c$ – заданные частоты таблицы сопряженности.
Если значение диагностики превышает значение 0,1, то это указывает на возможные проблемы с аппроксимацией χ^2. Значение диагностики более 1 указывает на серьезные проблемы с аппроксимацией.

См. материалы Хромова (Khromov) с соавт., Лаззаротто (Lazzarotto) с соавт.

Список использованной и рекомендуемой литературы

1. Ababneh F. Matched–pairs tests of homogeneity with applications to homologous nucleotide sequences / F. Ababneh, L.S. Jermiin, C. Ma et al. // Bioinformatics, 2006, vol. 22, no. 10, pp. 1225–1231.
2. Agresti A. An introduction to categorical data analysis. – New York, NY: John Wiley & Sons, 1996.
3. Agresti A. Categorical data analysis. – New York, NY: John Wiley & Sons, 2002.
4. Aoki S. Network algorithm for the exact test of Hardy–Weinberg proportion for multiple alleles // Department of Mathematical Engineering and Information Physics, The University of Tokyo, Technical Report METR 01–06, 2001.
5. Basu A. Improved power in multinomial goodness–of–fit tests / A. Basu, S. Ray, C. Park et al. // Journal of the Royal Statistical Society: Series D (The Statistician), 2002, vol. 51, pp. 381–393.
6. Basu A., Basu S. Penalized minimum disparity methods for multinominal models // Statistica Sinica, 1998, vol. 8, pp. 841–860.
7. Bhapkar V.P. A note on the equivalence of two test criteria for hypotheses in categorical data // Journal of the American Statistical Association, 1966, vol. 61, pp. 228–235.
8. Bhapkar V.P., Gore A.P. A distribution–free test for symmetry in hierarchical data // Journal of Multivariate Analysis, 1973, vol. 3, pp. 483–489.
9. Bhattacharya B., Basu A. Disparity based goodness–of–fit tests for and against isotonic order restrictions for multinomial models // Journal of Nonparametric Statistics, 2003, vol. 15, no. 1, pp. 1–10.
10. Bland J.M., Altman D.G. Statistics notes: Cronbach's alpha // BMJ (British Medical Journal), 1997, vol. 314, p. 572.
11. Bland M. An introduction to medical statistics. – Oxford, UK: Oxford University Press, 2000.
12. Borkowf C.B. An efficient algorithm for generating two–way contingency tables with fixed marginal totals and arbitrary mean proportions, with applications to permutation tests // Computational Statistics & Data Analysis, 2004, vol. 44, pp. 431–449.
13. Bowker A.H. A test for symmetry in contingency tables // Journal of the American Statistical Association, 1948, vol. 43, pp. 572–574.
14. Boyett J.M. Algorithm AS 144: Random R x C tables with given row and column totals // Applied Statistics, 1979, vol. 28, no. 3, pp. 329–332.
15. Bradley D.R. Type I error rate of the chi–square test of independence in R x C tables that have small expected frequencies / D.R. Bradley, T.D. Bradley, S.G. McGrath et al. // Psychological Bulletin, 1979, vol. 86, pp. 1290–1297.
16. Bratley P., Fox B., Schrage L. A guide to simulation. – New York, NY: Springer–Verlag, 1987.
17. Bravo F. Bartlett–type adjustments for empirical discrepancy test statistics // Economics Working Paper Archive at York, 2004, vol. 14.
18. Camilli G., Hopkins K. D. Applicability of chi–square to 2 x 2 contingency tables with small expected cell frequencies // Psychological Bulletin, 1978, vol. 85, pp. 163–167.
19. Chernick M.R. Friis R.H. Introductory biostatistics for the health sciences. Modern application including bootstrap. – New York, NY: John Wiley & Sons, 2003.
20. Corcoran C.D., Mehta C.R. Exact level and power of permutation, bootstrap and asymptotic tests of trend // Journal of Modern Statistical Methods, 2001.
21. Cressie N., Read T.R.C. Multinominal goodness–of–fit tests // Journal of the Royal Statistical Society: Series B (Statistical Methodology), 1984, vol. 46, no. 3, pp. 440–464.
22. Everitt B.S. The analysis of contingency tables. – London, UK: Chapman & Hall, 1977.

23. Fisher R.A. On the interpretation of χ^2 from contingency tables, and the calculation of P // Journal of the Royal Statistical Society, 1922, vol. 85, pp. 87–94.
24. Fisher R.A. Statistical tests of agreement between observation and hypothesis // Economica, 1923, vol. 3, pp. 139–147.
25. Fox B. Algorithm 647: Implementation and relative efficiency of quasirandom sequence generators // ACM Transactions on Mathematical Software, December 1986, vol. 12, no. 4, pp. 362–376.
26. Freeman G.H., Halton J.H. Note on an exact treatment of contingency, goodness of fit and other problems of significance // Biometrika, 1951, vol. 38, pp. 141–149.
27. Gail M., Mantel N. Counting the number of r x c contingency tables with fixed margins // Journal of the American Statistical Association, December 1977, vol. 72, no. 360, pp. 859–862.
28. Gokhale D.V., Kullback S. The information in contingency tables. – New York, NY: Marcel Dekker, 1978.
29. Goodman L.A. On methods for comparing contingency tables // Journal of the Royal Statistical Society: Series A (General), 1963, vol. 126, no. 1, pp. 94–108.
30. Greenwood P.E., Nikulin M.S. Guide to chi–squared testing. – New York, NY: John Wiley & Sons, 1996.
31. Haberman S.J. A warning on the use of chi–squared statistics with frequency tables with small expected cell counts // Journal of the American Statistical Association, 1988, vol. 82, no. 402, pp. 555–560.
32. Harshbarger T.R. Introductory statistics: A decision map. – New York, NY: Macmillan, 1971.
33. Karlis D., Xekalaki E. A simulation comparison of several procedures for testing the Poisson assumption // The Statistician, 2000, vol. 49, part 3, pp. 355–382.
34. Kastenbaum M.A., Lamphiear D.E. Calculation of chi–square to test the no three–factor interaction hypothesis // Boimetrics, March 1959, vol. 15, no.1, pp. 107–115.
35. Kendall M.G. Rank correlation methods. – London, UK: Griffin, 1970.
36. Khromov–Borisov N.N., Smolyanitsky A.G. Comprehensive catalog of statistical formulae, algorithms and software – step towards good statistics practice // Rechtsmedizin, 2003, No. 4, p. 278.
37. Klotz J., Teng J. One–way layout for counts and the exact enumeration of the Kruskal–Wallis H distribution with ties // Journal of the American Statistical Association, March 1977, vol. 72, no. 357, pp. 165–169.
38. Krampe A., Kuhnt S. Bowker's test for symmetry and modifications within the algebraic framework // Technical Report TR29–05, Universitat Dortmund, 2005.
39. Kroll N.E.A. Testing independence in 2 x 2 contingency tables // Journal of Educational and Behavioral Statistics, 1989, vol. 14, no. 1, pp. 47–79.
40. L'Ecuyer P. Random Number Generation // In Handbook of Simulation / Ed. by J. Banks. – New York, NY: John Wiley & Sons, 1998,
41. Lazzarotto G.B. SANCT – methodology and software for the structural analysis of forensic population data / G.B. Lazzarotto, N.N. Khoromov–Borisov, T.B.L. Kist et al. // Rechtsmedizin, 2003, No. 2, p. 279.
42. Lee R.P.–l. The use of correlational statistics in social survey research // The Chung Chi Journal, November 1969, vol. 9, no. 1, pp. 66–71.
43. Legendre P. Species associations: The Kendall coefficient of concordance revisited // The Journal of Agricultural, Biological, and Environmental Statistics, 2005, vol. 10, no. 2, pp. 226–245.
44. Lehmann E.L. Testing statistical hypotheses. – New York, NY: Chapman & Hall, 1994.
45. Lewis J., Baldwin J. Statistical package for improved analysis of hillslope monitoring data

collected as part of the board of forestry's long–term monitoring program. Agreement No. PSW–96–CL–032, CDF No. 8CA95056, Final Report, May 1997. California Department of Forestry & Fire Protection.
46. Ludbrook J. Computer–intensive statistical procedures // Critical Reviews in Biochemistry and Molekular Biology, 2000, vol. 35, no. 5, pp. 339–358.
47. Ludbrook J. Statistical techniques for comparing measurers and methods of measurement: A critical review // Clinical and Experimental Pharmacology and Physiology, 2002, vol. 29, pp. 527–536.
48. Lydersen S., Fagerland M.W., Laake P. Recommended tests for association in 2 x 2 tables // Statistics in Medicine, 2009, vol. 28, pp. 1159–1175.
49. March D.L Exact probabilities for R x C contingency tables [G2] // Communications of the ACM archive, November 1972, vol. 15, no. 11, pp. 991–992.
50. Martin Andres A., Tapia Garcia J.M. Optimal unconditional test in 2 x 2 multinomial trials // Computational Statistics & Data Analysis, 1999, vol. 31, pp. 311–321.
51. Maxwell A.E. Comparing the classification of subjects by two independent judges // British Journal of Psychiatry, 1970, vol. 116, pp. 651–655.
52. Mehta C.R., Patel N.R. A network algorithm for performing Fisher's exact test in r x c contingency tables // Journal of the American Statistical Association, 1983, vol. 78, pp. 427–434.
53. Mehta C.R., Patel N.R. Exact inference for categorical data // Biometrics, 1997, vol. 53, no. 1, 112–117.
54. Montgomery D.C., Runger G.C. Applied statistics and probability for engineers. – New York, NY: John Wiley & Sons, 2003.
55. Mudholkar G.S., Hutson A.D. Continuity corrected approximations for and «exact» inference with Pearson's X^2 // Journal of Statistical Planning and Inference, 1997, vol. 59, pp. 61–78.
56. Muller M.J. Exact tests for small sample 3 x 3 contingency tables with embedded fourfold tables: Rationale and application // The German Journal of Psychiatry, 2001, no. 4, pp. 57–62.
57. Neyman J. Contributions to the theory of the χ^2 test // Proceedings of the First Berkley Symposium on Mathematical Statistics and Probability, 1949.
58. Olsson U. Measuring correlation in ordered two–way contingency tables // Journal of Marketing Research, 1980, vol. 17, pp. 391–394.
59. Pagano M., Halvorsen K.T. An algorithm for finding the exact significance levels of r x c contingency tables // Journal of the American Statistical Association, 1981, vol. 76, pp. 931–934.
60. Park C., Basu A., Harris I.R. Tests of hypotheses in multiple samples based on penalized disparities // Pennsylvania State University, Department of Statistics, Technical Report No. 2001–02–03.
61. Patefield W.M. Algorithm AS 159: An efficient method of generating random R x C tables with given row and column totals // Applied Statistics, 1981, vol. 30, no. 1, pp. 91–97.
62. Perez T., Pardo J.A. On choosing a goodness–of–fit test for discrete multivariate data // Kybernetes, December 2003, vol. 32, no. 9/10, pp. 1405–1424.
63. Powers S., Gose K.C. A Basic program for calculating the Stuart–Maxwell test // Educational and Psychological Measurement, 1986, vol. 46, no. 3, pp. 651–653.
64. Read T.R.C. Small–sample comparisons for the power divergence goodness–of–fit statistics // Journal of the American Statistical Association, December 1984, vol. 79, no. 388, pp. 929–935.
65. Read T.R.C., Cressie N. Goodness–of–fit statistics for discrete multivariate data. – New York, NY: Springer–Verlag, 1988.

66. Rupp. T. Rough set methodology in meta–analysis: A comparative and exploratory analysis // Darmstadt Discussion Papers in Economics, no. 157. – Darmstadt: Darmstadt University of Technology, 2005.
67. Saunders I.W. Algorithm AS 205: Enumeration of R x C tables with repeated row totals // Applied Statistics, 1984, vol. 33, no. 3, pp. 340–352.
68. Simonoff J.S., Tsai C.–L. Assessing the influence of individual observations on a goodness–of–fit test based on nonparametric regression // Statistics & Probability Letters, July 1991, vol. 12, no. 1, pp. 9–17.
69. Simonoff J.S., Tsai C.–L. Higher order effects in log–linear and log–non–linear models for contingency tables with ordered categories // Journal of the Royal Statistical Society: Series C (Applied Statistics), 1991, vol. 40, no. 3, pp. 449–458.
70. Smith P., McDonald J. Simulate and reject Monte Carlo exact conditional test for quasi–independence // Proceedings of COMPSTAT, 1994.
71. Sokal R.R., Rohlf F.J. Biometry: the principles and practice of statistics in biological research. – New York, NY: W.H. Freeman, 1995.
72. Somers R.H. A new asymmetric measure of association for ordinal variables // American Sociological Review, 1962, vol. 27, pp. 799–811.
73. Sprent P., Smeeton N.C. Applied nonparametric statistical methods. – Boca Raton, FL: Chapman & Hall / CRC, 2001.
74. Stafford J.E. Exact cumulant calculations for Pearson X^2 and Zelterman statistics for r–way contingency tables // Journal of Computational and Graphical Statistics, 1995, vol. 4, no. 3, pp. 199–212.
75. Stuart A.A. A test for homogeneity of the marginal distributions in a two–way classification // Biometrika, 1955, vol. 42, pp. 412–416.
76. Van Belle G. Biostatistics: A methodology for the health sciences // G. van Belle, L.D. Fisher, P.J. Heagerty et al. – New York, NY: John Wiley & Sons, 2003.
77. Von Davier M. Bootstrapping goodness–of–fit statistics for sparse categorical data – results of a Monte Carlo study // Methods of Psychological Research Online, 1997, vol.2, no. 2.
78. Von Eye A., Schauerhuber M., Mair P. Significance tests for the measure of raw agreement // InterStat (Statistics on the Internet), January 2007, no. 1.
79. Williams D.A. Improved likelihood ratio tests for complete contingency tables // Biometrika, April 1976, vol. 63, no. 1, pp. 33–37.
80. Zelterman D. Approximating the distribution of goodness–of–fits tests for discrete data // Computational Statistics and Data Analysis, 1984, vol. 2, pp. 207–214.
81. Zelterman D. Discrete distributions: Applications in the health sciences. – New York, NY: John Wiley & Sons, 2004.
82. Zelterman D. Goodness–of–fit tests for large sparse multinomial distributions // Journal of the American Statistical Association, June 1987, vol. 82, no. 398, pp. 624–629.
83. Zelterman D. Models for discrete data. – Oxford, UK: Oxford Science Publications, 1999.
84. Zelterman D., Chan I.S.–F., Mielke P.W. Exact tests of significance in higher dimensional tables // The American Statistician, 1995, vol. 49, pp. 357–361.
85. Аптон Г. Анализ таблиц сопряженности. – М.: Финансы и статистика, 1982.
86. Афифи А., Эйзен С. Статистический анализ: Подход с использованием ЭВМ. – М.: Мир, 1982.
87. Бикел П., Доксам К. Математическая статистика. Выпуск 2. – М.: Финансы и статистика, 1983.
88. Брандт З. Анализ данных. Статистические и вычислительные методы для научных работников и инженеров. – М.: Мир, ООО «Издательство АСТ», 2003.
89. Браунли К.А. Статистическая теория и методология в науке и технике. – М.: Наука,

1977.
90. Кендалл М., Стьюарт А. Статистические выводы и связи. – М.: Наука, 1973.
91. Кендэл М. Ранговые корреляции. – М: Статистика, 1975.
92. Кулаичев А.П. Методы и средства анализа данных в среде Windows®. STADIA. – М.: Информатика и компьютеры, 1999.
93. Кулаичев А.П. Методы и средства комплексного анализа данных. – М.: ИНФРА–М, 2006.
94. Петрович М.Л., Давидович М.И. Статистическое оценивание и проверка гипотез на ЭВМ. – М.: Финансы и статистика, 1989.
95. Прохоров Ю.В. Вероятность и математическая статистика. Энциклопедия / Гл. ред. Ю.В. Прохоров. – М.: Научное издательство «Большая Российская энциклопедия», 1999.
96. Раушенбах Г.В. Меры близости и сходства // Анализ нечисловой информации в социологических исследованиях. – М.: Наука, 1985, с. 169–203.
97. Сборник научных программ на Фортране. Выпуск 1. Статистика. – М.: Статистика, 1974.
98. Сергиенко В.И., Бондарева И.Б. Математическая статистика в клинических исследованиях. – М.: ГЭОТАР–Медиа, 2006.
99. Тюрин Ю.Н., Макаров А.А. Статистический анализ данных на компьютере. – М.: ИНФРА–М, 1998.
100. Флейс Дж. Статистические методы для изучения таблиц долей и пропорций. – М.: Финансы и статистика, 1989.
101. Холлендер М., Вулф Д. Непараметрические методы статистики. – М.: Финансы и статистика, 1983.

Глава 7. Проверка нормальности распределения

7.1. Введение

Проверка типа распределения эмпирической выборки, частная задача которого – проверка нормальности, имеет важнейшее значение в прикладной статистике и является излюбленным сюжетом в статистической литературе. Перечислим только некоторые из задач, которые решаются с использованием данных методов:

- Для принятия решения, применять тот или иной метод статистической обработки данных, часто необходимо установить, является ли нормальным распределение количественной эмпирической выборки.
- Важной задачей анализа согласия распределения является тестирование датчиков случайных чисел, применяемых в моделировании методом Монте–Карло в различных областях науки и техники.
- По типу статистического распределения параметров технологического процесса можно сделать определенные выводы о качестве этого процесса и вовремя скорректировать процесс.

Можно указать и другие задачи, в которых необходима проверка типа распределения. Обратим внимание пользователей, что:

- методы представленного в настоящей главе программного обеспечения работают только с количественными эмпирическими выборками в одномерном и в многомерном случае;
- тестируется только нормальность (напомним, что нормальное распределение является непрерывным), но не распределение другого типа.

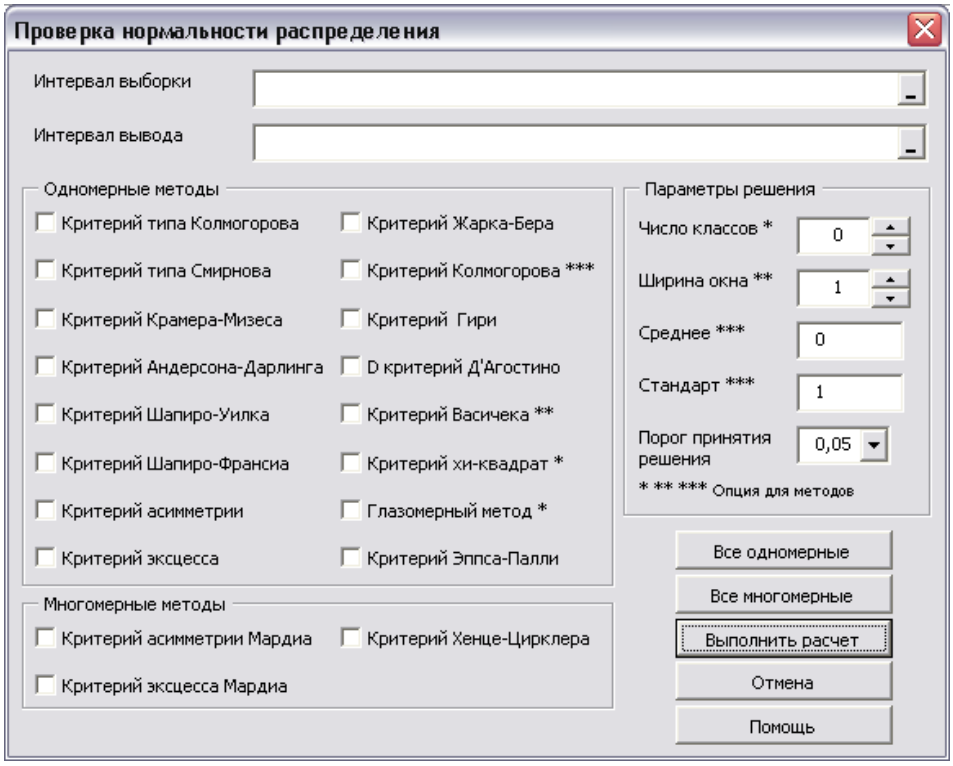
Применяются разнообразные методы, предназначенные для тестирования различных параметров распределения, в той или иной степени позволяющих исследовать его нормальность. Выводов бывает достаточно для принятия решения о выборе методов дальнейшего прикладного анализа, в частности, параметрической или непараметрической статистики.
Для проверки нормальности распределения реализации случайной одномерной величины, представленной в виде эмпирической выборки, программой предлагаются различные критерии. Методы реализуют почти все классические и современные подходы к проверке согласия распределения количественной эмпирической выборки с нормальным распределением.
Также представлены методы проверки согласия эмпирического многомерного распределения с нормальным теоретическим многомерным распределением. При проверке согласия многомерного распределения размерность эмпирической выборки может быть произвольной. Отметим, что в данном случае размерностью выборки называют число измерений, которым представлена каждая варианта многомерной выборки. Удобна геометрическая интерпретация данного параметра. Фактически каждая варианта (элемент) такой выборки представлена точкой в многомерном пространстве, размерность которого и есть размерность выборки вариант. Размерность не следует путать с численностью выборки, представляющей собой количество вариант.
Методы охватывают выборки практически любой численности. Однако показано (см. статью Селезнева с соавт., ссылки и другие работы), что для малых выборок (при численности выборки менее 50) и уровня значимости ≤ 0,05 все критерии проверки нормальности «работают» плохо вследствие малой мощности при малой численности выборки. Дополнительно о влиянии численности на мощность критериев см. главу «Введение». Тем не менее, критерий Шапиро–Уилка показывает для таких выборок лучшие результаты, чем другие тесты.

7.2. Работа с программным обеспечением

Выберите из меню программы пункт AtteStat | Проверка нормальности. На экране появится диалоговое окно, подобное изображенному на рисунке.
Затем проделайте следующие шаги:

- Выберите или введите интервал эмпирической количественной выборки. Для проверки согласия одномерного распределения выборка может занимать строку, столбец или прямоугольную область рабочего листа (в данном случае численность выборки будет получена программой перемножением количества строк на количество столбцов). Для проверки согласия многомерного распределения количество столбцов выбранного интервала будет означать размерность выборки.
- Выберите или введите интервал вывода. Можно указать только первую ячейку данного интервала.
- Выберите метод (или методы) проверки нормальности. Может быть выбрано любое число предлагаемых методов одновременно. Можно воспользоваться кнопками Все одномерные или Все многомерные. При нажатии соответствующей кнопки будет выбрана полностью группа одномерных или многомерных критериев.
- Для глазомерного метода и критерия хи–квадрат Фишера измените или оставьте по умолчанию число классов. Значение по умолчанию, равное «0», означает, что число классов будет вычислено автоматически.
- Для критерия Васичека измените или оставьте по умолчанию ширину окна.
- Для критерия Колмогорова введите или оставьте по умолчанию значения среднего значения и стандартного отклонения.

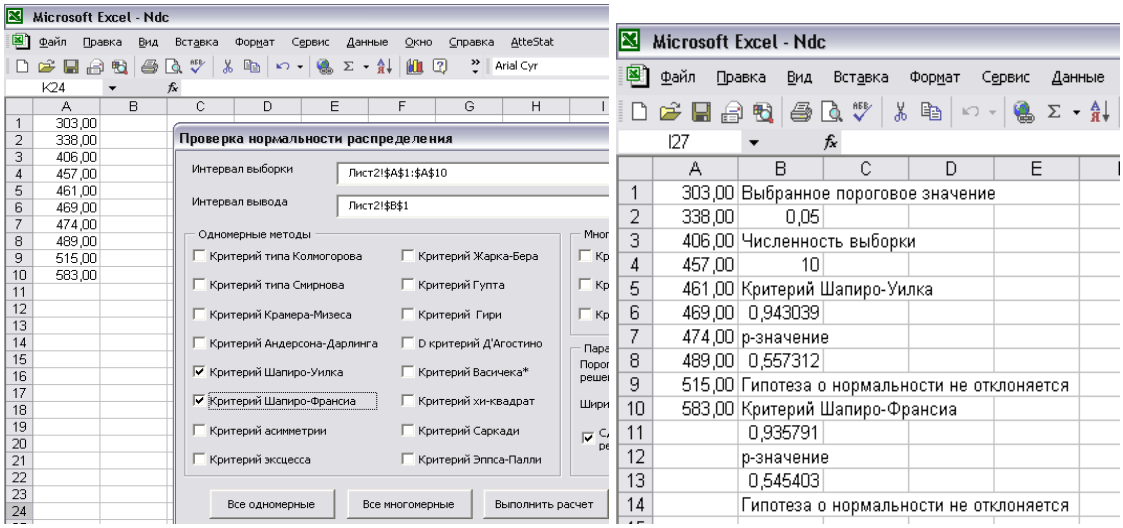
- Нажмите кнопку «Выполнить расчет».



После вычисления, начиная с ячейки, указанной на этапе 2 (интервал вывода), будет выдан результат расчета в виде значения статистики критерия, *P*–значения (для некоторых критериев – критического значения).
При выборе нескольких критериев одновременно результаты расчетов будут выданы друг за другом. Если по какой–либо причине стандартные пороговые значения не устраивают пользователя, их можно просто проигнорировать и сделать вывод по результатам расчета достигнутого уровня значимости по своему усмотрению.
При возникновении ошибок, вызванных неверными действиями пользователя, или ошибок периода выполнения, выдаются сообщения об ошибках.

7.2.1. Пример применения

В качестве примера протестируем выборку, приведенную на с. 338 монографии Хана и Шапиро. В ячейки A1:A10 листа электронных таблиц введем 10 вариант эмпирической выборки. Затем выберем из меню AtteStat «Проверка нормальности». В качестве интервала выборки методом протаскивания курсора выделяем введенную выборку. В качестве интервала вывода указываем ячейку B1. Начиная с данной ячейки, будет производиться вывод результатов расчета. Затем отмечаем один или несколько методов, с помощью которых будет тестироваться выборка. Выберем критерий, представленный в источнике (критерий Шапиро–Уилка) и родственный тест (критерий Шапиро–Франсиа). После перечисленных действий экран компьютера будет выглядеть примерно так, как показано на следующем фрагменте.



Нажатием кнопки «Выполнить расчет» будет запущена процедура расчета. После небольшого времени, зависящего от быстродействия компьютера, экран примет вид, подобный показанному на рисунке.

Получено полное совпадение результатов расчета с источником. Интерпретация результатов дана при описании соответствующих методов расчета. В ячейках **B6**, **B8**, **B11**, **B13** мы установили число знаков после десятичной точки, равное 6.

7.2.2. Сообщения об ошибках

При ошибках ввода и во время выполнения программы могут выдаваться диагностические сообщения следующих типов:

Ошибка	Комментарий
Не определен интервал переменной.	Вы не выбрали или неверно ввели интервал эмпирической выборки. Лучшим способом избежать ошибки является не ввод, а выделение интервала стандартным образом, т. е. протаскиванием курсора.
Пустая ячейка.	Проверьте исходные данные и заполните все ячейки, отмеченные Вами как входной интервал. Для избежания ошибок расчета, вызванных разногласиями, трактовать ли пустую ячейку как нуль, данное программное обеспечение требует заполнения всех ячеек. Если в ячейке не должно быть данных по физической природе исследуемого процесса, введите в данную ячейку нуль.
Нечисловой тип данных.	Проверьте типы ячеек входного интервала. Тип может быть только числовым. Проще всего выделить интервал ячеек и явно определить их тип как числовой стандартными средствами.
Не определена область вывода.	Не выбран или неверно введен выходной интервал. Лучшим способом избежать ошибки является не ввод, а выделение интервала стандартным образом, т. е. протаскиванием курсора.

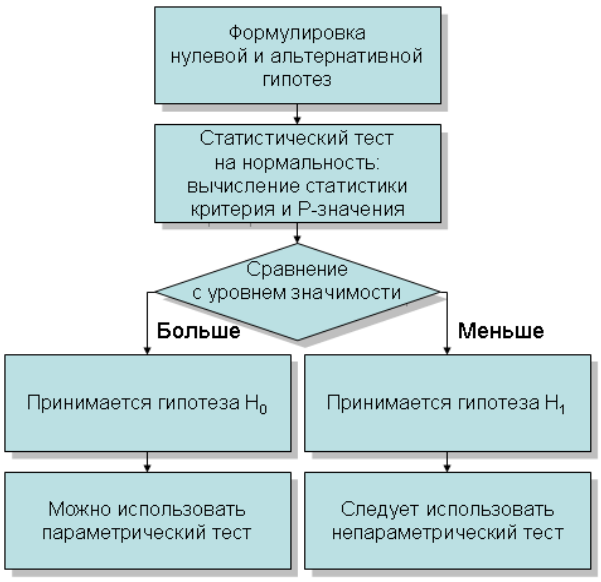
7.3. Теоретическое обоснование

Если тип распределения некоторой случайной величины нам неизвестен, располагая

случайной эмпирической выборкой (реализацией случайной величины), мы можем захотеть проверить, совпадает ли эмпирическая функция распределения случайной величины с некоторой заданной или вычисленной по выборочным параметрам теоретической функцией эмпирического распределения. При такой постановке говорят о проверке статистической гипотезы согласия.
Частным случаем данной задачи является установление нормальности распределения (соответствия эмпирической функции распределения непрерывной количественной случайной величины и нормальной функции распределения). Парадоксальным эпиграфом к данной главе могли быть слова Фишера: «Отклонения от нормальной формы распределения, если только они не представляются явными без всякой оценки, могут быть обнаружены только в случае большой выборки; при малых же выборках оказывается невозможным определение сколько–нибудь надежных статистических критериев для этих отклонений». К счастью, за полвека, прошедшие со времени данной публикации, были выполнены определенные исследования.
Все критерии проверки типа распределения (и, в частном случае, проверки нормальности) часто называют критериями согласия, хотя, по нашему мнению, критериями согласия справедливо называть только критерии, основанные на функциях распределения, названные так на основе термина «согласие распределений».

7.3.1. Процедура тестирования

Нормальность – основная предпосылка применения параметрических тестов, представленных в программах анализа данных. Поэтому часто исследователя интересует вопрос, соответствует ли распределение эмпирической выборки, измеренной в количественной шкале, нормальному распределению. На схеме показан алгоритм действий при проверке нормальности распределения.
Анализ выполняется стандартно. Сначала формулируют нулевую гипотезу и задаются уровнем значимости. Нулевая гипотеза обычно H_0 формулируется так: «нет статистически значимого различия». Альтернативная (конкурирующая) двусторонняя гипотеза H_1: «имеет место статистически значимое различие» (возможны варианты формулировки). Затем, с учетом, является ли гипотеза простой или сложной (в программе речь идет только о сложной гипотезе), проверяется согласие эмпирического распределения либо иных характеристик нормальному распределению, после чего по результатам проверки делается вывод.



Методы, представленные в данном программном обеспечении, выдают *P*–значение, позволяющее принять или отвергнуть нулевую гипотезу, поэтому отпадает необходимость использования статистических таблиц.

7.3.2. Типы тестов на нормальность

Важно представлять, для какой цели производится проверка нормальности распределения. К примеру, соответствие асимметрии или эксцесса эмпирического распределения тем же параметрам нормального теоретического распределения совсем не тождественно согласию эмпирической и теоретической функций эмпирического распределения. Авторами показано, что в ряде задач достаточно проверить лишь некоторые параметры распределения.
Считается, что для выборок, немного отличающихся от нормальных, результаты применения критерия Стьюдента (см. главу «Параметрическая статистика») будут близки к верным результатам, если эксцесс и коэффициент асимметрии анализируемых выборок, как у нормальных выборок.
Более подробную информацию по данному вопросу можно найти в работе Рейнеке (Reineke) с соавт. и в статье Д'Агостино (D'Agostino) с соавт. (1990 г.). Ряд критериев предназначен для тестирования нескольких параметров одновременно. Эти критерии называют омнибусными (в отечественных источниках принято наименование – составные). Название «омнибусный» заимствовано из социологии. В социологии омнибусным исследованием принято называть исследование, проводимое одновременно для нескольких клиентов и по нескольким темам. Такая организация исследования дает возможность каждому из клиентов за меньшие деньги и в более короткий срок получить оперативную информацию по интересующим вопросам, что позволяет снизить затраты на проведение самостоятельных исследований в несколько раз. Отметим, что омнибусные критерии относятся не к отдельному типу исследования согласия распределений, а к способу организации такого исследования. Поэтому омнибусные критерии могут иметь место не только в категории «Критерии моментов», но и в других категориях.
Проверка нормальности распределения может быть выполнена с помощью специальных статистических критериев, в зависимости от анализируемых характеристик эмпирической выборки. Современными авторами обычно выделяются критерии следующих типов:

- критерии функций распределения,
- критерии, основанные на регрессии,
- критерии моментов, включая составные тесты,
- информационные критерии,
- графические методы,
- Байесовские критерии.

Сводка основных идей проверки типа распределения (в т. ч. различные подходы к проверке нормальности) представлена Кобзарем. Подробный обзор типов критериев дан в диссертации Ли (Lee). В специальной литературе предложены и другие идеи по поводу проверки нормальности статистического распределения. См. работы Деклерка (Declercq) и Дюво (Duvaut), Лианг (Liang) и Бентлера (Bentler).

7.3.2.1. Простые и сложные гипотезы

При проверке согласия эмпирического и некоторого теоретического распределения различают простые и сложные гипотезы:

- простой гипотеза будет в том случае, если теоретическое распределение задано всеми своими параметрами;

- сложной гипотеза будет, если все или некоторые параметры теоретического распределения неизвестны и оцениваются по выборке.

Иначе, если распределение имеет l параметров и гипотеза утверждает, что k из них имеют заданные значения, то гипотеза будет:

- простой, если $k = l$,
- сложной, если $k < l$.

Разность $l – k$ называется числом степеней свободы гипотезы, а k будет числом ограничений, наложенных гипотезой.

В случае нормального распределения по выборке могут оцениваться математическое ожидание (его оценка – среднее значение) и дисперсия (для других типов распределения число оцениваемых по эмпирической выборке параметров может быть другим). Поэтому для нормального распределения сложная гипотеза может быть трех видов:

- по выборке оценивается математическое ожидание (его оценка – среднее значение), дисперсия задана,
- по выборке оценивается дисперсия, математическое ожидание задано,
- по выборке оцениваются и математическое ожидание, и дисперсия.

Хотя статистика критерия вычисляется во всех случаях по одним и тем же алгоритмам, необходимо наличие статистических таблиц или, лучше, формул вычисления критических значений либо *P*–значений, особых как для каждого типа распределения, так и для каждого случая сложной гипотезы. В литературе опубликованы формулы или таблицы для многих критериев и для различных гипотез.

В данном программном обеспечении для пользователей доступен один критерий для простой гипотезы – критерий Колмогорова. Остальные критерии применяются только для сложной гипотезы, когда математическое ожидание и дисперсия там, где это необходимо по условиям алгоритма, оцениваются по эмпирической выборке (см. также особенности для критерия хи–квадрат).

Некоторые критерии согласия изначально, по замыслу алгоритмов своего вычисления, представленному их авторами, не предполагают различие простой и сложной гипотез. Все параметры оцениваются по эмпирической выборке, поэтому данные критерии предназначены только для сложных гипотез.

7.3.3. Критерии функций распределения

Критерии, построенные на основе функций распределения, в зависимости от метрики подразделяются на следующие типы:

- критерии типа Колмогорова,
- критерии типа омега–квадрат,
- критерии типа Эппса–Палли.

Данные критерии являются эффективными методами проверки согласия распределений. Рассматриваются критерии нормальности, построенные на непосредственном сравнении эмпирической и теоретической функций эмпирического распределения и различающиеся метриками.

Эмпирической функцией распределения (empirical distribution function, EDF) называют такую функцию $F_n(x)$ от вариант упорядоченной в порядке возрастания выборки, что

$$F_n(x_i) = \frac{i}{n}, i = 1,...,n,$$

где x_i, $i = 1,2,...,n$ – варианты упорядоченной выборки,
n – численность выборки.

Таким образом, эмпирическая функция распределения от каждой варианты выборки показывает, сколько вариант выборки меньше данной варианты. График функции $F_n(x)$

является равномерным по оси ординат ступенчатым графиком с шагом ступеньки, в точности равным i / n. Весь график заключен в полосе, ограниченной сверху и снизу ординатами с численными значениями 0 и 1. По оси абсцисс график в общем случае равномерным не является.
Как показывает приведенная выше формула, эмпирическая функция распределения может строиться непосредственно по заданной эмпирической выборке, минуя какие–либо промежуточные вычисления.
В случае простой гипотезы теоретическая функция распределения полностью определена заданными параметрами. В случае сложной гипотезы, когда все параметры распределения оцениваются по выборке, теоретическая функция – это функция нормального распределения, определенная параметрами, вычисленными по эмпирической выборке.
Эмпирическая характеристическая функция (empirical characteristic function, ECF) распределения имеет вид

$$\psi_n(t) = \frac{1}{n}\sum_{j=1}^{n} e^{it\left(X_j - \bar{X}\right)S^{-1}},$$

i – мнимая единица,
t – нормированное отклонение,
n – численность выборки,
X_j, $j = 1,2,...,n$ – исходная выборка, в общем случае многомерная,
$\bar{X}$ – вектор средних значений,
S – матрица дисперсий–ковариаций.

Сравнительный обзор критериев, модификации статистик Колмогорова, Крамера – фон Мизеса, Койпера, Уотсона и Андерсона–Дарлинга для различных вариантов гипотез нормальности и экспоненциальности дал Стефенс (Stephens, 1974). Методика моделирования методом Монте–Карло представлена Стефенсом (1970).

7.3.3.1. Критерии типа Колмогорова

Представлены следующие критерии рассматриваемого типа:

- Критерий Колмогорова (классический, для простой гипотезы).
- Модифицированный критерий Колмогорова.
- Модифицированный критерий Смирнова.

Критерии типа Колмогорова предназначены для проверки согласия эмпирической и теоретической функций распределения и построены на модульной метрике. Статистика задается формулой

$$D_n = \sup_{|x|<\infty} \left|F_n(x) - F(x)\right|,$$

где $F_n(.)$ – эмпирическая функция распределения, построенная тем или иным способом по исходной эмпирической выборке,
$F(.)$ – теоретическая функция распределения.

Модифицированный критерий Колмогорова известен также под наименованием точного критерия Дарбина (Durbin’s exact test). См. также работу Дайера (Dyer). Интересным вариантом рассматриваемого теста можно считать критерий, предложенный Ляо (Liao) и Шимокава (Shimokawa) и описанный в аналитическом обзоре Хассана (Hassan), изученный для некоторых специальных типов распределений. См. также замечания к критерию Койпера (Kuiper), представленному в главе «Непараметрическая статистика».

7.3.3.1.1. Критерий Колмогорова

Статистика критерия Колмогорова представляет собой результат сравнения эмпирической и заданной теоретической функций распределения в модульной метрике

$$D_n = \sup_{|x|<\infty} |F_n(x) - F(x)|,$$

что эквивалентно

$$D_n = \sup_{-\infty<x<\infty} |F_n(x) - F(x)|,$$

где x – случайная величина,
$F_n(.)$ – эмпирическая функция распределения.
$F(.)$ – теоретическая функция распределения.
Предполагается, что теоретическая функция распределения полностью задана своими параметрами. Иначе, рассматривается простая гипотеза. Это означает, что параметры распределения не могут быть вычислены по эмпирической выборке.
Статистика критерия Колмогорова обладает тем интересным свойством, что для любой непрерывной теоретической функции распределения распределение статистики $\sqrt{n}D_n$ при $n \to \infty$ подчиняется λ–распределению (распределению Колмогорова):

$$\lim_{n\to\infty} P\left(\sqrt{n}D_n > x\right) = K(x),$$

где $K(x)$ – функция распределения Колмогорова.

Критерий Колмогорова представлен в оригинальной работе 1933 г. См. также статью Каца (Kac). Пример ошибочного вычисления см. в монографии Руниона. Для понимания принципа расчета может помочь графическая интерпретация в статьях Мэйджа (Mage), Аймэна (Iman), работах Шора с соавт., Мюллера с соавт.

7.3.3.1.2. Модифицированный критерий Колмогорова

Практическое вычисление статистики модифицированного критерия Колмогорова (критерия типа Колмогорова для сложной гипотезы) производится по формуле

$$D_n = \max(D_n^+, D_n^-),$$

где

$$D_n^+ = \max_{1\le m\le n}\left(\frac{m}{n} - F(\eta_m)\right),$$

$$D_n^- = \max_{1\le m\le n}\left(F(\eta_m) - \frac{m-1}{n}\right),$$

η_m, m = 1,2,...,n – эмпирическая выборка, отсортированная в порядке возрастания значений вариант,
n – численность выборки,
$F(.)$ – теоретическая функция распределения.
Все или некоторые параметры теоретической функции непрерывного распределения (в данном случае функции нормального распределения) для случая сложной гипотезы оцениваются по эмпирической выборке.
В Рекомендациях по стандартизации Р.50.1.037–2002 предложено вычислять модифицированную статистику

$$S_K = \frac{6nD_n + 1}{6\sqrt{n}},$$

хотя в программе для удобства и с целью сравнения с другими программами анализа данных

выдается значение статистики D_n. Если пользователя интересует значение модифицированной статистики, пересчет не вызовет затруднений.
Распределение рассматриваемого критерия не обладает свойством независимости от типа распределения, характерным для критерия Колмогорова. Поэтому для каждого тестируемого теоретического распределения и каждого случая сложной гипотезы распределение статистики критерия будет отличаться. В упомянутых Рекомендациях рассматриваются различные варианты критерия. В представленном же программном обеспечении рассматривается только случай проверки нормальности, когда все параметры распределения оцениваются по эмпирической выборке.
В данных Рекомендациях установлено, что *P*–значения статистики S_k для проверки нормальности в случае сложной гипотезы, когда оба параметра распределения оцениваются по эмпирической выборке (метод представлен в настоящей программе), могут быть аппроксимированы обобщенной функцией гамма–распределения с параметрами (4,9014; 0,0691; 0,2951).

Пример дан в книгах Кулаичева, а также Шора с соавт. Иные применяемые аппроксимации описаны в монографиях Тюрина и Тюрина с соавт., статье Лиллиефорса (в зарубежных источниках представленный критерий может называться Kolmogorov–Smirnov test with Lilliefors critical values или Kolmogorov–Smirnov test with Lilliefors correction) и других источниках.

7.3.3.1.3. Модифицированный критерий Смирнова

Вычисление статистики критерия типа Смирнова (модифицированного критерия Смирнова) производится по формуле

$$D_n^+ = \max_{1\le m\le n}\left(\frac{m}{n} - F(\eta_m)\right),$$

где η_m, $m = 1,2,...,n$ – эмпирическая выборка, отсортированная в порядке возрастания значений вариант,
n – численность выборки,
$F(.)$ – теоретическая функция распределения.
В Рекомендациях по стандартизации Р.50.1.037–2002 предложено вычислять модифицированную статистику

$$S_M = \frac{(6nD_n^+ + 1)^2}{9n}.$$

хотя в программе для удобства и с целью сравнения с другими программами анализа данных выдается значение статистики D_n^+. Если пользователя интересует значение модифицированной статистики, пересчет не вызовет затруднений.
Для каждого тестируемого теоретического распределения и каждого случая сложной гипотезы распределение статистики критерия будет отличаться. В упомянутых Рекомендациях рассматриваются различные варианты критерия. В представленном же программном обеспечении рассматривается только случай проверки нормальности, когда все параметры распределения оцениваются по эмпирической выборке. В данных Рекомендациях установлено, что *P*–значения статистики S_M для проверки нормальности в случае сложной гипотезы, когда оба параметра распределения оцениваются по эмпирической выборке, могут быть аппроксимированы функцией логнормального распределения с параметрами (0,1164; 0,5436).

О методе см. справочник Большева с соавт., также Руководство по пакету прикладных программ SSJ (Stochastic Simulation in Java), составленному Лекюйе (L'Ecuyer), и указанные в нем источники, в том числе относительно точного вычисления распределения статистики Смирнова.

7.3.3.2. Критерии типа омега–квадрат

Представлены следующие критерии рассматриваемого типа:

- Критерий Крамера–Мизеса.
- Критерий Андерсона–Дарлинга.
- Критерий хи–квадрат Фишера.

Критерии типа омега–квадрат основаны на идее сравнения эмпирической и теоретической функций распределения в квадратичной метрике

$$\omega^2 = \int_{-\infty}^{\infty} [F_n(x) - F(x)]^2 \psi[F(x)] dF(x),$$

где $F_n(.)$ – эмпирическая функция распределения, построенная тем или иным способом по исходной эмпирической выборке,
$F(.)$ – теоретическая функция распределения,
$\psi[.]$ – некоторая весовая функция.
Таблицы для определения критических значений критериев будут различаться для простой гипотезы и для каждого случая сложной гипотезы при оценке согласия эмпирического распределения с конкретным типом теоретического распределения.

Применение критериев типа омега–квадрат для проверки согласия различных распределений исследовано Г.В. Мартыновым. См. также Рекомендации по стандартизации Р.50.1.037–2002.

7.3.3.2.1. Критерий Крамера–Мизеса

При выборе весовой функции в критерии типа омега–квадрат в виде $\psi(t) = 1$ получается критерий Крамера–Мизеса (Мизеса, Крамера–Фон Мизеса, Крамера–Мизеса–Смирнова и др.). Как и в алгоритме вычисления критерия Колмогорова, функция распределения может строиться непосредственно по эмпирической выборке, без разнесения вариант по классам, поэтому практическое вычисление статистики критерия Крамера–Мизеса удобно производить по формуле

$$S_\omega = n\omega^2 = \frac{1}{12n} + \sum_{j=1}^{n} \left[F(\eta_j) - \frac{2j-1}{2n} \right]^2,$$

где η_m, $m = 1,2,...,n$ – эмпирическая выборка, отсортированная в порядке возрастания значений вариант,
n – численность выборки,
$F(.)$ – теоретическая функция распределения.
Для каждого тестируемого теоретического распределения и каждого случая сложной гипотезы распределение статистики критерия будет отличаться. В Рекомендациях по стандартизации Р.50.1.037–2002 рассматриваются различные варианты критерия. В представленном же программном обеспечении рассматривается только случай проверки нормальности, когда все параметры распределения оцениваются по эмпирической выборке. В упомянутых Рекомендациях установлено, что P–значения критерия для проверки нормальности в случае сложной гипотезы, когда оба параметра распределения оцениваются по эмпирической выборке, могут быть аппроксимированы функцией логнормального распределения с параметрами (0,1164; 0,5436).

Подробное исследование критерия см. в монографии Мартынова. Близок к рассматриваемому тесту критерий U^2 Уотсона (Watson), описанный в ряде зарубежных источников.

7.3.3.2.2. Критерий Андерсона–Дарлинга

При выборе весовой функции в критерии типа омега–квадрат в виде

$$\psi(t) = \frac{1}{t(1-t)}$$

получается критерий Андерсона–Дарлинга (A^2 критерий Андерсона–Дарлинга).
Практическое вычисление статистики критерия производится по формуле

$$A^2 = n\Omega^2 = -n - 2\sum_{j=1}^{n}\left\{\frac{2j-1}{2n}\ln F(\eta_j) + \left(1 - \frac{2j-1}{2n}\right)\ln\left[1 - F(\eta_j)\right]\right\},$$

где η_m, $m = 1,2,...,n$ – эмпирическая выборка, отсортированная в порядке возрастания значений вариант,
n – численность выборки,
$F(.)$ – теоретическая функция распределения.
Для каждого тестируемого теоретического распределения и каждого случая сложной гипотезы распределение статистики критерия будет отличаться. В Рекомендациях по стандартизации Р.50.1.037–2002 рассматриваются различные варианты критерия. В представленном же программном обеспечении рассматривается только случай проверки нормальности, когда все параметры распределения оцениваются по эмпирической выборке. В данных Рекомендациях установлено, что P–значения критерия для проверки нормальности в случае сложной гипотезы, когда оба параметра распределения оцениваются по эмпирической выборке, могут быть аппроксимированы функцией распределения S_U Джонсона с параметрами (–2,7057;1,7154;0,0925;0,1043). Обратим внимание пользователя на незначительное различие обозначений в программе (следуя Хану с соавт., см. главу «Введение») и в упомянутых Рекомендациях: последний и предпоследний параметры аппроксимации функцией распределения S_U Джонсона в Рекомендациях, по неизвестным нам демоническим причинам, поменяны местами.

Распределение статистики критерия для простой гипотезы теоретически исследовано Мартыновым. Описание дано в справочнике Степнова.

7.3.3.2.3. Критерий хи–квадрат Фишера

Критерий хи–квадрат Фишера (Пирсона–Фишера) является одним из старейших и самых популярных среди исследователей критериев согласия, применяемых для анализа выборок большой численности.
Критерий хи–квадрат Фишера предназначен для проверки сложных гипотез и является модификацией критерия хи–квадрат Пирсона, предназначенного для проверки простых гипотез. Вычисление статистики критерия хи–квадрат Фишера в случае проверки согласия непрерывного эмпирического распределения и непрерывного теоретического распределения производится по формуле

$$\chi^2 = \sum_{i=1}^{k}\frac{(\nu_i - nd_i p_i)^2}{nd_i p_i},$$

где ν_i, $i = 1,2,...,k$ – частоты наблюдаемых случаев в k классах,
nd_ip_i, $i = 1,2,...,k$ – соответствующие ожидаемые частоты,

p_i, $i = 1,2,...,k$ – теоретические вероятности, вычисленные по формуле плотности распределения (в данном частном случае – нормального),
k – число классов распределения,
n – общее число наблюдений, вычисляемое по формуле

$$n = \sum_{i=1}^{k} v_i,$$

d_i, $i = 1,2,...,k$ – величина классового интервала (разность соседних значений интервала); умножение на данную величину необходимо для непрерывных распределений, к которым принадлежит распределение нормальное.
При появлении интервалов с ожидаемыми частотами менее 5, по условным предпосылкам применения алгоритма, их рекомендуется объединять с соседними интервалами. Величины классовых интервалов при этом подлежат пересчету. Афифи с соавт. указывают, что некоторые ожидаемые частоты могут быть ≥ 2 (часто они располагаются на концах интервала), но при этом остальные обязательно должны быть ≥ 5. Программа имеет одно ограничение: если возникли несоответствующие интервалы, пересчета не производится, а результаты расчета данным критерием не следует воспринимать, как правильные. Нужно воспользоваться другим тестом. Данная ситуация возникает тем вернее, чем меньше численность выборки.
Статистика критерия хи–квадрат Фишера распределена как χ^2 с числом степеней свободы $k – s – 1$, где s – число оцениваемых параметров распределения. В рассматриваемом случае при проверке нормальности распределения, когда по выборке оцениваются среднее значение и дисперсия, $s = 0$, и таким образом, число степеней свободы будет $k – 3$. Здесь нужно отметить, что параметры нормального распределения для расчета теоретических вероятностей, используемых при расчете статистики рассматриваемого критерия, должны быть вычислены по эмпирическим частотам, а не по исходным выборкам. Поэтому для вычислений данных выборочных показателей используются формулы для среднего значения и дисперсии (смещенная оценка), соответственно, в следующей форме:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{k} b_i v_i,$$

$$\sigma^2 = \frac{1}{n}\left[\sum_{i=1}^{k} b_i^2 v_i - \frac{1}{n}\left(\sum_{i=1}^{k} b_i v_i\right)^2\right],$$

где b_i, $i = 1,2,...,k$ – середины классовых интервалов.
Осветим несколько нерешенных проблем, свойственных рассматриваемому критерию:

- В программе число классовых интервалов вычисляется по правилу Стержесса (см. главу «Описательная статистика»). От выбора числа классовых интервалов существенно зависит результат анализа рассматриваемым критерием, но нельзя сказать, что проблема выбора оптимального числа классов решена. По этой причине многие исследователи полагают, что использовать критерии типа хи–квадрат для обработки количественных данных нецелесообразно. Достаточно полный обзор методов выбора числа классов дан в книге Новицкого с соавт.
- На другую проблему указал проф. Воинов (цитируется по личной переписке): «... параметры должны быть оценены по эмпирическим частотам, а не по исходной выборке. Это условие необходимо, но не достаточно!!! Достаточным условием того, что критерий будет в пределе хи–квадрат с $k – s – 1$ степенью свободы и не зависеть от параметров, является то, что предельная ковариационная матрица стандартизованных частот будет такая же, как и в случае оценок, полученных по методу минимума хи–квадрат. Я не уверен, что это условие выполняется для выборочных среднего и

дисперсии по группированным данным ...». Данное утверждение может быть проверено с помощью методов, представленных в главах «Параметрическая статистика», «Непараметрическая статистика» и «Дисперсионный анализ».

- В руководствах по прикладной статистике обычно указывается, что числа классов должно быть достаточно для верной передачи характеристик эмпирической функции распределения. При этом никаких рекомендаций о проверке данного утверждения не приводится. Оно может быть проверено с помощью методов, представленных в главе «Непараметрическая статистика».

Выдача результатов включает дополнительные параметры:

- число классов,
- классовый интервал,
- середины классовых интервалов,
- численности классов,
- теоретические частоты.

Критерий представлен в книге Тюрина с соавт., работах Лемешко, Кобзаря, Рекомендациях по стандартизации Р 50.1.033–2001. Критерий *J* Ястремского, основанный на хи–квадрат, статистика которого имеет нормальное распределение, описывает Лакин. См. также работу Карлис (Karlis) с соавт. Вклад в развитие теории критериев типа хи–квадрат внесли Никулин, Мирвалиев, Воинов, Пя. Из важнейших результатов данных авторов нужно отметить группу критериев типа хи–квадрат, свободных от метода разбиения на классовые интервалы и от способа оценки неизвестных параметров распределения.

7.3.3.3. Критерии типа Эппса–Палли

В разделе рассмотрены:

- Критерий Эппса–Палли.
- Критерий Хенце–Цирклера.

Критерии типа Эппса–Палли (Epps–Pulley test) основаны на измерении расстояния эмпирической характеристической функции и модельной (теоретической) функции распределения

$$T_n = n\int_{-\infty}^{\infty}\left|\psi_n(t) - e^{-t^2/2}\right|^2 \varphi(t)dt,$$

где $\psi_n(t)$ – эмпирическая характеристическая функция,
t – нормированное отклонение,
n – численность выборки,
|.| означает модуль комплексного выражения.

Обзор критериев рассматриваемого типа, включая аппроксимации и результаты компьютерного моделирования, представлен Эппсом (Epps).

7.3.3.3.1. Критерий Эппса–Палли

Представив эмпирическую характеристическую функцию (обозначения выбраны таким образом, чтобы они совпадали с аналогичными обозначениями критерия Хенце–Цирклера)

$$\psi_n(t) = \frac{1}{n}\sum_{j=1}^{n} e^{it\left(X_j - \bar{X}\right)/S},$$

где i – мнимая единица,
S – дисперсия,

$\overline{X}$ – среднее значение выборки X_j, $j = 1,2,...,n$,
в тригонометрической форме и взяв выражение $\varphi(t)$ в виде плотности стандартного нормального распределения, несложно получить удобную формулу для вычисления статистики критерия Эппса–Палли

$$T_n = 1 + \frac{n}{\sqrt{3}} + \frac{2}{n}\sum_{j=2}^{n}\sum_{k=1}^{j-1} e^{-\frac{1}{2}(X_j - X_k)^2 / S^2} - \sqrt{2}\sum_{j=1}^{n} e^{-\frac{1}{4}(X_j - \overline{X})^2 / S^2}.$$

Согласно Хенце (Henze), *P*–значение для малых выборок берется по таблице, а для выборок численностью от 10 и выше вычисляется по формуле
$P = \Phi(z)$,
где $\Phi(z)$ – функция стандартного нормального распределения.
Величина $z = z(T_n^*)$ рассчитывается как

$$z = \gamma + \delta \log((T_n^* - \xi)/(\xi + \lambda - T_n^*)),$$

где $T_n^* = (T_n - 0{,}365/n + 1{,}34/n^2)(1 + 1{,}3/n)$,
а греческими буквами обозначены константы.
Минимальная численность выборки, анализируемой критерием Эппса–Палли, равна 4.
Максимальная численность равна 200.

См. статьи Эппса, Рекомендации по стандартизации Р.50.1.037–2002 Росстандарта России, статью Хенце (Henze). Многомерный аналог критерия Эппса–Палли представлен критерием Хенце–Цирклера.

7.3.3.3.2. Критерий Хенце–Цирклера

Существует аналог критерия Эппса–Палли, предназначенный для проверки нормальности многомерного распределения. Вычисление критерия Хенце–Цирклера (инвариантного теста Хенце–Цирклера, Henze–Zirkler test) производится по формуле

$$D_{n,\beta} = \frac{1}{n^2}\sum_{j=1}^{n}\sum_{k=1}^{n} e^{-\frac{\beta^2}{2}|Y_j - Y_k|^2} - 2(1+\beta^2)^{-d/2}\frac{1}{n}\sum_{j=1}^{n} e^{-\frac{\beta^2}{2(1+\beta^2)}|Y_j|^2} + (1+2\beta^2)^{-d/2},$$

где β – вычисляемый особым образом или задаваемый параметр,
d – размерность многомерной (d–мерной) выборки X_j, $j = 1,2,...,n$,
n – число вариант d–мерной выборки.
Многомерность эмпирической выборки при практическом вычислении в настоящем программном обеспечении означает, что она представлена таблицей чисел, строки которой являются вариантами (в данном случае – векторными) d–мерной выборки, число строк равно численности выборки, а число столбцов равно размерности («числу измерений»).
Остальные входящие в формулу параметры вычисляются как
$|Y_j - Y_k|^2 = (X_j - X_k)' S^{-1} (X_j - X_k)$,

$$|Y_j|^2 = (X_j - \overline{X})' S^{-1}(X_j - \overline{X}),$$

где S^{-1} – матрица, обратная дисперсионно–ковариационной матрице,
$\overline{X}$ – d–мерный вектор среднего значения, вычисленный по d–мерной выборке,
штрих означает операцию транспонирования.
P–значения критерия вычислены путем нормальной аппроксимации.

См. работу Свантессон (Svantesson) с соавт.

7.3.4. Критерии, основанные на регрессии

К тестам, основанным на регрессии и корреляции (иногда их называют критериями, основанными на регрессии порядковых статистик), относятся группа критериев типа Шапиро–Уилка и *D* критерий Д'Агостино.
В некоторых программных продуктах, в том числе в AtteStat, реализованы как оригинальный тест, так и различные расширения критерия Шапиро–Уилка:

- критерий Шапиро–Уилка (Shapiro–Wilk's *W* test),
- критерий Шапиро–Франсиа (Shapiro–Francia's *W′* test).

В данной программе не представлены следующие варианты:

- расширенный критерий Шапиро–Уилка для численности выборки до 2000, разработанный Ройстоном (Royston's extension of *W* for large samples),
- расширенный критерий Шапиро–Уилка для численности выборки до 5000, предложенный Рахманом и Говиндараджулу (Rahman, Govindarajulu),
- критерий Вайсберга и Бингхэма (Weisberg–Bingham's *W″* test).

Отметим, однако, что упомянутые, но нереализованные пока критерии могут быть заменены представленными тестами.

Представляют интерес исследования критериев типа Шапиро–Уилка, выполненные Райан (Ryan) и Джойнером (Joiner), Чен (Chen) и Шапиро. Обзор см. в статьях Баи (Bai) с соавт., Веррилла (Verrill) с соавт.

7.3.4.1. Критерий Шапиро–Уилка

В ряде опытов, особенно в экспериментальных и клинических биомедицинских исследованиях, часто возникает ситуация, когда численность выборки мала. Специально для проверки нормальности распределения малых, численностью от 3 до 50 вариант, выборок Шапиро (Shapiro) и Уилк (Wilk) разработали критерий. На основе формул оригинальной статьи критерий в принципе можно применять для любых по численности выборок, однако авторы табулировали константы, необходимые для вычисления статистики критерия и аппроксимации *P*–значения, только до 50 вариант.
Статистика критерия имеет вид

$$W = \frac{\left(\sum_{i=1}^{n} a_i x_i\right)^2}{\sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2},$$

где x_i, $i = 1,2,...,n$ – отсортированная в порядке возрастания выборка,
n – численность выборки,
a_i, $i = 1,2,...,n$ – константы.
В матричной форме формула вычисления констант имеет вид
$a = (m'V^{-1}V^{-1}m)^{-1/2}m'V^{-1}$,
где m и V – соответственно, вектор математических ожиданий и дисперсионно–ковариационная матрица массива упорядоченных сгенерированных выборок численностью n, распределенных по стандартному нормальному закону. Вычисление данных величин сопряжено с большими вычислительными сложностями, вызванными требованиями к объему (обычно используется от 2000 до 8000 выборок, и, если математические ожидания можно просто накапливать, для получения дисперсионно–ковариационной матрицы все выборки необходимо хранить) и адресации памяти, быстродействию. Так, в наших опытах решение задачи «в лоб» было вполне успешным, но, к сожалению, имело быстродействие,

драматичное для диалоговой системы. Методика вычислений также приводится в более поздних публикациях Ройстона (J.P. Royston) и Ройстона (P. Royston).
Поэтому практически вычисление статистики оригинального критерия производится по формуле, пригодной для быстрых вычислений,

$$W = \frac{\left(\sum_{i=1}^{k} a_{n-i+1}\left(x_{n-i+1} - x_i\right)\right)^2}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2},$$

где $k = n / 2$, если n – четное,
$k = (n - 1) / 2$, если n – нечетное,
a_{n-i+1}; $i = 1,2,...,k$; $n = 3,4,...,50$, – табулированные константы.
Для вычисления P–значений критерия применяется нормальная аппроксимация. Величина

$$Z = \gamma_n + \eta_n \ln\frac{W - \varepsilon_n}{1 - W},$$

где γ_n, η_n, ε_n – табулированные константы для соответствующих значений n,
распределена нормально как $N(0,1)$.

Другие аппроксимации, действительные для численности выборок до 5000, получены в работе Ройстона (P. Royston, 1993). Критерий реализован на основе монографии Хана с соавт. (Hahn et al., имеется русский перевод). См. также справочник Степнова. Ройстон (J.P. Royston) в 1983 году представил критерий H – многомерный аналог критерия Шапиро–Уилка. О критерии H Ройстона см. также работу Свантессон (Svantesson) с соавт. Очень простое многомерное обобщение критерия Шапиро–Уилка под наименованием маргинального алгоритма (marginals algorithm) предложили Петерсон (Peterson) с соавт.

7.3.4.2. Критерий Шапиро–Франсиа

Шапиро (Shapiro) и Франсиа (Francia) предположили, что для больших выборок статистика критерия W может быть вычислена менее трудоемко, чем это сделано в критерии Шапиро–Уилка. Она имеет другое обозначение, но похожую запись

$$W' = \frac{\left(\sum_{i=1}^{n} b_i x_i\right)^2}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2},$$

где x_i, $i = 1,2,...,n$ – отсортированная в порядке возрастания выборка,
n – численность выборки,
b_i, $i = 1,2,...,n$ – константы.
В матричной форме формула вычисления констант имеет совсем простой вид
$b = (m'm)^{-1/2}m$,
где m – вектор математических ожиданий, вычисленный на основе упорядоченных сгенерированных выборок численностью n, распределенных по стандартному нормальному закону. Определение данной величины сопряжено с большими вычислительными сложностями, вызванными требованиями к быстродействию компьютера. Поэтому авторы теста воспользовались тем, что ранее Блом (Blom, см. Дэйвида) записал простую в вычислении оценку компонент вектора математических ожиданий

$$\tilde{m}_i = \Psi\left[(i - 3/8)/(n + 1/4)\right], i = 1,2,...,n,$$

где $\Psi(.)$ – функция, обратная функции стандартного нормального распределения.

Статистика критерия не относится к какому–либо стандартному типу распределения, поэтому Ройстон (J.P. Royston, 1983) для практических вычислений предложил ее трансформацию с последующей аппроксимацией по стандартному нормальному закону. Другие аппроксимации, также действительные для численности выборок до 5000, даны в работе Ройстона (P. Royston, 1993).

7.3.4.3. Критерий Д'Агостино

D критерий Д'Агостино (D'Agostino's *D* test) построен, как и критерий Шапиро–Уилка, на порядковых статистиках. Вычисление статистики критерия производится по формуле

$$D = \frac{\sum_{i=1}^{n}\left(i - \frac{n+1}{2}\right)x_i}{n^2 s},$$

где x_i, $i = 1,2,...,n$ – отсортированная в порядке возрастания выборка,
n – численность выборки,

$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$ – смещенная оценка дисперсии,

где $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ – выборочное среднее значение.

Практически значимость может вычисляться посредством нормальной аппроксимации критического значения критерия. При этом модифицированная статистика

$$\frac{D - ED}{\sqrt{DD}},$$

где $ED = \frac{(n-1)}{2\sqrt{(2n\pi)}}\frac{\Gamma(n/2-1/2)}{\Gamma(n/2)} \approx \left(2\sqrt{\pi}\right)^{-1} \approx 0{,}28209479$ – математическое ожидание,

$\sqrt{DD} = \left(\frac{12\sqrt{3} - 27 + 2\pi}{24n\pi}\right)^{1/2} \approx 0{,}02998598/\sqrt{\mathrm{n}}$ – стандартное отклонение.

распределена по стандартному нормальному закону. По этой причине *D* критерий Д'Агостино полагается более удобным в вычислении, чем критерий Шапиро–Уилка, требующий для своего вычисления либо таблиц, либо довольно сложных трудоемких аппроксимаций, связанных с объемными вычислениями.

Формулы взяты из источников: Д'Агостино (D'Agostino, 1971), Донг (Dong) с соавт. и Уайт (White) с соавт. Во втором источнике асимптотические формулы для математического ожидания и стандартного отклонения записаны неправильно, причем опечатка в формуле для стандартного отклонения идет из оригинальной работы. В третьем источнике асимптотическая формула для стандартного отклонения не приводится. Мы исправили данные формулы и приводим их полностью.

7.3.5. Критерии моментов

Существует группа критериев, которые позволяют оценить отклонение некоторых параметров эмпирического распределения (обычно это – коэффициент асимметрии, эксцесс или и тот, и другой параметр одновременно) от тех же параметров нормального распределения. Подробнее о данных параметрах эмпирической выборки см. главу

«Описательная статистика». Рассматриваемые критерии принадлежат к группе критериев, основанных на обычных и абсолютных моментах распределения. По результатам применения данных критериев нельзя делать заключение о соответствии тестируемой выборки нормальному распределению. Данными критериями можно лишь проверить, что тестируемые параметры эмпирической выборки принимают определенные значения, соответствующие нормальному распределению.
Наиболее распространены следующие критерии, основанные на моментах распределения:

- критерий коэффициента асимметрии (третий нормированный центральный момент),
- критерий эксцесса (четвертый нормированный центральный момент),
- критерий Жарка–Бера, построенный на идее одновременного анализа коэффициента асимметрии и эксцесса,
- критерий Гири (первый нормированный центральный абсолютный момент),
- многомерный критерий асимметрии Мардиа,
- многомерный критерий эксцесса Мардиа.

Напомним, что центральные выборочные моменты определяются формулами

$$m_k = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^k, k = 1,2,...,$$

где x_i, $i = 1,2,...,n$ – эмпирическая выборка,
n – численность выборки,

$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ – выборочное среднее значение,

k – порядок момента.
Центральные абсолютные выборочные моменты определяются формулами

$$\beta_k = \frac{1}{n}\sum_{i=1}^{n}| x_i - \bar{x} |^k, k = 1,2,....$$

Абсолютные моменты четных порядков совпадают с обычными моментами. Первый центральный абсолютный момент называется средним арифметическим отклонением. Наряду со средним квадратическим отклонением, данный показатель может применяться в качестве характеристики рассеяния.
Хорошо проработаны многомерные аналоги критериев коэффициента асимметрии и эксцесса – критерии Мардиа, представленные в данном программном обеспечении. Многомерный критерий Мардиа–Фостера идейно близок к составным тестам типа Жарка–Бера (одновременно тестируются асимметрия и эксцесс), в настоящем программном обеспечении не реализован.
Применение критерия коэффициента асимметрии и критерия эксцесса рекомендуется для проверки отклонения от нормальности, например, при решении вопроса о применении критерия Стьюдента, представленного в главе «Параметрическая статистика».

7.3.5.1. Критерий коэффициента асимметрии

Коэффициент асимметрии (skewness) характеризует несимметричность распределения случайной величины. Для нормального распределения коэффициент асимметрии равен нулю. Коэффициент асимметрии – величина, не зависящая от выбора начала отсчета и от единиц измерения случайной величины. Выборочный коэффициент асимметрии (sample skewness) может вычисляться по формуле выборочных моментных отношений

$$b_1 = \frac{m_3}{S^3},$$

где S^2 – оценка выборочной дисперсии, вычисляемая по формуле

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2,$$

где x_i, $i = 1,2,...,n$ – варианты эмпирической выборки,

$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ – выборочное среднее значение.

n – численность выборки,

$m_3 = \frac{n}{(n-1)(n-2)}\sum_{i=1}^{n}(x_i - \bar{x})^3$ – выборочная оценка 3–го центрального момента.

Запишем модифицированную статистику

$$B_1 = \frac{\sqrt{n(n-1)}}{n-2} b_1.$$

Тогда статистика B_1 для большой численности выборки распределена асимптотически нормально с нулевым средним и дисперсией

$$DB_1 = \frac{6n(n-1)}{(n-2)(n+1)(n+3)}.$$

Минимальная численность выборки, анализируемой критерием коэффициента асимметрии, равна 3.

Описание см. у Крамера, Ван дер Вардена. В литературе имеются и другие формы записи статистики критерия, а также ее аппроксимации. См. Большева с соавт., Степнова, Стенгоса (Stengos) с соавт. О коэффициенте асимметрии см. также главу «Описательная статистика». Исследователями предложены следующие варианты критерия коэффициента асимметрии:

- критерий асимметрии Д'Агостино (D'Agostino's test for skewness), подробно представленный в статье Д'Агостино с соавт. (1990),
- критерий g_1 Фишера (Fisher g statistics for skewness), описанный там же.

7.3.5.2. Критерий эксцесса

Эксцесс (kurtosis, excess) характеризует степень выраженности хвостов распределения – частоту появления удаленных от среднего значений. Для нормального распределения эксцесс равен трем, поэтому при вычислении эксцесса от полученного значения часто отнимают число три, чтобы показать, насколько эксцесс эмпирической выборки отличается от эксцесса нормального распределения. Эксцесс – величина, не зависящая от выбора начала отсчета и от единиц измерения случайной величины. Выборочный эксцесс (sample kurtosis) может вычисляться по формуле выборочных моментных отношений

$$b_2 = \frac{m_4}{S^4},$$

где S^2 – оценка выборочной дисперсии, вычисляемая по формуле

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2,$$

где x_i, $i = 1,2,...,n$ – варианты эмпирической выборки,

$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ – выборочное среднее значение.

n – численность выборки,

$$m_4 = \frac{n(n+1)\sum_{i=1}^{n}(x_i-\bar{x})^4 - 3(n-1)\left\{\sum_{i=1}^{n}(x_i-\bar{x})^2\right\}^2}{(n-1)(n-2)(n-3)}$$ – выборочная оценка 4–го центрального момента.

Запишем модифицированную статистику

$$B_2 = \frac{n-1}{(n-2)(n-3)}\left[(n+1)b_2+6\right].$$

Тогда статистика B_2 для большой численности выборки распределена асимптотически нормально с нулевым средним и дисперсией

$$DB_2 = \frac{24n(n-1)^2}{(n-3)(n-2)(n+3)(n+5)}.$$

Минимальная численность выборки, анализируемой критерием эксцесса, равна 4.

Описание см. у Крамера, Ван дер Вардена. В литературе имеются и другие аппроксимации статистики критерия. См. Большева с соавт., Степнова. Об эксцессе см. также главу «Описательная статистика». Исследователями предложены следующие варианты критерия эксцесса, рекомендуемые в конкретных указанных случаях:

- критерий эксцесса Д'Агостино (D'Agostino's test for kurtosis), подробно представленный в статье Д'Агостино с соавт. (1990 г.) и ставший уже классическим,
- критерий g_2 Фишера (Fisher *g* statistics for kurtosis), описанный там же,
- критерий Анскомба–Глина (Anscombe–Glynn kurtosis test) – тестирование на нормальность против асимметричных распределений или распределений с тяжелыми хвостами,
- *I* критерий Мартинеса–Иглевича (Martinez–Iglewicz *I* test) – тестирование на нормальность против других альтернативных распределений с тяжелыми хвостами.

Данные критерии не представлены в настоящем программном обеспечении и упомянуты для полноты информации.

7.3.5.3. Критерий Жарка–Бера

Известным представителем составных тестов является широко применяемый (и широко критикуемый) критерий Жарка–Бера (Jarque–Bera test, он же Bowman–Shenton K^2 test). С помощью данного критерия производится одновременный анализ коэффициента асимметрии и эксцесса. Статистика критерия вычисляется по формуле

$$J = n\left(\frac{b_1^2}{6} + \frac{(b_2-3)^2}{24}\right),$$

где b_1 – коэффициент асимметрии,
b_2 – эксцесс,
n – численность выборки.
В соответствии с требованиями алгоритма, коэффициент асимметрии вычисляется по формуле
$b_1 = k_3 / S^3$,
где S^2 – оценка выборочной дисперсии, вычисляемая по формуле

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2,$$

где x_i, i = 1,2,...,n – варианты эмпирической выборки,

$\bar{x}$ – выборочное среднее значение.

$$k_3 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3.$$

Эксцесс вычисляется по формуле

$$b_2 = \frac{k_4}{S^4},$$

где $$k_4 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4.$$

Статистика критерия для большой численности выборки распределена асимптотически как χ^2 с параметром, равным 2.

Критерии описаны во множестве оригинальных источников. Гел и Гаствирт представили робастный вариант критерия (The Gel–Gastwirth robust Jarque–Bera test). См. также обзор Дурник (Doornik) с соавт., Ромао (Romao) с соавт., книгу Селезнева с соавт., справочник Степнова. Составной критерий Д'Агостино–Пирсона (D'Agostino–Pearson test) был подробно представлен в статье Д'Агостино с соавт. (1990 г.), но в настоящее время дезавуирован из–за обнаруженных теоретических проблем. Отечественным ГОСТом определен так называемый составной критерий, представляющий собой совокупность двух тестов, одним из которых является вариант критерия Гири.

7.3.5.4. Критерий Гири

Гири предложил серию критериев, построенных на соотношениях для центральных абсолютных моментов. Вместо обычного критерия эксцесса может применяться критерий Гири (Geary's kurtosis test), построенный на соотношении первого центрального абсолютного момента:

$$d = \frac{1}{nS}\sum_{i=1}^{n}|x_i - \bar{x}|,$$

где S^2 – смещенная оценка выборочной дисперсии, вычисляемая по формуле

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2,$$

где x_i, $i = 1,2,...,n$ – варианты эмпирической выборки,

$\bar{x}$ – выборочное среднее значение.

n – численность выборки.

Практически значимость может вычисляться посредством нормальной аппроксимации критического значения критерия. При этом статистика d распределена нормально с математическим ожиданием $\sqrt{2/\pi}$ и дисперсией $(1 – 3 / \pi) / n$.

Критерий изучен Большевым с соавт., Д'Агостино (D'Agostino) и Розман (Rosman), Чо (Cho) с соавт., Уолпоул (Walpole) с соавт. Родственным описанному тесту является критерий Бонетта–Сайера (Bonett–Seier test).

7.3.5.5. Критерий асимметрии Мардиа

Многомерный аналог критерия коэффициента асимметрии предложен Мардиа. Статистика критерия вычисляется по формуле

$$b_{1,d} = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left[(X_i - \bar{X})'S^{-1}(X_j - \bar{X})\right]^3,$$

где d – размерность многомерной (d–мерной) выборки X_j, j = 1,2,...,n,
n – число вариант d–мерной выборки,
S^{-1} – матрица, обратная дисперсионно–ковариационной матрице,
$\bar{X}$ – d–мерный вектор среднего значения, вычисленный по d–мерной выборке,
штрих означает операцию транспонирования.
Для практического исследователя–расчетчика многомерность эмпирической выборки означает, что она представлена таблицей чисел, строки которой являются вариантами (в данном случае – векторными) d–мерной выборки, число строк равно численности выборки, а число столбцов равно размерности («числу измерений»).

Статистика $\frac{n}{6}b_{1,d}$ распределена асимптотически как χ^2 с параметром $d(d + 1)(d + 2) / 6$.

О критериях Мардиа см. оригинальные работы (Mardia), а также статью и библиографию Канкайнена (Kankainen) с соавт. (Taskinen, Oja), справочник Родионова с соавт.

7.3.5.6. Критерий эксцесса Мардиа

Многомерный аналог критерия эксцесса предложен Мардиа. Статистика критерия вычисляется по формуле

$$b_{2,d} = \frac{1}{n}\sum_{i=1}^{n}\left[(X_i - \bar{X})'S^{-1}(X_i - \bar{X})\right]^2,$$

где d – размерность многомерной (d–мерной) выборки X_j, j = 1,2,...,n,
n – число вариант d–мерной выборки,
S^{-1} – матрица, обратная дисперсионно–ковариационной матрице,
$\bar{X}$ – d–мерный вектор среднего значения, вычисленный по d–мерной выборке,
штрих означает операцию транспонирования.
Для исследователя многомерность эмпирической выборки означает, что она представлена таблицей чисел, строки которой являются вариантами (в данном случае – векторными) d–мерной выборки, число строк равно численности выборки, а число столбцов равно размерности («числу измерений»).
Статистика $b_{2,d}$ распределена асимптотически нормально со средним $d(d + 2)$ и дисперсией $8d(d + 2) / n$.

О критериях Мардиа см. оригинальные работы (Mardia), а также статью и библиографию Канкайнена (Kankainen) с соавт. (Taskinen, Oja), справочник Родионова с соавт.

7.3.6. Информационные критерии

Информационные критерии согласия основаны на информационной мере – энтропии (см. «Информационный анализ»). Они основаны на том научном факте, что энтропия непрерывного распределения максимальна, если распределение нормальное.
Наиболее известным является классический критерий Васичека (Vasicek's test). Другие методы – это:

- критерий ван Эса (van Es' test),
- критерий Корреа (Correa's test),
- модифицированный критерий Васичека – критерий Вичорковкого–Гржегоржевского

(Wieczorkowski–Grzegorzewski's test).

Помимо оригинальных работ, все данные критерии описаны в обзоре Эстебана (Esteban) с соавт.

7.3.6.1. Критерий Васичека

Статистика критерия Васичека (Vasicek's test) вычисляется по формуле

$$K_{mn} = \frac{n}{2mS}\left\{\prod_{i=1}^{n}\left(x_{i+m} - x_{i-m}\right)\right\}^{1/n},$$

где x_i, $i = 1,2,...,n$ – варианты упорядоченной (от меньшего значения к большему значению) эмпирической выборки, причем условились, что при индексе варианты в данной формуле $(i + m) > n$ индекс берется n, при индексе $(i – m) < 1$ индекс берется 1,
m – ширина окна, положительное наименьшее целое значение из интервала от 1 до $(n – 1) / 2$,
n – численность выборки,
S^2 – смещенная оценка выборочной дисперсии, вычисляемая по формуле

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2,$$

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

где – выборочное среднее значение.
Гипотеза о нормальности распределения не отклоняется на заданном уровне значимости при выполнении условия
$K_{mn} \geq K^*$,
где K^* – критическое значение, взятое из таблицы, вычисленной методом компьютерного моделирования.

Таблица критических значений в оригинальном источнике отличается чрезмерной лаконичностью, поэтому мы используем более подробную таблицу, вычисленную Эстебаном (Esteban) с соавт. Как и в оригинальной статье Васичека, вычисления выполнены для n = 1,2,...,50, поэтому при большей численности выборки критерий не применяется. Кроме того, таблицы получены для значений $1 \leq m \leq 9$ с учетом представленного выше правила выбора ширины окна и только для уровня значимости 0,05, что учтено в программе. См. также книги Кобзаря, Тику (Tiku) с соавт., статью Мудхолкара (Mudholkar) и Тиань (Tian). В последнем источнике разъясняется роль такого важного параметра алгоритма критерия Васичека, как ширина окна m. При конкретных альтернативных распределениях эмпирической выборки и фиксированной ее численности максимальная мощность критерия (см. главу «Введение») достигается при определенной ширине окна.

7.3.7. Графические методы

Простейшим из графических методов является глазомерный метод, когда визуально сравниваются график функции распределения или плотности распределения эмпирической и график наложенной на нее теоретической. В практической реализации графических методов может оказаться полезным использование инструмента «Гистограмма», представленного в главе «Описательная статистика».

О чтении гистограмм, в числе огромного числа источников, см. монографию под ред. Кумэ.

Кроме того, некоторые представленные в программе методы имеют очевидную графическую интерпретацию. См., например, статьи Мэйджа (Mage), Аймэна (Iman).

7.3.7.1. Глазомерный метод

Простейшим из графических методов является так называемый глазомерный метод, когда визуально сравниваются график плотности распределения эмпирической и график наложенной на нее соответствующей теоретической функции. Сравнение производится пользователем, который играет в данном случае роль эксперта.
Выдача результатов анализа рассматриваемым методом в программе включает параметры:

- число классов,
- номера классов,
- численности классов,
- теоретические частоты нормального распределения,
- диаграмму, на которой гистограмма представляет собой отображение эмпирического распределения, а точечная диаграмма со значениями, соединенными сглаживающими линиями, отображает теоретическое нормальное распределение.

Число классов может быть задано либо вычислено автоматически, как указано в разделе Работа с программным обеспечением. См. также замечания в разделе, посвященном критерию хи–квадрат. О вычислении оптимального числа классов см. главу «Описательная статистика».

7.3.8. Байесовские критерии

Обсуждение Байесовских критериев см. в работах Шпигельхальтера (Spiegelhalter) 1977 и 1980 гг.

Список использованной и рекомендуемой литературы

1. Ahmad I.A. Modification of some goodness of fit statistics II: two–sample and symmetry testing // Sankhya: The Indian Journal of Statistics, 1996, vol. 58, series A, part 3, pp. 464–472.
2. Ahmad I.A., Mugdadi A.R. Testing normality using the kernel methods // Journal of Nonparametric Statistics, 2003, vol. 15, no. 3, pp. 273–288.
3. Anderson T.W., Darling D.A. A test of goodness of fit // Journal of the American Statistical Association, 1954, vol. 49, pp. 765–769.
4. Anderson T.W., Darling D.A. Asymptotic theory of certain «Goodness of fit» criteria based on stochastic processes // The Annals of Mathematical Statistics, 1952, vol. 23, no. 2, pp. 193–212.
5. Arizono I., Ohta H. A test for normality based on Kullback–Leibler information // The American Statistician, February 1989, vol. 43, no. 1, pp. 20–22.
6. Babu G.J., Rao C.R. Goodness–of–fit tests when parameters are estimated // The Indian Journal of Statistics, 2004, vol. 66, part 1, pp. 63–74.
7. Bai Z.D., Chen L. Weighted W test for normality and asymptotics a revisit of Chen–Shapiro test for normality // Journal of Statistical Planning and Inference, 1 May 2003, vol. 113, no. 2, pp. 485–503.
8. Baringhaus L., Danchke R., Henze N. Recent and classical tests for normality – A comparative study // Communications of Statistics – Simulation, 1989, vol. 18, pp. 363–379.
9. Biining H. Kolmogorov–Smirnov and Cramer–von Mises type two–sample tests with various weight functions // Communications in Statistics: Simulation and Computation, 2001, vol. 30, no. 4, pp. 847–866.

10. Brown B.M., Hettmannsperger T.P. Normal scores, normal plots, and tests for normality // Journal of the American Statistical Association, December 1996, vol. 91, no. 436, pp. 1668–1675.
11. Brys G., Hubert M., Struyf A. A robustification of the Jarque–Bera test of normality // COMPSTAT 2004 – Proceedings in Computational Statistics, 16th Symposium Held in Prague, Czech Republic, 2004. – Physica–Verlag / Springer, 2004, pp. 729–736.
12. Cabana A. Transformations of the empirical measure and Kolmogorov–Smirnov tests // The Annals of Statistics, 1996, vol.25, no. 5, pp. 2020–2035.
13. Chen E.H. The power of the Shapiro–Wilk W test for normality in samples contaminated normal distribution // Journal of the American Statistical Association, December 1971, vol. 66, no. 336, pp. 760–762.
14. Cho D., Im K.S. Test of normality using Geary's skewness and kurtosis statistics // Faculty Working Papers, 2002, No. 02–32. – Department of Economics, College of Business and Administration, University of Central Florida.
15. Conover W.J. Practical nonparametric statistics. – New York, NY: John Wiley & Sons, 1999.
16. Csorgo M., Seshadri V., Yalovsky M. Some exact tests for normality in the presence of unknown parameters // Journal of the Royal Statistical Society: Series B (Statistical Methodology), 1973, vol. 35, no. 3, pp. 507–522.
17. D'Agostino R., Pearson E.S. Tests for departure from normality. Empirical results for the distributions of b_2 and $\sqrt{b_1}$ // Biometrika, December 1973, vol. 60, no. 3, pp. 613–622.
18. D'Agostino R.B. An omnibus test of normality for moderate and large size samples // Biometrika, August 1971, vol. 58, no. 2, pp. 341–348.
19. D'Agostino R.B. Goodness–of–fit techniques / Ed. by R.B. D'Agostino, M.S. Stephens. – New York, NY: Marcel Dekker, 1986.
20. D'Agostino R.B. Simple compact portable test of normality: Geary's test revisited // Psychological Bulletin, 1970, vol. 74, pp. 138–140.
21. D'Agostino R.B. Small sample probability points for the D test of normality // Biometrika, April 1972, vol. 59, no. 1, pp. 219–221.
22. D'Agostino R.B. Transformations to normality of the null distribution of g_1 // Biometrika, December 1970, vol. 57, no. 3, pp. 679–681.
23. D'Agostino R.B., Belanger A., D'Agostino R.B.Jr. A suggestion for using powerful and informative tests of normality // The American Statistician, November 1990, vol. 44, no. 4, pp. 316–321.
24. D'Agostino R.B., Rosman B. The power of Geary's test of normality // Biometrika, April 1974, vol. 61, no. 1, pp. 181–184.
25. D'Agostino R.B., Tietjen G.L. Approaches to the null distribution of $\sqrt{b_1}$ // Biometrika, April 1973, vol. 60, no. 1, pp. 169–173.
26. D'Agostino R.B., Tietjen G.L. Simulation probability points of b_2 for small samples // Biometrika, December 1971, vol. 58, no. 3, pp. 669–672.
27. Dallal G.E., Wilkinson L. An analytic approximation to the distribution of Lilliefors's test statistic for normality // The American Statistician, November 1986, vol. 40, no. 4, pp. 294–296.
28. Davis C.S., Stephens M.A. Algorithm AS 248: Empirical distribution function goodness–of–fit tests // Journal of the Royal Statistical Society: Series C (Applied Statistics), 1989, vol. 38, no. 3, pp. 535–543.
29. De Wet T. Goodness–of–fit tests for location and scale families based on a weighted L_2 – Wasserstein distance // Sociedad de Estadistica e Investigacion Operativa Test, 2002, vol. 11, no. 1, pp. 89–107.

30. De Wet T., Venter J.H. Asymptotic distributions of certain test criteria of normality // South African Statistical Journal, 1972, vol. 6, pp. 135–149.
31. Del Barrio E. Tests of goodness of fit based on the L_2–Wasserstein distance / E. del Barrio, J.A. Cuesta–Albertos, C. Matran et al. // The Annals of Statistics, 1999, vol. 27, no. 4, pp. 1230–1239.
32. Del Barrio E., Cuesta–Albertos J.A., Matran C. Contributions of empirical and quantile processes to the asymptotic theory of goodness–of–fit tests // Sociedad de Estadistica e Investigacion Operativa Test, 2000, vol. 9, no. 1, pp. 1–96.
33. Dong L.B., Giles D.E.A. An empirical likelihood ratio test for normality // Econometrics Working Paper EWP0401. Department of Economics, University of Victoria, Canada, 2004.
34. Doornik J.A. Hansen H. An omnibus test for univariate and multivariate normality // Working paper. Nuffield College, Oxford, 1994.
35. Ducharme G.R., de Micheaux P.L. Goodness–of–fit tests of normality for the innovations in ARMA models // Journal of Time Series Analysis, May 2004, vol. 25, no. 3, pp. 373–395.
36. Ducharme G.R., Frichot B. Quasi most powerful invariant goodness–of–fit tests // Scandinavian Journal of Statistics: Theory and applications, June 2003, vol. 30, no. 2, pp. 399–414.
37. Dufour J.–M. Simulation–based finite sample normality tests in linear regressions / J.–M. Dufour, A. Farhat, L. Gardiol et al. // The Econometrics Journal, 1998, vol. 1, no. 1, pp. 154–173.
38. Dyer A.R. Comparisons of tests for normality with a cautionary note // Biometrika, April 1974, vol. 61, no. 1, pp. 185–189.
39. Epps T.W. Tests for location–scale families based on the empirical characteristic function // Metrika, September 2005, vol. 62, no. 1, pp. 99–114.
40. Epps T.W., Pulley L.B. A test for normality based on the empirical characteristic function // Biometrika, December 1983, vol. 70, pp. 723–726.
41. Esteban M.D. Monte Carlo comparison of four normality tests using different entropy estimates / M.D. Esteban, M.E. Castellanos, D. Morales et al. // Communications in Statistics: Simulation and Computation, 2001, vol. 30, no. 4, pp. 761–786.
42. Feltz C.J., Goldin G.A. Partition–based goodness–of–fit tests on the line and the circle // Australian & New Zealand Journal of Statistics, June 2001, vol. 43, no. 2, pp. 207–220.
43. Filliben J.J. The probability plot correlation coefficient test for normality // Technometrics, 1975, vol. 17, no. 1, pp. 111–117.
44. Foirentini G., Sentana E., Calzolari G. On the validity of the Jarque–Bera normality test in conditionally heteroskedastic dynamic regression models // CEMFI Working Paper No. 0306, January 2003, Madrid.
45. Freund J.E. Mathematical Statistics. – Prentice–Hall, 1992.
46. Gastwirth J.L., Owens M.E.B., On classical tests of normality // Biometrika, April 1977, vol. 64, no. 1, pp. 135–139.
47. Geary R.C. Testing for normality // Biometrika, December 1947, vol. 34, no. 3/4, pp. 209–242.
48. Geary R.C. Tests de la normalite // Annales de l'institut Henri Poincare, 1956, vol. 15, no. 1, pp. 35–65.
49. Geary R.C. The ratio of the mean deviation to the standard deviation as a test of normality // Biometrica, 1935, vol. 27, pp. 310–332.
50. Giles D.E.A. A saddlepoint approximation to the distribution function of the Anderson–Darling test statistic // Communications in Statistics: Simulation and Computation, 2001, vol. 30, no. 4, pp. 899–906.
51. Glivenko V. Sulla determinazione empirica di probabilita // Giornale dell'Istituto Italiano degli

Attuari, 1933, vol. 4, no. 1. pp. 92–99.
52. Gokhale D.V. On entropy–based goodness–of–fit tests // Computational Statistics & Data Analysis, March 1983, vol. 1, pp. 157–165.
53. Gosh S. A new graphical tool to detect non–normality // Journal of the Royal Statistical Society: Series B (Statistical Methodology), 1996, vol. 58, no. 4, pp. 691–702.
54. Guidance for data quality assessment. Practical methods for data analysis. EPA QA/G–9. – Washington, DC: United States Environmental Protection Agency, 2000.
55. Gupta A.K., Chen T. Goodness–of–fit tests for the skew–normal distribution // Communications in Statistics: Simulation and Computation, 2001, vol. 30, no. 4, pp. 907–930.
56. Hahn G.J., Shapiro S.S. Statistical models in engineering. – New York, NY: John Wiley & Sons, 1994.
57. Hall P., Welsh A.H. A test for normality based on the empirical characteristic function // Biometrika, August 1983, vol. 70, no. 2, pp. 485–489.
58. Hall P., Welsh A.H. Amendments and corrections: A test for normality based on the empirical characteristic function // Biometrika, December 1984, vol. 71, no. 3, p. 655.
59. Hassan A.S. Goodness–of–fit for the generalized exponential distribution // InterStat (Statistics on the Internet), July 2005, No. 1.
60. Hegazy Y.A.S., Green J.R. Some new goodness–of–fit tests using order statistics // Journal of the Royal Statistical Society: Series C (Applied Statistics), 1975, vol. 24, no. 3, pp. 299–308.
61. Henze N. An approximation to the limit distribution of the Epps–Pulley test statistic for normality // Metrica, December 1990, vol. 37, no. 1, pp. 7–18.
62. Henze N., Zirkler B. A class of invariant consistent tests for multivariate normality // Communications in Statistics: Theory and Methods, 1990, vol. 19, no. 10, pp. 3595–3617.
63. Iman R.L. Graphs for use with the Lilliefors test for normality and exponential distributions // The American Statistician, May 1982, vol. 36, no. 2, pp. 109–112.
64. Jarque C.M., Bera A.K. A test for normality of observation and regression residuals // International Statistical Review, 1987, vol. 55, pp. 163–172.
65. Jarque C.M., Bera A.K. Efficient tests for normality, heteroscedasticity and serial independence of regression residuals // Economic Letters, 1980, vol. 6, pp. 255–259.
66. Kac M. On deviations between theoretical and empirical distributions // Proceedings of the National Academy of Sciences USA, May 1949, vol. 35, no. 5, pp. 252–257.
67. Kac M., Kiefer J., Wolfowitz J. On tests of normality and other tests of goodness of fit based on distance methods // The Annals of Mathematical Statistics, 1955, v. 26, pp. 189–211.
68. Kankainen A., Oja H., Taskinen S. On Mardia's tests of multinormality // Theory and applications of recent robust methods / Ed. by M. Hubert, G. Pison, A. Stryuf et al. – Basel: Birkhauser, 2003.
69. Karlis D., Xekalaki E. A simulation comparison of several procedures for testing the Poisson assumption // The Statistician, 2000, vol. 49, part 3, pp. 355–382.
70. Kiefer J. K–sample analogues of the Kolmogorov–Smirnov and Cramer–von Mises tests // The Annals of Mathematical Statistics, 1959, vol. 30, pp. 420–447.
71. Kim N., Bickel P.J. The limit distribution of a test statistic for bivariate normality // Statistica Sinica, 2003, vol. 13, pp. 327–349.
72. Klar B. Klassische und neue statistishe Anpassungstests. Zur Erlangung des akademischen Grades eines Doctors der Naturwissenschaften. Universitat Karlsruhe, 1998.
73. Kolmogoroff A.N. Sulla determinazione empirica di una legge di distribuzione // Giornale dell'Istituto Italiano degli Attuari, 1933, vol. 4, no. 1, pp. 83–91.
74. Koziol J.A. Assessing multivariate normality: A compendium // Communications in Statistics

– Theory and Methods, 1986, vol. 15, pp. 2763–2783.
75. Krumbholz W., Lassahn R. Exact percentage points for the Kolmogorov test on truncated versions of known continuous distributions with unknown truncation parameters // Statistical Papers, 1999, vol. 40, pp. 221–231.
76. L'Ecuyer P. SSJ User's Guide. Package gof. Goodness–of–fit test statistics. – Universite de Montreal, 2006.
77. L'Ecuyer P., Cordeau J.–F., Compagner A. Entropy–based tests for random number generators // unpublished manuscript, 1997. Based on paper L'Ecuyer P., Compagner A., Cordeau J.–F. Entropy tests for random number generators // Les cahiers du GERAD, Septembre 1996, no. G–96–41.
78. LaRiccia V.N. Asymptotical chi–squared distributed tests of normality for type II censored samples // Journal of the American Statistical Association, December 1986, vol. 81, no. 396, pp. 1026–1031.
79. Lassahn R. Die exakte Berechnung der Quantile des Kolmogoroffschen Anpassungstests auf Gleichverteilung mit Hilfe der Steck–Determinante // Discussion Papers in Statistics and Quantitative Economics, 1996, Nr. 70.
80. Lassahn R. Exakte Quantile einiger Anpassungstests vom Kolmogoroff–Smirnowschen Typ im Fall nicht vollig spezifizierter Verteilungshypothesen. Dissertation an der Universitat der Bundeswehr, Hamburg, 1999.
81. Lee Y.H.Jr. Fisher information test of normality. Ph.D. dissertation. ETD–82198–9530. – Virginia Polytechnic Institute, USA, 1998.
82. Liang J.J., Bentler P.M. A t–distribution plot to detect non–multinormality // Computational Statistics & Data Analysis, 1999, vol. 30, pp. 31–44.
83. Lilliefors H.W. Corrigenda: On the Kolmogorov–Smirnov test for normality with mean and variance unknown // Journal of the American Statistical Association, December 1969, vol. 64, no. 328, pp. 1702.
84. Lilliefors H.W. On the Kolmogorov–Smirnov test for normality with mean and variance unknown // Journal of the American Statistical Association, June 1967, vol. 62, no. 318, pp. 399–402.
85. Lin C.–C., Mudholkar G.S. A simple test for normality against asymmetric alternatives // Biomatrika, August 1980, vol. 67, no. 2, pp. 455–461.
86. Linnet K. Testing normality of transformed data // Journal of the Royal Statistical Society: Series C (Applied Statistics), 1988, vol. 37, no. 2, pp. 180–186.
87. Locke C., Spurrier J.D. The use of U–statistics for testing normality against nonsymmetric alternatives // Biometrika, April 1976, vol. 63, no. 1, pp. 143–147.
88. Looney S.W. How to use tests for univariate normality to assess multivariate normality // The American Statistician, February 1995, vol. 49, no. 1, pp. 64–70.
89. Lund U., Jammalamadaka S.R. An entropy–based test for goodness of fit statistic for the von Mises distribution // Journal of Statistical Computation and Simulation, 2000, vol. 67, pp. 319–332 // InterStat (Statistics on the Internet), January 1999, No. 1.
90. Mage D.T. An objective graphical method for testing normal distributional assumptions using probability plots // The American Statistician, May 1982, vol. 36, no. 2, pp. 116–120.
91. Mardia K.V. Application of some measures of multivariate skewness and kurtosis in testing normality and robustness studies // Sankhya: The Indian Journal of Statistics, 1974, vol. 36, series B, pt. 2, pp. 115–128.
92. Mardia K.V. Assessment of multinormality and the robustness of Hotelling's T–squared test // Journal of the Royal Statistical Society: Series C (Applied Statistics), 1975, vol. 24(2).
93. Mardia K.V. Measures of multivariate skewness and kurtosis with application // Biometrica, 1970, vol. 57, pp. 519–530.

94. Mardia K.V. Tests of univariate and multivariate normality // Handbook of statistics, vol. 1, pp. 279–320 / Ed. by S. Kotz et al. – New York: John Wiley & Sons, 1980.
95. Mardia K.V., Kent J.T., Bibby J.M. Multivariate analysis. – New York, NY: Academic Press, 1979.
96. Marsaglia G., Marsaglia J. Evaluating the Anderson–Darling distribution // Journal of Statistical Software, February 2004, vol. 9, no. 2, pp. 1–5.
97. Marsaglia G., Tsang W.W., Wang J. Evaluating Kolmogorov's distribution // Journal of Statistical Software, November 2003, vol. 8, no. 18, pp. 1–4.
98. Martinez J., Iglewicz B. A test for departure from normality based on a biweight estimator of scale // Biometrika, 1981, vol. 68, no. 1, pp. 331–333.
99. Massey F.J.Jr. The Kolmogorov–Smirnov test for goodness of fit // Journal of the American Statistical Association, 1951, vol. 46, pp. 68–78.
100. Mateu–Figueras G., Puig P., Pewsey A. Goodness–of–fit tests for the skew–normal distribution when the parameters are estimated from the data // Communications in Statistics: Theory and Methods, 2007, vol. 36, no. 9, pp. 1735–1755.
101. Mecklin C.J., Mundfrom D.J. An appraisal and bibliography of tests for multivariate normality // International Statistical Revue, 2004, vol. 72, no. 1, pp. 123–138.
102. Mecklin C.J., Mundfrom, D.J. On using asymptotic critical values in testing for multivariate normality // InterStat (Statistics on the Internet), 2003.
103. Mendes M., Pala A. Type I error rate and power of three normality tests // Pakistan Journal of Information and Technology, 2003, vol. 2, no. 2, pp. 135–139.
104. Mittnik S., Rachev S.T., Samorodnitsky G. The distribution of test statistics for outlier detection in heavy–tailed samples // Technical report TR001248, August 1999, Cornell University Operations Research and Industrial Engineering.
105. Morris K.W., Szynal D. Goodness–of–fit tests based on characterizations in terms of moments of order statistics // Applicationes Mathematicae, 2002, vol. 29, no. 3, pp. 251–283.
106. Morris K.W., Szynal D. Goodness–of–fit tests using characterizations of continuous distributions // Applicationes Mathematicae, 2001, vol. 25, no. 2, pp. 151–168.
107. Mudholkar G.S. A graphical procedure for comparing goodness–of–fit tests / G.S. Mudholkar, G.D. Kollia, C.T. Lin et al. // Journal of the Royal Statistical Society: Series B (Statistical Methodology), 1991, vol. 53, no. 1, pp. 221–232.
108. Mudholkar G.S., McDermott M., Srivastava D.K. A test of p–variate normality // Biometrika, December 1992, vol. 79, no. 4, pp. 850–854.
109. Mudholkar G.S., Natarajan R., Chaubey Y.P. A goodness–of–fit test for the inverse Gaussian distribution using its independence characterization // Sankhya: The Indian Journal of Statistics, 2001, vol. 63, series B, pt. 3, pp. 362–374.
110. Mudholkar G.S., Tian L. On the null distributions of the entropy tests for the Gaussian and inverse Gaussian models // Communications in Statistics: Theory and Methods, 2001, vol. 30, no. 8–9, pp. 1507–1520.
111. Nikulin M.S. Some recent results on chi–squared tests. – Kingston, Ontario: Queen's University, 1991.
112. NIST/SEMATECH e–Handbook of statistical methods (NIST Handbook 151, ver. 1/27/2005). – Gaithersburg, MD: National Institute of Standards and Technology, 2005.
113. Oja H. New tests for normality // Biometrika, April 1983, vol. 70, no. 1, pp. 297–299.
114. Ojeda R., Cardoso J.–F., Moulines E. Asymptotically invariant Gaussianity test for causal invertible time series // 1997 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'97), Munich, Germany, April 21–24, 1997, vol. 05, pp. 3713–3716.

115. Park S. A goodness–of–fit test for normality based on the sample entropy of order statistics // Statistics & Probability Letters, 1 October 1999, vol. 44, no. 4, pp. 359–363.
116. Pearson E.S. Note on tests for normality // Biometrika, May 1931, vol. 22, no. 3/4, pp. 423–424.
117. Peterson P., Stromberg A.J. A simple test for departures from multivariate normality // University of Kentucky, Lexington, Technical Report 373, March, 1998.
118. Pettitt A.N. A Cramer–von Mises type goodness of fit statistic related to $\sqrt{b_1}$ and b_2 // Journal of the Royal Statistical Society: Series B (Statistical Methodology), 1977, vol. 39, no. 3, pp. 364–370.
119. Pettitt A.N., Stephens M.A. The Kolmogorov–Smirnov goodness–of–fit statistic with discrete and grouped data // Technometrics, May 1977, vol. 19, no. 2, pp. 205–210.
120. Pinto J.V., Ng P., Allen D.S. Logical extremes, beta, and the power of the test // Journal of Statistics Education, 2003, vol. 11, no. 1.
121. Poitras G. More on the correct use of omnibus tests for normality // Economics Letters, 2006, vol. 90, pp. 304–309.
122. Prescott P. Comparison of tests for normality using stylized sensitivity surfaces // Biometrika, August 1976, vol. 63, no. 2, pp. 285–289.
123. Prescott P. On a test for normality based on sample entropy // Journal of the Royal Statistical Society: Series B (Statistical Methodology), 1976, vol. 38, no. 3, pp. 254–256.
124. Rahman M.M., Govindarajulu Z. A modification of the test of Shapiro and Wilk for normality // Journal of Applied Statistics, April 1, 1997, Vol. 24, num. 2, pp. 219–236.
125. Rayner J.C.W., Best D.J. Goodness–of–fit tests and diagnostics // International Encyclopedia of the Social & Behavioral Sciences. – Elsevier Science, 2001, pp. 6305–6310.
126. Razali N.M., Wah Y.B. Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests // Journal of Statistical Modeling and Analytics , 2011 , vol. 2, no. 1, pp. 21-33.
127. Reineke D.M., Baggett J., Elfessi A. A note on the effect of skewness, kurtosis, and shifting on one–sample *t* and sign tests // Journal of Statistics Education, 2003, vol. 11, no. 3.
128. Rhiel S.G., Chaffin W.W. An investigation of the large–sample/small–sample approach to the one–sample test for a mean (sigma unknown) // Journal of Statistics Education 1996, vol. 4, no. 3.
129. Romao X., Delgado R., Costa A. An empirical power comparison of univariate goodness–of–fit tests for normality // Journal of Statistical Computation and Simulation, May 2010, vol. 80, no. 5, pp. 545–591.
130. Royston J.P. A simple method for evaluating the Shapiro–Francia W' test for non–normality // The Statistician, September 1983, vol. 32, pp. 297–230.
131. Royston J.P. Algorithm AS 181: The W test for normality // Journal of the Royal Statistical Society: Series C (Applied Statistics), 1982, vol. 31, no. 2, pp. 176–180.
132. Royston J.P. An extension of Shapiro and Wilk's W test for normality to large samples // Journal of the Royal Statistical Society: Series C (Applied Statistics), 1982, vol. 31, no. 2, pp. 115–124.
133. Royston J.P. Approximating the Shapiro–Wilk's W–test for non–normality // Statistics and Computing, 1992, no. 2, pp. 117–119.
134. Royston J.P. Correction: Algorithm AS 181: The W test for normality // Journal of the Royal Statistical Society: Series C (Applied Statistics), 1983, vol. 32, no. 2, p. 224.
135. Royston J.P. Remark ASR 63: A remark on AS 181. The W test for normality // Journal of the Royal Statistical Society: Series C (Applied Statistics), 1986, vol. 35, no. 2, pp. 232–234.
136. Royston J.P. Some techniques for assessing multivariate normality based on the

Shapiro–Wilk W // Journal of the Royal Statistical Society: Series C (Applied Statistics), 1983, vol. 32, no. 2, pp. 121–133.
137. Royston P. A pocket–calculator algorithm for the Shapiro–Francia test for non–normality: An application to medicine // Statistics in Medicine, January 1993, vol. 12, no. 2, pp. 181–184.
138. Royston P. A remark on algorithm AS 181: The W test for normality // Journal of the Royal Statistical Society: Series C (Applied Statistics), 1995, vol. 44, pp. 547–551.
139. Royston P. A Simple method for evaluating the Shapiro–Francia W' test of non–normality // The Statistician, 1983, vol. 32, pp. 297–300.
140. Royston P. A toolkit for testing nonnormality in complete and censored samples // The Statistician, 1993, vol. 42, no. 1, pp.37–43.
141. Royston P. Algorithm AS R94 // Journal of the Royal Statistical Society: Series C (Applied Statistics), 1995, vol. 44, no. 4, pp. 547–551.
142. Royston P. An extension of Shapiro and Wilk's W test for normality to large samples // Journal of the Royal Statistical Society: Series C (Applied Statistics), 1982, vol. 31, pp. 115–124.
143. Royston P. Estimating departure from normality // Statistics in Medicine, August 1991, vol. 10, no. 8, pp. 1283–1293.
144. Royston P. Graphical detection of non–normality by using Michael's statistic // Journal of the Royal Statistical Society: Series C (Applied Statistics), 1993, vol. 42, no. 1, pp. 153–158.
145. Royston P. Remark AS R94: A remark on algorithm AS 181: The W test for normality // Journal of the Royal Statistical Society: Series C (Applied Statistics), 1995, vol. 44, no. 4, pp. 547–551.
146. Royston P., Altman D.G. Approximating statistical functions by using fractional polynomial regression // Journal of the Royal Statistical Society: Series D (The Statistican), September 1997, vol. 46, no. 3, pp. 411–422.
147. Ryan B.F., Joiner B.L., Cryer J.D. MINITAB Handbook. – Pacific Grove, CA: Duxbury Press, 2005.
148. Ryan T.A.Jr., Joiner B.L. Normal probability plots and tests for normality // Technical Paper, 1976.
149. Sainz de Rozas G.P. Using Mathematica to build non–parametric statistical tables // Journal of Statistical Software, 2003, vol. 8, no. 4.
150. Saniga E.M., Miles J.A. Power of some standard goodness–of–fit tests of normality against asymmetric stable alternatives // Journal of the American Statistical Association, December 1979, vol. 74, no. 368, pp. 861–865.
151. Sarkadi K. On testing for normality // Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, June 21 – July 18, 1965 and December 27, 1965–January 7, 1966, vol. 1: Statistics / Ed. by L.M. Le Cam, J. Neyman. – Berkeley, CA: University of California Press, 1967, pp. 373–387.
152. Sarkadi K. Testing for normality // Mathematical Statistics Banach Center Publications, 1980, vol. 6, pp. 281–287.
153. Sarkadi K. The consistency of the Shapiro–Francia test // Biometrika, 1975, vol. 62, pp. 445–450.
154. Seier E. Comparison of tests of univariate normality // InterStat (Statistics on the Internet), January 2002.
155. Sen P.K., Jureckova J., Picek J. Goodness–of–fit test of Shapiro–Wilk type with nuisance regression and scale // Austrian Journal of Statistics, 2003, vol. 32, no. 1–2, pp. 163–177.

156. Shannon C.E. A mathematical theory of communication // The Bell System Technical Journal, July, October 1948, vol. 27, pp. 379–423, 623–656.
157. Shapiro S.S., Francia R.S. An approximate analysis of variance test for normality // Journal of the American Statistical Association, March 1972, vol. 67, no. 337, pp. 215–216.
158. Shapiro S.S., Wilk M.B. An analysis of variance test for normality (complete samples) // Biometrika, December 1965, vol. 52, no. 3/4, pp. 591–611.
159. Shapiro S.S., Wilk M.B., Chen H.J. A comparative study of various tests for normality // Journal of the American Statistical Association, December 1968, vol. 63, no. 324, pp. 1343–1372.
160. Shawky A.I., Bakoban R.A. Modified goodness–of–fit tests for exponentiated gamma distribution with unknown shape parameter // InterStat (Statistics on the Internet), July 2009.
161. Spiegelhalter D.J. A test for normality against alternatives // Biometrika, 1977, vol. 64, pp. 415–418.
162. Spiegelhalter D.J. An omnibus test for normality for small samples // Biometrika, August 1980, vol. 67, no. 2, pp. 493–496.
163. Srivastava M.S., Hui T.K. On assessing multivariate normality based on Shapiro–Wilk W statistic // Statistics & Probability Letters, January 1987, vol. 5, no. 1, pp. 15–18.
164. Stengos T., Wu X. Information–theoretic distribution tests with application to symmetry and normality // SSRN Electronic Paper Collection (March 4, 2004), 21st Canadian Econometrics Study Group Conference Financial Econometrics, September 24–26, 2004, York University, Toronto, Canada.
165. Stephens M.A. Asymptotic results for goodness–of–fit statistics with unknown parameters // The Annals of Statistics, 1976, vol. 4, pp. 357–369.
166. Stephens M.A. EDF statistics for goodness of fit and some comparisons // Journal of the American Statistical Association, September 1974, vol. 69, pp. 730–737.
167. Stephens M.A. Use of Kolmogorov–Smirnov, Cramer–von Mises and related statistics without extensive table // Journal of the Royal Statistical Society: Series B (Statistical Methodology), 1970, vol. 32, no. 1, pp. 115–122.
168. Svantesson T., Wallace J. Tests for assessing multivariate normality and the covariance structure of MIMO data // IEEE International Conference on Acoustics, Speech and Signal Processing, Hong Kong, April 6–10, 2003 (ICASSP'03).
169. Thas O. Nonparametrical tests based on sample space partitions. Thesis for the degree of Ph.D. in applied biological sciences. – Gent, Belgien: University Gent, 2001.
170. Thode H.C. Testing for normality. – New York, NY: Marcel Dekker, 2002.
171. Thomas D.R., Rao J.N.K. On the power of some goodness–of–fit tests under cluster sampling // Proceedings of the Survey Research Methods Section, American Statistical Association, 1985, pp. 291–296.
172. Tiku M.L., Akkaya A.D. Robust estimation and hypothesis testing. – New Delhi: New Age International, 2004.
173. Tsang W.W., Wang J. Evaluating the CDF of the Kolmogorov statistic for normality testing // COMPSTAT 2004 – Proceedings in Computational Statistics, 16th Symposium Held in Prague, Czech Republic, 2004. – Physica–Verlag / Springer, 2004, pp. 1869–1876.
174. Uthoff V.A. The most powerful scale and location invariant test of the normal versus the double exponential // The Annals of Statistics, January 1973, vol. 1, no. 1, pp. 170–174.
175. Van Es B. Estimating functionals related to a density by a class of statistics based on spacings // Scandinavian Journal of Statistics, 1992, vol. 19, pp. 61–72.
176. Vasicek O. A test for normality based on sample entropy // Journal of the Royal Statistical Society: Series B (Statistical Methodology), 1976, vol. 38, no. 1, pp. 54–59.

177. Verrill S., Johnson R.A. The asymptotic equivalence of some modified Shapiro–Wilk statistics – complete and censored sample cases // The Annals of Statistics, March 1987, vol. 15, no. 1, pp. 413–419.
178. Von Eye A. Comparing tests of multinormality – A Monte Carlo study // InterStat (Statistics on the Internet), October 2005, No. 1.
179. Walpole R.E. Probability and statistics for engineers and scientists / R.E. Walpole, R.H. Mayers, S.L. Mayers et al. – Upper Saddle River, NJ: Prentice Hall, 2002.
180. Weisberg S. An empirical comparison of the percentage points of W and W' // Biometrika, December 1974, vol. 61, no. 3, pp. 644–646.
181. Weisberg S. Comment on «Some large–sample tests for nonnormality in the linear regression model» // Journal of the American Statistical Association, 1980, vol. 75, pp. 28–31.
182. Weisberg S., Bingham C. An approximate analysis of variance test for non–normality suitable for machine calculation // Technometrics, February 1975, vol. 17, no. 1, pp. 133–134.
183. White H., MacDonald G.M. Some large–sample tests for nonnormality in the linear regression model // Journal of the American Statistical Association, March 1980, vol. 75, no. 369, pp. 17–28.
184. Wieczorkowski R., Grzegorzewski P. Entropy estimators improvements and comparisons // Communications in Statistics: Simulation and Computation, 1999, vol. 28, pp. 541–567.
185. Wilcox R.R. Fundamentals of modern statistical methods. – New York, NY: Springer, 2001.
186. Xu D.F. Normality test and procedure for calculating skewness and kurtosis // Chinese Journal of Preventive Medicine (Zhonghua Yu Fang Yi Xue Za Zhi), November 1983, vol. 17, no. 6, pp. 321–323.
187. Zhang J., Wu Y. Likelihood–ratio tests for normality // Computational Statistics & Data Analysis, 2005, vol. 49, pp. 709–721.
188. Zhu L.X., Wong H.L., Fang K.T. A test for multivariate normality based on sample entropy and projection pursuit // Journal of Statistical Planning and Inference, June 1995, vol. 45, no. 3, pp. 373–385.
189. Большев Л.Н., Смирнов Н.В. Таблицы математической статистики. – М.: Наука, 1983.
190. Боровков А.А. Математическая статистика. Оценка параметров. Проверка гипотез. – М.: Наука, 1984.
191. Брандт З. Анализ данных. Статистические и вычислительные методы для научных работников и инженеров. – М.: Мир, ООО «Издательство АСТ», 2003.
192. Ван дер Варден Б.Л. Математическая статистика. – М.: Издательство иностранной литературы, 1960.
193. Вентцель Е.С. Теория вероятностей. – М.: Высшая школа, 1999.
194. Воинов В.Г. Об оптимальных свойствах критерия Рао–Робсон–Никулина // Заводская лаборатория. Диагностика материалов, 2006, № 3, с. 65–70.
195. Гайдышев И. Анализ и обработка данных: Специальный справочник. – СПб: Питер, 2001.
196. Голенко Д.И. Моделирование и статистический анализ псевдослучайных чисел на электронных вычислительных машинах. – М.: Наука, 1965.
197. ГОСТ 8.207–76. Государственная система обеспечения единства измерений. Прямые измерения с многократными наблюдениями. Методы обработки результатов наблюдений. – М.: ИПК Издательство стандартов, 2001.

198. ГОСТ Р ИСО 5479–2002. Статистические методы. Проверка отклонения распределения вероятностей от нормального распределения. – М.: Издательство стандартов, 2002.
199. Дерффель К. Статистика в аналитической химии. – М.: Мир, 1994.
200. Дэйвид Г. Порядковые статистики. – М.: Наука, 1979.
201. Ермаков С.М. Метод Монте–Карло и смежные вопросы. – М.: Наука, 1975.
202. Золотухина Л.А., Винник Е.В. Эмпирическое исследование мощности критерия Саркади и его модификации // Заводская лаборатория. Диагностика материалов, 1985, № 1, с. 51–55.
203. Кендалл М., Стьюарт А. Статистические выводы и связи. – М.: Наука, 1973.
204. Кобзарь А.И. Прикладная математическая статистика. Для инженеров и научных работников. – М.: ФИЗМАТЛИТ, 2006.
205. Крамер Г. Математические методы статистики. – М.: Мир, 1975.
206. Кулаичев А.П. Компьютерный контроль процессов и анализ сигналов. – М.: Информатика и компьютеры, 1999.
207. Кулаичев А.П. Методы и средства анализа данных в среде Windows®. STADIA. – М.: Информатика и компьютеры, 1999.
208. Кулаичев А.П. Методы и средства комплексного анализа данных. – М.: ИНФРА–М, 2006.
209. Кумэ Х. Статистические методы повышения качества / Под ред. Х. Кумэ. – М.: Финансы и статистика, 1990.
210. Лакин Г.Ф. Биометрия. – М.: Высшая школа, 1990.
211. Лемешко Б.Ю. Асимптотически оптимальное группирование наблюдений в критериях согласия // Заводская лаборатория. Диагностика материалов, 1998, т. 64, №1, с. 56–64.
212. Лемешко Б.Ю., Лемешко С.Б. Сравнительный анализ критериев проверки отклонения распределения от нормального закона // Метрология, 2005, № 2, с.3–23.
213. Лемешко Б.Ю., Лемешко С.Б., Постовалов С.Н. Сравнительный анализ мощности критериев согласия при близких альтернативах. II. Проверка сложных гипотез // Сибирский журнал индустриальной математики, 2008, т. 11, № 4 (36), с. 78–93.
214. Лемешко Б.Ю., Лемешко С.Б., Постовалов С.Н. Сравнительный анализ мощности критериев согласия при близких конкурирующих гипотезах. I. Проверка простых гипотез // Сибирский журнал индустриальной математики, 2008, т. 11, № 2 (34), с. 96–111.
215. Лемешко Б.Ю., Постовалов С.Н., Чимитова Е.В. О распределениях статистики и мощности критерия типа χ^2 Никулина // Заводская лаборатория. Диагностика материалов, 2001, т. 67, № 3, с. 52–58.
216. Лемешко Б.Ю., Рогожников А.П. Исследование методами статистического моделирования свойств некоторых критериев нормальности // Девятая международная научно–техническая конференция по актуальным проблемам электронного приборостроения, Новосибирский государственный технический университет, 24–26 сентября 2008 г.
217. Лемешко Б.Ю., Чимитова Е.В. Максимизация мощности критериев типа χ^2 // Доклады СО АН ВШ, Новосибирск, 2000, № 2, с. 53–61.
218. Лемешко Б.Ю., Чимитова Е.В. О выборе числа интервалов в критериях согласия типа χ^2 // Заводская лаборатория. Диагностика материалов, 2003, т. 69, № 1, с. 61–67.
219. Лемешко Б.Ю., Чимитова Е.В. Численное сравнение оценок максимального

правдоподобия с одношаговыми и влияние точности оценивания на распределения статистик критериев согласия // Заводская лаборатория. Диагностика материалов, 2003, т. 69, № 5, с. 62–68.

220. Лыхмус К.Н. Информационный критерий гомогенности выборки / Биометрический анализ в биологии. – М.: Издательство Московского университета, 1982, с. 51–57.
221. Мартынов Г.В. Критерии омега–квадрат. – М.: Наука, 1978.
222. Мирвалиев М., Никулин М.С. Критерии согласия типа хи–квадрат // Заводская лаборатория. Диагностика материалов, 1992 , т. 58, № 3, с. 52–58.
223. Мюллер П., Нойман П., Шторм Р. Таблицы по математической статистике. – М.: Финансы и статистика, 1982.
224. Никитин Я.Ю. Асимптотическая эффективность непараметрических критериев. – М.: Наука, 1995.
225. Никулин М.С. Критерий хи–квадрат для непрерывных распределений с параметрами сдвига и масштаба // Теория вероятностей и ее применение, 1973, т. XVIII, № 3, с. 583–591.
226. Никулин М.С. О критерии хи–квадрат для непрерывных распределений // Теория вероятностей и ее применение, 1973, т. XVIII, № 3, с. 675–676.
227. Никулин М.С., Воинов В.Г. Критерий согласованности Чи–квадрата для экспонентного распределения первого порядка. – Л.: ЛОМИ, 1987.
228. Новицкий П.В., Зограф И.А. Оценка погрешностей результатов измерений. – Л.: Энергоатомиздат, 1985.
229. Прохоров Ю.В. Вероятность и математическая статистика. Энциклопедия / Гл. ред. Ю.В. Прохоров. – М.: Научное издательство «Большая Российская энциклопедия», 1999.
230. Пя Н.Е. Модифицированные критерии хи–квадрат, основанные на классах Неймана–Пирсона, для нормального распределения // Известия НАН РК, серия физико–математическая, 2004, № 5, с. 92–98.
231. Рекомендации по стандартизации Р 50.1.033–2001. Прикладная статистика. Правила проверки согласия опытного распределения с теоретическим. Часть I. Критерии типа хи–квадрат. – М.: Издательство стандартов, 2002.
232. Рекомендации по стандартизации Р 50.1.037–2002. Прикладная статистика. Правила проверки согласия опытного распределения с теоретическим. Часть II. Непараметрические критерии. – М.: Издательство стандартов, 2002.
233. Родионов Д.А. Справочник по математическим методам в геологии / Д.А. Родионов, Р.И. Коган, В.А. Голубева и др. – М.: Недра, 1987.
234. Романовский В.И. Математическая статистика. Кн.2. Оперативные методы математической статистики. – Ташкент: Издательство Академии наук УзССР, 1963.
235. Селезнев В.Д., Денисов К.С. Исследование свойств критериев согласия функции распределения данных с гауссовой методом Монте–Карло для малых выборок // Заводская лаборатория. Диагностика материалов, 2005, № 1, с. 68–73.
236. Степнов М.Н. Статистические методы обработки результатов механических испытаний: Справочник. – М.: Машиностроение, 1985.
237. Тейлор Дж. Введение в теорию ошибок. – М.: Мир, 1985.
238. Тюрин Ю.Н. Непараметрические методы статистики. – М.: Знание, 1978.
239. Тюрин Ю.Н., Макаров А.А. Статистический анализ данных на компьютере. – М.: ИНФРА–М, 1999.
240. Уилкс С. Математическая статистика. – М.: Наука, 1967.
241. Фишер Р.А. Статистические методы для исследователей. – М.: Госстатиздат,

1958.
242. Хан Г., Шапиро С. Статистические модели в инженерных задачах. – М.: Мир, 1969.
243. Хромов–Борисов Н.Н., Лаззаротто Г.Б., Ледур Кист Т.Б. Биометрические задачи в популяционных исследованиях // VII Всероссийский популяционный семинар «Методы популяционной биологии», 16–21 февраля 2004, Сыктывкар.
244. Шеннон К. Работы по теории информации и кибернетике. – М.: Издательство иностранной литературы, 1963.
245. Шор Я.Б., Кузьмин Ф.И. Таблицы для анализа и контроля надежности. – М.: Советское радио, 1968.

Глава 8. Дисперсионный анализ

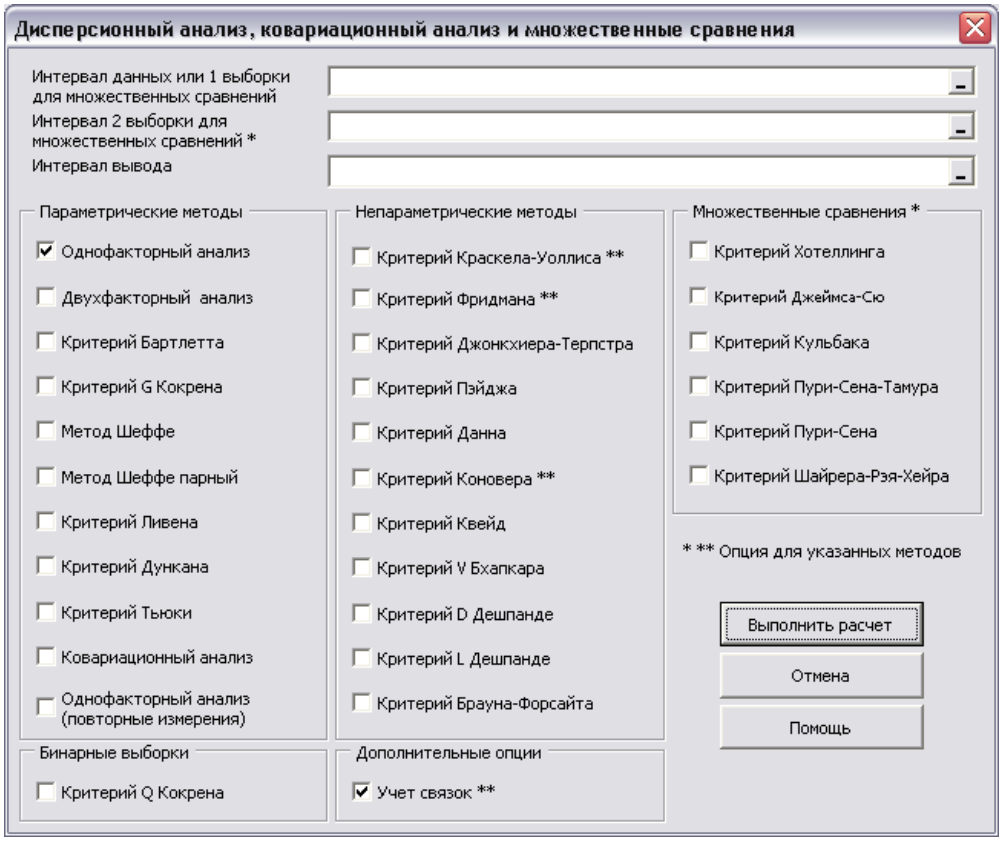
8.1. Введение

Назначение представленных в данной главе дисперсионного анализа, множественных сравнений и ковариационного анализа подробно разъясняется в соответствующих теоретических разделах.
Методы дисперсионного анализа и множественных сравнений могут быть предназначены для нормально распределенных совокупностей (т. е. будут многомерными аналогами параметрических тестов) и для выборок, свободных от предположения о типе распределения (т. е. будут многомерными аналогами непараметрических тестов). Методы ковариационного анализа предполагают нормальность распределения ошибок (относительно линейной регрессии). Нормальность распределения произвольных по численности и «числу измерений» выборок может быть проверена с помощью методов главы «Проверка нормальности распределения».

8.2. Работа с программным обеспечением

Выберите из меню программы пункт AtteStat | Дисперсионный анализ. На экране появится диалоговое окно, изображенное на рисунке:
Затем проделайте следующие шаги:

- Выберите или введите интервалы матрицы исходных данных, трактуемой в ряде методов также в качестве многомерной выборки. Считывание программой значений выборки (столбца таблицы), если указанный пользователем интервал содержит пустые значения, обрывается, как только встречается пустое значение. Данная особенность вызвана необходимостью обеспечения возможности работы с выборками разных численностей. Так, например, если в столбце будет пять значений, потом пустое значение, потом еще четыре значения, программой будет считана только выборка из первых пяти значений. Если предполагается использовать методы множественных сравнений, здесь следует выбрать интервал первой многомерной выборки. Особенности представления исходных данных для дисперсионного анализа см. в главе «Дисперсионный анализ». Особенности представления исходных данных для множественных сравнений см. в разделе «Множественные сравнения».

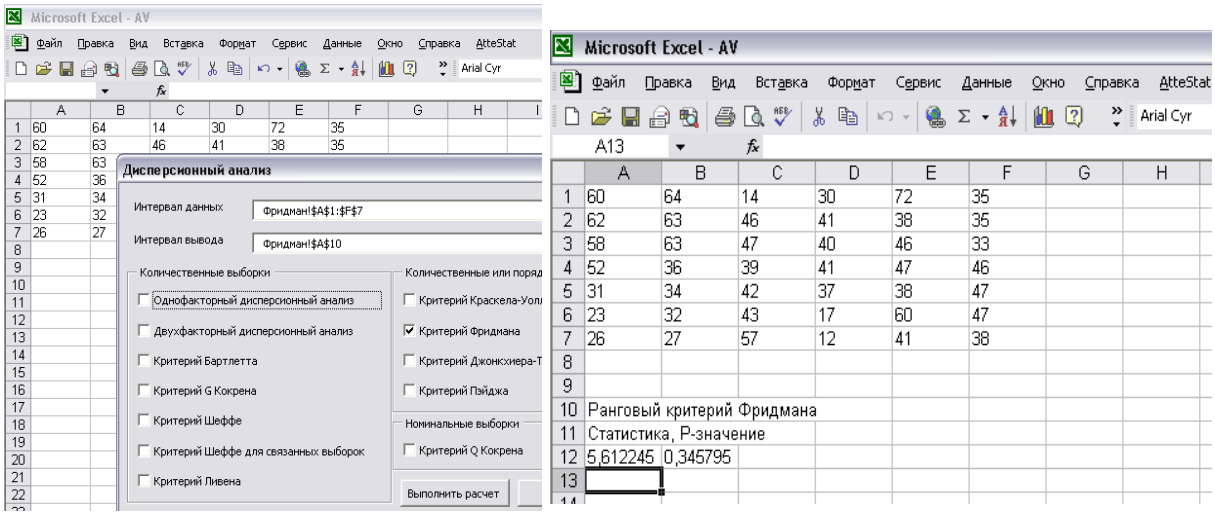


- Для множественных сравнений выберите или введите интервал второй многомерной выборки. Для дисперсионного анализа содержимое данного поля значения не имеет.
- Оставьте по умолчанию или измените дополнительные опции.
- Выберите или введите выходной интервал. Начиная с первой ячейки выходного интервала (следовательно, можно указать только одну ячейку, т.к. остальные ячейки интервала игнорируются), будут выведены результаты вычислений.
- Выберите или оставьте по умолчанию метод анализа.
- Нажмите кнопку «Выполнить расчет».

После выполнения вычислений будет, начиная с первой ячейки выходного интервала, выведено название метода и результаты расчета: номера выборок (для некоторых методов), статистика критерия, вычисленное *P*–значение. Интерпретация полученных результатов статистических расчетов подробно рассмотрена в разделах, посвященных методам расчета. За выбор адекватного исходным данным метода расчета несет ответственность пользователь. Программное обеспечение берет на себя верификацию исходных данных, выдавая подробную диагностику. При неверных действиях пользователя или ошибках периода выполнения выдаются сообщения об ошибках.

8.2.1. Пример применения

В качестве примера исследуем массив исходных данных, приведенных на с. 244 монографии Браунли. Как и в источнике, воспользуемся критерием Фридмана.
В интервал ячеек **A2**:**F8** введем исходные данные. В качестве интервала вывода (начала интервала) укажем ячейку **A10**. Выберем нужный метод дисперсионного анализа. Экран компьютера при выполнении данных манипуляций будет выглядеть примерно так.



После нажатия кнопки «Выполнить расчет» экран примет вид, показанный на фрагменте. Результаты совпадают с источником. Нулевая гипотеза может быть принята. Подробную интерпретацию результатов см. в описании критерия и источнике.

8.2.2. Сообщения об ошибках

При ошибках ввода данных могут выдаваться диагностические сообщения следующих типов:

Ошибка	Комментарий
Не определен интервал переменной.	Вы не выбрали или неверно ввели интервал эмпирической выборки. Лучшим способом избежать ошибки является не ввод, а выделение интервала стандартным образом, т. е. протаскиванием курсора.
Не определена область вывода.	Вы не выбрали или неверно ввели выходной интервал. Лучшим способом избежать ошибки является не ввод, а выделение интервала стандартным образом, т. е. протаскиванием курсора.
Нечисловой тип данных.	Тип данных может быть только числовым. Проще всего выделить интервал ячеек и явно определить их тип как числовой стандартными средствами. Убедитесь, что в заданном интервале не содержится нечисловых значений.
Мала размерность выборки.	Количество выборок, трактуемое в программе также как размерность многомерной выборки, для дисперсионного анализа должно быть не менее двух.
Мала численность выборки.	Численность каждой выборки не может быть меньше двух. Укажите интервал матрицы исходных данных, содержащих выборки численностью от двух и более.

8.3. Теоретическое обоснование

8.3.1. Дисперсионный анализ

Дисперсионным анализом называют совокупность статистических методов,

предназначенных для обработки данных экспериментов, целью которых являлось не установление каких–то свойств и параметров, а сравнение эффектов различных воздействий на каком–либо экспериментальном материале. Методы дисперсионного анализа используются для проверки гипотез о наличии связи между результативным признаком и исследуемыми факторами, а также для установления силы влияния факторов и их взаимодействий.
Из представленных критериев одна многочисленная группа тестов является параметрическими и требуют нормальности распределения исходных выборок. Данные методы предназначены только для нормально распределенных количественных данных. Исследованию свойств некоторых параметрических методов при нарушении предположений о нормальности посвящены работы Лемешко с соавт. Проверить нормальность распределения, включая многомерный случай, можно с помощью методов главы «Проверка нормальности распределения». Другие методы являются непараметрическими и не требуют предположений относительно вида исходного распределения.
Критерий *Q* Кокрена предназначен для бинарных (дихотомических) данных.

Для параметрических и непараметрических методов проверки гипотез (см. «Параметрическая статистика» и «Непараметрическая статистика») существуют многомерные аналоги в дисперсионном анализе, как показано в таблице.

Метод проверки гипотезы для двух выборок	«Функциональный аналог» из дисперсионного анализа
Параметрические тесты	
Критерий Стьюдента	Однофакторный дисперсионный анализ
	Критерий Шеффе
	Критерий Пейджа
	Критерий Дункана
	Критерий Тьюки
Критерий Стьюдента парный	Однофакторный дисперсионный анализ с повторными измерениями
	Многофакторный дисперсионный анализ
	Критерий Шеффе для связанных выборок
F–критерий	Критерий Бартлетта
	Критерий G Кокрена
	Критерий Ливена
Непараметрические тесты	
Критерий Вилкоксона	Критерий Джонкхиера–Терпстра
	Критерий Краскела и Уоллиса
	Критерий Данна
	Критерий Коновера
	Критерий Кьюзика
Критерий Вилкоксона парный	Ранговый критерий Фридмана
	Критерий Квейд
Точный метод Фишера	Критерий *Q* Кокрена
	Критерий *V* Бхапкара
	Критерий *D* Дешпанде
	Критерий *L* Дешпанде
	Критерий Брауна–Форсайта

Методы дисперсионного анализа следует использовать, когда число выборок больше двух. Нельзя применять критерии, предназначенные для сравнения выборок попарно, а затем делать какие–либо выводы относительно всей совокупности.
В дисперсионном анализе, как и в других областях анализа данных, сложилась определенная терминология. Фактором называют величину, определяющую свойства исследуемого объекта или системы, иначе – причину, влияющую на конечный результат. Конкретную реализацию фактора называют уровнем фактора или способом обработки. Значение измеряемого признака называют откликом.

См. книги Браунли, Кобзаря, Холлендера с соавт., нормативный документ EPA QA/G–9.

8.3.1.1. Однофакторный дисперсионный анализ

Исходные данные для однофакторного дисперсионного анализа представлены в виде таблицы (прямоугольной матрицы), причем число столбцов (выборок) соответствует числу уровней фактора (уровней обработки), число строк равно числу наблюдений. При этом выборки могут иметь как одинаковое число вариант (равные объемы), так и различное, в зависимости от требований применяемого метода.
Предлагаются методы однофакторного дисперсионного анализа:

- Однофакторный дисперсионный анализ (ANOVA).
- Однофакторный дисперсионный анализ с повторными измерениями.
- Ранговый однофакторный анализ Краскела–Уоллиса.
- Критерий Данна.
- Критерий Коновера.
- Критерий Джонкхиера–Терпстра.
- Критерий Бартлетта.
- G–критерий Кокрена.
- Критерий Шеффе.
- Критерий Дункана.
- Критерий Тьюки.
- Критерий Ливена.
- Критерий Брауна–Форсайта.
- Критерий V Бхапкара.
- Критерий D Дешпанде.
- Критерий L Дешпанде.

8.3.1.1.1. Однофакторный дисперсионный анализ

При однофакторном дисперсионном анализе (дисперсионном анализе по одному признаку, analysis of variance, ANOVA) предполагается, что результаты наблюдений для разных уровней представляют собой выборки из нормально распределенных генеральных совокупностей. Эти совокупности имеют свои средние и дисперсии, которые полагаются одинаковыми. Задачей анализа является проверка нулевой гипотезы о равенстве средних рассматриваемых совокупностей. Вычисление критерия производится по формуле

$$t = \frac{N-k}{k-1} \frac{\sum_{i=1}^{k} n_i \left(\bar{x}_{i.} - \bar{x}_{..} \right)^2}{\sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(x_{ij} - \bar{x}_{i.} \right)^2},$$

$N = \sum_{i=1}^{k} n_i$
где – общая численность,
n_i, $i = 1,2,...,k$ – численность i–й выборки,
$\bar{x}_{i.} = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij}, i = 1,2,...,k$ – среднее значение i–й выборки,
$\bar{x}_{..} = \frac{1}{N} \sum_{i=1}^{k} \sum_{j=1}^{n_i} x_{ij}$ – общее среднее значение,
x_{ij}, $i = 1,2,...,k$; $j = 1,2,...,n_i$ – варианты выборки,
k – число столбцов (выборок).
Сумма, стоящая в числителе формулы вычисления критерия, служит приближенной мерой вариации между анализируемыми выборками, а двойная сумма, стоящая в знаменателе, служит мерой вариации внутри выборок.
Статистика критерия имеет F–распределение с параметрами $k - 1$ и $N - k$.

См. монографию Шеффе.

8.3.1.1.2. Однофакторный дисперсионный анализ (повторные измерения)

При однофакторном дисперсионном анализе с повторными измерениями (repeated measurements ANOVA) предполагается, что результаты наблюдений одного и того же процесса для разных временных уровней представляют собой выборки из нормально распределенных генеральных совокупностей. Эти совокупности имеют свои средние и дисперсии, которые полагаются одинаковыми. Задачей анализа является проверка нулевой гипотезы о равенстве средних рассматриваемых совокупностей.
Вычисления производятся по формулам:
$t = \frac{D_{col}}{D}$,
$D_{col} = \frac{SS_{col}}{c - 1}$
где – дисперсия, объясняемая столбцами,
$D = \frac{SS}{(r-1)(c-1)}$ – остаточная дисперсия,
$SS_{col} = r \sum_{j=1}^{c} (T_{.j} - T_{..})^2$ – средний квадрат столбцов,
$SS = \sum_{i=1}^{r} \sum_{j=1}^{c} (x_{ij} - T_{i.} - T_{.j} + T_{..})^2$ – средний квадрат погрешности,
$T_{i.} = \frac{1}{c} \sum_{j=1}^{c} x_{ij}, i = 1,2,...,r$ – средние суммы строк,
$T_{.j} = \frac{1}{r} \sum_{i=1}^{r} x_{ij}, j = 1,2,...,c$ – средние суммы столбцов,
$T_{..} = \frac{1}{rc} \sum_{i=1}^{r} \sum_{j=1}^{c} x_{ij}$ – общее среднее,
c – число столбцов (выборок),

r – число строк (параметров).
Статистика критерия имеет F–распределение с параметрами $r - 1$ и $(r - 1)(c - 1)$.
Результаты расчета совпадают с эффектом столбцов в двухфакторном дисперсионном анализе.

Описание см. в монографии Дэйвиса (Davis).

8.3.1.1.4. Критерий Данна

Ранговый однофакторный анализ Краскела и Уоллиса может показать, что параметры положения совокупностей различаются. Однако данный критерий не позволяет узнать, параметры каких совокупностей действительно различаются между собой. Для решения проблемы применяется непараметрический критерий Данна (Bonferroni–Dunn post hoc test, Dunn's multiple comparison post–test). Критерий применим для независимых групп как равной, так и различной численности. Вычисление критерия производится по формуле

$$Q_{ij} = \frac{\left|\overline{R}_i - \overline{R}_j\right|}{\sqrt{\frac{N(N+1)}{12}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}}, i = 1,2,...,k; j = i+1,...,k,$$

где $\overline{R}_i = \frac{1}{n_i}\sum_{l=1}^{n_i} R_{il}, i = 1,2,...,k$ – средний ранг i–й выборки,
$R_{i.}$, $i = 1,2,...,k$ – ранги наблюдений i–ой выборки,
$N = \sum_{i=1}^{k} n_i$ – общая численность,
n_i, $i = 1,2,...,k$ – численность i–й выборки,
k – количество столбцов (групп).
P–значения критерия p_{ij}, $i = 1,2,...,k$; $j = i + 1,...,k$, являются решениями нелинейных уравнений

$$Q_{ij} = \Psi\left(\frac{p_{ij}}{k(k-1)}\right), i = 1,2,...,k; j = i+1,...,k,$$

где $\Psi(.)$ – обратная функция стандартного нормального распределения.
Уравнения могут быть решены одним из методов локальной оптимизации. В простейшем случае используется метод деления отрезка пополам.

Описание см. у Гланца, Даниэла (Daniel), Зигеля (Siegel) с соавт., Холлендера с соавт. В литературе описаны родственные рассматриваемому методу критерий Райана (Ryan–Einot–Gabriel–Welsch test) и критерий Бартоломью (Bartholomew test). См. также критерий Шаича–Хамерле (Schaich–Hamerle post hoc test), представленный в монографиях Шаич (Schaich) с соавт., Бортц (Bortz) с соавт.

8.3.1.1.3. Ранговый однофакторный анализ Краскела и Уоллиса

Критерий Краскела–Уоллиса (ранговый однофакторный анализ Краскела–Уоллиса) является непараметрическим аналогом однофакторного дисперсионного анализа и предназначен для проверки нулевой гипотезы о равенстве эффектов обработки (воздействия) на выборки с неизвестными, но равными средними. Нулевая гипотеза заключается в том, что все совокупности одинаково распределены. Вычисление критерия производится по формуле

$$H = \frac{12}{N(N+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(N+1),$$

где R_i, $i = 1,2,...,k$ – сумма рангов наблюдений i–ой выборки,

$N = \sum_{i=1}^{k} n_i$ – общая численность,

n_i, $i = 1,2,...,k$ – численность i–й выборки,

k – количество столбцов (групп).

В программе введена поправка на объединение рангов

$$b = 1 - \frac{1}{N(N^2-1)} \sum_{j=1}^{g} t_j (t_j^2 - 1),$$

где t_j, $j = 1,2,...,g$ – численность связки,

g – число связок.

Тогда модифицированная статистика, выводимая программой, будет записана как $H' = H / b$.

Статистика критерия (равно и модифицированная статистика) имеет χ^2–распределение с параметром $k - 1$.

См. книги Бикела с соавт., Петровича с соавт., Холлендера с соавт. Точное вычисление критерия см. в работе Клотца (Klotz) с соавт.

8.3.1.1.5. Критерий Коновера

Ранговый однофакторный анализ Краскела и Уоллиса может показать, что параметры положения совокупностей различаются. Однако данный критерий не позволяет узнать, параметры каких совокупностей действительно различаются между собой. Для решения проблемы применяется непараметрический критерий Коновера (Conover post hoc test). Критерий применим для независимых групп как равной, так и различной численности. Вычисление критерия производится по формуле

$$C_{ij} = \frac{\left|\overline{R}_i - \overline{R}_j\right|}{\sqrt{\frac{N(N+1)}{12} \cdot \frac{N-1-H}{N-k} \left(\frac{1}{n_i} + \frac{1}{n_j}\right)}}, i = 1,2,...,k;\ j = i+1,...,k,$$

где $\overline{R}_i = \frac{1}{n_i} \sum_{l=1}^{n_i} R_{il}, i = 1,2,...,k$ – средний ранг i–й выборки,

$R_{i.}$, $i = 1,2,...,k$ – ранги наблюдений i–ой выборки,

$N = \sum_{i=1}^{k} n_i$ – общая численность,

n_i, $i = 1,2,...,k$ – численность i–й выборки,

H – статистика критерия Краскела–Уоллиса,

k – количество столбцов (групп).

P–значения критерия p_{ij}, $i = 1,2,...,k$; $j = i + 1,...,k$, подчиняются t–распределению с параметром $N - k$.

Описание см. у Бортц (Bortz) с соавт.

8.3.1.1.6. Критерий Джонкхиера и Терпстра

Критерий Джонкхиера–Терпстра (критерий Джонкхиера) представляет собой многомерное обобщение критерия Манна–Уитни (см. главу «Непараметрическая статистика) и предназначен для проверки нулевой гипотезы о равенстве эффектов обработки (воздействия) на выборки с неизвестными, но равными средними. Вычисление критерия производится по формуле

$$J = \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} U_{ij},$$

где U_{ij}, $i = 1,2,...,k-1$; $j = 2,3,...,k$ – статистика критерия Манна–Уитни для выборок с номерами i и j,
k – число столбцов (выборок).
Для больших выборок распределение преобразованной статистики

$$\frac{J - MJ}{\sqrt{DJ}}$$

является приближенно нормальным. Здесь математическое ожидание и дисперсия рассчитываются по формулам, соответственно:

$$MJ = \frac{1}{4}\left(N^2 - \sum_{i=1}^{k} n_i^2 \right),$$

$$DJ = \frac{1}{72}\left(N^2(2N+3) - \sum_{i=1}^{k} n_i^2 (2n_i + 3) \right),$$

$N = \sum_{i=1}^{k} n_i$ – общая численность,
где n_i, $i = 1,2,...,k$ – численность i–й выборки.

Описание см. в книгах Тюрина с соавт., Холлендера с соавт., в работе Кьюзика (Cuzick).

8.3.1.1.7. Критерий Бартлетта

Критерий Бартлетта (M–критерий Бартлетта) служит для проверки нулевой гипотезы о равенстве дисперсий нормальных генеральных совокупностей. Вычисления статистики критерия производится по формуле

$$M = \left[1 + \frac{1}{3(k-1)} \left(\sum_{i=1}^{k} \frac{1}{n_i - 1} - \frac{1}{N-k} \right) \right]^{-1} \left[(N-k)\ln s^2 - \sum_{i=1}^{k} (n_i - 1)\ln s_i^2 \right],$$

где n_i, $i = 1,2,...,k$ – численность i–й выборки,

$$s^2 = \frac{\sum_{i=1}^{k} (n_i - 1) s_i^2}{N-k},$$

$s_i^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} \left(x_{ij} - \bar{x}_{i.} \right)^2, i = 1,2,...,k$ – выборочная дисперсия i–й выборки,

$\bar{x}_{i.} = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij}, i = 1,2,...,k$ – среднее значение i–й выборки,

$N = \sum_{i=1}^{k} n_i$ – суммарная численность всех выборок,
k – число столбцов (выборок).
Статистика критерия имеет χ^2–распределение с параметром $k - 1$.

Описание см. у Браунли, Когана с соавт.

8.3.1.1.8. Критерий G Кокрена

Критерий G Кокрена (статистика Кокрена, критерий Кохрана) используется для проверки нулевой гипотезы о равенстве дисперсий нормальных генеральных совокупностей по независимым выборкам с одинаковыми численностями. Вычисление статистики критерия производится по формуле

$$G = \frac{\max\limits_{1 \le i \le k} \sigma_i^2}{\sum\limits_{i=1}^{k} \sigma_i^2},$$

где $\sigma_i^2, i = 1,2,...,k$ – выборочные дисперсии совокупностей,
k – число выборочных совокупностей.
P–значение модифицированной статистики

$$G' = \frac{G(k-1)}{1-G}$$

является решением нелинейного уравнения

$$G' = F^{-1}_{(n-1),(n-1)(k-1)}\left(\frac{p}{k}\right),$$

где $F^{-1}_{...}(.)$ – обратная функция F–распределения,
n – численность каждой совокупности.
Уравнение может быть решено одним из методов локальной оптимизации. В простейшем случае используется метод деления отрезка пополам. В программе данная функция для удобства оформлена как стандартная функция распределения статистики критерия G Кокрена.

Описание см. в монографиях Мюллера с соавт., Налимова, Зигеля (Siegel) с соавт.

8.3.1.1.9. Критерий Шеффе

Однофакторный анализ может показать, что средние значения совокупностей различаются. Однако он не позволяет узнать, средние значения каких совокупностей действительно различаются между собой. Для решения проблемы применяется метод множественного сравнения Шеффе (критерий Шеффе). Критерий Шеффе предназначен для проверки так называемой гипотезы о линейном контрасте. Линейный контраст

$$L = \sum_{i=1}^{k} c_i \mu_i$$

представляет собой линейную функцию от средних значений μ_i, $i = 1,2,...,k$, k независимых нормальных выборок с неизвестными, но равными дисперсиями, и известных констант c_i, $i = 1,2,...,k$, удовлетворяющих условию

$$\sum_{i=1}^{k} c_i = 0.$$

В частном случае проверяется серия гипотез о простых линейных контрастах вида $L_0 = \mu_i - \mu_j$, $i = 1,2,...,k-1$; $j = i+1,...,k$.
Вычисление критерия производится по формуле

$$t = \frac{\sum_{i=1}^{k} c_i \bar{x}_i}{\sqrt{M \sum_{i=1}^{k} \frac{c_i^2}{n_i}}},$$

где $M = \frac{1}{N-k} \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(x_{ij} - \bar{x}_{i.}\right)^2$ – средний квадратичный остаток,

$\bar{x}_{i.} = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij}, i = 1,2,...,k$ – среднее значение i–й выборки,

$N = \sum_{i=1}^{k} n_i$ – общая численность,

n_i, $i = 1,2,...,k$ – численность i–й выборки,
k – число столбцов (выборок).
Статистика критерия имеет F–распределение с параметрами $k-1$ и $N-k$.

Обсуждение см. в книгах Полларда (Pollard), Полларда, Шеффе, Мюллера с соавт., Ликеша с соавт., Бикела с соавт., Браунли.

8.3.1.1.10. Критерий Дункана

Однофакторный анализ может показать, что средние значения совокупностей различаются. Однако он не позволяет узнать, средние значения каких совокупностей действительно различаются между собой. Для решения проблемы применяется критерий Дункана (Duncan's test). Вычисление критерия производится по формуле

$$d = \frac{\left|\bar{x}_i - \bar{x}_j\right|}{\sqrt{\frac{M}{2}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}}, i = 1,2,...,k; j = i+1,...,k,$$

где $M = \frac{1}{N-k} \sum_{l=1}^{k} \sum_{m=1}^{n_i} \left(x_{lm} - \bar{x}_l\right)^2$ – средний квадратичный остаток,

$\bar{x}_i = \frac{1}{n_i} \sum_{l=1}^{n_i} x_{il}, i = 1,2,...,k$ – среднее значение i–й выборки,

$N = \sum_{l=1}^{k} n_l$ – общая численность,

n_i, $i = 1,2,...,k$ – численность i–й выборки,
k – число столбцов (выборок).
P–значение критерия p является решением нелинейного уравнения
$P_{r+2,\,N-k}(d) = (1-p)^{r+1}$,
где $P_{\cdot,\cdot}(.)$ – функция распределения стьюдентизированного размаха,

r – количество средних значений, расположенных между $\bar{x}_i$ и $\bar{x}_j$ в упорядоченном по возрастанию ряду k средних.
Благодаря простой структуре уравнения не представляет большого труда вычислить обратную функцию распределения рассматриваемого критерия

$$p = 1 - \exp \frac{\ln P_{r+2,N-k}(d)}{r+1}.$$

Описание см. в сборниках таблиц Оуэна, Мюллера с соавт. См. также описание критериев Ньюмена–Кейлса (Student–Newman–Kuels test) и его варианта для сравнения с контрольной группой – критерия Даннета (Dunnett test) в книге Гланца.

8.3.1.1.11. Критерий Тьюки

Если независимые выборки имеют равные численности, гипотезы о простых линейных контрастах могут быть проверены с помощью критерия Тьюки (метода Тьюки). Критерий Тьюки имеет аналогичные критерию Шеффе предпосылки для своего применения.
Линейный контраст

$$L = \sum_{i=1}^{k} c_i \mu_i$$

представляет собой линейную функцию от средних значений μ_i, $i = 1,2,...,k$, k независимых нормальных выборок с неизвестными, но равными дисперсиями, и известных констант c_i, $i = 1,2,...,k$, удовлетворяющих условию

$$\sum_{i=1}^{k} c_i = 0.$$

В частном случае проверяется серия гипотез о простых линейных контрастах вида $L = \mu_i - \mu_j$, $i = 1,2,...,k-1$; $j = i+1,...,k$.
Вычисление критерия при проверке нулевой гипотезы $L = L_0$ производится по формуле

$$t = \frac{\sum_{i=1}^{k} c_i \bar{x}_i - L_0}{\frac{1}{2}\sqrt{M}\sum_{i=1}^{k}|c_i|}\sqrt{m},$$

где $M = \frac{1}{k(m-1)}\sum_{i=1}^{k}\sum_{j=1}^{m}\left(x_{ij} - \bar{x}_{i.}\right)^2$ – средний квадратичный остаток,

$\bar{x}_{i.} = \frac{1}{m}\sum_{j=1}^{m} x_{ij}, i = 1,2,...,k$ – среднее значение i–й выборки,

m – численность каждой выборки,
k – число столбцов (выборок).
Статистика критерия Тьюки подчиняется распределению стьюдентизированного размаха с параметрами k и $k(m-1)$.

Обсуждение см. в книгах Мюллера с соавт., Ликеша с соавт., Бикела с соавт., Гланца, Афифи с соавт.

8.3.1.1.12. Критерий Ливена

Критерий Ливена (Levene’s test for equality of variance) является аналогом критерия Бартлетта.

Перед вычислением статистики критерия выполняется преобразование исходных данных по формуле

$$z_{ij} = \left| x_{ij} - \bar{x}_{i.} \right|, i = 1,2,...,k; j = 1,2,...,n_i,$$

k – число столбцов (выборок),

n_i, $i = 1,2,...,k$ – численность i–й выборки,

$\bar{x}_{i.} = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij}, i = 1,2,...,k$ – среднее значение i–й выборки (групповое среднее значение).

Вычисление статистики критерия производится по формуле, аналогичной статистике однофакторного дисперсионного анализа,

$$W = \frac{N-k}{k-1} \frac{\sum_{i=1}^{k} n_i \left(\bar{z}_{i.} - \bar{z}_{..} \right)^2}{\sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(z_{ij} - \bar{z}_{i.} \right)^2},$$

где $N = \sum_{i=1}^{k} n_i$ – общая численность,

$\bar{z}_{i.} = \frac{1}{n_i} \sum_{j=1}^{n_i} z_{ij}, i = 1,2,...,k$ – среднее значение i–й выборки,

$\bar{z}_{..} = \frac{1}{N} \sum_{i=1}^{k} \sum_{j=1}^{n_i} z_{ij}$ – общее среднее значение.

Статистика критерия имеет F–распределение с параметрами $k - 1$ и $N - k$.

См. монографии Шукри (Shoukri) с соавт.

8.3.1.1.13. Критерий Брауна–Форсайта

Критерий Брауна–Форсайта (Brown–Forsythe test for equality of group variances) является вариантом критерия Ливена. Перед вычислением статистики критерия выполняется преобразование исходных данных по формуле

$$z_{ij} = \left| x_{ij} - \tilde{x}_{i.} \right|, i = 1,2,...,k; j = 1,2,...,n_i,$$

k – число столбцов (выборок),

n_i, $i = 1,2,...,k$ – численность i–й выборки,

$\tilde{x}_{i.}, i = 1,2,...,k$ – медиана i–й выборки (групповая медиана).

Вычисление статистики критерия производится по формуле, аналогичной статистике однофакторного дисперсионного анализа,

$$W = \frac{N-k}{k-1} \frac{\sum_{i=1}^{k} n_i \left(\bar{z}_{i.} - \bar{z}_{..} \right)^2}{\sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(z_{ij} - \bar{z}_{i.} \right)^2},$$

где $N = \sum_{i=1}^{k} n_i$ – общая численность,

$\bar{z}_{i.} = \frac{1}{n_i}\sum_{j=1}^{n_i} z_{ij}, i = 1,2,...,k$ – среднее значение *i*–й выборки,

$\bar{z}_{..} = \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{n_i} z_{ij}$ – общее среднее значение.

Статистика критерия имеет *F*–распределение с параметрами $k - 1$ и $N - k$.

См. руководство NIST/SEMATECH.

8.3.1.1.14. Критерий V Бхапкара

Критерий Бхапкара (*V*–критерий Бхапкара) предназначен для проверки нулевой гипотезы о равенстве параметров положения (сдвига в средних) и масштаба (сдвига в дисперсиях). Вычисление статистики критерия производится по формуле

$$V = (2k-1)\prod_{i=1}^{k} n_i \left\{ \frac{1}{\prod_{i=1}^{k} n_i} \sum_{i=1}^{k} n_i \left(u_i - \frac{1}{k} \right)^2 - \left[\frac{1}{\prod_{i=1}^{k} n_i} \sum_{i=1}^{k} \left(u_i - \frac{1}{k} \right) \right]^2 \right\},$$

где n_i, $i = 1,2,...,k$ – численность *i*–й выборки,
k – число столбцов (выборок),
u_i, $i = 1,2,...,k$ – количества подвыборок в сгенерированных выборках, *i*–ая варианта в которых меньше остальных $k - 1$ вариант; при этом $\prod_{i=1}^{k} n_i$ подвыборок генерируются из исходных выборок таким образом, чтобы в каждой подвыборке была представлена одна варианта каждой из *k* выборок.
Статистика критерия имеет χ^2–распределение с параметром $k - 1$.

См. монографию Кобзаря.

8.3.1.1.15. Критерий D Дешпанде

Критерий Дешпанде (*D*–критерий Дешпанде, Дюфора и Люнга) предназначен для проверки нулевой гипотезы о равенстве параметров масштаба (сдвига в дисперсиях). Вычисление статистики критерия производится по формуле

$$D = \frac{(2k-1)(k-1)^2 C_{2(k-1)}^{k-1} \prod_{i=1}^{k} n_i}{2\left[k^2 + \left(k^2 + 4k + 2\right) C_{2(k-1)}^{k-1}\right]} \left\{ \frac{1}{\prod_{i=1}^{k} n_i} \sum_{i=1}^{k} n_i \left(u_i + v_i\right)^2 - \left[\frac{1}{\prod_{i=1}^{k} n_i} \sum_{i=1}^{k} \left(u_i + v_i\right) \right]^2 \right\},$$

где n_i, $i = 1,2,...,k$ – численность *i*–й выборки,
k – число столбцов (выборок),
u_i, $i = 1,2,...,k$ – количества подвыборок в сгенерированных выборках, *i*–ая варианта в которых меньше остальных $k - 1$ вариант; при этом $\prod_{i=1}^{k} n_i$ подвыборок генерируются из исходных выборок таким образом, чтобы в каждой подвыборке была представлена одна варианта каждой из *k* выборок,

v_i, $i = 1,2,...,k$ – количества подвыборок в сгенерированных выборках, i–ая варианта в которых больше остальных $k - 1$ вариант.
Статистика критерия имеет χ^2–распределение с параметром $k - 1$.

См. монографию Кобзаря.

8.3.1.1.16. Критерий L Дешпанде

Критерий Дешпанде (L–критерий Дешпанде, Дюфора и Люнга) предназначен для проверки нулевой гипотезы о равенстве параметров положения (сдвига в средних). Вычисление статистики критерия производится по формуле

$$L = \frac{(2k-1)(k-1)^2 C_{2(k-1)}^{k-1} \prod_{i=1}^{k} n_i}{2k^2 \left[C_{2(k-1)}^{k-1} - 1\right]} \left\{ \frac{1}{\prod_{i=1}^{k} n_i} \sum_{i=1}^{k} n_i \left(-u_i + v_i\right)^2 - \left[\frac{1}{\prod_{i=1}^{k} n_i} \sum_{i=1}^{k} \left(-u_i + v_i\right) \right]^2 \right\},$$

где n_i, $i = 1,2,...,k$ – численность i–й выборки,
k – число столбцов (выборок),
u_i, $i = 1,2,...,k$ – количества подвыборок в сгенерированных выборках, i–ая варианта в которых меньше остальных $k - 1$ вариант; при этом $\prod_{i=1}^{k} n_i$ подвыборок генерируются из исходных выборок таким образом, чтобы в каждой подвыборке была представлена одна варианта каждой из k выборок,
v_i, $i = 1,2,...,k$ – количества подвыборок в сгенерированных выборках, i–я варианта в которых больше остальных $k - 1$ вариант.
Статистика критерия имеет χ^2–распределение с параметром $k - 1$.

См. монографию Кобзаря.

8.3.1.2. Многофакторный дисперсионный анализ

Исходные данные для двухфакторного дисперсионного анализа представлены в виде таблицы (прямоугольной матрицы), причем число столбцов соответствует числу уровней первого фактора (уровней обработки), число строк равно числу уровней второго фактора (уровней обработки). n строк – блоков наблюдений параметров объектов – расположены в k столбцах, соответствующих видам обработки (видам воздействия на объекты). При этом каждый блок может быть результатом измерений параметров как на одном объекте, так и на группе объектов, например в виде среднего значения какого–либо параметра, вычисленного по всем объектам исследуемой группы при определенном виде воздействия на группу. Следующий блок, таким образом, будет средним значением другого параметра по всем объектам группы при том же виде воздействия.
Предлагаются методы двухфакторного дисперсионного анализа:

- двухфакторный дисперсионный анализ (MANOVA),
- ранговый двухфакторный анализ Фридмана,
- критерий Квейд,
- критерий Пэйджа,
- Q–критерий Кокрена,
- критерий Шеффе для связанных выборок.

8.3.1.2.1. Двухфакторный дисперсионный анализ

Результаты опытов никогда в точности не соответствуют степени влияния на них того или иного признака. Происходит это потому, что на результаты оказывают влияние и неучтенные в условиях эксперимента факторы. При включении в дисперсионный анализ двух и более факторов имеет место многофакторный дисперсионный анализ (MANOVA).
Двухфакторный дисперсионный анализ, иначе называемый дисперсионным анализом по двум признакам (двухфакторный дисперсионный анализ без повторений), применяется для зависимых нормально распределенных выборок. Нулевая гипотеза состоит в утверждении о равенстве эффектов строк между собой и равенстве эффектов столбцов между собой.
Вычисления производятся по формулам:

эффект строк $t_{row} = \frac{D_{row}}{D}$,

эффект столбцов $t_{col} = \frac{D_{col}}{D}$,

где $D_{row} = \frac{1}{r-1} SS_{row}$ – дисперсия, объясняемая строками,

$D_{col} = \frac{1}{c-1} SS_{col}$ – дисперсия, объясняемая столбцами,

$D = \frac{SS - SS_{row} - SS_{col}}{(r-1)(c-1)}$ – остаточная дисперсия,

$SS_{row} = \frac{1}{c} \sum_{i=1}^{r} T_{i.}^2 - \frac{T_{..}^2}{rc}$ – средний квадрат строк,

$SS_{col} = \frac{1}{r} \sum_{j=1}^{c} T_{.j}^2 - \frac{T_{..}^2}{rc}$ – средний квадрат столбцов,

$SS = \sum_{i=1}^{r} \sum_{j=1}^{c} x_{ij}^2 - \frac{T_{..}^2}{rc}$ – средний квадрат погрешности,

$T_{i.} = \sum_{j=1}^{c} x_{ij}, i = 1,2,...,r$ – суммы строк,

$T_{.j} = \sum_{i=1}^{r} x_{ij}, j = 1,2,...,c$ – суммы столбцов,

$T_{..} = \sum_{i=1}^{r} \sum_{j=1}^{c} x_{ij}$ – общая сумма,

c – число столбцов (выборок),
r – число строк (параметров).
Статистика критерия имеет F–распределение с параметрами $r - 1$ и $(r - 1)(c - 1)$ в случае исследования эффекта строк и с параметрами с – 1 и $(r - 1)(c - 1)$ в случае исследования эффекта столбцов.

8.3.1.2.2. Ранговый критерий Фридмана

Если не выполнены предположения, позволяющие провести двухфакторный дисперсионный анализ, применяется свободный от типа распределения непараметрический критерий

Фридмана. Ранговый двухфакторный анализ Фридмана (ранговый критерий Фридмана, Кендалла и Бэбингтона Смита) применяется для проверки нулевой гипотезы о том, что различные методы обработки или иных воздействий на изучаемый объект (процесс) дают одинаковые результаты. Таким образом, нулевая гипотеза состоит в отсутствии эффектов столбцов (эффектов обработки). Критерий может также применяться в качестве непараметрического аналога однофакторного дисперсионного анализа с повторными измерениями.
Вычисление статистики критерия производится по формуле

$$S = \frac{12\sum_{j=1}^{k}(R_j - nR_{\cdot\cdot})^2}{nk(k+1) - \frac{1}{k-1}\sum_{i=1}^{n}\left(\sum_{j=1}^{g_i} t_{ij}^3 - k\right)},$$

где R_j, $j = 1,2,...,k$ – соответствующие суммы рангов в строках,
n – численность каждой совокупности,
k – число эффектов обработки (воздействий, уровней фактора),

$$R_{\cdot\cdot} = \frac{k+1}{2},$$

g_i, $i = 1,2,...,n$ – число связок в блоке,
t_{ij}, $i = 1,2,...,n$; $j = 1,2,...,g_i$ – численность соответствующей связки, равная 1 при отсутствии связок в блоке.
Суммы рангов вычисляются по формуле

$$R_j = \sum_{i=1}^{n} r_{ij},\ j = 1,2,...,k,$$

где r_{ij}, $i = 1,2,...,n$; $j = 1,2,...,k$ – ранги, причем ранжирование производится по каждой строке отдельно.
Статистика критерия имеет χ^2–распределение с параметром $k - 1$.

Описание см. в книгах Хотеллингом с соавт., Браунли, Петровича с соавт., в справочнике Оуэна.

8.3.1.2.3. Критерий Квейд

Если не выполнены предположения, позволяющие провести двухфакторный дисперсионный анализ, применяется свободный от типа распределения непараметрический ранговый критерий Квейд (Quade's test). Нулевая гипотеза состоит в отсутствии эффектов столбцов.
Вычисление статистики критерия производится по формуле

$$S = \frac{n-1}{n}\sum_{j=1}^{k} T_j^2 \left[\sum_{i=1}^{n}\sum_{j=1}^{k} R_{ij}^2 - \frac{1}{n}\sum_{j=1}^{k} T_j^2\right]^{-1},$$

где R_{ij}, $i = 1,2,...,n$; $j = 1,2,...,k$ – скорректированные ранги,
T_j, $j = 1,2,...,k$ – суммы столбцов матрицы скорректированных рангов,
n – численность каждой совокупности,
k – число эффектов обработки (воздействий, уровней фактора).
Скорректированные ранги рассчитываются по формуле

$$R_{ij} = Q_i \cdot \left(r_{ij} - \frac{k+1}{2}\right), i = 1,2,...,n;\ j = 1,2,...,k,$$

где r_{ij}, $i = 1,2,...,n$; $j = 1,2,...,k$ – ранги, причем ранжирование производится отдельно по

каждой строке матрицы исходных данных,
Q_j, $j = 1,2,...,n$ – ранги размахов строк матрицы исходных данных.
Статистика критерия имеет F–распределение с параметрами $k - 1$ и $(k - 1)(n - 1)$.

См. книгу Петровича с соавт., работы Понтеса (Pontes, 2000), Солиани (Soliani) с соавт., статьи Кемпбелл (Campbell), Теодорссон–Норхайм (Theodorsson–Norheim).

8.3.1.2.4. Критерий Пэйджа

Критерий Пэйджа (критерий L Пэйджа, дисперсионный анализ Пэйджа) предназначен для проверки нулевой гипотезы о равенстве эффектов обработки (воздействия) на выборки с неизвестными, но равными средними. Нулевая гипотеза состоит в утверждении о равенстве эффектов строк между собой и равенстве эффектов столбцов между собой. Статистика критерия вычисляется по формуле

$$L = \sum_{i=1}^{k} iR_i,$$

где R_i, $i = 1,2,...,k$ – упорядоченные по возрастанию суммы рангов блоков,
k – число эффектов обработки (воздействий, уровней фактора).

Для больших выборок распределение преобразованной статистики $\frac{L - ML}{\sqrt{DL}}$ является приближенно нормальным. Здесь математическое ожидание и дисперсия рассчитываются по формулам, соответственно:

$$ML = \frac{1}{4} nk(k+1)^2,$$

$$DL = \frac{n\left(k^3 - k\right)^2}{144(k-1)},$$

где n – численность каждой совокупности.

См. источники: Лисенков, Тюрин с соавт., Холлендер с соавт.

8.3.1.2.5. Критерий Q Кокрена

Критерий Q Кокрена используется в случае, если группы однородных субъектов подвергаются более чем двум экспериментальным воздействиям, и их ответы носят двухвариантный (бинарный, дихотомический) характер. Предполагается, что 0 означает отрицательный ответ, 1 – положительный. Каждая выборка представляет собой измерения одного условия по всем группам. Варианты выборки, таким образом – это измерения в рассматриваемых группах по данному условию. Нулевая гипотеза состоит в том, что в генеральной совокупности доли всех экспериментальных условий равны. Вычисление производится по формуле

$$Q = \frac{(c-1)\left(c\sum_{j=1}^{c} T_{\cdot j}^2 - \left(\sum_{j=1}^{c} T_{\cdot j}\right)^2\right)}{c\sum_{i=1}^{r} T_{i\cdot} - \sum_{i=1}^{r} T_{i\cdot}^2},$$

где $T_{\cdot j} = \sum_{i=1}^{r} x_{ij}, j = 1,2,...,c$ – суммы столбцов,

$T_{i.} = \sum_{j=1}^{c} x_{ij}, i = 1,2,...,r$ – суммы строк,
c – число столбцов (выборок),
r – число строк (параметров).
Статистика критерия имеет χ^2–распределение с параметром $c - 1$.

Описание см. у Браунли.

8.3.1.2.6. Критерий Шеффе для связанных выборок

Двухфакторный позволяет обнаружить существование эффектов столбцов (эффектов обработки) в таблице дисперсионного анализа. Однако он не дает возможности точно указать столбцы, которые обладают нулевыми эффектами. Для решения проблемы применяется метод множественного сравнения Шеффе для связанных выборок (парный критерий Шеффе). Критерий Шеффе для связанных выборок предназначен для проверки так называемой гипотезы о линейном контрасте. Линейный контраст

$$L = \sum_{i=1}^{k} c_i \mu_i$$

представляет собой линейную функцию от средних значений μ_i, $i = 1,2,...,k$, k независимых нормальных выборок с неизвестными, но равными дисперсиями, и известных констант c_i, $i = 1,2,...,k$, удовлетворяющих условию

$$\sum_{i=1}^{k} c_i = 0.$$

В частном случае проверяется серия гипотез о простых линейных контрастах вида
$L_0 = \mu_i - \mu_j$, $i = 1,2,...,k-1$; $j = i + 1,...,k$.
Вычисление статистики критерия производится по формуле

$$t = \frac{\left(\sum_{i=1}^{r} c_i \bar{x}_i\right)^2}{\frac{(r-1)S}{c}\sum_{i=1}^{r} c_i^2},$$

где $S = \sum_{i=1}^{r}\sum_{j=1}^{c} x_{ij}^2 - \frac{T_{..}^2}{rc}$ – остаточный средний квадрат,

$T_{..} = \sum_{i=1}^{r}\sum_{j=1}^{c} x_{ij}$ – общая сумма,

$\bar{x}_{i.} = \frac{1}{n_i}\sum_{j=1}^{n_i} x_{ij}, i = 1,2,...,k$ – среднее значение i–й выборки,

c – число столбцов (выборок),
r – число строк (параметров).
Статистика критерия имеет F–распределение с параметрами $r - 1$ и $(r - 1)(c - 1)$ в случае исследования эффекта строк и с параметрами $c - 1$ и $(r - 1)(c - 1)$ в случае исследования эффекта столбцов.

См. справочники Полларда (Pollard), Полларда.

8.3.2. Множественные сравнения

Методы множественного сравнения применяются, если исходные данные представлены многомерными выборками. В разделе предлагаются несколько популярных методов множественных сравнений, представляющих собой обобщения методов проверки гипотез (в т.ч. дисперсионного анализа) на многомерные выборки.
Для параметрических и непараметрических методов проверки гипотез (см. главы «Параметрическая статистика» и «Непараметрическая статистика») и дисперсионного анализа существуют многомерные аналоги в множественном сравнении, как показано в таблице.

Метод проверки гипотезы для двух выборок и дисперсионного анализа	Многомерный «функциональный аналог» из множественных сравнений
Параметрические тесты	
Критерий Стьюдента	Критерий Хотеллинга
Критерий Уэлча	Критерий Джеймса–Сю
F–критерий	Критерий Кульбака (2 многомерные выборки)
Критерий Бартлетта	Критерий Кульбака ($k > 2$ выборок)
	Критерий Уилкса
Непараметрические тесты	
Критерий Вилкоксона	Критерий Пури–Сена–Тамура
Критерий Муда	Критерий Пури–Сена
Критерий Краскела–Уоллиса	Критерий Шейрера–Рэя–Хэйра (2 многомерные выборки)

Из методов данного класса в программе представлены:
- критерий Хотеллинга,
- критерий Джеймса–Сю,
- критерий Кульбака,
- критерий Пури–Сена–Тамура,
- критерий Пури–Сена.
- критерий Шейрера–Рэя–Хэйра.

В главе для полноты информации описан также критерий Уилкса и даны рекомендации по его самостоятельному вычислению.
Исходные данные для множественных сравнений представлены в виде таблиц (прямоугольных матриц). Каждой выборке соответствует одна матрица, причем число столбцов каждой матрицы соответствует размерности многомерной выборки, число строк равно числу наблюдений. При этом выборки могут иметь как одинаковое число вариант (равные объемы), так и различное, в зависимости от требований применяемого метода. Размерности сравниваемых многомерных выборок должны быть одинаковы.

Обзор см. в диссертации Понтеса (Pontes, 2005). См. также источники: Родионов с соавт., Коган с соавт., Пури (Puri) с соавт., Сен (Sen) с соавт., Тамура (Tamura) и Цвик (Zwick).

8.3.2.1. Критерий Хотеллинга

Критерий T^2 (критерий следа, критерий Хотеллинга, критерий Лоули и Хотеллинга), для случая двух многомерных выборок предложенный Хоттелингом, применяется в задаче статистической проверки гипотезы о равенстве векторов средних двух многомерных совокупностей. Предполагается, что многомерные выборки извлечены из нормальных многомерных распределений с равными между собой ковариационными матрицами.
Статистика критерия Хотеллинга вычисляется по формуле

$$T^2 = \frac{n_1 n_2}{n_1 + n_2}\left(\bar{x}_1 - \bar{x}_2\right)' S^{-1}\left(\bar{x}_1 - \bar{x}_2\right),$$

где n_1 – количество многомерных вариант первой многомерной выборки,
n_2 – количество многомерных вариант второй многомерной выборки,
$\bar{x}_1$ и $\bar{x}_2$ – векторы средних двух многомерных совокупностей,
S – дисперсионно–ковариационная матрица совокупности.
Если дисперсионно–ковариационная матрица совокупности неизвестна, она вычисляется через выборочные дисперсионно–ковариационные матрицы совокупностей по формуле:

$$S = \frac{1}{n_1 + n_2 - 2}\left[\left(n_1 - 1\right)S_1 + \left(n_2 - 1\right)S_2\right],$$

где S_1 и S_2 – выборочные дисперсионно–ковариационные матрицы многомерных совокупностей.
Модифицированная статистика критерия $\frac{n_1 + n_2 - m - 1}{m(n_1 + n_2 - 2)}T^2$ имеет F–распределение с параметрами m и $n_1 + n_2 - m - 1$, где m – размерность каждой выборки.

Описание см. у Андерсона, Афифи с соавт., Джонсона с соавт., Кульбака, Мэйндоналда, Хальда, в справочнике под редакцией Ллойда с соавт. Связь с расстоянием Махаланобиса выведена Уилксом.

8.3.2.2. Критерий Джеймса–Сю

Критерий Джеймса–Сю предназначен для проверки гипотезы о равенстве векторов средних двух многомерных совокупностей. Предполагается, что многомерные выборки извлечены из нормальных многомерных распределений с неизвестными или неравными между собой ковариационными матрицами. Критерий является решением многомерной проблемы Беренса–Фишера (см. главу «Параметрическая статистика»).
Статистика критерия вычисляется по формуле

$$2I = \left(\bar{x}_1 - \bar{x}_2\right)' S^{-1}\left(\bar{x}_1 - \bar{x}_2\right),$$

где $\bar{x}_1$ и $\bar{x}_2$ – векторы средних двух многомерных совокупностей,
S – дисперсионно–ковариационная матрица совокупности.
Дисперсионно–ковариационная матрица совокупности вычисляется через выборочные дисперсионно–ковариационные матрицы совокупностей по формуле:
$S = S_1 / n_1 + S_2 / n_2$,
где n_1 – количество многомерных вариант первой многомерной выборки,
n_2 – количество многомерных вариант второй многомерной выборки,
S_1 и S_2 – выборочные дисперсионно–ковариационные матрицы многомерных совокупностей.
Статистика критерия $2I$ подчиняется асимптотически распределению χ^2 с m степенями свободы.

Критерий описан Родионовым с соавт.

8.3.2.3. Критерий Кульбака

Критерий Кульбака предназначен для проверки равенства ковариационных матриц двух или более многомерных совокупностей. Предполагается, что многомерные выборки извлечены из совокупностей, подчиняющихся нормальным многомерным распределениям. Для двух выборок статистика критерия может вычисляться по формуле

$$2I_0 = \sum_{i=1}^{2}(n_i - 1)\ln\frac{|S|}{|S_i|},$$

где n_1 и n_2 – количества многомерных вариант сравниваемых совокупностей,
S_1 и S_2 – выборочные дисперсионно–ковариационные матрицы многомерных совокупностей,
|.| – определитель матрицы.
Статистика критерия $2I_0$ подчиняется асимптотически B–распределению Фишера, иначе нецентральному распределению χ^2 с параметром нецентральности

$$\lambda = \frac{(2m^3 + 2m^2 - m)}{12}\left(\frac{1}{n_1 - 1} + \frac{1}{n_2 - 1} - \frac{1}{n_1 + n_2 - 2}\right)$$

и $m(m + 1) / 2$ степенями свободы, где m – размерность каждой выборки.

Критерий представил Кульбак, подробно описали Родионов с соавт.

8.3.2.4. Критерий Пури–Сена–Тамура

Ранговый непараметрический критерий Пури–Сена–Тамура предназначен для проверки гипотезы о равенстве векторов средних двух многомерных совокупностей.
Статистика критерия вычисляется по формуле

$$\Lambda = \sum_{i=1}^{2}(\bar{r}_i - \bar{r})' S^{-1}(\bar{r}_i - \bar{r}),$$

где $\bar{r}_1$ и $\bar{r}_2$ – векторы средних рангов двух многомерных совокупностей,
$\bar{r}$ – вектор средних рангов объединенной совокупности,
S – дисперсионно–ковариационная матрица рангов объединенной совокупности.
Статистика критерия подчиняется асимптотически распределению χ^2 с m степенями свободы, где m – размерность каждой выборки.

Критерий описан Родионовым с соавт., где рассмотрены также случаи использования иных ранговых отметок.

8.3.2.5. Критерий Пури–Сена

Ранговый непараметрический критерий Пури–Сена предназначен для проверки равенства ковариационных матриц двух многомерных совокупностей.
Статистика критерия вычисляется по формуле

$$\Lambda = \sum_{i=1}^{2}(\bar{e}_i - \bar{e})' S^{-1}(\bar{e}_i - \bar{e}),$$

где $\bar{e}_1$ и $\bar{e}_2$ – векторы средних ранговых отметок двух многомерных совокупностей,
$\bar{e}$ – вектор средних ранговых отметок объединенной совокупности,

S – дисперсионно–ковариационная матрица ранговых отметок объединенной совокупности. При этом ранговые отметки вычисляются как

$$E_{ij} = \left(\frac{R_{ij}}{N+1} - 0{,}5 \right)^2, i = 1,2,...,N;\ j = 1,2,...,m,$$

где R_{ij}, $i = 1,2,...,N$; $j = 1,2,...,m$ – ранги соответствующей выборки,
N – численность соответствующей выборки: n_1 или n_2 – многомерных вариант сравниваемых совокупностей, $n_1 + n_2$ – объединенной совокупности,
m – размерность каждой выборки.
Статистика критерия подчиняется асимптотически распределению χ^2 с m степенями свободы.

Критерий описан Родионовым с соавт., где рассмотрены также случаи использования иных ранговых отметок.

8.3.2.6. Критерий Шейрера–Рэя–Хэйра

Критерий Шейрера–Рэя–Хэйра представляет собой многомерное расширение критерия Краскела–Уоллиса. Критерий парный, т. е. формально представленные для анализа выборки должны иметь равные количества строк и столбцов. При этом предполагается, что по n строкам располагаются значения k–мерных выборочных значений.
В программе представлен вариант критерия для анализа двух многомерных выборок. Алгоритм также описан для двух выборок, хотя может быть обобщен на произвольное количество многомерных выборок. Пусть даны две многомерных выборки: X_{ij}, $i = 1,2,...,n$; $j = 1,2,...,k$, и Y_{ij}, $i = 1,2,...,n$; $j = 1,2,...,k$. Многомерные выборки совместно ранжируются по убыванию. При этом совпадающим значениям присваиваются средние (по связке) ранги. В результате ранжирования получаются массивы рангов, соответственно, R_{ij}, $i = 1,2,...,n$; $j = 1,2,...,k$, и S_{ij}, $i = 1,2,...,n$; $j = 1,2,...,k$.
Затем составляется таблица размером 2 x k. В ячейки таблицы записываются суммы рангов, вычисленные по формулам, соответственно,

$$T_{1j} = \sum_{i=1}^{n} R_{ij},\ j = 1,2,...,k,$$

$$T_{2j} = \sum_{i=1}^{n} S_{ij},\ j = 1,2,...,k.$$

Вычисления статистик критерия производятся по формулам:
эффект столбцов $H_c = \frac{RSS_c}{RMS_{total}}$,
эффект строк $H_r = \frac{RSS_r}{RMS_{total}}$,
эффект взаимодействия строк и столбцов $H_{rc} = \frac{RSS_{rc}}{RMS_{total}}$,
где квадратичные остатки вычисляются по формулам:

$$RSS_c = \frac{1}{2k} \sum_{j=1}^{k} \left(\sum_{i=1}^{2} T_{ij} \right)^2 - \frac{N(N+1)^2}{4},$$

$$RSS_r = \frac{1}{nk} \sum_{i=1}^{2} \left(\sum_{j=1}^{k} T_{ij} \right)^2 - \frac{N(N+1)^2}{4},$$

$$RSS_{rc} = \frac{1}{n}\sum_{i=1}^{2}\sum_{j=1}^{k}T_{ij}^2 - \frac{N(N+1)^2}{4} - RSS_c - RSS_r,$$

$$RMS_{total} = \frac{N(N+1)}{12},$$

где $N = 2nk$ – общая численность представленных выборок.
Статистика H_c имеет χ^2–распределение с параметром $k - 1$. Статистика H_r имеет χ^2 – распределение с параметром 1. Статистика H_{rc} (в случае двух выборок) имеет χ^2 – распределение с параметром $k - 1$.

См. монографию Сокала (Sokal) с соавт., статью Шайрера (Scheirer) с соавт.

8.3.2.7. Критерий Уилкса

Критерий λ Уилкса предназначен для выполнения однофакторного многомерного дисперсионного анализа. Его можно считать обобщением множественного критерия Хотеллинга на случай $k > 2$ многомерных выборок. Предполагается, что многомерные выборки извлечены из нормальных многомерных распределений с равными между собой ковариационными матрицами.
Статистика критерия вычисляется по формуле

$$\lambda = \frac{|W|}{|W+B|},$$

где W – общая матрица внутригруппового разброса,
B – матрица межгруппового разброса,
$|.|$ – операция вычисления определителя.
Элемент w_{ij} матрицы W вычисляется как

$$w_{ij} = \sum_{r=1}^{k} s_{ij}^{(r)}, i = 1,2,...,m;\ j = 1,2,...,m,$$

где $s_{ij}^{(r)}, r = 1,2,...,k,$ – элемент т.н. Матрицы $S^{(r)}$ остаточных сумм квадратов и произведений выборки r,
k – количество многомерных выборок,
m – число переменных (размерность) каждой многомерной выборки,
r, $r = 1,2,...,k$ – верхний индекс, означающий номер многомерной выборки.
Элемент $s_{ij}^{(r)}$ матрицы $S^{(r)}$ вычисляется как

$$s_{ij}^{(r)} = \frac{1}{n-k}\sum_{l=1}^{n_r}(x_{il}^{(r)} - \bar{x}_i^{(r)})^T(x_{jl}^{(r)} - \bar{x}_j^{(r)}), i = 1,2,...,m;\ j = 1,2,...,m,$$

где $x_{il}^{(r)}, i = 1,2,...,n_r,$ – значение варианты переменной i,
$x_{jl}^{(r)}, j = 1,2,...,n_r,$ – значение варианты переменной j,
$\bar{x}_i^{(r)}, i = 1,2,...,m,$ – среднее значение переменной i,
$\bar{x}_j^{(r)}, j = 1,2,...,m,$ – среднее значение переменной j,
n_r, $r = 1,2,...,k$ – численность выборки r (число m–мерных вариант в каждой многомерной выборке),

$n = \sum_{r=1}^{k} n_r$ – общее количество многомерных выборок.
Элемент b_{ij} матрицы B вычисляется как

$$b_{ij} = \sum_{r=1}^{k} n_r \bar{x}_i^{(r)} \bar{x}_j^{(r)} - n\bar{x}_i \bar{x}_j, i = 1,2,...,m; j = 1,2,...,m,$$

где $\bar{x}_i, i = 1,2,...,m,$ – среднее значение переменной i по всем k выборкам,
$\bar{x}_j, j = 1,2,...,m,$ – среднее значение переменной j по всем k выборкам.
Модифицированная статистика

$$\hat{\lambda} = -\left(n - 1 - \frac{m+k}{2}\right)\ln\lambda$$

подчиняется χ^2–распределению с $m(k-1)$ степенями свободы (аппроксимация Бартлетта).
В программе критерий λ Уилкса непосредственно не реализован. Однако не следует думать, что он представляет только теоретический интерес, т. к. программное обеспечение содержит весь набор необходимых инструментов. Если пользователю необходимо применить критерий Уилкса, рекомендуем произвести вычисления с помощью методов главы «Матричная и линейная алгебра». Ковариационные матрицы можно вычислить с помощью соответствующего метода, представленного в главе «Корреляционный анализ».

Метод представлен Андерсоном, Афифи с соавт., Петровичем с соавт. Андерсон указал точное распределение статистики критерия, а в последних двух источниках представлена также аппроксимация статистики F–распределением, предложенная Рао. Кульбак представил свой тест также и для $k > 2$ многомерных выборок.

8.3.3. Ковариационный анализ

Однофакторный ковариационный анализ использует концепции однофакторного дисперсионного анализа и линейной регрессии. Предполагается, что исходные данные представляют собой совокупность регрессий («предиктор–зависимая переменная» или, в смысле построения графика регрессии, «абсцисса–ордината»), соответствующих различным уровням значения качественного признака. При этом сами значения качественного признака не вводятся.
Метод позволяет протестировать ряд статистических гипотез, как указано в соответствующем разделе.
Предпосылки применения ковариационного анализа:

- нормальность распределения ошибок (относительно линейной регрессии),
- однородность дисперсии ошибок,
- зависимость отклика от количественного предиктора линейна,
- равный наклон регрессий на уровнях качественного фактора.

Если данные не удовлетворяют представленным требованиям, они могут быть преобразованы соответствующими методами.

8.3.3.1. Однофакторный ковариационный анализ

Однофакторный ковариационный анализ (one–way ANCOVA) использует концепции однофакторного дисперсионного анализа, линейной регрессии и множественных сравнений. Представление исходных данных для расчета имеет свою особенность. Массив вводится в виде совокупности k регрессий (иначе – уровней, соответствующих градациям качественного фактора), как показано на следующей иллюстрации, причем численности n_i, $i = 1,2,...,k$, пар

предикторов x_{ij}, $i = 1,2,...,k$; $j = 1,2,...,n_i$, и зависимых переменных y_{ij}, $i = 1,2,...,k$; $j = 1,2,...,n_i$, могут полагаться как равными, так и различными:

Уровни качественного фактора (не вводятся)							
Уровень 1		Уровень 2		...		Уровень k	
x_{11}	y_{11}	x_{21}	y_{21}	...	...	x_{k1}	y_{k1}
x_{12}	y_{12}	x_{22}	y_{22}	...	...	x_{k2}	y_{k2}
...	...	...	...	...	...	...	...
y_{1n_1}	y_{1n_1}	x_{2n_2}	y_{2n_2}	...	...	x_{kn_k}	y_{kn_k}

Как и в некоторых методах дисперсионного анализа, оперирующими неравными по численности столбцами данных, в представленном методе, оперирующим неравными по численности парами столбцов, выделение интервала исходных данных производится точно таким же образом, чтобы охватить максимальную по численности регрессию. При этом пустые ячейки для регрессий меньшей численности будут учтены автоматически численностями n_i, $i = 1,2,...,k$, ибо ввод каждой регрессии будет прекращен, как только встретится пустая ячейка, и программа перейдет к считыванию следующей регрессии. Обозначим:

$\bar{x}_{i.} = \frac{1}{n_i}\sum_{j=1}^{n_i} x_{ij}, i = 1,2,...,k,$ – среднее значение предиктора на уровне i, $i = 1,2,...,$k,

$\bar{y}_{i.} = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij}, i = 1,2,...,k,$ – среднее значение зависимой переменной на уровне i, $i = 1,2,...,$k,

$\bar{x}_{..} = \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{n_i} x_{ij}$ – общее среднее значение предикторов на всех уровнях,

$\bar{y}_{..} = \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij}$ – общее среднее значение зависимых переменных на всех уровнях.

где k – количество уровней качественного фактора,
N – общая численность пар предикторов и зависимых переменных, вычисляемая как

$N = \sum_{i=1}^{k} n_i$.

С учетом введенных обозначений суммы квадратов и смешанные произведения между уровнями вычисляются как:

$M_{xx} = \sum_{i=1}^{k} n_i(\bar{x}_{i.} - \bar{x}_{..})^2$, $M_{xy} = \sum_{i=1}^{k} n_i(\bar{x}_{i.} - \bar{x}_{..})(\bar{y}_{i.} - \bar{y}_{..})$ и $M_{yy} = \sum_{i=1}^{k} n_i(\bar{y}_{i.} - \bar{y}_{..})^2$.

Соответственно, суммы квадратов и смешанные произведения внутри всех уровней находятся как:

$E_{xx} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_{i.})^2$, $E_{xy} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_{i.})(y_{ij} - \bar{y}_{i.})$ и $E_{yy} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i.})^2$.

Полные суммы квадратов и смешанные произведения будут:

$T_{xx} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_{..})^2$, $T_{xy} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_{..})(y_{ij} - \bar{y}_{..})$ и $T_{yy} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{..})^2$.

С учетом введенных обозначений, некоторые средние квадраты, необходимые для

дальнейшего конструирования статистик, вычисляются как:

$$MS_M = \frac{1}{k-1}\left(M_{yy} - \frac{T_{xy}^2}{T_{xx}} + \frac{E_{xy}^2}{E_{xx}}\right), \quad MS_Z = \frac{E_{xy}^2}{E_{xx}}, \quad MS_E = \frac{1}{N-k-1}\left(E_{yy} - \frac{E_{xy}^2}{E_{xx}}\right),$$

$$MS_B = \frac{1}{k-1}\left(B_M - \frac{E_{xy}^2}{E_{xx}}\right), \quad MS_R = \frac{1}{N-2k}\left(E_{yy} - B_M\right),$$

где B_M – сумма средних квадратов внутри уровней, вычисляемая по формуле

$$B_M = \sum_{i=1}^{k} b_i,$$

где b_i, $i = 1,2,...,k$ – групповые коэффициенты регрессии.

Групповые коэффициенты регрессии вычисляются как

$$b_i = \frac{E_{xy(i)}^2}{E_{xx(i)}}, i = 1,2,...,k,$$

где $E_{xy(i)}$, $i = 1,2,...,k$ – смешанные произведения внутри уровней,

$E_{xx(i)}$, $i = 1,2,...,k$ – суммы квадратов внутри уровней.

Данные параметры вычисляются, соответственно, по формулам:

$$E_{xy(i)} = \sum_{j=1}^{n_i}(x_{ij} - \bar{x}_{i.})(y_{ij} - \bar{y}_{i.}), i = 1,2,...,k, \quad \text{и} \quad E_{xx(i)} = \sum_{j=1}^{n_i}(x_{ij} - \bar{x}_{i.})^2, i = 1,2,...,k.$$

Программой стандартно выводятся вычисленные ранее средние значения на уровнях, групповые коэффициенты регрессии, а также скорректированные (adjusted) групповые средние значения на уровнях, вычисляемые по формуле

$$\bar{y}_{i(adj)} = \bar{y}_{i.} - \hat{\beta}(\bar{x}_{i.} - \bar{x}_{..}), i = 1,2,...,k,$$

где $\hat{\beta}$ – оценка коэффициента регрессии, вычисляемая по формуле

$$\hat{\beta} = \frac{E_{xy}^2}{E_{xx}}.$$

Отметим, что данный параметр выводится программой под наименованием «B».

Соответствующие стандартные ошибки скорректированных групповых средних значений вычисляются по формуле

$$S\bar{y}_{i(adj)} = \sqrt{MS_E\left(\frac{1}{n_i} + \frac{(\bar{x}_{i.} - \bar{x}_{..})^2}{E_{xx}}\right)}, i = 1,2,...,k.$$

Доверительные интервалы оцениваемых скорректированных групповых средних значений вычисляются по формуле

$$I_{\bar{y}_{i(adj)}} = \left(\bar{y}_{i(adj)} - \sqrt{0{,}5\,|\,m_{p,k',f}\,|}S\bar{y}_{i(adj)}; \bar{y}_{i(adj)} + \sqrt{0{,}5\,|\,m_{p,k',f}\,|}S\bar{y}_{i(adj)}\right), i = 1,2,...,k,$$

где $|m_{p,k',f}|$ – значение обратной функции распределения стьюдентизированного максимума модулей,

$p = 0{,}95$ – строится 95% доверительный интервал,

$k' = k(k-1)/2$,

$f = N - k - 1$.

Далее остановимся на регрессиях, которые могут быть построены по данным ковариационного анализа. Групповые регрессии задаются уравнениями

$$\hat{y} = \bar{y}_{i.} + b_i(x - \bar{x}_{i.}), i = 1,2,...,k.$$

Можно построить групповые регрессии, используя оценку коэффициента регрессии $\hat{\beta}$ (при этом мы получим параллельные групповые регрессии):

$\hat{y} = \bar{y}_{i.} + \hat{\beta}(x - \bar{x}_{i.}), i = 1,2,...,k.$

Возможно построение регрессии средних значений

$\hat{y} = \bar{y}_{..} + \hat{\beta}_M (x - \bar{x}_{..}), i = 1,2,...,k,$

где $\hat{\beta}_M$ – оценка коэффициента регрессии средних значений, вычисляемая по формуле

$$\hat{\beta}_M = \frac{M_{xy}}{M_{xx}}.$$

Возможно также построение полной регрессии

$\hat{y} = \bar{y}_{..} + \hat{\beta}_T (x - \bar{x}_{..}), i = 1,2,...,k,$

где $\hat{\beta}_T$ – оценка полного коэффициента регрессии, вычисляемая по формуле

$$\hat{\beta}_T = \frac{T_{xy}}{T_{xx}}.$$

Все указанные выше параметры выводятся программой. Кроме того, для всех групповых регрессий, регрессии средних значений и полной регрессии выводится свободный член уравнения соответствующей регрессии в стандартной форме

$y = a + bx,$

где a – свободный член уравнения,

b – коэффициент регрессии (здесь и далее подставить соответствующее значение из показанных выше уравнений регрессий),

x – значение предиктора,

y – значение зависимой переменной.

Свободный член вычисляется как

$a = \bar{y} - b\bar{x},$

где $\bar{x}$ – среднее значение предиктора,

$\bar{y}$ – среднее значение зависимой переменной.

На основе результатов предыдущих вычислений в программе последовательно выводятся значения статистик и двусторонних P-значений для проверки следующих статистических гипотез, обычно исследуемых в данном типе анализа.

Гипотеза о равенстве скорректированных групповых средних значений: статистика

$$S_m = \frac{MS_M}{MS_E}$$

подчиняется F-распределению со степенями свободы $k - 1$ и $N - k - 1$.

Гипотеза о равенстве наклона регрессии средних значений нулю: статистика

$$S_g = \frac{MS_Z}{MS_E}$$

подчиняется F-распределению со степенями свободы 1 и $N - k - 1$.

Гипотеза о равенстве наклонов групповых регрессий: статистика

$$S_b = \frac{MS_B}{MS_R}$$

подчиняется F-распределению со степенями свободы $k - 1$ и $N - 2k$.

Данные статистики выводятся программой под наименованиями, соответственно, «Sm», «Sg» и «Sb». Выводятся также двусторонние P-значения для данных статистик.

В источниках встречаются и другие статистики для проверки различных гипотез, сформулированных для рассматриваемого типа статистического анализа.

См. монографии Афифи с соавт., Вайлдта (Wildt) с соавт., Милликена (Milliken) с соавт., Сокала (Sokal) с соавт., Вестфалла (Westfall) с соавт., Сахаи (Sahai) с соавт., статью и монографию Хсу (Hsu), монографию под ред. Эдвардс (Edwards).

Список использованной и рекомендуемой литературы

1. Adichie J.N. Ranking in analysis of covariance tests // Communications in Statistics – Theory and Methods, 1975, vol. 4, no. 9, pp. 883–890.
2. Algina J., Keselman H.J. Testing repeated measures hypotheses when covariance matrices are heterogeneous: Revisiting the robustness of the Welch–James test // Multivariate Behavioral Research, 1997, vol. 32, no. 3, pp. 255–274.
3. Amini S.B., Woolson R.F. Rank covariance methods for the analysis of survival data // Biometrical Journal, 18 January 2007, vol. 33, no. 4, pp. 429–439.
4. Antony N.M., Brown, T.A., Barlow D.H. Current perspectives on panic and panic disorder // Current Directions in Psychological Science, 1992, vol. 1, pp. 79–82.
5. Armstrong R., Hilton A. The use of analysis of variance (ANOVA) in applied microbiology // Microbiologist, December 2004, pp. 18–21.
6. Atil H., Unver Y. Multiple comparisons // OnLine Journal of Biological Sciences, 2001, vol. 1, no. 8, pp. 723–727.
7. Bao P., Ananda M.M.A. Performance of two–way ANOVA procedures when cell frequencies and variances are unequal // Communications in Statistics – Simulation and Computation, 2001, vol. 30, no. 4, pp. 805–830.
8. Barankin E.W. Extension of the Romanovsky–Bartlett–Scheffe test // Proceedings of the Berkeley symposium on mathematical statistics and probability, August 13–18, 1945 and January 27–29, 1946 / Ed. by J. Neyman. – Berkeley, CA: University of California Press, 1949, pp. 433–449.
9. Bartholomew D.J. A test of homogeneity for ordered alternatives // Biometrika, 1959, vol. 41, pp. 36–48.
10. Bartholomew D.J. Ordered tests in the analysis of variance // Biometrika, 1961, vol. 48, pp. 325–332.
11. Bartlett M.S. Properties of sufficiency and statistical tests // Proceedings of the Royal Society of London, 1937, Series A 160, pp. 268 –282.
12. Benjamini Y., Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing // Journal of the Royal Statistical Society: Series B (Statistical Methodology), 1995, vol. 57, pp. 125–133.
13. Bilodeau M., Brenner D. Theory of multivariate statistics. – New York, NY: Springer–Verlag, 1961.
14. Bock R.D., Haggard E. A. The use of multivariate analysis of variance in behavioral research // Handbook of Measurement and Assessment in Behavioral Sciences / Ed. by D.K. Whitla. – Reading, MA: Addison–Wesley 1968.
15. Bortz J., Lienert G.A., Boehnke K. Verteilungsfreie methoden in der biostatistik. – Heidelberg: Springer Medizin Verlag, 2008.
16. Bradley R.A., Patel K.M., Wackerly D.D. Approximate small–sample distributions for multivariate two–sample nonparametric tests // Biometrics, Sep., 1971, vol. 27, no. 3, pp. 515–530.
17. Bray J.H., Maxwell S.E. Analyzing and interpreting significant MANOVAs // Review of Educational Research, 1982, vol. 52, pp. 340–367.

18. Brown M.B., Feng S. Williams' test comparing multiple dose groups to a control // Bulletin of the International Statistical Institute, 52nd Session, Proceedings, Tome LVIII, Finland 1999. Contributed Paper Meeting 57: Statistics in medicine.
19. Brown M.B., Forsythe A.B. Robust tests for equality of variances // Journal of the American Statistical Association, 1974, vol. 69, pp. 364–367.
20. Brown M.B., Forsythe A.B. The small sample behavior of some statistics which test the equality of several means // Technometrics, 1974, vol. 16, pp. 129–132.
21. Butar F.B. Permutation tests for more than two samples // Journal of Mathematical Sciences and Mathematics Education, 2007, vol. 2, no. 2, pp. 20–29.
22. Campbell D.T., Stanley J.C. Experimental and quasi–experimental designs for research. – Boston, MA: Houghton Mifflin, 1963.
23. Campbell R.A. A comparison of the Quade and Friedman tests to the unbalanced two–way analysis of variance with biomedical data // Computers in Biology and Medicine, 1988, vol. 18, no. 6, pp. 441–447.
24. Cochran W.F. The comparison of percentages in matched samples // Biometrika, 1950, vol. 37, pp. 256–266.
25. Cohen A.D. Robust multivariate analysis for the comparison of several samples // Multivariate Behavioral Research, 1983, vol. 18, no. 3, pp. 259–274.
26. Cohen J. Statistical power analysis for the behavioral sciences. – Hillsdale, NJ: Lawrence Erlbaum, 1988.
27. Conover W.J. Practical nonparametric statistics. – New York, NY: John Wiley & Sons, 1999.
28. Conover W.J., Iman R.L. Analysis of covariance using the rank transformation // Biometrics, September 1982, vol. 38, no. 3, pp. 715–724.
29. Conover W.J., Iman R.L. Rank transformations as a bridge between parametric and nonparametric statistics // The American Statistician, 1981, vol. 35, pp. 124–129.
30. Cortina J., Nouri H. Effect size for ANOVA designs. – Newbury Park, CA: Sage Publications, 2000.
31. Cowles M., Davis C. On the origins of the .05 level of statistical significance // Psychological Bulletin, 1982, vol. 89, pp. 553–558.
32. Cuzick J. A Wilcoxon–type test for trend // Statistics in Medicine, January/March 1985, vol. 4, no. 1, pp. 87–90.
33. Daniel W.W. Applied nonparametric statistics. – PWS–Kent publishing company, 1990.
34. Davis C.S. Statistical methods for the analysis of repeated measurements. – New York, NY: Springer–Verlag, 2002.
35. De Cani J.S. Balancing type I risk and loss of power in ordered Bonferroni procedures // Journal of Educational Psychology, 1984, vol. 76, pp. 1035–1037.
36. Dunn O.J. Multiple comparisons among means // Journal of the American Statistical Association, 1961, vol. 56, pp. 52–54.
37. Dunn O.J. Multiple comparisons using rank sums // Technometrics, 1964, vol. 6, pp. 241–252.
38. Dunnett C.W. A multiple comparisons procedure for comparing several treatments with a control // Journal of American Statistical Association, 1955, vol. 50, pp. 1096–1121.
39. Dunnett C.W. New tables for multiple comparisons with a control // Biometrics, 1964, vol. 20, pp. 482–491.
40. Edwards A. Experimental design in psychological research. – New York, NY: Holt, Rinehart & Winston, 1972.
41. Edwards L.K. Applied analysis of variance in behavioral science / Ed. by L.K. Edwards. – New York, NY: Marcel Dekker, 1993.
42. Erdman L.W. Studies to determine if antibiosis occurs among rhizobia // Journal of the

American Society of Agronomy, 1946, vol. 38, pp. 251–258.
43. Fisher R.A. Statistical methods for research workers. – Edinburgh: Oliver & Boyd, 1925.
44. Fisher R.A. The design of experiments. – Edinburgh: Oliver & Boyd, 1942.
45. Fleiss J.L. The design and analysis of clinical experiments. – New York, NY: John Wiley & Sons, 1986.
46. Gary G. A review of some statistical methods for covariance analysis of categorical data / G. Gary, G.G. Koch, I.A. Amara et al. // Biometrics, September 1982, vol. 38, no. 3, Special No.: Analysis of Covariance, pp. 563–595.
47. Girden E.R. ANOVA: Repeated measures. Quantitative applications in the social sciences, series No. 84. – Thousand Oaks, CA: Sage Publications, 1992.
48. Graybill F.A. An Introduction to linear statistical models. Volume I. – New York, NY: McGraw–Hill, 1961.
49. Graybill F.A. Theory and applications of the linear model. – North Scituate, MA: Duxbury Press, 1976.
50. Green P.E. Analyzing multivariate data. – Hinsdale, IL: Dryden Press, 1978.
51. Grimm L. Statistical applications for the behavioral sciences. – New York, NY: John Wiley & Sons, 1993.
52. Guidance for data quality assessment. Practical methods for data analysis. EPA QA/G–9. – Washington, DC: United States Environmental Protection Agency, 2000.
53. Guilford J.P. Fundamental statistics in psychology and education. – New York, NY: McGraw–Hill, 1965.
54. Harris R.J. A primer of multivariate statistics. – Mahwah, NJ: Lawrence Erlbaum Associates, 2001.
55. Henderson C.R. Estimation of variance and covariance components // Biometrics, 1953, vol. 9, pp. 226–252.
56. Hochberg Y., Tamhane A.C. Multiple comparison procedures. – New York, NY: John Wiley & Sons, 1987.
57. Hocking R.R. Analysis of linear models. – Monterey, CA: Brooks–Cole Publishing, 1984.
58. Holland B.S., Copenhaver M.D. Improved Bonferroni–type multiple testing procedures // Psychological Bulletin, 1988, vol. 104, pp. 145–149.
59. Holm S. A simple sequentially rejective multiple test procedure // Scandanavian Journal of Statistics, 1979, vol. 6, pp. 65–70.
60. Hotelling H. A generalized T test and measure of multivariate dispersion // Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability, July 31–August 12, 1950 / Ed. by J. Neyman. – Berkeley, CA: University of California Press, 1951, pp. 23–41.
61. House D.E. A nonparametric version of Williams' test for a randomized block design // Biometrics, March 1986, vol. 42, no. 1, pp. 187–190.
62. Hsu J.C. Constrained simultaneous confidence intervals for multiple comparisons with the best // The Annals of Statistics, 1984, vol. 12, no. 3, pp. 1136–1144.
63. Hsu J.C. Multiple comparisons: Theory and methods. – London: Chapman & Hall, 1996.
64. Huberty C.J. No.s in the use and interpretation of discrimination analysis // Psychological Bulletin, 1984, vol. 95, pp. 156–171.
65. Huberty C.J., Morris J.D. Multivariate analysis versus multiple univariate analyses // Psychological Bulletin, 1989, vol. 105, pp. 302–308.
66. Huberty C.J., Smith J.D. The study of effects in MANOVA // Multivariate Behavioral Research, 1982, vol. 17, pp. 417–482.
67. Hummel T.J., Sligo J.R. Empirical comparison of univariate and multivariate analysis of variance procedures // Psychological Bulletin, 1971, vol. 76, pp. 49–57.
68. Huynh H., Mandeville G.K. Validity conditions in repeated measures designs // Psychological

Bulletin, 1979, vol. 86, pp. 964–973.
69. Iverson G.R. Analysis of variance. Quantitative Applications in the Social Sciences, series No. 1. – Thousand Oaks, CA: Sage Publications, 1987.
70. Jaccard J. Interaction effects in factorial analysis of variance. Quantitative Applications in the Social Sciences, series No. 118. – Thousand Oaks, CA: Sage Publications, 1998.
71. Jackson S. Random factors in ANOVA. Quantitative Applications in the Social Sciences, series no. 98. – Thousand Oaks, CA: Sage Publications, 1994.
72. John P. Statistical design and analysis of experiments. – New York, NY: Macmillan Publishing, 1971.
73. Johnson R.A., Wichern D.W. Applied multivariate statistical analysis. – Upper Saddle River, NJ: Prentice–Hall, 1999.
74. Jonckheere A.R. A distribution–free k–sample test against ordered alternatives // Biometrika, 1954, vol. 41, pp. 133–145.
75. Kaplan R.M., Saccuzzo D.P. Psychological testing: Principles, applications, and applications. – Pacific Grove, CA: Brooks/Cole, 1989.
76. Kennedy W.J.Jr., Gentle J.E. Statistical computing. – New York, NY: Marcel Dekker, 1980.
77. Kerlinger F.N. Foundations of behavioral research. – New York, NY: Holt, Rinehart & Winston, 1986.
78. Keselman H.J., Keselman J.C., Games P.A. Maximum family wise type I error rate: The least significant difference, Newman–Keuls, and other multiple comparison procedures // Psychological Bulletin, 1991, vol. 110, pp. 155–161.
79. Kirk R. Experimental designs: Procedures for the behavioral sciences. – Belmont, CA: Brooks/Cole, 1968.
80. Kladopoulos C.N., Ramsey P.H. A more robust procedure for testing the null hypothesis in MANOVA // InterStat (Statistics on the Internet), September 2005, No. 1.
81. Kleinbaum D.G., Kupper L.L., Muller K.E. Applied regression analysis and other multivariable methods. – Boston, MA: PWS–KENT, 1988.
82. Klotz J., Teng J. One–way layout for counts and the exact enumeration of the Kruskal–Wallis H distribution with ties // Journal of the American Statistical Association, March 1977, vol. 72, no. 357, pp. 165–169.
83. Krishnaiah P.R. Handbook of statistics. Vol. 1. Analysis of variance / Ed. by P.R. Krishnaiah. – New York, NY: Elsevier, 1980.
84. Larsen R.J., Diener E. Affect intensity as an individual differences characteristic: A review // Journal of Research in Personality, 1987, vol. 21, pp. 1–39.
85. Lawson A. Rank analysis of covariance: Alternative approaches // Journal of the Royal Statistical Society. Series D (The Statistician), September 1983, vol. 32, no. 3, pp. 331–337.
86. Lehmann E.L. Testing statistical hypotheses. – New York, NY: John Wiley & Sons, 1986.
87. Levene H. Robust tests for the equality of variance // Contributions to probability and statistics / Ed. by I. Olkin. – Palo Alto, CA: Stanford University Press, 1960, pp. 278–292.
88. Levin I.P. Relating statistics and experimental design: An introduction. Quantitative Applications in the Social Sciences, series no. 125. – Thousand Oaks, CA: Sage Publications, 1999.
89. Lix L.M. Multivariate tests of means in independent groups designs // Evaluation & the Health Professions, 2004, vol. 27, no. 1, pp. 45–69.
90. Mahfoud Z.R., Randles R.H. Practical tests for randomized complete block designs // Journal of Multivariate Analysis archive, September 2005, vol. 96, no. 1, pp. 73–92.
91. Mardia K.V. Assessment of multinormality and the robustness of Hotelling's T–squared test // Journal of the Royal Statistical Society: Series C (Applied Statistics), 1975, vol. 24, no. 2, pp. 163–171.

92. Maxwell S.E., Delaney H.D. Designing experiments and analyzing data: A model comparison perspective. – Mahwah, NJ: Lawrence Erlbaum Associates, 2004.
93. McDonald B.J., Thompson W.A.Jr. Rank sum multiple comparisons in one– and two–way classifications // Biometrika, 1967, vol. 54, pp. 487–497.
94. McNemar Q. Psychological statistics. – New York, NY: John Wiley & Sons, 1966.
95. Mead R., Curnov R.N., Hasted A.M. Statistical methods in Agriculture and Experimental Biology. – London, UK: Chapman & Hall, 1993.
96. Mendenhall W. Introduction to linear models and the design and analysis of experiments. – Belmont, CA: Duxbury Press, 1968.
97. Mendoza J.L., Graziano W.G. The statistical analysis of dyadic social behavior: A multivariate approach // Psychological Bulletin, 1982, vol. 92, pp. 532–540.
98. Miller R.G. Simultaneous statistical inference. – New York, NY: Springer Verlag, 1981.
99. Milliken G.A., Johnson D.E. Analysis of messy data. Volume I: Designed experiments. – Belmont, CA: Lifetime Learning Publications, 1984.
100. Milliken G.A., Johnson D.E. Analysis of messy data. Volume III: Analysis of covariance. – Boca Raton, FL: Chapman & Hall / CRC, 2002.
101. Miwa T., Hayter A.J., Liu W. Calculations of level probabilities for normal random variables with unequal variances with applications to Bartholomew's test in unbalanced one–way models // Computational Statistics and Data Analysis, 2000, vol. 34, no. 1, pp. 17–32.
102. Morrison D.F. Multivariate statistical methods. – New York, NY: McGraw–Hill, 1976.
103. Motulsky H.J. Intuitive biostatistics. – New York, NY: Oxford University Press, 1995.
104. Neter J., Wasserman W., Kutner M.H. Applied linear statistical models: Regression, analysis of variance, and experimental designs. – Homewood, IL: Richard D. Irwin, 1990.
105. NIST/SEMATECH e–Handbook of statistical methods (NIST Handbook 151, ver. 1/27/2005). – Gaithersburg, MD: National Institute of Standards and Technology, 2005.
106. O'Brien R.G. A general ANOVA method for robust tests of additive models for variances // Journal of the American Statistical Association, 1979, vol. 74, pp. 877–880.
107. O'Brien R.G. A simple test for variance effects in experimental designs // Psychological Bulletin, 1981, vol. 89, no. 3, pp. 570–574.
108. Oja H., Randles R.H. Multivariate nonparametric tests // Statistical Science, 2004, vol. 19, pp. 598–605.
109. Oktaba W. Note on the ANOVA of a completely confounded factorial experiment // Applicationes Mathematicae, 2005, vol. 32, no. 2, pp. 119–132.
110. Olson C.L. Comparative robustness of six tests in multivariate analysis of variance // Journal of the American Statistical Association, 1974, vol. 69, pp. 894–908.
111. Olson C.L. On choosing a test statistic in multivariate analysis of variance // Psychological Bulletin, 1976, vol. 83, pp. 579–586.
112. Ott L. Introduction to statistical methods and data analysis. – Belmont, CA: Duxbury Press, 1977.
113. Page E.B. Ordered hypotheses for multiple treatments: a significance test for linear ranks // Journal of the American Statistical Association, 1963, vol. 58, pp. 216–230.
114. Panchapakesan S., Balakrishnan N., Gupta S.S. Advances in statistical decision theory and applications. – Boston, MA: Birkhauser, 1977.
115. Pedhazur E.J. Multiple regression in behavioral research. – Fort Worth, TX: Holt, Rinehart & Winston, 1982.
116. Pollard J.H. A handbook of numerical and statistical techniques. – New York, NY: Cambridge University Press, 1977.
117. Pontes A.C.F. Analise de variancia multivariada com a utilizacao de testes nao–

parametricos e componentes principais baseados em matrizes de postos. – Sao Paulo: Paracicaba, 2005.
118. Pontes A.C.F. Obtencao dos niveis de significancia para os testes de Kruskal–Wallis, Friedman e comparacoes multiplas nao–parametricas. – Sao Paulo: Paracicaba, 2000.
119. Puri M.L., Sen P.K. Nonparametric methods in multivariate analysis. – New York, NY: John Wiley & Sons, 1971.
120. Quade D. Analyzing randomized blocks by weighted rankings // Report SW 18/72 of the Mathematical Center, Amsterdam, 1972.
121. Quade D. Nonparametric analysis of covariance by matching // Biometrics, 1982, vol. 38, pp. 597–611.
122. Quade D. Rank analysis of covariance // Journal of the American Statistical Association, 1967, vol. 62, pp. 1187–1200.
123. Quade D. Using weighted rankings in the analysis of complete blocks with additive block effects // Journal of the American Statistical Association, September 1979, vol. 74, no. 367, pp. 680–683.
124. Raghavachari M. Multi–sample tests for scale // Journal Annals of the Institute of Statistical Mathematics, December 1970, vol. 22, no. 1, pp. 459–464.
125. Ramaswamy R., Koch G.G., Amara I.A. Application of rank analysis of covariance methods to analysis of multiple anatomical regions with treatment for seborrhea dermatitis // Journal of Biopharmaceutical Statistics, 1997, vol. 7, no. 3, pp. 403–416.
126. Rao C.D. Linear statistical inference and its applications. – New York, NY: John Wiley & Sons, 2002.
127. Rao K.A., Badade M. A robust test for equality of variances of k normal populations // Bulletin of the International Statistical Institute, 52nd Session, Proceedings, Tome LVIII, Finland 1999. Contributed Paper Meeting 7: Statistical tests.
128. Rayner J.C.W., Best D.J. Nonparametric tests for data in randomized blocks with ordered alternatives // Journal of Applied Mathematics & Decision Sciences, 1999, vol. 3, no. 2, pp. 143–153.
129. Remington R.D., Schork M.A. Statistics with applications to the biological and health sciences. – Englewood Cliffs, NJ: Prentice–Hall, 1970.
130. Rousson V. On distribution–free tests for the multivariate two–sample location–scale model // Journal of Multivariate Analysis, January 2002, vol. 80, no. 1, pp. 43–57.
131. Rucci A., Tweney R. Analysis of variance and the «Second Discipline» of scientific psychology: A historical account // Psychological Bulletin, 1980, vol. 87, pp. 166–184.
132. Rutherford A. Introducing ANOVA and ANCOVA: A GLM approach. – London: Sage Publications, 2001.
133. Ryan T.A. Multiple comparisons in psychological research // Psychological Bulletin, 1959, vol. 56, pp. 26–47.
134. Sahai H., Ageel M.I. The analysis of variance: fixed, random and mixed models. – Boston, MA: Birkhauser, 2000.
135. Schaich H.E., Hamerle A. Verteilungsfreie statistische Prufverfahren. – Berlin: Springer, 1984.
136. Scheffe H. The analysis of variance. – New York, NY: John Wiley & Sons, 1959.
137. Scheirer C.J., Ray W.S., Hare N. The analysis of ranked data derived from completely randomized factorial designs // Biometrics, June 1976, vol. 32, no. 2, pp. 429–434.
138. Seaman M.A., Levin J.R., Serlin, R. New developments in pairwise multiple comparisons: Some powerful and practicable procedures // Psychological Bulletin, 1991, vol. 110, pp. 577–586.
139. Searle S.R. Linear models. – New York, NY: John Wiley & Sons, 1971.

140. Sen P.K., Puri M.L. On a class of multivariate multisample rank order tests, II: Test for homogeneity of dispersion matrices // Sankhya, Series A, 1968, vol. 30, part I, pp. 1–22.
141. Shaughnessy J.J., Zechmeister E.B. Research methods in psychology. – New York, NY: McGraw–Hill, 1990.
142. Shoukri M.M., Pause C.A. Statistical methods for health sciences. – New York, NY: CRC Press, 1998.
143. Siegel S., Casttelan Jr. N.J. Non–parametric statistics. – New York, NY: McGraw–Hill, 1988.
144. Siotani M. On the distributions of the Hotelling's T^2–statistics // Annals of the Institute of Statistical Mathematics, 1956, vol. 8, no. 1, pp. 1–14.
145. Siskind V. Approximate probability integrals and critical values for Bartholomew's test for ordered means // Biometrika, December 1976, vol. 63, no. 3, pp. 647–654.
146. Snedecor G.W., Cochran W.G. Statistical methods. – Ames, IA: Iowa State University Press, 1980.
147. Sokal R.R., Rohlf F.J. Biometry: the principles and practice of statistics in biological research. – New York, NY: WH Freeman and Company, 1995.
148. Soliani L. Manuale di statistica per la ricerca e la professione. Statistica unuvariata e bivariata, parametrica e non-parametrica per le discipline ambientali e biologiche / L. Soliani, F. Sartore, E. Siri. – Parma: Universita di Parma, 2005.
149. Solorzano E. Nonparametric multiple comparisons with more than one control using normal scores and Savage statistics // InterStat (Statistics on the Internet), November 2004.
150. Steel R.G.D., Torrie J.H. Principles and procedures of statistics. – New York, NY: McGraw–Hill, 1980.
151. Stevens J. Applied multivariate statistics for the social sciences. – Mahwah, NJ: Lawrence Erlbaum Associates, 2002.
152. Stevens J. Intermediate statistics: A modern approach. – Hillsdale, NJ: Erlbaum, 1990.
153. Stevens J.P. Power of the multivariate analysis of variance tests // Psychological Bulletin, 1980, vol. 88, pp. 728–737.
154. Sturm–Beiss R. A visualization tool for one– and two–way analysis of variance // Journal of Statistics Education, 2005, vol. 13, no. 1.
155. Subbaiah P., Mudholkar G.S. On a multivariate analog of Studentized range test // Journal of the American Statistical Association, 1981, vol. 76, no. 375, pp. 725–728.
156. Tamura R. Multivariate nonparametric several–sample tests // The Annals of Mathematical Statistics, June 1966, vol. 37, no. 3, pp. 611–618.
157. Tatsuoka M.M. Multivariate analysis: Techniques for educational and psychological research. – New York, NY: John Wiley & Sons, 1971.
158. Theodorsson–Norheim E. Friedman and Quade tests: BASIC computer program to perform nonparametric two–way analysis of variance and multiple comparisons on ranks of several related samples // Computers in Biology and Medicine, 1987, vol. 17, no. 2, pp. 85–99.
159. Turner J.R. Introduction to analysis of variance: Design, analysis, and interpretation. Quantitative Applications in the Social Sciences, series No. 129. – Thousand Oaks, CA: Sage Publications, 2001.
160. Van Belle G. Biostatistics: A methodology for the health sciences // G. van Belle, L.D. Fisher, P.J. Heagerty et al. – New York, NY: John Wiley & Sons, 2003.
161. Wang L., Zhou X.–H. A fully nonparametric diagnostic test for homogeneity of variances // The Canadian Journal of Statistics, 2005, vol. 33, no. 7.
162. Wildt A.R., Ahtola O. Analysis of covariance. – Newbury Park, CA: Sage Publications, 1978.

163. Wilkinson L. Response variable hypothesis in the multivariate analysis of variance // Psychological Bulletin, 1975, vol. 82, pp. 408–412.
164. Williams D.A. A test for differences between treatment means when several dose levels are compared with a zero dose control // Biometrics, 1971, vol. 27, pp. 103–177.
165. Zwick R. Rank and normal scores alternatives to Hotelling's T^2 // Multivariate Behavioral Research, 1986, vol. 21, no. 2, pp. 169–186.
166. Андерсон Т. Введение в многомерный статистический анализ. – М.: Государственное издательство физико–математической литературы, 1963.
167. Аренс Х., Лейтер Ю. Многомерный дисперсионный анализ. – М.: Финансы и статистика, 1985.
168. Афифи А., Эйзен С. Статистический анализ: Подход с использованием ЭВМ. – М.: Мир, 1982.
169. Бейли Н. Статистические методы в биологии. – М.: Мир, 1963.
170. Бикел П., Доксам К. Математическая статистика. Выпуск 2. – М.: Финансы и статистика, 1983.
171. Браунли К.А. Статистическая теория и методология в науке и технике. – М.: Наука, 1977.
172. Ветров А.А., Ломовацкий Г.И. Дисперсионный анализ в экономике. – М.: Статистика, 1975 г.
173. Гланц С. Медико–биологическая статистика. – М.: Практика, 1998.
174. ГОСТ Р ИСО 5725–2–2002. Точность (правильность и прецизионность) методов и результатов измерений. Часть 2. Основной метод определения повторяемости и воспроизводимости стандартного метода измерений. – М.: Издательство стандартов, 2002.
175. Гудман С.Н. На пути к доказательной биостатистике. Часть 1: обманчивость величины p // Международный журнал медицинской практики, 2002, № 1, с. 8–17.
176. Гудман С.Н. На пути к доказательной биостатистике. Часть 2: байесовский критерий // Международный журнал медицинской практики, 2002, № 2, с. 5–14.
177. Джонсон Н., Лион Ф., Статистика и планирование эксперимента в технике и науке. Методы обработки данных. – М.: Мир, 1980.
178. Дронов С.В. Многомерный статистический анализ: Учебное пособие. – Барнаул: Издательство Алтайского государственного университета, 2003.
179. Дюк В. Обработка данных на ПК в примерах. – СПб.: Питер, 1997.
180. Кобзарь А.И. Прикладная математическая статистика. Для инженеров и научных работников. – М.: ФИЗМАТЛИТ, 2006.
181. Коган Р.И., Белов Ю.П., Родионов Д.А. Статистические ранговые критерии в геологии. – М.: Недра, 1983.
182. Кулаичев А.П. Методы и средства анализа данных в среде Windows®. STADIA. – М.: Информатика и компьютеры, 1999.
183. Кулаичев А.П. Методы и средства комплексного анализа данных. – М.: ИНФРА–М, 2006.
184. Кульбак С. Теория информации и статистика. – М.: Наука, 1967.
185. Лакин Г.Ф. Биометрия. – М.: Высшая школа, 1990.
186. Леман Э. Проверка статистических гипотез. – М.: Наука, 1979.
187. Лемешко Б.Ю., Миркин Е.П. Критерии Бартлетта и Кокрена в измерительных задачах при вероятностных законах, отличающихся от нормального // Измерительная техника, 2004, № 10, с. 10–16.
188. Лемешко Б.Ю., Пономаренко В.М. Проверка гипотез в моделях дисперсионного анализа со случайными факторами при нарушении предположений о

нормальности // Доклады академии наук высшей школы России, 2005, № 2, с. 26–39.
189. Лемешко Б.Ю., Пономаренко В.М., Трушина Е.А. К проверке статистических гипотез в регрессионном и дисперсионном анализах при нарушении предположений о нормальности ошибок // Материалы 6–й всероссийской НТК «Информационные технологии в науке, проектировании и производстве», Н. Новгород, 2002, с. 1–5.
190. Лисенков А.Н. Математические методы планирования многофакторных медико–биологических экспериментов. – М.: Медицина, 1979.
191. Ллойд Э. Справочник по прикладной статистике. В 2–х т. Т. 1. / Под ред. Э. Ллойда, У. Ледермана, Ю.Н. Тюрина. – М.: Финансы и статистика, 1989.
192. Ллойд Э. Справочник по прикладной статистике. В 2–х т. Т. 2. / Под ред. Э. Ллойда, У. Ледермана. – М.: Финансы и статистика, 1990.
193. Мэйндоналд Дж. Вычислительные алгоритмы в прикладной статистике. – М.: Финансы и статистика, 1988.
194. Мюллер П., Нойман П., Шторм Р. Таблицы по математической статистике. – М.: Финансы и статистика, 1982.
195. Налимов В.В. Применение математической статистики при анализе вещества. – М.: Государственное издательство физико–математической литературы, 1960.
196. Оуэн Д.Б. Сборник статистических таблиц. – М.: ВЦ АН СССР, 1966.
197. Петрович М.Л., Давидович М.И. Статистическое оценивание и проверка гипотез на ЭВМ. – М.: Финансы и статистика, 1989.
198. Поллард Дж. Справочник по вычислительным методам статистики. – М.: Финансы и статистика, 1982.
199. Прохоров Ю.В. Вероятность и математическая статистика. Энциклопедия / Гл. ред. Ю.В. Прохоров. – М.: Научное издательство «Большая Российская энциклопедия», 1999.
200. Родионов Д.А. Справочник по математическим методам в геологии / Д.А. Родионов, Р.И. Коган, В.А. Голубева и др. – М.: Недра, 1987.
201. Родионов Д.А. Статистические методы разграничения геологических объектов по комплексу признаков. – М.: Недра, 1968.
202. Родионов Д.А. Статистические решения в геологии. – М.: Недра, 1981.
203. Рокицкий П.Ф. Биологическая статистика. – Мн.: Вышэйшая школа, 1973.
204. Сидоренко Е.В. Методы математической обработки в психологии. – СПб.: ООО «Речь», 2001.
205. Снедекор Дж.У. Статистические методы в применении к исследованиям в сельском хозяйстве и биологии. – М.: Сельхозгиз, 1961.
206. Тюрин Ю.Н., Макаров А.А. Статистический анализ данных на компьютере. – М.: ИНФРА–М, 1999.
207. Уилкс С. Математическая статистика. – М.: Наука, 1967.
208. Хальд А. Математическая статистика с техническими приложениями. – М.: Издательство иностранной литературы, 1956.
209. Холлендер М., Вулф Д. Непараметрические методы статистики. – М.: Финансы и статистика, 1983.
210. Хьютсон А. Дисперсионный анализ. – М.: Статистика, 1971.
211. Шеффе Г. Дисперсионный анализ. – М.: Наука, 1980.

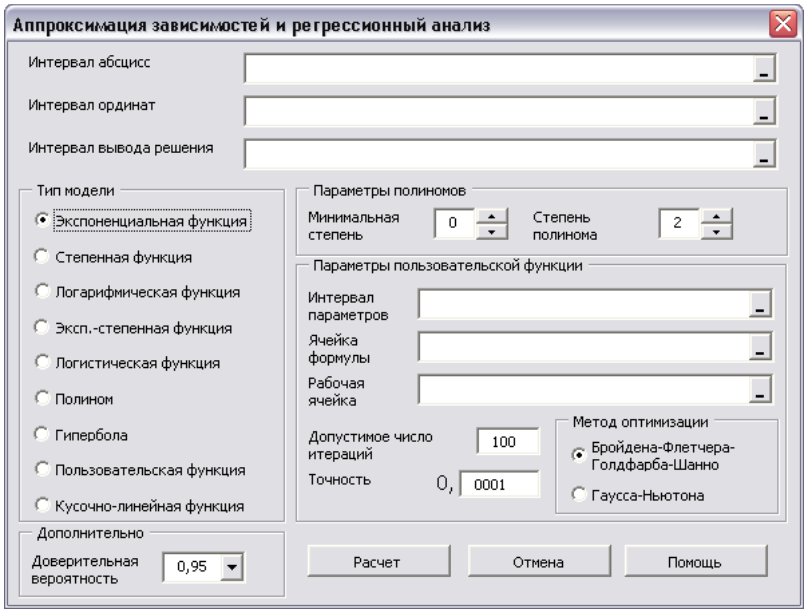
Глава 9. Регрессионный анализ

9.1. Введение

Программное обеспечение регрессионного анализа и аппроксимации зависимостей предназначено для вычисления параметров аппроксимирующих функций различными методами и их статистических оценок.
Номенклатура методов насчитывает несколько различных аппроксимирующих функций. Если пользователь не найдет необходимую функцию в предлагаемом перечне, можно воспользоваться универсальным методом – пользовательской функцией. Выбор параметров для данного метода не очевиден, поэтому порядок работы с данным методом представлен в виде подробного примера.

9.2. Работа с программным обеспечением

Выберите из меню программы пункт AtteStat | Регрессионный анализ. На экране появится диалоговое окно, подобное окну, изображенному на рисунке.



Затем:

- Выберите или введите интервал абсцисс (альтернативные наименования: аргумента, предиктора, контролируемой переменной). Данное наименование вызвано тем, что ось на плоском (двумерном) графике, соответствующая аргументу, называется абсциссой; в противоположность ей ось, соответствующая функции, называется ординатой.
- Выберите или введите интервал ординат (альтернативные наименования: функции выхода эксперимента, зависимой переменной).
- Выберите или введите выходной интервал (интервал вывода). Начиная с первой ячейки выходного интервала (следовательно, можно указать только одну ячейку, т.к. остальные ячейки интервала игнорируются), будут выведены результаты вычислений.
- Выберите или оставьте по умолчанию метод аппроксимации.
- Выберите или оставьте по умолчанию параметры решения, если выбор параметров предусмотрен методом. Для полинома можно указать минимальную степень и степень

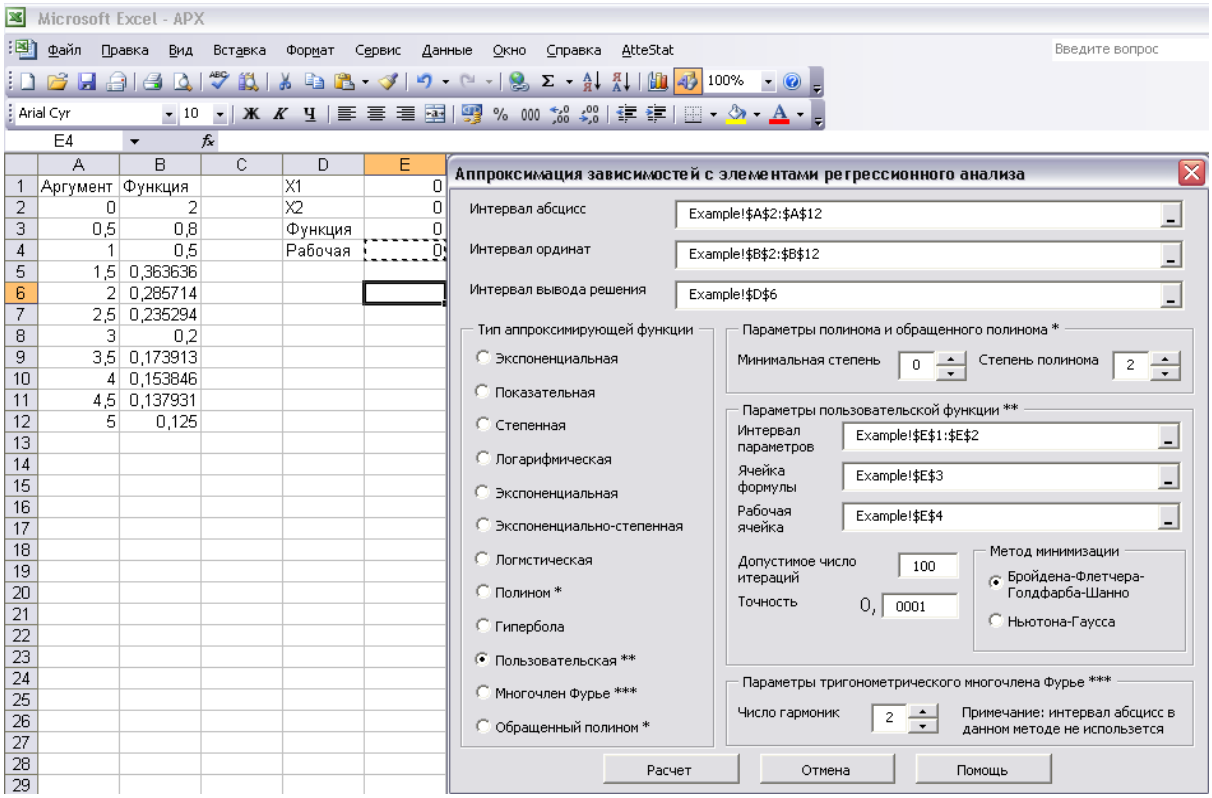
полинома. Для пользовательской функции следует указать интервал параметров модели, предварительно введя в него произвольные (но по возможности максимально близкие к точным) начальные значения. Также потребуется выбрать метод решения и указать ячейку с корректно введенной пользовательской функцией и произвольную рабочую ячейку, требуемую данным методом.

- Выберите или оставьте по умолчанию доверительную вероятность, необходимую для расчета статистических оценок.
- Нажмите кнопку Расчет.

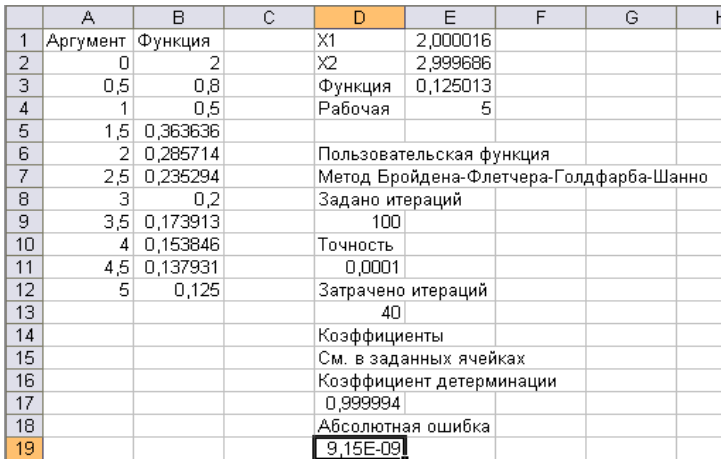
После выполнения вычислений будет, начиная с первой ячейки выходного интервала, выведено название метода и результаты расчета, включая график с изображением модели, функции и доверительных интервалов.
Программное обеспечение берет на себя верификацию исходных данных, выдавая подробную диагностику. При неверных действиях пользователя выдаются сообщения об ошибках. При решении задачи аппроксимации зависимости нужно быть готовым к тому, что для конкретных исходных данных не все методы аппроксимации применимы. Если аппроксимация представленной пользователем зависимости не может быть выполнена желаемым методом, выдается сообщение об ошибке в вычислениях.

9.2.1. Пример применения

Расчеты методами, представленными в настоящем программном обеспечении, проиллюстрируем следующим примером, заимствованным со с. 337–338 монографии Носача.
Пусть решается задача аппроксимации зависимости, представляющей собой 11 пар значений аргумент–функция, дробно–рациональной функцией вида
$y(t) = \theta_1 / (1 + \theta_2 t)$,
где θ_1, θ_2 – подлежащие определению неизвестные коэффициенты,
t – аргумент.
Аппроксимация производится методом «Пользовательская функция».
Для расчета введем в ячейки **A2:A12** и **B2:B12** значения аргумента и функции, соответственно. Далее, в ячейки **E1** и **E2** введем начальные приближения искомых коэффициентов (пусть нулевые, хотя можно и как в источнике). В ячейку **E3** введем аппроксимирующую функцию в виде следующей строки =E1/(1+E2*E4), соответствующей показанной выше математической формуле. Ячейку **E4** будем использовать в качестве рабочей ячейки для хранения текущего значения аргумента, можно поместить в нее любое значение или не помещать ничего. Ячейкой D6 укажем начальную позицию, с которой будет выводиться решение. Перед началом расчета экран компьютера будет выглядеть примерно так, как показано на рисунке.



Нажимаем кнопку Расчет. После определенного непродолжительного времени, которое зависит от производительности компьютера, начиная с позиции, указанной ячейкой **D6**, будут выведены результаты расчета, как показано на иллюстрации.



После окончания расчета в ячейки **E1** и **E2** будут помещены вычисленные значения коэффициентов аппроксимирующей функции. Дополнительно выводится информация о затраченных итерациях и точности подгонки функции.
Нами рассмотрен наиболее сложный пример использования настоящего программного обеспечения. Другие методы использовать значительно проще.

9.2.2. Сообщения об ошибках

При ошибках ввода данных могут выдаваться диагностические сообщения следующих типов:

Ошибка	Комментарий
Не определен интервал абсцисс.	Вы не выбрали или неверно ввели интервал абсцисс. Лучшим способом избежать ошибки является не ввод, а выделение интервала ячеек стандартным образом.
Не определен интервал ординат.	Вы не выбрали или неверно ввели интервал ординат. Лучшим способом избежать ошибки является не ввод, а выделение интервала ячеек стандартным образом.
Не определен интервал вывода.	Вы не выбрали или неверно ввели интервал вывода. Лучшим способом избежать ошибки является не ввод, а выделение интервала ячеек стандартным образом.
Пустая ячейка.	В данных, указанных в качестве интервала абсцисс или ординат, встретилась пустая ячейка. Необходимо заполнить или устранить все пустые ячейки.
Нечисловой тип данных.	В данных, указанных в качестве интервала абсцисс или ординат, встретилась нечисловая ячейка. Необходимо устранить данную ошибку. Лучшим способом является выделение ячейки или группы ячеек и объявление их числовыми с помощью стандартных средств.
Разные численности.	Указанные интервалы абсцисс и ординат имеют разные численности выборок. Методы аппроксимации требуют, чтобы каждой абсциссе соответствовала ордината. Устраните ошибку и повторите расчет.
Степень не соответствует численности.	Каждый метод аппроксимации требует, чтобы число пар абсцисс и ординат было большим, чем количество вычисляемых параметров аппроксимирующей кривой. Проконсультируйтесь с описанием соответствующего метода расчета.
Не определен интервал параметров.	Вы не выбрали или неверно ввели интервал параметров для пользовательской функции. Лучшим способом избежать ошибки является не ввод, а выделение интервала ячеек стандартным образом. Кроме того, интервал параметров до вычислений содержит начальные значения параметров, после успешного вычисления – полученные значения параметров.
Не определена ячейка формулы.	Вы не выбрали или неверно ввели ячейку формулы для пользовательской функции. Лучшим способом избежать ошибки является не ввод, а выделение ячейки стандартным образом.
Не определена рабочая ячейка.	Вы не выбрали или неверно ввели рабочую ячейку для пользовательской функции. Данная произвольная ячейка необходима для вычисления пользовательской функции с помощью машины вычислений. Лучшим способом избежать ошибки является не ввод, а выделение интервала ячеек стандартным образом.
Неверная формула.	Вы неверно ввели формулу для пользовательской функции. Формула вводится в соответствии со стандартными правилами.

Ошибка	Комментарий
Ошибка вычисления модели.	Получение данной ошибки сигнализирует пользователю, что его данные не могут быть аппроксимированы выбранным методом аппроксимации зависимости. Попробуйте воспользоваться другим методом.

9.3. Теоретическое обоснование

Аппроксимацией называется замена одних математических объектов другими, в том или ином смысле близкими исходным. В более узком смысле аппроксимация – вычисление (подбор) неизвестных параметров алгебраических уравнений, в том числе приближение одних функций другими, причем аналитическое выражение для аппроксимируемой функции может быть известно или неизвестно.

9.3.1. Оценка качества аппроксимации

В программе качество аппроксимации оценивается коэффициентом детерминации, вычисляемым по формуле

$$R^2 = 1 - \frac{\sigma_E^2}{\sigma_Y^2},$$

где σ_E^2 – дисперсия остатков,

σ_Y^2 – дисперсия функции выхода эксперимента (далее – функции).

Дисперсия остатков вычисляется как

$$\sigma_E^2 = \frac{1}{N}\sum_{i=1}^{N}(e_i - \bar{e})^2,$$

где N – число экспериментальных значений (пар аргумент–функция),

e_i, $i = 1,2,...,N$ – остатки,

$\bar{e}$ – среднее значение остатков.

Остатки рассчитываются по формуле

$$e_i = \hat{y}_i - y_i, i = 1,2,..., N,$$

где $\hat{y}_i, i = 1,2,..., N,$ – рассчитанные значения модели (модельной оценки),

y_i, $i = 1,2,...,N$ – заданные значения функции.

Среднее значение остатков вычисляется по формуле

$$\bar{e} = \frac{1}{N}\sum_{i=1}^{N} e_i.$$

Дисперсия функции вычисляется как

$$\sigma_Y^2 = \frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{y})^2,$$

где $\bar{y}$ – среднее значение функции.

Среднее значение функции рассчитывается по формуле

$$\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i.$$

В источниках при записи формулы коэффициента детерминации часто опускают множители $1 / N$ в выражениях для дисперсий, поэтому в числителе и знаменателе второго слагаемого

помещают только соответствующие квадраты.
Программой рассчитывается также исправленный коэффициент детерминации, вычисление которого производится по формуле

$$R^2_{(adj)} = 1 - \frac{\sigma^2_{E(adj)}}{\sigma^2_{Y(adj)}},$$

где $\sigma^2_{E(adj)}$ – исправленная дисперсия остатков,
$\sigma^2_{Y(adj)}$ – исправленная дисперсия функции.
Исправленная дисперсия остатков вычисляется как

$$\sigma^2_{E(adj)} = \frac{1}{(N-k)} \sum_{i=1}^{N} (e_i - \bar{e})^2,$$

где k – количество оцениваемых параметров модели.
Исправленная дисперсия функции вычисляется как

$$\sigma^2_{Y(adj)} = \frac{1}{(N-1)} \sum_{i=1}^{N} (y_i - \bar{y})^2.$$

Чем ближе вычисленное значение коэффициента детерминации или исправленного коэффициента детерминации к 1, тем лучше модель аппроксимирует представленные экспериментальные данные. И наоборот, чем меньше 1 вычисленное значение, тем хуже аппроксимирующая функция соответствует представленным данным.
Для проверки значимости коэффициента детерминации рассчитывается статистика

$$F = \frac{R^2 / k}{(1 - R^2)/(N - k - 1)},$$

где k – число оцениваемых параметров модели.
Статистика подчиняется F–распределению с параметрами k и $N - k - 1$.
Для проверки автокорреляции остатков применяется статистика Дарбина–Уотсона

$$d = \frac{\sum_{i=2}^{n} (e_i - e_{i-1})^2}{\sum_{i=1}^{2} e_i^2}.$$

В грубом приближении автокорреляция отсутствует при $1{,}5 \le d \le 2{,}5$.

9.3.2. Регрессионный анализ

Основные задачи регрессионного анализа:

1. Выбор наилучшей регрессионной модели (функции) для заданного набора экспериментальных данных (независимой переменной – выхода эксперимента).
2. Вычисление оптимальных параметров модели. Наиболее популярным для рассматриваемого типа задач является метод наименьших квадратов.
3. Оценка значимости и вычисление доверительных интервалов параметров модели.
4. Оценка значимости и вычисление доверительных интервалов выхода модели.

Решение первой задачи иногда находится вне вычислительных методов. К формулировке данной задачи могут привести математическое моделирование или просто интуиция исследователя. Также можно предположить вид аппроксимирующей функции, исходя из вида графика, построенного по результатам эксперимента. Для решения второй задачи используются методы аппроксимации (без статистических оценок). Решение третьей и четвертой задачи производится с помощью алгоритмов прикладной статистики. При этом

оптимальные оценки параметров модели уже получены при решении третьей задачи.
Программа дополнительно выводит стандартные отклонения вычисленных оценок параметров модели

$$SE(\hat{\theta}) = \sqrt{diag\left[\left(P^T(X,\hat{\theta})P(X,\hat{\theta})\right)^{-1} MSE\right]},$$

где $P(.,.)$ – матрица частных производных модели по параметрам,
$\hat{\theta}$ – вектор оценок параметров,
X – заданный вектор аргументов,
MSE – средняя квадратичная ошибка (дисперсия ошибки регрессии).
Матрица частных производных функции модели по параметрам (опуская номер итерации) имеет вид

$$P(X,\theta) = \begin{bmatrix} \frac{\partial f(x_1,\theta)}{\partial \theta_1} & \frac{\partial f(x_1,\theta)}{\partial \theta_2} & \cdots & \frac{\partial f(x_1,\theta)}{\partial \theta_k} \\ \frac{\partial f(x_2,\theta)}{\partial \theta_1} & \frac{\partial f(x_2,\theta)}{\partial \theta_2} & \cdots & \frac{\partial f(x_2,\theta)}{\partial \theta_k} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{\partial f(x_N,\theta)}{\partial \theta_1} & \frac{\partial f(x_N,\theta)}{\partial \theta_2} & \cdots & \frac{\partial f(x_N,\theta)}{\partial \theta_k} \end{bmatrix},$$

где x_i, $i = 1,2,...,N$ – элементы вектора аргумента,
$f(.,\theta)$– выход модели, получающийся подстановкой в функцию модели заданного аргумента, при фиксированном значении вектора параметров.
Практически производные вычисляются либо методом конечных разностей (если вид модели заранее неизвестен), либо задаются в явном виде (если вид модели задан).
Дисперсия ошибки регрессии вычисляется по формуле

$$MSE = \frac{1}{N-k}\sum_{i=1}^{N}(y_i - f(x_i,\hat{\theta}))^2.$$

Также выводятся доверительные интервалы оценок параметров, которые вычисляются по формуле, опуская индексы,

$$I_\theta = \left[\hat{\theta} - \Psi((1+\beta)/2)\cdot SE(\hat{\theta}); \hat{\theta} + \Psi((1+\beta)/2)\cdot SE(\hat{\theta})\right],$$

где $\Psi(.)$ – обратная функция стандартного нормального распределения,
β – доверительный уровень, выраженный в долях.
Дополнительно выводятся значения t-статистики (при проверке гипотезы о равенстве нулю) и P–значения оценок параметров.
Доверительные интервалы оценок модели $\hat{y}_i = f(x_i,\hat{\theta}), i = 1,2,...,N,$ вычисляются как, опуская индексы,

$$I_{\hat{Y}} = [\hat{y} - \Psi((1+\beta)/2)\cdot SD(\hat{y}); \hat{y} + \Psi((1+\beta)/2)\cdot SD(\hat{y})],$$

где $SD(.)$ – стандартное отклонение ошибки регрессии – корень квадратный из дисперсии ошибки регрессии.
Дополнительно выводятся значения t-статистики (при проверке гипотезы о равенстве нулю) и P–значения стандартизованных остатков оценок модели. P–значения, не превышающие 0,05, отмечаются красным цветом, чтобы сигнализировать пользователю о возможном выбросе.

См. монографии Кулаичева (т. 1), Айвазяна с соавт., Полларда, Доугерти, Ферстера с соавт., Сергиенко с соавт., Райана (Ryan), Бородича.

9.3.3. Метод наименьших квадратов

Методы аппроксимации обычно основаны на следующих идеях:

- метод наименьших квадратов,
- метод наименьших модулей,
- метод максимального правдоподобия.

Для рассматриваемого типа задач методы приводят к аналогичным результатам. Рассмотрим теоретические основы построения модели с использованием метода наименьших квадратов (method of least squares).

Пусть записан функционал

$$F(\theta) = \sum_{i=1}^{N} \left(y_i - f(x_i, \theta) \right)^2,$$

где y_i, $i = 1,2,...,N$ – заданное экспериментальное значение, соответствующее значению абсциссы x_i, $i = 1,2,...,N$,

$f(.,.)$, $i = 1,2,...,N$,– модельное значение, рассчитанное по теоретической формуле, заданной с точностью до параметров θ,

N – число пар экспериментальных значений,

θ – вектор параметров, состоящий из подлежащих определению компонент θ_i, $i = 1,2,...,r$,

r – число параметров, зависящее от вида теоретической формулы.

Минимизация функционала $F(\theta)$ по вектору параметров θ_i, $i = 1,2,...,r$,

$$F(\theta) \underset{\theta}{\rightarrow} \min$$

приводит к системе r алгебраических уравнений:

$$\frac{\partial F(\theta)}{\partial \theta_j} = 0,\ j = 1,2,...,r,$$

где в левой части уравнения находятся частные производные функционала $F(\theta)$ по параметрам θ_i, $i = 1,2,...,r$.

Решив полученную линейную или нелинейную систему алгебраических уравнений, получим искомый оптимальный вектор параметров.

9.3.4. Полиномиальные модели

Модель представлена в виде полинома:

$$z(\theta, x) = \sum_{j=0}^{r} \theta_j x^j,$$

где θ_i, $i = 0,1,...,r$ – коэффициенты полинома,

x – заданная абсцисса,

r – степень полинома.

Аналитически выразив частные производные функционала $F(\theta)$ по параметрам, получаем θ_i, $i = 0,1,...,r$, систему $r + 1$ алгебраических уравнений, линейных относительно параметров:

$$\sum_{j=0}^{r} (\theta_j \sum_{i=1}^{N} x_i^{j+k}) = \sum_{i=1}^{N} y_i x_i^k, k = 0,1,...,r.$$

Представленным программным обеспечением, по желанию пользователя, может решаться более общая задача: наряду со степенью полинома можно указать минимальное значение степени члена полинома u (по умолчанию равно нулю):

$$z(\theta, x) = \sum_{j=u}^{r} \theta_j x^j,$$

Матрица частных производных функции модели по параметрам в общем случае будет

$$P(X,\theta)=\begin{bmatrix} x_1^u\, x_1^{u+1} \ldots x_1^r \\ x_2^u\, x_2^{u+1} \ldots x_2^r \\ \ldots \quad \ldots \quad \ldots \ldots \\ x_N^u\, x_N^{u+1} \ldots x_N^r \end{bmatrix}.$$

9.3.5. Экспоненциально–степенная аппроксимация

Экспоненциальной функцией называется зависимость

$z(x)=\theta_1 e^{\theta_2 x}$,

где θ_i, i = 1,2 – параметры,

x – заданная абсцисса.

Матрица частных производных функции модели по параметрам будет

$$P(X,\theta)=\begin{bmatrix} e^{\theta_2 x_1} & \theta_1 x_1 e^{\theta_2 x_1} \\ e^{\theta_2 x_2} & \theta_1 x_2 e^{\theta_2 x_2} \\ \ldots & \ldots \\ e^{\theta_2 x_N} & \theta_1 x_N e^{\theta_2 x_N} \end{bmatrix}.$$

Степенной функцией называется зависимость

$z(x)=\theta_1 x^{\theta_2}, x>0$,

где θ_i, i = 1,2 – параметры,

x – заданная абсцисса.

Матрица частных производных функции модели по параметрам будет

$$P(X,\theta)=\begin{bmatrix} x_1^{\theta_2} & \theta_1\theta_2 x_1^{\theta_2-1} \\ x_2^{\theta_2} & \theta_1\theta_2 x_2^{\theta_2-1} \\ \ldots & \ldots \\ x_N^{\theta_2} & \theta_1\theta_2 x_N^{\theta_2-1} \end{bmatrix}.$$

Гиперболой называется зависимость

$z(x)=\theta_1 x^{-1}+\theta_2, x\neq 0$,

где θ_i, i = 1,2 – параметры,

x – заданная абсцисса.

Матрица частных производных функции модели по параметрам будет

$$P(X,\theta)=\begin{bmatrix} x_1^{-1} & 1 \\ x_2^{-1} & 1 \\ \ldots & \ldots \\ x_N^{-1} & 1 \end{bmatrix}.$$

Экспоненциально–степенной называется зависимость

$z(x)=e^{\theta_1 x} x^{\theta_2}, x>0$,

где θ_i, i = 1,2 – параметры.

Матрица частных производных функции модели по параметрам будет

$$P(X,\theta)=\begin{bmatrix} e^{\theta_1 x_1} x_1^{\theta_2+1} & \theta_2 e^{\theta_1 x_1} x_1^{\theta_2-1} \\ e^{\theta_1 x_2} x_2^{\theta_2+1} & \theta_2 e^{\theta_1 x_2} x_2^{\theta_2-1} \\ \cdots & \cdots \\ e^{\theta_N x_1} x_N^{\theta_2+1} & \theta_2 e^{\theta_1 x_N} x_N^{\theta_2-1} \end{bmatrix}.$$

При значениях аргумента, выходящих за указанные ограничения, соответствующие функции могут выдавать ошибку типа деления на нуль или выхода значений из допустимой области определения.

9.3.6. Логарифмическая функция

Логарифмической функцией называется зависимость

$z(x)=\theta_1+\theta_2 x+\theta_3 \ln(x), x>0,$

где θ_i, i = 1,2,3 – параметры,

x – заданная абсцисса.

Матрица частных производных функции модели по параметрам будет

$$P(X,\theta)=\begin{bmatrix} 1 & x_1 & \ln(x_1) \\ 1 & x_2 & \ln(x_2) \\ \cdots\cdots & \cdots & \\ 1 & x_N & \ln(x_N) \end{bmatrix}.$$

При значениях аргумента, выходящих за указанные ограничения, представленная функция может выдавать ошибку выхода значений из допустимой области определения.

9.3.7. Логистический анализ

Пусть аргумент x означает время или величину растущего объекта, влияющего на размер y наблюдаемого явления. Тогда скорость роста может быть охарактеризована дифференциальным уравнением

$$\frac{dy}{dx}=f(x,y).$$

В частном случае введенная зависимость может иметь вид

$$\frac{dy}{dx}=f(y)g(x).$$

Аналитически можно получить различные примеры кривых роста, например

$z(x)=\theta_1\left[1+e^{\theta_2+\theta_3 x}\right]^{-1}$,

где θ_i, i = 1,2,3 – параметры, определяющие характер кривой,

x – заданная абсцисса.

Матрица частных производных функции модели по параметрам будет

$$P(X,\theta)=\begin{bmatrix} \left[1+e^{\theta_2+\theta_3 x_1}\right]^{-1} & -\theta_1 e^{\theta_2+\theta_3 x_1}\left[1+e^{\theta_2+\theta_3 x_1}\right]^{-2} & -\theta_1 x_1 e^{\theta_2+\theta_3 x_1}\left[1+e^{\theta_2+\theta_3 x_1}\right]^{-2} \\ \left[1+e^{\theta_2+\theta_3 x_2}\right]^{-1} & -\theta_1 e^{\theta_2+\theta_3 x_2}\left[1+e^{\theta_2+\theta_3 x_2}\right]^{-2} & -\theta_1 x_2 e^{\theta_2+\theta_3 x_2}\left[1+e^{\theta_2+\theta_3 x_2}\right]^{-2} \\ \cdots & \cdots & \cdots \\ \left[1+e^{\theta_2+\theta_3 x_N}\right]^{-1} & -\theta_1 e^{\theta_2+\theta_3 x_N}\left[1+e^{\theta_2+\theta_3 x_N}\right]^{-2} & -\theta_1 x_N e^{\theta_2+\theta_3 x_N}\left[1+e^{\theta_2+\theta_3 x_N}\right]^{-2} \end{bmatrix}.$$

9.3.8. Пользовательская функция

В программе реализована возможность аппроксимации опытной зависимости с помощью функции, заданной пользователем.
Требования к пользовательской функции:

- Допустимость с точки зрения машины вычислений.
- Функция может содержать произвольное число параметров и стандартных элементарных функций.
- Аргумент в области допустимых значений. Некоторые варианты пользовательских функций могут содержать корректные формулы, но при некоторых значениях аргумента могут быть получены ошибки времени выполнения типа деления на нуль, переполнения.

Для аппроксимации пользовательской функцией в программе применяются:

- Метод переменной метрики.
- Метод Гаусса– Ньютона.

Сравнивая данные методы, заметим, что иногда методы переменной метрики требуют меньшее число итераций, но время исполнения каждой итерации существенно выше (за счет вычисления оптимального параметра шага итераций). Достоинством метода переменной метрики является более широкая область сходимости (т. е. начальные значения параметров можно задать более удаленными от истинных их значений), если решение удается получить вообще, причем метод Гаусса– Ньютона для тех же самых данных иногда дает решение. Поэтому для каждого набора данных и каждой модели может оказаться оптимальным свой метод. Может также встретиться случай, когда решение не удастся получить ни одним из представленных методов.

9.3.8.1. Метод Бройдена–Флетчера–Голдфарба–Шанно

Применяется один из вариантов метода переменной метрики, а именно, метод Бройдена–Флетчера–Голдфарба–Шанно (метод BFGS). Согласно схеме метода, очередное приближение искомого вектора θ решения нелинейной системы можно найти как

$$\theta^{(i+1)} = \theta^{(i)} + \rho_i d_i, i = 0,1,2,...,$$

где i, i = 0,1,2,... – номер итерации,
ρ_i – параметр шага итераций,
d_i – направление антиградиента (градиентом называют вектор, показывающий направление наибольшего роста скалярной функции n переменных), вычисляемое как

$$d_i = -A_i \nabla F(\theta^{(i)}),$$

где A_i – симметрическая положительно определенная матрица, аппроксимирующая матрицу, обратную к матрице Гессе системы $\left[\nabla^2 F(\theta^{(i)})\right]^{-1}$,
$F(\theta)$ – квадратичный функционал невязок, построенный на основе выхода экспериментальных и модельных значений заданной пользователем функции.
Входящие в выражение градиента производные вычисляются методом конечных разностей.
Параметр ρ_i определяется из условия минимума

$$\varphi(\rho) = F(\theta^{(i)} + \rho_i d_i).$$

Для решения задачи минимизации

$$\varphi(\rho) \underset{\rho}{\rightarrow} \min$$

использован метод деления отрезка пополам.
Следующее приближение обращенной матрицы Гессе вычисляется по формуле

$$A_{i+1} = A_i + \frac{r_i r_i^T}{r_i^T g_i} - \frac{A_i g_i g_i^T A_i}{g_i^T A_i g_i},$$

где $r_i = \theta^{(i+1)} - \theta^{(i)}$ – разность приближений,
$g_i = \nabla F(\theta^{(i+1)}) - \nabla F(\theta^{(i)})$ – разность градиентов.

Начальные значения входящих в формулы переменных берутся как $g_0 = \nabla F(\theta^{(0)})$, $A_0 = I$, а $\theta^{(0)}$ задано пользователем.
Вычисления прекращаются, если евклидова норма приращения очередного приближения вектора параметров меньше некоторого заранее заданного малого положительного числа ε.

Методика и численные примеры представлены в монографии Носача.

9.3.8.2. Метод Гаусса– Ньютона

Метод Гаусса– Ньютона является одним из популярных квазиньютоновских методов.
Согласно схеме метода, очередное приближение искомого вектора θ решения нелинейной системы можно найти как

$$\theta^{(i+1)} = \theta^{(i)} + \left[P_i^T(X,\theta)P_i(X,\theta)\right]^{-1} P_i^T(X,\theta)(Y - f(X,\theta^{(i)}), i = 0,1,2,...,$$

где i, i = 0,1,2,... – номер итерации,
$P_i(.,.)$, i = 0,1,2,... – матрица частных производных модели по параметрам,
X – заданный вектор независимой переменной (аргумента),
Y – заданный вектор функции выхода эксперимента,
$f(.,.)$ – вектор выхода модели, получающийся подстановкой в функцию модели заданного вектора аргумента, при фиксированном значении вектора параметров.
Начальное значение $\theta^{(0)}$ задано пользователем.
Вычисления прекращаются, если евклидова норма приращения очередного приближения вектора параметров меньше некоторого заранее заданного малого положительного числа ε.
Практически матрица частных производных вычисляется методом конечных разностей, т. к. вид модели заранее неизвестен.

Методика и численные примеры представлены в монографии Носача. Эффективные идеи даны в книгах Дэнниса (Dennis) с соавт., Дрейпера (Draper) с соавт.

9.3.9. Кусочно–линейная аппроксимация

Модель представлена (интерполирована) в виде кусочно–линейной функции

$$z(x) = \begin{cases} a_1 + b_1 x, x_1 \le x \le x_2, \\ a_2 + b_2 x, x_2 \le x \le x_3, \\ \dots \\ a_{N-1} + b_{N-1} x, x_{N-1} \le x \le x_N, \end{cases}$$

где a_i, i = 1,2,...,N – 1 – массив вычисленных свободных членов,
b_i, i = 1,2,...,N – 1 – массив вычисленных коэффициентов.
Для $x < x_1$ и для $x > x_N$ модель не определена.
Вычисления коэффициентов на основе представленных опытных данных производятся по формулам:

$$a_i = \frac{y_i x_{i+1} - y_{i+1} x_i}{x_{i+1} - x_i}, i = 1,2,..., N-1,$$

$$b_i = \frac{y_{i+1} - y_i}{x_{i+1} - x_i}, i = 1,2,..., N-1.$$

Очевидно, что вычисления производятся «точно», поэтому вывод статистических характеристик для данной модели подобно тому, как это сделано для других моделей, не имеет смысла.

В приложениях иногда возникает необходимость вычисления значения ординаты по обращенной модели. Для расчета требуемого значения следует воспользоваться кусочной формулой

$$x(y) = \begin{cases} \frac{y - a_1}{b_1}, y_1 \le y \le y_2, \\ \frac{y - a_2}{b_2}, y_2 \le y \le y_3, \\ \dots \\ \frac{y - a_{N-1}}{b_{N-1}}, y_{N-1} \le y \le y_N. \end{cases}$$

Список использованной и рекомендуемой литературы

1. Afifi A.A., Clark V. Computer–aided multivariate analysis. – Boca Raton, FL: Chapman & Hall / CRC, 1997.
2. Bates D.M., Watts D.G. Nonlinear regression analysis and its application. – New York, NY: John Wiley & Sons, 1988.
3. Chatterjee S., Hadi A.S. Regression analysis by example. – New York, NY: John Wiley & Sons, 2006.
4. Cizek P. (Non) linear regression modeling // Handbook of computational statistics: Concepts and methods / Ed. by J.E. Gentle, W. Hardle, Y. Mori. – New York, NY: Springer, 2004, pp. 621–654.
5. Dastidar S.G. Gompertz: A Scilab program for estimating Gompertz curve using Gauss–Newton method of least squares // Journal of Statistical Software, April 2006, vol. 15, no. 12.
6. Dennis J.E., Jr., Schnabel R.B. Numerical methods for unconstrained optimization and nonlinear equation. – Philadelphia, PA: The Society for Industrial and Applied Mathematics, 1996.
7. Draper N.R., Smith H. Applied regression analysis. – New York, NY: Wiley & Sons, 1998.
8. Manly B.F.J. Statistics for environmental science and management. – Boca Raton, FL: Chapman & Hall / CRC, 2001.
9. Motulsky H., Christopoulos A. Fitting models to biological data using linear and nonlinear regression. – San Diego, CA: GraphPad Software, 2003.
10. Pollard J.H. A handbook of numerical and statistical techniques with examples mainly from the life sciences. – Cambridge, NY: Cambridge University Press, 1977.
11. Rawlings J.O., Pantula S.G., Dickey D.A. Applied regression analysis: a research tool. – New York, NY: Springer–Verlag, 1998.
12. Rice J.A. Mathematical statistics and data analysis. – Belmont, CA: Wadsworth, 1995.
13. Ryan T.P. Modern regression methods. – New York, NY: John Wiley & Sons, 1997.
14. Seber G.A.F., Wild C.J. Nonlinear regression. – Hoboken, NJ: John Wiley & Sons, 2003.

15. Uusipaikka E. Confidence intervals in generalized regression models. – Boca Raton, FL: Chapman & Hall / CRC, 2009.
16. Von Eye A., Schuster C. Regression analysis for social sciences. – San Diego, CA: Academic Press, 1998.
17. Адлер Ю.П., Маркова Е.В., Грановский Ю.В. Планирование эксперимента при поиске оптимальных условий. – М.: Наука, 1976.
18. Айвазян С.А., Мхитарян В.С. Прикладная статистика и основы эконометрики. – М.: ЮНИТИ, 1998.
19. Афифи А., Эйзен С. Статистический анализ: Подход с использованием ЭВМ. – М.: Мир, 1982.
20. Банди Б. Методы оптимизации. Вводный курс. – М.: Радио и связь, 1988.
21. Бородич С.А. Вводный курс эконометрики. – Мн.: БГУ, 2000.
22. Брандт З. Анализ данных. Статистические и инженерные методы для научных работников и инженеров. – М.: Мир, ООО «Издательство АСТ», 2003.
23. Бронштейн И.Н., Семендяев К.А. Справочник по математике. – М.: Наука, 1981.
24. Васильков Ю.В., Василькова Н.Н. Компьютерные технологии вычислений в математическом моделировании. – М.: Финансы и статистика, 2002.
25. Вольтерра В. Математическая теория борьбы за существование. – М.: Наука, 1976.
26. Вучков И., Бояджиева Л., Солаков Е. Прикладной линейный Регрессионный анализ. – М.: Финансы и статистика, 1987.
27. Гайдышев И. Анализ и обработка данных: специальный справочник. – СПб: Питер, 2001.
28. Гилл Ф., Мюррей У., Райт М. Практическая оптимизация. – М.: Мир, 1985.
29. Джонстон Дж. Эконометрические методы. – М.: Статистика, 1980.
30. Доугерти К. Введение в эконометрику. – М.: ИНФРА–М, 1999.
31. Дрейпер Н., Смит Г. Прикладной регрессионный анализ. Кн. 1. – М.: Финансы и статистика, 1986.
32. Дрейпер Н., Смит Г. Прикладной регрессионный анализ. Кн. 2. – М.: Финансы и статистика, 1987.
33. Дэннис Дж., мл., Шнабель Р. Численные методы безусловной оптимизации и решения нелинейных уравнений. – М.: Мир, 1988.
34. Дюк В. Обработка данных на ПК в примерах. – СПб: Питер, 1997.
35. Каханер Д., Моулер К., Нэш С. Численные методы и программное обеспечение. – М.: Мир, 2001.
36. Корн Г., Корн Т. Справочник по математике для научных работников и инженеров. – М.: Наука, 1973.
37. Крутов В.И. Основы научных исследований / Под ред. В.И. Крутова, В.В. Попова. – М.: Высшая школа, 1989.
38. Кулаичев А.П. Компьютерный контроль процессов и анализ сигналов. – М.: Информатика и компьютеры, 1999.
39. Кулаичев А.П. Полное собрание сочинений в трех томах. Том 1. Методы и средства анализа данных в среде Windows®. STADIA. – М.: Информатика и компьютеры, 1999.
40. Ларичев О.И., Горвиц Г.Г. Методы поиска локального экстремума овражных функций. – М.: Наука, 1989.
41. Львовский Е.Н. Статистические методы построения эмпирических формул: Учебное пособие для втузов. – М.: Высшая школа,1982.
42. Молчанов И.Н. Машинные методы решения прикладных задач. Алгебра, приближение функций. – Киев: Наукова думка, 1987.
43. Мостеллер Ф., Тьюки Дж. Анализ данных и регрессия. Вып. 1. – М.: Финансы и

статистика, 1982.
44. Мостеллер Ф., Тьюки Дж. Анализ данных и регрессия. Вып. 2. – М.: Финансы и статистика, 1982.
45. Мудров А.Е. Численные методы для ПЭВМ на языках Бейсик, Фортран и Паскаль. – Томск: МП «РАСКО», 1991.
46. Носач В.В. Решение задач аппроксимации с помощью персональных компьютеров. – М.: МИКАП, 1994.
47. Ортега Дж., Рейнболдт В. Итерационные методы решения нелинейных систем уравнений со многими неизвестными. – М.: Мир, 1975.
48. Осовский С. Нейронные сети для обработки информации. – М.: Финансы и статистика, 2002.
49. Пен Р.З. Статистические методы моделирования и оптимизации процессов целлюлозно–бумажного производства: Учебное пособие. – Красноярск: Издательство КГУ, 1982.
50. Петрович М.Л. Регрессионный анализ и его математическое обеспечение на ЕС ЭВМ: Практическое руководство. – М.: Финансы и статистика, 1982.
51. Поллард Дж. Справочник по вычислительным методам статистики. – М.: Финансы и статистика, 1982.
52. Прохоров Ю.В. Вероятность и математическая статистика. Энциклопедия / Гл. ред. Ю.В. Прохоров. – М.: Научное издательство «Большая Российская энциклопедия», 1999.
53. Прохоров Ю.В. Математический энциклопедический словарь / Гл. ред. Ю.В. Прохоров. – М.: Научное издательство «Большая Российская энциклопедия», 1995.
54. Прохоров Ю.В. Физический энциклопедический словарь / Гл. ред. Ю.В. Прохоров. – М.: Научное издательство «Большая Российская энциклопедия», 1995.
55. Родионов Д.А. Справочник по математическим методам в геологии / Д.А. Родионов, Р.И. Коган, В.А. Голубева и др. – М.: Недра, 1987.
56. Сергиенко В.И., Бондарева И.Б. Математическая статистика в клинических исследованиях. – М.: ГЭОТАР–Медиа, 2006.
57. Ферстер Э., Ренц Б. Методы корреляционного и регрессионного анализа. Руководство для экономистов. – М.: Финансы и статистика, 1983.
58. Шуп Т. Решение инженерных задач на ЭВМ: Практическое руководство. – М.: Мир, 1982.

Глава 10. Корреляционный анализ

10.1. Введение

В программном обеспечении исследуется корреляция и связи типа корреляции:

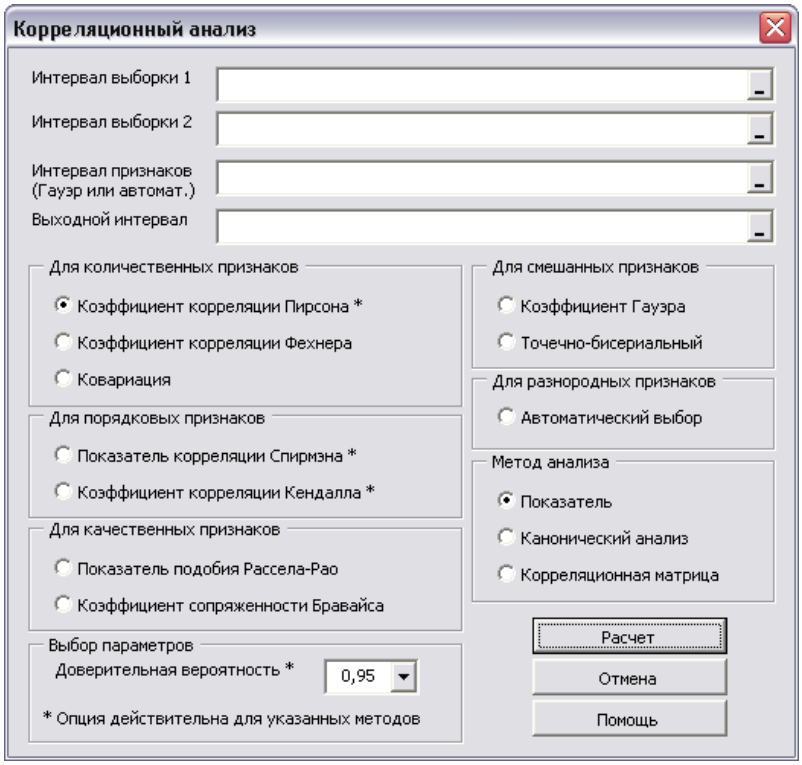
- количественных признаков,
- порядковых признаков,
- номинальных признаков,
- смешанных признаков,
- разнородных признаков.

Также выполняется канонический корреляционный анализ.

10.2. Работа с программным обеспечением

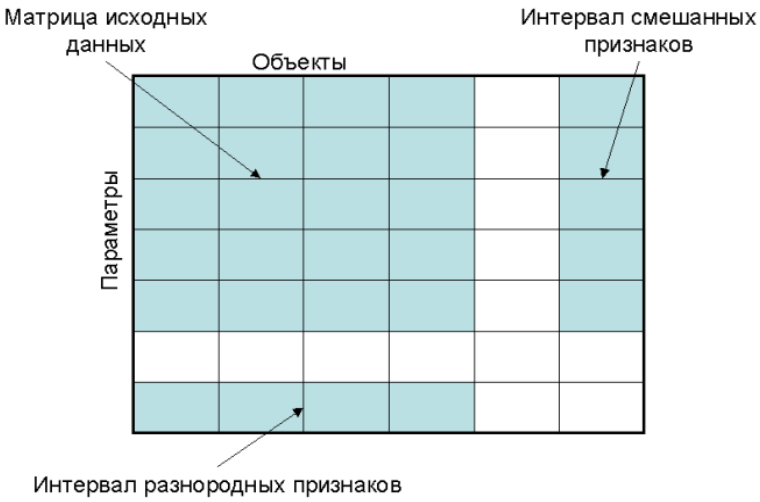
Выберите из меню программы пункт AtteStat | Корреляционный анализ. На экране появится

диалоговое окно, изображенное на рисунке:



Затем проделайте следующие шаги:

- Выберите или введите интервалы вектора или матрицы исходных данных (выборки 1).
- Выберите или введите интервалы вектора или матрицы исходных данных (выборки 2) для методов «Исследование корреляции» и «Канонический анализ».
- Для вычисления корреляционной матрицы разнородных или смешанных признаков выберите или введите интервал типов признаков, равный, соответственно методу, числу столбцов или строк в матрице данных. При этом значение признаков в данном интервале соответствует: 0 для количественного признака, 1 для порядкового, 2 для дихотомического. Показанный ниже рисунок поясняет, каким образом соотносятся матрица исходных данных и интервалы разнородных или смешанных признаков. Обозначения параметров и объектов условны – они могут поменяться местами в зависимости от условий задачи. Важно только знать, что программой строится матрица корреляций отмеченных пользователем столбцов электронной таблицы (порядок построенной корреляционной матрицы равен числу столбцов) между собой. Поэтому, если требуется транспонировать матрицу исходных данных, воспользуйтесь специальным инструментом из главы «Матричная и линейная алгебра».



Выберите или введите выходной интервал. Начиная с первой ячейки выходного интервала (следовательно, можно указать только одну ячейку, т.к. остальные ячейки интервала игнорируются), будут выведены результаты вычислений.

- Выберите или оставьте по умолчанию коэффициент корреляции в зависимости от характера исходных данных. Программное обеспечение само не определяет автоматически тип исходных данных, поэтому пользователь должен выбрать коэффициент корреляции, адекватный его исходным данным. Обратите внимание, что для качественных (дихотомических) признаков предусмотрены только два значения вариант выборок: 0 при отсутствии признака, 1 при наличии признака. Типы признаков для разнородных или смешанных данных определяются, как рассказано выше. Для разнородных признаков коэффициент корреляции (типа корреляции) определяется автоматически в зависимости от типов признаков.
- Выберите или оставьте по умолчанию метод анализа и дополнительные параметры, относящиеся к данному методу.
- Нажмите кнопку Расчет.

После выполнения вычислений будет, начиная с первой ячейки выходного интервала, выведено название метода и результаты расчета.
За выбор адекватного исходным данным метода расчета несет ответственность пользователь. Программа берет на себя верификацию исходных данных, выдавая подробную диагностику. При неверных действиях пользователя выдаются сообщения об ошибках.

10.2.1. Сообщения об ошибках

При ошибках ввода данных могут выдаваться диагностические сообщения следующих указанных типов.

Ошибка	Комментарий
Не определен интервал переменной.	Вы не выбрали или неверно ввели входной интервал. Лучшим способом избежать ошибки является не ввод, а выделение интервала стандартным образом, т. е. протаскиванием курсора. Ошибка может быть вызвана также тем обстоятельством, что для метода «Исследование корреляции» требуется указание двух выборок, а для метода «Канонический анализ» требуется указание двух матриц исходных данных.

Ошибка	Комментарий
Пустая ячейка в области данных.	Проверьте исходные данные и заполните все ячейки, отмеченные Вами как входной интервал. Для избежания ошибок расчета, вызванных разногласиями, трактовать ли пустую ячейку как нуль, программное обеспечение требует заполнения всех ячеек. Если в ячейке не должно быть данных по физической природе исследуемого процесса, введите в данную ячейку нуль.
Нечисловой тип данных.	Проверьте типы ячеек входного интервала. Тип может быть только числовым. Проще всего выделить интервал ячеек и явно определить их тип как числовой стандартными средствами.
Разные численности выборок.	Методы «Исследование корреляции» и «Канонический анализ» оперируют, соответственно, векторами или матрицами данных. Вектора или матрицы состоят из выборок, содержащих равные количества наблюдений. Убедитесь, что все выборки исходных данных содержат равные количества наблюдений.
Не определена область вывода.	Не выбран или неверно введен выходной интервал. Лучшим способом избежать ошибки является не ввод, а выделение интервала стандартным образом.
Мало выборок в 1 группе.	Аналогичное сообщение может быть выдано и для группы 2. Для метода «Канонический анализ» число выборок в матрице данных не должно быть менее двух. Если это не так, используйте метод «Исследование корреляции» для признаков соответствующего типа.
Мало данных.	Сообщение выдается в случае, если число вариант (наблюдений) анализируемой исходной выборки меньше двух. Для применения любых методов корреляционного анализа число вариант выборки не должно быть менее двух. Максимальные численности выборок ограничены только требованиями.
Мало наблюдений в матрице данных.	При построении корреляционной матрицы число наблюдений не должно быть меньше, чем это определяется требованиями алгоритмов.
Несоответствие меры и метода.	Некоторые сочетания мер (коэффициентов корреляции) и реализованных в программе методов не являются допустимыми. Например, нельзя выбирать точечно–бисеральный коэффициент корреляции и строить с его помощью корреляционную матрицу.
Несовместимость массивов.	При использовании смешанных признаков (исследование с помощью коэффициента Гауэра) численность интервала типов признаков должна быть равна числу строк в матрице данных. При исследовании корреляции разнородных признаков численность интервала признаков должна быть равна числу столбцов в матрице данных. При этом значение признаков в данном интервале соответствует: 0 для количественного признака, 1 – для порядкового, 2 – для дихотомического.
Ошибка в исходных данных.	Недопустимое значение в интервале признаков при исследовании корреляции смешанных или разнородных признаков. Допустимые значения: 0 для количественного признака, 1 – для порядкового, 2 – для дихотомического.

Ошибка	Комментарий
Данные не номинального типа.	Была сделана попытка расчета с помощью метода, разработанного только для номинальных признаков, однако представленные данные не являются номинальными.

10.3. Теоретическое обоснование

В практических наблюдениях часто бывают случаи, когда зависимости не имеют функционального характера – равномерному изменению одного признака соответствует изменение величины другого признака в среднем. Такой вид соотношений называется корреляционной зависимостью, или корреляцией. Корреляционным анализом называется совокупность методов обнаружения корреляционной зависимости между случайными величинами или признаками. Считается, что исследование взаимной зависимости приводит к теории корреляции, тогда как изучение зависимости ведет к теории регрессии.
Выделяется также случай функциональной зависимости между величинами, измерения которых, возможно, подвержены ошибкам наблюдений или измерений. Под функциональной связью понимается такой род соотношения между двумя признаками, когда любому значению одного признака всегда соответствует определенное одно и то же значение другого. Функциональная зависимость отражает физические взаимосвязи изучаемого явления и может изучаться методами математического моделирования. Подробнее данные вопросы рассматриваются в главе «Регрессионный анализ».
В программе по умолчанию предполагается, что объекты располагаются по столбцам электронной таблицы. По строкам располагаются параметры, описывающие объекты. Это существенно для многомерных методов, оперирующих матрицами исходных данных. Если требуется повести исследование транспонированной матрицы исходных данных, для быстрого выполнения данной операции можно воспользоваться методом главы «Матричная и линейная алгебра».
Отметим, что возможность вычисления корреляционной матрицы, в том числе для признаков различных и смешанных типов, позволяет использовать корреляционную матрицу для факторного анализа указанных типов признаков, реализованного в главе «Факторный анализ».

10.3.1. Корреляция количественных признаков

В данном разделе представлены методы исследования корреляции количественных признаков:

- коэффициент корреляционного отношения Пирсона, применяемый для измерения тесноты связи при прямолинейной корреляции,
- коэффициент корреляции Фехнера.

Дополнительно предоставлена возможность расчета ковариации и ковариационной матрицы. Данный показатель может быть необходим для применения в других методах, например, для «ручного» расчета критерия Уилкса, описанного в главе «Дисперсионный анализ».
Коэффициенты ранговой корреляции, которые исследуют корреляцию порядковых признаков (рангов), пусть и полученных из признаков количественных (путем применения операции присвоения рангов), помещены в раздел «Корреляция порядковых признаков».
Полученные в результате применения линейных методов корреляционного анализа выводы могут подтвердить или опровергнуть гипотезу о существовании линейной зависимости между рядами, но не связи другого типа. Вывод в этом случае такой: чем ближе вычисленная величина корреляционного отношения к 0, тем слабее сила линейной связи между рядами, чем ближе вычисленная величина к значению +1 (полная положительная корреляция) или к

значению –1 (полная отрицательная корреляция), тем сильнее сила линейной связи.

10.3.1.1. Коэффициент корреляционного отношения Пирсона

Коэффициент корреляционного отношения Пирсона (коэффициент корреляции, выборочный коэффициент корреляции, коэффициент корреляции Бравайса – Пирсона) измеряет силу линейной корреляционной связи количественных признаков. Выборочная оценка коэффициента корреляции вычисляется по формуле

$$\hat{r} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}},$$

где $x_i, i = 1,2,...,n$ – варианты первой компоненты 2–мерной выборки,

$\bar{x}$ – соответствующее среднее значение,

$y_i, i = 1,2,...,n$ – варианты второй компоненты 2–мерной выборки,

$\bar{y}$ – соответствующее среднее значение,

n – численность 2–мерной выборки.

Иначе коэффициент корреляции может оказаться удобным вычислить как

$$\hat{r} = \frac{Cov(X,X)Cov(Y,Y)}{\sqrt{Cov(X,Y)}},$$

где X – первая компонента,

Y – вторая компонента,

$Cov(.,.)$– выборочная ковариация.

Использование коэффициента корреляции оправдано лишь тогда, когда совместное распределение пары количественных признаков соответствует 2–мерному нормальному распределению. Частой грубой ошибкой в публикациях является игнорирование этой предпосылки применения рассматриваемого показателя, поэтому перед вычислением коэффициента Пирсона следует проверить нормальность 2–мерной выборки с помощью методов из главы «Проверка нормальности распределения».

При $|\hat{r}| < 1$ программой вычисляются еще ряд параметров. Доверительный интервал оцениваемого коэффициента корреляции нормальной двумерной генеральной совокупности вычисляется как

$$r \in \left[\tanh\left(z(\hat{r}) - \frac{N_{(1+p)/2}}{\sqrt{n-3}}\right); \tanh\left(z(\hat{r}) + \frac{N_{(1+p)/2}}{\sqrt{n-3}}\right)\right],$$

где r – истинное значение коэффициента корреляции,

$N_{(1+p)/2}$ – квантиль нормального распределения,

p – стандартное значение доверительного уровня,

$z(.)$ – z–преобразование выборочного коэффициента корреляции.

Нормализующее z–преобразование выборочного коэффициента корреляции вычисляется как гиперболический арктангенс действительной переменной по формуле

$$z(\hat{r}) = Arth(\hat{r}) = \frac{1}{2}\ln\frac{1+\hat{r}}{1-\hat{r}} \text{ при } |\hat{r}| < 1.$$

Коэффициент корреляции может применяться для проверки гипотезы независимости признаков (некоррелированности, значимости связи) следующим образом. Гипотеза о

независимости признаков отвергается на выбранном уровне значимости, если вычисленное по опытным данным значение коэффициента корреляции превосходит (по модулю) критическое. В случае нормального распределения исходных данных величина выборочного коэффициента корреляции считается значимо отличной от нуля, если выполняется неравенство

$$r^2 > \left[1 + (n-2)/t_\alpha^2\right]^{-1},$$

где t_α – критическое значение *t*–распределения с $n - 2$ степенями свободы.
Иначе говоря, величина

$$t_r = |r| \sqrt{\frac{n-2}{1-r^2}}$$

имеет *t*–распределение с $n - 2$ степенями свободы.
Кроме того, распределение величины $z(r)$ уже при небольших значениях *n* приближается нормальным распределением с математическим ожиданием, равным

$$Mz = \frac{1}{2}\ln\frac{1+r}{1-r} + \frac{r}{2(n-1)} + \ldots$$

и дисперсией

$$Dz = \frac{1}{n-3} + \ldots,$$

где опущены слагаемые, малые по сравнению с оставленными слагаемыми.
Таким образом, случайная величина

$$\frac{z(r) - Mz}{\sqrt{Dz}}$$

распределена приближенно по стандартному нормальному закону $N(0,1)$.
При исследовании многомерной совокупности случайных величин из коэффициентов корреляции, вычисленных попарно между случайными величинами, составляется квадратная симметрическая корреляционная матрица с единицами на главной диагонали, которая служит основным элементом при построении многих алгоритмов многомерной статистики, например, в факторном анализе.

Вывод см. в учебном пособии Львовского. О вычислении коэффициента корреляции, включая доверительные интервалы, см. монографию Айвазяна с соавт., работы Альтмана (Altman) с соавт., таблицы Большева с соавт., Мюллера с соавт. О проверке значимости см. также монографии Ферстера с соавт., Зайцева. О сравнении коэффициентов корреляции двух независимых совокупностей см. также монографии Мюллера с соавт., Родионова, а также работы Уильямса (Williams), Вольфе (Wolfe), Лемешко с соавт.

10.3.1.2. Коэффициент корреляции Фехнера

Коэффициент корреляции Фехнера (фехнеровский коэффициент корреляции, индекс Фехнера) был предложен для изучения корреляции количественных признаков. При вычислении коэффициента происходит понижение количественной шкалы до номинальной шкалы. В расчетах участвуют только количественные признаки (по ним вычисляются средние значения), поэтому метод представлен в разделе, посвященном количественным признакам. Вычисления производятся по формуле

$$r_F = \frac{C - H}{C + H},$$

где C – число совпадений знаков отклонений вариант от соответствующих средних значений,
H – число несовпадающих знаков.
Коэффициент корреляции Фехнера с успехом применяется также и для изучения корреляции «чисто» номинальных признаков. В этом случае, в соответствующих обозначениях, приведенная формула может быть записана как

$$r_F = \frac{a + d - b - c}{a + b + c + d},$$

где a, b, c, d – значения в клетках таблицы 2 x 2.
Коэффициент корреляции Фехнера может применяться для проверки гипотезы независимости признаков (некоррелированности, значимости связи). В этом случае можно произвести вычисление по формуле

$$t_F = \frac{a + d - b - c - 1}{\sqrt{a + b + c + d}} = \frac{r_F n - 1}{\sqrt{n}},$$

где $n = a + b + c + d$.
Критические значения статистики t_F приближенно распределены по стандартному нормальному закону $N(0,1)$.

В литературе встречаются и иные формулировки коэффициента корреляции Фехнера. См. монографии Лакина, Ферстера с соавт.

10.3.1.3. Ковариация

Ковариация (covariance) – числовая характеристика совместного распределения двух случайных величин X и Y. Иногда говорят о 2–мерной случайной величине, причем под X понимают первую компоненту, а под Y – вторую компоненту указанной случайной величины.
Ковариация определяется формулой
$Cov(X,Y) = E|(X - EX)(Y - EY)|$,
где E – символ математического ожидания.
Значение $Cov(X,X)$ по определению является дисперсией случайной величины X.
Выборочная ковариация вычисляется как

$$Cov(X,Y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}),$$

где $x_i, i = 1,2,...,n$ – варианты первой компоненты 2–мерной выборки,
$\bar{x}$ – соответствующее среднее значение,
$y_i, i = 1,2,...,n$ – варианты второй компоненты 2–мерной выборки,
$\bar{y}$ – соответствующее среднее значение,
n – численность 2–мерной выборки.
Ковариация не может служить в качестве показателя типа корреляции, т.к. не обладает свойствами данного показателя. В частности, ковариация не является безразмерной величиной. Ее максимальное значение не ограничивается единицей.
Ковариация предлагается здесь из–за технического удобства вычисления. Необходимость вычисления ковариации в настоящем программном обеспечении вызвана тем, что в ряде разновидностей множественного статистического анализа (дисперсионный анализ, факторный анализ) находит применение ковариационная матрица, элементами которой служат попарные ковариации компонент случайного вектора.

Теоретическое обоснование происхождения ковариации лучше всего посмотреть у Ван дер

Вардена.

10.3.2. Корреляция порядковых признаков

В данном разделе рассмотрены методы исследования связи типа корреляции признаков, измеренных в порядковой шкале, либо признаков, приведенных к порядковой шкале, (ранговой корреляции). В программе исследованы:

- показатель ранговой корреляции Спирмэна,
- коэффициент ранговой корреляции Кендалла.

Обзор коэффициентов ранговой корреляции (включая проверку значимости) см. в работах Филлера (Fieller) с соавт.

10.3.2.1. Показатель ранговой корреляции Спирмэна

Показатель ранговой корреляции Спирмэна (показатель корреляции рангов Спирмэна, коэффициент корреляции рангов, коэффициент корреляции Спирмэна, коэффициент ранговой корреляции ρ, Spearman rank correlation) применяется в случае, если изучается линейная связь между рядами, представленными в количественной или порядковой шкале. Следует заметить, что при анализе количественных признаков применять показатель Спирмэна вместо коэффициента корреляционного отношения Пирсона не следует, если для этого не существует веских оснований, так как при его вычислении происходит понижение количественной шкалы до порядковой шкалы. Поэтому наиболее широкое применение показатель Спирмэна нашел при анализе корреляции порядковых признаков.
Расчет выборочной оценки показателя ранговой корреляции ведется по формуле

$$\hat{\rho}_S = 1 - \frac{6(S_\rho + B_x + B_y)}{n^3 - n}; S_\rho = \sum_{i=1}^{n} (r_i - s_i)^2 ,$$

где $r_i, s_i, i = 1,2,...,n$ – массивы рангов анализируемых рядов,
n – число пар вариант исследуемых рядов,
B_x, B_y – поправки на объединение рангов в соответствующих рядах, вычисляемые по формуле

$$B_{\cdot} = \frac{1}{12} \sum_{i=1}^{m} n_i (n_i^2 - 1) ,$$

где m – число групп объединенных рангов в ряду,
$n_i, i = 1,2,...,m$ – число рангов в i–ой группе.
Доверительный интервал оцениваемого показателя Спирмэна вычисляется аналогично коэффициенту Пирсона.
Показатель ранговой корреляции Спирмэна может применяться для проверки гипотезы независимости признаков (некоррелированности, значимости связи) следующим образом. Гипотеза о независимости признаков отвергается на выбранном уровне значимости, если вычисленное по опытным данным значение коэффициента корреляции превосходит (по модулю) критическое. Можно также произвести вычисление по формуле

$$t_\rho = |\rho_S| \sqrt{\frac{n-2}{1-\rho_S^2}} ,$$

где критические значения статистики t_ρ имеют t–распределение с n – 2 степенями свободы.

См. т. 2 Справочника под ред. Э. Ллойда и др., монографии Ферстера с соавт., Зайцева, Лакина, Малета с соавт., статью Артузи (Artusi) с соавт.

10.3.2.2. Коэффициент ранговой корреляции Кендалла

Коэффициент ранговой корреляции Кендалла (коэффициент корреляции рангов, ранговый коэффициент корреляции, коэффициент корреляции Кендэла, τ Кендалла, Kendall rank correlation) предназначен для вычисления силы корреляционной связи между двумя рядами при тех же условиях, что и рассмотренный выше показатель Спирмэна. Коэффициент Кендалла считается более строгой оценкой по сравнению с показателем ранговой корреляции Спирмэна.
Все основные замечания, данные при описании показателя Спирмэна, справедливы и в отношении коэффициента Кендалла.
Расчет выборочной оценки коэффициента ранговой корреляции ведется по формуле

$$\hat{\tau} = \frac{S_\tau}{\sqrt{\left(\frac{n(n-1)}{2} - B_x\right)\left(\frac{n(n-1)}{2} - B_y\right)}}, \quad S_\tau = \sum_{i=1}^{n}\sum_{j=i+1}^{n} sign(r_j - s_i),$$

где $r_i, s_i, i = 1,2,...,n$ – массивы рангов анализируемых рядов,
n – число пар вариант исследуемых рядов,
B_x, B_y – поправки на объединение рангов в соответствующих рядах, вычисляемые по формуле

$$B_\cdot = \frac{1}{2}\sum_{i=1}^{m} n_i(n_i - 1),$$

где m – число групп объединенных рангов в ряду,
$n_i, i = 1,2,...,m$ – число рангов в i–ой группе.
Доверительный интервал оцениваемого коэффициента Кендалла может вычисляться разными методами. В программе доверительный интервал вычисляется по формуле Нётера

$$\tau \in \left[\hat{\tau} - \frac{2\sigma\Psi(1-(1-p)/2)}{n(n-1)}; \hat{\tau} + \frac{2\sigma\Psi(1-(1-p)/2)}{n(n-1)}\right],$$

где Ψ(.) – функция, обратная функции стандартного нормального распределения,
p – стандартное значение доверительного уровня,
σ – величина, определяемая из формулы:

$$\sigma^2 = 4\sum_{i=1}^{n} C_i^2 - 2\sum_{i=1}^{n} C_i - \frac{2(2n-3)}{n(n-1)}\left(\sum_{i=1}^{n} C_i\right)^2,$$

где $C_i, i = 1,2,...,n,$ – вспомогательные величины, вычисляемые как

$$C_i = \sum_{\substack{j=1 \\ i \neq j}}^{n} \delta(r_i, r_j, s_i, s_j), i = 1,2,...,n,$$

где δ(.,.,.,.) – величины, вычисляемые по формуле

$$\delta(a,b,c,d) = \begin{cases} 1, (a-b)(c-d) > 0, \\ 0, (a-b)(c-d) < 0. \end{cases}$$

Проблема, однако, заключается в том, что для некоторых наборов данных величина σ², рассчитанная по показанной выше формуле, может оказаться отрицательной. Пример таких данных:

1,059	1,242
1,091	1,237
1,849	1,11
1,943	2,691
2,416	1,352
5,134	5,705
5,29	4,055
7,344	3,257
7,435	3,772

В этом случае метод Нётера оказывается несостоятельным и доверительный интервал вычисляется программой стандартно как

$$\tau \in \left(\hat{\tau} - t_{(1+\beta)/2} \frac{\sigma}{\sqrt{n}}; \hat{\tau} + t_{(1+\beta)/2} \frac{\sigma}{\sqrt{n}} \right),$$

где σ – стандартное отклонение,
$t_{(1+\beta)/2}$ – значение обратной функции *t*–распределения Стьюдента с параметрами $n - 1$ и $(1 + \beta) / 2$,
β – доверительный уровень, выраженный в долях.
При этом стандартное отклонение считается как

$$\sigma = \sqrt{\frac{2(2n+5)}{9n(n-1)}}.$$

Коэффициент ранговой корреляции Кендалла может применяться для проверки гипотезы независимости признаков (некоррелированности, значимости связи) следующим образом. Гипотеза о независимости признаков отвергается на выбранном уровне значимости, если вычисленное по опытным данным значение коэффициента корреляции превосходит (по модулю) критическое. В случае больших выборок можно произвести вычисление по формуле

$$t_\tau = |\tau| \sqrt{\frac{9n(n-1)}{2(2n+5)}},$$

где критические значения статистики t_τ приближенно распределены по стандартному нормальному закону $N(0,1)$.

См. т. 2 Справочника под ред. Э. Ллойда и др., монографии Холлендера с соавт., Зайцева, статью Самара (Samara) с соавт.

10.3.3. Корреляция номинальных признаков

В данном разделе представлены методы исследования связи типа корреляции для признаков, измеренных в номинальной шкале либо приведенных к номинальной шкале. Особо отметим введенную выше поправку «типа корреляции», т.к. обычная корреляция для номинальных признаков не определена.
Рассмотрены коэффициенты (показатели подобия)

- Рассела–Рао,
- Бравайса.

Коэффициенты предназначены для оценки связи между дихотомическими (номинальными с числом градаций, равным двум, иначе качественными) признаками. Эти и другие показатели находят широкое применение в кластерном анализе, где они именуются также мерами сходства типа корреляции. Подробнее см. главу «Кластерный анализ».
Для исследования корреляции признаков, измеренных в номинальной шкале с числом градаций признаков больше двух (категоризированных данных), используются методы

анализа двумерных таблиц сопряженности (кросстабуляции), выполняемого с помощью методов главы «Кросстабуляция». Представленные показатели принято именовать мерами связи.
В программе предполагается, что дихотомическая переменная может принимать только значения 1 (верхний уровень) и 0 (нижний уровень). Значение варианты выборки, равное 0, указывает на отсутствие переменной или признака, значение, равное 1 – на наличие.
Например, в ячейку (клетку) *a* записано число пар элементов массивов 1 и 2, одновременно имеющих признак, равный 1. В ячейку *c* записано число пар элементов массивов 1 и 2, в которых значение элемента массива 1 равно 1, а значение элемента массива 2 равно 0 и т. д.
Отметим, что при вычислении представленных показателей путаница между ячейками *b* и *c* не ведет к каким–либо неприятностям.

10.3.3.1. Коэффициент Рассела–Рао

Коэффициент Рассела–Рао (показатель подобия Рассела–Рао) вычисляется по формуле

$$r = \frac{a}{a+b+c+d},$$

где *a*, *b*, *c*, *d* – значения в клетках таблицы 2 x 2.
Коэффициент Рассела–Рао может применяться для проверки гипотезы независимости признаков (значимости связи). В этом случае статистика

$$t_r = \frac{r \cdot n - 1}{\sqrt{n}},$$

где $n = a + b + c + d$ – сумма таблицы,
приближенно распределена по стандартному нормальному закону $N(0,1)$.

10.3.3.2. Коэффициент сопряженности Бравайса

Специальная форма коэффициента корреляции – коэффициент сопряженности Бравайса (φ–коэффициент ассоциации Пирсона, коэффициент контингенции Пирсона, тетрахорический показатель связи) – рассчитывается по формуле

$$\varphi = \frac{ad - bc - 0{,}5 \cdot n}{\sqrt{(a+b)(a+c)(d+b)(d+c)}},$$

где *a*, *b*, *c*, *d* – значения в клетках таблицы 2 x 2.
$0{,}5 \cdot n$ – поправка на непрерывность Йэйтса,
$n = a + b + c + d$ – сумма таблицы,
В литературе встречаются и иные, эквивалентные, формулировки рассматриваемого коэффициента. Например, показанная формула фактически совпадает с формулой

$$\varphi = \sqrt{\frac{\chi^2}{n}},$$

где χ^2 – статистика хи–квадрат Пирсона.
Коэффициент сопряженности Бравайса может применяться для проверки гипотезы независимости признаков (значимости связи). В этом случае статистика

$$t_\varphi = |\varphi| \sqrt{\frac{n-2}{1-\varphi^2}}$$

приближенно имеет *t*–распределение с $n - 2$ степенями свободы.

См. монографии Лакина, Ферстера с соавт., Малета с соавт.

10.3.4. Корреляция признаков, измеренных в различных шкалах

Настоящий раздел посвящен исследованию корреляции признаков, измеренных в различных (смешанных) шкалах. Рассмотрены:

- коэффициент Гауэра,
- точечно–бисериальная корреляция, позволяющая исследовать корреляцию в некоторых частных случаях.

Еще одной интересной возможностью программы является исследование корреляции разнородных признаков.

Проблема исследования корреляции в таблицах данных, полученных измерением параметров, относящихся к различным шкалам, довольно часто возникает в практике, особенно в медико–биологических исследованиях. Полученные результаты решения ценны как сами по себе, так и в качестве исходных данных для других методов исследования.

10.3.4.1. Коэффициент Гауэра

Коэффициент Гауэра допускает одновременное использование признаков, измеренных в шкалах: количественной, порядковой и дихотомической. Могут анализироваться выборки (например, описывающие параметры пациента), содержащие в себе признаки различных типов. Так, часть параметров может быть количественной (например, результаты инструментальных измерений), часть – порядковой (например, результаты исследований в баллах), часть – дихотомической (например, наличие или отсутствие некоторых симптомов). Вычисление элемента матрицы сходства, построенной на основе коэффициента Гауэра, производится по формуле:

$$s_{ij} = \frac{\sum_{k=1}^{p} S_{ijk}}{\sum_{k=1}^{p} W_{ijk}}, i = 1,2,...,n,\ j = 1,2,...,n,$$

где S_{ijk}, $i = 1,2,...,n$; $j = 1,2,...,n$; $k = 1,2,...,p$ – вклад признака в сходство объектов,
W_{ijk}, $i = 1,2,...,n$; $j = 1,2,...,n$; $k = 1,2,...,p$ – весовая переменная признака,
p – число признаков, характеризующих объект,
n – число объектов.

10.3.4.1.1. Расчет вклада признаков

- Для дихотомических признаков алгоритм подсчета вклада признака и взятия весовых переменных совпадает с коэффициентом Жаккара

$$J = \frac{a}{a+b+c},$$

где a, b, c – значения в клетках таблицы 2 x 2.

- Для порядковых признаков алгоритм вычисления вклада признака совпадает с хемминговым расстоянием, если последнее мысленно обобщить на порядковые переменные, а весовые переменные берутся равными 1 для каждого участвующего в расчете порядкового признака.

$$H = a + d,$$

где a, b – значения в клетках таблицы 2 x 2.

- Для количественных признаков

$$S_{ijk} = 1 - \frac{\left|x_{ik} - x_{jk}\right|}{R_k},$$

где x_{ik} и x_{jk} – значения k–й переменной для объектов i и j,
R_k – размах k–го признака, вычисленный по всем объектам,
а весовые переменные берутся аналогично случаю порядковых признаков.
Чтобы программа имела возможность различить шкалы измерения признаков, на этапе выбора исходных данных требуется ввести интервал признаков, как это подробно проиллюстрировано в разделе, посвященном работе с программным обеспечением.

Коэффициент Гауэра реализован только для метода «Матрица», поэтому в рассматриваемом случае следует выделить матрицу исходных данных в виде столбцов равной численности.

10.3.4.2. Точечно–бисериальная корреляция

Если одна переменная дихотомизирована, а другая измерена в количественной шкале, вычисляется точечно–бисериальный коэффициент корреляции (точечный двухсерийный коэффициент корреляции). Имеют место несколько эквивалентных формул вычисления выборочной оценки коэффициента, например,

$$\hat{r}_{pb} = \frac{(\bar{x}_1 - \bar{x}_0)}{s_n} \cdot \frac{\sqrt{n_1 n_0}}{n},$$

где $\bar{x}_1$ – среднее вариант количественной выборки, соответствующих событиям верхнего уровня дихотомической выборки,
$\bar{x}_0$ – среднее вариант количественной выборки, соответствующих событиям нижнего уровня дихотомической выборки,
s_n – среднее квадратичное значение количественной выборки,
n_1 – число событий верхнего уровня,
n_0 – число событий нижнего уровня.
Средние значения вычисляются по формулам, соответственно,

$$\bar{x}_1 = \frac{1}{n_1} \sum_{\substack{i=1 \\ a_i=1}}^{n} x_i \quad \bar{x}_0 = \frac{1}{n_0} \sum_{\substack{i=1 \\ a_i=0}}^{n} x_i,$$

и

где x_i, $i = 1,2,...,n$ – количественная выборка,
a_i, $i = 1,2,...,n$ – дихотомическая выборка,
$n = n_1 + n_0$ – численность пар анализируемых выборок.
В программе предполагается, что дихотомическая переменная может принимать только значения 1 (верхний уровень) и 0 (нижний уровень).
Выборочное среднее квадратичное отклонение вычисляется по формуле

$$s_n = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2},$$

где $\bar{x}$ – выборочное среднее, которое вычисляется по формуле

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

При $|\hat{r}_{pb}| < 1$ программой вычисляются еще ряд параметров. Доверительный интервал оцениваемой точечно–бисериального коэффициента корреляции вычисляется как

$$r_{pb} \in \left[\tanh\left(z(\hat{r}_{pb}) - \frac{N_{(1+p)/2}}{\sqrt{n-3}} \right); \tanh\left(z(\hat{r}_{pb}) + \frac{N_{(1+p)/2}}{\sqrt{n-3}} \right) \right],$$

где $N_{(1+p)/2}$ – квантиль нормального распределения,
p – стандартное значение доверительного уровня,
$z(.)$ – z–преобразование выборочного коэффициента корреляции.
Нормализующее z–преобразование выборочного точечно–бисериального коэффициента корреляции вычисляется как гиперболический арктангенс действительной переменной по формуле

$$z(\hat{r}_{pb}) = Arth(\hat{r}_{pb}) = \frac{1}{2}\ln\frac{1+\hat{r}_{pb}}{1-\hat{r}_{pb}} \quad \text{при} \quad |\hat{r}_{pb}| < 1.$$

Точечно–бисериальный коэффициент корреляции может применяться для проверки гипотезы независимости признаков (некоррелированности, значимости связи) следующим образом. Гипотеза о независимости признаков отвергается на выбранном уровне значимости, если вычисленное по опытным данным значение коэффициента корреляции превосходит (по модулю) критическое. В случае нормального распределения исходных данных величина выборочного коэффициента корреляции считается значимо отличной от нуля, если выполняется неравенство

$$r_{pb}^{\ 2} > \left[1+(n-2)/t_{\alpha}^{2}\right]^{-1},$$

где t_{α} – критическое значение t–распределения с $n-2$ степенями свободы.
Иначе говоря, величина

$$t_r = | r_{pb} | \sqrt{\frac{n-2}{1-r_{pb}^2}}$$

имеет t–распределение с $n-2$ степенями свободы.
Отметим, что результаты вычисления точечно–бисериального коэффициента корреляции и коэффициента корреляции Пирсона, хотя и при различных исходных предпосылках, в случае формальной подстановки в формулу последнего тех же числовых данных, совпадают.

См. монографии Лакина, Зайцева, Мак–Немара (McNemar).

10.3.5. Корреляция разнородных признаков

Представленное программное обеспечение обладает уникальными в своем роде возможностями, важнейшими из которых является исследование корреляции признаков, измеренных в различных шкалах. Для исследования корреляции признаков, измеренных в смешанных шкалах (один объект описывается вектором данных, принадлежащих к различным шкалам), применяется коэффициент Гауэра. Метод применяется для исследования корреляции объектов. Не менее часто возникает задача исследования корреляции разнородных признаков, для решения которых предназначена описываемая ниже опция.
При построении корреляционной матрицы разнородных признаков программное обеспечение «Корреляционный анализ» автоматически выбирает следующие правила вычисления коэффициентов корреляции:

- При вычислении корреляции двух количественных параметров – коэффициент Пирсона.
- При вычислении корреляции порядковых/количественных и порядковых параметров – коэффициент ранговой корреляции Кендалла.
- При вычислении корреляции двух дихотомических признаков – коэффициент сопряженности Бравайса.
- При вычислении корреляции количественных/порядковых и дихотомических

признаков – точечно–бисериальная корреляция.
Из вычисленных корреляций формируется общая корреляционная матрица. Чтобы программа имела возможность различить шкалы измерения признаков, на этапе выбора исходных данных требуется ввести интервал признаков, как это подробно проиллюстрировано в разделе, посвященном работе с программным обеспечением.

10.3.6. Канонический корреляционный анализ

Канонический корреляционный анализ выполняется между двумя совокупностями (группами) выборок и предназначен для определения линейной функции от первых p компонент и линейной функции от остальных q компонент так, чтобы коэффициент корреляции между этими линейными функциями принял наибольшее из возможных значений. Численности групп (количество выборок в первой и второй группах, обозначены как p и q) могут различаться, однако необходимым требованием является равное количество вариант во всех выборках, составляющих обе группы. Матрица взаимной корреляции двух групп выборок имеет вид

$$R = \begin{bmatrix} R_{11} & R_{12} \\ R_{12}^T & R_{22} \end{bmatrix},$$

где R_{11} – матрица взаимной корреляции p переменных 1–й группы, размер p x p,
R_{22} – матрица взаимной корреляции q переменных 2–й группы, размер q x q,
R_{12} – матрица взаимной корреляции переменных 1–й и 2–й группы, размер p x q.
Решение задачи сводится к обобщенной проблеме собственных значений

$$R_{12}^T R_{11}^{-1} R_{12} \cdot v = \lambda \cdot R_{22} \cdot v,$$

где λ – вектор q собственных значений.
Так называемые канонические корреляции представляют собой квадратные корни из собственных значений. Программой выводятся значения критерия χ^2 (массив длиной q) и соответствующие степени свободы (массив длиной q), а также коэффициенты правой (массив размером q x q) и левой (массив размером q x p) стороны.

См. сборник научных программ на Фортране, главу 10 монографии Итона (Eaton).

Список использованной и рекомендуемой литературы

1. Akaike H. Factor analysis and AIC // Psychometrika, 1987, vol. 52, pp. 317–332.
2. Altman D.G. Statistics with confidence // Ed. by D.G. Altman, D. Machin, T.N. Bryant et al. – London: BMJ Publishing Group, 2000.
3. Altman D.G., Gardner M.J. Calculating confidence intervals for regression and correlation // British Medical Journal, 1988, vol. 296, pp. 1238–1242.
4. Artusi R., Verderio P., Marubini E. Bravais–Pearson and Spearman correlation coefficients: meaning, test of hypothesis and confidence interval // The International journal of biological markers, 2002, vol. 17, no. 2, pp. 148–151.
5. Barcikowski R., Stevens J.P. A Monte Carlo study of the stability of canonical correlations, canonical weights, and canonical variate–variable correlations // Multivariate Behavioral Research, 1975, vol. 10, pp. 353–364.
6. Bentler P.M., Bonett D.G. Significance tests and goodness of fit in the analysis of covariance structures // Psychological Bulletin, 1980, vol. 88, pp. 588–606.
7. Bollen K.A. Sample size and Bentler and Bonett’s nonnormed fit index // Psychometrika, 1986, vol. 51, pp. 375 –377.
8. Bollen K.A. Structural equations with latent variables – New York, NY: John Wiley & Sons,

1989.
9. Bonett D.G., Wright T.A. Sample size requirements for estimating Pearson, Kendall and Spearman correlations // Psychometrika, March 2000, vol. 65, no. 1, pp. 23–28.
10. Boomsma A. On the robustness of LISREL (maximum likelihood estimation) against small sample size and nonnormality. – Amsterdam: Sociometric Research Foundation, 1983.
11. Chatfield C., Collins A. Introduction to multivariate analysis. – New York, NY: Chapman & Hall / CRC, 2000.
12. Chernick M.R. Friis R.H. Introductory biostatistics for the health sciences. Modern application including bootstrap. – New York, NY: John Wiley & Sons, 2003.
13. Deshpande J.V., Gore A.P., Shanubhogue A. Statistical analysis of nonnormal data. – New York, NY: John Wiley & Sons, 1995.
14. Duncan O.D., Haller A.O., Portes A. Peer Influences on aspirations: A reinterpretation // American Journal of Sociology, 1968, vol. 74, pp. 119 –137.
15. Eaton M.L. Multivariate statistics: A vector space approach (Lecture notes – Monograph series, vol. 53). – Beachwood, OH: Institute of Mathematical Statistics, 2007.
16. Fieller E.C., Hartley H.O., Pearson E.S. Tests for rank correlation. I // Biometrica, December 1957, vol. 44, no. 3/4, pp. 470–481.
17. Fieller E.C., Pearson E.S. Tests for rank correlation. II // Biometrica, June 1961, vol. 48, no. 1/2, pp. 29–40.
18. Fung W.K. Dimension reduction based on canonical correlation / W.K. Fung, X. He, L. Liu et al. // Statistica Sinica, 2002, vol. 12, no. 4, pp. 1093–1114.
19. Gifi A. Non–linear multivariate analysis. – Chichester, UK: John Wiley & Sons, 1990.
20. Gonzalez I. CCA: An R package to extend canonical correlation analysis / I. Gonzalez, S. Dejean, P.G.P. Martin et al. // Journal of Statistical Software // January 2008, vol. 23, no. 12.
21. Green P.E., Halbert M.H., Robinson P.J. Canonical analysis: An exposition and illustrative application // Journal of Marketing Research, February 1966, vol. 3, pp. 32–39.
22. Grimm L.G. Reading and understanding more multivariate statistics / Ed. by L.G. Grimm, P.R. Yarnold. – Washington, DC: American Psychological Association, 2000.
23. Guidance for data quality assessment. Practical methods for data analysis. EPA QA/G–9. – Washington, DC: United States Environmental Protection Agency, 2000.
24. Guyatt G. Basic statistics for clinicians: 4. Correlation and regression / G. Guyatt, S. Walter, H. Shannon et al. // Canadian Medical Association Journal, February 1995, vol. 152, no. 4, pp. 497–504.
25. Haller A.O., Butterworth C.E. Peer Influences on levels of occupational and educational aspiration // Social Forces, 1960, vol. 38, pp. 289–295.
26. Hampel F.R. Robust Statistics / F.R. Hampel, E.M. Ronchetti, P.J. Rousseeuw et al. – New York: John Wiley & Sons, 1986.
27. Hardle W., Simar L. Applied multivariate statistical analysis. – New York, NY: Springer, 2003.
28. Harris R.J. A primer of multivariate statistics. – Mahwah, NJ: Lawrence Erlbaum Associates, 2001.
29. Hoelter J.W. The analysis of covariance structures: goodness–of–fit indices // Sociological Methods and Research, 1983, vol. 11, pp. 325 –344.
30. Hotelling H. The most predictable criterion // Journal of Educational Psychology, 1935, vol. 26, pp. 139–142.
31. Huber P.J. Robust statistics. – New York, NY: John Wiley & Sons, 1981.
32. Jaeschke R. Basic statistics for clinicians: 3. Assessing the effects of treatment: measures of association / R. Jaeschke, G. Guyatt, H. Shannon et al. // Canadian Medical Association Journal, February 1995, vol. 152, no. 3, pp. 351–357.

33. James L.R., Mulaik S.A., Brett J.M. Causal analysis. – Beverly Hills, CA: SAGE Publications, 1982.
34. Joreskog K.G. A general method for estimating a linear structural equation system // In Structural equation models in the social sciences / Ed. by A.S. Goldberger and O.D. Duncan. – New York: Academic Press, 1973.
35. Keesling J.W. Maximum likelihood approaches to causal analysis. Ph.D. dissertation. – Chicago: University of Chicago, 1972.
36. Le C.T. Introductory biostatistics. – New York, NY: John Wiley & Sons, 2003.
37. Lee S.Y. Analysis of covariance and correlation structures // Computational Statistics and Data Analysis, 1985, vol. 2, pp. 279–295.
38. Levine M.S. Canonical analysis and factor comparison. Quantitative Applications in the Social Sciences Series, no. 6. – Thousand Oaks, CA: SAGE Publications, 1977.
39. Loehlin J.C. Latent variable models. – Hillsdale, NJ: Lawrence Erlbaum Associates, 1987.
40. Long J.S. Regression models for categorical and limited dependent variables. Advanced Quantitative Techniques in the Social Sciences, vol. 7. – Thousand Oaks, CA: SAGE Publications, 1997.
41. Lord F.M. A significance test for the hypothesis that two variables measure the same trait except for errors of measurement // Psychometrika, 1957, vol. 22, pp. 207–220.
42. Lucy D. Introduction to statistics for forensic scientists. – Chichester, UK: John Wiley & Sons, 2005.
43. McArdle J.J., McDonald R.P. Some algebraic properties of the reticular action model // British Journal of Mathematical and Statistical Psychology, 1984, vol. 37, pp. 234–251.
44. McDonald R.P. An index of goodness–of–fit based on noncentrality // Journal of Classification, 1989, vol. 6, pp. 97–103.
45. McNemar Q. Psychological statistics. – New York, NY: John Wiley & Sons, 1966.
46. Motulsky H.J. InStat guide to choosing and interpreting statistical tests. – San Diego, CA: GraphPad Software, 1998.
47. Motulsky H.J. Intuitive biostatistics. – New York: Oxford University Press, 1995.
48. Neter J., Wasserman W., Kutner M.H. Applied linear statistical models: Regression, analysis of variance, and experimental designs. – Homewood, IL: Richard D. Irwin, 1990.
49. Rencher A.C. Methods of multivariate analysis. – New York, NY: John Wiley & Sons, 2002.
50. Samara B., Randles R.H. A test for correlation based on Kendall's tau // Communications in Statistics – Theory and Methods, 1988, vol. 17, pp. 3191–3205.
51. Schwarz G. Estimating the dimension of a model // Annals of Statistics, 1978, vol. 6, pp. 461–464.
52. Siegel S., Castellan Jr. N.J. Nonparametric statistics for the behavioral sciences. – London: McGraw–Hill, 1988.
53. Stevens J. Applied multivariate statistics for the social sciences. – Mahwah, NJ: Lawrence Erlbaum Associates, 2002.
54. Tabachnick B.G., Fidell L.S. Using multivariate statistics. – Boston, MA: Allyn & Bacon, 2000.
55. Thompson B. Canonical correlation analysis: Uses and interpretation. Quantitative Applications in the Social Sciences Series, no. 47. – Thousand Oaks, CA: SAGE Publications, 1984.
56. Wilcox R.R. Fundamentals of modern statistical methods. – New York, NY: Springer, 2001.
57. Williams E.J. Significance of difference between two non–independent correlation coefficients // Biometrics, March 1959, vol. 15, no. 1, pp. 135–136.
58. Wolfe D.A. On testing equality of related correlation coefficients // Biometrica, April 1976, vol. 63, no. 1, pp. 214–215.

59. Айвазян С.А., Енюков И.С., Мешалкин Л.Д. Прикладная статистика: Исследование зависимостей: Справочное издание. – М.: Финансы и статистика, 1985.
60. Афифи А., Эйзен С. Статистический анализ: Подход с использованием ЭВМ. – М.: Мир, 1982.
61. Белов Е.Б. Компьютеризованный статистический анализ для историков. Учебное пособие / Е.Б. Белова, Л.И. Бородкин, И.М. Гарскова и др. – М.: МГУ, 1999.
62. Большев Л.Н., Смирнов Н.В. Таблицы математической статистики. – М.: Наука, 1983.
63. Браунли К.А. Статистическая теория и методология в науке и технике. – М.: Наука, 1977.
64. Ван дер Варден Б.Л. Математическая статистика. – М.: Издательство иностранной литературы, 1960.
65. Вучков И., Бояджиева Л., Солаков Е. Прикладной линейный регрессионный анализ. – М.: Финансы и статистика, 1987.
66. Гаек Я., Шидак З. Теория ранговых критериев. – М.: Наука, 1971.
67. Гайдышев И. Анализ и обработка данных: специальный справочник. – СПб: Питер, 2001.
68. Гласс Дж., Стэнли Дж. Статистические методы в педагогике и психологии. – М.: Прогресс, 1976.
69. Дронов С.В. Многомерный статистический анализ: Учебное пособие. – Барнаул: Издательство Алтайского государственного университета, 2003.
70. Дубров А.М., Мхитарян В.С., Трошин Л.И. Многомерные статистические методы. – М.: Финансы и статистика, 2000.
71. Дюк В. Обработка данных на ПК в примерах. – СПб.: Питер, 1997.
72. Зайцев Г.Н. Математическая статистика в экспериментальной ботанике. – М.: Наука, 1984.
73. Кендэл М. Ранговые корреляции. – М: Статистика, 1975.
74. Кудлаев Э.М., Орлов А.И. Вероятностно–статистические методы исследования в работах А.Н. Колмогорова // Заводская лаборатория. Диагностика материалов, 2003, т. 69, № 5, с. 55–61.
75. Кулаичев А.П. Методы и средства комплексного анализа данных. – М.: ИНФРА–М, 2006.
76. Лакин Г.Ф. Биометрия. – М.: Высшая школа, 1990.
77. Лемешко Б.Ю., Помадин С.С. Корреляционный анализ наблюдений многомерных случайных величин при нарушении предположений о нормальности // Сибирский журнал индустриальной математики, 2002, т. 5, № 3, с. 115–130.
78. Ллойд Э. Справочник по прикладной статистике. В 2–х т. Т. 2. / Под ред. Э. Ллойда, У. Ледермана, С.А. Айвазяна и др. – М.: Финансы и статистика, 1990.
79. Львовский Е.Н. Статистические методы построения эмпирических формул. – М.: Высшая школа, 1988.
80. Малета Ю.С., Тарасов В.В. Непараметрические методы статистического анализа в биологии и медицине. – М.: Издательство Московского университета, 1982.
81. Медик В.А., Токмачев М.С., Фишман Б.Б. Статистика в медицине и биологии: Руководство. В 2–х томах / Под ред. Ю.М. Комарова. Т. 1. Теоретическая статистика. – М.: Медицина, 2000.
82. Мюллер П., Нойман П., Шторм Р. Таблицы по математической статистике. – М.: Финансы и статистика, 1982.
83. Налимов В.В. Применение математической статистики при анализе вещества. – М.: Государственное издательство физико–математической литературы, 1960.
84. Новиков Д.А., Новочадов В.В. Статистические методы в медико–биологическом

эксперименте (типовые случаи). – Волгоград: Издательство ВолГМУ, 2005.
85. Прохоров Ю.В. Вероятность и математическая статистика. Энциклопедия / Гл. ред. Ю.В. Прохоров. – М.: Научное издательство «Большая Российская энциклопедия», 1999.
86. Родионов Д.А. Справочник по математическим методам в геологии / Д.А. Родионов, Р.И.Коган, В.А. Голубева и др. – М.: Недра, 1987.
87. Родионов Д.А. Статистические решения в геологии. – М.: Недра, 1981.
88. Рокицкий П.Ф. Биологическая статистика. – Мн.: Вышэйшая школа, 1973.
89. Сборник научных программ на Фортране. Выпуск 1. Статистика. – М.: Статистика, 1974.
90. Уилкс С. Математическая статистика. – М.: Наука, 1967.
91. Ферстер Э., Ренц Б. Методы корреляционного и регрессионного анализа. Руководство для экономистов. – М.: Финансы и статистика, 1983.
92. Фукс В. По всем правилам искусства. Точные методы в исследованиях литературы, музыки и изобразительного искусства // В кн. Моль А., Фукс В., Касслер М. Искусство и ЭВМ. – М.: Мир, 1975.
93. Холлендер М., Вулф Д. Непараметрические методы статистики. – М.: Финансы и статистика, 1983.
94. Хьюбер П. Робастность в статистике. – М.: Мир, 1984.
95. Юл Дж.Э., Кендэл М.Дж. Теория статистики. – М.: Госстатиздат ЦСУ СССР, 1960.

Глава 11. Факторный анализ

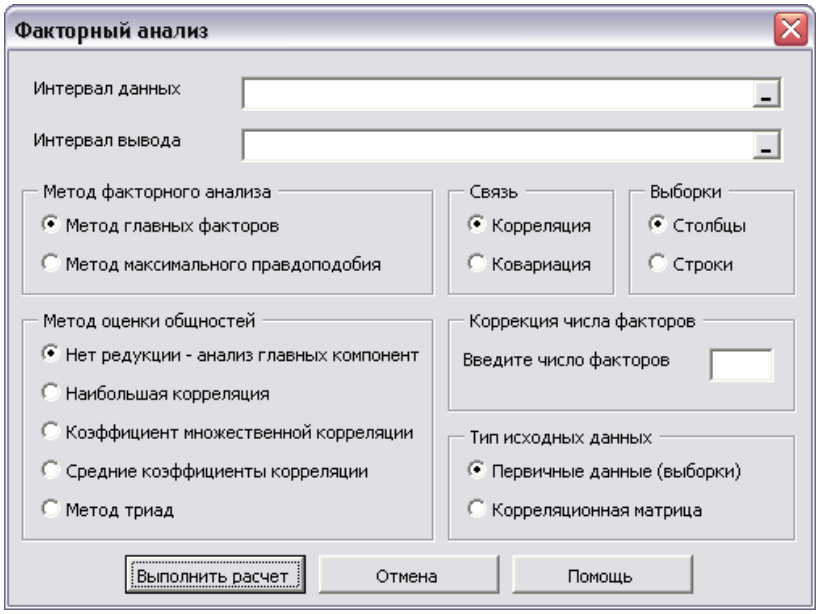
11.1. Введение

В программном обеспечении применяются методы факторного анализа:

- метод главных факторов (если корреляционная матрица не редуцируется – метод главных компонент или компонентный анализ),
- метод максимального правдоподобия.

11.2. Работа с программным обеспечением

Выберите из меню программы пункт AtteStat | Факторный анализ. На экране появится диалоговое окно, изображенное на рисунке:

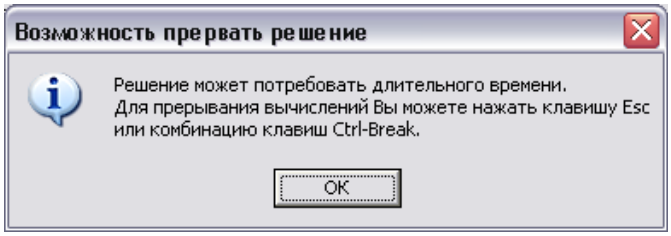


Затем проделайте следующие шаги:

- Выберите или введите интервал матрицы исходных данных (первичных выборок) или корреляционной матрицы (см. п.8 настоящего перечня).
- Выберите или введите выходной интервал. Начиная с первой ячейки выходного интервала (следовательно, можно указать только одну ячейку, т.к. остальные ячейки интервала игнорируются), будут выведены результаты вычислений.
- Выберите или оставьте по умолчанию метод факторного анализа.
- Выберите или оставьте по умолчанию метод оценки общностей.
- Выберите или оставьте по умолчанию тип связи.
- Укажите или оставьте по умолчанию, как расположены выборки.
- Введите желаемое число факторов. Оно не должно превышать число параметров. Параметр может быть опущен.
- Выберите или оставьте по умолчанию тип исходных данных. Программа может работать с исходными данными – первичными выборками или с уже вычисленной корреляционной матрицей. Данная опция позволяет выполнять факторный анализ не только количественных данных, но и данных в любой другой шкале измерения, а также в смешанных шкалах. Нужно только рассчитать заранее корреляционную матрицу с помощью методов, реализованных в главе «Корреляционный анализ».
- Нажмите кнопку «Выполнить расчет».

После выполнения вычислений будет, начиная с первой ячейки выходного интервала, выведено название метода и результаты расчета.

Время решения для больших задач может быть длительным и сильно зависеть от производительности компьютерной системы, поэтому в программу заложена возможность прерывания решения по желанию пользователя до нормального окончания с заданными параметрами. О данной возможности пользователю сообщается в специальном информационном окне, показанном на рисунке, перед любым производством самого решения.



Для начала решения следует нажать кнопку OK.
За выбор адекватного исходным данным метода расчета несет ответственность пользователь. Программное обеспечение берет на себя верификацию исходных данных, выдавая подробную диагностику. При неверных действиях пользователя выдаются сообщения об ошибках. Обратим внимание, что не все сочетания параметров допустимы. Если встречается недопустимое сочетание параметров, программное обеспечение, как правило, само принимает решение о выполнении расчета. Например, если пользователь захочет выполнить редукцию матрицы дисперсий–ковариаций, программное обеспечение пропустит данный этап безо всякой диагностики.

11.2.1. Сообщения об ошибках

При ошибках ввода данных могут выдаваться диагностические сообщения следующих типов:

Ошибка	Комментарий
Не определен интервал переменной.	Вы не выбрали или неверно ввели интервал эмпирической выборки. Лучшим способом избежать ошибки является не ввод, а выделение интервала стандартным образом, т. е. протаскиванием курсора.
Пустая ячейка в области данных.	Проверьте исходные данные и заполните все ячейки, отмеченные Вами как входной интервал. Для избежания ошибок расчета, вызванных разногласиями, трактовать ли пустую ячейку как нуль, программное обеспечение требует заполнения всех ячеек. Если в ячейке не должно быть данных по физической природе исследуемого явления, введите в данную ячейку нуль.
Нечисловой тип данных.	Проверьте типы ячеек входного интервала. Тип может быть только числовым. Проще всего выделить интервал ячеек и явно определить их тип как числовой стандартными средствами.
Не определена область вывода.	Вы не выбрали или неверно ввели выходной интервал. Лучшим способом избежать ошибки является не ввод, а выделение интервала стандартным образом, т. е. протаскиванием курсора.
Корреляционная матрица не квадратная	Если выбраны исходные данные в виде корреляционной матрицы, следует выделить диапазон ячеек, содержащих данную матрицу. Корреляционная матрица должна быть квадратной, т. е. число ее строк должно равняться числу столбцов. Кроме того, корреляционная матрица должна быть симметричной относительно главной диагонали. Проверка этого утверждения не производится программой. Следовательно, ответственность за качество выводов, если это не так, лежит на пользователе программы.

В процессе решения задачи факторного анализа нужно быть готовым к тому, что иногда решение получить не удается. Это вызвано сложностью решаемой проблемы собственных значений корреляционной матрицы. Для матриц высокого порядка может произойти потеря значимости в процессе вычислений. Поэтому теоретически нельзя исключить ситуацию, когда методы факторного анализа, к сожалению, окажутся неприменимы.

11.3. Теоретическое обоснование

Методами факторного анализа решаются три основных вида задач:

- отыскание скрытых, но предполагаемых закономерностей, которые определяются воздействием внутренних или внешних причин (факторов) на изучаемый процесс;
- выявление и изучение статистической связи признаков с факторами или главными компонентами;
- сжатие информации путем описания процесса при помощи общих факторов или главных компонент, число которых меньше количества первоначально взятых признаков (параметров), однако с той или иной степенью точности обеспечивающих воспроизводимость корреляционной матрицы.

Следует пояснить, что в факторном анализе понимается под сжатием информации. Дело в том, что корреляционная матрица получается путем обработки исходного массива данных. Корреляционная матрица образована из попарных коэффициентов корреляции компонент случайного вектора. Предполагается, что та же самая корреляционная матрица может быть получена с использованием тех же объектов, но описанных меньшим числом параметров. Таким образом, якобы происходит уменьшение размерности задачи, хотя на самом деле это не так. Это не сжатие информации в общепринятом смысле – восстановить исходные данные по корреляционной матрице нельзя.

Основное требование к исходным данным для факторного анализа – это то, что они должны подчиняться многомерному нормальному распределению. По крайней мере, должно быть сделано допущение о многомерном нормальном распределении совокупности.

Нормальность распределения может быть проверена с помощью методов, представленных в главе «Проверка нормальности распределения».

Основным объектом исследования методами факторного анализа является корреляционная матрица, построенная с использованием коэффициента корреляции (корреляционного отношения) Пирсона, разработанного для количественных признаков. Напомним, что коэффициентом корреляции называется безразмерная числовая характеристика совместного распределения двух случайных величин, выражающая их взаимосвязь. Чем ближе коэффициент корреляции к 1 или –1, тем сильнее эта взаимосвязь. Чем ближе к 0, тем взаимосвязь слабее. Подробнее см. главу «Корреляционный анализ».

Некоторые авторы предлагают использовать для факторного анализа дисперсионно–ковариационную матрицу, построенную из дисперсий–ковариаций. Дисперсионно–ковариационная (ковариационная) матрица образована из попарных ковариаций компонент случайного вектора. Ковариация случайной величины сама с собой, как известно, является дисперсией. Напомним, что ковариацией называется числовая характеристика совместного распределения двух случайных величин. В отличие от коэффициента корреляции, определяемого через ковариацию, последняя не является безразмерной величиной, поэтому менее удобна для применения в факторном анализе. В дальнейших рассуждениях, за исключением некоторых моментов (например, проблемы общности), говоря о корреляции, будем иметь в виду также и ковариацию. В данном программном обеспечении реализованы оба типа связи: коэффициент корреляции Пирсона и ковариация.

В литературе предлагается также использование других коэффициентов типа корреляции, предназначенных для порядковых, качественных и смешанных признаков. В данной версии

программного обеспечения эта возможность реализована следующим образом.
Программа может работать с исходными данными – первичными выборками или с уже вычисленной корреляционной матрицей. Специальная опция позволяет выполнять факторный анализ не только количественных данных, но и данных в любой другой шкале измерения, а также в смешанных шкалах. Нужно только рассчитать заранее корреляционную матрицу с помощью методов, реализованных в главе «Корреляционный анализ».
Дальнейший анализ – стандартный.
Основным требованием к построенной матрице является ее положительная полуопределенность. Свойства матриц подробно рассмотрены в соответствующих источниках. Из свойства положительной полуопределенности как раз и следует неотрицательность всех собственных значений.
Коэффициенты корреляции, составляющие корреляционную матрицу, по умолчанию вычисляются между параметрами (признаками, тестами), а не между объектами (индивидуумами, лицами), поэтому размерность корреляционной матрицы равна числу параметров. Это так называемая техника *R*. Однако может быть, например, изучена корреляция между объектами (точнее, их состояниями, описываемыми векторами параметров). Эта методика называется техникой *Q*. Проведение факторного анализа техникой *Q* обосновано тем, что состояния объектов могут иметь общую побудительную причину (причины), которая (которые) может быть выявлена с помощью факторного анализа. Существует также техника *P*, предполагающая факторный анализ результатов экспериментальных исследований, выполненных на одном и том же индивидууме в различные промежутки времени («объекты» – один и тот же индивидуум в различные промежутки времени), причем изучаются корреляции между состояниями индивидуума. Аналог техники *Q* для последнего случая составляет предмет исследования техники *O*.
Применение техники *R* (*P*) или техники *Q* (*O*) или выбор техники *R* (*Q*) или *P* (*O*) осуществлены нами с помощью одних и тех же алгоритмов в пределах одной программы путем простого указания, находятся выборки, соответственно, в столбцах или строках, а параметры, рассуждая аналогично, в строках или столбцах.
Получение матрицы факторного отображения в принципе является целью факторного анализа. Ее строки представляют собой координаты концов векторов, соответствующих *m* переменным в *r*–мерном факторном пространстве. Близость концов этих векторов дает представление о взаимной зависимости переменных. Каждый вектор в сжатой, концентрированной форме несет информацию о процессе. Близость этих векторов дает представление о взаимной зависимости переменных. Дополнительно, если число выделенных факторов больше единицы, обычно производится вращение матрицы факторного отображения с целью получения так называемой простой структуры.
Для наглядности результаты можно изобразить графически, что, однако, проблематично для трех и более выделенных факторов. Поэтому обычно дают изображение *r*–мерного факторного пространства в двумерных срезах.
В настоящем программном обеспечении реализованы методы факторного анализа:

- метод главных факторов (если корреляционная матрица не редуцируется – он же метод главных компонент). Иначе метод главных компонент называют просто компонентным анализом.
- метод максимального правдоподобия.

При решении задачи факторного анализа возможна ситуация, когда вектора исходных данных коллинеарны (параметры линейно зависимы). Напомним, что два вектора называются коллинеарными, если они лежат на параллельных прямых или на одной прямой. В таком случае при решении возможно получение различных вычислительных проблем. Корреляционная матрица для таких данных может оказаться вырожденной. Применяемый

для определения собственных значений метод дает решение и в этом случае. При этом часть собственных значений, равная разности порядка матрицы и ее ранга, будет нулевой в вычислительном смысле, что делает метод главных факторов более устойчивым к таким «нехорошим» данным, чем метод максимума правдоподобия. Однако метод главных факторов уступает методу максимума правдоподобия в том, что он не позволяет получить точной оценки общности.
Для выявления мультиколлинеарности специально разработаны эффективные статистические методы (см. главу «Матричная и линейная алгебра»), позволяющие выявить, при ее наличии, коллинеарность векторов исходных данных. После обнаружения таких параметров рекомендуется оставить в исходных данных только один из группы линейно зависимых параметров.

Лучшим руководством по факторному анализу является монография Хармана. Пример применения факторного анализа для исходных данных, измеренных не в количественной шкале, см. в работе Каплана. Теорию см. в статье Уткина с соавт.

11.3.1. Метод главных факторов

Рассмотрим подробнее метод главных компонент (компонентный анализ, principal components analysis), который по определению Лоули с соавт. представляет собой вариант метода главных факторов (когда корреляционная матрица не редуцируется), а затем сам метод главных факторов (principal factor analysis). В методе главных компонент в качестве исходного элемента анализа может быть использована как корреляционная, так и дисперсионно–ковариационная матрица, причем выводы по результатам анализа тождественны.

11.3.1.1. Компонентный анализ

Основная модель метода главных компонент Хотеллинга записывается в матричном виде следующим образом:
$Z = AP$,
где Z – матрица стандартизованных исходных данных, ее размер m x n,
A – матрица факторного отображения, ее размер m x r,
P – матрица значений факторов, ее размер r x n,
m – количество переменных (векторов данных),
n – количество индивидуумов (элементов одного вектора),
r, $r \le m$– количество выделенных факторов.
Как видно из приведенного выше выражения, модель компонентного анализа содержит только общие для имеющихся векторов факторы.
Матрица стандартизованных исходных данных определяется из матрицы исходных данных Y (ее размер m x n) по формуле

$$z_{ij} = \frac{y_{ij} - \bar{y}_{\bar{i}}}{s_i}, i = 1,2,...m, j = 1,2,...,n,$$

где y_{ij} – элемент матрицы исходных данных,
$\bar{y}_{ij}$ – среднее значение,
s_j – стандартное отклонение.
Для вычисления корреляционной матрицы – основного элемента факторного анализа – имеет место простое соотношение:

$$\frac{1}{n-1}ZZ' = R,$$

где R – корреляционная матрица, ее размер m x m,
$'$ – символ транспонирования.
На главной диагонали матрицы R стоят значения, равные 1. Эти значения называются общностями и обозначаются как h_i^2, являясь мерой полной дисперсии переменной. Для метода главных факторов общности отличны от 1 и вычисляются определенным образом.
Неизвестными являются матрицы A и P. Матрица A может быть найдена из основной теоремы факторного анализа
$R = ACA'$,
где C – корреляционная матрица, отражающая связь между факторами.
Если $C = I$, то говорят об ортогональных факторах, если матрица C не равна I, говорят о косоугольных факторах. Здесь I – единичная матрица.
Для матрицы C справедливо соотношение

$$\frac{1}{n-1}PP' = C.$$

Нами рассматривается только случай ортогональных факторов, для которых
$R = AA'$.
Модель классического факторного анализа содержит ряд общих факторов и по одному характерному фактору на каждую переменную. Число главных компонент всегда меньше либо равно числу переменных.

11.3.1.2. Факторный анализ методом главных факторов

По утверждению Хармана, «под методом главных факторов понимают приложение метода главных компонент к редуцированной корреляционной матрице (т. е. к матрице, у которой на главной диагонали вместо единиц стоят значения общностей)». Для метода главных факторов (факторного анализа методом главных факторов Томсона) основная модель записывается в виде
$Z = FP^+$,
где F – полная факторная матрица, ее размер m x (r + m),
P^+ – матрица значений факторов, включая значения характерных факторов, ее размер (r + m) x n.
Матрица F может быть представлена в виде суммы двух матриц
$F = A + U$,
где A – матрица нагрузок общих факторов,
U – матрица нагрузок характерных факторов.
Очевидно, что матрицы A и U имеют размер матрицы F. Матрица A (а именно ее часть размером m x r, остальная же ее часть размером m x m является нулевой) понимается как матрица факторного отображения.
Полная дисперсия переменной складывается из общности h_i^2, значение которой меньше либо равно 1 и означает часть полной дисперсии переменной, приходящейся на главные факторы, и характерности, обозначаемой как u_i^2, приходящейся на характерные факторы.
Следовательно,
$u_i^2 = 1 - h_i^2$.
Часть размером m x r матрицы U является нулевой, остальная ее часть (размером m x m) представляет собой диагональную матрицу с квадратными корнями из характерностей на

главной диагонали, которые уже вычислены из общностей. Таким образом, может быть определена матрица U, а следовательно, и матрица F, если известна матрица A.
Используя введенную выше основную теорему для случая ортогональных факторов, можно записать
$R = FF'$.
Развернув выражение, получим:
$R = R_h + U^2$,
где R – корреляционная матрица с единицами на главной диагонали,
R_h – корреляционная матрица с общностями на главной диагонали, определяемая выражением
$R_h = AA'$,
U^2 означает UU'.

11.3.1.3. Проблема общности

Для метода главных факторов имеет место проблема общности, то есть на главной диагонали корреляционной матрицы, в отличие от метода главных компонент, необходимо проставить значения общностей, чтобы получить корреляционную матрицу R_h.
По определению, общность – сумма квадратов факторных нагрузок. Общность данной переменной – та часть ее дисперсии, которая обусловлена общими факторами. Это вытекает из предположения, что полная дисперсия складывается из общей дисперсии, обусловленной общими для всех переменных факторами, а также специфичной дисперсии, обусловленной факторами, специфичными только для данной переменной, и дисперсии, обусловленной ошибкой. Мы рассматриваем только методы, оперирующие общностями, не превышающими единицу.
Редукцией (редуцированием) корреляционной матрицы в методе главных факторов называется процесс замены единиц на главной диагонали корреляционной матрицы некоторыми величинами, называемыми общностями. Без редукции, то есть с единицами на главной диагонали корреляционной матрицы, мы получаем широко известный компонентный анализ (метод главных компонент).
В программном обеспечении реализованы следующие способы оценки общностей:

1. Способ наибольшей корреляции.
2. Коэффициент множественной корреляции, при этом общности вычисляются с помощью выражения

$$h_i^2 = 1 - \frac{1}{r^{ii}},$$

где в знаменателе стоит диагональный элемент матрицы, обратной к матрице R_h. Этот метод, однако, осложняется тем, что, как показывает опыт расчетов, полученная в результате редукции корреляционная матрица обычно не является матрицей Грама. Замена же диагональных членов оценками общностей считается допустимой, только если сохраняются свойства матрицы Грама. Напомним, что Эрмитова матрица называется положительно полуопределенной (матрицей Грама), если все ее главные миноры неотрицательны (см. главу «Матричная и линейная алгебра»).

3. Средние по столбцу корреляционной матрицы коэффициенты корреляции.
4. Метод триад.

Если редуцированная корреляционная матрица не является матрицей Грама, программой будет выдано предупреждение, не препятствующее дальнейшему выполнению расчета, но служащее предупреждением исследователю.

11.3.1.4. Проблема факторов

Матрица факторного отображения определяется для компонентного анализа или для метода главных факторов методом множителей Лагранжа (максимизация функции, связанная с дополнительными условиями) из решения проблемы собственных значений матрицы R или R_h. Для простоты записи значок h далее опускаем, имея в виду R или R_h, в зависимости от применяемой разновидности метода факторного анализа.
Факторы пропорциональны собственным векторам матрицы R. Стандартная проблема собственных значений матрицы R записывается в виде:
$(R - \lambda_l I) = 0$,
где λ_l, $l = 1,2,...,m$ – l–е собственное значение матрицы R,
l – номер собственного значения.
Результатом расчета будет матрица факторного отображения A размером m x r, m – количество переменных (векторов данных), r, $r \le m$ – количество выделенных факторов.
Данная матрица состоит из элементов (векторов длиной m) a_l, $l = 1,2,...,r$; $r \le m$ – соответствующих l–му собственному значению собственных векторов матрицы R.
Задача упрощается тем, что матрица R является действительной и симметрической, поэтому для решения проблемы собственных значений применимы хорошо разработанные эффективные устойчивые алгоритмы.

11.3.1.5. Измерение факторов

Оценка значений факторов (так называемое измерение факторов) не является необходимой для интерпретации результатов процедурой. Остановимся на ней для полноты изложения.
Способ измерения главных компонент основан на применении основной модели факторного анализа:
$Z = AP$.
Умножив обе части равенства на A', а затем на $(A'A)^{-}$, получим
$P = A^{+}Z$,
где A^{+} – матрица, определяемая по формуле
$A^{+} = (A'A)^{-}A'$.
Способ измерения главных факторов основан множественном регрессионном анализе (см. главу «Распознавание образов с обучением»).

11.3.2. Метод максимума правдоподобия

В факторном анализе может применяться метод максимума правдоподобия (метод максимального правдоподобия Лоули, maximum–likelihood solution). В методе максимума правдоподобия в качестве исходного элемента анализа может быть использована корреляционная, но не дисперсионно–ковариационная матрица, хотя мы предоставили пользователям возможность поэкспериментировать.
Оценка общностей до применения метода не производится – если исследователь отметит данную опцию, этап редуцирования корреляционной матрицы будет проигнорирован.
Общности находятся в результате вычислений из условия полной воспроизводимости, с точностью до ошибки вычислений, редуцированной корреляционной матрицы (не путать с воспроизводимостью матрицы исходных данных!), причем процесс редукции и составляет суть итерационного процесса метода. В этом заключается основное преимущество метода максимума правдоподобия перед методом главных факторов.
Все основные выкладки рассматриваемого метода выполнены Лоули, однако мы дадим основные шаги алгоритма так, как они представлены Харманом:

1. Методом главных компонент вычисляется матрица факторного отображения

(удерживаются заданное пользователем количество главных компонент), которая в схеме алгоритма обозначена $A_{½}$, причем индекс имеет смысл только удобного обозначения матрицы в итерационном процессе.

2. Вычисляется диагональная матрица характерностей $D_i^2 = \mathrm{diag}(I - A_{i-½}A_{i-½}')$, где $i = 1,2,...$ номер итерации, I – единичная матрица.
3. Вычисляется матрица $J_{i-½} = A_{i-½}'D_i^{-2}A_{i-½}$, предварительно матрица характерностей обращается.
4. Вычисляется диагональная матрица $J_i = Q_i'J_{i-½}Q_i$, применяя метод вращения Якоби, где Q_i – матрица вращения.
5. Вычисляется факторная матрица $A_i = A_{i-½}Q_i$.
6. Вычисляется следующее приближение матрицы факторного отображения $A_{i+½} = (RD_i^{-2} - I)A_iJ_i^{-1}$, предварительно матрица J_i обращается.
7. Итерации повторяются, начиная с шага 2, пока не будет выполнено условие $|A_{i+½} - A_{i-½}| < \varepsilon$, где ε – заранее заданное малое положительное число, например, 0,001.

Программой выдаются вычисленные оценки общностей. Если задать число факторов равным числу параметров, то оценки общности будут совпадать с общностями нередуцированной корреляционной матрицы, то есть будут равны единице. За счет итерационного подбора общностей любое заданное пользователем число факторов обеспечит полное выделение дисперсий. Максимальное число удерживаемых факторов можно приблизительно установить из анализа процента дисперсии, выдаваемой программой для каждого фактора.

Основной недостаток рассматриваемого метода – неустойчивость к данным, содержащим совпадающие или линейно зависимые выборки (коллинеарные вектора исходных данных). С точки зрения многомерной статистики (в широком смысле) проблема заключается в том, что в данном случае, даже если исходные данные показывают многомерное нормальное распределение (см. главу «Проверка нормальности распределения»), оно будет вырожденным. С точки зрения факторного анализа (в узком смысле), будет вырожденной матрица характерностей. Вряд ли можно избежать численной неустойчивости в данном случае до того, как будет устранена мультиколлинеарность. Об исследовании мультиколлинеарности рассказано в одноименном разделе главы «Матричная и линейная алгебра».

См. источники: Донг (Dong) и де Лью (de Leeuw).

11.3.3. Проблема вращения

Оси координат, соответствующие выделенным факторам, ортогональны, и их направления устанавливаются последовательно, по максимуму оставшейся дисперсии. Но полученные таким образом координатные оси большей частью содержательно не интерпретируются. Поэтому получают более предпочтительное положение системы координат путем вращения этой системы вокруг ее начала. Пространственная конфигурация векторов в результате применения этой процедуры остается неизменной. Целью вращения является нахождение одной из возможных систем координат для получения так называемой простой факторной структуры. Обычно применяют популярный метод вращения VARIMAX.

Результатом расчета является матрица факторного отображения A размером m x r, m – количество переменных (векторов данных), r, $r \le m$ – количество выделенных факторов. Данная матрица состоит из элементов (векторов длиной m) a_l, $l = 1,2,...,r$; $r \le m$ – соответствующих l–му собственному значению собственных векторов матрицы R. В дальнейших рассуждениях более удобным обозначением элементов матрицы факторного отображения будет поэлементная запись a_{ij}, $i = 1,2,...,m$; $j = 1,2,...,r$.

Метод VARIMAX выполняет ортогональное вращение матрицы факторного отображения таким образом, чтобы удовлетворить выражение (нормальный критерий Кайзера, варимакс–критерий)

$$V = \sum_{j=1}^{r}\left\{ m\sum_{i=1}^{m}\left(\frac{a_{ij}^2}{h_i^2}\right)^2 - \left[\sum_{i=1}^{m}\left(\frac{a_{ij}^2}{h_i^2}\right)\right]^2\right\} \to \max,$$

где a_{ij}, $i = 1,2,...,m$; $j = 1,2,...,r$ – факторные нагрузки,

$h_i^2, i = 1,2,...,m,$ – суммарная факторная нагрузка, вычисляемая по формуле

$$h_i^2 = \sum_{j=1}^{p} a_{ij}^2, i = 1,2,...,m.$$

Результатом работы метода VARIMAX будет так называемая повернутая матрица факторного отображения, отличающаяся от исходной пространственной конфигурации переменных в пространстве выделенных факторов тем, что «гроздья» точек, описываемых матрицей, будут располагаться ближе к осям факторного пространства, сохраняя свое взаимное расположение. Считается, что такое вращение помогает улучшить интерпретируемость решения.

Подробное описание метода VARIMAX приводится в первом выпуске «Сборника научных программ на Фортране» и в монографии Хармана. Метод VARIMAX применяется также в многомерном шкалировании, представленном в одноименной главе, в том смысле, что речь там идет не о пространстве факторов, а о пространстве шкал.

11.3.4. Критерии максимального числа факторов

Существует несколько критериев оценки максимального числа удерживаемых (значимых) факторов. Эффективные критерии, основанные на величине собственных значений корреляционной матрицы, в конечном счете, приводят к анализу процента дисперсии, выделенной факторами. Все общие факторы, число которых равно числу параметров, выделяют 100% дисперсии. Данное утверждение справедливо для всех методов факторного анализа. Если сумма процентов дисперсии превышает величину 100%, то это означает: при вычислении собственных значений корреляционной матрицы были получены отрицательные собственные значения и, как следствие, комплексные собственные вектора, что может означать некорректную редукцию исходной корреляционной матрицы.

11.3.4.1. Адекватность метода главных факторов

Для методов семейства главных факторов максимальное число удерживаемых факторов можно приблизительно установить из анализа процента дисперсии, выдаваемой программой для каждого фактора. Резюмируя сказанное, рекомендуем такой порядок действий:

- сначала пользователь проводит «разведочный» факторный анализ без указания максимального числа факторов,
- затем по величине дисперсий приблизительно оценивает необходимое число факторов,
- задавая число факторов, проводит повторный анализ, используя его результаты как окончательные.

11.3.4.2. Значимость числа факторов метода максимума правдоподобия

Для метода максимума правдоподобия с целью оценки значимости числа выделенных факторов предложен критерий Уилкса, статистика которого вычисляется по формуле:

$$U_m = n \ln \frac{|R|}{|R_h|},$$

где $|R|$ – определитель корреляционной матрицы с единицами на главной диагонали,
$|R_h|$ – определитель корреляционной матрицы с общностями на главной диагонали,
n – количество индивидуумов (элементов одного вектора).
Таким образом, из–за необходимости знания количества индивидуумов, статистика Уилкса вычисляется программой только в том случае, если исходные данные представляют собой первичные выборки, а не заранее вычисленную корреляционную матрицу, без доступа к исходным данным. Подробнее см. главу «Введение».
Распределение статистики Уилкса при больших значениях n стремится к распределению χ^2 с числом степеней свободы, равном
$\nu = ½[(m - r)^2 + m - r]$,
где m – количество переменных (векторов данных),
r, $r \le m$ – количество выделенных (или назначенных пользователем) факторов.

Критерий Уилкса применяется также в методе минимальных остатков, подробно описанном Харманом.

Список использованной и рекомендуемой литературы

1. Akaike H. A new look at the statistical identification model // IEEE Transactions on Automatic Control, 1974, vol. 19, pp. 716 –723.
2. Akaike H. Factor analysis and AIC // Psychometrika, 1987, vol. 52, pp. 317–332.
3. Akaike H. Information theory and the extension of the maximum likelihood principle // Second International Symposium on Information Theory / Ed. by V.N. Petrov, F. Csaki. – Budapest: Akailseoniai–Kiudo, 1973, pp. 267–281.
4. Alwin D.F., Jackson D.J. Applications of simultaneous factor analysis to issues of factorial invariance // Factor analysis and measurement in sociological research / Ed. by D.J. Jackson. – London: Sage, 1979, pp. 249–279.
5. Alwin D.F., Jackson D.J. Measurement models for response errors in surveys: Issues and applications // Sociological methodology / Ed. by K.F. Schuessler. – San Francisco, CA: Jossey–Bass, 1980, pp. 68–119.
6. Amick D.J., Walberg H.J. Introductory multivariate analysis for educational, psychological and social research. – Chicago, IL: University of Illinois at Chicago Press, 1975.
7. Anderson T.W., Rubin H. Statistical inference in factor analysis // Proceedings of the third Berkeley symposium on mathematical statistics and probability, December 1954 and July–August 1955, vol. 5: Contributions to econometrics, industrial research, and psychometry / Ed. by J. Neyman. – Berkeley, CA: University of California Press, 1956, pp. 111–150.
8. Bagozzi R.P., Yi Y. On the evaluation of structural equation models // Journal of the Academy of Marketing Science, 1988, vol. 16, pp. 74–94.
9. Bar–Hen A. Generalized principal component analysis of continuous and discrete variables // InterStat (Statistics on the Internet), September 2002, No. 1.
10. Bartholomew D.J. Latent variable models and factor analysis. – New York, NY: Oxford University Press, 1987.
11. Basilevsky A.T. Statistical factor analysis and related methods: Theory and applications. – New York, NY: John Wiley & Sons, 1994.

12. Bentler P.M. Comparative fit indexes in structural models // Psychological Bulletin, 1990, vol. 107, pp. 238–246.
13. Bentler P.M. EQS structural equations program manual. – Los Angeles, CA: BMDP Statistical Software, 1989.
14. Bentler P.M., Bonett D.G. Significance tests and goodness of fit in the analysis of covariance structures // Psychological Bulletin, 1980, vol. 88, pp. 588–606.
15. Bickel P.J., Doksum K.A. Mathematical Statistics. – San Francisco, CA: Holden–Day, 1977.
16. Biostatistics: A methodology for the health sciences // G. van Belle, L.D. Fisher, P.J. Heagerty et al. – New York, NY: John Wiley & Sons, 2003.
17. Blyth C.R. On Simpson's paradox and the sure–thing principle // Journal of the American Statistical Association, 1972, vol. 67, pp. 364–366.
18. Bock R.D., Gibbons R., Muraki E. Full–information item factor analysis // Applied Psychological Measurement, 1988, vol. 12, pp. 261–280.
19. Bollen K.A. Structural equations with latent variables. – New York, NY: John Wiley & Sons, 1989.
20. Bradburn N.M. The structure of psychological well–being. – Chicago, IL: Aldine, 1969.
21. Browne M.W. Asymptotically distribution–free methods for the analysis of variance structures // British Journal of Mathematical and Statistical Psychology, 1984, vol. 37, pp. 62–83.
22. Bryant F.B. A four–factor model of perceived control: Avoiding, coping, obtaining, and savoring // Journal of Personality, 1989, vol. 57, pp. 773–797.
23. Bryant F.B., Veroff J. Dimensions of subjective mental health in American men and women // Journal of Health and Social Behavior, 1984, vol. 25, pp. 116–135.
24. Bryant F.B., Veroff J. The structure of psychological well–being: A sociohistorical analysis // Journal of Personality and Social Psychology, 1982, vol. 43, pp. 653–673.
25. Bryant F.B., Yarnold P.R. A measurement model for the short form of the Student Jenkins Activity Survey // Journal of Personality Assessment, 1989, vol. 53, pp. 188–191.
26. Bugli C., Lambert P. Comparison between principal component analysis and independent component analysis in electroencephalograms modelling // Biometrical Journal, April 2007, vol. 49, no. 2, pp. 312–327.
27. Campbell A. The sense of well–being in America. – New York, NY: McGraw–Hill, 1980.
28. Cattell R.B. The meaning and strategic use of factor analysis // Handbook of multivariate experimental psychology / Ed. by R.B. Cattell. – Chicago, IL: Rand McNally, 1966, pp. 174–243.
29. Cattell R.B. The scientific use of factor analysis. – New York, NY: Plenum, 1978.
30. Cattell R.B. The scree test for the number of factors // Multivariate Behavioral Research, 1966, vol. 1, pp. 245 –276.
31. Cattell R.B. The three basic factor–analytic research designs–their interrelations and derivatives // Psychological Bulletin, 1952, vol. 49, pp. 499–520.
32. Cattell R.B., Vogelman S. A comprehensive trial of the scree and KG criteria for determining the number of factors // Multivariate Behavioral Research, 1977, vol. 12, pp. 289–325.
33. Cerny B.A., Kaiser H.F. A study of a measure of sampling adequacy for factor–analytic correlation matrices // Multivariate Behavioral Research, 1977, vol. 12, pp. 43–47.
34. Chatfield C., Collins A. Introduction to multivariate analysis. – New York, NY: Chapman & Hall / CRC, 2000.
35. Choi Y.–S. A modified resistant principal factor analysis // Bulletin of the International Statistical Institute, 52nd Session, Tome LVIII, Finland, 1999.
36. Cooley W.W., Lohnes P.R. Multivariate data analysis. – New York, NY: John Wiley & Sons, 1971.

37. Cunningham W.R. Principles for the identification of structural differences // Journal of Gerontology, 1978, vol. 33, pp. 82–86.
38. Cureton E.E. A factor analysis of project TALENT tests and four other test batteries, (Interim Report 4 to the U.S. Office of Education, Cooperative Research Project No. 3051.) Palo Alto: Project TALENT Office, American Institutes for Research and University of Pittsburgh, 1968.
39. Cureton E.E., Mulaik S.A. The weighted Varimax rotation and the Promax rotation // Psychometrika, 1975, vol. 40, pp. 183–195.
40. De Leeuw J. Factor analysis as matrix decomposition // UCLA Statistics Series, 2003, no. 344. – Los Angeles, CA: UCLA, 2003.
41. De Leeuw J. Principal component analysis of binary data by iterated singular value decomposition // Computational Statistics & Data Analysis, 10 January 2006, vol. 50, no. 1, pp. 21–39.
42. Dillion W.R., Goldstein M. Multivariate analysis: Methods and applications. – Berkley: University of California Press, 1984.
43. Dong H.–K. Non–Gramian and singular matrices in maximum likelihood factor analysis // Applied Psychological Measurement, 1985, vol. 9, no. 4, pp. 363–366.
44. Dunteman G.H. Principal components analysis. – Newbury Park, CA: Sage Publications, 1989.
45. Dziuban C.D., Harris C.W. On the extraction of components and the applicability of the factor model // American Educational Research Journal, 1973, vol. 10, pp. 93–99.
46. Engineering and Scientific Subroutine Library for Linux on POWER, version 4, release 2. – International Business Machines Corporation, September 2003.
47. Fruchter B. Introduction to factor analysis. – New York, NY: Van Nostrand, 1954.
48. Fuller W.A. Measurement error models. – New York, NY: John Wiley & Sons, 1987.
49. Geweke J.F., Singleton K.J. Interpreting the likelihood ratio statistic in factor models when sample size is small // Journal of the American Statistical Association, 1980, vol. 75, pp. 133 –137.
50. Glass D.C. Behavior patterns, stress, and coronary disease. – Hillsdale, NJ: Erlbaum, 1977.
51. Gnanadesikan R. Methods for statistical data analysis of multivariate observations. – New York, NY: John Wiley & Sons, 1977.
52. Gorsuch R.L. Factor analysis. – Hillsdale, NJ: Erlbaum, 1983.
53. Green P.E. Analyzing multivariate data. – Hinsdale, IL: Dryden Press, 1978.
54. Guadagnoli E., Velicer W.F. Relation of sample size to the stability of component patterns // Psychological Bulletin, 1988, vol. 103, pp. 265–275.
55. Hair J.F. Multivariate data analysis with readings / J.F. Hair, R.E. Anderson, R.L. Tatham et al. – New York, NY: Macmillan, 1992.
56. Hammarling S. The singular value decomposition in multivariate statistics // ACM SIGNUM Newsletter, 1985, vol. 20, no. 3, pp. 2–25.
57. Hardle W., Simar L. Applied multivariate statistical analysis. – New York, NY: Springer Verlag, 2003.
58. Harman H.H. Modern factor analysis. – Chicago, IL: University of Chicago Press, 1976.
59. Harris C.W. Some Rao–Guttman relationships // Psychometrika, 1962, vol. 27, pp. 247–263.
60. Harris R.J. A primer of multivariate statistics. – Mahwah, NJ: Lawrence Erlbaum Associates, 2001.
61. Higuchi I. Robust principal component analysis with adaptive selection for tuning parameters // Journal of Machine Research, 2004, vol. 5, pp. 453–471.
62. Hoelter J.W. The analysis of covariance structures: Goodness of fit indices // Sociological Methods and Research, 1983, vol. 11, pp. 325–344.

63. Horn J.L., Engstrom R. Cattell's scree test in relation to Bartlett's chi–square test and other observations on the number of factors problem // Multivariate Behavioral Research, 1979, vol. 14, pp. 283–300.
64. Horst P. Factor analysis of data matrices. – New York, NY: Holt, 1965.
65. Hotelling H. Analysis of a complex of statistical variables into principal components // Journal of Educational Psychology, 1933, vol. 24, pp. 417–441, 498–520.
66. Hu L., Bentler P.M., Kano Y. Can test statistics in covariance structure analysis be trusted? // Psychological Bulletin, 1992, vol. 112, pp. 351–362.
67. Jackson J.E. A user's guide to principal components. – New York, NY: John Wiley & Sons, 1991.
68. Jahn W., Vahle H. Die Faktorenanalyse und ihre Anwendung. – Berlin: Die Wirtschaft, 1970.
69. Johnson R.A., Wichern D.W. Applied multivariate statistical analysis. – Upper Saddle River, NJ: Prentice–Hall, 1999.
70. Jolliffe I.T. Principal component analysis. – New York, NY: Springer–Verlag, 1986.
71. Joreskog K.G. Factor analysis by least–squares and maximum likelihood methods // Statistical methods for digital computers / Ed. by K. Enslein, A. Ralston, H.S. Wilf. – New York, NY: John Wiley & Sons, 1977.
72. Joreskog K.G. On the statistical treatment of residuals in factor analysis // Psychometrika, 1962, vol. 27, pp. 335–354.
73. Joreskog K.G. Simultaneous factor analysis in several populations // Psychometrika, 1971, vol. 36, pp. 409–426.
74. Joreskog K.G. Statistical analysis of sets of congeneric tests // Psychometrika, 1971, vol. 36, pp. 109–133.
75. Joreskog K.G. Structural analysis of covariance and correlation matrices // Psychometrika, 1978, vol. 43, pp. 443–477.
76. Joreskog K.G., Sorbom, D. G. Advances in factor analysis and structural equation models. – Cambridge, MA: Abt Books, 1979.
77. Kahaner D.K., Moler C., Nash S.G. Numerical methods and software. – Englewood Cliffs, NJ: Prentice Hall, 1989.
78. Kaiser H.F. A second generation Little Jiffy // Psychometrika, 1970, vol. 35, pp. 401–415.
79. Kaiser H.F. Image analysis // Problems in Measuring Change / Ed. by C.W. Harris, Madison, WI: University of Wisconsin Press, 1963.
80. Kaiser H.F. The application of electronic computers to factor analysis // Educational and Psychological Measurement, 1960, vol. 20, pp. 141–151.
81. Kaiser H.F., Cerny B.A. Factor analysis of the image correlation matrix // Educational and Psychological Measurement, 1979, vol. 39, pp. 711–714.
82. Kaiser H.F., Rice J. Little Jiffy, Mark IV // Educational and Psychological Measurement, 1974, vol. 34, pp. 111–117.
83. Kaplan E.H., Small C.A. Anti–Israel sentiment predicts anti–semitism in Europe // Journal of Conflict Resolution, 2006, vol. 50, no. 4, pp. 548–561.
84. Kerlinger F.N., Pedhazur E.J. Multiple regression in behavioral research. – New York, NY: Holt, Rinehart & Winston, 1973.
85. Kim J.O., Mueller C.W. Factor analysis: statistical methods and practical issues. Sage University Paper Series on Quantitative Applications in the Social Sciences, series no. 07–014 – Beverly Hills, CA: Sage Publications, 1978.
86. Kim J.O., Mueller C.W. Introduction to factor analysis: What it is and how to do it. Sage University Paper Series on Quantitative Applications in the Social Sciences, series no. 07–013. – Beverly Hills, CA: Sage Publications, 1978.
87. Kleinbaum D.G., Kupper L.L., Muller K.E. Applied regression analysis and other

multivariable methods. – Boston: PWS–Kent, 1988.
88. Knol D.L., Berger M.P. Empirical comparison between factor analysis and multidimensional item response models // Multivariate Behavioral Research, 1991, vol. 26, 457–477.
89. Kshirsagar A.M. Multivariate analysis. – New York, NY: Marcel Dekker, 1972.
90. La Du T.J., Tanaka J.S. Influence of sample size, estimation method, and model specification on goodness–of–fit assessment in structural equation models // Journal of Applied Psychology, 1989, vol. 74, pp. 625–635.
91. Lawley D.N., Maxwell A.E. Factor analysis as a statistical method. – New York, NY: Elsevier Science, 1971.
92. Lee H.B., Comrey A.L. Distortions in a commonly used factor analytic procedure // Multivariate Behavioral Research, 1979, vol. 14, pp. 301–321.
93. Maiti S.S., Mukherjee B.N. Two new goodness–of–fit indices for covariance matrices with linear structures // British Journal of Mathematical and Statistical Psychology, 1991, vol. 44, pp. 153–180.
94. Malinowski E.R. Factor analysis in chemistry // Technometrics, 2003, vol. 45, no. 2, pp. 180–181.
95. Mardia K.V., Kent J.T., Bibby J.M. Multivariate analysis. – London: Academic Press, 1979.
96. Maxwell A.E. Statistical methods on factor analysis // Psychological Bulletin, 1959, vol. 56, pp. 228–235.
97. McDonald R.P. A note on Rippe's test of significance in common factor analysis // Psychometrika, 1975, vol. 40, pp. 117–119.
98. McDonald R.P. Factor analysis and related methods. – Hillsdale, NJ: Lawrence Erlbaum Associates, 1985.
99. McKennell A.C., Andrews F.M. Measures of cognition and affect in perceptions of well–being // Social Indicators Research, 1980, vol. 8, pp. 257–298.
100. Morrison D.F. Multivariate statistical methods. – New York, NY: McGraw–Hill, 1976.
101. Mulaik S.A. Evaluation of goodness–of–fit indices for structural equation models / S.A. Mulaik, L.R. James, J. Van Alstine et al. // Psychological Bulletin, 1989, vol. 105, pp. 430–445.
102. Mulaik S.A. The foundations of factor analysis. – New York, NY: McGraw–Hill, 1972.
103. Muthen B.O. Contributions to factor analysis of dichotomized variables // Psychometrika, 1978, vol. 43, pp. 551–560.
104. Muthen B.O. Goodness of fit with categorical and other nonnormal variables // Testing structural equation models / Ed. by K.A. Bollen, J.S. Long. – London: Sage, 1993, pp.205–234.
105. Papanastasiou E.C. Factor structure of the «attitudes toward research» scale // Statistics Education Research Journal, May 2005, vol. 4, no. 1, pp. 16–26.
106. Parry C.D., McArdle J.J. An applied comparison of methods for least–squares factor analysis of dichotomous variables // Applied Psychological Measurement, 1991, vol. 15, pp. 35–46.
107. Pearson K. On lines and planes of closest fit to systems of points in space // Philosophical Magazine, 1901, vol. 6, no. 2, pp. 559–572.
108. Rao C.R. Estimation and tests of significance in factor analysis // Psychometrika, 1955, vol. 20, pp. 93–111.
109. Rao C.R. The use and interpretation of principal component analysis in applied research // Sankhya A, 1964, vol. 26, pp. 329–358.
110. Rencher A.C. Methods of multivariate analysis. – New York, NY: John Wiley & Sons,

2002.
111. Rousson V., Gasser T. Simple component analysis // Journal of the Royal Statistical Society: Series C (Applied Statistics), 2004, vol. 53, part 4, pp. 539–555.
112. Schwarz G. Estimating the dimension of a model // Annals of Statistics, 1978, vol. 6, pp. 461–464.
113. Smith L.I. A tutorial on principal components analysis: Student tutorial. – Dunedin, New Zealand: University of Otago, 2002.
114. Spearman C. General Intelligence objectively determined and measured // American Journal of Psychology, 1904, vol. 15, pp. 201–293.
115. Steiger J.H. Factor indeterminacy in the 1930's and the 1970's: Some interesting parallels // Psychometrika, 1979, vol. 44, pp. 157–167.
116. Stevens J.P. Applied multivariate statistics for the social sciences. – Hillsdale, NJ: Erlbaum, 1986.
117. Stewart D.W. The application and misapplication of factor analysis in marketing research // Journal of Marketing Research, February 1981, vol. 18, no. 1, pp. 51–62.
118. Tabachnick B.G., Fidell L.S. Using multivariate statistics. – New York, NY: Harper & Row, 1983.
119. Takane Y., de Leeuw J. On the relationship between item response theory and factor analysis of discretized variables // Psychometrika, 1987, vol. 52, pp. 393–408.
120. Tanaka J.S. Multifaceted conception of fit in structural equation models // Testing structural equation models / Ed. by K.A. Bollen, J.S. Long. – London: Sage, 1993, pp. 10–39.
121. Thurstone L.L. Multiple–factor analysis: A development and expansion of the vectors of mind. – Chicago, IL: University of Chicago Press, 1947.
122. Tucker L.R., Lewis C. A reliability coefficient for maximum likelihood factor analysis // Psychometrika, 1973, vol. 38, pp. 1–10.
123. Weinfurt K.P., Bryant F.B., Yarnold P.R. The factor structure of the affect intensity measure: In search of a measurement model // Journal of Research in Personality, 1994, vol. 28, pp. 314–331.
124. Yarnold P.R. Note on the multidisciplinary scope of psychological androgyny theory // Psychological Reports, 1984, vol. 55, pp. 936–938.
125. Yeung K.Y., Ruzzo W.L. Principal component analysis for clustering gene expression data // Bioinformatics, 2001, vol. 17, no. 9, pp. 763–774.
126. Белова Е.Б. Компьютеризованный статистический анализ для историков. Учебное пособие / Е.Б. Белова, Л.И. Бородкин, И.М. Гарскова и др. – М.: МГУ, 1999.
127. Бессокирная Г.П. Факторный анализ: традиции использования и новые возможности // Социология: методология, методы, математические модели (Социология: 4М), 2000, № 12, с. 142–153.
128. Благуш П. Факторный анализ с обобщениями. – М.: Финансы и статистика, 1988.
129. Блюмин С.Л., Суханов В.Ф., Чеботарев С.В. Экономический факторный анализ. – Липецк: ЛЭГИ, 2004.
130. Большаков А.А., Каримов Р.Н. Методы обработки многомерных данных и временных рядов: Учебное пособие для вузов. – М.: Горячая линия – Телеком, 2007.
131. Вайшля О.Б. Факторный анализ показателей фотосинтеза, дыхания и продуктивности у гетерозисных гибридов и родительских линий PISUM SATIVUM L. // Электронный журнал «Исследовано в России», 2004, том. 7, с. 144–163.
132. Ватанабе С. Разложение Карунена–Лоэва и факторный анализ. Теория и приложения. // В сб. «Автоматический анализ сложных изображений» – М.: Мир,

1969, с.239–253.
133. Гайдышев И. Анализ и обработка данных: специальный справочник. – СПб: Питер, 2001.
134. Дронов С.В. Многомерный статистический анализ: Учебное пособие. – Барнаул: Издательство Алтайского государственного университета, 2003.
135. Дубров А.М. Обработка статистических данных методом главных компонент. – М.: Финансы и статистика, 1978.
136. Дубров А.М., Мхитарян В.С., Трошин Л.И. Многомерные статистические методы. – М.: Финансы и статистика, 2000.
137. Дюк В. Обработка данных на ПК в примерах. – СПб.: Питер, 1997.
138. Жуковская В.М., Мучник И.Б. Факторный анализ в социально–экономических исследованиях. – М.: Статистика, 1976.
139. Иберла К. Факторный анализ. – М.: Статистика, 1980.
140. Калинкина Д. Проблема подавления шума на изображениях и видео и различные подходы к ее решению // On–line журнал «Графика и мультимедиа», 2005, вып. 9.
141. Каханер Д., Моулер К., Нэш С. Численные методы и программное обеспечение. – М.: Мир, 2001.
142. Ким Дж.О., Мьюллер Ч.У. Факторный анализ: статистические методы и практические вопросы // Факторный, дискриминантный и кластерный анализ / Дж.О. Ким, Ч.У. Мюллер, У.Р. Клекка и др. – М.: Финансы и статистика, 1989.
143. Клишина Ю.Н. Применение анализа соответствий в обработке нечисловой информации // Социология: методология, методы, математические модели (Социология: 4М), 1991, № 2, с. 105–118.
144. Кулаичев А.П. Методы и средства комплексного анализа данных. – М.: ИНФРА–М, 2006.
145. Кэттел Р.Б., Ханна Д.К. Принципы и процедуры однозначного поворота в факторном анализе // Статистические методы для ЭВМ / Под ред. К. Энслейна, Э. Рэлстона, Г.С. Уилфа. – М.: Наука, 1986, с. 184–218.
146. Леонов В.П. Факторный анализ: Основные положения и ошибки применения // Международный журнал медицинской практики, 2005, № 3, с. 14–16.
147. Лиепа И.Я. Математические методы в биологических исследованиях: факторный и компонентный анализы: Учебное пособие для студентов биологического факультета. – Рига: Латвийский государственный университет им. П. Стучки, 1980.
148. Лоули Д., Максвелл А. Факторный анализ как статистический метод. – М.: Мир, 1967.
149. Максимов Г.К., Синицын А.Н. Статистическое моделирование многомерных систем в медицине. – М.: Медицина, 1983.
150. Молчанов И.Н. Машинные методы решения прикладных задач. Алгебра, приближение функций. – Киев: Наукова думка, 1987.
151. Окунь Я. Факторный анализ. – М.: Статистика, 1974.
152. Осипов Г.В. Методы измерения в социологии. – М.: Наука, 2003.
153. Пен Р.З. Статистические методы моделирования и оптимизации процессов целлюлозно–бумажного производства: Учебное пособие. – Красноярск: Издательство КГУ, 1982.
154. Петренко В., Сергеева М. Фактор рекламы в представлении курильщиком рынка табачных изделий // Лаборатория рекламы, маркетинга и public relations, 2003, № 1 (20), с. 30–33.
155. Поттосина С.А. Экономико–математические модели и методы. – Мн.: БГУИР,

2003.
156. Прохоров Ю.В. Вероятность и математическая статистика. Энциклопедия / Гл. ред. Ю.В. Прохоров. – М.: Научное издательство «Большая Российская энциклопедия», 1999.
157. Сборник научных программ на Фортране. Выпуск 1. Статистика. – М.: Статистика, 1974.
158. Сборник научных программ на Фортране. Выпуск 2. Матричная алгебра и линейная алгебра. – М.: Статистика, 1974.
159. Тронева Н.В., Тронева М.А. Электронно–зондовый микроанализ неоднородных поверхностей (в свете теории распознавания образов). – М.: Металлургия, 1996.
160. Уилкинсон Дж.Х., Алгебраическая проблема собственных значений. – М.: Наука, 1970.
161. Уилкинсон, Райнш. Справочник алгоритмов на языке АЛГОЛ. Линейная алгебра. – М.: Машиностроение, 1976.
162. Уткин В.А., Гайдышев И.П., Кобазева О.М. О возможном единообразии приложений и условий реализации факторного анализа // Наука и образование Зауралья, 2001, №1, с. 33–38.
163. Харман Г. Современный факторный анализ. – М.: Статистика, 1972.
164. Харман Г.Дж. Метод минимальных остатков в факторном анализе // Статистические методы для ЭВМ / Под ред. К. Энслейна, Э. Рэлстона, Г.С. Уилфа. – М.: Наука, 1986, с. 169–183.

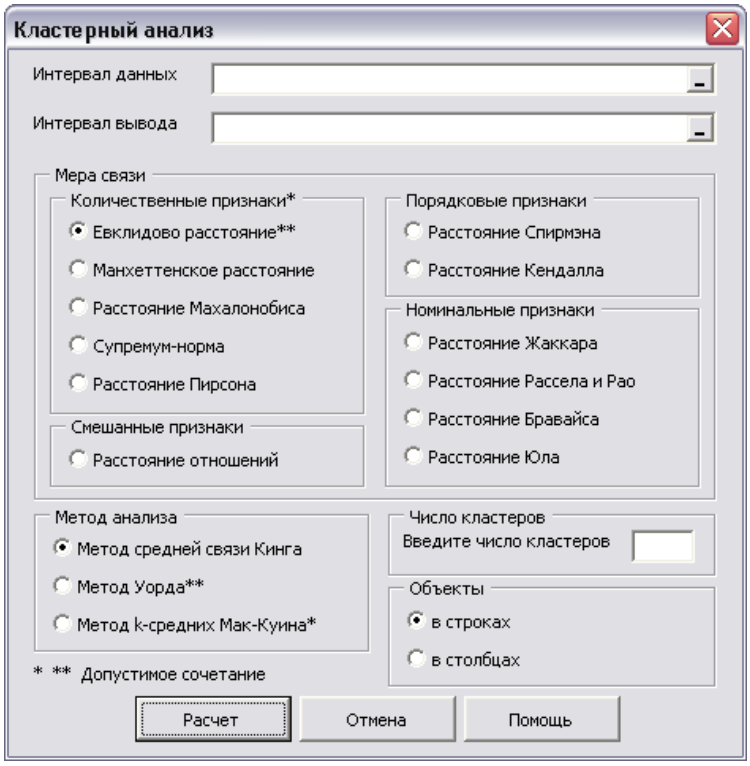
Глава 12. Кластерный анализ

12.1. Введение

В программном обеспечении применяются методы кластерного анализа, относящиеся к категории методов обучения без учителя (автоматической классификации).

12.2. Работа с программным обеспечением

Выберите из меню программы пункт AtteStat | Кластерный анализ. На экране появится диалоговое окно, изображенное на рисунке:



Затем проделайте следующие шаги:

- Выберите или введите интервалы матрицы исходных данных.
- Выберите или введите выходной интервал. Начиная с первой ячейки выходного интервала (следовательно, можно указать только одну ячейку, т.к. остальные ячейки интервала игнорируются), будут выведены результаты вычислений.
- Выберите или оставьте по умолчанию меру различия (кроме метода Уорда).
- Выберите или оставьте по умолчанию метод анализа.
- Введите число кластеров.
- Укажите или оставьте по умолчанию, как расположены классифицируемые объекты. Данная опция, кроме удобства по классификации объектов, расположенных в строках либо в столбцах, позволяет также выбрать, по желанию пользователя, что классифицировать – объекты или параметры.
- Нажмите кнопку «Выполнить расчет».

После выполнения вычислений будет, начиная с первой ячейки выходного интервала, выведено название метода и результаты расчета. Интерпретация полученных результатов расчетов подробно рассмотрена ниже.
За выбор адекватного исходным данным метода расчета несет ответственность пользователь. Программное обеспечение берет на себя верификацию исходных данных, выдавая подробную диагностику. При неверных действиях пользователя выдаются сообщения об ошибках.

12.2.1. Сообщения об ошибках

При ошибках ввода данных могут выдаваться диагностические сообщения следующих типов:

Ошибка	Комментарий
Не определен интервал переменной.	Вы не выбрали или неверно ввели входной интервал. Лучшим способом избежать ошибки является не ввод, а выделение интервала стандартным образом, т. е. протаскиванием курсора.
Пустая ячейка в области данных.	Проверьте исходные данные и заполните все ячейки, отмеченные Вами как входной интервал. Для избежания ошибок расчета, вызванных разногласиями, трактовать ли пустую ячейку как нуль, программное обеспечение требует заполнения всех ячеек. Если в ячейке не должно быть данных по физической природе исследуемого процесса, введите в данную ячейку нуль.
Нечисловой тип данных.	Проверьте типы ячеек входного интервала. Тип может быть только числовым. Проще всего выделить интервал ячеек и явно определить их тип как числовой стандартными средствами.
Не определена область вывода.	Вы не выбрали или неверно ввели выходной интервал. Лучшим способом избежать ошибки является не ввод, а выделение интервала стандартным образом, т. е. протаскиванием курсора.
Не задано число кластеров.	Не задано число кластеров для метода k–средних Мак–Куина. Введите число кластеров.

При решении могут также возникнуть:

- ошибки типа нехватки памяти для вычислений.
- ошибки в работе алгоритма.

Эти ошибки не позволят выполнить требуемое решение для заданных наборов исходных данных, причем причина кроется именно в количестве и структуре исходных данных. Наибольшее неудобство для пользователя представляет вторая из названных ошибок, т.к. она сигнализирует о принципиальной невозможности применения данного метода расчета к заданному набору данных. Например, при использовании расстояния Махалонобиса, в соответствии с алгоритмом, вычисляется матрица, обратная к дисперсионно–ковариационной, из чего вытекает требование невырожденности данной матрицы, которое иногда не соблюдается. Более полную информацию см. в главе «Матричная и линейная алгебра».

12.3. Теоретическое обоснование

Методами кластерного анализа решается задача разбиения (классификации, кластеризации) множества объектов таким образом, чтобы все объекты, принадлежащие одному кластеру (классу, группе) были более похожи друг на друга, чем на объекты других кластеров. В отечественной литературе синонимом термина «кластерный анализ» является термин «таксономия». В иностранной литературе под таксономией традиционно понимается классификация видов животных и растений.
Нами рассматриваются следующие методы кластерного анализа:

- Иерархические методы:
  - метод средней связи Кинга,
  - метод Уорда.
- Итеративные методы группировки:
  - метод k–средних Мак–Куина.

Классифицируемыми могут быть как параметры, так и объекты, поэтому по ходу изложения

там, где идет речь о классификации объектов, вполне можно говорить о классификации параметров, и наоборот. В данном программном обеспечении такая возможность предусмотрена.
Меры различия накладывают жесткие ограничения на применяемые методы кластерного анализа:

- метод средней связи Кинга можно применять для признаков любых типов: количественных, порядковых, номинальных (как частный случай – экспертных ранжировок) и смешанных признаков
- метод Уорда можно применять только для количественных признаков, т.к. в его схеме применяется только евклидово расстояние
- метод k–средних Мак–Куина можно применять только для количественных признаков. Для использования метода с целью классификации данных в шкалах, отличной от количественной, требуется модификация метода.

Применяя формально метод, не соответствующий типу данных, пользователь рискует получить результаты, лишенные смысла.

12.3.1. Меры различия

Виды используемых в кластерном анализе мер сходства и различия перекликаются с философской дилеммой: «ищите сходство» или «ищите различие». Меры для кластерного анализа могут быть следующих видов:

- Мера сходства типа расстояния (функции расстояния), называемая также мерой различия. В этом случае объекты считаются тем более похожими, чем меньше расстояние между ними, поэтому некоторые авторы называют меры сходства типа расстояния мерами различия. При определенных условиях данная мера будет метрикой.
- Мера сходства типа корреляции, называемая связью, является мерой, определяющей похожесть объектов. В этом случае объекты считаются тем более похожими, чем больше связь между ними. Данные меры с помощью элементарных преобразований могут быть сведены к мерам сходства типа расстояния, что и сделано в данном программном обеспечении с целью единообразия.

Применяются меры различия, в зависимости от принадлежности параметров, описывающих объекты в различных шкалах измерения:

1. для количественных признаков:
    - евклидово расстояние
    - манхеттенское расстояние,
    - супремум–норма,
    - расстояние Махаланобиса,
    - расстояние Пирсона,
2. для порядковых признаков:
    - расстояние Спирмэна,
    - расстояние Кендалла,
3. для номинальных признаков:
    - расстояние Жаккара,
    - расстояние Рассела–Рао,
    - расстояние Бравайса,
    - расстояние Юла,
4. для смешанных и произвольных данных:
    - расстояние отношений.

Рассмотренными мерами могут оперировать различные методы кластерного анализа.

Следует, однако, понимать, что меры различия накладывают жесткие ограничения на применяемые методы кластерного анализа:

- Для признаков любых типов: количественных, порядковых, номинальных (как частный случай – экспертных ранжировок) и смешанных можно применять метод средней связи Кинга.
- Только для количественных признаков можно применять метод k–средних Мак–Куина. Для использования метода с целью классификации данных в шкалах, отличной от количественной, требуется модификация метода.
- Только для количественных признаков можно применять метод Уорда, т.к. в его схеме применяется только евклидово расстояние.

Применяя формально метод, не соответствующий типу данных, пользователь рискует получить результаты, лишенные смысла.
Мера сходства типа расстояния называется метрикой, если она удовлетворяет определенным условиям:

- симметрии,
- неравенству треугольника,
- различимости нетождественных объектов,
- неразличимости тождественных объектов.

Термин «метрика» следует использовать с учетом данных условий.

12.3.1.1. Евклидово расстояние

Наиболее общей мерой является метрика Минковского

$$d_{ij} = \sqrt[r]{\sum_{k=1}^{n} \left| x_{ik} - x_{jk} \right|^r},$$

где x_{ij}, x_{jk}, $k = 1,2,...,n$ – выборочные совокупности i и j, соответственно,
n – численность каждой выборки.
Если в метрике Минковского положить $r = 2$, мы получим стандартное евклидово расстояние (евклидову метрику)

$$d_{ij} = \sqrt{\sum_{k=1}^{n} \left(x_{ik} - x_{jk} \right)^2}.$$

12.3.1.2. Манхеттенское расстояние

При $r = 1$ метрика Минковского дает манхеттенское расстояние (метрику города, city block, Manhattan distance)

$$d_{ij} = \sum_{k=1}^{n} \left| x_{ik} - x_{jk} \right|,$$

где x_{ij}, x_{jk}, $k = 1,2,...,n$ – выборочные совокупности i и j, соответственно,
n – численность каждой выборки.

12.3.1.3. Супремум–норма

При $r \to \infty$ метрика Минковского дает метрику доминирования

$$d_{ij} = \max_k \left| x_{ik} - x_{jk} \right|, k = 1,2,...,n,$$

где x_{ij}, x_{jk}, $k = 1,2,...,n$ – выборочные совокупности i и j, соответственно,
n – численность каждой выборки.
что совпадает с супремум–нормой (∞–нормой)

$$d_{ij} = \sup\left\{\left|x_{ik} - x_{jk}\right|\right\}, k = 1,2,...,n.$$

12.3.1.4. Расстояние Махаланобиса

Для различных ковариационных матриц в случае произвольного распределения дивергенция, оценивающая расхождение между статистическими распределениями *i* и *j*, выражается формулой

$$J_{ij} = \int_x \left[p_i(x) - p_j(x)\right] \ln \frac{p_i(x)}{p_j(x)} dx.$$

Иначе дивергенция называется полной средней информационной мерой различия двух классов или, более коротко, средней различающей информацией.
Практически дивергенция может быть вычислена по формуле

$$J_{ij} = \frac{1}{2} tr\left[\left(C_i - C_j\right)\left(C_j^{-1} - C_i^{-1}\right)\right] + \frac{1}{2} tr\left[\left(C_i^{-1} + C_j^{-1}\right)\left(m_i - m_j\right)\left(m_i - m_j\right)'\right],$$

где C_i, C_j – дисперсионно–ковариационные матрицы совокупностей *i* и *j*,
m_i, m_j – вектора средних совокупностей *i* и *j*.
Мера Махаланобиса (расстояние Махалонобиса, обобщенное Евклидово расстояние, обобщенное расстояние) является дивергенцией в предположении, что ковариационные матрицы классов равны

$$C_i = C_j = \Sigma,$$

а многомерная совокупность подчиняется многомерному нормальному распределению. Соответствие совокупности многомерному нормальному распределению может быть протестировано с помощью методов главы «Проверка нормальности распределения», а равенство ковариационных матриц – с помощью методов главы «Дисперсионный анализ». Указанные условия накладывают ограничения на применение рассматриваемой меры различия, что часто ошибочно не принимается во внимание исследователями, но несомненно, должно учитываться в практических расчетах. После должных преобразований получим, что мера Махаланобиса вычисляется как

$$d_{ij} = \left(X_i - X_j\right)' \Sigma^{-1} \left(X_i - X_j\right),$$

где Σ – общая внутригрупповая дисперсионно–ковариационная матрица.

12.3.1.5. Расстояние Пирсона

Коэффициент корреляционного отношения Пирсона (коэффициент корреляции, выборочный коэффициент корреляции, коэффициент корреляции Бравайса–Пирсона) измеряет силу линейной корреляционной связи количественных признаков. Коэффициент корреляции вычисляется по формуле

$$r = \frac{\sum_{i=1}^{n} \left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sqrt{\sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2} \sqrt{\sum_{i=1}^{n} \left(y_i - \bar{y}\right)^2}},$$

где где x_i, y_i, i = 1,2,...,n – выборочные совокупности *i* и *j*, соответственно,
$\bar{x}, \bar{y}$ – соответствующие выборочные средние значения,
n – численность каждой выборки.
Использование коэффициента корреляции в качестве меры связи оправдано лишь тогда,

когда совместное распределение пары признаков нормально или приближенно нормально. Расстояние Пирсона как мера сходства может быть получено из рассмотренного коэффициента путем отнятия от единицы.

12.3.1.6. Расстояние Спирмэна

Показатель ранговой корреляции Спирмэна (показатель корреляции рангов Спирмэна, коэффициент корреляции рангов, коэффициент корреляции Спирмэна, коэффициент ранговой корреляции ρ, Spearman rank correlation) применяется в случае, если изучается линейная связь между рядами, представленными в количественной или порядковой шкале. Практически при анализе количественных признаков применять показатель Спирмэна вместо коэффициента корреляционного отношения Пирсона не следует, так как при его вычислении происходит понижение количественной шкалы до порядковой. Поэтому наиболее широкое применение показатель Спирмэна нашел при анализе корреляции порядковых признаков. Расчет ведется по формуле

$$\hat{\rho}_S = 1 - \frac{6(S_\rho + B_x + B_y)}{n^3 - n}; \quad S_\rho = \sum_{i=1}^{n} (r_i - s_i)^2,$$

где r_i, s_i, $i = 1,2,...,n$ – массивы рангов выборочных совокупностей,
n – численность каждой выборки.
B_x, B_y – поправки на объединение рангов в соответствующих совокупностях, вычисляемые по формуле

$$B_{\cdot} = \frac{1}{12}\sum_{i=1}^{m} n_i(n_i^2 - 1),$$

где m – число групп объединенных рангов,
n_i, $i = 1,2,...,m$ – число рангов в i–ой группе.
Расстояние Спирмэна как мера сходства может быть получено из рассмотренного коэффициента путем отнятия от единицы.

12.3.1.7. Расстояние Кендалла

Коэффициент ранговой корреляции Кендалла (коэффициент корреляции рангов, ранговый коэффициент корреляции, коэффициент корреляции Кендэла, Kendall rank correlation) предназначен для вычисления силы корреляционной связи между двумя рядами при тех же условиях, что и рассмотренный выше показатель Спирмэна. Коэффициент Кендалла считается более строгой оценкой по сравнению с показателем ранговой корреляции Спирмэна.
Все основные положения и замечания, данные при описании показателя Спирмэна, справедливы и в отношении коэффициента Кендалла. Расчет ведется по формуле:

$$\tau = \frac{S_\tau}{\sqrt{\left(\frac{n(n-1)}{2} - B_x\right)\left(\frac{n(n-1)}{2} - B_y\right)}}; S_\tau = \sum_{i=1}^{n}\sum_{j=i+1}^{n} sign(r_j - s_i),$$

где r_i, s_i, $i = 1,2,...,n$ – массивы рангов выборочных совокупностей,
n – численность каждой выборки.
B_x, B_y – поправки на объединение рангов в соответствующих совокупностях, вычисляемые по формуле

$$B_{\cdot} = \frac{1}{2}\sum_{i=1}^{m} n_i(n_i - 1),$$

где m – число групп объединенных рангов,

n_i, i = 1,2,...,m – число рангов в i–ой группе.
Расстояние Кендалла как мера сходства может быть получено из рассмотренного коэффициента путем отнятия от единицы.

12.3.1.8. Расстояние Жаккара

Показатель подобия Жаккара (коэффициент Жаккара) вычисляется по формуле
$J = a / (a + b + c)$,
где a, b, c – значения в клетках таблицы 2 x 2.
Расстояние Жаккара как мера сходства может быть получено из рассмотренного коэффициента путем отнятия от единицы.

12.3.1.9. Расстояние Рассела–Рао

Показатель подобия Рассела и Рао вычисляется по формуле
$J = a / (a + b + c + d)$,
где a, b, c, d – значения в клетках таблицы 2 x 2.
Расстояние Рассела–Рао как мера сходства может быть получено из рассмотренного коэффициента путем отнятия от единицы.

12.3.1.10. Расстояние Бравайса

Специальная форма коэффициента корреляции – коэффициент сопряженности Бравайса (φ–коэффициент Пирсона) – рассчитывается по формуле

$$C = \frac{ad - bc}{\sqrt{(a+b)(a+c)(d+b)(d+c))}},$$

где a, b, c, d – значения в клетках таблицы 2 x 2.
Расстояние Бравайса как мера сходства может быть получено из рассмотренного коэффициента путем отнятия от единицы.

12.3.1.11. Расстояние Юла

Ориентировочную оценку корреляционной связи в случае исследования таблиц сопряженности 2 x 2 может дать коэффициент ассоциации Юла. Вычисления производятся по формуле

$$Q = \frac{ad - bc}{ad + bc},$$

ггде a, b, c, d – значения в клетках таблицы 2 x 2.
Расстояние Юла как мера сходства может быть получено из рассмотренного коэффициента путем отнятия от единицы.

12.3.1.12. Расстояние отношений

Расстояние отношений (расстояние между матрицами отношений) построено на результатах исследования в теории множеств и может быть применено к объектам, измеренным в различных, в том числе смешанных, шкалах. Примерами такого рода объектов могут быть экспертные ранжировки, для обработки которых и была первоначально представлена рассматриваемая опция программы. Смешанные данные часто возникают в медицинских исследованиях, когда вектор, описывающий объект (пациента), представляет собой совокупность количественных (результаты инструментальных исследований) и качественных (наличие–отсутствие симптомов) данных.
Расстояние отношений между объектами k и l определяется по формуле

$$d(P_k, P_l) = \sum_{i=1}^{n}\sum_{j=1}^{n}\left|p_{ij}^{(k)} - p_{ij}^{(l)}\right|,$$

где $p_{ij}^{(\cdot)}, i = 1,2,...,n;\ j = 1,2,...,n,$ – элементы матрицы отношений частичного порядка, которые автоматически вычисляются программой на основе матрицы исходных данных (в экспертных оценках – матрицы опроса), как рассмотрено в главе «Обработка экспертных оценок».
При использовании представленной меры из методов кластерного анализа, реализованных в настоящем программным обеспечении, можно использовать только метод средней связи Кинга. Метод Уорда неприменим по понятной причине. Метод *k*–средних для использования предлагаемой меры нуждается в модификации.
Обратим внимание на одну специфическую особенность взаимодействия методов настоящей главы и главы «Обработка экспертных оценок». При классификации экспертов для выявления их однородных групп в качестве объектов классификации мы подразумеваем самих экспертов. Поэтому, если при обработке экспертных ранжировок объекты расположены в строках, а эксперты – в столбцах, то при классификации экспертов (матрица экспертных ранжировок – та же самая) следует выбрать опцию «Объекты в столбцах».

Подробно о данной мере рассказано в книге Литвака. О применении новых результатов теории множеств в кластерном анализе см. работы Петровского.

12.3.2. Метод средней связи Кинга

Метод средней связи Кинга (King) является одним из важнейших иерархических агломеративных методов кластерного анализа. Процесс классификации состоит из элементарных шагов:

- Поиск и объединение двух наиболее похожих объектов в матрице сходства.
- Основанием для помещения объекта в кластер является близость двух объектов, в зависимости от меры сходства.
- На каком–либо этапе ранее объединенные в один кластер объекты считаются одним объектом с усредненными по кластеру параметрами.
- На следующем этапе находятся два очередных наиболее похожих объекта, и процедура повторяется с шага 2 до полного исчерпания матрицы сходства.

При использовании представленного здесь метода не возникает проблемы возможного несоответствия применяемой меры и шкалы измерения, т.к. метод оперирует не исходными объектами, а построенной матрицей сходства, по определению являющейся количественной. Координаты центра тяжести кластера вычисляются не по исходным данным – они являются продуктом манипуляций с матрицей сходства.
В качестве меры различия для метода средней связи используется любая из представленных в программе мер, чем и определяется универсальность метода для любых типов данных, в том числе для смешанных данных.
Помимо общей информации (число объектов, число параметров, тип связи), программа выдает таблицу номеров объединенных объектов и уровней соответствующих связей. После объединения пары объектов второй объект каждой пары исключается из рассмотрения и делается перенумерация остальных объектов.
Программа также выдает таблицу принадлежности объектов кластерам. Данная таблица строится на основе следующей идеи. Если взять построенную по результатам анализа дендрограмму (см. ниже раздел «Графическое представление результатов кластерного анализа»), то мысленно двигая воображаемую горизонтальную линию от самого верхнего значения ординаты, соответствующего максимальному уровню связи, вниз, мы последовательно пересекаем 1 (верхний уровень), 2, 3, … вертикальных частей линий,

соединяющих объекты. Как только мы достигаем заданного пользователем числа пересечений, не представляет никакого труда, начиная с данного уровня связи, «размотать» дендрограмму в обратном, нисходящем направлении (дендрограмма строилась снизу вверх) и установить, какие объекты принадлежат той или иной ветви. Гроздья данных объектов и будут составлять кластеры.

12.3.3. Метод Уорда

Метод Уорда (Ward) является одним из иерархических агломеративных методов кластерного анализа. Процесс классификации состоит из элементарных шагов:

- Поиск и объединение двух наиболее похожих объектов в матрице сходства.
- Основанием для помещения объекта в кластер является минимум дисперсии внутри кластера при помещении в него текущего классифицируемого объекта.
- На каком–либо этапе ранее объединенные в один кластер объекты считаются одним объектом с усредненными по кластеру параметрами.
- На следующем этапе находятся два очередных наиболее похожих объекта, и процедура повторяется с шага 2 до полного исчерпания матрицы сходства.

В качестве меры различия для метода Уорда используется только евклидово расстояние. Этим фактом вызвано ограничение области применения программы только количественной шкалой.

Помимо общей информации (число объектов, число параметров), программа выдает таблицу номеров объединенных объектов и уровней соответствующих связей. После объединения пары объектов второй объект каждой пары исключается из рассмотрения и делается перенумерация остальных объектов. Программа также выдает таблицу принадлежности объектов кластерам.

Программа также выдает таблицу принадлежности объектов кластерам. Данная таблица строится на основе следующей идеи. Если взять построенную по результатам анализа дендрограмму (см. ниже раздел «Графическое представление результатов кластерного анализа»), то мысленно двигая воображаемую горизонтальную линию от самого верхнего значения ординаты, соответствующего максимальному уровню связи, вниз, мы последовательно пересекаем 1 (верхний уровень), 2, 3, … вертикальных частей линий, соединяющих объекты. Как только мы достигаем заданного пользователем числа пересечений, не представляет никакого труда, начиная с данного уровня связи, «размотать» дендрограмму в обратном, нисходящем направлении (дендрограмма строилась снизу вверх) и установить, какие объекты принадлежат той или иной ветви. Гроздья данных объектов и будут составлять кластеры.

12.3.4. Метод k–средних Мак–Куина

Теоретическое обоснование метода *k*–средних (*k* внутригрупповых средних) Мак–Куина (McQueen) сравнительно просто, логично и может быть найдено во многих источниках. Принцип классификации сводится к следующим элементарным шагам:

- Некоторое, возможно, случайное, исходное разбиение множества объектов на заданное число кластеров (классов, групп, популяций). Расчет «центров тяжести» кластеров.
- Отнесение остальных объектов к ближайшим кластерам.
- Пересчет новых «центров тяжести» кластеров.
- Переход к шагу 2, пока новые «центры тяжести» кластеров не перестанут отличаться от старых.
- Получено оптимальное разбиение.

В качестве меры различия для метода средней связи используется любая из представленных в программе мер, предназначенных для количественных данных.
Помимо общей информации (число объектов, число параметров, тип связи), программа выдает координаты «центров тяжести» кластеров и таблицу принадлежности объектов кластерам. Отметим, что в результате расчета может быть получено, что часть кластеров окажется пустой. Это – следствие того, что пользователем задано слишком много кластеров, причем это число превышает естественное количество кластеров, существующее в представленных исходных данных. Результаты анализа могут быть использованы, а пустые кластеры следует просто не принимать во внимание.
Результаты расчета могут быть использованы для графического построения пространственных эллипсоидов.

12.3.5. Модифицированный метод k–средних

С целью использования метода *k*–средних для кластерного анализа данных, измеренных в шкалах, отличной от количественной, требуется модификация метода, заключающаяся в использовании в качестве центра тяжести кластера не среднего значения, вычисление которого корректно может быть выполнено только для количественных данных, а медианы. Универсальным решением может быть медиана множества, представленная в главе «Описательная статистика».
Принцип классификации сводится к следующим элементарным шагам:

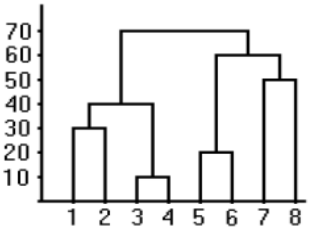
1. Некоторое, возможно, случайное, исходное разбиение множества объектов на заданное число кластеров (классов, групп, популяций). Расчет «центров тяжести» кластеров, в качестве которых для смешанных и произвольных шкал может быть использована медиана Ойя (median Oja), для данных типа экспертных оценок может быть использовано среднее Кемени или медиана Кемени, представленные в главе «Обработка экспертных оценок».
2. Отнесение остальных объектов к ближайшим кластерам.
3. Пересчет новых «центров тяжести» кластеров.
4. Переход к шагу 2, пока новые «центры тяжести» кластеров не перестанут отличаться от старых.
5. Получено оптимальное разбиение.

В качестве меры различия для метода средней связи используется любая из представленных в программе мер, а также данные, измеренные в смешанных шкалах.
Помимо общей информации (число объектов, число параметров, тип связи), программа выдает координаты «центров тяжести» кластеров и таблицу принадлежности объектов кластерам. Отметим, что в результате расчета может быть получено, что часть кластеров окажется пустой. Это – следствие того, что пользователем задано слишком много кластеров, причем это число превышает естественное количество кластеров, существующее в представленных исходных данных. Результаты анализа могут быть использованы, а пустые кластеры следует просто не принимать во внимание.

12.3.6. Графическое представление результатов кластерного анализа

Результаты кластерного анализа могут быть более наглядными, если их представить в виде графиков.
Результаты расчета методом *k*–средних могут быть очевидно использованы для графического построения пространственных эллипсоидов. Для этого достаточно изобразить координаты объектов в *k*–мерном пространстве (при числе «измерений», превышающем 2 или 3, можно изобразить 2 или 3-мерные срезы данного пространства).

По результатам расчета иерархическими методами можно построить специальный график, называемый дендрограммой (дендограммой). Предположим, после применения одного из иерархических методов получены результаты классификации в виде величин связи для пар объектов. Идея построения дендрограммы состоит в том, что пары объектов, отложенных по оси абсцисс, соединяются в соответствии с уровнем связи, отложенным по оси ординат. Ниже показан пример дендрограммы.



Список использованной и рекомендуемой литературы

1. Anderberg M.R. Cluster analysis for applications. – New York, NY: Academic Press, 1973.
2. Arnold S.J. A test for clusters // Journal of Marketing Research, 1979, vol. 16, pp. 545–551.
3. Art D., Gnanadesikan R., Kettenring R. Data–based metrics for cluster analysis // Utilitas Mathematica, 1982, vol. 21A, pp. 75–99.
4. Assael H. Segmenting markets by group purchasing behavior: An application of the AID technique // Journal of Marketing Research, May 1970, vol. 7 no. 2, pp. 153–158.
5. Banfield J.D., Raftery A.E. Model–based Gaussian and non–Gaussian clustering // Biometrics, 1993, vol. 49, pp. 803–821.
6. Bar–Hillel A. Learning a Mahalanobis metric from equivalence constraints / A. Bar–Hillel, T. Hertz, N. Shental et al. // Journal of Machine Learning Research, 2005, vol. 6, pp. 937–965.
7. Bensmail H. Inference in model–based cluster analysis / H. Bensmail, G. Celeux, A.E. Raftery et al. // Statistics and Computing, 1997, vol. 7, pp. 1–10.
8. Bezdek J.C. Pattern recognition with fuzzy objective function algorithms. – New York, NY: Plenum Press, 1981.
9. Bezdek J.C., Pal S.K. Fuzzy models for pattern recognition. – New York, NY: IEEE Press, 1992.
10. Binder D.A. Approximations to Bayesian clustering rules // Biometrika, 1981, vol. 68, pp. 275–285.
11. Binder D.A. Bayesian cluster analysis // Biometrika, 1978, vol. 65, pp. 31–38.
12. Blashfield R.K., Aldenderfer M.S. The literature on cluster analysis // Multivariate Behavioral Research, 1978, vol. 13, pp. 271–295.
13. Bock H.H. On some significance tests in cluster analysis // Journal of Classification, 1985, vol. 2, pp. 77–108.
14. Borgatti S.P. How to explain hierarchical clustering // Connections, 1994, vol. 17, no. 2, pp. 78–80.
15. Breiman L. Classification and regression trees / L. Breiman, J.H. Friedman, R.A. Olshen et al. – Monterey, CA: Wadsworth & Brooks/Cole,1984.
16. Calinski T., Harabasz J. A dendrite method for cluster analysis // Communications in Statistics, 1974, vol. 3, pp. 1–27.
17. Chatfield C., Collins A. Introduction to multivariate analysis. – New York, NY: Chapman & Hall / CRC, 2000.
18. Cooper M.C., Milligan G.W. The effect of error on determining the number of clusters // Proceedings of the International Workshop on Data Analysis, Decision Support and Expert

Knowledge Representation in Marketing and Related Areas of Research, 1988, pp. 319–328.
19. Duda R.O., Hart P.E. Pattern classification and scene analysis. – New York, NY: John Wiley & Sons, 1973.
20. Dudoit S., Frydlyand J., Speed T.P. Comparison of discrimination methods for the classification of tumors using gene expression data // Technical report No. 40, December 1984, University of California, Berkeley, CA.
21. Dunn J.C. A fuzzy relative of the ISODATA process and its use in detecting compact well–separated clusters // Journal of Cybernetics, 1973, vol. 3, pp. 32–57.
22. Duran B.S., Odell P.L. Cluster analysis. – New York, NY: Springer–Verlag, 1974.
23. Englemann L., Hartigan J.A. Percentage points of a test for clusters // Journal of the American Statistical Association, 1969, vol. 64, pp. 1647–1648.
24. Everitt B.S. A Monte Carlo investigation of the likelihood ratio test for the number of components in a mixture of normal distributions // Multivariate Behavioral Research, 1981, vol. 16, pp. 171–80.
25. Everitt B.S. Cluster analysis. – New York, NY: & Sons, 1993.
26. Everitt B.S. Unresolved problems in cluster analysis // Biometrics, 1979, vol. 35, pp. 169–181.
27. Everitt B.S., Hand D.J. Finite mixture distributions. – New York, NY: Chapman & Hall, 1981.
28. Everitt B.S., Landau S., Leese M. Cluster analysis. – New York, NY: Edward Arnold, 2001.
29. Gan G., Ma C., Wu J. Data clustering: Theory, algorithms, and applications. – Philadelphia, PA: The Society for Industrial and Applied Mathematics, 1979.
Girman C.J. Cluster analysis and classification tree methodology as an aid to improve understanding of benign prostatic hyperplasia. Ph.D. Thesis. – Chapel Hill, NC: Department of Biostatistics, University of North Carolina, 1994.
30. Gordon A.D. Classification. – Chapman & Hall / CRC, 1999.
31. Gower J.C. A comparison of some methods of cluster analysis // Biometrics, 1967, vol. 23, pp. 623–628.
32. Griffiths P. Applied Statistics Algorithms / Ed. by P. Griffiths, I.D. Hill. – Chichester, UK: Ellis Horwood Limited, 1985.
33. Grimm L.G. Reading and understanding more multivariate statistics / Ed. by L.G. Grimm, P.R. Yarnold. – Washington, DC: American Psychological Association, 2000.
34. Hardle W., Klinke S., Turlach B.A. XploRe: An interactive statistical computing environment. – New York, NY: Springer Verlag, 1995.
35. Hardle W., Simar L. Applied multivariate statistical analysis. – New York, NY: Springer, 2003.
36. Harman H.H. Modern factor analysis. – Chicago, IL: University of Chicago Press, 1976.
37. Hartigan J.A. Asymptotic distributions for clustering criteria // Annals of Statistics, 1978, vol. 6, pp. 117–131.
38. Hartigan J.A. Clustering algorithms. – New York, NY: John Wiley & Sons, 1975.
39. Hartigan J.A. Consistency of single linkage for high–density clusters // Journal of the American Statistical Association, 1981, vol. 76, pp. 388–394.
40. Hartigan J.A. Statistical theory in clustering // Journal of Classification, 1985, vol. 2, pp. 63–76.
41. Hartigan J.A., Hartigan P.M. The dip test of unimodality // Annals of Statistics, 1985, vol. 13, pp. 70–84.
42. Hartigan J.A., Wong M.A. Algorithm AS136: A K–means clustering algorithm // Applied Statistics,1979, vol. 28, pp. 100–108.
43. Hartigan P.M. Computation of the dip statistic to test for unimodality // Applied Statistics,

1985, vol. 34, pp. 320–325.
44. Hawkins D.M., Muller M.W., ten Krooden J.A. Cluster analysis // In Topics in applied multivariate analysis / Ed. by D.M. Hawkins. – Cambridge, UK: Cambridge University Press, 1982.
45. Hubert L. Approximate evaluation techniques for the single–link and complete–link hierarchical clustering procedures // Journal of the American Statistical Association, 1974, vol. 69, pp. 698–704.
46. Jain A.K., Dubes R.C. Algorithms for clustering data. – Upper Saddle River, NJ: Prentice–Hall, 1988.
47. Jardine N., Sibson R. Mathematical taxonomy. – London: John Wiley & Sons, 1971.
48. Kaufman L., Rousseeuw P.J. Finding groups in data: An introduction to cluster analysis. – New York, NY: John Wiley & Sons, 1990.
49. Kendall M.G., Stuart A. The advanced theory of statistics. Volume 3. – Charles Griffin, 1976.
50. Klastorin T.D. Assessing cluster analysis results // Journal of Marketing Research, 1983, vol. 20, pp. 92–98.
51. Krzanowski W.J. Principles of multivariate analysis. – Ney York, NY: Oxford University Press, 1990.
52. Lee K.L. Multivariate tests for clusters // Journal of the American Statistical Association, 1979, vol. 74, pp. 708–714.
53. Ling R.F. A probability theory of cluster analysis // Journal of the American Statistical Association, 1973, vol. 68, pp. 159–169.
54. Liu X., Krishnan A., Mondry A. An entropy–based gene selection method for cancer classification using microarray data // BMC Bioinformatics, 2005, vol. 6:76.
55. Lucy D. Introduction to statistics for forensic scientists. – Chichester, West Sussex: John Wiley & Sons, 2005.
56. MacQueen J.B. Some methods for classification and analysis of multivariate observations // Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 1967, vol. 1, pp. 281–297.
57. Marriott F.H.C. Practical problems in a method of cluster analysis // Biometrics, 1971, vol. 27, pp. 501–514.
58. Marriott F.H.C. Separating mixtures of normal distributions // Biometrics, 1975, vol. 31, pp. 767–769.
59. Massart D.L., Kaufman, L. The interpretation of analytical chemical data by the use of cluster analysis. – New York, NY: John Wiley & Sons, 1983.
60. McClain J.O., Rao V.R. CLUSTISZ: A program to test for the quality of clustering of a set of objects // Journal of Marketing Research, 1975, vol. 12, pp. 456–460.
61. McLachlan G.J., Basford K.E. Mixture models. – New York, NY: Marcel Dekker, 1988.
62. Mezzich J.E., Solomon H. Taxonomy and behavioral science. – New York, NY: Academic Press, 1980.
63. Milligan G.W. A review of Monte Carlo tests of cluster analysis // Multivariate Behavioral Research, 1981, vol. 16, pp. 379–407.
64. Milligan G.W. An examination of the effect of six types of error perturbation on fifteen clustering algorithms // Psychometrika, 1980, vol. 45, pp. 325–342.
65. Milligan G.W., Cooper M.C. An examination of procedures for determining the number of clusters in a data set // Psychometrika, 1985, vol. 50, pp. 159–179.
66. Minnotte M.C. A test of mode existence with applications to multimodality. Ph.D. Thesis. – Rice University, Department of Statistics, 1992.
67. Mucha H.–J. Clusteranalyse mit microcomputern. – Berlin: Akademie Verlag, 1992.
68. Mueller D.W., Sawitzki G. Excess mass estimates and tests for multimodality // Journal of the

American Statistical Association, 1991, 86, 738 –746.
69. Nevalainen J., Larocque D., Oja H. On the multivariate spatial median for clustered data // The Canadian Journal of Statistics, 2007, vol. 35, pp. 215–231.
70. Niinimaa A., Oja H. Multivariate median // Encyclopedia of Statistical Sciences. – New York, NY: John Wiley & Sons, 2006.
71. Oja H. Descriptive statistics for multivariate distributions // Statistics and Probability Letters, 1983, vol. 1, pp. 327–332.
72. Pollard D. Strong consistency of k–means clustering // Annals of Statistics, 1981, vol. 9, pp. 135–140.
73. Polonik W. Measuring mass concentrations and estimating density contour clusters – An excess mass approach // Technical Report, Universitaet Heidelberg, Beitraege zur Statistik, Nr. 7, 1993.
74. Punj G., Stewart D.W. Cluster analysis in marketing research: Review and Suggestions for application // Journal of Marketing Research, May 1983, vol. 20, no. 2, pp. 134–148.
75. Ronkainen T., Oja H., Orponen P. Computation of the multivariate Oja median // Developments in Robust Statistics / Ed. by R. Dutter, P. Filzmoser, U. Gather et al. – Heidelberg: Springer–Verlag, 2003, pp. 344–359.
76. Rousseeuw P.J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis // Journal of Computational and Applied Mathematics, 1987, vol. 20, pp. 53–65.
77. Ruspini E.H. A new approach to clustering // Information Control, 1969, vol. 15, pp. 22–32.
78. Scott A.J., Symons M.J. Clustering methods based on likelihood ratio criteria // Biometrics, 1971, vol. 27, pp. 387–397.
79. Silverman B.W. Density estimation. – New York, NY: Chapman & Hall, 1986.
80. Sneath P.H.A., Sokal R.R. Numerical taxonomy. – San Francisco, CA: W.H. Freeman, 1973.
81. Spath H. Cluster analysis algorithms. – Chichester, England: Ellis Horwood, 1980.
82. Sugar C.A., Lenert L.A., Olshen R.A. An application of cluster analysis to health services research: Empirically defined health states for depression from the SF–12. Technical Report No. 203, 1999. Division of Biostatistics, Stanford University School of Medicine.
83. Symons M.J. Clustering criteria and multivariate normal mixtures // Biometrics, 1981, vol. 37, pp. 35–43.
84. Tabachnick B.G., Fidell L.S. Using multivariate statistics. – Boston, MA: Allyn & Bacon, 2000.
85. Thode H.C.Jr., Mendell N.R., Finch S.J. Simulated percentage points for the null distribution of the likelihood ratio test for a mixture of two normals // Biometrics, 1988, vol. 44, pp. 1195–1201.
86. Tibshirani R., Walther G., Hastie T. Estimating the number of clusters in a dataset via the gap statistic // Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2001, vol. 63, pp. 411–423.
87. Titterington D.M., Smith A.F.M., Makov U.E. Statistical analysis of finite mixture distributions. – New York, NY: John Wiley & Sons, 1985.
88. Valente de Oliveira J., Pedrycz W. Advances in fuzzy clustering and its application. – Hoboken, NJ: John Wiley & Sons, 2007.
89. Van Belle G. Biostatistics: A methodology for the health sciences // G. van Belle, L.D. Fisher, P.J. Heagerty et al. – New York, NY: John Wiley & Sons, 2003.
90. Van Ryzin J. Classification and clustering / Ed. by J. Van Ryzin. – New York, NY: Academic Press, 1977.
91. Ward J.H. Hierarchical grouping to optimize an objective function // Journal of the American Statistical Association, 1963, vol. 58, pp. 236–244.
92. Webb A.R. Statistical pattern recognition. – New York, NY: & Sons, 2002.

93. Wind Y.J. No.s and advances in segmentation research // Journal of Marketing Research, August 1978, vol. 15 no. 3, pp. 317–337.
94. Wolfe J.H. Comparative cluster analysis of patterns of vocational interest // Multivariate Behavioral Research, 1978, vol. 13, pp. 33–44.
95. Wolfe J.H. Pattern clustering by multivariate mixture analysis // Multivariate Behavioral Research, 1970, vol. 5, pp. 329–350.
96. Wong M.A. A hybrid clustering method for identifying high–density clusters // Journal of the American Statistical Association, 1982, vol. 77, pp. 841–847.
97. Wong M.A., Lane T. A kth nearest neighbor clustering procedure // Journal of the Royal Statistical Society: Series B (Statistical Methodology), 1983, vol. 45, pp. 362–368.
98. Wong M.A., Schaack C. Using the kth nearest neighbor clustering procedure to determine the number of subpopulations // American Statistical Association Proceedings of the Statistical Computing Section, 1982, pp. 40–48.
99. Xiong H., Chen X. Kernel–based distance metric learning for microarray data classification // BMC Bioinformatics, 14 June 2006, 7:299.
100. Айвазян С.А. Прикладная статистика: Классификация и снижение размерности: Справочное издание / С.А. Айвазян, В.М. Бухштабер, И.С. Енюков и др. / Под ред. С.А. Айвазяна. – М.: Финансы и статистика, 1989.
101. Айвазян С.А., Мхитарян В.С. Прикладная статистика и основы эконометрики. – М.: ЮНИТИ, 1998
102. Белова Е.Б. Компьютеризованный статистический анализ для историков. Учебное пособие / Е.Б. Белова, Л.И. Бородкин, И.М. Гарскова и др. – М.: МГУ, 1999.
103. Браверман Э.М. Методы экстремальной группировки параметров и задача выделения существенных факторов // Автоматика и телемеханика, 1970, № 1, с. 123–132.
104. Браверман Э.М., Мучник И.Б. Структурные методы обработки эмпирических данных. – М.: Наука, 1983.
105. Вятченин Д.А. Нечеткие методы автоматической классификации. – Минск: УП «Технопринт», 2004.
106. Гайдышев И. Анализ и обработка данных: специальный справочник. – СПб: Питер, 2001.
107. Горский В.Г., Гриценко А.А., Орлов А.И. Метод согласования кластеризованных ранжировок // Автоматика и телемеханика, 2000, № 3, с. 179–187.
108. Двоенко С.Д. Алгоритмы автоматической классификации (обзор) // Автоматика и телемеханика, 1971, № 12, с. 78–113.
109. Двоенко С.Д. Неиерархический дивизимный алгоритм группировки // Автоматика и телемеханика, 1999, № 9, с. 47–57.
110. Дерябин В.Е. Критерий для определения таксономической ценности признака // В сб. Биометрический анализ в биологии. – М.: Издательство Московского университета, 1982, с. 118–130.
111. Дюк В. Обработка данных на ПК в примерах. – СПб.: Питер, 1997.
112. Дюран Б., Оделл П. Кластерный анализ. – М.: Финансы и статистика, 1977.
113. Жамбю М. Иерархический кластер–анализ и соответствия. – М.: Финансы и статистика, 1988.
114. Кендалл М., Стьюарт А. Многомерный статистический анализ и временные ряды. – М.: Наука, 1976.
115. Ким Дж.О. Факторный, дискриминантный и кластерный анализ / Дж.О. Ким, Ч.У. Мюллер, У.Р. Клекка и др. – М.: Финансы и статистика, 1989.
116. Крускал Дж. Взаимосвязь между многомерным шкалированием и кластер–

анализом // Классификация и кластер / Под ред. Дж. Вэн Райзина. – М.: Мир, 1980, с. 20–41.
117. Кулаичев А.П. Методы и средства комплексного анализа данных. – М.: ИНФРА–М, 2006.
118. Литвак Б.Г. Экспертная информация. Методы получения и анализа. – М.: Радио и связь, 1982.
119. Лумельский В.Я. Группировка параметров на основе квадратной матрицы связей // Автоматика и телемеханика, 1970, № 1, с. 133–143.
120. Мандель И.Д. Кластерный анализ. – М.: Финансы и статистика, 1988.
121. Миркин Б.Г. Анализ качественных признаков и структур. – М.: Статистика, 1980.
122. Петровский А.Б. Кластерный анализ объектов с противоречивыми свойствами // Десятая национальная конференция по искусственному интеллекту с международным участием КИИ–2006, 25–28 сентября 2006 г., Обнинск: Труды конференции. В 3–т. – М: Физматлит, 2006.
123. Петровский А.Б. Пространства множеств и мультимножеств. – М.: Едиториал УРСС, 2003.
124. Попов И.В., Фролкина Н.А. Анализ связанных объектов и визуализация результатов // Доклады международной конференции Диалог 2004.
125. Прохоров Ю.В. Вероятность и математическая статистика. Энциклопедия / Гл. ред. Ю.В. Прохоров. – М.: Научное издательство «Большая Российская энциклопедия», 1999.
126. Рао С.Р. Кластер–анализ в применении к изучению перемешивания рас в популяции людей // Классификация и кластер / Под ред. Дж. Вэн Райзина. – М.: Мир, 1980, с. 148–167.
127. Раушенбах Г.В. Проблемы измерения близости в задачах анализа данных // Программно–алгоритмическое обеспечение анализа данных в медико–биологических исследованиях. – М.: Наука, 1987.
128. Родионов Д.А. Справочник по математическим методам в геологии / Д.А. Родионов, Р.И.Коган, В.А. Голубева и др. – М.: Недра, 1987.
129. Родионов Д.А. Статистические методы разграничения геологических объектов по комплексу признаков. – М: Недра, 1968.
130. Смирнов Е.С. Таксономический анализ. – М.: Издательство Московского университета, 1969.
131. Сокал Р.Р. Кластер–анализ: предпосылки и основные направления // Классификация и кластер / Под ред. Дж. Вэн Райзина. – М.: Мир, 1980, с. 7–19.
132. Соломон Г. Зависящие от данных методы кластер–анализа // Классификация и кластер / Под ред. Дж. Вэн Райзина. – М.: Мир, 1980, с. 129–147.
133. Уиллиамс У.Т., Ланс Дж.Н. Методы иерархической классификации // Статистические методы для ЭВМ / Под ред. К. Энслейна, Э. Рэлстона, Г.С. Уилфа. – М.: Наука, 1986, с. 348–372.
134. Хант Э. Искусственный интеллект. – М.: Мир, 1978.
135. Холл Д.Д., Ханна Д. ISODATA: метод анализа сходств и различий в сложных реальных данных // Статистические методы для ЭВМ / Под ред. К. Энслейна, Э. Рэлстона, Г.С. Уилфа. – М.: Наука, 1986, с. 268–301.
136. Черныш М.Ф. Опыт применения кластерного анализа // Социология: методология, методы, математические модели (Социология: 4М), 2000, № 12.
137. Шрейдер Ю.А. Что такое расстояние? – М.: Физматгиз, 1963.

Глава 13. Информационный анализ

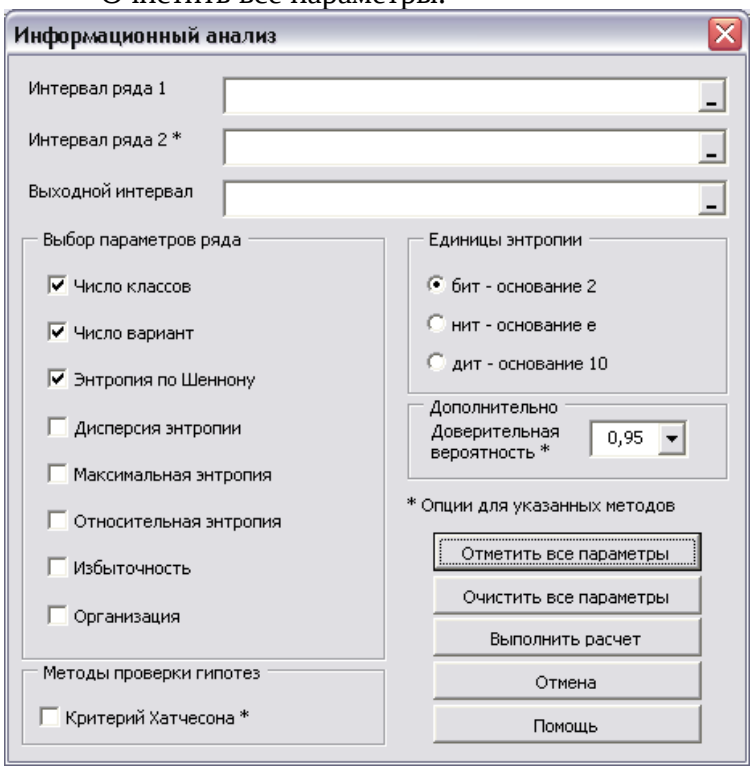
13.1. Введение

Программное обеспечение обеспечивает вычисление основных показателей разведочного информационного анализа.

13.2. Работа с программным обеспечением

Выберите из меню программы пункт AtteStat | Информационный анализ. На экране появится диалоговое окно, изображенное на рисунке.
Затем проделайте следующие шаги:

- Выберите или введите интервал переменной. Обратите особое внимание! Исходными данными для расчета является не сама эмпирическая выборка, а построенные на ее основе дискретный или интервальный вариационные ряды. Они представляют собой таблицы распределения количеств вариант по классам (частоты) и различаются тем, что в первом случае количества относятся к определенным, возможно нечисловым, значениям признаков, а во втором случае – к интервалам изменения признака (классовым интервалам).
- Если предполагается использовать методы, оперирующие двумя рядами, выберите или введите интервал второй переменной.
- Выберите или введите выходной интервал. Начиная с первой ячейки выходного интервала (следовательно, можно указать только одну ячейку, т.к. остальные ячейки интервала игнорируются), будут выведены вычисленные отмеченные Вами параметры информационного анализа.
- Отметьте необходимые информационные параметры расчета (только при расчете по одному ряду), при необходимости пользуясь кнопками Отметить все параметры или Очистить все параметры.



- Оставьте по умолчанию или измените единицы измерения энтропии и связанных с ней информационных параметров. Отметим, что данный выбор затрагивает одновременно все информационные параметры, для которых данные единицы измерения применимы.
- Нажмите кнопку «Выполнить расчет».

При ошибках, вызванных неверными действиями пользователя, выдаются сообщения об ошибках.

13.2.1. Сообщения об ошибках

При ошибках ввода и во время выполнения программы могут выдаваться диагностические сообщения следующих типов:

Ошибка	Комментарий
Пустая ячейка в области данных.	Проверьте исходные данные и заполните все ячейки, отмеченные Вами как входной интервал. Для избежания ошибок расчета, вызванных разногласиями, трактовать ли пустую ячейку как нуль, программное обеспечение требует заполнения всех ячеек. Если в ячейке не должно быть данных по физической природе исследуемого процесса, введите в данную ячейку нуль.
Нечисловой тип данных.	Проверьте типы ячеек входного интервала. Тип может быть только числовым. Проще всего выделить интервал ячеек и явно определить их тип как числовой стандартными средствами.
Мало данных для выбранного метода.	Для расчета необходимо выбрать интервал, содержащий хотя бы две ячейки с числовым значением.
Не определена область данных.	Вы не выбрали или неверно ввели входной интервал. Лучшим способом избежать ошибки является не ввод, а выделение интервала стандартным образом, т. е. протаскиванием курсора.
Не определена область вывода.	Вы не выбрали или неверно ввели выходной интервал. Лучшим способом избежать ошибки является не ввод, а выделение интервала стандартным образом, т. е. протаскиванием курсора.
Ошибочные данные.	Таблица сопряженности содержит ошибочные данные. Таблица сопряженности должна содержать только неотрицательные целые числа, т.к. в клетки таблицы помещаются количества вариант, обладающих данными признаками.

13.3. Теоретическое обоснование

Методы информационного анализа находят применение в различных научно–технических областях (примеры приводятся ниже). Программа обеспечивает вычисление основных показателей информационного анализа:

- Число классов.
- Число вариант.
- Энтропия.
- Дисперсия энтропии.

- Максимальная энтропия.
- Относительная энтропия.
- Избыточность.
- Организация.

Исходными данными для программы являются дискретный или интервальный вариационные ряды. Они представляют собой таблицы распределения количеств вариант по классам и различаются тем, что в первом случае количества относятся к определенным, возможно нечисловым, значениям признаков, а во втором случае – к интервалам изменения признака (классовым интервалам). Из исходной эмпирической выборки вариационный ряд удобно построить с помощью инструмента «Гистограмма» главы «Описательная статистика».

13.3.1. Число классов

Под группировкой (классификацией, разнесением вариант по классам) понимается некоторое разбиение эмпирической выборки, содержащей все N наблюдавшихся вариант $x_1,x_2,...,x_N$, на s интервалов (классов). Результатом группировки является вариационный ряд.
Исходными данными для расчета методами информационного анализа является не сама эмпирическая выборка, а построенные на ее основе, в зависимости от шкалы измерения исходных данных (см. главу «Введение»), дискретный или интервальный вариационные ряды. Они представляют собой таблицы распределения количеств вариант по классам и различаются тем, что в первом случае количества относятся к определенным, возможно нечисловым, значениям признаков, а во втором случае – к интервалам изменения признака (классовым интервалам).
Количества вариант, попавших в тот или иной класс в результате классификации, называют частотами. Частоты, отнесенные к численности N, называют частостями. Считается, что частости могут служить оценками вероятностей. Это утверждение тем вернее, чем больше численность изучаемой выборочной совокупности.
Часто группировка является естественной (например, виды растений). В данном случае для дискретного вариационного ряда число классов s равно числу градаций переменной (в рассмотренном примере – числу видов). Отметим, что при анализе реальных данных может оказаться, что в конкретной совокупности некоторые классы окажутся с нулевыми частотами. Учитывать данные нулевые частоты (считать их нулями без уменьшения числа классов s на количество классов с нулевыми частотами) либо не учитывать (считать только классы с ненулевыми частотами, соответственно уменьшая s), зависит от конкретной задачи.
Для практического построения интервального вариационного ряда на основе численных данных, прежде всего, необходимо определить либо задать число классов (групп, интервалов) s. Субъективным критерием правильности выбора числа классов является верная передача типа распределения эмпирических частот данной совокупности. Если выбрано слишком мало классов, можно потерять характерную картину эмпирического распределения. При слишком подробном делении на классы можно затушевать реальную картину распределения частот случайными отклонениями. Все подробности относительно оптимальной группировки численных выборок даны в главе «Описательная статистика».

13.3.2. Число вариант ряда

Совокупности состоят из отдельных элементов (объектов), которые объединены общностью некоторых свойств (признаков, переменных). В статистическом анализе данные объекты принято называть вариантами. Количество элементов (вариант) совокупности можно называть по–разному. Так, если речь идет о выборке, количество ее элементов может называться численностью, величиной или размером.

Численность вариационного ряда определяется по формуле

$$N = \sum_{i=1}^{s} n_i,$$

где n_i, $i = 1,2,...,s$ – численности классов группировки,
s – число классов группировки.
Численности классов группировки принято называть также частотами.
Знание числа вариант может быть полезно для элементарного пересчета исходных данных, если исходные данные заданы в виде частот распределения, а данное программное обеспечение требует, чтобы исходные данные были заданы в виде вариационного ряда, представляющего собой частоты в соответствующих классах.

13.3.3. Энтропия

В теории информации в качестве меры количества информации, возможности выбора (количества разнообразия) и неопределенности применяется величина, определяемая по формуле Шеннона

$$H = -\sum_{i=1}^{s} p_i \log_a p_i,$$

где p_i, $i = 1,2,...,s$ – вероятность появления дискретного события,
s – число классов группировки,
a – основание логарифма – единица, выбранная для оценки величины энтропии, обычно равная числу 2.
Величины p_i, $i = 1,2,...,s$, образуют множество вероятностей, но для практических вычислений вероятности допустимо заменить частотами распределений:

$$p_i \approx \frac{n_i}{N}, i = 1,2,..., s,$$

n_i, $i = 1,2,...,s$ – численности классов группировки (частоты),
s – число классов группировки,
N – число вариант ряда.
В практических вычислениях некоторые частоты могут оказаться нулевыми, поэтому условились считать, что $0 \cdot \log_a 0 = 0$.
Величина H называется энтропией дискретного множества вероятностей (энтропией дискретной случайной величины, средней собственной информацией, энтропией Шеннона, энтропией Шеннона–Винера) и в источниках иногда обозначается как I, чтобы отличить ее от энтропии непрерывного распределения. Энтропия представляет собой количественную меру степени неопределенности исхода случайного опыта, зависящую не от индивидуальных свойств результатов опыта, а от соответствующих вероятностей. В дискретном случае энтропия равна нулю, когда одна из вероятностей равна 1, а остальные нулю.
Энтропия непрерывного распределения с функцией плотности распределения $p(x)$ определяется как

$$H = -\int_{-\infty}^{\infty} p(x) \log_a p(x) dx$$

и называется также относительной или дифференциальной энтропией.
Если в качестве основания логарифма a выбрано число 2, энтропия вычисляется в битах, если число e – в нитах, если число 10 – в дитах. Если энтропия вычислена в нитах, для вычисления энтропии в битах нужно разделить значение в нитах на ln2 (это бывает необходимо, если в системе программирования стандартная функция вычисления логарифма по основанию 2 может быть не представлена). Утверждение вытекает из известной формулы

замены основания логарифмов

$$\log_a c = \frac{\log_b c}{\log_b a}$$

при $a > 0$ и $a \neq 1$.

Исходные данные для вычисления энтропии системы представляют собой дискретный или интервальный вариационный ряд.

Некоторыми авторами энтропия называется количеством информации или двоичной энтропией. Так называемая энтропия Колмогорова, характеризующая хаотическое движение в фазовом пространстве произвольной размерности, также определяется по формуле Шеннона. Введены обобщения энтропии, такие как энтропия Реньи порядка α:

$$H_\alpha(p_1,...,p_n) = \frac{1}{1-\alpha}\log_2 \sum_{i=1}^{s} p_i^\alpha,$$

при α = 1 обычно называемая энтропией Шеннона,

при α = 0 – энтропией Хартли.

Двусторонний доверительный интервал оцениваемой энтропии вычисляется по формуле

$$I_H = \left(H - \Psi((1+\beta)/2)\frac{D_H}{\sqrt{s}}; H + \Psi((1+\beta)/2)\frac{D_H}{\sqrt{s}} \right),$$

где Ψ(.) – обратная функция стандартного нормального распределения,

D_H – дисперсия энтропии.

Подробное теоретическое обоснование см. в книге Дмитриева.

13.3.4. Дисперсия энтропии

Дисперсия энтропии вычисляется по формуле

$$D_H = \frac{1}{N}\left[\sum_{i=1}^{s} p_i \log_a^2 p_i - \left(\sum_{i=1}^{s} p_i \log_a p_i \right)^2 \right] + \frac{s-1}{2N^2},$$

где p_i, $i = 1,2,...,s$ – вероятность появления дискретного события,

s – число классов группировки,

N – число вариант ряда,

a – основание логарифма, обычно равное числу 2.

Дисперсия энтропии находит применение при оценке значимости различий индексов Шеннона, характеризующих видовое разнообразие, вычисленных для двух совокупностей, а также для вычисления доверительных интервалов оцениваемой энтропии. Соответствующий критерий предложен Хатчесоном (Hutchenson).

См. главу «Параметрическая статистика», статьи Хатчесона, Грениер (Grenier) с соавт., работу Шитикова с соавт., книгу Зара (Zar).

13.3.5. Максимальная энтропия

Максимальное разнообразие системы вычисляется по формуле Хартли

$H_{max} = \log_2 s$,

где H_{max} – максимальная энтропия,

s – число классов группировки.

Таким образом, справедлива формула

$0 \le H \le H_{max}$,

где H – энтропия.

Максимум энтропии соответствует наибольшей неопределенности или равенству вероятностей всех возможностей. Опыт имеет наибольшую энтропию при k равновероятных исходах с вероятностями (в практических вычислениях – частотами) $1 / k$. Степень неопределенности опыта тем больше, чем больше число k его исходов.
Энтропия графического изображения зависит от количества уровней, а при одинаковом числе уровней – от закона распределения. Равномерный закон распределения соответствует полной хаотичности. В этом случае энтропия достигает максимума, который зависит только от количества уровней
$H_{max} = \log_a(h_{max} – h_{min} + 1)$,
где (.) – размах дискретной случайной величины,
h_{max} – максимальное значение уровня изображения,
h_{min} – минимальное значение уровня изображения.
Напомним, что размах выборки (размах вариации, амплитуда ряда) – это разность между максимумом и минимумом вариант выборки.

13.3.6. Относительная энтропия

Для сравнения систем, различающихся по количеству (!) элементов кода, простое сопоставление энтропий не будет корректным. Для решения задачи применяется относительная энтропия (коэффициент сжатия информации), определяемая как
$h = H / H_{max}$,
где H – энтропия,
H_{max} – максимальная энтропия.
Относительная энтропия определяет относительную степень информационной загруженности системы по отношению к возможной максимальной нагрузке. Кроме того, относительная энтропия, как и избыточность, может характеризовать степень близости закона распределения к равномерному.

13.3.7. Избыточность

Избыточность показывает, какая доля или процент передаваемой информации является избыточной. Она дает соотношение между полным количеством информации, шумом и сохранившейся упорядоченностью системы. Избыточность (в источниках обозначается также как R) вычисляется по формуле
$D = 1 – H / H_{max}$,
где H – энтропия,
H_{max} – максимальная энтропия.
Избыточность характеризует степень близости закона распределения к равномерному распределению и может быть выражена в долях, как в показанной формуле, либо в процентах. Например, для полутонового изображения с 256 уровнями энтропия не может превышать 8, а избыточность при равномерном законе распределения равна нулю.

13.3.8. Организация системы

Под организацией системы понимают реализованную в ней неопределенность. Абсолютная организация системы вычисляется по формуле
$O = H_{max} – H$,
где H_{max} – максимальная энтропия,
H – энтропия.
При организации, равной максимальной энтропии, система становится детерминированной, полностью стабильной.

13.3.9. Примеры информационного анализа

Ниже представлены несколько практических приложений информационного анализа:

- Разведочный информационный анализ.
- Исследование структурной перестройки объекта.
- Сравнение групп по индексу межвидового разнообразия.

13.3.9.1. Разведочный информационный анализ

В математической статистике и математическом моделировании важно точно обосновать возможность применения тех или иных методов анализа и адекватность модели.
Напомним, что математическим моделированием называют приближенное описание явлений, выраженное с помощью математической символики, а также процесс их изучения с помощью математических моделей. Математическое моделирование включает этапы:

- Формулировка законов, связывающих основные объекты модели, на основе изучения явлений и проникновения в их взаимосвязи.
- Составление уравнений модели. Исследование и математическое решение задачи, к которой приводит математическая модель.
- Сопоставление результатов расчета математической модели с данными изучаемого явления, полученными в результате наблюдения за этим явлением. При наличии в модели параметров, неизвестных на этапе составления или недоступных для прямого измерения, производится идентификация модели на основе экспериментальных данных.
- Исследование изучаемого явления с помощью математической модели. Уточнение модели на основе новых данных об изучаемом явлении.

Математическое моделирование строит модели в виде различных уравнений или систем уравнений. Уравнения математической модели могут быть алгебраическими, дифференциальными, интегральными и их совокупностями.
Математической же статистикой называют раздел математики, посвященный математическим методам сбора, систематизации, обработки и интерпретации статистических данных, а также использование их для научных и практических выводов. Напомним, что под статистическими данными понимают любую систему данных: числовую информацию, извлекаемую из результатов выборочных обследований; [эмпирические] выборки из любых генеральных совокупностей; результаты измерений и т.п. Математическая статистика строит вероятностно–статистическую, а не математическую в указанном выше смысле, модель.
Статистической моделью называют описание выборочного пространства всех мыслимых исходов наблюдаемого случайного явления, выделение семейства распределений вероятностей этих исходов и определение другой априорной информации об этом семействе.
Используя разведочный информационный анализ, в первом приближении проверку адекватности типа модели можно выполнить путем вычисления информационных показателей представленных статистических данных. Сделать выводы по вычисленным показателям необходимо в соответствии со следующей таблицей:

Величина избыточности, %	Характеристика системы	Адекватный тип модели
0–10	Вероятностная	Вероятностно–статистическая
10–30	Вероятностно–детерминированная	Дифференциальные уравнения
30–100	Детерминированная	Дифференциальные или

		интегральные уравнения

Это очень важные результаты. Например, при вычисленной избыточности, равной 25%, вероятностно–статистическая модель не будет адекватно описывать исследуемое явление. Исследователю придется заняться составлением математической модели в виде дифференциальных уравнений или сменить работу.

13.3.9.2. Исследование структурной перестройки объекта

Разность избыточности в норме и при патологии приводит к понятию ненадежности (эквивокации) передачи информации, что дает количественную характеристику структурной перестройки исследуемого объекта (системы). Вычисление эквивокации производится по формуле

$$D = R_{norm} - R_{pat} = \frac{H_{pat} - H_{norm}}{H_{\max}},$$

где R_{norm} – избыточность в норме,
R_{pat} – избыточность в патологии,
H_{norm} – энтропия в норме,
H_{pat} – энтропия в патологии.

13.3.9.3. Сравнение групп по индексам межвидового разнообразия

В качестве примера применения информационного анализа (в том числе его использования для проверки статистических гипотез) укажем так называемые [информационные] индексы видового разнообразия. В литературе часто используется индекс Шеннона (Shannon index), называемый авторами также индексом Шеннона–Винера (Shannon–Wiener diversity index) или индексом Шеннона–Уивера (Shannon–Weaver diversity index).
Индекс Шеннона представляет собой энтропию, вычисленную с использованием того или иного основания логарифма. Обычно используется основание 2, но встречаются публикации, в которых используется основание e = 2,718281828... , а также 10. Настоящее программное обеспечение позволяет оперировать всеми данными вариантами.
Для сравнения групп по индексам Шеннона применяется специальная модификация критерия Стьюдента, предложенная Хатчесоном (Hutcheson). Статистика критерия вычисляется по формуле

$$t = \frac{\left|H_1 - H_2\right|}{\sqrt{D_{H_1}^2 / n_1 + D_{H_2}^2 / n_2}},$$

где H_1 и H_2 – индексы Шеннона (энтропии) совокупностей,
D_{H_1} и D_{H_2} – соответствующие оценки дисперсий индексов Шеннона,
n_1 и n_2 – соответствующие численности совокупностей.
Распределение статистики критерия Хатчесона близко к t–распределению Стьюдента при числе степеней свободы, равном

$$\nu = \frac{\left(D_1 + D_2\right)^2}{D_{H_1}^2 / n_1 + D_{H_2}^2 / n_2}.$$

В специальной литературе представлены и другие индексы (многие не имеющие отношения к теории информации).

См. работы Хатчесона, Грениер (Grenier) с соавт., Шитикова с соавт., Кейлок (Keylock), Магурран (Magurran), Розенцвейг (Rosenzweig), Пилу (Pielou), книгу Зара (Zar).

Список использованной и рекомендуемой литературы

1. Dudok de Wit T. When do finite sample effects significantly affect entropy estimates? // The European Physical Journal B, 1999, vol. 11, no. 3, pp. 513–516.
2. Ebrahimi N., Pflughoeft K., Soofi E.S. Two measures of sample entropy // Statistics & Probability Letters, 22 June 1994, vol. 20, no. 3, pp. 225–234.
3. Gray R.M. Entropy and information theory. – New York, NY: Springer–Verlag, 1991.
4. Grenier C., Hamon P., Bramel–Cox P.J. Core collection of sorghum: II. Comparison of three random sampling strategies // Crop Science, January–February 2001, vol. 41, no. 1, pp. 241–246.
5. Hutcheson K. A test for comparing diversities based on the Shannon formula // Journal of Theoretical Biology, October 1970, vol. 29, no. 1, pp. 151–154.
6. Ikeda S. Continuity and characterization of Shannon–Wiener information measure for continuous probability distributions // Annals of the Institute of Statistical Mathematics, 1959, vol. 11, no. 2, pp. 131–144.
7. Keylock C.J. Simpson diversity and the Shannon–Wiener index as special cases of a generalized entropy // Oikos, April 2005, vol. 109, no. 1, p. 203.
8. Krippendorf K. Information theory. Structural models for qualitative data. – Beverly Hills, CA: Sage Publications, 1986.
9. Learned–Miller E.G., Fisher III J.W. ICA using spacings estimates of entropy // Journal of Machine Learning Research, 2003, vol. 4, pp. 1271–1295.
10. Liu X., Krishnan A., Mondry A. An entropy–based gene selection method for cancer classification using microarray data // BMC Bioinformatics, 2005, vol. 6:76.
11. MacKay D.J.C. Information theory, inference and learning algorithms. – New York, NY: Cambridge University Press, 2003.
12. Magurran A.E. Ecological diversity and its measurement. – Princeton, NJ: Princeton University Press, 1988.
13. Pielou E.C. Shannon's formula as a measure of specific diversity: Its use and misuse // The American Naturalist, September–October, 1966, vol. 100, no. 914, pp. 463–465.
14. Pincus S.M. Approximate entropy as a measure of system complexity // Proceedings of the National Academy of Sciences of the United States of America, March 1991, vol. 88, pp. 2297–2301.
15. Richman J.S., Lake D.E., Moorman J.R. Sample entropy // Methods in Enzymology, 2004, vol. 384, Numerical Computer Methods, part E, pp. 172–184.
16. Rosenzweig M.L. Species diversity in space and time. – New York, NY: Cambridge University Press, 1995.
17. Shannon C.E. A mathematical theory of communication // The Bell System Technical Journal, July, October 1948, vol. 27, pp. 379–423, 623–656.
18. Weaver W., Shannon C.E. The mathematical theory of communication. – Urbana, IL: University of Illinois Press, 1949.
19. Zar J.H. Biostatistical analysis. – Englewood Cliffs, NJ: Prentice Hall, 1999.
20. Zografos K., Nadarajah S. Expressions for Renyi and Shannon entropies for multivariate distributions // Statistics & Probability Letters, 2005, vol. 71, pp. 71–84.
21. Бандарин В.А. Теория информации в медицине / Под ред. В.А. Бандарина. – Минск: Беларусь, 1974.
22. Бикел П., Доксам К. Математическая статистика. – М.: Финансы и статистика, 1983.
23. Биргер И.А. Техническая диагностика. – М.: Машиностроение, 1978.
24. Блюменфельд Л.А. Информация, термодинамика и конструкция биологических систем // Соросовский образовательный журнал, 1996, т. 2, № 7, с. 88–92.
25. Бриллюэн Л. Наука и теория информации. – М.: Государственное издательство

физико–математической литературы, 1960.
26. Вентцель Е.С. Теория вероятностей. – М.: Высшая школа, 1999.
27. Гайдышев И. Анализ и обработка данных: специальный справочник. – СПб: Питер, 2001.
28. Голдман С. Теория информации. – М.: Иностранная литература, 1957.
29. Григорович В.Г. Информационные методы в управлении качеством / В.Г. Григорович, Н.О. Козлова, В.В. Шильдин и др. – М.: РИА Стандарты и качество, 2001.
30. Дмитриев В.И. Прикладная теория информации. – М.: Высшая школа, 1989.
31. Кадыров Х.К., Антомонов Ю.Г. Синтез математических моделей биологических и медицинских систем. – Киев: Наукова думка, 1974.
32. Козлова Н.О., Шильдин В.В., Литвинов О.В. Информационные методы в управлении качеством. – М.: РИА «Стандарты и качество», 2001.
33. Колмогоров А.Н. Теория информации и теория алгоритмов. – М.: Наука, 1987.
34. Колмогоров А.Н. Три подхода к определению понятия «количество информации» // Проблемы передачи информации / Под ред. П.С. Яглома, 1965, т. 1, вып. 1.
35. Кошкин Г.М. Энтропия и информация // Соросовский образовательный журнал, 2001, т. 7, № 11, с. 122–127.
36. Краснощеков П.С., Петров А.А. Принципы построения моделей. – М.: ФАЗИС: ВЦ РАН, 2000.
37. Кульбак С. Теория информации и статистика. – М.: Наука, 1967.
38. Леонтюк А.С., Леонтюк Л.А.., Сыкало А.И. Информационный анализ в морфологических исследованиях. – Минск: Наука и техника, 1981.
39. Лидовский В.В. Теория информации. – М.: Компания Спутник+, 2004.
40. Лоскутов А.Ю., Михайлов А.С. Введение в синергетику: Учебное руководство. – М.: Наука, 1990.
41. Луценко Е.В. Математический метод СК–анализа в свете идей интервальной бутстрепной робастной статистики объектов нечисловой природы // Научный электронный журнал КубГАУ, 2004, № 01(3).
42. Мартин Н., Ингленд Дж. Математическая теория энтропии. – М.: Мир, 1988.
43. Миллер Р., Канн Д. Статистический анализ в геологических науках. – М.: Мир, 1967.
44. Паниотто В.И., Максименко В.С. Количественные методы в социологических исследованиях. – Киев: Наукова думка, 1982.
45. Прохоров Ю.В. Вероятность и математическая статистика. Энциклопедия / Гл. ред. Ю.В. Прохоров. – М.: Научное издательство «Большая Российская энциклопедия», 1999.
46. Рубин А.Б. Термодинамика биологических процессов // Соросовский образовательный журнал, 1998, т. 4, № 10, с. 77–83.
47. Стратонович Р.Л. Теория информации. – М.: Советское радио, 1975.
48. Фано Р.М. Передача информации. Статистическая теория связи. – М.: Мир, 1965.
49. Хакен Г. Информация и самоорганизация. Макроскопический подход к сложным системам. – М.: Едиториал УРСС, 2005.
50. Хартли Р.В.Л. Передача информации // Теория информации и ее приложения / Под ред. А.А. Харкевича. – М.: Физматгиз, 1959.
51. Шеннон К.Э. Некоторые задачи теории информации // Информационное общество: Сборник / Сост. А. Лактионов. – М.: Издательство «АСТ», 2004, с. 41–44.
52. Шеннон К.Э. Работы по теории информации и кибернетике. – М.: Издательство иностранной литературы, 1963.
53. Шитиков В.К., Розенберг Г.С. Оценка биоразнообразия: попытка формального обобщения // Количественные методы экологии и гидробиологии / Под ред. Г.С.

Розенберга. – Тольятти: ИЭВБ РАН, 2005, с. 91–129.
54. Шустер Г. Детерминированный хаос: Введение. – М.: Мир, 1988.
55. Яглом А.М., Яглом И.М. Вероятность и информация. – М.: Наука, 1973.

Глава 14. Распознавание образов с обучением

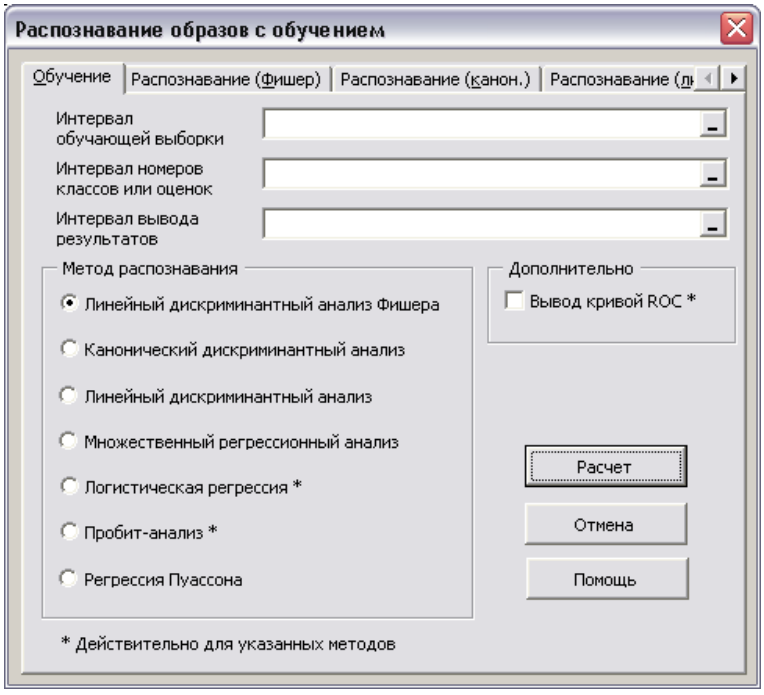
14.1. Введение

В программном обеспечении распознавания образов с обучением (с учителем) применяются следующие статистические методы, которые могут быть интерпретированы как методы распознавания образов с обучением:

- Линейный дискриминантный анализ Фишера.
- Канонический дискриминантный анализ.
- Линейный дискриминантный анализ.
- Линейный множественный регрессионный анализ.
- Логистическая регрессия.
- Пробит анализ.
- Регрессия Пуассона.

14.2. Работа с программным обеспечением

Выберите из меню программы пункт AtteStat | Распознавание образов. На экране появится диалоговое окно, подобное окну, изображенному на рисунке:

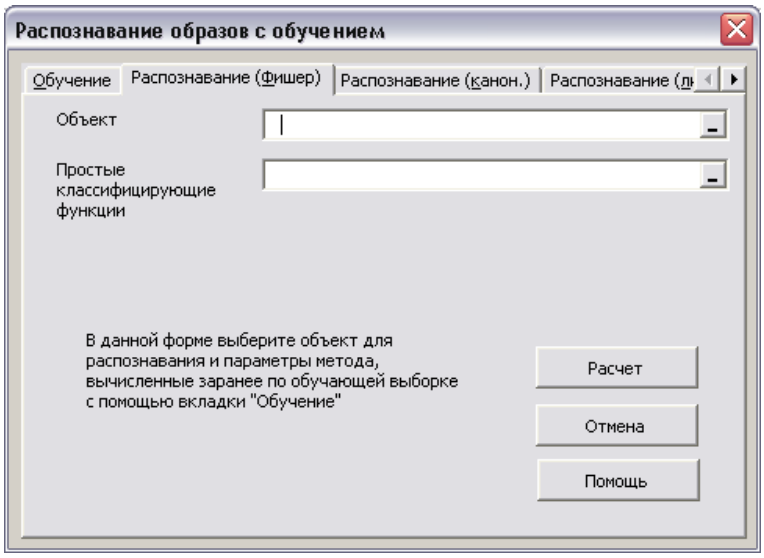


Диалоговое окно содержит набор закладок, что вызвано спецификой решаемой задачи, и обеспечивает как обучение на основе обучающей выборки, так распознавание вновь введенных объектов всеми применяемыми методами. Подобная организация интерфейса гарантирует компактное представление возможностей программы и удобство пользователя при работе с ней.
Режим «Обучение» обеспечивается одной и той же закладкой для всех представленных в

программе методов расчета. Представление закладки «Обучение» см. на рисунке выше.
Для решения задачи обучения:

- Выберите или введите интервалы обучающей выборки. Объекты описываются своими параметрами, и нами принято для определенности, что объекты располагаются по вертикали (в строках), а параметры по горизонтали (в столбцах). Например, один объект, описываемый 4 параметрами, занимает 4 ячейки строки электронной таблицы.
- Выберите или введите интервалы номеров классов (для методов дискриминантного анализа) или оценок (для множественной или логистической регрессии). Численность данного массива должна равняться числу объектов и не содержать пропусков. Например, если имеется в наличии 6 классов, то каждый объект будет принадлежать одному из 6–ти классов: 1, 2, 3, 4, 5 или 6. Ситуация, когда какой–либо класс не содержит объектов (например, класс 2 не присутствует в списке), недопустима.
- Выберите или введите интервал вывода результатов. Начиная с первой ячейки выходного интервала (следовательно, можно указать только одну ячейку, т.к. остальные ячейки интервала игнорируются), будут выведены результаты вычислений.
- Выберите или оставьте по умолчанию метод анализа.

Нажмите кнопку Расчет.
После выполнения вычислений будет, начиная с первой ячейки выходного интервала, выведено название метода и результаты обучения.
Режим «Распознавание» имеет отдельную закладку для каждого из методов распознавания. Ниже описаны приемы практической работы в одном из вариантов режима «Распознавание».



Для решения задачи распознавания методом линейного дискриминантного анализа Фишера:

- Выберите или введите интервал параметров объекта. Объект описывается своими параметрами, число которых должно быть тем же, что и при обучении.
- Выберите или введите интервалы простых классифицирующих функций, вычисленных на этапе обучения.
- Нажмите кнопку Расчет.

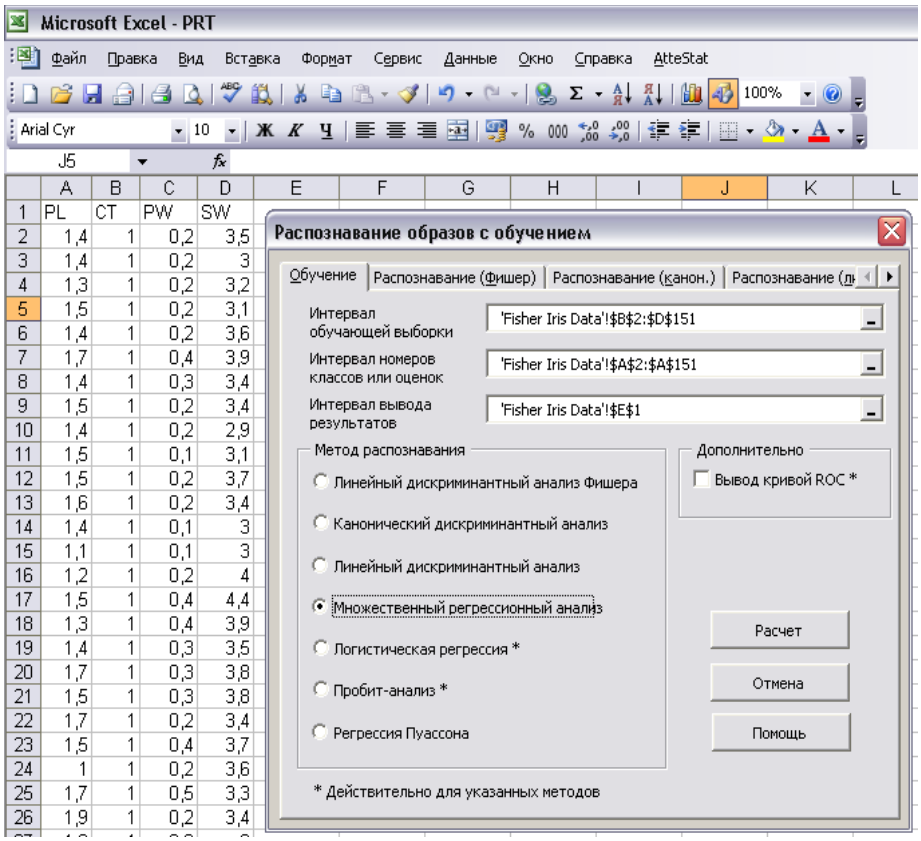
После выполнения расчета будет выдано информационное окно, содержащее наиболее вероятный номер класса, которому принадлежит объект, и вероятность принадлежности объекта этому классу.

Программное обеспечение берет на себя верификацию исходных данных, включая проверку

соответствия всех параметров друг другу для всех используемых методов, выдавая подробную диагностику. При неверных действиях пользователя выдаются сообщения об ошибках.

14.2.1. Пример применения

В качестве примера рассмотрим классический массив данных, называемых «Ирисы Фишера». Эти исходные данные стандартно применяют для тестирования различных алгоритмов распознавания. Полный набор данных можно взять в «Википедии».
Итак, имеются данные измерений для 150 экземпляров ирисов, в равных частях (по 50 штук) принадлежащих к трем видам (iris setosa, iris versicolor, iris virginica). Для каждого экземпляра ириса известны 4 величины: длина чашелистика (Sepal Length), ширина чашелистика (Sepal Width), длина лепестка (Petal Length), ширина лепестка (Petal Width). Входной файл состоит из 150 строк (по 50 для каждого сорта).
Следуя монографии Фон Ай (von Eye) с соавт. (с. 50), построим множественную линейную регрессию зависимости параметра PL (Petal Length) от SW (Sepal Width) и PW (Petal Width). Через CT (Constant Term) обозначим свободный член (константу в уравнении регрессии). Ввод данных стандартный. Отметим только, что для построения модели со свободным членом используется стандартный прием – введение дополнительного регрессора, имеющего значение 1 в каждом наблюдении. Перед выполнением расчета со всеми выбранными параметрами экран компьютера выглядит примерно так, как показано на фрагменте снимка с экрана.



После выполнения расчета экран будет выглядеть примерно так, как показано ниже на фрагменте снимка с экрана.

	A	B	C	D	E	F	G	H	I	J	K
1	PL	CT	PW	SW	Число объектов обучающей выборки						
2	1,4	1	0,2	3,5	150						
3	1,4	1	0,2	3	Число параметров						
4	1,3	1	0,2	3,2	3						
5	1,5	1	0,2	3,1	Множественный линейный регрессионный анализ						
6	1,4	1	0,2	3,6	Число параметров меньше числа объектов.						
7	1,7	1	0,4	3,9	Выполняется решение переопределенной системы.						
8	1,4	1	0,3	3,4	Среднее квадратичное отклонение						
9	1,5	1	0,2	3,4	5,545716						
10	1,4	1	0,2	2,9	Сумма квадратичных остатков						
11	1,5	1	0,1	3,1	30,75497						
12	1,5	1	0,2	3,7	Дисперсия						
13	1,6	1	0,2	3,4	0,209217						
14	1,4	1	0,1	3	Порог Кука						
15	1,1	1	0,1	3	0,027211						
16	1,2	1	0,2	4	Порог DFFITS						
17	1,5	1	0,4	4,4	0,282843						
18	1,3	1	0,4	3,9	Порог DFBETAS						
19	1,4	1	0,3	3,5	0,164957						
20	1,7	1	0,3	3,8	Коэффици	Дисперси	t-статисти	P-значение			
21	1,5	1	0,3	3,8	2,258164	0,098294	7,202639	2,84E-11			
22	1,7	1	0,2	3,4	2,155611	0,002791	40,8039	0			
23	1,5	1	0,4	3,7	-0,35503	0,008535	-3,84295	0,00018			
24	1	1	0,2	3,6	Номер	Наблюден	Предсказа	Остаток	Показател	Стандарти	Стьюдент
25	1,7	1	0,5	3,3	1	1,4	1,446665	-0,04666	0,020426	-0,10202	-0,10273

Все рассчитанные параметры совпадают с источником. Кроме того, выводится ряд дополнительных результатов, интерпретация которых дана в соответствующем разделе.

14.2.2. Сообщения об ошибках

При ошибках ввода данных могут выдаваться диагностические сообщения следующих указанных типов.

Ошибка	Комментарий
Не определен интервал переменной.	Вы не выбрали или неверно ввели интервал эмпирической выборки. Лучшим способом избежать ошибки является не ввод, а выделение интервала стандартным образом, т. е. протаскиванием курсора.
Пустая ячейка в области данных.	Проверьте исходные данные и заполните все ячейки, отмеченные Вами как входной интервал. Для избежания ошибок расчета, вызванных разногласиями, трактовать ли пустую ячейку как нуль, программное обеспечение требует заполнения всех ячеек. Если в ячейке не должно быть данных по физической природе исследуемого процесса, введите в данную ячейку нуль.
Нечисловой тип данных.	Проверьте типы ячеек входного интервала. Тип может быть только числовым. Проще всего выделить интервал ячеек и явно определить их тип как числовой стандартными средствами.
Не определена область вывода.	Не выбран или неверно введен выходной интервал. Лучшим способом избежать ошибки является не ввод, а выделение интервала стандартным образом.
Несоответствие числа объектов.	Число объектов обучающей выборки не соответствует численности массива классов.
Пропущен номер	Номера классов содержат один или несколько пропущенных номеров.

класса.	Введите номера классов без пропусков, начиная с 1. Например, классы могут быть следующие: 1, 2, 3, 4, 5. Но ни в коем случае не 1, 2, 3, 5, 6. В последнем случае пропущен класс номер 4. Иначе – класс номер 4 не содержит объектов.
Несоответствие числа параметров.	Число параметров распознаваемого объекта не соответствует параметрам распознавания, вычисленным на этапе обучения. Число параметров распознаваемого объекта должно равняться числу параметров объекта из обучающей выборки. Кроме того, все массивы, вычисленные при обучении, должны быть выбраны точно. Для минимизации ошибок при выборе интервалов ячеек электронной таблицы рекомендуется метод протаскивания курсора.
Мало коэффициентов.	Число коэффициентов множественной или логистической регрессии должно соответствовать числу параметров распознаваемого объекта.
Неверная оценка.	При использовании метода логистической регрессии оценки имеют бинарный характер. Они должны состоять только из нулей и единиц.
Ошибка в вычислениях.	Применяемый алгоритм вызвал ошибку периода выполнения. Вероятной причиной ошибки может быть особенность исходных данных. Например, данную ошибку вызывает столбец матрицы данных, обладающий очень малой дисперсией. Для локализации ошибки возможно применение методов исследования мультиколлинеарности, представленных в главе «Матричная и линейная алгебра»

14.3. Теоретическое обоснование

Дискриминантный анализ представляет собой линейный метод распознавания данных с обучением. Линейным он является потому, что модель метода линейна относительно дискриминантных функций. Пользователь должен задать некоторое число объектов, указав их принадлежность к так называемым обучающим группам (классам, кластерам, популяциям). Поэтому применению методов распознавания обязательно должны предшествовать исследования методами классификации без обучения (кластерного анализа или эмпирической классификации, когда, например, врач на основании своего опыта выполняет отнесение диагноза того или иного пациента к определенному классу). В задачах, построенных на реальных экспериментальных данных, классы могут пересекаться, особенно если обучение производится на основании эмпирической классификации. Пересечение классов обычно ухудшает качество классификации, а наилучшие для данного набора данных результаты распознавания получаются в случае непересекающихся классов.

Методы распознавания образов с обучением используют в качестве обучающих выборок объекты, заранее классифицированные тем или иным способом. Качество процедуры дискриминации определяется вероятностью правильной классификации. Очень хорошие результаты для распознавания образов с обучением дает предварительное применение метода (см. главу «Кластерный анализ») ближней связи, а применение метода *k*–средних для предварительного отнесения объектов классам дает практически 100% качество распознавания для любого метода дискриминантного анализа. Это обусловлено тем, что из рассмотренных нами методов распознавания без обучения только метод *k*–средних на основе выбранной метрики гарантированно строит непересекающиеся кластеры.

Методы распознавания образов с обучением вырабатывают некоторые решающие правила, позволяющие отнести предлагаемые объекты к заданным классам. Решающие правила могут

быть получены:

- в виде простых классифицирующих функций, как это сделано в линейном дискриминантном анализе Фишера,
- в виде дискриминантных функций, как это сделано в каноническом дискриминантном анализе,
- в виде некоторых характеристик (групповая ковариационная матрица, групповой вектор средних и определитель ковариационной матрицы), как это сделано в линейном дискриминантном анализе,
- в виде набора коэффициентов регрессии (без свободного члена или со свободным членом), как это сделано в методе линейного множественного регрессионного анализа, логистической регрессии, пробит анализа или регрессии Пуассона.

14.3.1. Оценка качества моделей

Качество регрессионной модели зависит от ее типа и в литературе оценивается различными способами, в том числе особыми статистическими критериями, некоторые из которых представлены в данном разделе.

14.3.1.1. Количественные классификаторы

Наиболее простым и понятным способом, которым интуитивно пользуются, является оценка качества дискриминации относительным процентным содержанием, иначе числом верно классифицированных объектов обучающей выборки, отнесенным к объему обучающей выборки, выраженным в процентах. Хотя этот подход недостаточно правомерен, но он является единственно возможным для малого объема данных.
Корректно же качество регрессионных моделей принято оценивать так.
Обучающий массив данных (выборок) случайным образом (см. главу «Рандомизация и генерация случайных последовательностей») делится на две части. Одна часть используется для построения модели. Другая часть – для проверки ее качества. Данный подход – стандартный, однако он может применяться, если представленный массив данных имеет достаточно большую (порядка сотен объектов) численность.
Качество построения множественной регрессионной модели удобно оценивать также средним квадратичным отклонением.

14.3.1.2. Бинарные классификаторы

Для оценки качества модели бинарного классификатора (в программе представлены логистическая регрессия и пробит анализ) предложен ряд параметров. Дальнейшие обозначения проще всего пояснить с помощью таблицы 2 x 2, естественной для бинарных откликов.

Модель	Опыт	
	Положительный исход	Отрицательный исход
Положительный исход	T_P	F_P
Отрицательный исход	F_N	T_N

Суть обозначений ясна из первых букв английских терминов:

- True – истинно,
- False – ложно,
- Positive – положительный,
- Negative – отрицательный.

В программе рассчитываются такие параметры качества, как чувствительность и специфичность. Рассмотрим их подробнее.
Чувствительность показывает долю истинно положительных случаев, правильно идентифицированных моделью,

$$Se = \frac{T_P}{T_P + F_N} \cdot 100\%.$$

Специфичность показывает долю истинно отрицательных случаев, правильно идентифицированных моделью,

$$Sp = \frac{T_N}{T_N + F_P} \cdot 100\%.$$

Программа рассчитывает оптимальные величины порога отсечения двумя наиболее широко применяемыми методами. При этом используются два критерия:

- Для метода 1 достигается минимум величины $|Se - Sp|$,
- Для метода 2 достигается максимум величины $Se + Sp$. Данный критерий предложен Юденом (Youden).

Выбор того или иного оптимального порога отсечения (а также любого другого желаемого порога, в том числе предлагаемой некоторыми авторами величины 0,5) производится на основе требований, предъявляемых исследователем к прогностическим характеристикам модели.
По желанию пользователя выводятся таблица и график так называемой ROC кривой (Receiver Operating Characteristic Curve), отображающей величины *Se* и 1 – *Sp* в зависимости от порога отсечения (параметрическая кривая) и стандартно применяемой для оценки бинарных классификаторов. Объективную оценку качества модели может показать также площадь под ROC кривой, в литературе называемая как AUC (Area Under Curve).

О сравнении ROC кривых см. статьи Вергара (Vergara) с соавт., Хэнли (Hanley) с соавт. О пороге отсечения см. статью Клотше (Klotsche) с соавт.

14.3.2. Оценка значимости модели

В данном разделе представлены методы оценки значимости бинарного классификатора.

14.3.2.1. Статистика Вальда

Статистическая значимость весовых коэффициентов бинарного классификатора может проверяться с помощью статистик Вальда, хотя этот способ и не рассматривается некоторыми авторами как абсолютно надежный. Статистики Вальда записываются как

$$W_i = \frac{\left|\hat{b}_i\right|}{\sqrt{Var\,(\hat{b}_i)}}, i = 1,2,...,m,$$

где $\hat{b}_i, i = 1,2,...,m,$ – вычисленные оценки весовых коэффициентов,
$Var\,(\hat{b}_i), i = 1,2,...,m,$ – дисперсии оценок весовых коэффициентов.
m – количество измеряемых в эксперименте параметров объекта.
Дисперсии оценок находятся как диагональные члены матрицы $I^{-1}(B)$, обратной к информационной матрице Фишера $I(B)$. Информационная матрица Фишера в данном случае представляет собой матрицу, элементы которой являются взятыми с обратным знаком элементами матрицы Гессе функции максимального правдоподобия (далее – ФМП)
$I(B) = – H(B)$,

где $B = \{b_i\}$, $i = 1,2,...,m$ – вектор–столбец весовых коэффициентов.
Статистики Вальда имеют стандартное нормальное распределение, что позволяет установить значимость вычисленных оценок весовых коэффициентов модели.

14.3.2.2. Статистика G

Более надежный способ оценки значимости весовых коэффициентов бинарного классификатора основан на статистиках

$$G_i = -2\ln\frac{L_i}{L}, i = 1,2,...,m,$$

где L_i, $i = 1,2,...,m$ – ФМП системы с исключенным параметром i,
L – ФМП полной системы представленных данных.
Данный метод не представлен в программе, однако может быть сконструирован пользователем самостоятельно на основе реализованных методов путем простой манипуляции с исходными данными.
Статистики G_i, $i = 1,2,...,m$, имеют χ^2 распределение со степенью свободы 1.

Относительно представленных и иных способов оценки значимости весовых коэффициентов бинарного классификатора см. Хосмер (Hosmer), Давнис с соавт.

14.3.3. Линейный дискриминантный анализ Фишера

Метод линейного дискриминантного анализа (линейная дискриминация Фишера, дискриминаторный анализ) предложен Фишером, который предположил, что классификация должна проводиться с помощью линейной комбинации дискриминантных (различающих) переменных. Основанием отнесения объекта к кластеру (классу, популяции) является наибольшее значение так называемой простой классифицирующей функции h_k для k–го класса, являющейся линейной комбинацией дискриминантных переменных:

$$h_k = b_{k0} + \sum_{i=1}^{p} b_{ki} X_i,$$

где p – число дискриминантных переменных,
b_{ki} – коэффициент для i–й переменной k–го класса, определяемый как

$$b_{ki} = (n_{\cdot} - g)\sum_{j=1}^{p} a_{ij} X_{jk\cdot},$$

где $n_{\cdot}$ – общее число наблюдений по всем классам,
a_{ij} – элементы матрицы, обратной к матрице W разброса внутри классов (внутригрупповая матрица сумм попарных произведений), вычисляемой по формуле

$$w_{ij} = \sum_{k=1}^{g}\sum_{m=1}^{n_k}\left(X_{ikm} - X_{ik\cdot}\right)\left(X_{jkm} - X_{jk\cdot}\right),$$

где g – число классов,
n_k – число наблюдений в k–м классе,
X_{ikm} – значение i–ой дискриминантной переменной (величина i–й переменной m–го наблюдения k–го класса),
$X_{jk\cdot}$ – среднее i–й переменной k–го класса.
В основе метода лежат два предположения:

1. Популяции, среди которых производится дискриминация, подчиняются многомерному нормальному распределению. Данное предположение проверяется с помощью методов главы «Проверка нормальности распределения».

2. Популяции, среди которых производится дискриминация, имеют статистически неразличимые ковариационные матрицы.

При искусственном объявлении ковариационных матриц статистически неразличимыми могут оказаться отброшенными наиболее важные индивидуальные черты, имеющие большое значение для хорошей дискриминации. Однако введенное предположение позволяет получить решение и в случае, когда количество обучающих выборок в кластере оказывается меньшим количества дискриминантных функций – то есть при тех условиях, когда более точный линейный дискриминантный анализ не работает.
Результаты линейного дискриминантного анализа Фишера совпадают в смысле качества классификации с результатами более сложного в реализации канонического дискриминантного анализа.

14.3.4. Канонический дискриминантный анализ

Канонический дискриминантный анализ основан на определении так называемых дискриминантных функций, количество которых меньше либо равно числу параметров объектов:

$$f_{km} = u_o + \sum_{i=1}^{p} u_i X_{ikm},$$

где f_{km} – значение канонической дискриминантной функции для m–го объекта k–го класса,
u_i – коэффициенты, определяемые по формуле

$$u_i = v_i \sqrt{n_{\cdot} - g}, u_0 = -\sum_{i=1}^{p} u_i X_{i\cdot\cdot},$$

где $X_{i\cdot\cdot}$ – среднее i–й переменной по всем классам,
v_i, $i = 1,2,...,p$ – коэффициенты, вычисляемые как компоненты собственных векторов решения обобщенной проблемы собственных значений:
$B\nu = \lambda W\nu$,
где B – межгрупповая сумма квадратов отклонений,
ν – собственный вектор,
остальные обозначения те же, что и в предыдущем разделе.
Матрица B определяется как
$B = T – W$,
где T – матрица сумм квадратов и попарных произведений, элементы которой вычисляются как

$$t_{ij} = \sum_{k=1}^{g} \sum_{m=1}^{n_k} \left(X_{ikm} - X_{i\cdot\cdot} \right) \left(X_{jkm} - X_{j\cdot\cdot} \right).$$

Отнесение новых, неклассифицированных объектов к заданным кластерам производится после вычисления дискриминантных функций на основе Евклидовой метрики.
Количество дискриминантных функций в каноническом дискриминантном анализе может быть равным или меньшим количества параметров, описывающих каждый объект.
Результаты распознавания методом канонического дискриминантного анализа совпадают с результатами линейного дискриминантного анализа Фишера.

14.3.5. Линейный дискриминантный анализ

Недостатком описанного метода линейной дискриминации Фишера является предположение о равенстве ковариационных матриц рассматриваемых выборок, вследствие чего могут оказаться отброшенными важные индивидуальные черты, имеющие большое значение для хорошей дискриминации. В методе линейного дискриминантного анализа, напротив,

ковариационные матрицы для различных классов считаются различными.
Отказ от предположения о статистической неразличимости ковариационных матриц, лежащего в основе линейной дискриминации Фишера, для обучающих кластеров не позволяет получить решение в случае, когда количество обучающих выборок в кластере оказывается меньше количества дискриминантных функций. В этом случае программа выдаст сообщение об ошибке при вычислении, а пользователю придется применить линейный дискриминантный анализ Фишера или канонический дискриминантный анализ.
В рассматриваемом методе основанием отнесения объекта к классу является наибольшее значение для данного объекта функции плотности нормального распределения среди всех классов. Вектор средних значений, входящих в формулу функции плотности нормального распределения, а также дисперсионно–ковариационная матрица для каждого обучающего класса оцениваются по исходным данным на этапе обучения.
Отсутствие простых решающих правил (для получения результата нужно проделать довольно объемные вычисления) было некоторым препятствием для широкого применения этого мощного метода в период, предшествовавший распространению персональных компьютеров. Метод использовался фактически редко, но здесь введен нами как серьезная альтернатива линейному дискриминантному анализу Фишера. В наших расчетах метод линейного дискриминантного анализа показал более высокое качество распознавания по сравнению с линейной дискриминацией Фишера и каноническим дискриминантным анализом.

14.3.6. Линейный множественный регрессионный анализ

В ходе нетривиального эксперимента абсолютно точные измерения параметров, как правило, невозможны. Чтобы уменьшить влияние ошибок, производится большое число измерений. Каждое измерение дает нам уравнение с известной из теоретических соображений структурой с точностью до коэффициентов, подлежащих определению. При числе измерений большем, меньшем или равном числу параметров, мы приходим к необходимости решения системы, число уравнений которой больше, меньше либо равно числу неизвестных параметров, соответственно.
Для решения задачи в первом приближении положим зависимость результата эксперимента от параметров линейной модели (примем линейную модель) и сформулируем задачу математически следующим образом. Требуется решить систему линейных уравнений:

$$\sum_{j=1}^{k} a_{ij} x_j \approx b_i, i = 1,2,...,n,$$

(в поэлементной записи) либо
$AX = B$ (в матричной записи),
где a_{ij}, $i = 1,2,...n$; $j = 1,2,...k$ – элемент матрицы экспериментальных данных A (матрицы регрессоров) размером n x k,
x_j, $j = 1,2,...k$ – элемент подлежащего определению вектора весовых коэффициентов X длиной k,
b_i, $i = 1,2,...n$ – элемент вектора результатов эксперимента B длиной k,
k – количество измеряемых в эксперименте параметров,
n – количество опытов.
В формуле знак ≈ применен вместо знака равенства, чтобы подчеркнуть неточность определения результата эксперимента.
Случай 1. При $k = n$ система в случае, если матрица системы не вырождена, имеет одно решение, поиск которого не вызывает трудностей и может быть осуществлен методом Гаусса или одним из итерационных методов.
Случай 2. При $k > n$ система имеет бесчисленное множество решений. Ранг n матрицы

системы меньше порядка системы. Число линейно независимых уравнений меньше количества неизвестных, поэтому возникает неопределенная (недоопределенная) система линейных уравнений порядка *k*.
Если систему удается решить, в общем случае полученный вектор решения системы уравнений не будет точно удовлетворять ни одному уравнению системы, однако можно получить решение, в каком–то смысле наилучшее. Существует бесчисленное множество решений рассматриваемой системы, однако из них можно выделить одно решение, наложив на систему дополнительные условия. Так, можно найти решение, обладающее минимальной Евклидовой нормой
$||B - AX||_2 = \min$.
Размерность матрицы *A* суть *n* x *k*, причем $n < k$. Сначала нужно так преобразовать запись системы, чтобы *k* и *n* поменять местами, то есть создать предпосылки для поиска псевдообратной к A^T матрицы. После элементарных выкладок с применением свойств псевдообратной матрицы получаем формулу для нахождения решения, обладающего минимальной нормой, в виде

$$\hat{X}^T = B^T \left(A^T\right)^+,$$

где $\hat{X}$ – вектор оценок весовых коэффициентов.
Случай 3. Данный случай $k < n$ является наиболее важным практически. Система при этом является несовместной и может быть решена приближенно.
Ранг *n* матрицы системы больше порядка системы. Число линейно независимых уравнений больше количества неизвестных, поэтому возникает переопределенная система линейных уравнений порядка *k*.
Для решения системы достаточно домножить левую и правую части уравнения на матрицу A^T слева. Затем домножить левую и правую части полученного уравнения на матрицу $(A^TA)^{-1}$ также слева. В результате получим готовую формулу решения

$$\hat{X} = (A^T A)^{-1} A^T B.$$

В поэлементной записи решение системы может быть представлено следующим образом. Введем вектор ошибок ε_j, $j = 1,2,...,n$. Тогда исходную систему можно переписать:

$$\sum_{j=1}^{k} a_{ij} x_j + \varepsilon_i = b_i, i = 1,2,...,n.$$

Точное решение системы получить не удается, поэтому обычно применяют метод наименьших квадратов, в котором минимизируют сумму квадратов ошибок ε_i, $i = 1,2,...,n$. Составим квадратичный функционал

$$I(x_1, x_2, ..., x_k) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} \left[b_i - \sum_{j=1}^{k} a_{ij} x_j \right]^2.$$

Потребуем минимума функционала $I = I(x_1,x_2,...,x_k)$ по элементам вектора x_j, $j = 1,2,...,k$:

$$I(x_1, x_2, ..., x_k) \underset{x_j}{\rightarrow} \min, j = 1,2,...,k.$$

Данное требование удовлетворяется в случае равенства нулю всех частных производных функционала *I* по элементам x_j, $j = 1,2,...,k$, что приводит к системе *k* линейных алгебраических уравнений, решив которые, мы найдем неизвестные величины x_j, $j = 1,2,...,k$:

$$\frac{\partial I}{\partial x_j} = 0, \; j = 1,2,...,k.$$

Подставив в последнюю формулу выражение для функционала, после элементарных преобразований получим:

$$\sum_{j=1}^{l}\left[\sum_{i=1}^{n} a_{ij}a_{il}\right] x_j = \sum_{i=1}^{n} b_i\, a_{il}, l=1,2,...,k.$$

Это выражение представляет собой запись системы k линейных алгебраических уравнений для k неизвестных. Выражение в квадратных скобках – это запись элемента матрицы системы уравнений, а правая часть формулы – запись элемента столбца свободных членов.
Лучшие результаты в смысле более точного описания реального объекта исследований часто, хотя и не всегда, можно получить, если использовать уравнения со свободным членом:

$$x_0 + \sum_{j=1}^{k} a_{ij}x_j + \varepsilon_i = b_i, i = 1,2,...,n.$$

Тогда остальные уравнения примут, соответственно, вид:

$$I(x_0, x_1, x_2, ..., x_k) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n}\left[b_i - x_0 - \sum_{j=1}^{k} a_{ij}x_j\right]^2,$$

$$I(x_0, x_1, x_2, ..., x_k) \underset{x_j}{\rightarrow} \min,\ j = 0,1,...,k,$$

$$\frac{\partial I}{\partial x_j} = 0,\ j=0,1,...,k,$$

$$nx_0 + \sum_{j=1}^{k}\left[\sum_{i=1}^{n} a_{ij}\right] x_j = \sum_{i=1}^{n} b_i,$$

$$\sum_{i=1}^{n} a_{il}x_0 + \sum_{j=1}^{l}\left[\sum_{i=1}^{n} a_{ij}a_{il}\right] x_j = \sum_{i=1}^{n} b_i\, a_{il}, l=1,2,...,k.$$

Последнее выражение будет системой $k + 1$ линейных алгебраических уравнений для $k + 1$ неизвестных. Практически, однако, в применении последней формулы нет необходимости. Свободный член может быть введен, не меняя структуры уравнений регрессии (для всех случаев). При расчете достаточно ввести дополнительный ($k + 1$)–й параметр, которому будет соответствовать регрессор 1 в каждом векторе данных. Для определенности, а именно такой порядок коэффициентов принят в блоке распознавания представленной программы, следует добавить единичный вектор в качестве первого столбца матрицы данных. Тогда первым коэффициентом в столбце коэффициентов регрессии, выводимом в рассматриваемом случае, как раз и будет вычисленный свободный член.
Программой также выводятся дисперсии $D\hat{x}_j,\ j = 1,2,...,k,$ оценок коэффициентов $\hat{x}_j,\ j = 1,2,...,k,$ являющиеся диагональными элементами дисперсионно–ковариационной матрицы

$$C(\hat{X}) = MSE \cdot (A^T A)^{-1},$$

где MSE – средняя квадратичная ошибка (дисперсия ошибки регрессии), вычисляемая по формуле

$$MSE = \frac{1}{n-k}\sum_{i=1}^{n} e_i^2,$$

где e_i, $i = 1,2,...,n$ – остатки, вычисляемые как

$$e_i = b_i - \hat{b}_i, i = 1,2,...,n,$$

где $\hat{b}_i, i = 1,2,...,n,$ – предсказанные результаты эксперимента.
Зная оценки коэффициентов и их дисперсии, можно вычислить статистики

$$t_j = \frac{\hat{x}_j}{\sqrt{D\hat{x}_j}}, j = 1,2,...,k,$$

которые асимптотически имеют *t*–распределение с *n* – *k* степенями свободы, что позволяет сделать вывод о значимости оценок коэффициентов (значимом отличии от нуля).
Стандартное отклонение ошибки регрессии вычисляется как

$$SE = \sqrt{MSE}.$$

В дальнейших вычислениях понадобятся также стандартные ошибки остатков

$$m_i = SE\sqrt{1-h_i}, i = 1,2,...,n,$$

где h_i, i = 1,2,...,*n* – показатели влияния (leverage, условные корреляции наблюдения и прогноза), вычисляемые как модули диагональных элементов матрицы

$$H = A(A^T A)^{-1} A^T.$$

Стандартизованные остатки (standardized residual) вычисляются как

$$E_i = \frac{e_i}{SE}, i = 1,2,...,n.$$

Считается, что более точные оценки стандартизованных остатков дают выводимые программой стьюдентизированные остатки (studentized deleted residual), которые вычисляются как

$$E_i^* = \frac{e_i}{m_{(i)}}, i = 1,2,...,n,$$

где индекс (*i*) здесь и далее означает, что вычисление показателя произведено при исключенном наблюдении с номером *i*.
В литературе выведена простая функциональная связь величин E_i, i = 1,2,...,*n*, и $E_i^*, i = 1,2,...,n,$ поэтому последние можно вычислить через первые, теоретически намного уменьшив объем вычислений. Практически, однако, в этом нет необходимости, т.к. указанные вычисления все–таки необходимы для оценки других параметров решения.

14.3.6.1. Обработка выбросов

Анализ стандартизованных или стьюдентизированных (в программе) остатков может применяться для выявления выбросов наблюдений относительно статистической модели.
Стьюдентизированные остатки $E_i^*, i = 1,2,...,n,$ асимптотически подчиняются *t*–распределению с *n* – *k* степенями свободы. Для удобства пользователей *P*–значения, не превышающие принятый для данного типа задачи стандартный порог 0,05 (что свидетельствует о значимости различий наблюдения и модельной оценки), выделяются красным цветом аналогично тому, как это сделано в главе «Обработка выбросов».

14.3.6.2. Выявление влияющих наблюдений

Для каждого наблюдения выводится мера Кука (Cook's distance, Cook's measure)

$$Q_i = \frac{E_i^2}{k}\frac{h_i}{1-h_i}, i = 1,2,...,n.$$

Значения $Q_i > \frac{4}{n-k}, i = 1,2,...,n,$ т. е. превышающие порог тревожности, свидетельствуют о сильном влиянии данного наблюдения на смещение гиперплоскости регрессии. Данные экстремальные значения выделяются в выводе программы синим цветом, чтобы можно было

визуально идентифицировать влияющее наблюдение.
Также для каждого наблюдения выводится значение меры Велча–Куха (Welsch–Kuh's distance, Welsch–Kuh's measure, *DFFITS*, *DFITS*)

$$DFFITS_i = E_i^* \sqrt{\frac{h_i}{1-h_i}}, i = 1,2,...,n.$$

Значения $|DFFITS_i| > 2\sqrt{\frac{k}{n}}, i = 1,2,...,n,$ т. е. превышающие порог тревожности, свидетельствуют о сильном влиянии данного наблюдения на смещение гиперплоскости регрессии. Данные экстремальные значения выделяются в выводе программы синим цветом, чтобы можно было визуально идентифицировать влияющее наблюдение.
Как уже упоминалось выше, в литературе выведена простая функциональная связь величин E_i, $i = 1,2,...,n$, и $E_i^*, i = 1,2,...,n,$ поэтому меры Кука и Велча–Куха эквивалентны. Это означает, что результаты анализа данными методами в большинстве случаев должны давать одинаковые выводы.
Также для каждого наблюдения выводятся k значений меры *DFBETAS*

$$DFBETAS_{ij} = \frac{\hat{x}_j - \hat{x}_{j(i)}}{SE_{(i)}\sqrt{c_j}}, i = 1,2,...,n; j = 1,2,...,k,$$

где c_j, $j = 1,2,...,k$ – диагональные элементы матрицы $C = (A^TA)^{-1}$.

Значения $|DFBETAS_{ij}| > \frac{2}{\sqrt{n-k}}, i = 1,2,...,n; j = 1,2,...,k,$ т. е. превышающие порог тревожности, свидетельствуют о сильном влиянии данного наблюдения на оценки весовых коэффициентов. Данные экстремальные значения выделяются в программе синим цветом, чтобы можно было визуально идентифицировать влияющее на данный коэффициент наблюдение.

14.3.6.3. Автокорреляция остатков

Для оценки автокорреляции остатков множественной линейной регрессии, построенной на основе упорядоченных исходных данных (например, по латентной переменной – времени), разработан критерий Дарбина–Уотсона. Вычисление статистики критерия производится по формуле

$$d = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2}.$$

Эквивалентная (матричная) форма статистики критерия

$$d = \frac{e^T A e}{e^T e},$$

где e – вектор остатков с элементами $e_i, i = 1,2,...,n$,
A – матрица размером n x n вида

$$A = \begin{bmatrix} 1 & -1 & 0 & \dots & 0 & 0 & 0 \\ -1 & 2 & -1 & \dots & 0 & 0 & 0 \\ 0 & -1 & 2 & \dots & 0 & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & 2 & -1 & 0 \\ 0 & 0 & 0 & \dots & -1 & 2 & -1 \\ 0 & 0 & 0 & \dots & 0 & -1 & 1 \end{bmatrix}.$$

Модифицированная статистика критерия d / 4 асимптотически подчиняется бета–распределению с параметрами, определяемыми следующим образом.
Сначала вычисляются, соответственно, математическое ожидание и дисперсия статистики Дарбина–Уотсона по формулам

$$Ed = \frac{P}{n-k-1},$$

$$Dd = \frac{2(Q - P \cdot Ed)}{(n-k-2)(n-k+1)},$$

где вспомогательные величины P и Q вычисляются, соответственно, как
$P = trA - \mathrm{tr}[X^T A X (X^T X)^{-1}]$,
$Q = trA^2 - 2tr[X^T A^2 X (X^T X)^{-1}] + tr\{[X^T A X (X^T X)^{-1}]^2\}$.
Идея состоит в том, что математическое ожидание и дисперсия статистики d / 4 должны быть равны, соответственно, математическому ожиданию и дисперсии функции бета–распределения. Функция бета–распределения с параметрами p и q имеет математическое ожидание и дисперсию, соответственно,

$$Eb = \frac{4p}{p+q},$$

$$Db = \frac{16pq}{(p+q)^2(p+q+1)}.$$

Приравнивая соответствующие выражения для Ed и Eb, а также Dd и Db, получим

$$(p+q) = \frac{Ed(4-Ed)}{Dd} - 1,$$

$$p = \frac{1}{4}(p+q)Ed.$$

После вычисления из второго выражения величины p можно вычислить q из первого выражения как

$$q = \frac{Ed(4-Ed)}{Dd} - 1 - p.$$

См. монографии Аллисона (Allison), Белсли (Belsley) с соавт., Дрейпера (Draper) соавт., Коэна (Cohen) с соавт., Кука (Cook) соавт., Райана (Ryan), Фон Ай (von Eye) с соавт., Уусипайкка (Uusipaikka), Чаттерджи (Chatterjee) с соавт., доклад Галченковой. О выбросах и влияющих наблюдениях см. монографии Кука (Cook) с соавт., Смита (Smith), Руссо (Rousseeuw) с соавт., статьи Кука. О статистике Дарбина–Уотсона см. статью Дарбина (Durbin) с соавт., монографию Хеннана.

14.3.7. Логистическая регрессия

В практических приложениях возникает ситуация, когда отклик эксперимента представлен в бинарном виде (1 – наличие признака, 0 – отсутствие признака). Множественная линейная регрессия не учитывает данное ограничение на выход модели. Для решения задачи может использоваться логит анализ (множественная логистическая регрессия).
Множественная логистическая регрессия может быть представлена в виде следующей модельной формулы

$$P_j(B) = Logit(X_j B) = \frac{1}{1+e^{-X_j B}},\ j = 1,2,...,n,$$

где $P_j(B)$, $i = 1,2,...,n$ – выход модели,
$B = \{b_i\}$, $i = 1,2,...,m$ – вектор–столбец весовых коэффициентов,
$X_j = \{x_i\}$, $i = 1,2,...,m$; $j = 1,2,...,n$ – вектор–строка параметров объекта j, измеренных в эксперименте,
X_jB, $j = 1,2,...,n$ – множественная линейная регрессия (замечание о модели со свободным членом см. ниже),
m – количество измеряемых в эксперименте параметров объекта,
n – численность обучающей выборки (число объектов).
Значение $P_j(.)$ может быть интерпретировано как вероятность получения логитом значения 1 при подстановке в уравнение определенного вектора X_j, $j = 1,2,...,n$, измеренного в эксперименте.
Оптимальные значения весовых коэффициентов могут быть найдены путем максимизации логарифмической функции максимального правдоподобия (далее – ФМП)

$$\ln L = \sum_{j=1}^{n} \left[Y_j \ln P_j(B) + (1 - Y_j) \ln(1 - P_j(B)) \right],$$

где Y_j, $j = 1,2,...,n$ – выход эксперимента, соответствующий измеренному в эксперименте вектору параметров X_j, $j = 1,2,...,n$.
Задача сводится к системе нелинейных алгебраических уравнений

$$\frac{\partial \ln L}{\partial B} = 0,$$

для решения которой в программе использован стандартный метод Ньютона–Рафсона.
Итерационная схема метода записывается формулой
$B_{k+1} = B_k - [H(B_k)]^{-1} g(B_k)$, $k = 0,1,...,$
где k – номер итерации,
$H(.)$ – матрица Гессе (матрица вторых производных) ФМП,
$g(.)$ – градиент (вектор производных) ФМП.
Вследствие аналитически доказанной сходимости итерационной схемы на всей области определения аргумента вектор начальных значений можно брать произвольным, для определенности и удобства вычислений – нулевым.
Решение задачи радикально упрощается благодаря известным выражениям вектора градиента и матрицы Гессе ФМП. Градиент ФМП имеет явное представление в виде (опуская номер итерации)

$$g(B) = \sum_{j=1}^{n} (Y_j - P_j(B)) X_j.$$

Матрица Гессе ФМП имеет явное представление в виде (опуская номер итерации)

$$H(B) = -\sum_{j=1}^{n} P_j(B)(1 - P_j(B)) X_j^T X_j.$$

Оценка качества регрессионной модели описана в одноименном разделе. Оценка значимости весовых коэффициентов также представлена.
Некоторыми авторами (Паклин, Хайкин) проводится параллель между логистической регрессией и однослойным персептроном из математической теории нейронных сетей. В теории нейронных сетей предполагается, что вид сигмоидальной функции активации нейрона значения практически не имеет (это не совсем так). Важны лишь такие ее свойства, как ограниченность на участке [0,1] и нелинейность. Собственно, это и объясняет почти полную идентичность результатов логит анализа и пробит анализа. Различие данных методов – лишь в скорости сходимости и объеме вычислений, который для логистической регрессии ниже.
Заметим, что если по условиям задачи требуется логит множественной линейной регрессии со свободным членом, в режиме обучения просто добавьте в массив исходных данных столбец из одних единиц, а при распознавании не забудьте в векторе каждого распознаваемого объекта также установить в данной позиции вектора единицу. При этом количество параметров, выводимое программой, будет показано с учетом данного фиктивного единичного параметра.

См. монографии Хосмер (Hosmer) с соавт., Шукри (Shoukri) с соавт., статьи Давнис с соавт., Дэвис (Davis) с соавт., Цвейг (Zweig) с соавт., Дэйвидсон (Davidson) с соавт., пособие Цыплакова. Введение в рассматриваемый предмет дано в обзорных статьях Паклина на сайте BaseGroup Labs. О решении систем нелинейных уравнений см. книгу Носача. О персептроне см. книги Осовского, Мандик (Mandic) с соавт., Минского с соавт., Круглова с соавт., Дюка с соавт.

14.3.8. Пробит анализ

В практических приложениях возникает ситуация, когда отклик эксперимента представлен в бинарном виде (1 – наличие признака, 0 – отсутствие признака). Множественная линейная регрессия не учитывает данное ограничение на выход модели. Для решения задачи может использоваться пробит анализ.
Пробит может быть представлен в виде следующей модельной формулы

$$P_j(B) = probit(X_j B) = \Phi(X_j B),\ j = 1,2,...,n,$$

где $P_j(B)$, $i = 1,2,...,n$ – выход модели,
$\Phi(.)$ – функция нормального распределения,
$B = \{b_i\}$, $i = 1,2,...,m$ – вектор–столбец весовых коэффициентов,
$X_j = \{x_i\}$, $i = 1,2,...,m$; $j = 1,2,...,n$ – вектор–строка параметров объекта j, измеренных в эксперименте,
$X_j B$, $j = 1,2,...,n$ – множественная линейная регрессия (замечание о модели со свободным членом см. ниже),
m – количество измеряемых в эксперименте параметров объекта,
n – численность обучающей выборки (число объектов).
Значение $P_j(.)$ может быть интерпретировано как вероятность получения пробитом значения 1 при подстановке в уравнение определенного вектора X_j, $j = 1,2,...,n$, измеренного в эксперименте.
Оптимальные значения весовых коэффициентов могут быть найдены путем максимизации логарифмической функции максимального правдоподобия (далее – ФМП)

$$\ln L = \sum_{j=1}^{n} \left[Y_j \ln P_j(B) + (1 - Y_j) \ln(1 - P_j(B)) \right],$$

где Y_j, $j = 1,2,...,n$ – выход эксперимента, соответствующий измеренному в эксперименте

вектору параметров X_j, $j = 1,2,...,n$.
Задача сводится к системе нелинейных алгебраических уравнений

$$\frac{\partial \ln L}{\partial B} = 0,$$

для решения которой в программе использован стандартный метод Ньютона–Рафсона.
Итерационная схема метода записывается формулой
$B_{k+1} = B_k - [H(B_k)]^{-1} g(B_k)$, $k = 0,1,...,$
где k – номер итерации,
$H(.)$ – матрица Гессе (матрица вторых производных) ФМП,
$g(.)$ – градиент (вектор производных) ФМП.
Вследствие аналитически доказанной сходимости итерационной схемы на всей области определения аргумента вектор начальных значений можно брать произвольным, для определенности и удобства вычислений – нулевым.
Решение задачи радикально упрощается благодаря известным выражениям вектора градиента и матрицы Гессе ФМП. Градиент ФМП имеет явное представление в виде (опуская номер итерации)

$$g(B) = \sum_{j=1}^{n} V_j(B) X_j,$$

где $V_j(B) = f(X_j B) \frac{Y_j - P_j(B)}{P_j(B)(1 - P_j(B))}, j = 1,2,...,n,$ – вспомогательные величины,
$f(X_j B)$, $j = 1,2,...,n$ – плотность вероятности нормальной случайной величины.
Матрица Гессе ФМП имеет явное представление в виде (опуская номер итерации)

$$H(B) = -\sum_{j=1}^{n} V_j(B)(V_j(B) + X_j B) X_j^T X_j.$$

Оценка качества регрессионной модели описана в одноименном разделе. Оценка значимости весовых коэффициентов также представлена.
Заметим, что если по условиям задачи требуется пробит со свободным членом, в режиме обучения просто добавьте в массив исходных данных столбец из одних единиц, а при распознавании не забудьте в векторе каждого распознаваемого объекта также установить в данной позиции вектора единицу. При этом количество параметров, выдаваемое программой, будет показано с учетом данного фиктивного единичного параметра.

См. монографию и статью Кэмерон (Cameron) с соавт., работу Анселин (Anselin), монографии Дэвидсона (Davidson) с соавт., Лонга (Long), пособие Цыплакова. О решении систем нелинейных уравнений см. книгу Носача.

14.3.9. Регрессия Пуассона

Регрессия Пуассона является методом распознавания так называемых счетных данных, возникающих при подсчете количества каких–либо экспериментальных сущностей (например, числа бактерий в чашке Петри). Она может быть представлена в виде следующей модельной формулы
$\mu_j(B) = \exp(X_j B)$, $j = 1,2,...,n$,
где $\mu_j(B)$, $j = 1,2,...,n$ – выход модели,
$B = \{b_i\}$, $i = 1,2,...,m$ – вектор–столбец весовых коэффициентов,
$X_j = \{x_i\}$, $i = 1,2,...,m$; $j = 1,2,...,n$ – вектор–строка параметров объекта j, измеренных в эксперименте,

X_jB, $j = 1,2,...,n$ – множественная линейная регрессия (замечание о модели со свободным членом см. ниже),
m – количество измеряемых в эксперименте параметров объекта (регрессоров),
n – численность обучающей выборки (число объектов).
В модели регрессии Пуассона параметр μ интерпретируется как счетное количество, соответствующее вектору регрессоров X. При этом μ является параметром распределения Пуассона, плотность которого имеет вид

$$P(Y = y) = \frac{e^{-\mu}\mu^{y}}{y!}, y = 0,1,2,....$$

Функция максимального правдоподобия (далее – ФМП) запишется как

$$L = \prod_{j=1}^{n} P(Y_j = y) = \prod_{j=1}^{n} \frac{\mu^{Y_j}}{Y_j!} e^{-\mu},$$

где Y_j, $j = 1,2,...,n$ – выход эксперимента, соответствующий измеренному в эксперименте вектору параметров X_j, $j = 1,2,...,n$.
Тогда логарифмическая ФМП будет записана, с учетом модельной формулы регрессии, как

$$\ln L = \sum_{j=1}^{n} \left[-\exp(X_j B) + Y_j X_j B - \ln Y_j! \right].$$

Оптимальные значения весовых коэффициентов могут быть найдены путем максимизации логарифмической ФМП. Задача сводится к системе нелинейных алгебраических уравнений

$$\frac{\partial \ln L}{\partial B} = 0,$$

для решения которой в программе использован стандартный метод Ньютона–Рафсона.
Итерационная схема метода записывается формулой
$B_{k+1} = B_k - [H(B_k)]^{-1} g(B_k)$, $k = 0,1,...,$
где k – номер итерации,
$H(.)$ – матрица Гессе (матрица вторых производных) ФМП,
$g(.)$ – градиент (вектор производных) ФМП.
Вследствие аналитически доказанной сходимости итерационной схемы на всей области определения аргумента вектор начальных значений можно брать произвольным, для определенности и удобства вычислений – нулевым.
Решение задачи радикально упрощается благодаря известным выражениям вектора градиента и матрицы Гессе ФМП. Градиент ФМП имеет явное представление в виде (опуская номер итерации)

$$g(B) = \sum_{j=1}^{n} X_j (Y_j - X_j B).$$

Матрица Гессе ФМП имеет явное представление в виде (опуская номер итерации)

$$H(B) = -\sum_{j=1}^{n} \exp(X_j B) X_j^T X_j.$$

Оценка качества регрессионной модели описана в одноименном разделе. Оценка значимости весовых коэффициентов также представлена.
Если по условиям задачи требуется регрессия Пуассона со свободным членом, в режиме обучения просто добавьте в массив исходных данных столбец из одних единиц, а при распознавании не забудьте в векторе каждого распознаваемого объекта также установить в данной позиции вектора единицу. При этом количество параметров, выдаваемое программой, будет показано с учетом данного фиктивного единичного параметра.

См. монографию и статью Кэмерона (Cameron) с соавт., работу Анселин (Anselin), монографии Дэвидсона (Davidson) с соавт., Лонга (Long), пособие Цыплакова. О решении систем нелинейных уравнений см. книгу Носача. Большое число источников посвящено изучению явления сверхдисперсии (overdispersion), иногда возникающего при исследовании представленным методом и заключающегося в превышении дисперсии модельной оценки регрессии Пуассона над самой оценкой, т. е. $D\lambda > E\lambda$. О сверхдисперсии см. статьи Баррона (Barron), Бонинга (Bohning), Бреслоу (Breslow), Дина (Dean) и Дина с соавт., Дугласа (Douglas).

14.3.10. Оценка прогностической ценности параметров

Сравнительная оценка прогностической ценности параметров (применительно к логистической регрессии) представлена в работе Плавинской с соавт., причем в качестве альтернативы классической *m*–мерной множественной логистической регрессии использованы *m* логистических регрессий для каждого параметра объекта в отдельности. Данный эффективный прием может быть реализован пользователем настоящего программного обеспечения путем простой манипуляции с исходными данными, в том числе и для других родственных методов распознавания.
Можно также упомянуть еще один эффективный способ оценки прогностической ценности, основанный на алгоритме Фаррара-Глаубера, представленном в главе «Матричная и линейная алгебра» (а именно, способ, который основан на вычислении коэффициентов детерминации, там же см. необходимые ссылки). Данный способ исследования мультиколлинеарности может оказаться практически полезным при решении проблемы оценки влияния того или иного параметра, по предположению исследователя, характеризующего объект, в рассмотренных в настоящей главе методах распознавания образов с обучением. Параметры, имеющие значимые коэффициенты детерминации, рекомендуется исключить из рассмотрения как имеющие малое влияние на результаты распознавания. Исключение данных параметров помогает не только сократить объем вычислений и уменьшить вычислительную сложность решения (часто обеспечить саму возможность решения конкретным методом распознавания), но и, что самое важное, интерпретировать результат распознавания образов с привлечением существенно меньшего числа параметров (отбросив параметры, мало влияющие на качество распознавания), т. е. снизить размерность (а следовательно, и стоимость решения) задачи.

Список использованной и рекомендуемой литературы

1. Abreu M.N.S., Siqueira A.L., Caiaffa W.T. Regressao logistica ordinal em estudos epidemiologicos // Revista de Saude Publica, 2009, vol. 43, no. 1, pp. 183–194.
2. Aitchison J., Silvey S.D. The generalization of probit analysis to the case of multiple responses // Biometrika, June 1957, vol. 44, no. 1/2, pp. 131–140.
3. Aldrich J.H., Nelson F.D. Quantitative applications in the social sciences: Vol. 45. Linear probability, logit, and probit models. – Beverly Hills, CA: Sage, 1989.
4. Allison P.D. Multiple regression. – Thousand Oaks, CA: Pine Forge Press, 1999.
5. Altman E.I. Application of classification techniques in business, banking and finance / E.I. Altman, R.B. Avery, R.A. Eisenbeis et al. – Greenwich, CT: JAI Press, 1981.
6. Altman E.I. Financial ratios, discriminant analysis, and the prediction of corporate bankruptcy // Journal of Finance, 1968, vol. 23, pp. 589–609.
7. Altman E.I., Haldeman R.G., Narayanan P. Zeta analysis: A new model to identify bankruptcy risk of corporations // Journal of Banking and Finance, 1977, vol. 1, pp. 29–51.
8. Ananth C.V., Kleinbaum D.G. Regression models for ordinal responses: A review of methods

and applications // International Journal of Epidemiology, 1997, vol. 26, no.6, pp. 1323–1333.
9. Anderson T.W. An introduction to multivariate statistical analysis. – New York, NY: John Wiley & Sons, 1984.
10. Barron D.N. The analysis of count data: Overdispersion and autocorrelation // Sociological Methodology, 1992, vol. 22, pp. 179–220.
11. Bartlett M.S. Properties of sufficiency and statistical tests // Proceedings of the Royal Society of London, 1937, A160, pp. 268–282.
12. Belsley D.A., Kuh E., Welsch R.E. Regression Diagnostics. – New York, NY: John Wiley & Sons, 1980.
13. Bender R., Grouven U. Ordinal logistic regression in medical research // Journal of the Royal College of Physicians of London, September/October 1997, vol. 31, no. 5, pp. 546–551.
14. Bender R., Grouven U. Using binary logistic regression models for ordinal data with non–proportional odds // Journal of Clinical Epidemiology, 1998, vol. 51, no. 10, pp. 809–816.
15. Blanchard G., Geman D. Hierarchical testing designs for pattern recognition // The Annals of Statistics, June 2005, vol. 33, no. 3.
16. Bohning D. A note on a test for Poisson overdispersion // Biometrika, June 1994, vol. 81, no. 2, pp. 418–419.
17. Box G.E.P. A general distribution theory for a class of likelihood criteria // Biometrika, 1949, vol. 36, pp. 317–346.
18. Breiman L., Ihaka R. Nonlinear discriminant analysis via scaling and ace // Technical report No. 40, December 1984, University of California, Berkeley, CA.
19. Breslow N. Tests of hypotheses in overdispersed Poisson regression and other quasi–likelihood models // Journal of the American Statistical Association, June 1990, vol. 85, no. 410, pp. 565–571.
20. Cameron A.C., Trivedi P.K. Regression analysis of count data. – Cambridge, NY: Cambridge University Press, 1998.
21. Cameron A.C., Trivedi P.K. Regression–based tests for overdispersion in the Poisson model // Journal of Econometrics, December 1990, vol. 46, no. 3, pp. 347–364.
22. Chatfield C., Collins A. Introduction to multivariate analysis. – New York, NY: Chapman & Hall / CRC, 2000.
23. Chatterjee S., Hadi A.S. Regression analysis by example. – New York, NY: John Wiley & Sons, 2006.
24. Chatterjee S., Hadi A.S. Sensitivity analysis in linear regression. – New York, NY: John Wiley & Sons, 1988.
25. Clarke W.R., Lachenbruch P.A., Broffit B. How nonnormality affects the quadratic discrimination function // Communications in Statistics – Theory and Methods, 1979, A8, pp. 1285–1301.
26. Cohen J. Applied multiple regression/correlation analysis for the behavioral sciences / J. Cohen, P. Cohen, S.G. West et al. – Mahwah, NJ: Lawrence Erlbaum Associates, 2003.
27. Cook R.D. Detection of influential observations in linear regression // Technometrics, 1977, vol. 19, no. 1, pp. 15–18.
28. Cook R.D. Influential observations in linear regression // Journal of the American Statistical Association, 1979, vol. 74, pp. 169–174.
29. Cook R.D., Weisberg S. Applied regression including computing and graphics. – New York, NY: Wiley, 1999.
30. Cook R.D., Weisberg S. Residuals and influence in regression. – London: Chapman & Hall, 1982.
31. Cooley W.W., Lohnes P.R. Multivariate data analysis. – New York, NY: John Wiley & Sons,

1971.
32. Davidson R., MacKinnon J.G. Estimation and inference in econometrics. – Oxford, NY: Oxford University Press, 1993.
33. Davis J., Goadrich M. The relationship between precision–recall and ROC curves // Proceedings of the 23rd International Conference of Machine Learning, Pittsburgh, PA, 2006.
34. Dean C.B. Testing for overdispersion in Poisson and binomial regression models // Journal of the American Statistical Association, June 1992, vol. 87, no. 418, pp. 451–457.
35. Dean C.B., Balshaw R. Efficiency lost by analyzing counts rather than event times in Poisson and overdispersion Poisson regression models // Journal of the American Statistical Association, December 1997, vol. 92, no. 440, pp. 1387–1398.
36. Dean C.B., Lawless J.F. Tests for detecting overdispersion in Poisson regression models // Journal of the American Statistical Association, June 1989, vol. 84, no. 406, pp. 467–472.
37. Dillion W., Goldstein M. Multivariate analysis: Methods and applications. – New York, NY: John Wiley & Sons, 1984.
38. Douglas J.B. Likelihood analyses of overdispersion Poisson models // Biometrics, December 1997, vol. 53, no. 4, pp. 1547–1551.
39. Draper N.R., Smith H. Applied regression analysis. – New York, NY: Wiley, 1998.
40. Dudoit S., Frydlyand J., Speed T.P. Comparison of discrimination methods for the classification of tumors using gene expression data // Technical report No. 576, June 2000, University of California, Berkeley, CA.
41. Durbin J., Watson G.S. Testing for serial correlation in least squares regression. II // Biometrika, June 1951, vol. 38, no. 1/2, pp. 159–177.
42. Efron B. Bootstrap methods: Another look at the jackknife // Annals of Statistics, 1979, vol. 7, pp. 1–26.
43. Efron B. Estimating the error rate of a prediction rule: Improvement on crossvalidation // Journal of American Statistical Association, 1983, vol. 78, pp. 316–331.
44. Efron B. The efficiency of logistic regression compared to normal discriminant analysis // Journal of the American Statistical Association, 1975, vol. 70, pp. 892–898.
45. Famoye F., Wulu J.T.Jr., Singh K.P. On the generalized Poisson regression model with an application to accident data // Journal of Data Science, 2004, vol. 2. pp. 287–295.
46. Finney D.J. Probit analysis: A statistical treatment of the sigmoid response curve. – Cambridge, UK: Cambridge University Press, 1971.
47. Fisher R.A. The use of multiple measurements in taxonomic problems // Annals of Eugenics, 1936, vol. 7, pp. 179–188.
48. Flury B., Reidwyl H. Multivariate statistics: A practical approach. – New York, NY: Chapman & Hall, 1988.
49. Fukunaga K., Kessell D.L. Nonparametric Bayes error estimation using unclassified samples // IEEE Transactions on Information Theory, 1973, 1T–19, pp. 434–440.
50. Gareen I.F., Gatsonis C. Primer on multiple regression models for diagnostic imaging research // Radiology, 2003, vol. 229, no. 2, pp. 305–310.
51. Glick N. Additive estimators for probabilities of correct classification // Pattern Recognition, 1978, vol. 10, pp. 211–222.
52. Gnanadesikan R. Methods for statistical data analysis of multivariate observations. – New York, NY: John Wiley & Sons, 1977.
53. Habbema J.D.F., Hermans J. Selection of variables in discriminant analysis by F–statistic and error rate // Technometrics, 1977, vol. 19, pp. 487–493.
54. Hamer M. Failure prediction: Sensitivity of classification accuracy to alternative statistical methods and variable sets // Journal of Accounting and Public Policy, 1983, vol. 2, pp. 289–

307.
55. Hand D.J. Discrimination and classification. – New York, NY: John Wiley & Sons, 1981.
56. Hand D.J. Kernel discriminant analysis. – New York, NY: Research Studies Press, 1982.
57. Hanley J., McNeil B. A method of comparing the areas under receiver operating characteristic curves derived from the same cases // Radiology, September 1983, vol. 148, pp. 839–843.
58. Hardle W., Simar L. Applied multivariate statistical analysis. – New York, NY: Springer, 2003.
59. Harris R.J. A primer of multivariate statistics. – Mahwah, NJ: Lawrence Erlbaum Associates, 2001.
60. Hausman J., Wise D. A conditional probit model for qualitative choice: Discrete decisions recognizing interdependence and heterogeneous Preferences // Econometrica, 1978, vol. 48, no. 2, pp. 403–426.
61. Hora S.C., Wilcox J.B. Estimation of error rates in several–population discriminant analysis // Journal of Marketing Research, 1982, vol. 19, pp. 57–61.
62. Hosmer D.W., Lemeshow S. Applied logistic regression. – New York, NY: John Wiley & Sons, 2000.
63. Huberty C.J. Applied discriminant analysis. – New York, NY: John Wiley & Sons, 1994.
64. Huberty C.J. Issues in the use and interpretation of discriminant analysis // Psychological Bulletin, 1984, vol. 95, pp. 156–171.
65. Huberty C.J., Wisenbaker J.M., Smith J.C. Assessing predictive accuracy in discriminant analysis // Multivariate Behavioral Research, 1987, vol. 22, pp. 307–329.
66. Joachimsthaler E.A., Stam A. Mathematical programming approaches for the classification problem in two–group discriminant analysis // Multivariate Behavioral Research, 1990, vol. 25, pp. 427–454.
67. Johnson R.A., Wichern D.W. Applied multivariate statistical analysis. – Englewood Cliffs, NJ: Prentice Hall, 1988.
68. Kendall M.G. A course in multivariate analysis. – London: Griffin, 1957.
69. Klecka W.R. Discriminant analysis. – Beverly Hills, CA: Sage Publications, 1980.
70. Klecka W.R. Quantitative applications in the social sciences. Vol. 19. Discriminant analysis. – Beverly Hills, CA: Sage, 1980.
71. Klotsche J. A novel nonparametric approach for estimating cut–offs in continuous risk indicators with application to diabetes epidemiology / J. Klotsche, D. Ferger, L. Pieper et al. // BMC Medical Research Methodology 2009, vol. 9, no. 63.
72. Konishi S., Honda M. Comparison of procedures for estimation of errors rates in discriminant analysis under nonormal populations // Journal of Statistical Computing and Simulation, 1990, vol. 36, pp. 105–115.
73. Kshirsagar A.M. Multivariate analysis. – New York, NY: Marcel Dekker, 1972.
74. Kukush A., Schneeweiss H., Wolf R. Three estimators for the Poisson regression model with measurement errors // Collaborative Research Center 386, Discussion Paper 243, 2001.
75. Lachenbruch P.A. An almost unbiased method of obtaining confidence intervals for the probability of misclassification in discriminant analysis // Biometrics, 1967, vol. 23, pp. 639–645.
76. Lachenbruch P.A. Discriminant analysis // Biometrics, 1979, vol. 35, pp. 69–85.
77. Lachenbruch P.A. Discriminant analysis. – New York, NY: Hafner Press, 1975.
78. Lachenbruch P.A., Sneeringer C., Revo L.T. Robustness of the linear and quadratic discriminant function to certain types of non–normality // Communications in Statistics, 1973, vol. 1, pp. 39–57.
79. Lenox C. Identifying failing companies: A re–evaluation of the logit, probit and DA

approaches // Journal of Economics and Business, July 1999, vol. 51, no. 4, pp. 347–364.
80. Long J.S. Regression models for categorical and limited dependent variables. – Thousand Oaks, CA: Sage Publications, 1997.
81. Mahalanobis P.C. On the generalized distance in statistics // Proceedings of the National Institute of Science of India, 1936, vol. 12, pp. 49–55.
82. Mandic D.P., Chambers J.A. Recurrent neural networks for prediction: learning algorithms, architectures, and stability. – New York, NY: John Wiley & Sons, 2001.
83. Marks S., Dunn O. J. Discriminant functions when covariance matrices are unequal // Journal of the American Statistical Association, 1974, vol. 69, pp. 555–559.
84. McGrath J.J. Improving credit evaluation with a weighted application blank // Journal of Applied Psychology, 1960, vol. 44, pp. 325–328.
85. McKay R.J., Campbell N.A. Variable selection techniques in discriminant analysis: I. Description // British Journal of Mathematical and Statistical Psychology, 1982, vol. 35, pp. 1–29.
86. McLachlan G.J. Discriminant analysis and statistical patter recognition. – New York, NY: John Wiley & Sons, 2004.
87. Mehrmann V., Rath W. Numerical methods for the computation of analytic singular value decompositions // Electronic Transactions on Numerical Analysis, 1993, vol. 1, pp. 72–88.
88. Morrison D.F. Multivariate statistical methods. – New York, NY: McGraw–Hill, 1990.
89. Morrison D.G. On the interpretation of discriminant analysis // Journal of Marketing Research, May 1969, vol. 6, no. 2, pp. 156–163.
90. Motulsky H., Christopoulos A. Fitting models to biological data using linear and nonlinear regression. A practical guide to curve fitting. – San Diego, CA: GraphPad Software Inc., 2003.
91. Muthen B. A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators // Psychometrica, March 1984, vol. 49, no. 1, pp. 115–132.
92. Neter J., Wasserman W., Kutner M.H. Applied linear regression models. – Homewood, IL: Irwin, 1998.
93. Neter J., Wasserman W., Whitmore G.A. Applied statistics. – Newton, MA: Allyn & Bacon, 1988.
94. Nocairi H. Discrimination on latent components with respect to patterns. Application to multicollinear data / H. Nocairi, E.M. Qannari, E. Vigneau et al. // Computational Statistics & Data Analysis, 2005, vol. 48, pp. 139–147.
95. Parodi S., Bottarell E. Poisson regression model in epidemiology – An introduction // Annali della Facolta di Medicina Veterinaria di Parma, 2006, vol. XXVI, pp. 25–44.
96. Perreault W.D., Behrman D.N., Armstrong G.M. Alternative approaches for interpretation of multiple discriminant analysis in marketing research // Journal of Business Research, 1979, vol. 7, pp. 151–173.
97. Pindyck R.S., Rubinfeld D.L. Econometric models and economic forecasts. – New York, NY: McGraw–Hill, 1981.
98. Porebski O.R. Discriminatory and canonical analysis of technical college data // British Journal of Mathematical and Statistical Psychological, 1966, vol. 19, pp. 215–236.
99. Press S.J., Wilson S. Choosing between logistic regression and discriminant analysis // Journal of the American Statistical Association, 1978, vol. 73, pp. 699–705.
100. Ragsdale C.T., Stam A. Introducing discriminant analysis to the business statistics curriculum // Decision Sciences, 1992, vol. 23, pp. 724–745.
101. Rao M.M. Discriminant analysis // Annals of the Institute of Statistical Mathematics, 1963, vol. 15, no. 1, pp. 11–24.

102. Rencher A.C. Methods of multivariate analysis. – New York, NY: John Wiley & Sons, 2002.
103. Rousseeuw P.J., Leroy A.M. Robust regression and outlier detection. – New York, NY: John Wiley & Sons, 1987.
104. Rulon P.J. Multivariate statistics for personnel classification / P.J. Rulon, D.V. Tiedeman, M.M. Tatsuoka et al. – New York, NY: John Wiley & Sons, 1967.
105. Ryan T.P. Modern regression methods. – New York, NY: Wiley, 1997.
106. Shoukri M.M., Pause C.A. Statistical methods for health sciences. – New York, NY: CRC Press, 1998.
107. Shubin H. Objective index of hemodynamic status for quantitation of severity and prognosis of shock complicating myocardial infarction / H. Shubin, A. Afifi, W.M. Rand et al. // Cardiovascular Research, 1968, vol. 2, p. 329.
108. Silverman B.W. Density estimation for statistics and data analysis. – New York, NY: Chapman & Hall, 1986.
109. Smith C.A.B. Some examples of discrimination // Annals of Eugenics, 1947, vol. 13, pp. 272–282.
110. Smith W.F. Experimental design for formulation. – Alexandria, VA: Society for Industrial and Applied Mathematics, 2005.
111. Sorum M.J. Estimating the expected and the optimal probabilities of misclassification // Technometrics, 1972, vol. 13, pp. 935–943.
112. Stevens J. Applied multivariate statistics for the social sciences. – Mahwah, NJ: Lawrence Erlbaum Associates, 2002.
113. Tacq J. Multivariate analysis techniques in social science research. – Thousand Oaks, CA: Sage Publications, 1996.
114. Tatsuoka M.M. Multivariate analysis techniques for educational and psychological research. – New York, NY: Macmillan, 1988.
115. Tatsuoka M.M. Multivariate analysis. – New York, NY: John Wiley & Sons, 1971.
116. Uusipaikka E. Confidence Intervals in generalized regression models. – New York, NY: Chapman & Hall / CRC, 2009.
117. Vergara I. StAR: a simple tool for the statistical comparison of ROC curves // I. Vergara, T. Norambuena, E. Ferrada et al. // BMC Bioinformatics, 2008, vol. 9, no. 265.
118. Vittinghoff E. Regression methods in biostatistics: Linear, logistic, survival, and repeated measures models / E. Vittinghoff, S.C. Shiboski, D.V. Glidden et al. – New York, NY: Springer, 1995.
119. Von Eye A., Schuster C. Regression analysis for social sciences. – San Diego, CA: Academic Press, 1998.
120. Walker S.H., Duncan D.B. Estimation of the probability of an event as a function of several independent variables // Biometrika, 1967, vol. 54, pp. 167–179.
121. Webb A.R. Statistical pattern recognition. – New York, NY: John Wiley & Sons, 2002.
122. Wilks S.S. Certain generalizations in the analysis of variance // Biometrika, 1932, vol. 24, pp. 471–494.
123. Youden W.J. Index for rating diagnostic tests // Cancer, 1950, vol. 3, no. 1, pp. 32–35.
124. Zeger S.L., Liang K.–Y. Longitudinal data analysis for discrete and continuous outcomes // Biometrics, March 1986, vol. 42, no. 1, pp. 121–130.
125. Zweig M.H., Campbell G. ROC Plots: A fundamental evaluation tool in clinical medicine // Clinical Chemistry, 1993, vol. 39, no. 4, pp. 561–577.
126. Гайдышев И. Анализ и обработка данных: специальный справочник. – СПб.:

Питер, 2001.
127. Гантмахер Ф.Р. Теория матриц. – М.: Наука, 1988.
128. Голуб Дж., Ван Лоун Ч. Матричные вычисления. – М.: Мир, 1999.
129. Давнис В.В., Тинякова В.И. Прогнозные модели субъективных предпочтений // Вестник ВГУ, Серия: Экономика и управление, 2005, № 1, с. 159–167.
130. Деммель Дж. Вычислительная линейная алгебра. Теория и приложения. – М.: Мир, 2001.
131. Дженнрич Р.И. Пошаговый дискриминантный анализ // Статистические методы для ЭВМ / Под ред. К. Энслейна, Э. Рэлстона, Г.С. Уилфа. – М.: Наука, 1986, с.94–113.
132. Дюк В. Обработка данных на ПК в примерах. – СПб: Питер, 1997.
133. Дюк В., Самойленко А. Data mining: Учебный курс. – СПб.: Питер, 2001.
134. Загоруйко Н.Г. Прикладные методы анализа данных и знаний. – Новосибирск: Издательство института математики, 1999.
135. Каримов Р.Н. Дискриминантный анализ. Методические указания к выполнению лабораторной работы по курсу «Обработка экспериментальной информации» для студентов специальности 220200. – Саратов: СГТУ, 2000.
136. Каримов Р.Н. Обработка экспериментальной информации. Учебное пособие. Ч. 3. Многомерный анализ. – Саратов: СГТУ, 2000.
137. Ким Дж.О. Факторный, дискриминантный и кластерный анализ / Дж.О. Ким, Ч.У. Мюллер, У.Р. Клекка и др. – М.: Финансы и статистика, 1989.
138. Кравец О.Я. Гибридные алгоритмы оптимизации моделей множественной регрессии на основе кросскорреляции // Информационные технологии моделирования и управления, 2005, № 4 (22), с. 548–554.
139. Круглов В.В., Борисов В.В. Искусственные нейронные сети. Теория и практика. – М.: Горячая линия – Телеком, 2001.
140. Кульбак С. Теория информации и статистика. – М.: Наука, 1967.
141. Лошадкин Н.А. Пробит–метод в оценке эффектов физиологически активных веществ при низких уровнях воздействия / Н.А. Лошадкин, В.Д. Гладких, В.А. Голденков и др. // Журнал Российского химического общества им. Д.И. Менделеева, 2002, т. XLVI,№ 6, с. 63–67.
142. Минский М., Пейперт С. Персептроны. – М.: Мир, 1971.
143. Неймарк Ю.И. Многомерная геометрия и распознавание образов // Соросовский образовательный журнал, 1996, т. 2, № 7, с. 119–123.
144. Носач В.В. Решение задач аппроксимации с помощью персональных компьютеров. – М.: МИКАП, 1994.
145. Осовский С. Нейронные сети для обработки информации. – М.: Финансы и статистика, 2002.
146. Плавинская С.И., Плавинский С.Л., Шестов Д.Б. Прогностическая значимость основных факторов риска у женщин по данным популяционного исследования и шкала риска смерти от ССЗ // Российский семейный врач, 2006, № 4, с. 4–9.
147. Поттосина С.А. Экономико–математические модели и методы. – Мн.: БГУИР, 2003.
148. Сборник научных программ на Фортране. Выпуск 1. Статистика. – М.: Статистика, 1974.
149. Уилкинсон Дж. Х., Алгебраическая проблема собственных значений. – М.: Наука, 1970.
150. Уилкинсон, Райнш. Справочник алгоритмов на языке АЛГОЛ. Линейная алгебра. – М.: Машиностроение, 1976.
151. Урбах В.Ю. Дискриминантный анализ и его применение в биологической

систематике и медицинской диагностике // Применение математических методов в биологии. Выпуск 3. – Л.: Издательство ЛГУ, 1964, с. 67–87.
152. Урбах В.Ю. Дискриминантный анализ: основные идеи и приложения (обзор и библиография) // Статистические методы классификации. – М.: Издательство МГУ, 1969, с. 79–173.
153. Фукунага К. Введение в статистическую теорию распознавания образов. – М.: Наука, 1979.
154. Хайкин С. Нейронные сети: полный курс. – М.: Издательский дом «Вильямс», 2006.
155. Хеннан Э. Многомерные временные ряды. – М.: Мир, 1974.
156. Цыплаков А.А. Некоторые эконометрические методы. Метод максимального правдоподобия в эконометрии. Методическое пособие. – Новосибирск: НГУ, 1997.
157. Штремель М.А., Кудря А.В., Иващенко А.В. Непараметрический дискриминантный анализ в задачах управления качеством // Заводская лаборатория. Диагностика материалов, 2006, № 5, с. 53–63.

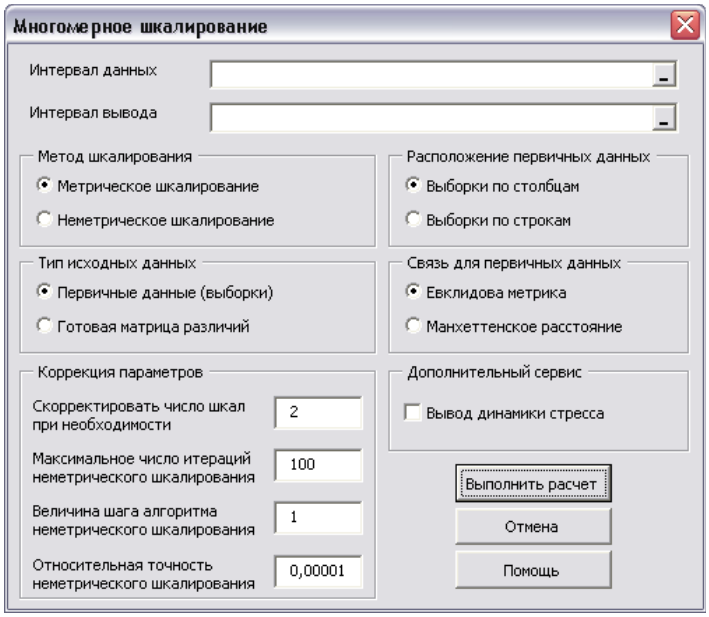
Глава 15. Многомерное шкалирование

15.1. Введение

В программном обеспечении многомерного шкалирования реализованы классические методы многомерного шкалирования.

15.2. Работа с программным обеспечением

Выберите из меню программы пункт AtteStat | Многомерное шкалирование. На экране появится диалоговое окно, изображенное на рисунке.

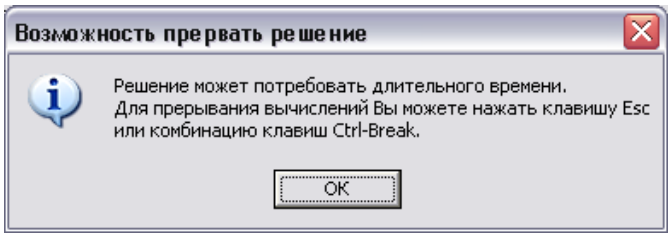


Затем проделайте следующие шаги:

- Выберите или введите интервал матрицы исходных данных (первичных выборок) или матрицы различий.
- Выберите или введите выходной интервал. Начиная с первой ячейки выходного

интервала (следовательно, можно указать только одну ячейку, т.к. остальные ячейки интервала игнорируются), будут выведены результаты вычислений.
- Выберите или оставьте по умолчанию метод многомерного шкалирования.
- Выберите или оставьте по умолчанию тип связи.
- Укажите или оставьте по умолчанию, как расположены выборки.
- Скорректируйте или оставьте по умолчанию число шкал.
- Выберите или оставьте по умолчанию тип исходных данных. Программа, кроме неметрического шкалирования, может работать с исходными данными – первичными выборками или с уже вычисленной матрицей различий.Для неметрического шкалирования задайте или оставьте по умолчанию число итераций, величину шага алгоритма и точность.
- Для неметрического шкалирования в разделе «Дополнительный сервис» можно отметить пункт «Вывод динамики стресса». Данная возможность интересна при исследовании сходимости алгоритма, однако следует помнить, что при большом количестве затраченных итераций, число которых до производства расчета установить затруднительно, объем выдачи может быть значительным (одна ячейка электронной таблицы на одну итерацию).
- Нажмите кнопку «Выполнить расчет».

Время решения для больших задач может оказаться длительным и сильно зависеть от производительности компьютерной системы, поэтому в программу заложена возможность прерывания решения по желанию пользователя до нормального окончания с заданными параметрами. О данной возможности пользователю сообщается в специальном информационном окне, показанном на рисунке, перед любым производством самого решения.



Для начала решения следует нажать кнопку OK.
После выполнения вычислений будет, начиная с первой ячейки выходного интервала, выведено название метода и результаты расчета. Интерпретация полученных результатов статистических расчетов рассмотрена ниже.
За выбор адекватного исходным данным метода расчета несет ответственность пользователь. Программное обеспечение берет на себя верификацию исходных данных, выдавая подробную диагностику. При неверных действиях пользователя выдаются сообщения об ошибках.

15.2.1. Сообщения об ошибках

При ошибках ввода данных могут выдаваться диагностические сообщения следующих типов:

Ошибка	Комментарий
Не определен интервал	Вы не выбрали или неверно ввели входной интервал. Лучшим способом избежать ошибки является не ввод, а выделение интервала стандартным

Ошибка	Комментарий
переменной.	образом, т. е. протаскиванием курсора.
Пустая ячейка.	Проверьте исходные данные и заполните все ячейки, отмеченные Вами как входной интервал. Для избежания ошибок расчета, вызванных разногласиями, трактовать ли пустую ячейку как нуль, требуется заполнение всех ячеек. Если в ячейке не должно быть данных по физической природе исследуемого процесса, введите в данную ячейку нуль.
Нечисловой тип данных.	Проверьте типы ячеек входного интервала. Тип может быть только числовым. Проще всего выделить интервал ячеек и явно определить их тип как числовой стандартными средствами.
Не определена область выводы.	Не выбран или неверно введен выходной интервал. Лучшим способом избежать ошибки является не ввод, а выделение интервала стандартным образом.
Матрица данных не квадратная.	Если выбраны исходные данные в виде матрицы различий, следует выделить диапазон ячеек, содержащих данную матрицу. Матрица различий должна быть квадратной, т. е. число ее строк должно равняться числу столбцов. Кроме того, матрица различий должна быть симметричной относительно главной диагонали. Проверка этого утверждения не производится программой. Следовательно, ответственность за качество выводов, если это не так, лежит на пользователе программы.
Метод работает только с первичными выборками.	Ввиду особенности алгоритма метод неметрического шкалирования работает только с первичными выборками. Задайте интервал данных, содержащий первичные выборки, и повторите расчет.
Неверное число шкал.	Число шкал не может быть меньшим 2 и не может превышать число параметров. Задайте верное число шкал.
Не задано число шкал. Расчет невоможен.	Число шкал не задано. Задайте верное число шкал или при выборе параметров расчета оставьте параметры по умолчанию.
Неверное число итераций.	Задано неверное число итераций для метода неметрического шкалирования. Число итераций не может быть меньшим 2. Максимальное число итераций не ограничено.
Не задано число итераций.	Число итераций для метода неметрического шкалирования не задано. Задайте верное число итераций или при выборе параметров расчета оставьте параметры по умолчанию.
Неверный шаг алгоритма.	Задан неверный шаг алгоритма неметрического шкалирования. Шаг алгоритма выбирается из соображений улучшения скорости его сходимости. При неуверенности, какой шаг следует задать, оставьте данный параметр по умолчанию.
Не задан шаг алгоритма.	Шаг алгоритма для метода неметрического шкалирования не задан. Задайте верный шаг алгоритма или оставьте данный параметр по умолчанию.

Ошибка	Комментарий
Неверная точность алгоритма.	Задана неверная точность алгоритма неметрического шкалирования. Точность алгоритма выбирается из соображений улучшения скорости его сходимости. При неуверенности, какую точность следует задать, оставьте данный параметр по умолчанию.
Не задана точность алгоритма.	Точность алгоритма для метода неметрического шкалирования не задана. Задайте верную точность алгоритма или оставьте данный параметр по умолчанию.

15.3. Теоретическое обоснование

Подобно методам факторного анализа, методы многомерного шкалирования используются для поиска структуры объектов, по терминологии многомерного шкалирования – стимулов – в многомерном пространстве (стимулом в многомерном шкалировании называют объект исследования, что соответствует понятию эмпирической выборки в прикладном статистическом анализе). Подобно методам кластерного анализа, методами многомерного шкалирования изучаются группировки объектов в многомерном пространстве. Методы многомерного шкалирования по цели расчета ближе к кластерному анализу – установлению пространственной конфигурации исследуемых стимулов.
В данном программном обеспечении применяются следующие методы многомерного шкалирования:

- метрический метод Торгерсона,
- неметрический метод Краскела.

Методы отражают различные идеологические подходы к решению проблемы многомерного шкалирования.

15.3.1. Метрики

Мера сходства d_{ij} между объектами i и j называется метрикой, если она удовлетворяет определенным условиям:

- симметрии $d_{ij} = d_{ji}$,
- неравенству треугольника $d_{ij} \le d_{ik} + d_{kj}$,
- различимости нетождественных объектов и неразличимости тождественных объектов,
- иногда также ставят требование максимальной схожести объекта с «самим собой» $d_{ii} = \min_{i,j} d_{ij}$, причем для рассматриваемых метрик всегда $d_{ii} = 0$.

Метрика представляет собой меру сходства типа расстояния между стимулами, вычисленную по определенной формуле. В процессе попарного вычисления метрик между всеми объектами, составляющими матрицу исходных данных, получается так называемая матрица различий.
Основным элементом исследования в многомерном шкалировании является квадратная матрица различий D между стимулами, которая содержит p строк и столбцов, где p – количество стимулов, вычисленная на основе одной из метрик из матрицы первичных исходных данных. По диагонали матрицы различий обязательно располагаются нулевые значения (расстояния между стимулами «сами с собой»).
Представленное программное обеспечение допускает использовать готовые матрицы различий между стимулами, вычисленные заранее на основе любых других метрик или иных мер связи, поэтому возможности расчетов не ограничены используемыми метриками. Так,

например, возможно получение элементов d_{ij}, i = 1,2,...,p; j = 1,2,...,p, матрицы различий D из корреляционной матрицы R по формуле

$$d_{ij} = \sqrt{1 - r_{ij}}, i = 1,2,..., p; j = 1,2,..., p,$$

где r_{ij}, i = 1,2,...,p; j = 1,2,...,p – коэффициент корреляции между стимулами i, i = 1,2,...,p, и j, j = 1,2,...,p.

Корреляционная матрица может быть построена с помощью методов главы «Корреляционный анализ», а все необходимые трансформации корреляционной матрицы – стандартными средствами.

В метрическом методе Торгерсона используется либо готовая матрица различий, либо матрица первичных исходных данных, по выбору. Напротив, неметрический метод Краскела работает только с матрицей первичных исходных данных и не работает с готовой матрицей различий.

15.3.1.1. Метрика Минковского

Наиболее общей классической мерой типа расстояния является метрика Минковского, которая определяет расстояние между стимулами i, i = 1,2,...,p, и j, j = 1,2,...,p.

$$d_{ij} = \sqrt[r]{\sum_{k=1}^{n} \left| x_{ik} - x_{jk} \right|^r},$$

где r – некоторая величина, причем $r \geq 1$,

n – размерность пространства,

x_{ik}, i = 1,2,...,p – проекция точки i, i = 1,2,...,p, на ось k, k = 1,2,...,p,

x_{jk}, i = 1,2,...,p – проекция точки j, j = 1,2,...,p, на ось k, k = 1,2,...,p.

Практически метрика вычисляется как мера между двумя выборками, численность каждой из которых равна n.

В данном программном обеспечении используются только некоторые частные случаи метрики Минковского:

- евклидова метрика,
- манхеттенское расстояние.

15.3.1.2. Евклидова метрика

Если в метрике Минковского положить r = 2, получим стандартное евклидово расстояние (евклидову метрику)

$$d_{ij} = \sqrt{\sum_{k=1}^{n} \left(x_{ik} - x_{jk} \right)^2},$$

где n – размерность пространства,

x_{ik}, i = 1,2,...,p – проекция точки i, i = 1,2,...,p, на ось k, k = 1,2,...,p,

x_{jk}, i = 1,2,...,p – проекция точки j, j = 1,2,...,p, на ось k, k = 1,2,...,p.

Практически метрика вычисляется как расстояние между двумя выборками, численность каждой из которых равна n. Геометрически евклидово расстояние представляет собой обычное расстояние между двумя точками в n–мерном пространстве.

15.3.1.3. Манхеттенское расстояние

При r = 1 метрика Минковского дает манхеттенское расстояние (метрику города, city block, Manhattan distance)

$$d_{ij} = \sum_{k=1}^{n} \left| x_{ik} - x_{jk} \right|,$$

где n – размерность пространства,
x_{ik}, $i = 1,2,...,p$ – проекция точки i, $i = 1,2,...,p$, на ось k, $k = 1,2,...,p$,
x_{jk}, $i = 1,2,...,p$ – проекция точки j, $j = 1,2,...,p$, на ось k, $k = 1,2,...,p$.
Практически метрика вычисляется как мера между двумя выборками, численность каждой из которых равна n.
В многомерном шкалировании метрики, отличные от евклидова расстояния, применяются неохотно. Расчеты показывают, что матрицы различий, построенные на основе данных метрик, приводят к появлению отрицательных собственных значений при решении проблемы собственных значений матрицы скалярных произведений, что автоматически приводит к наличию комплексных собственных векторов, которые с трудом поддаются интерпретации в терминах многомерного шкалирования.

15.3.2. Метрический метод Торгерсона

Метрический метод многомерного шкалирования Торгерсона исходит в своих предпосылках из той идеи, что исходные данные об исследуемых объектах (параметры, посредством которых описаны объекты) являются результатами точных измерений, свободных от ошибок измерения. Задачей метода является представление конфигурации объектов в пространстве шкал меньшей размерности.
Метод основан на анализе так называемой матрицы скалярных произведений B. Данная матрица строится на основе матрицы различий D Метрики. Вычисления элементов матрицы скалярных произведений B производятся по формуле

$$b_{ij} = \frac{1}{2}\left(-d_{ij}^2 + \frac{1}{p}\sum_{i=1}^{p} d_{ij}^2 + \frac{1}{p}\sum_{j=1}^{p} d_{ij}^2 - \sum_{i=1}^{p}\sum_{j=1}^{p} d_{ij}^2 \right), i = 1,2,..., p; j = 1,2,..., p,$$

где d_{ij}, $i = 1,2,...,p$; $j = 1,2,...,p$, расстояния между стимулами i, $i = 1,2,...,p$, и j, $j = 1,2,...,p$, представляющие собой одну из разновидностей метрики Минковского, составляющие матрицу D,
p – количество стимулов.
Из решения стандартной проблемы собственных значений действительной симметрической матрицы скалярных произведений, которая не представляет вычислительных сложностей, устанавливается представление стимулов в координатном пространстве осей шкальных значений. Подробнее о решении проблемы собственных значений и применяемых алгоритмах можно узнать из главы «Матричная и линейная алгебра».
Полученные координатные оси почти всегда могут быть содержательно интерпретированы. Кроме того, результаты анализа могут быть использованы для классификации стимулов.
Основным результатом расчета является матрица шкальных значений Z (координат стимулов в пространстве шкал). Элементы матрицы координат стимулов вычисляются по формуле

$$z_{pq} = \sqrt{\lambda_q}\, y_{pq},$$

где q, $1 \le q \le n$ – количество удерживаемых максимальных собственных значений,
n – размерность исходного пространства,
λ_q – q–е собственное значение матрицы скалярных произведений,
y_{pq} – собственный вектор матрицы скалярных произведений, соответствующий собственному значению λ_q.
Максимальное число координат стимулов q (количество осей шкальных значений, размерность пространства шкал) будет равно количеству p собственных значений матрицы скалярных произведений. Однако часть собственных значений λ_q на этапе вычислений может

оказаться нулевой в вычислительном смысле. Кроме того, количество осей шкальных значений может быть скорректировано самим пользователем программы, исходя из анализа величин выдаваемых программой собственных значений (также их процентного содержания).
В реальных расчетах, как правило, происходит следующее: из величины собственных значений λ_q и их процентного содержания сразу ясно, сколько собственных значений нужно оставить, так как отброшенные значения обычно являются очень малыми величинами, по модулю порядка 10^{-10}.
Хотя в литературе описаны и другие методы выбора количества осей шкальных значений, общее правило состоит в том, что осей должно быть достаточно для содержательной интерпретации пространственной конфигурации стимулов. На практике часто ограничиваются двумя или тремя осями.
Если число осей шкальных значений больше или равно двум, дополнительно производится объективное вращение решения методом VARIMAX подобно тому, как это сделано в факторном анализе, который подробно рассмотрен в одноименной главе. Процедура вращения не изменяет взаимную пространственную координацию стимулов, но часто улучшает интерпретируемость решения путем сдвига гроздей стимулов в координатном пространстве ближе к той или иной оси шкал.
Хотя это и не является необходимым этапом решения, результаты анализа рекомендуется изобразить графически, а для размерности пространства более двух или трех графики должны быть представлены двумерными срезами пространства.

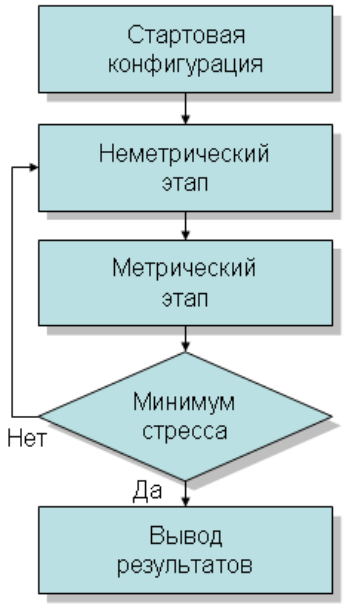
15.3.3. Неметрический метод Краскела

Все методы неметрического многомерного шкалирования, в отличие от метрического многомерного шкалирования, исходят в своих предпосылках из той идеи, что данные об исследуемых объектах являются результатами измерений, искаженных ошибками. Поэтому причинами возможного неудовлетворительного представления конфигурации объектов в пространстве меньшей размерности в неметрическом шкалировании считаются ошибки в исходных данных. Параметры, описывающие объекты и заданные матрицей исходных данных, в неметрическом шкалировании в процессе решения изменяются таким образом, чтобы получить лучшее представление конфигурации стимулов в пространстве шкал меньшей размерности.
Стрессом в неметрическом многомерном шкалировании называют функционал, подлежащий минимизации. Имеются различные формы стресса. Здесь мы используем стресс в одной из классических форм

$$S = \sqrt{\frac{\sum_{i=1}^{p}\sum_{j=1}^{p}\left(d_{ij} - \hat{d}_{ij}\right)^2}{\sum_{i=1}^{p}\sum_{j=1}^{p} d_{ij}^2}},$$

где d_{ij}, $i = 1,2,...,p$; $j = 1,2,...,p$, элементы матрицы различий, представляющие собой расстояния между стимулами i, $i = 1,2,...,p$, и j, $j = 1,2,...,p$, вычисленные по формуле одной из разновидности метрики Минковского по полной матрице исходных данных, элементами которой являются величины x_{kl}, $k = 1,2,...,p$; $l = 1,2,...,n$,
n – размерность исходного пространства,
$\hat{d}_{ij}, i = 1,2,..., p; j = 1,2,..., p,$ элементы оценки матрицы различий, представляющие собой оценки расстояний d_{ij}, $i = 1,2,...,p$; $j = 1,2,...,p$, вычисленные в пространстве шкал меньшей размерности, элементами которой являются величины x_{kl}, $k = 1,2,...,p$; $l = 1,2,...,q$,

q – размерность пространства шкал,
p – количество стимулов.
Алгоритм представляет собой итерационный процесс. Каждая итерация включает в себя метрический этап, применяемый для получения оценки матрицы различий, и неметрический этап, заключающийся в подборе оценки матрицы исходных данных, которая будет минимизировать стресс. Таким образом, целью вычислений является подбор такой конфигурации матрицы исходных данных, чтобы обеспечить минимум стресса по величинам x_{kl}, $k = 1,2,...,p$; $l = 1,2,...,n$. Общая схема алгоритма показана на рисунке.



Метрический этап подробно описан в разделе, посвященном методу Торгерсона, поэтому подробнее остановимся на неметрическом этапе.
Ряд современных авторов предлагает для минимизации стресса использовать различные универсальные методы оптимизации, от градиентных методов и до нейронных сетей. Универсальные алгоритмы, использующие численное определение градиентов и других параметров схемы алгоритмов безусловной оптимизации, эффективно можно использовать, когда вид целевой функции неизвестен заранее. Если же вид целевой функции известен и, более того, весьма прост, то более плодотворным оказывается подход Краскела, который предполагает аналитическое вычисление градиента. Минимизация рассматриваемой формы стресса приводит к формуле, называемой градиентным алгоритмом Краскела. Алгоритм записывается в рекурсивном виде как

$$\hat{x}_{kl}^{(c+1)} = \hat{x}_{kl}^{c} - \frac{2\alpha}{B}\sum_{\substack{i=1 \\ i\neq k}}^{p}\left[\frac{\hat{d}_{ik}^{(c)} - \hat{d}_{ik}^{(c+1)}}{\hat{d}_{ik}^{r-1}}\left|\hat{x}_{il}^{(c)} - \hat{x}_{kl}^{(c)}\right|^{r-1} sign\left(\hat{x}_{il}^{(c)} - \hat{x}_{kl}^{(c)}\right)\right], k = 1,2,...,p; l = 1,2,...,n,$$

где c – номер итерации, при $c = 0$ стартовая конфигурация алгоритма задается метрическим методом Торгерсона, максимально допустимое число итераций задается,
знак «крышечка» означает оценки величин, вычисленные на этапе итерации,
α – параметр, обычно называемый шагом итерации, от которого может зависеть как скорость сходимости итерационного процесса, так и сама сходимость алгоритма; значение параметра «по умолчанию» может быть изменено пользователем,
r – величина, применяемая при вычислении метрики Минковского в той или иной форме, причем $r \geq 1$, в программе реализованы только случаи евклидова расстояния ($r = 2$) и

манхеттенского расстояния ($r = 1$),
sign(.)– знак выражения,

$$B = \sum_{i=1}^{p}\sum_{j=1}^{p}\hat{d}_{ij}^{2}.$$

Преобразования, выполненные над матрицей исходных данных в ходе итерационного процесса, представляют собой применение к исходным данным некоторой неизвестной монотонной функции. Итерационный процесс завершается при достижении минимума стресса, о чем свидетельствует отличие между стрессом на текущей и предыдущей итерациях на некоторую заранее заданную пользователем малую величину ε, например, 0,00001. Таким образом, нормальное условие останова может быть записано в виде
$|S^{(c+1)} - S^{(c)}| \le \varepsilon$,
где c – номер итерации.
Аварийным завершением процесса считают его завершение не по минимуму стресса, а по достижении определенного заранее заданного пользователем числа итераций. В данном случае программа выдает полученную в результате расчета информацию по состоянию на момент останова.
Если число осей шкальных значений больше или равно двум, дополнительно производится объективное вращение решения методом VARIMAX подобно тому, как это сделано в факторном анализе. Процедура вращения не изменяет взаимную пространственную координацию стимулов, но часто улучшает интерпретируемость решения путем сдвига гроздей стимулов в координатном пространстве ближе к той или иной оси шкал.
Кроме возможности пользовательского ввода параметров, предусмотрен вывод динамики стресса по итерациям. Данная возможность может быть полезна при исследовании сходимости процесса.
Хотя это и не является необходимым этапом решения, результаты анализа рекомендуется изобразить графически, а для размерности пространства более двух или трех графики должны быть представлены двумерными срезами пространства. Динамика стресса по итерациям, при необходимости, также представляется графически.

15.3.4. Проблема вращения

Оси координат пространства шкал ортогональны, и их направления устанавливаются последовательно, по максимуму оставшейся дисперсии. Более предпочтительное положение системы координат получают путем вращения этой системы вокруг ее начала.
Пространственная конфигурация стимулов в результате применения этой процедуры остается неизменной. Целью вращения является нахождение одной из возможных систем координат для получения так называемой простой структуры. Обычно применяют популярный метод вращения VARIMAX, который, помимо методов многомерного шкалирования, используется также в методах, представленных в главе «Факторный анализ». В главе приводится краткое описание идеи метода VARIMAX. Такого описания, конечно, недостаточно для численной реализации алгоритма. Подробное описание метода VARIMAX приводится во многих источниках по вычислительным методам статистики, например, у Магнуса, в первом выпуске «Сборника научных программ на Фортране».

Список использованной и рекомендуемой литературы

1. Arabie P., Carroll J.D., DeSarbo W.S. Three–way scaling and clustering. – Newbury Park, CA : Sage Publications, 1987.
2. Blalock H.M.Jr. Social statistics. – New York, NY: McGraw–Hill, 1979.
3. Bock R.D. Multivariate statistical methods in behavioral research. – New York, NY:

McGraw–Hill, 1975.
4. Borg I., Groenen P. Modern multidimensional scaling: Theory and applications. – New York, NY: Springer–Verlag, 1997.
5. Borg I., Lingoes J.C. A model and algorithm for multidimensional scaling with external constraints on the distances // Psychometrika, 1980, vol. 45, pp. 25–38.
6. Carroll J.D., Arabie P. Multidimensional scaling // Annual Review of Psychology, January 1980, vol. 31, pp. 607–649.
7. Carroll J.D., Chang J.J. Analysis of individual differences in multidimensional scaling via an N–way generalization of «Eckart–Young» decomposition // Psychometrika, 1970, vol. 35, pp. 283–319.
8. Carroll J.D., Pruzansky S., Kruskal J.B. CANDELINC: a general approach to multidimensional analysis of many–way arrays with linear constraints on parameters // Psychometrika, 1980, vol. 45, pp. 3–24.
9. Carroll J.D., Wish M. Multidimensional scaling: Models, methods and relations to Delphi // The Delphi Method: Techniques and Applications / Ed. by H. Linstone, M. Turoff. – Reading, MA: Addison–Wesley, 1975.
10. Cooley W.W., Lohnes P.R. Multivariate data analysis. – New York, NY: John Wiley & Sons, 1971.
11. Cox T., Cox M. Multidimensional scaling. – London: Chapman & Hall, 1994.
12. Davison M.L. Multidimensional scaling. – Melbourne: Krieger Publishing Company, 1992.
13. DeLeeuw J. Applications of convex analysis to multidimensional scaling // Recent Developments in Statistics / Ed. by J.R. Barm et al. – North Holland, 1977, pp. 133–145.
14. DeLeeuw J. Convergence of the majorization method for multidimensional scaling // Journal of Classification, 1988, vol. 5, pp. 163–180.
15. Green P.E., Carmone F.J.Jr., Smith S.M. Multidimensional scaling: Concepts and applications. – Boston: Allyn & Bacon, 1989.
16. Groenen P. The majorization approach to multidimensional scaling: Some problems and extensions. Ph. D. thesis. – Leiden: DSWO Press, 1993.
17. Guttman L. A general nonmetric technique for finding the smallest coordinate space for a configuration of points // Psychometrika, 1968, vol. 33, pp. 469–506.
18. Harris R.J. A primer of multivariate statistics. – New York, NY: Academic Press, 1975.
19. Hays W.L. Statistics for the social sciences. – New York, NY: Holt Rinehart and Winston, 1973.
20. Hofmann T., Buhmann J. Multidimensional scaling and data clustering // Advances in Neural Information Processing Systems 7 (NIPS'94). – Morgan Kaufmann Publishers, 1995, pp. 104–111.
21. Jackson J.E. A user's guide to principal components. – New York, NY: John Wiley & Sons, 1991.
22. Kearsley A.J, Tapia R.A, Trosset M.W. The solution of the metric STRESS and SSTRESS problems in multidimensional scaling using Newton's method // Computational Statistics, 1998, vol. 13, no. 3, pp. 369–396.
23. Klock H., Buhmann J.M. Data visualization by multidimensional scaling: A deterministic annealing approach // Pattern Recognition, 1999, vol. 33, no. 4, pp. 651–669.
24. Kruskal J.B. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis // Psychometrika, 1964, vol. 29, pp. 1–27.
25. Kruskal J.B. Nonmetric multidimensional scaling: A numerical method // Psychometrika, 1964, vol. 29, pp. 115–130.
26. Kruskal J.B. Transformations of data // International Encyclopedia of Statistics, vol. 2 / Ed. by W.H. Kruskal, J.M. Tanur. – New York, NY: The Free Press, 1978.

27. Kruskal J.B., Wish M. Multidimensional scaling // International Handbooks of Quantitative Applications in the Social Sciences, vol.4 / Ed. by M.S. Lewis–Beck. – Thousand Oaks, CA: Sage Publications, 1994, pp. 301–387.
28. Kruskal J.B., Wish M. Multidimensional scaling. – Beverly Hills, CA: Sage Publications, 1978.
29. Lipovetsky S., Conkin W.M. Thurstone scaling via binary response regression // Statistical Methodology, 2004, vol. 1, pp. 93–104.
30. Martens J.–B. Multidimensional modeling of image quality // Proceedings of the IEEE, January 2002, vol. 90, no. 1, pp. 133–153.
31. Mathar R.A. Hybrid global optimization algorithm for multidimensional scaling. Classification and knowledge optimization // Proceedings of the 20th Annual Conference of the Gesellschaft fur Klassification e.V., University of Freiburg, 1996, pp. 63–71.
32. McQuaid M. Multidimensional scaling for group memory / M. McQuaid, T.–H. Ong, H. Chen et al. // Decision Support Systems, November 1999, vol. 27, no. 1, pp. 163–176.
33. Morrison D.F. Multivariate statistical methods. – New York, NY: McGraw–Hill, 1967.
34. Nunnally J.C. Psychometric theory. – New York, NY: McGraw–Hill, 1978.
35. Overall J.E., Klett C.J. Applied multivariate analysis. – New York, NY: McGraw–Hill, 1972.
36. Rabinowitz G. An introduction to nonmetric multidimensional scaling // American Journal of Political Science, 1975, vol. 19, pp. 343–390.
37. Rabinowitz G. Nonmetric multidimensional scaling // New tools for social scientists. Advances and applications in research methods / Ed. by W.D. Berry, M.S. Lewis–Beck. – Newbury Park, CA: Sage, 1986, pp. 77–107.
38. Ramsay J.O. Maximum likelihood estimation in multidimensional scaling // Psychometrika, 1977, vol. 42, pp. 241–266.
39. Rohde D.L.T. Methods for binary multidimensional scaling // Neural Computation, 1 May 2002, vol. 14, no. 5, pp. 1195–1232.
40. Schiffman S., Reynolds M., Young F. Introduction to multidimensional scaling: Theory, methods, and applications. – New York, NY: Academic Press, 1981.
41. Schiffman S.S. Introduction to multidimensional scaling: Theory, methods, and applications / Ed. by S.S. Schiffman, M.L. Reynolds, F.W. Young. – New York, NY: Academic Press, 1981.
42. Shepard R.N. Analysis of proximities: Multidimensional scaling with unknown distance function // Psychometrica, 1962, vol. 27, pp. 125–140, 219–246.
43. Shepard R.N. Multidimensional scaling: Theory and applications in the behavioral sciences / Ed. by R.N. Shepard, A.K. Romney, S.B. Nerlove. – New York, NY: Seminar Press, 1972.
44. Snedecor G.W., Cochran W.G. Statistical methods. – Ames, lA: The Iowa State University Press, 1967.
45. Spence I., Graef J. The determination of the underlying dimensionality of an empirically obtained matrix of proximities // Multivariate Behavioral Research, 1974, vol. 9, pp. 331–342.
46. Spence I., Ogilvie J.C. A table of expected stress values for random rankings in nonmetric multidimensional scaling // Multivariate Behavioral Research, 1973, vol. 8, pp. 511–517.
47. Takane Y., Young F.W., DeLeeuw J. Nonmetric individual differences multidimensional scaling: An alternating least squares method with optimal scaling features // Psychometrika, 1977, vol. 42, pp. 7–67.
48. Torgerson W.S. Theory and methods of scaling. – New York, NY: John Wiley & Sons,1958.
49. Tsogo L., Masson M.H., Bardot A. Multidimensional scaling methods for many–object sets: A review // Multivariate Behavioral Research, 2000, vol. 35, no. 3, pp. 307–319.
50. Tukey J.W. Exploratory data analysis. – Reading, MA: Addison–Wesley, 1977.
51. Tzeng J., Lu H.H–S., Li W.–H. Multidimensional scaling for large genomic data sets // BMC

Bioinformatics, 2008, vol. 9, issue 179.
52. Van Wezel M.C. Nonmetric multidimensional scaling with neural networks / M.C. van Wezel, W.A. Kosters, P. van der Putten et al. // Lecture Notes in Computer Science, 2001, vol. 2189, pp.145–155.
53. Van Wezel M.C., Kosters W.A. Nonmetric multidimensional scaling: Neural networks versus traditional techniques // Intelligent Data Analysis, 2004, vol. 8, pp. 601–613.
54. Wish M., Carroll J.D. Applications of individual differences scaling to studies of human perception and judgment // Handbook of perception, vol. 2 / Ed. by E.C. Carterettee, M.P. Friedman. – San Diego, CA: Academic Press, 1974, pp. 449–491.
55. Young F.W. Multidimensional scaling // Encyclopedia of Statistical Sciences, vol. 5. – New York, NY: John Wiley & Sons, 1985, pp. 649–658.
56. Young F.W., Hamer R.M. Multidimensional scaling: History, theory and applications. – Hillsdale, NJ: Erlbaum Associates, 1987.
57. Young F.W., Torgerson W.S. TORSCA a FORTRAN IV program for Shepard–Kruskal multidimensional scaling analysis // Behavioral Science, 1976, vol. 12, p. 498.
58. Айвазян С.А. Прикладная статистика: Классификация и снижение размерности: Справочное издание / С.А. Айвазян, В.М. Бухштабер, И.С. Енюков и др. – М.: Финансы и статистика, 1989.
59. Андерсон Т. Введение в многомерный статистический анализ. – М.: Государственное издательство физико–математической литературы, 1963.
60. Андрукевич П.Ф. Сравнение моделей одномерного и многомерного методов шкалирования // В сб. Статистические методы анализа экспертных оценок / Под ред. Ю.Н. Тюрина, А.А. Френкель. – М.: Наука, 1977, с. 267–280.
61. Большаков А.А., Каримов Р.Н. Методы обработки многомерных данных и временных рядов: Учебное пособие для вузов. – М.: Горячая линия – Телеком, 2007.
62. Гайдышев И.П. Анализ и обработка данных: специальный справочник. – СПб: Питер, 2001.
63. Герганов Е.Н., Терехина Л.Ю., Фрумкина Р.Х. Анализ восприятия звуковых стимулов индивидами – носителями разных фонетических систем // Вопросы кибернетики: Экспертные оценки. – М.: Научный совет по комплексной проблеме «Кибернетика» АН СССР, 1979, с.180–189.
64. Дронов С.В. Многомерный статистический анализ: Учебное пособие. – Барнаул: Издательство Алтайского государственного университета, 2003.
65. Дэйвисон М. Многомерное шкалирование: Методы наглядного представления данных. – М.: Финансы и статистика, 1988.
66. Дэннис Дж., мл., Шнабель Р. Численные методы безусловной оптимизации и решения нелинейных уравнений. – М.: Мир, 1988.
67. Дюк В., Самойленко А. Data mining: Учебный курс. – СПб: Питер, 2001.
68. Жижикин А.В. Использование полуметрического метода многомерного шкалирования при исследовании данных социологических опросов // Современные проблемы математики и информатики, 2000, вып. 3, с. 190–195.
69. Закускин С. Анализ потребительских предпочтений методами многомерного шкалирования // Лаборатория рекламы, маркетинга и public relations, 2003, №5.
70. Клигер С.А., Косолапов М.С., Толстова Ю.Н. Шкалирование при сборе и анализе социологической информации. – М.: Наука, 1978.
71. Косолапов М.С. Классификация методов пространственного представления структуры исходных данных // Социологические исследования, 1976, № 2, с. 98–109.
72. Краскэл Дж.Б. Многомерное шкалирование и другие методы поиска структуры // Статистические методы для ЭВМ / Под ред. К. Энслейна, Э. Рэлстона, Г.С. Уилфа. –

М.: Наука, 1986, с.301–347.
73. Крускал Дж. Взаимосвязь между многомерным шкалированием и кластер–анализом // Классификация и кластер / Под ред. Дж. Вэн Райзина. – М.: Мир, 1980, с. 20–41.
74. Ллойд Э. Справочник по прикладной статистике. В 2–х томах. Т. 2. / Под ред. Э. Ллойда, У. Ледермана, С.А. Айвазяна и др. – М.: Финансы и статистика, 1990.
75. Магнус Я.Р. Матричное дифференциальное исчисление с приложениями к статистике и эконометрике. – М.: Физматлит, 2002.
76. Мудров А.Е. Численные методы для ПЭВМ на языках Бейсик, Фортран и Паскаль. – Томск: МП «РАСКО», 1991.
77. Носач В.В. Решение задач аппроксимации с помощью персональных компьютеров. – М.: МИКАП, 1994.
78. Парамей Г.В. Применение многомерного шкалирования в психологических исследованиях // Вестник МГУ, серия 14 «Психология», 1983, № 2, с. 57–69.
79. Перекрест В.Т. Нелинейный типологический анализ социально–экономической информации: Математические и вычислительные методы. – Л.: Наука, 1983.
80. Петров В.М. Опыт применения неметрического многомерного шкалирования при изучении предпочтений молодежи в области авторской песни // Социология: методология, методы, математические модели (Социология: 4М), 1991, № 1, с. 99–114.
81. Прохоров Ю.В. Математический энциклопедический словарь / Гл. ред. Ю.В. Прохоров. – М.: Большая Российская Энциклопедия, 1995.
82. Сатаров Г.А. Многомерное шкалирование и другие методы при комплексном анализе данных // Анализ нечисловой информации в социологических исследованиях. – М.: Наука, 1985, с.132–140.
83. Сатаров Г.А., Каменский В.С. Общий подход к анализу экспертных оценок методами неметрического многомерного шкалирования // В сб. Статистические методы анализа экспертных оценок / Под ред. Ю.Н. Тюрина, А.А. Френкель. – М.: Наука, 1977, с. 251–266.
84. Сборник научных программ на Фортране. Выпуск 1. Статистика. – М.: Статистика, 1974.
85. Терехина А.Ю. Анализ данных методами многомерного шкалирования. – М.: Наука, 1986.
86. Терехина А.Ю. Методы многомерного шкалирования и визуализации данных // Автоматика и телемеханика, 1973, № 7, с. 86–94.
87. Терехина А.Ю. Многомерное шкалирование в психологии // Психологический журнал, 1983, т. 4, № 1, с. 76–88.
88. Терехина А.Ю. О двух задачах индивидуального многомерного шкалирования // Автоматика и телемеханика, 1974, № 4, с. 135–142.
89. Толстова Ю.Н. Принципы анализа данных в социологии // Социология: методология, методы, математические модели (Социология: 4М), 1991, №1, с. 51–61.
90. Торгерсон У.С. Многомерное шкалирование. Теория и метод // Статистическое измерение качественных характеристик / Под ред. Е.М. Четыркина. – М.: Статистика, 1972, с. 94–118.
91. Шепард Р.Н. Многомерное шкалирование и безразмерное представление различий // Психологический журнал, 1980, т. 1, № 4, с. 72–83.
92. Шуп Т. Решение инженерных задач на ЭВМ: Практическое руководство. – М.: Мир, 1982.

Глава 16. Обработка экспертных оценок

16.1. Введение

В программном обеспечении обработки экспертных оценок применяются различные методы обработки. Отметим, что представленные методы не исчерпывают всех возможностей обработки экспертных оценок. Некоторые другие методы программы также могут быть использованы для обработки экспертных оценок. Примеры таких решений:

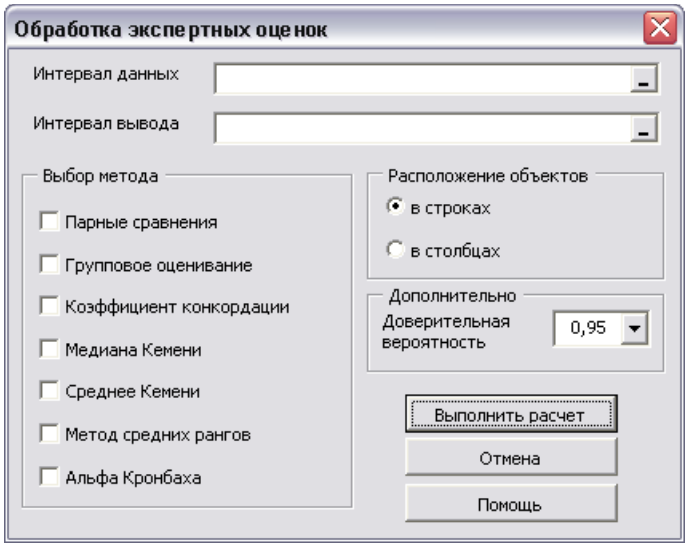
- Выявление однородных групп экспертов может быть выполнено с помощью соответствующих методов кластерного анализа, многомерного шкалирования, факторного анализа.
- Нахождение согласованного мнения группы экспертов может быть выполнено с помощью соответствующих методов дисперсионного анализа.
- Исследование корреляции экспертных оценок может быть выполнено с помощью соответствующих методов корреляционного анализа.

Остановимся более подробно на выявлении однородных групп экспертов. Для этого рекомендуется воспользоваться методами главы «Кластерный анализ». Из представленных методов классификации можно применять только метод средней связи Кинга в комбинации с мерой различия «Расстояние отношений», вычисляемое на основе матриц отношений частичного порядка. Обратим внимание на одну специфическую особенность взаимодействия представленных методов и методов главы «Кластерный анализ». При классификации экспертов для выявления их однородных групп мы в качестве объектов классификации подразумеваем самих экспертов. Поэтому, если при обработке экспертных ранжировок объекты расположены в строках, а эксперты – в столбцах, то при классификации экспертов (матрица экспертных ранжировок – та же самая) в главе «Кластерный анализ» следует выбрать опцию «Объекты в столбцах».

См. книги Бешелева с соавт., Нейлора. Обзоры см. в книгах Литвака (1982), Тюрина, статье Шмерлинга с соавт. Об организации экспертной работы см. книгу Литвака (2004).

16.2. Работа с программным обеспечением

Выберите из меню программы пункт AtteStat | Экспертные оценки. На экране появится диалоговое окно, изображенное на рисунке:



Затем проделайте следующие шаги:

- Выберите или введите интервалы матрицы исходных данных.
- Выберите или введите выходной интервал. Начиная с первой ячейки выходного интервала (следовательно, можно указать только одну ячейку, т.к. остальные ячейки интервала игнорируются), будут выведены результаты вычислений.
- Выберите метод анализа.
- Если это необходимо для выбранного метода, укажите или оставьте по умолчанию, как расположены объекты в матрице опроса. По умолчанию объекты расположены в строках, эксперты в столбцах. К примеру, в этом случае один столбец электронной таблицы может представлять собой ранжировку объектов, представленную одним экспертом.
- Нажмите кнопку «Выполнить расчет».

После выполнения вычислений будет, начиная с первой ячейки выходного интервала, выведено название метода и результаты расчета.
За выбор адекватного исходным данным метода расчета несет ответственность пользователь. Программное обеспечение берет на себя верификацию исходных данных, выдавая подробную диагностику. При неверных действиях пользователя выдаются сообщения об ошибках.

16.2.1. Сообщения об ошибках

При ошибках ввода данных могут выдаваться диагностические сообщения следующих типов:

Ошибка	Комментарий
Пустая ячейка в области данных.	Проверьте исходные данные и заполните все ячейки, отмеченные Вами как входной интервал. Для избежания ошибок расчета, вызванных разногласиями, трактовать ли пустую ячейку как нуль, программное обеспечение требует заполнения всех ячеек. Если в ячейке не должно быть данных по физической природе исследуемого процесса, введите в данную ячейку нуль.
Нечисловой тип данных.	Проверьте типы ячеек входного интервала. Тип может быть только числовым. Проще всего выделить интервал ячеек и явно определить их тип как числовой стандартными средствами.
Не определен интервал переменной.	Вы не выбрали или неверно ввели входной интервал. Лучшим способом избежать ошибки является не ввод, а выделение интервала стандартным образом, т. е. протаскиванием курсора.
Не определена область вывода.	Вы не выбрали или неверно ввели выходной интервал. Лучшим способом избежать ошибки является не ввод, а выделение интервала стандартным образом, т. е. протаскиванием курсора.

16.3. Теоретическое обоснование

Математико–статистические методы обработки экспертных оценок необходимо применять во всех случаях, когда исходные данные представляют собой результаты работы экспертов или экспертных комиссий, и требуется найти обоснованное согласованное мнение группы экспертов для представления результатов лицу, принимающему решение. Попутно могут

быть решены и другие, частные, задачи типа получения весовых коэффициентов объектов и весовых коэффициентов компетентности экспертов.
В программном обеспечении применяются различные методы обработки экспертных оценок, предназначенные для решения следующих основных задач:
Методы оценивания:

- Метод парных сравнений Терстоуна.
- Метод группового оценивания.

Методы исследования согласованности мнений экспертов:

- Коэффициент конкордации.
- Альфа Кронбаха.

Методы получения коллективного мнения:

- Метод средних рангов.
- Медиана Кемени.
- Среднее Кемени.

Представленные методы отражают различные подходы к решению однотипных задач и при определенных условиях могут давать одинаковые результаты.
Специфические для каждого метода алгоритмы требуют, чтобы исходные данные имели определенную структуру. Исходные данные, содержащие экспертные оценки, могут быть различных видов. Методы, применяемые в настоящем программном обеспечении, рассчитаны на обработку исходной матрицы определенной структуры, поэтому при обработке тех или иных исходных данных следует убедиться, что применяются адекватные методы. Конкретные требования к исходным данным представлены при описании методов расчета. Ниже рассмотрены основные матрицы исходных данных, допустимые в настоящем программном обеспечении.
Матрица парных сравнений (матрица предпочтений) A должна иметь следующий вид:

Объекты	A_1	A_2	...	A_n
A_1	a_{11}	a_{12}	...	a_{1n}
A_2	a_{21}	a_{22}	...	a_{2n}
...	...	...	...	...
A_n	a_{n1}	a_{n2}	...	a_{nn}

Элементы матрицы парных сравнений получаются как:

$$a_{ij}=\begin{cases} 0, & A_i \prec A_j, \\ 1, & A_i \sim A_j, \\ 2, & A_i \succ A_j, \end{cases} \quad i, j = 1,2,...,n,$$

где n – количество объектов,
$\prec$, $\sim$, $\succ$ – расширения на множества математических операций, соответственно, $<$, $\approx$, $>$.
Операция $A_i > A_j$ означает, что объект A_i превосходит, по мнению эксперта, объект A_j. Иначе говоря, запись $A_i > A_j$ означает предпочтение объекта i над объектом j. Если эксперт затрудняется в выборе варианта предпочтений, имеет место так называемая связь, обозначаемая как $\sim$.
Матрица опроса P должна иметь следующий вид:

Объекты	Эксперты			
	E_1	E_2	...	E_m
A_1	p_{11}	p_{12}	...	p_{1m}

A_2	p_{21}	p_{22}	...	p_{2m}
...	...	...	...	...
A_n	p_{n1}	p_{n2}	...	p_{nm}

Элементы p_{ij}, $i = 1,2,...,n$; $j = 1,2,...,m$, где n – количество объектов, m – количество экспертов, матрицы опроса могут быть как численными значениями (в том числе весами, приписанными экспертами данным объектам), так и ранжировками. Однако некоторые методы анализа могут иметь специфические требования к виду матрицы опроса, поэтому необходимо проявлять внимательность.
Ранжировка, данная одним экспертом, может быть представлена матрицей отношений частичного порядка P (существуют и другие типы матрицы отношений), элементы которой вычисляются по формуле, в терминологии и обозначениях множеств,

$$p_{ij} = \begin{cases} 1, если\ \_(a_i, a_j) \in P, (a_j, a_i) \notin P, \\ 0, если\ \_(a_i, a_j) \notin P, (a_j, a_i) \notin P, \\ -1, если\ \_(a_i, a_j) \notin P, (a_j, a_i) \in P. \end{cases}$$

Таким образом, m матриц отношений эквивалентны рассмотренной выше матрице опроса. Использование матриц отношений вызвано их удобством при изложении некоторых методов обработки экспертных оценок. Практически, по причине того, что в настоящем программном обеспечении объекты заданы числами, вычисление производится по формуле

$$p_{ij} = \begin{cases} 1, если\ \_\ a_i > a_j, \\ 0, если\ \_\ a_i = a_j, \\ -1, если\ \_\ a_i < a_j. \end{cases}$$

В литературе применяются также иные формы матрицы предпочтений. См. обзор Шмерлинга с соавт., книгу Ливака (1982). См. книгу Тюрина, учебные пособия Тиняковой, Эйтингона с соавт., статью Шмерлинга.

16.3.1. Парные сравнения

Метод парных сравнений Терстоуна в качестве исходных данных для анализа использует матрицу парных сравнений, применительно к результатам опроса одного эксперта.
Данная квадратная матрица обладает следующими основными свойствами:

- Матрица несимметрическая.
- Матрица действительная.
- Матрица неотрицательная.

Для таких матриц, согласно теореме Перрона и ее обобщению – теореме Фробениуса (Гантмахер, с. 334), всегда имеется действительное положительное собственное число. Этому положительному числу, превосходящему по модулю все остальные собственные числа, соответствует собственный вектор с положительными координатами. Данный вектор может быть интерпретирован в качестве весового вектора (вектора весовых коэффициентов), служащего решением задачи.
Для нахождения максимального по модулю собственного значения и соответствующего собственного вектора матрицы с указанными свойствами может быть применен достаточно просто реализуемый степенной метод (Деммель, с. 165), алгоритм которого записывается как $y_{i+1} = Ax_i$,

$x_{i+1} = y_{i+1} / \|y_{i+1}\|_2$,

$\lambda_{i+1} = x_{i+1}^T A x_{i+1}$,

где i – номер итерации,
x – искомый собственный вектор весовых коэффициентов,
λ – соответствующее собственное число (в решении задачи не используется),
A – матрица парных сравнений.
Итерационный процесс повторяется циклически до достижения требуемой точности, заданной малой величиной типа 0,000001. Для оценки точности обычно используется сравнение Евклидовых норм собственных векторов, вычисленных на текущей и на предыдущей итерации.

См. учебное пособие Тиняковой.

16.3.2. Групповое оценивание

Метод группового оценивания в качестве исходных данных для анализа использует матрицу опроса.
Пусть P – матрица опроса, имеющая размеры n строк на m столбцов, где n – количество объектов, m – количество экспертов.
Квадратные матрицы PP^T и P^TP обладают следующими основными свойствами:

- Матрицы симметрические.
- Матрицы действительные.
- Матрицы положительно полуопределенные.

Для таких матриц все собственные значения неотрицательны, в силу чего все собственные вектора действительные.
Собственный вектор p размером n, соответствующий максимальному собственному числу матрицы PP^T, может быть интерпретирован в качестве вектора групповой оценки (весовых коэффициентов объектов).
Собственный вектор ν размером m, соответствующий максимальному собственному числу матрицы P^TP, может быть интерпретирован в качестве вектора компетентности экспертов (весовых коэффициентов компетентности). К данному параметру – коэффициентам компетентности, как и к прочим результатам анализа, следует относиться внимательно. Не следует понимать большое значение коэффициента как признак профессиональной компетентности эксперта. Данный весовой коэффициент всего лишь означает близость оценки этого эксперта к некоторой согласованной оценке всей группы экспертов. Не нужно пояснять, что большинство по объективным и субъективным причинам может ошибаться весьма часто.
Для нахождения всех собственных значений и соответствующих собственных векторов матрицы с указанными свойствами может быть применен один из многочисленных алгоритмов, например, достаточно эффективный для данного типа задач метод Якоби (Уилкинсон с соавт., с. 182).
Все изложенные в настоящем разделе методики анализа могут быть реализованы непосредственно с помощью методов главы «Матричная и линейная алгебра».

См. учебное пособие Тиняковой.

16.3.3. Коэффициент конкордации

Коэффициент конкордации (согласованности) Кендалла предназначен для исследования,

хорошо ли согласуются друг с другом представленные экспертами ранжировки. Вычисление коэффициента конкордации производится по формуле

$$W = \frac{\sum_{i=1}^{n}\left(\sum_{j=1}^{k} x_{ij} - \frac{k(n+1)}{2}\right)^2}{\frac{1}{12}k^2 n(n^2-1) - k\sum_{j=1}^{k} B_j},$$

где x_{ij}, $i = 1,2,...,n$; $j = 1,2,...,k$ – массив ранговых оценок,
n – число объектов,
k – число экспертов.
B_j, $j = 1,2,...,k$ – поправки на объединение рангов в оценках экспертов, вычисляемые по формуле

$$B_{\cdot} = \frac{1}{12}\sum_{l=1}^{m} n_l(n_l^2 - 1),$$

где m – число групп объединенных рангов в данной экспертной оценке,
n_i, $i = 1,2,...,l$ – число рангов в i–ой группе.
Программа вычисления коэффициента конкордации составлена таким образом, что можно вводить исходные данные как в виде массива ранговых оценок, так и в виде массива количественных экспертных оценок, если такие будут иметь место. В последнем случае ранжирование будет выполнено автоматически.
Программой оценивается значимость вычисленного показателя на том основании, что величина $k(n-1)W$ распределена как χ^2 с числом степеней свободы, равном $n-1$. Малая величина вычисленного программой P–значения означает, что представленные экспертами ранжировки хорошо согласованы. В противном случае можно предположить, что ранжировки неоднородны. В этом случае рекомендуется применить методы кластерного анализа (см. главу «Кластерный анализ») с использованием соответствующей случаю мерой различия для выявления согласованных групп экспертов.

См. в книги Айвазяна с соавт., Большева с соавт., Джонсона с соавт., работе Шмерлинга. Связь со статистикой Фридмана рассмотрена Тюриным.

16.3.4. Метод средних рангов

Метод средних рангов (средних арифметических рангов, упорядочение по сумме рангов) представляет собой разумный выбор согласованного мнения группы экспертов, матрица опроса которых представляет собой ранжировки.
Суть метода заключается в следующем:

- Мнения экспертов ранжируются (если это не было сделано заранее).
- Подсчитывается сумма рангов для каждого объекта.
- Массив сумм рангов объектов ранжируется, представляя решение задачи.

См. монографию Кендалла (Кендэла), статью Шмерлинга.

16.3.5. Медиана Кемени

Медиана Кемени (медиана Кемени–Снелла, Kemeny–Snell median) представляет собой выбор согласованного мнения группы экспертов, матрица опроса которых представляет собой ранжировки.
Пусть матрица опроса имеет размеры n строк на m столбцов, где n – количество объектов, m

– количество экспертов. Запишем заданное множество ранжировок как $\{P_1, P_2,...,P_m\}$. Пусть $d(P,P_i)$ – расстояние между произвольной ранжировкой P и ранжировкой P_i, $i = 1,2,...,m$. Тогда некоторая ранжировка P, принадлежащая множеству заданных ранжировок и удовлетворяющая выражению

$$M\{P_1, P_2,..., P_m\} = \arg\min_P \sum_{i=1}^{m} d(P, P_i),$$

называется медианой Кемени. Напомним, что обобщение понятия медианы (медианы множества) на произвольные шкалы введено нами в главе «Описательная статистика». Расстояние между ранжировками k и l определяется по формуле

$$d(P_k, P_l) = \sum_{i=1}^{n}\sum_{j=1}^{n}\left|p_{ij}^{(k)} - p_{ij}^{(l)}\right|,$$

где $p_{ij}^{(.)}, i = 1,2,...,n; j = 1,2,...,n,$ – элементы матриц отношений частичного порядка ранжировок k и l, соответственно, которые автоматически вычисляются программой на основе матрицы опроса, как это описано в разделе «Обработка экспертных оценок».
По определению медиана Кемени ищется только среди ранжировок, заданных анализируемой матрицей опроса. Решением, таким образом, будет ранжировка, представленная одним из экспертов, поэтому дополнительно программа выдает номер этого эксперта.

Предыстория вопроса и основные методы рассмотрены в докладе Буры (Bury). О медиане Кемени см. монографии Литвака, книгу Тюрина. Примеры практического применения даны в статьях Богомолова, Глухова с соавт. Связь медианы Кемени с другими показателями (например, коэффициентами ранговой корреляции) рассмотрена в брошюре Тюрина, обзоре Шмерлинга с соавт. (см. также указанные там ссылки).

16.3.6. Среднее Кемени

Среднее значение Кемени (среднее Кемени–Снелла, Kemeny–Snell mean) представляет собой выбор согласованного мнения группы экспертов, матрица опроса которых представляет собой ранжировки.
Пусть матрица опроса имеет размеры n строк на m столбцов, где n – количество объектов, m – количество экспертов. Запишем заданное множество ранжировок как $\{P_1, P_2,...,P_m\}$. Пусть $d(P,P_i)$ – расстояние между произвольной ранжировкой P и ранжировкой P_i, $i = 1,2,...,m$. Тогда некоторую произвольную ранжировку (без связей, иначе, без совпадающих вариант) P, удовлетворяющую выражению

$$M\{P_1, P_2,..., P_m\} = \arg\min_P \sum_{i=1}^{m} d(P, P_i),$$

назовем средним Кемени. Напомним, что обобщение понятия среднего значения на произвольные шкалы измерения введено нами в главе «Описательная статистика». Расстояние между ранжировками k и l определяется по формуле

$$d(P_k, P_l) = \sum_{i=1}^{n}\sum_{j=1}^{n}\left|p_{ij}^{(k)} - p_{ij}^{(l)}\right|,$$

где $p_{ij}^{(.)}, i = 1,2,...,n; j = 1,2,...,n,$ – элементы матриц отношений частичного порядка ранжировок k и l, соответственно, которые автоматически вычисляются программой на основе матрицы опроса.
По определению, среднее Кемени ищется среди всех $n!$ возможных ранжировок весьма

неэффективным методом полного перебора, поэтому ввиду трудоемкости вычислений число вариант искусственно ограничено нами величиной 8. Данное ограничение – предельное для диалоговой системы с использованием метода генерации перестановок стандартным антилексикографическим методом и для современного уровня развития вычислительной техники. Данной величины достаточно для многих практических применений. О сложности задачи см. статью Вакабаяси (Wakabayashi). Отметим также разработки Литвака (1982), посвященные решению проблемы вычислительной сложности, а также Тюрина.
Для ранжировок без связей среднее Кемени, как правило, совпадает с медианой Кемени, для которой программой не вводится никаких ограничений на численность выборки. Принципиальное различие между данными показателями будет наблюдаться в случае обработки ранжировок со связями, поэтому применение среднего Кемени рекомендуется в задачах такого рода.

Предыстория вопроса и основные методы решения рассмотрены в докладе Буры (Bury). См. также Тюрина, Богомолова, Глухова с соавт., обзор Шмерлинга с соавт. Обзор популярных методов генерации перестановок дал Липский (включая тексты алгоритмов на псевдокоде). О генерации перестановок см. также учебник Новикова.

16.3.7. Альфа Кронбаха

Статистика альфа Кронбаха (Cronbach’s alpha) применяется для оценки надежности статистических тестов. В литературе представлены несколько методов расчета альфы. В представленной программе альфа рассчитывается по формуле

$$\alpha = \frac{SS_{row} - SS}{SS_{row}},$$

где $SS_{row} = \frac{1}{c}\sum_{i=1}^{r} T_{i.}^2 - \frac{T_{..}^2}{rc}$ – средний квадрат строк,

$SS = \sum_{i=1}^{r}\sum_{j=1}^{c} x_{ij}^2 - \frac{T_{..}^2}{rc}$ – средний квадрат погрешности,

$T_{i.} = \sum_{j=1}^{c} x_{ij}, i = 1,2,...,r$ – суммы строк,

$T_{..} = \sum_{i=1}^{r}\sum_{j=1}^{c} x_{ij}$ – общая сумма,

c – число столбцов (выборок),
r – число строк (параметров).
Рассчитываются также доверительные интервалы оцениваемой альфы. Нижняя граница доверительного интервала оцениваемой альфы считается как

$$L_\alpha = 1 - (1-\alpha) F_{r-1,(r-1)(c-1)}^{-1}(1-(1-\beta)/2),$$

где $F_{...}^{-1}(.)$ – обратная функция *F*–распределения,
β – доверительный уровень, выраженный в долях.
Верхняя граница доверительного интервала оцениваемой альфы считается как

$$H_\alpha = 1 - (1-\alpha) F_{r-1,(r-1)(c-1)}^{-1}((1-\beta)/2).$$

См. монографии Аванесова, Цыганова, Суен (Suen) и Суен с соавт., отчеты Фельдт (Feldt),

Конинг (Koning) с соавт., Ли (Li) с соавт., Гутри (Guthrie), статьи Мягкова, Фельдт с соавт., Ван Зил (Van Zyl) с соавт., Якобуччи (Iacobucci) с соавт., Бланд (Bland) с соавт., Кистнер (Kistner) с соавт., Бонетт (Bonett), Панди (Pandey) с соавт., Суен с соавт., Аванесова.

Список использованной и рекомендуемой литературы

1. Bland J.M., Altman D.G. Statistics notes: Cronbach's alpha // British Medical Journal, 22 February 1997, vol. 314, pp. 572.
2. Bogart K.P. Preference structures I: Distances between transitive preference relations // Journal of Mathematical Sociology, 1973, vol. 3, pp. 49–67.
3. Bogart K.P. Preference structures II: Distances between asymmetric relations // SIAM Journal on Applied Mathematics, September 1975, vol. 29, no. 2, pp. 254–262.
4. Bonett D.G. Sample size requirements for testing and estimating coefficient alpha // Journal of Educational and Behavioral Statistics, 2002, vol. 27, no. 4, pp. 335–340.
5. Bury H. Kemeny's median algorithm: Application for determining group judgement // 16th JISR–IIASA Workshop on Methodologies and Tools for Complex System Modeling and Integrated Policy Assessment, July 15–17, 2002, IIASA, Laxenburg, Austria.
6. Chebotarev P.Yu., Shamis E. Characterizations of scoring methods for preference aggregation // Annals of Operations Research, January 1998, no. 0, pp. 299–332.
7. Chebotarev P.Yu., Shamis E. Preference fusion when the number of alternatives exceed two: Indirect Scoring procedures // Journal of the Franklin Institute, 1999, vol. 336, no. 2, pp. 205–226.
8. Feldt L.S. Statistical tests and confidence intervals for Cronbach's coefficient alpha // Iowa Testing Programs Occasional Papers Number 33.
9. Feldt L.S., Woodruff D.J., Salih F.A. Statistical Inference for coefficient alpha // Applied Psychological Measurement, 1987, vol. 11, no. 1, pp. 93–103.
10. Genest Ch., Rivest L.A. Statistical look at Saaty's method of estimating pairwise preferences expressed on a ratio scale // Journal of Mathematical Psychology, 1994, vol. 38, pp. 477–496.
11. Guthrie A.C. A review of coefficient alpha and some basic tenets of classical measurement theory // Proceedings of the Annual Meeting of the Southwest Educational Research Association, Dallas, TX, 27–29 January 2000.
12. Iacobucci D., Duhachek A. Advancing alpha: Measuring reliability with confidence // Journal of Consumer Psychology, 2003, vol. 13, no. 4, pp. 478–487.
13. Kemeny J. Mathematics without numbers // Daedalus, 1959, vol. 88, pp. 577–591.
14. Kistner E.O., Muller K.E. Exact distributions of intraclass correlation and Cronbach's alpha with Gaussian data and general covariance // Psychometrika, September 2004, vol. 69, no. 3, pp. 459–474.
15. Klamler C. Kemeny's rule and Slater's rule: A binary comparison // Economic Bulletin, 2003, vol. 4, no. 35, pp. 1–7.
16. Koning A.J., Franses P.H. Confidence intervals for Cronbach's coefficient alpha values // ERIM Report Series Reference No. ERS–2003–041–MKT.
17. Li J.C., Woodruff D.J. Bayesian statistical inference for coefficient alpha // ACT Research Report Series, January 2002, no. 2.
18. Linstone H.A. The Delphi method: Techniques and application / Ed. by H.A. Linstone, M. Turoff. – Reading, MA: Addison Wesley, 1975.
19. Osgood C.E., Suci G.J., Tannenbaum P.H. The measurement of meaning. – Chicago, IL: University of Illinois Press, 1967.
20. Pandey T.N., Hubert L. An empirical comparison of several interval estimation procedures for coefficient alpha // Psychometrika, June 1975, vol. 40, no. 2, pp. 169–181.

21. Saari D.G., Merlin V.R. A geometric examination of Kemeny's rule // Social Choice and Welfare, 2000, vol. 17, pp. 403–438.
22. Suen H.K. Principles of test theories. – Hillsdale, NJ: Erlbaum, 1990.
23. Suen H.K., Ary D. Analyzing quantitative behavioral observation data. – Hillsdale, NJ: Erlbaum, 1989.
24. Suen H.K., Lei P.W. Classical versus generalizability theory of measurement // Educational Measurement, 2007, vol. 4, pp. 3–20.
25. Van Zyl J.M., Neudecker H., Nel D.G. On the distribution of the maximum likelihood estimator of Cronbach's alpha // Psychometrika, Septembr, 2000, vol. 65, no. 3, pp. 271–280.
26. Wakabayashi Y. The complexity of computing medians of relations // Resenhas, 1998, vol. 3, no. 3, pp. 323–349.
27. Аванесов В.С. Введение в статистические и математические методы педагогических измерений // Педагогические измерения, 2005, № 4, с. 91–116.
28. Аванесов В.С. Тесты в социологическом исследовании. – М.: Наука, 1982.
29. Айвазян С.А., Енюков И.С., Мешалкин Л.Д. Прикладная статистика: Исследование зависимостей. – М.: Финансы и статистика, 1985.
30. Бешелев С.Г., Гурвич Ф.Г. Экспертные оценки. – М.: Наука, 1973.
31. Бешелев С.Д., Гурвич Ф.Г. Математико–статистические методы экспертных оценок. – М.: Статистика, 1980.
32. Богомолов А.В. Использование лингвистических переменных и методов обработки экспертной информации для автоматизированного распознавания ранних стадий нарушений функционального состояния человека // Информационные технологии, 2000, № 8, с. 50–54.
33. Большев Л.Н., Смирнов Н.В. Таблицы математической статистики. – М.: Наука, 1983.
34. Гаек Я., Шидак З. Теория ранговых критериев. – М.: Наука, 1971.
35. Гантмахер Ф.Р. Теория матриц. – М.: Наука, 1988.
36. Гитис В.Г., Ермаков Б.В. Основы пространственно–временного прогнозирования в геоинформатике. – М.: Физматлит, 2004.
37. Гладышевский А.М. Методы и модели экономического прогнозирования. – М.: Экономика, 1977.
38. Глухов А.И., Погодаев А.К. Медиана Кемени в определении приоритетов развития предприятий // Управление большими системами, 2006, выпуск 14, с. 40–45.
39. Голанский М.М. Экономическое прогнозирование. – М.: Наука, 1983.
40. Громов Л.М. Руководство по научно–техническому прогнозированию / Под ред. Л.М. Громова. – М.: Прогресс, 1977.
41. Деммель Дж. Вычислительная линейная алгебра. Теория и приложения. – М.: Мир, 2001.
42. Дерзский В.Г., Нечай Т.А. Прогнозирование технико–экономических параметров новой техники. – Киев: Наукова думка, 1982.
43. Джонсон Н., Лион Ф., Статистика и планирование эксперимента в технике и науке. Методы обработки данных. – М.: Мир, 1980.
44. Диков Э. Квалиметрия // Юный техник, 1970, № 2, с. 19.
45. Добров Г.М. Прогнозирование для промышленности и правительственных учреждений / Под ред. Г.М. Доброва. – М.: Прогресс, 1972.
46. Добров Г.М. Прогнозирование науки и техники. – М.: Наука, 1977.
47. Завлин П.Н., Казанцев А.К. Экономика и управление в отраслевых НТО. – М.: Экономика, 1990.
48. Канторович Л.В. Математические методы организации и планирования производства.

– Л.: Издательство ЛГУ, 1939.
49. Кемени Дж., Снелл Дж. Кибернетическое моделирование: Некоторые приложения. – М.: Советское радио, 1972.
50. Кендэл М. Ранговые корреляции. – М: Статистика, 1975.
51. Кини Р.Л., Райфа Х. Принятие решений при многих критериях: предпочтения и замещения. – М.: Радио и связь, 1981.
52. Крымский С.Б. Экспертные оценки в социологических исследованиях / Под ред. С.Б. Крымского. – Киев: Наукова Думка, 1990.
53. Лимер Э.Э. Статистический анализ неэкспериментальных данных: Выбор формы связи. – М.: Финансы и статистика, 1983.
54. Липский В. Комбинаторика для программистов. – М.: Мир, 1988.
55. Литвак Б.Г. Экспертная информация. Методы получения и анализа. – М.: Радио и связь, 1982.
56. Литвак Б.Г. Экспертные технологии в управлении. – М.: Дело, 2004.
57. Марселлус Д. Программирование экспертных систем на Турбо Прологе. – М.: Финансы и статистика, 1994.
58. Мартино Г. Технологическое прогнозирование. – М.: Прогресс, 1977.
59. Мосин В.Н., Крук Д.М. Основы экономического и социального прогноза. – Л.: Высшая школа, 1985.
60. Мягков А.Ю. Шкалы лжи из опросника MMPI: Опыт экспериментальной валидации // Методика и техника социологических вычислений, 2002, с. 117–130.
61. Нейлор К. Как построить свою экспертную систему. – М.: Энергоатомиздат, 1991.
62. Новиков Ф.А. Дискретная математика для программистов. Учебник для вузов. – СПб.: Питер, 2005.
63. Райхман Э.П., Азгальдов Г.Г. Экспертные методы в оценке качества товаров. – М.: Экономика, 1974.
64. Саати Т., Кернс К. Аналитическое планирование. Организация систем. – М.: Радио и связь, 1991.
65. Сифоров В.И. Прогностика. – М.: Наука, 1990.
66. Сойер Б., Фостер Д.Л. Программирование экспертных систем на Паскале. – М.: Финансы и статистика, 1990.
67. Стрижов В.В. Согласование экспертных оценок для биосистем в экстремальных условиях. Сообщения по прикладной математике. Научное издание. – М.: ВЦ РАН 2002.
68. Стрижов В.В. Уточнение экспертных оценок с помощью измеряемых данных // Заводская лаборатория. Диагностика материалов, 2006, № 7, с. 59–64.
69. Твисс Б. Управление научно–техническими нововведениями. – М.: Экономика, 1989.
70. Тинякова В.И. Математические методы обработки экспертной информации: Учебное пособие. – Воронеж: Издательство ВГУ, 2006.
71. Тюрин Ю.Н. Непараметрические методы статистики. – М.: Знание, 1978.
72. Уилкинсон, Райнш. Справочник алгоритмов на языке АЛГОЛ. Линейная алгебра. – М.: Машиностроение, 1976.
73. Фишберн Р. Теория полезности для принятия решений. – М.: Наука, 1978.
74. Хамханова Д.Н. Теоретические основы обеспечения единства экспертных измерений. – Улан–Удэ: Издательство ВСГТУ, 2006.
75. Цыганов Ш.И. Математическая обработка результатов педагогического тестирования. – Уфа: РИО БашГУ, 2007.
76. Цыганов Ш.И. Математические теории педагогических измерений. – Уфа: Эдвис, 2007.

77. Чабровский В.А. Прогнозирование развития науки и техники. – М.: Экономика, 1983.
78. Четыргин Е.Н. Статистические методы прогнозирования. – М.: Статистика, 1975.
79. Шмерлинг Д.С. О проверке согласованности экспертных оценок // В сб. Статистические методы анализа экспертных оценок. Ученые записки по статистике, т. 29 / Под ред. Ю.Н. Тюрина, А.А. Френкель. – М.: Наука, 1977, с. 77–83.
80. Шмерлинг Д.С. Экспертные оценки. Методы и применение (обзор) / Д.С. Шмерлинг, С.А. Дубровский, Т.Д. Аржанова и др. // В сб. Статистические методы анализа экспертных оценок. Ученые записки по статистике, т. 29 / Под ред. Ю.Н. Тюрина, А.А. Френкель. – М.: Наука, 1977, с. 290–382.
81. Эйтингон В.Н. Методы организации экспертизы и обработки экспертных оценок в менеджменте: Учебно–методическое пособие / В.Н. Эйтингон, М.А. Кравец, Н.П. Панкратова и др. – Воронеж: Издательство ВГУ, 2004.
82. Элти Дж., Кумбс М. Экспертные системы: концепции и примеры. – М.: Финансы и статистика, 1987.
83. Эрох Я. Прогнозирование НТП. – М.: Прогресс, 1974.

Глава 17. Анализ выживаемости

17.1. Введение

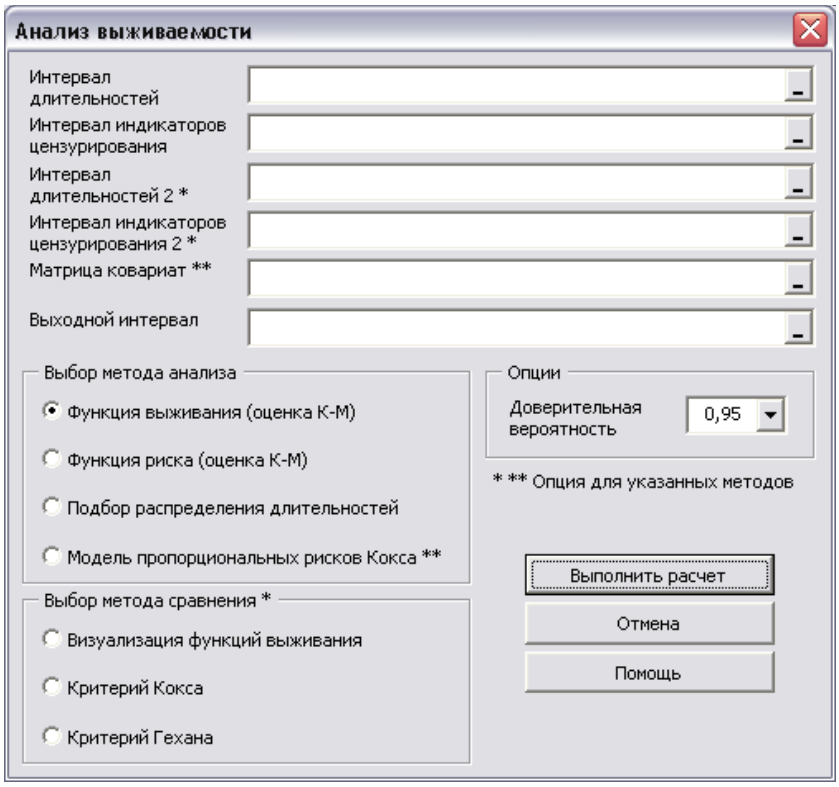
Программное обеспечение анализа выживаемости (анализа данных типа времени жизни) исследует особые объекты, имеющие в различных отраслях знаний следующие наименования:

- в медико–биологических науках – время жизни,
- в общественных науках – длительность до момента прекращения,
- в технических науках – наработка до отказа.

Представленные методы применимы во всех перечисленных областях, хотя изначально разработка выполнялась для медицинских приложений. Вследствие этого некоторые специфические отраслевые особенности могут не учитываться. Например, при испытаниях технических систем отказавшие элементы могут быть заменены новыми элементами или отремонтированы, после чего испытания продолжены. Данная ситуация, однако, невозможна при исследовании выживаемости пациентов с определенной критической для жизни патологией. Далее, в технической диагностике испытания могут быть выполнены и повторены в любое удобное время. При исследовании же, к примеру, длительности времени забастовок организовать их специально в научных целях не представляется возможным.

17.2. Работа с программным обеспечением

Выберите из меню программы пункт AtteStat | Анализ выживаемости. На экране появится диалоговое окно, изображенное на рисунке:



Затем проделайте следующие шаги:

- Выберите или введите интервал длительностей.
- Выберите или введите интервал индикаторов цензурирования. Индикаторы могут принимать только значения 0 (пациент цензурирован, т. е. выбыл из исследования, и его состояние неизвестно, или умер по причине, не связанной с исследуемой патологией) или 1 (пациент умер по причине, связанной с исследуемой патологией).
- Выберите или оставьте назначенный по умолчанию стандартный доверительный уровень, необходимый для построения доверительных интервалов.
- Если предполагается использовать методы сравнения, выберите или введите интервал длительностей и интервал индикаторов цензурирования для второй выборки подобно тому, как это сделано для анализа единственной выборки.
- Если предполагается использовать модель пропорциональных рисков Кокса, дополнительно выберите или введите интервал матрицы ковариат. При этом число строк данного интервала должно совпадать с числом строк интервала длительностей и индикаторов цензурирования, а число столбцов должно равняться числу ковариат.
- Выберите или введите выходной интервал. Начиная с первой ячейки выходного интервала (следовательно, можно указать только одну ячейку, т.к. остальные ячейки интервала игнорируются), будут выведены результаты анализа.
- Отметьте требуемый метод расчета.
- Нажмите кнопку «Выполнить расчет».

При ошибках, вызванных неверными действиями пользователя, выдаются сообщения об ошибках.

17.2.1. Сообщения об ошибках

При ошибках ввода и во время выполнения программы могут выдаваться диагностические сообщения следующих типов:

Ошибка	Комментарий
Пустая ячейка в области данных.	Проверьте исходные данные и заполните все ячейки, отмеченные Вами как входной интервал. Во избежание ошибок расчета, вызванных разногласиями, трактовать ли пустую ячейку как нуль, программа требует заполнения всех ячеек. Если в ячейке не должно быть данных по физической природе исследуемого процесса, введите в данную ячейку нуль.
Нечисловой тип данных.	Проверьте типы ячеек входного интервала. Тип может быть только числовым. Проще всего выделить интервал ячеек и явно определить их тип как числовой стандартными средствами.
Не определена область данных.	Вы не выбрали или неверно ввели входной интервал. Лучшим способом избежать ошибки является не ввод, а выделение интервала стандартным образом, т. е. протаскиванием курсора. Обратим внимание, что некоторые представленные методы требуют не только ввода обязательного набора исходных данных (это интервал длительностей и интервал индикаторов цензурирования), но и аналогичный набор для второй выборки (в случае выбора методов сравнения) либо матрицы ковариат (для модели Кокса).
Не определена область вывода.	Вы не выбрали или неверно ввели выходной интервал. Лучшим способом избежать ошибки является не ввод, а выделение интервала стандартным образом, т. е. протаскиванием курсора.
Неверные индикаторы.	Неверное число в интервале индикаторов цензурирования. Индикаторы могут принимать только значения 0 (пациент цензурирован, т. е. выбыл из исследования, и его состояние неизвестно, или умер по причине, не связанной с исследуемой патологией) или 1 (пациент умер по причине, связанной с исследуемой патологией).
Все данные цензурированы.	Индикаторы цензурирования указывают, что все данные цензурированы. Цензурирование 100% данных препятствует проведению статистического анализа представленными методами.

17.3. Теоретическое обоснование

Методы анализа выживаемости могут применяться как в клинических исследованиях, так и при оценке надежности технических систем по цензурированным выборкам.
Предлагаются следующие методы расчета, обычно применяемые в анализе выживаемости:

- Вычисление оценки функции выживания.
- Вычисление оценки функции риска.
- Подбор теоретического распределения.
- Критерий Кокса.
- Критерий Гехана.
- Модель пропорциональных рисков Кокса.

Все методы предполагают использование особых величин, называемых индикаторами цензурирования. Индикаторы могут принимать только следующие стандартные значения, «понимаемые» программой:

- 1 – пациент умер по причине, связанной с исследуемой патологией,
- 0 – пациент цензурирован, т. е. выбыл из исследования, и его состояние либо неизвестно в момент исследования, либо он умер по причине, не связанной с

исследуемой патологией.
Медицинская терминология, использованная в предыдущем абзаце, естественно может быть обобщена на технические, социальные и любые другие системы.

17.3.1. Функция выживания

Оценка Каплана–Мейера функции выживания, в источниках называемая также множительной оценкой, вычисляется по формуле

$$\hat{S}(t \geq t_j) = \hat{S}(t_j) = \prod_{i=1}^{j} \left(1 - \frac{d_i}{r_i}\right), j = 1,2,...,K,$$

где d_i – количество наблюдений, моменты прекращения которых наблюдались с длительностью t_i, $i = 1,2,...,K$,
K – число моментов прекращения,
r_i – количество наблюдений, незаконченных либо цензурированных к моменту t_i, $i = 1,2,...,K$, причем

$$r_j = \sum_{i=j}^{K} (m_i + d_i), j = 1,2,...,K,$$

где m_i – количество наблюдений, цензурированных между моментами t_i и t_{i+1}.
Дисперсия оценки Каплана–Мейера функции выживания вычисляется по формуле Гринвуда

$$D\hat{S}(t_j) = \left[\hat{S}(t_j)\right]^2 \sum_{i=1}^{j} \frac{d_i}{r_i(r_i - d_i)}, j = 1,2,...,K.$$

На хвосте распределения оценка по формуле Гринвуда может не существовать, поэтому в данном случае используется формула Пето

$$D\hat{S}(t_j) = \left[\hat{S}(t_j)\right]^2 \frac{1 - \hat{S}(t_j)}{r_j}.$$

Доверительный интервал оцениваемой функции выживания вычисляется по асимптотической формуле

$$I_S(t_j) = \left(\hat{S}(t_j) - \Psi((1+\beta)/2)\sqrt{D\hat{S}(t_j)}; \hat{S}(t_j) + \Psi((1+\beta)/2)\sqrt{D\hat{S}(t_j)}\right), j = 1,2,...,K,$$

где Ψ(.) – обратная функция стандартного нормального распределения,
β – доверительный уровень, выраженный в долях.
В случае малых выборок для вычисления доверительного интервала предлагается использовать уточненную формулу

$$I_S(t_j) = \left(\hat{S}(t_j)^{\exp\left[-\Psi((1+\beta)/2)\sqrt{D\hat{S}(t_j)}/\left(\hat{S}(t_j)\log\hat{S}(t_j)\right)\right]}; \hat{S}(t_j)^{\exp\left[\Psi((1+\beta)/2)\sqrt{D\hat{S}(t_j)}/\left(\hat{S}(t_j)\log\hat{S}(t_j)\right)\right]}\right), j = 1,2,...,K.$$

Функция выживания изображается в виде ступенчатого графика. Введение необходимого числа фиктивных точек в программе позволило получить данный тип графика. Эти точки перечислены в разделе «Данные для графика функции выживания» листинга результатов расчета на рабочем листе программы и предназначены только для построения графика.

См. монографии Кокса с соавт., Власова, Аален (Aalen) с соавт.

17.3.2. Функция риска

Оценка Каплана–Мейера функции риска вычисляется по формуле

$$\hat{H}(t \ge t_j) = \hat{H}(t_j) = \sum_{i=1}^{j} \frac{d_i}{r_i},\ j = 1,2,...,K,$$

где d_i – количество наблюдений, моменты прекращения которых наблюдались с длительностью t_i, $i = 1,2,...,K$,
K – число моментов прекращения,
r_i – количество наблюдений, незаконченных либо цензурированных к моменту t_i, $i = 1,2,...,K$, причем

$$r_j = \sum_{i=j}^{K} (m_i + d_i),\ j = 1,2,...,K,$$

где m_i – количество наблюдений, цензурированных между моментами t_i и t_{i+1}.
Дисперсия оценки Каплана–Мейера функции риска вычисляется по формуле

$$D\hat{H}(t_j) = \frac{D\hat{S}(t_j)}{\hat{S}(t_j)},\ j = 1,2,...,K,$$

где $D\hat{S}(t_j), j = 1,2,...,K,$ – дисперсия функции выживания,
$\hat{S}(t_j), j = 1,2,...,K,$ – оценка Каплана–Мейера функции выживания.
Доверительный интервал оцениваемой функции риска вычисляется как

$$I_H(t_j) = \left(\hat{H}(t_j) - \Psi((1+\beta)/2)\sqrt{D\hat{H}(t_j)}; \hat{H}(t_j) + \Psi((1+\beta)/2)\sqrt{D\hat{H}(t_j)} \right),\ j = 1,2,...,K,$$

где $\Psi(.)$ – обратная функция стандартного нормального распределения,
β – доверительный уровень, выраженный в долях.
Функция риска обычно изображается либо в виде ступенчатого графика (см. раздел, посвященный функции выживания) либо в виде ломаной линии, соединяющей заданные точки, соответствующие моментам прекращения. В программе был выбран второй из указанных типов.

См. монографии Кокса с соавт., Власова, Аален (Aalen) с соавт., статью Аален.

17.3.3. Оценка параметра положения

Среднее значение (параметр положения) и дисперсия (параметр масштаба) длительности в задаче анализа данных типа времени жизни, как параметрические оценки параметра положения и параметра масштаба (если он имеет смысл), зависят от выбранного типа теоретического распределения представленных эмпирических данных.
Поэтому точечные оценки данных параметров см. в разделе «Подбор распределения».
Отметим, что в случае цензурированных выборок задача существенно сложнее обычной постановки. Для некоторых моделей оценок параметров в виде простых формул не существует – оценка возможна лишь вычислительными методами.
Непараметрические оценки параметра положения – точечная оценка медианы и ее интервальная оценка – обычно изучаются после изучения раздела, посвященного непараметрической оценке Каплана–Мейера. Точечная оценка медианы имеет непосредственное отношение к оценке Каплана–Мейера, вычисляется через нее. Поэтому в меню предлагаемого программного обеспечения отдельная позиция для оценки медианы длительности отсутствует, а вывод вычисленного значения данной оценки производится сразу после вывода функции выживания.
Точечная оценка медианы имеет вид

$\hat{m}_{0,5} = \inf\{t : \hat{S}(t) \le 0{,}5\}$,

где $\hat{S}(t)$ – оценка Каплана–Мейера функции выживания,
t – длительность.
Последняя формула означает, что в качестве точечной оценки медианы берется такая минимальная длительность t (из представленных эмпирических длительностей), для которой выполняется неравенство $\hat{S}(t) \le 0{,}5$.
Доверительный интервал оцениваемой медианы задается формулой
$I_m = (y_c; y_{n+1-c})$,
где y_i, $i = 1,2,...,n$, т. е. упорядоченные по возрастанию длительности,
c – параметр, вычисляемый по формуле
$c = [n / 2 - \Psi((1 + \beta) / 2)n^{1/2} / 2]$,
где [.] – целая часть числа,
$\Psi(.)$ – обратная функция стандартного нормального распределения,
β – доверительный уровень, выраженный в долях.

См. монографию Дезу (Desu) с соавт. О вычислении точечной и интервальной оценки медианы см. статью Орлова, книгу Холлендера с соавт. Предлагаемая в статьях Брукмейер (Brookmeyer) с соавт., Баркер (Barker) методика вычисления непараметрических доверительных интервалов оцениваемой медианы длительности через доверительные интервалы оцениваемой функции выживания представляется сомнительной.

17.3.4. Подбор распределения

В разделе изучается проблема подбора теоретического распределения (подгонка, fitting distribution) для эмпирического распределения длительностей, в том числе цензурированных, следующими типами подходящих по теоретическим соображениям распределений, обычно применяемыми для данной задачи:

- логнормальное распределение,
- логлогистическое распределение,
- гамма–распределение,
- распределение Вейбулла,
- экспоненциальное распределение,
- распределение Рэлея,
- распределение Гомпертца.

Функция распределения (с точностью до параметров) может быть известна также из теоретических соображений. В таком случае задача вычисления параметров распределений может трактоваться как идентификация математической модели. Параметрическая статистическая модель выживания описывается с помощью следующих функций:

- функция плотности распределения $f(t)$,
- [кумулятивная или интегральная] функция распределения $F(t) = P(T \le t)$ – функция распределения длительностей до момента отказа t,
- функция выживания $S(t) = 1 - F(t)$ – вероятность безотказной работы до момента t,
- функция интенсивности отказов (функция риска) $h(t) = \frac{f(t)}{1 - F(t)}$.

Под длительностью t может пониматься время жизни, количество циклов до отказа и т.п., в зависимости от конкретной задачи.
Методика практического применения данных функций рассмотрена в монографии Хана с

соавт. Отметим, что в источниках все упомянутые функции могут быть записаны в отличных друг от друга, однако, эквивалентных формах. Поэтому в описании каждого распределения указаны названия параметров (традиционно записанных в формулах греческими литерами), данные на листинге.
Теоретические значения частот [непрерывных] распределений элементарно вычисляются как произведение плотности теоретического распределения на численность выборки и на длину соответствующего классового интервала.
В программе теоретические распределения всех указанных типов выводятся на том же поле графика, что и эмпирическое распределение, кривыми различных цветов, как указано легендой графика.

Задача подгонки теоретических распределений к цензурированным данным существенно сложнее обычной постановки задачи, с которой можно ознакомиться по монографиям Бьюри (Bury), Эванс (Evans) с соавт. Сводка параметров распределений и обзор алгоритмов подгонки теоретических распределений к экспериментальным данным представлена в монографии Кобзаря. См. также книги Кришнамурти (Krishnamoorthy), Кляйбер (Kleiber) с соавт., статьи Лу (Lu) с соавт., Айтчисон (Aitchison) с соавт.

17.3.4.1. Общая методика

Аппроксимация эмпирического распределения опытных данных любым теоретическим стандартным распределением сводится к вычислению одним из математических методов (метод максимального правдоподобия, метод моментов, реже – метод наименьших квадратов) параметров теоретического распределения, ответственных за форму, масштаб и положение кривой распределения. Метод максимального правдоподобия является наиболее популярным методом решения рассматриваемого типа задач благодаря хорошей вычислительной устойчивости.
Реализация метода начинается с составления функции максимального правдоподобия (ФМП) в виде (возможны эквивалентные записи ФМП, с учетом введенных выше функций статистической модели, в зависимости от того, какие результаты интересуют автора исследования)

$$L(\theta)=\prod_{i=1}^{n}\{f(t_i)\}^{\delta_i}\{S(t_i)\}^{1-\delta_i},$$

где θ – вектор неизвестных параметров статистической модели,
t_i, i = 1,2,...,n – массив длительностей,
δ_i, i = 1,2,...,n – соответствующий массив индикаторов цензурирования (для конкретной длительности индикатор равен 1, если выборка нецензурирована, и 0, если цензурирована),
n – численность массива длительностей.
Отметим, что в дальнейших выкладках число нецензурированных длительностей стандартно обозначено как

$$r=\sum_{i=1}^{n}\delta_i.$$

Вектор искомых параметров находится из условия максимума ФМП:

$$L(\theta)\underset{\theta}{\rightarrow}\max.$$

Максимум ФМП находится из условия равенства нулю частных производных ФМП по искомым параметрам, т. е. искомые параметры удовлетворяют уравнениям

$$\frac{\partial L(\theta)}{\partial\theta}=0.$$

Для упрощения максимизируют не саму ФМП, а логарифм ФМП. Эта возможность основана на том факте, что ФМП и логарифм достигают максимума при одних и тех же значениях искомых параметров, однако работать с логарифмом ФМП значительно проще:

$$\frac{\partial \ln L(\theta)}{\partial \theta} = 0.$$

Задача сводится, таким образом, к аналитическому либо численному (одним из методов оптимизации) решению полученной линейной или нелинейной системы уравнений.

17.3.4.2. Логарифмические модели

Решение статистической модели может быть выполнено различными методами. Однако в любом случае стараются использовать наиболее эффективную модификацию общего метода, иногда позволяющую радикально упростить решение. Не конкретизируя тип модели (в программе предлагаются две логарифмических модели – логнормальная и логлогистическая), представим общий метод решения логарифмической двухпараметрической модели.

ФМП логарифмической двухпараметрической модели общего вида имеет вид

$$L(\theta) = \prod_{i=1}^{n} \left\{ \sigma^{-1} f\left(\frac{y_i - \mu}{\sigma} \right) \right\}^{\delta_i} \left\{ S\left(\frac{y_i - \mu}{\sigma} \right) \right\}^{1-\delta_i},$$

где $y_i = \ln t_i$, $i = 1,2,...,n$ – массив логарифмов длительностей,

$\theta = \{\mu,\sigma\}$– вектор параметров,

μ – параметр положения,

σ – параметр масштаба.

Логарифмическая ФМП может быть записана как

$$\ln L(\theta) = -r \ln \sigma + \sum_{i=1}^{n} [\delta_i \ln f(z_i) + (1-\delta_i) \ln S(z_i)],$$

где $z_i = \frac{y_i - \mu}{\sigma}, i = 1,2,...,n,$ – массив стандартизированных логарифмов длительностей.

В дальнейших выкладках понадобятся следующие очевидные выражения для производных

$$\frac{\partial z}{\partial \mu} = -\frac{1}{\sigma}, \quad \frac{\partial z}{\partial \sigma} = -\frac{z}{\sigma}.$$

Тогда компоненты вектора градиента $G(\theta)$ логарифмической ФМП по параметрам

$$g_1 = \frac{\partial \ln L(\theta)}{\partial \mu} = -\frac{1}{\sigma} \sum_{i=1}^{n} \left[\delta_i \frac{\partial \ln f(z_i)}{\partial z_i} + (1-\delta_i) \frac{\partial \ln S(z_i)}{\partial z_i} \right],$$

$$g_2 = \frac{\partial \ln L(\theta)}{\partial \sigma} = -\frac{r}{\sigma} - \frac{1}{\sigma} \sum_{i=1}^{n} \left[\delta_i z_i \frac{\partial \ln f(z_i)}{\partial z_i} + (1-\delta_i) z_i \frac{\partial \ln S(z_i)}{\partial z_i} \right].$$

Компоненты матрицы вторых производных $H(\theta)$ логарифмической ФМП по параметрам (матрицы Гессе) вычисляются как

$$h_{11} = \frac{\partial^2 \ln L(\theta)}{\partial \mu^2} = \frac{1}{\sigma^2} \sum_{i=1}^{n} \left[\delta_i \frac{\partial^2 \ln f(z_i)}{\partial z_i^2} + (1-\delta_i) \frac{\partial^2 \ln S(z_i)}{\partial z_i^2} \right],$$

$$h_{12} = h_{21} = \frac{\partial^2 \ln L(\theta)}{\partial \mu \partial \sigma} = \frac{1}{\sigma^2} \sum_{i=1}^{n} \left[\delta_i \frac{\partial \ln f(z_i)}{\partial z_i} + (1-\delta_i) \frac{\partial \ln S(z_i)}{\partial z_i} \right] +$$

$$+\frac{1}{\sigma^2}\sum_{i=1}^{n}\left[\delta_i z_i\frac{\partial^2 \ln f(z_i)}{\partial z_i^2}+(1-\delta_i)z_i\frac{\partial^2 \ln S(z_i)}{\partial z_i^2}\right],$$

$$h_{22}=\frac{\partial^2 \ln L(\theta)}{\partial \sigma^2}=\frac{r}{\sigma^2}+\frac{2}{\sigma^2}\sum_{i=1}^{n}\left[\delta_i z_i\frac{\partial \ln f(z_i)}{\partial z_i}+(1-\delta_i)z_i\frac{\partial \ln S(z_i)}{\delta z_i}\right]+$$

$$+\frac{1}{\sigma^2}\sum_{i=1}^{n}\left[\delta_i z_i^2\frac{\partial^2 \ln f(z_i)}{\partial z_i^2}+(1-\delta_i)z_i^2\frac{\partial^2 \ln S(z_i)}{\partial z_i^2}\right].$$

С учетом введенных обозначений итерационная схема максимизации логарифмической ФМП алгоритма метода Ньютона–Рафсона может быть записана как

$\theta^{j+1}=\theta^j-\left[H(\theta)\right]^{-1}G(\theta),\ j=0,1,2,...,$

где j, j = 0,1,2,... – номер итерации.

При численной реализации метода Ньютона–Рафсона должна быть учтена особенность данного метода при решении рассматриваемой задачи, заключающаяся в весьма узкой области сходимости. Поэтому начальные приближения параметров должны быть заданы достаточно близкими к оптимальному решению. В этом случае метод сходится очень быстро. Для грубой же локализации начальных приближений может применяться один из глобальных методов. В простейшем случае можно применить метод перебора с небольшим шагом по разумной области определения параметров либо, для упрощения численной реализации, один из вариантов метода спуска. Низкое быстродействие данных примитивных, но надежных методов компенсируется высоким быстродействием современных компьютеров.

Общая методика и конкретные применения метода максимального правдоподобия для подгонки статистических моделей к цензурированным данным типа времени жизни подробно рассмотрены в монографиях Коллетт (Collett) и Лелесс (Lawless), статье Юзеф (Yousef). См. также пособие Цыплакова.

17.3.4.2.1. Логнормальное распределение

Плотность логнормального (логарифмически нормального) распределения с двумя параметрами имеет вид

$$f(t)=\frac{1}{\sigma t\sqrt{2\pi}}\exp\left[-\frac{(\ln t-\mu)^2}{\sigma^2}\right],\sigma>0.$$

Введем нормированную величину

$$z=\frac{\ln t-\mu}{\sigma}.$$

Тогда можно записать

$$f(z)=\varphi(z)=\frac{1}{\sqrt{2\pi}}e^{-z^2/2},$$

где φ(.) – функция плотности стандартного нормального распределения.

Соответствующая функция выживания

$S(z) = 1 – Φ(z)$,

где Φ(.) – функция стандартного нормального распределения.

С учетом нормировки функция максимального правдоподобия (ФМП) запишется как

$$L(\mu,\sigma)=\prod_{i=1}^{n}\left\{\frac{1}{\sigma}\varphi(z_i)\right\}^{\delta_i}\left\{1-\Phi(z_i)\right\}^{1-\delta_i}.$$

Соответствующая логарифмическая ФМП имеет вид

$$\ln L(\mu,\sigma) = -r\ln\sigma - \frac{1}{2}\sum_{i=1}^{n}\delta_i z_i^2 + \sum_{i=1}^{n}(1-\delta_i)\ln[1-\Phi(z_i)].$$

Производные, необходимые для итерационной схемы метода Ньютона–Рафсона, представленной в разделе «Общая методика для логарифмических моделей», запишутся как

$$\frac{\partial \ln f(z)}{\partial z} = -z, \quad \frac{\partial \ln S(z)}{\partial z} = -\frac{f(z)}{S(z)},$$

$$\frac{\partial^2 \ln f(z)}{\partial z^2} = -1, \quad \frac{\partial^2 \ln S(z)}{\partial z^2} = \frac{zf(z)}{S(z)} - \left[\frac{f(z)}{S(z)}\right]^2.$$

17.3.4.2.2. Логлогистическое распределение

Плотность логлогистического (логарифмически логистического) распределения с двумя параметрами имеет вид

$$f(t) = \frac{1}{\sigma t}\exp\left(\frac{\ln t-\mu}{\sigma}\right)\left[1+\exp\left(\frac{\ln t-\mu}{\sigma}\right)\right]^{-2}, \sigma > 0.$$

Введем нормированную величину

$$z = \frac{\ln t-\mu}{\sigma}.$$

Тогда можно записать

$$f(z) = \frac{e^z}{(1+e^z)^2}.$$

Соответствующая функция выживания

$$S(z) = \frac{1}{1+e^z}.$$

С учетом нормировки функция максимального правдоподобия (ФМП) запишется как

$$L(\mu,\sigma) = \prod_{i=1}^{n}\left\{\frac{1}{\sigma}\frac{e^z}{(1+e^z)^2}\right\}^{\delta_i}\left\{\frac{1}{1+e^z}\right\}^{1-\delta_i}.$$

Соответствующая логарифмическая ФМП имеет вид

$$\ln L(\mu,\sigma) = -r\ln\sigma + \sum_{i=1}^{n}\delta_i\left[z_i - 2\ln(1+e^{z_i})\right] - \sum_{i=1}^{n}(1-\delta_i)\ln(1+e^{z_i}).$$

Производные, необходимые для итерационной схемы метода Ньютона–Рафсона, представленной в разделе «Общая методика для логарифмических моделей», запишутся как

$$\frac{\partial \ln f(z)}{\partial z} = 1 - \frac{2e^z}{1+e^z}, \quad \frac{\partial \ln S(z)}{\partial z} = -\frac{e^z}{1+e^z},$$

$$\frac{\partial^2 \ln f(z)}{\partial z^2} = -\frac{2e^z}{(1+e^z)^2}, \quad \frac{\partial^2 \ln S(z)}{\partial z^2} = -\frac{e^z}{(1+e^z)^2}.$$

17.3.4.3. Гамма– распределение

Плотность гамма–распределения имеет вид

$$f(t) = \frac{1}{\alpha\Gamma(\kappa)}\left(\frac{t}{\alpha}\right)^{\kappa-1}e^{-t/\alpha}, t \geq 0, \alpha > 0, \kappa > 0,$$

Соответствующая функция выживания
$S(t) = 1 - I(\kappa, t / \alpha)$,
где $I(.,.)$ – неполная гамма–функция.
Поэтому функция максимального правдоподобия (ФМП) запишется как

$$L(\alpha,\kappa) = \prod_{i=1}^{n} \left\{ \frac{1}{\alpha\Gamma(\kappa)} \left(\frac{t_i}{\alpha} \right)^{\kappa-1} e^{-t_i/\alpha} \right\}^{\delta_i} \left\{1 - I(\kappa, t_i/\alpha)\right\}^{1-\delta_i}.$$

Соответствующая логарифмическая ФМП имеет вид

$$\ln L(\alpha,\kappa) = -r\kappa\ln\alpha - r\ln\Gamma(\kappa) + (\kappa-1)\sum_{i=1}^{n}\delta_i \ln t_i - \frac{1}{\alpha}\sum_{i=1}^{n}\delta_i t_i + \sum_{i=1}^{h}(1-\delta_i)\ln\left[1 - I(\kappa, t_i/\alpha)\right].$$

Аналитическое представление производных логарифмической ФМП выполнить сложно, поэтому задача решается численно одним из вариантов метода спуска, не использующим производных.

17.3.4.4. Распределение Вейбулла

Плотность распределения Вейбулла с двумя параметрами имеет вид
$f(t) = \lambda\gamma t^{\gamma-1}\exp(-\lambda t^{\gamma})$, $t \ge 0$, $\lambda > 0$, $\gamma > 0$.
Соответствующая функция выживания
$S(t) = \exp(-\lambda t^{\gamma})$.
Поэтому функция максимального правдоподобия (ФМП) запишется как

$$L(\lambda,\gamma) = \prod_{i=1}^{n} \left\{\lambda\gamma t_i^{\gamma-1}\exp(-\lambda t_i^{\gamma})\right\}^{\delta_i} \left\{\exp(-\lambda t_i^{\gamma})\right\}^{1-\delta_i}.$$

После преобразований окончательно получаем

$$L(\lambda,\gamma) = \prod_{i=1}^{n} \left\{\lambda\gamma t_i^{\gamma-1}\right\}^{\delta_i} \exp(-\lambda t_i^{\gamma}).$$

Соответствующая логарифмическая ФМП имеет вид

$$\ln L(\lambda,\gamma) = \ln(\lambda\gamma)\sum_{i=1}^{n}\delta_i + (\gamma-1)\sum_{i=1}^{n}\delta_i \ln t_i - \lambda\sum_{i=1}^{h} t_i^{\gamma}.$$

Логарифмическая ФМП примет окончательный вид

$$\ln L(\lambda,\gamma) = r\ln(\lambda\gamma) + (\gamma-1)\sum_{i=1}^{n}\delta_i \ln t_i - \lambda\sum_{i=1}^{h} t_i^{\gamma}.$$

Для вычисления значений искомых параметров найдем частные производные логарифмической ФМП по искомым параметрам и приравняем их нулю. Сначала найдем производную по λ:

$$\frac{\partial \ln L(\lambda,\gamma)}{\partial\lambda} = \frac{r}{\lambda} - \sum_{i=1}^{n} t_i^{\gamma} = 0.$$

Отсюда уравнение для вычисления параметра λ получается как

$$\lambda = r\left[\sum_{i=1}^{n} t_i^{\gamma}\right]^{-1}.$$

Вычислив производную по параметру γ,

$$\frac{\partial \ln L(\lambda,\gamma)}{\partial\gamma} = \frac{r}{\gamma} + \sum_{i=1}^{n}\delta_i \ln t_i - \lambda\sum_{i=1}^{n} t_i^{\gamma}\ln t_i = 0,$$

с учетом выражения для параметра λ, получаем нелинейное уравнение для поиска параметра γ в виде:

$$\frac{r}{\gamma}+\sum_{i=1}^{n}\delta_i \ln t_i - r\left[\sum_{i=1}^{n} t_i^{\gamma}\right]^{-1}\sum_{i=1}^{n} t_i^{\gamma}\ln t_i = 0.$$

Решение уравнения может быть произведено одним из методов оптимизации – в простейшем случае методом деления отрезка пополам.

17.3.4.5. Экспоненциальное распределение

Плотность экспоненциального распределения имеет вид
$f(t) = \lambda e^{-\lambda t}$, $t \geq 0$, $\lambda > 0$.
Соответствующая функция выживания
$S(t) = e^{-\lambda t}$.
Поэтому функция максимального правдоподобия (ФМП) запишется как

$$L(\lambda) = \prod_{i=1}^{n}\left\{\lambda e^{-\lambda t_i}\right\}^{\delta_i}\left\{e^{-\lambda t_i}\right\}^{1-\delta_i}.$$

После преобразований окончательно получаем

$$L(\lambda) = \prod_{i=1}^{n}\lambda^{\delta_i} e^{-\lambda t_i}.$$

Соответствующая логарифмическая ФМП имеет вид

$$\ln L(\lambda) = r\ln\lambda - \lambda\sum_{i=1}^{n} t_i.$$

Для вычисления значения искомого параметра найдем производную логарифмической ФМП по данному параметру и приравняем ее нулю. Производная по параметру λ имеет вид:

$$\frac{\partial \ln L(\lambda)}{\partial\lambda} = \frac{r}{\lambda} - \sum_{i=1}^{n} t_i = 0.$$

Отсюда уравнение для вычисления параметра λ получается как

$$\lambda = r\left[\sum_{i=1}^{n} t_i\right]^{-1}.$$

17.3.4.6. Распределение Рэлея

Плотность распределения Рэлея имеет вид

$$f(t) = \frac{t}{\beta^2}\exp\left(-\frac{t^2}{2\beta^2}\right), t \geq 0, \beta > 0.$$

Соответствующая функция выживания

$$S(t) = \exp\left(-\frac{t^2}{2\beta^2}\right).$$

Поэтому функция максимального правдоподобия (ФМП) запишется как

$$L(\beta) = \prod_{i=1}^{n}\left\{\frac{t_i}{\beta^2}\exp\left(-\frac{t_i^2}{2\beta^2}\right)\right\}^{\delta_i}\left\{\exp\left(-\frac{t_i^2}{2\beta^2}\right)\right\}^{1-\delta_i}.$$

После преобразований окончательно получаем

$$L(\beta) = \prod_{i=1}^{n}\left(\frac{t_i}{\beta^2}\right)^{\delta_i}\exp\left(-\frac{t_i^2}{2\beta^2}\right).$$

Соответствующая логарифмическая ФМП имеет вид

$$\ln L(\beta) = \sum_{i=1}^{n} \delta_i \ln \frac{t_i}{\beta^2} - \frac{1}{2}\beta^{-2}\sum_{i=1}^{n} t_i^2.$$

Для вычисления значения искомого параметра найдем производную логарифмической ФМП по данному параметру и приравняем ее нулю. Производная по параметру β имеет вид:

$$\frac{\partial \ln L(\beta)}{\partial \beta} = -2\beta^{-1}\sum_{i=1}^{n} \delta_i + \beta^{-3}\sum_{i=1}^{n} t_i^2 = 0.$$

Уравнение для вычисления параметра β получается как

$$\beta = \left[\frac{1}{2r}\sum_{i=1}^{n} t_i^2\right]^{-2}.$$

17.3.4.7. Распределение Гомпертца

Плотность распределения Гомпертца имеет вид

$$f(t) = \beta e^{\alpha t}\exp\left[\frac{\beta}{\alpha}\left(1 - e^{\alpha t}\right)\right], t \geq 0, \beta > 0, \alpha \in \left]-\infty;0\right[\cup \left]0;\infty\right[.$$

Соответствующая функция выживания

$$S(t) = \exp\left[\frac{\beta}{\alpha}\left(1 - e^{\alpha t}\right)\right].$$

Поэтому функция максимального правдоподобия (ФМП) запишется как

$$L(\alpha,\beta) = \prod_{i=1}^{n}\left\{\beta e^{\alpha t_i}\exp\left[\frac{\beta}{\alpha}\left(1 - e^{\alpha t_i}\right)\right]\right\}^{\delta_i}\left\{\exp\left[\frac{\beta}{\alpha}\left(1 - e^{\alpha t_i}\right)\right]\right\}^{1-\delta_i}.$$

После преобразований окончательно получаем

$$L(\alpha,\beta) = \prod_{i=1}^{n}\left\{\beta e^{\alpha t_i}\right\}^{\delta_i}\exp\left[\frac{\beta}{\alpha}\left(1 - e^{\alpha t_i}\right)\right].$$

Соответствующая логарифмическая ФМП имеет вид

$$\ln L(\alpha,\beta) = r\ln\beta + \alpha\sum_{i=1}^{n}\delta_i t_i + \frac{\beta}{\alpha}\left(n - \sum_{i=1}^{n} e^{\alpha t_i}\right).$$

Для вычисления значений искомых параметров найдем частные производные логарифмической ФМП по искомым параметрам и приравняем их нулю. Сначала найдем производную по β:

$$\frac{\partial \ln L(\alpha,\beta)}{\partial \beta} = \frac{r}{\beta} + \frac{1}{\alpha}\left(n - \sum_{i=1}^{n} e^{\alpha t_i}\right) = 0.$$

Отсюда уравнение для вычисления параметра β получается как

$$\beta = \alpha r\left[\sum_{i=1}^{n} e^{\alpha t_i} - n\right]^{-1}.$$

Вычислив производную по параметру α,

$$\frac{\partial \ln L(\alpha,\beta)}{\partial \alpha} = \sum_{i=1}^{n}\delta_i t_i - \frac{\beta}{\alpha^2}\left[n + (1-\alpha^2)\sum_{i=1}^{n} e^{\alpha t_i}\right] = 0,$$

с учетом выражения для параметра β, получаем нелинейное уравнение для поиска параметра α в виде:

$$\alpha\sum_{i=1}^{n}\delta_i t_i - r\left[\sum_{i=1}^{n} e^{\alpha t_i} - n\right]^{-1}\left[n + (1-\alpha^2)\sum_{i=1}^{n} e^{\alpha t_i}\right] = 0.$$

Решение уравнения может быть произведено одним из методов оптимизации – в простейшем случае методом деления отрезка пополам.

17.3.4.8. Оценка качества подгонки модели

Программой выводятся результаты расчета объективными критериями согласия для каждого распределения. Адекватной является модель с *P*–значением, большим 0,05. В этом случае теоретическое и эмпирическое распределения значимо не различаются.
Качество статистической модели можно оценить (также сравнить между собой различные модели), используя информационный критерий Акаике (Akaike's information criterion, AIC)

$$AIC = -2\ln L(\hat{\theta}) + 2k + \frac{2k(k+1)}{(n-k-1)},$$

где $\ln L(\hat{\theta})$ – оценка логарифма функции максимального правдоподобия,
$\hat{\theta}$ – вектор оценок параметров статистической модели,
k – число параметров модели.
Последний член в уравнении для AIC призван скорректировать значение статистики критерия для малых выборок и некоторыми авторами не используется.
Оценка логарифма функции максимального правдоподобия в рассматриваемом случае имеет теоретический вид

$$\ln L(\hat{\theta}) = \sum_{i=1}^{n} \delta_i \ln\left[f(t_i, \hat{\theta}) \right] + \sum_{i=1}^{n} (1-\delta_i) \ln\left[S(t_i, \hat{\theta}) \right],$$

где t_i, $i = 1,2,...,n$ – эмпирический массив длительностей,
δ_i, $i = 1,2,...,n$ – соответствующий массив индикаторов цензурирования,
n – численность массива длительностей,
$f(.,.)$ – оценка функции плотности теоретического распределения,
$S(.,.)$ – оценка соответствующей функции выживания.
При расчете здесь нет необходимости в явном выписывании упомянутых функций, т. к. формулы для логарифмов функций максимального правдоподобия всех изучаемых теоретических распределений известны из предыдущих выкладок (см. выше).
При сравнении нескольких статистических моделей лучшей считается модель с наименьшим значением AIC.
В литературе представлены и другие информационные критерии, имеющие интерпретацию, аналогичную AIC.

См. монографию и статью Бернхэм (Burnham) с соавт., статьи Акаике (Akaike), Аль–Фозан (Al–Fawzan), Анжиллетта (Angilletta), Боздоган (Bozdogan), Бидюк с соавт., руководство Мотулски (Motulsky) с соавт.

17.3.5. Критерий Кокса

Критерий Кокса (логарифмический ранговый критерий, обобщенный критерий Сэвиджа) является обобщением критерия Сэвиджа (см. главу «Непараметрическая статистика») на цензурированные выборки и вычисляется по формуле

$$S = \sum_{i=1}^{N} \left[I(A_i) \sum_{j=1}^{i} \frac{1}{N+1-j} \right],$$

где $N = n_1 + n_2$ – численность объединенной выборки,
n_1 – численность первой сравниваемой выборки (обычно с наибольшей численностью),
n_2 – численность второй сравниваемой выборки (с наименьшей численностью),

$I(A_i)$, i =1,2,...,N – индикатор, что i–й член вариационного ряда, построенного по объединенной выборке, является нецензурированным (наработкой до отказа) и принадлежит первой из сравниваемых выборок; в этом случае значение индикатора равно 1, в противном случае – нулю.
Практически значимость может вычисляться посредством нормальной аппроксимации критического значения критерия. При этом модифицированная статистика

$$\frac{S - ES}{\sqrt{DS}},$$

где $ES = r_1$ – математическое ожидание,

$DS = \frac{r_1 r_2}{r_1 + r_2 - 1}\left(1 - \frac{1}{N}\sum_{j=1}^{N}\frac{1}{j}\right)$ – дисперсия,

r_1 и r_2 – количества нецензурированных элементов первой и второй сравниваемых выборок, соответственно,
распределена по стандартному нормальному закону.

См. монографию Скрипника с соавт.

17.3.6. Критерий Гехана

Критерий Гехана (обобщенный критерий Вилкоксона) является обобщением критерия Вилкоксона (см. главу «Непараметрическая статистика») на цензурированные выборки. В источниках могут быть даны различные эквивалентные формулы и схемы (часто оптимизированные для «ручного» счета) вычисления критерия. В программе статистика критерия вычисляется по наиболее простой формуле

$$W = \sum_{i=1}^{n_1}\sum_{j=1}^{n_2} u_{ij},$$

где n_1 – численность первой сравниваемой выборки,
n_2 – численность второй сравниваемой выборки.
Величины под знаками суммы вычисляются как

$$u_{ij} = \begin{cases} 1, x_i > y_j, \\ 1, x_i^* \geq y_j, \\ -1, x_i < y_j, \\ -1, x_i \leq y_j^*, \\ 0, \textit{если\_иначе}, \end{cases} \quad i = 1,2,...,n_1; j = 1,2,...,n_2,$$

где x_i, i = 1,2,...,n_1 – элементы первой сравниваемой выборки,
y_i, i = 1,2,...,n_2 – элементы второй сравниваемой выборки,
* – знак, означающий, что элемент выборки цензурирован.
Практически значимость может вычисляться посредством нормальной аппроксимации критического значения критерия. При этом модифицированная статистика

$$\frac{|W|}{\sqrt{DW}},$$

где DW – дисперсия, вычисляемая по формуле

$$DW = \frac{n_1 n_2 \sum_{i=1}^{N}\sum_{j=1}^{N} U_{ij}^2}{N(N-1)},$$

$$U_{ij} = \begin{cases} 1, X_i > X_j, \\ 1, X_i^* \geq X_j, \\ -1, X_i < X_j, \quad i, j = 1,2,...,N, \\ -1, X_i \leq Y_j^*, \\ 0, если\_иначе, \end{cases}$$

X_i, i = 1,2,...,N – элементы объединенной выборки,
$N = n_1 + n_2$ – численность объединенной выборки,
распределена по стандартному нормальному закону.

См. монографию Ли (Lee) с соавт.

17.3.7. Модель пропорциональных рисков Кокса

Полупараметрическая (semi–parametric) модель пропорциональных рисков (пропорциональных интенсивностей) Кокса может быть записана в виде

$$\frac{h(t)}{h_0(t)} = \exp(\beta^T x),$$

где $h(t)$ – функция риска,
$h_0(t)$ – базовая функция риска (функция риска при нулевых ковариатах),
β – вектор (длиной m) коэффициентов модели (коэффициентов при ковариатах),
x – вектор (длиной m) ковариат (предикторов независимых переменных) модели.
С помощью модели Кокса обычно исследуется отношение $h(t)$ / $h_0(t)$, поэтому базовая функция риска в модели Кокса не оценивается.
Для обучения модели пропорциональных рисков, помимо длительностей (n–мерный вектор t) и индикаторов цензурирования (n–мерный вектор δ), должна быть представлена n x m матрица X ковариат, представляющая собой n m–мерных векторов x_i, i = 1,2,...,n, ковариат для каждого индивидуума обучающей выборки, где n – численность выборки (количество пациентов). Целью обучения является определение оптимальных значений компонент m–мерного вектора коэффициентов модели β при ковариатах.
Предложенный Коксом метод частичного правдоподобия (partial likelihood) позволяет оценить значения компонент вектора коэффициентов модели, доставляющие максимум так называемой частичной функции максимального правдоподобия (ФМП). Частичная ФМП имеет вид

$$PL(\beta) = \prod_{i=1}^{n} \left[\exp(\beta^T x_i) \Big/ \sum_{j \in R_i} \exp(\beta^T x_j) \right]^{\delta_i},$$

где $R_i = \{j: t_j \geq t_i\}$ – множество таких длительностей j, для которых $t_j \geq t_i$.
Соответствующая логарифмическая ФМП имеет вид

$$\ln PL(\beta) = \sum_{i=1}^{n} \delta_i \left[\beta^T x_i - \ln\left(\sum_{j=1}^{n} Y_j(t_i) \exp(\beta^T x_j) \right) \right],$$

где $Y_j(t_i)$, i = 1,2,...,n; j = 1,2,...,n – индикатор, равный 1 для $t_j \geq t_i$ (выборка упорядочена) или равный 0 в иных случаях.

Для дальнейших расчетов необходимы аналитические представления вектора первых производных (градиент) и матрицы вторых производных (матрица Гессе) логарифмической ФМП по искомым коэффициентам модели.
Градиент, т. е. *m*–мерный вектор первых производных логарифмической ФМП, вычисляется по формуле

$$G(\beta) = \frac{\partial \ln PL(\beta)}{\partial \beta} = \sum_{i=1}^{n} \delta_i \left[x_i - \bar{x}(t_i, \beta) \right],$$

где $\bar{x}(.,.)$ – вектор взвешенных средних значений для длительности *i*, *i* =1,2,...,*n*, вычисляется как

$$\bar{x}(t_i, \beta) = \frac{\sum_{j=1}^{n} Y_j(t_i) \exp(\beta^T x_j) x_j}{\sum_{j=1}^{n} Y_j(t_i) \exp(\beta^T x_j)}, i = 1,2,...,n.$$

Фактически совокупность векторов взвешенных средних представляет собой *n* x *m* матрицу $\bar{X}$, по структуре аналогичную заданной матрице ковариат.
Матрица Гессе, т. е. *m* x *m* матрица вторых производных логарифмической ФМП, вычисляется по формуле

$$H(\beta) = \frac{\partial^2 \ln PL(\beta)}{\partial \beta \partial \beta^T} = -\sum_{i=1}^{n} \delta_i \left\{ \frac{\sum_{j=1}^{n} Y_j(t_i) \exp(\beta^T x_j) \left[x_j - \bar{x}(t_i, \beta) \right] \left[x_j - \bar{x}(t_i, \beta) \right]^T}{\sum_{j=1}^{n} Y_j(t_i) \exp(\beta^T x_j)} \right\}.$$

Модель может быть решена с помощью метода Ньютона–Рафсона. Итерационная схема метода имеет вид
$\beta^{l+1} = \beta^l - [H(\beta)]^{-1} G(\beta)$, $l = 0,1,2,...,$
где *l*, *l* = 0,1,2,... – номер итерации.
Начальные приближения для итерационного процесса можно взять нулевыми.
Дисперсии $D\hat{\beta}_j, j = 1,2,...,m,$ оценок коэффициентов при ковариатах, вычисленных, как показано выше, представляют собой соответствующие диагональные элементы матрицы Гессе (на последней итерации), взятые с обратным знаком. Доверительные интервалы данных оцениваемых коэффициентов считаются стандартно как

$$\beta_j = \left(\hat{\beta}_j - \Psi((1+\beta)/2) \sqrt{D\hat{\beta}_j}; \hat{\beta}_j + \Psi((1+\beta)/2) \sqrt{D\hat{\beta}_j} \right), j = 1,2,...,m,$$

где $\hat{\beta}_j, j = 1,2,...,m,$ – оценки коэффициентов при ковариатах,
$\Psi(.)$ – обратная функция стандартного нормального распределения,
β – доверительный уровень, выраженный в долях.
Построенную модель пропорциональных рисков применяют также для вычисления регрессии Кокса, которая записывается в виде
$h(t) = h_0(t)\exp(\beta^T x)$.
При необходимости явного построения функции риска с помощью данной регрессии для конкретного индивидуума может быть использована любая подходящая параметрическая модель из числа представленных в разделе «Подбор распределения». Получающаяся в результате регрессионная модель часто называется не регрессией Кокса, а по имени регрессии базовой функции риска (например, Вейбулла или Гомпертца).

Теоретическое обоснование см. в монографиях Кокса с соавт., Кляйн (Klein) с соавт., Дюпон (Dupont), Фортхофер (Forthofer) с соавт., Кемпбелл (Campbell). Техника вычислений представлена О'Квигли (O'Quigley), Лелесс (Lawless). См. также статьи Фан (Fan) с соавт., Гош (Ghosh D.), работы Биндер (Binder), Кларксон (Clarkson) с соавт., доклад Ю (Yu) с соавт., соответствующую статью энциклопедии под ред. Армитейдж (Armitage) с соавт. Примеры отбора ковариат и валидации модели см. в статьях Ли (Lee M.S.) с соавт., Ле (Le), Ван Хоувелинген (Van Houwelingen), Ван Хоувелинген с соавт., Нейджелкерк (Nagelkerke) с соавт.

В модели пропорциональных рисков индикаторы цензурирования могут принимать только значения 0 (пациент цензурирован, т. е. выбыл из исследования, и его состояние неизвестно, или умер по причине, не связанной с исследуемой патологией) или 1 (пациент умер по причине, связанной с исследуемой патологией). Однако возможен такой случай, что причин смерти, связанной с исследуемой патологией, может быть выявлено более 1. Для такого случая разработана модель конкурирующих рисков (competing risks). Метод не представлен в данном программном обеспечении. Модель рассмотрена в монографиях Краудера (Crowder), Хедекера (Hedeker) с соавт., Диггля (Diggle) с соавт., Лачина (Lachin), Кальбфляйша (Kalbfleisch) с соавт., Пинтили (Pintilie), Штейерберга (Steyerberg), Фитцмауриса (Fitzmaurice) с соавт., Твиска (Twisk), статьях Фюрстовой (Fürstová) с соавт., Брауна (Brown), Чианг (Chiang), Фиокко (Fiocco) с соавт.

Список использованной и рекомендуемой литературы

1. Aalen O.O. Nonparametric inference for a family of counting processes // The Annals of Statistics, 1978, vol. 6, no. 4, pp. 701–726.
2. Aalen O.O., Borgan O., Gjessing H.K. Survival and event history analysis: A process point of view. – New York, NY: Springer, 2008.
3. Ahn H. Estimating the mean and variance of censored phosphorus concentrations in Florida rainfall // Journal of the American Water Resources Association, June 1998, vol. 34, no.3, pp. 583–593.
4. Aitchison J., Brown J.A.C. The lognormal distribution. – Cambridge, UK: Cambridge University Press, 1963.
5. Akaike H. A Bayesian analysis of the minimum AIC procedure // Annals of the Institute of Statistical Mathematics, 1978, vol. 30, part A, no. 1, pp. 9–14.
6. Al–Fawzan M.A. Algorithms for estimating the parameters of the Weibull distribution // Statistics on the Internet (InterStat), October 2000, no. 1.
7. Altman D.G., Bland J.M. Time to event (survival) data // British Medical Journal, 1998, vol. 317, pp. 468–469.
8. Andersen P.K., Gill R.D. Cox's regression model for counting processes: A large sample study // The Annals of Statistics, 1982, vol. 10, no. 4, pp. 1100–1120.
9. Angilletta M.J.Jr., Oufiero C.E., Leache A.D. Direct and indirect effects of environmental temperature on the evolution of reproductive strategies: An information–theoretic approach // The American Naturalist, October 2006, vol. 168, no. 4, pp. 123–135.
10. Annest A. Iterative Bayesian model averaging: A method for the application of survival analysis to high–dimensional microarray data / A. Annest, R.E. Bumgarner, A.E. Raftery et al. // BMC Bioinformatics, 26 February 2009, vol. 10, p. 72.
11. Armitage P. Encyclopedia of biostatistics / Ed. by P. Armitage, T. Colton. – New York, NY: John Wiley & Sons, 2005.
12. Auget J.–L. Advances in statistical methods for the health sciences applications to cancer and AIDS studies, genome sequence analysis, and survival analysis / Ed. by J.–L. Auget, N. Balakrishnan, M. Mesbah et al. – Boston, MA: Birkhauser, 2007.

13. Balakrishnan N. Handbook of statistics. Vol. 20. Advances in reliability / Ed. by N. Balakrishnan and C.R. Rao. – New York, NY: Elsevier, 2001.
14. Balakrishnan N. Handbook of statistics. Vol. 23. Advances in survival analysis / Ed. by N. Balakrishnan, C.R. Rao. – New York, NY: Elsevier, 2003.
15. Barber S., Jennison C. A review of inferential methods for the Kaplan–Meier estimator // Research report 98–02, 1998, Statistics group, University of Bath, U.K.
16. Barber S., Jennison C. Bootstrapping the Kaplan–Meier estimator // Proceedings in Computational Statistics Compstat 1998 / Ed. by R. Payne, P.J. Green, 1998, vol. 2, pp. 139–140.
17. Barber S., Jennison C. Symmetric tests and confidence intervals for survival probabilities and quantiles of censored survival data // Biometrics, 1999, vol. 55, pp. 430–436.
18. Barker C. The mean, median, and confidence Intervals of the Kaplan–Meier survival estimate – Computations and applications // The American Statistician, 1 February 2009, vol. 63, no. 1, pp. 78–80.
19. Barros A.J.D., Hirakata V.N. Alternatives for logistic regression in cross–sectional studies: an empirical comparison of models that directly estimate the prevalence ratio // BMC Medical Research Methodology, 2003, vol. 3, no. 21.
20. Bertholon H., Bousquet N., Celeux G. An alternative competing risk model to the Weibull distribution in lifetime data analysis // Rapport de recherche no. 5265, Juillet 2004. Institut National de Recherche en Informatique et Automatique. Rocquencourt, France.
21. Bewick V., Cheek L., Ball J. Statistics review 12: Survival analysis // Critical Care, 2004, vol. 8, pp. 389–394.
22. Bhattacharjee A. A simple test for the absence of covariate dependence in hazard regression models // Munich Personal RePEc Archive, MPRA Paper, November 2007, no. 3937.
23. Billingham L.J., Abrams K.R., Jones D.R. Methods for the analysis of quality–of–life and survival data in health technology assessment // Health Technology Assessment, 1999, vol. 3, no. 10.
24. Binder D.A. Fitting Cox's proportional hazards models from survey data // Proceedings of the Survey Research Methods Section, American Statistical Association, 1990, pp. 342–347.
25. Biswas A. Statistical advances in the biomedical sciences: Clinical trials, epidemiology, survival analysis, and bioinformatics / Ed. by A. Biswas, S. Datta, J.P. Fine et al. – Hoboken, NJ: John Wiley & Sons, 2008.
26. Bland J.M., Altman D.G. Survival probabilities (the Kaplan–Meier method) // British Medical Journal, 5 December 1998, vol. 317, pp. 1572–1580.
27. Bohoris G.A. Calculating the comparative two–sample tests for censored reliability data // International Journal of Quality & Reliability Management, 1997, vol. 14, issue 1, pp. 82–88.
28. Borgan O., Goldstein L., Langholz B. Methods for the analysis of sampled cohort data in the Cox proportional hazards model // The Annals of Statistics, 1995, vol. 23, no. 5, pp. 1749–1778.
29. Bozdogan H. Akaike's information criterion and recent developments in information complexity // Journal of Mathematical Psychology, 2000, vol. 44, pp. 62–91.
30. Breslow N., Crowley J. A large sample study of the life table and product limit estimates under random censorship // The Annals of Statistics, 1974, vol. 2, no. 3, pp. 437–453.
31. Brookmeyer R., Crowley J. A confidence interval for the median survival time // Biometrics, March 1982, vol. 38, no. 1, pp. 29–41.
32. Brown C.C. The statistical comparison of relative survival rates // Biometrics, December 1983, vol. 39, pp. 941–948.
33. Burnham K.P., Anderson D.R. Model selection and multimodel inference: A practical information–theoretic approach. – New York, NY: Springer, 1998.

34. Burnham K.P., Anderson D.R. Multimodel inference: Understanding AIC and BIC in model selection // Sociological Methods & Research, November 2004, vol. 33, no. 2, pp. 261–304.
35. Bury K. Statistical distributions in engineering. – Cambridge, UK: Cambridge University Press, 1999.
36. Campbell M.J. Statistics at square two. Understanding modern statistical applications in medicine. – Malden, MA: Blackwell Publishing, 2006.
37. Chiang C.L. Competing risks in mortality analysis // Annual Review of Public Health, May 1991, vol. 12, pp. 281–307.
38. Chiang C.L. On the probability of death from specific causes in the presence of competing risks // Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, vol. 4 (University of California Press, 1961), pp. 169–180.
39. Choi T., Cole S.R. A family of ordered logistic regression models fit by data expansion // International Journal of Epidemiology, 2004, vol. 33, no. 6, p. 1413.
40. Clarkson D.B., Jennrich R.I. Computing extended maximum likelihood estimates for Cox proportional–hazards models // Game theory, optimal stopping, probability and statistics: Papers in honor of Thomas S. Ferguson / Ed. by F.T. Bruss, L. Le Cam. – Beachwood, OH: Institute of Mathematical Statistics, 2000, pp. 205–217.
41. Cohen A.C. Truncated and censored samples: Theory and applications. – New York, NY: Marcel Dekker, 1991.
42. Collett D. Modelling survival data in medical research. – Boca Raton, FL: Chapman & Hall / CRC, 1993.
43. Cox D.R. Regression models and life–tables // Journal of the Royal Statistical Society, Series B (Methodological), 1972, vol. 34, no. 2, pp. 187–220.
44. Crowder M.J. Classical competing risks. – Boca Raton, FL: Chapman & Hall / CRC, 2001.
45. Datta S. Estimating the mean life time using right censored data // Statistical Methodology, 2005, vol. 2, pp. 65–69.
46. Desu M.M., Raghavarao D. Nonparametric statistical methods for complete and censored data. – Boca Raton, FL: Chapman & Hall / CRC, 2004.
47. Diggle P.J. Analysis of longitudinal data / P.J. Diggle, P. Heagerty, K.–Y. Liang et al. – Oxford, NY: Oxford University Press, 2004.
48. Dupont W.D. Statistical modeling for biomedical researchers. A simple introduction to the analysis of complex data. – New York, NY: Cambridge University Press, 2002.
49. Elandt–Johnson R., Johnson N. Survival models and data analysis. – New York, NY: John Wiley & Sons, 1999.
50. Esteve J., Benhamou E., Raymond L. Statistical methods in cancer research. Vol. IV. Descriptive epidemiology. – Lyon, France: International Agency for Research on Cancer, 1994.
51. Evans M., Hastings N., Peacock B. Statistical distributions. – New York, NY: John Wiley & Sons, 2000.
52. Fan J., Gijbels I., King M. Local likelihood and local partial likelihood in hazard regression // The Annals of Statistics, 1997, vol. 25, no. 4, pp. 1661–1690.
53. Fan J., Li R. Variable selection for Cox's proportional hazards model and frailty model // The Annals of Statistics, 2002, vol. 30, no. 1, pp. 74–99.
54. Fine J., Gray R. A proportional hazards model for the subdistribution of a competing risk // Journal of the American Statistical Association, 1999, vol. 94, pp. 496–509.
55. Fiocco M., Putter H., van Houwelingen J.C. Reduced rank proportional hazards model for competing risks // Biostatistics, 2005, vol. 6, no. 3, pp. 465–478.
56. Fitzmaurice G. Longitudinal data analysis / Ed. by G. Fitzmaurice, M. Davidian, G. Verbeke et al. – Boca Raton, FL: Chapman & Hall / CRC, 2009.

57. Fletcher R.H., Fletcher S.W., Wagner E.H. Clinical epidemiology: the essentials. – Baltimore, Maryland: Williams & Wilkins, 1996.
58. Forthofer R.N., Lee E.S., Hernandez M. Biostatistics: A guide to design, analysis, and discovery. – New York, NY: Elsevier, 2007.
59. Freedman A.N. Cancer risk prediction models: A workshop on development, evaluation, and application / A.N. Freedman, D. Seminara, M.H. Gail et al. // JNCI Journal of the National Cancer Institute, 2005, vol. 97, no. 10, pp. 715–723.
60. Freedman D.A. Survival analysis: A primer // The American Statistician, May 2008, vol. 62, no. 2, pp. 110–119.
61. Fürstová J., Valenta Z. Statistical analysis of competing risks: Overall survival in a group of chronic myeloid leukemia patients // The European Journal for Biomedical Informatics, 2011, vol. 7, issue 1, pp. en1-en10.
62. Ghosh A. A FORTRAN program for fitting Weibull distribution and generating samples // Computers & Geosciences, 1999, vol. 25, no. 7, pp. 729–738.
63. Ghosh D. Proportional hazards regression for cancer studies // Biometrics, March 2008, vol. 64, issue 1, pp. 141–148.
64. Greenwood P.E., Wefelmeyer W. Cox's factoring of regression model likelihoods for continuous–time processes // Bernoulli, 1998, vol. 4, no. 1, pp. 65–80.
65. Gu M., Zheng Z. On the Bartlett adjustment for the partial likelihood ratio test in the Cox regression model // Statistica Sinica, 1993, no. 3, pp. 543–555.
66. Hedeker D., Gibbons R.D. Longitudinal data analysis. – Hoboken, NJ: John Wiley & Sons, 2006.
67. Heritier S. Robust methods in biostatistics / S. Heritier, E. Cantoni, S. Copt et al. – Chichester, West Sussex: John Wiley & Sons, 2009.
68. Hewett P., Ganser G.H. A comparison of several methods for analyzing censored data // Annals of Occupational Hygiene, 2007, vol. 51, no. 7, pp. 611–632.
69. Hosmer D.W., Jr., Lemeshow S. Applied survival analysis: Regression modeling of time to event data. – New York, NY: John Wiley & Sons, 1999.
70. Ismail A.A. On the optimal design of step–stress partially accelerated life tests for the Gompertz distribution with type–I censoring // Statistics on the Internet (InterStat), June 2006, no. 1.
71. Jimenez F., Jodra P. A Note on the moments and computer generation of the shifted Gompertz distribution // Communications in Statistics: Theory and Methods, January 2009, vol. 38, no. 1, pp. 75–89.
72. Johnson N.L., Kotz S., Balakrishnan N. Continuous univariate distribution. Vol. 1. – New York, NY: John Wiley & Sons, 1994.
73. Jones M.P., Crowley J. Asymptotic properties of a general class of nonparametric tests for survival analysis // The Annals of Statistics, 1990, vol. 18, no. 3, pp. 1203–1220.
74. Joyce K.A., Ghosn F., Bayer R. Competing outcomes: A competing risk model of military and political outcome of interstate wars // The annual meeting of the International Studies Association, Hilton Hawaiian Village, Honolulu, Hawaii, March 05, 2005.
75. Kalbfleisch J.D., Prentice R.L. The statistical analysis of failure time data. – Hoboken, NJ: John Wiley & Sons, 2002.
76. Kamakura T. Computational methods in survival analysis // Research Paper, Humboldt–Universitat Berlin, Center for Applied Statistics and Economics, 2007.
77. Kaplan E.L., Meier P. Nonparametric estimation from incomplete observations // Journal of the American Statistical Association, June 1958, vol. 53, no. 282, pp. 457–481.
78. Kleiber C., Kotz S. Statistical size distributions in economics and actuarial sciences. – New York, NY: John Wiley & Sons, 2003.

79. Klein J.P., Moeschberger M.L. Survival analysis: Techniques for censored and truncated data. – New York, NY: Springer, 2003.
80. Klein, J.P. Small sample moments of some estimators of the variance of the Kaplan– Meier and Nelson–Aalen estimators // Scandinavian Journal of Statistics, 1991, vol. 18, no. 4, pp. 333–340.
81. Kleinbaum D.G., Klein M. Survival analysis: A self–learning text. – New York, NY: Springer, 2005.
82. Krishnamoorthy K. Handbook of statistical distributions with applications. – Boca Raton, FL: Chapman & Hall / CRC, 2006.
83. Kubler H. On the fitting of the three–parameter distributions lognormal, gamma, and Weibull // Statistical Papers, June 1979, vol. 20, no. 2, pp. 68–125.
84. Kunimura D. The Gompertz distribution–estimation of parameters // Actuarial Research Clearing House, 1998, vol. 2, pp. 65–76.
85. Lachin J.M. Biostatistical methods. The assessment of relative risks – Hoboken, NJ: John Wiley & Sons, 2000.
86. Lai D, Hardy RJ. Potential gains in life expectancy or years of potential life lost: impact of competing risks of death // International Journal of Epidemiology, 1999, vol. 28, pp. 894–898.
87. Laird N.M., Olivier D. Covariance analysis of censored survival data using log–linear analysis techniques // Journal of the American Statistical Association, 1981, vol. 76, pp. 231–240.
88. Langberg N., Proschan F., Quinzi A.J. Estimating dependent life lengths, with applications to the theory of competing risks // The Annals of Statistics, 1981, vol. 9, no. 1, pp. 157–167.
89. Langova K. Survival analysis for clinical studies // Biomedical papers of the Medical Faculty of the University Palacky, Olomouc, Czech Republic, 2008, vol. 152, no. 2, pp. 303–307.
90. Lau B., Cole S.R., Gange S.J. Competing risk regression models for epidemiologic data // American Journal of Epidemiology, 2009, vol. 170, no. 2, pp. 244–256.
91. Lawless J.F. Statistical models and methods for lifetime data. – Hoboken, NJ: John Wiley & Sons, 2003.
92. Le C.T., Zelterman D. Goodness of fit tests for proportional hazards regression models // Biometrical Journal, 1992, vol. 34, no. 5, pp. 557–566.
93. Lee E.T., Wang J.W. Statistical methods for survival data analysis. – Hoboken, NJ: John Wiley & Sons, 2003.
94. Lee M.S., Proportionality assumption test of Cox's proportional hazard model in survival analysis / M.S. Lee, K.Y. Yoo, D.Y. Noh et al. // Journal of Korean Cancer Association, December 1991, vol. 23, no. 4, pp. 852–859.
95. Lepeule J. Survival analysis to estimate association between short–term mortality and air pollution / J. Lepeule, V. Rondeau, L. Filleul et al. // Environmental Health Perspectives, February 2006, vol. 114, no. 2, pp. 242–247.
96. Li G. Nonparametric likelihood ratio confidence bands for quantile functions from incomplete survival data / G. Li, M. Hollander, I.W. McKeague et al. // The Annals of Statistics, 1996, vol. 24, no. 2, pp. 628–640.
97. Limpert E., Stahel W., Abbt M. Log–normal distributions across the sciences: Keys and clues // BioScience, 2001, vol. 51, no. 5, pp. 341–352.
98. Lin D.Y., Wei L.J. Goodness–of–fit tests for the general Cox regression model // Statistica Sinica, 1991, vol. 1, pp. 1–17.
99. Lu J.Z. Fitting Weibull and lognormal distributions to medium–density fiberboard fiber and wood particle length / J.Z. Lu, C.J. Monlezun, Q. Wu et al. // Wood and Fiber Science, 2007, vol. 39, no. 1, pp. 82–94.
100. Lunn M., McNeil D. Applying Cox regression to competing risks // Biometrics, June

1995, vol. 51, no. 2, pp. 524–532.
101. Ma Z., Krings A. Competing risks analysis of reliability, survivability, and prognostics and health management (PHM) // Proceedings IEEE Aerospace Conference, March 1–8, Big Sky, MT, 2008.
102. Machin D., Cheung Y.B., Parmar M.K.B. Survival analysis: A practical approach. – Hoboken, NJ: John Wiley & Sons, 2006.
103. Martinussen T., Scheike T.H. Dynamic regression models for survival data. – New York, NY: Springer, 2006.
104. Matthews D.E., Farewell V.T. Using and understanding medical statistics. – Basel, Switzerland: Karger, 2007.
105. Motulsky H., Christopoulos A. Fitting models to biological data using linear and nonlinear regression. A practical guide to curve fitting. – San Diego, CA: GraphPad Software Inc., 2003.
106. Nagelkerke N.J.D., Oosting J., Hart A.A.M. A simple test for goodness of fit of Cox's proportional hazards model // Biometrics, June 1984, vol. 40, no. 2, pp. 483–486.
107. Nelson W. Applied life data analysis. – New York, NY: John Wiley & Sons, 1982.
108. O'Quigley J. Proportional hazards regression. – New York, NY: Springer, 2008.
109. Pintilie M. Competing risks: A practical perspective. – Chichester, UK: John Wiley & Sons, 2006.
110. Pons O. Estimation in a Cox regression model with a change–point according to a threshold in a covariate // The Annals of Statistics, 2003, vol. 31, no. 2, pp. 442–463.
111. Porta N. Competing risks methods / N. Porta, G. Gomez, M.L. Calle et al. // Technical Report DR 2007/14. Department of Statistics and Operations Research, Universitat Politecnica de Catalunya, 2007.
112. Porta N., Calle M.L., Gomez G. The role of survival functions in competing risks. Technical Report DR 2008/06. Department of Statistics and Operations Research, Universitat Politecnica de Catalunya, 2008.
113. Prentice R.L. Relative risk regression analysis of epidemiologic data // Environmental Health Perspectives, 1985, vol. 63, pp. 225–234.
114. Pyke D.A. Statistical analysis of survival and removal rate experiments // Ecology, 1986, vol. 67, no. 1, pp. 240–245.
115. Qiguang W., Jianhua L. Sampling inspection of reliability in (log)normal case with type I censoring // Acta Mathematica Scientia, 2006, vol. 26B, no.2, pp. 331–343.
116. Resche–Rigon M., Azoulay E., Chevret S. Evaluating mortality in intensive care units: Contribution of competing risks analyses // Critical Care, 2006, vol. 10, no. 1, p. R5.
117. Ricklefsa R.E., Scheuerleina A. Biological implications of the Weibull and Gompertz models of aging // The Journals of Gerontology Series A: Biological Sciences and Medical Sciences, 2002, vol. 57A, no. 2, pp. B69–B76.
118. Royston P., Parmar M.K.B., Altman D.G. Visualizing length of survival in time–to–event studies: A complement to Kaplan Meier plots // JNCI Journal of the National Cancer Institute, 2008, vol. 100, no. 2, pp. 92–97.
119. Satagopan J.M. A note on competing risks in survival data analysis / J.M. Satagopan, L. Ben–Porat, M. Berwick et al. // British Journal of Cancer, 2004, vol. 91, pp. 1229–1235.
120. Selvin S. Survival analysis for epidemiologic and medical research: A practical guide. – Cambridge, UK: Cambridge University Press, 2008.
121. Serfling R. Efficient and robust fitting of lognormal distributions // North American Actuarial Journal, 2002, vol. 6, no.4, pp. 95–116.
122. Steyerberg E.W. Clinical prediction models: A practical approach to development, validation, and updating. – New York, NY: Springer, 2009.

123. Stute W. The jackknife estimate of variance of a Kaplan–Meier integral // The Annals of Statistics, 1996, vol. 24, no. 6, pp. 2679–2704.
124. Sultan K.S., Mahmoud M.R., Saleh H.M. Estimation of parameters of the Weibull distribution based on progressively censored data // International Mathematical Forum, 2007, no. 41, pp. 2031–2043.
125. Tian L., Zucker D., Wei L.J. On the Cox model with time–varying regression coefficients // Journal of the American Statistical Association, March 2005, vol. 100, no. 469, pp. 172–183.
126. Tsodikov A.D., Ibrahim J.G., Yakovlev A.Y. Estimating cure rates from survival data // Journal of the American Statistical Association, December 2003, vol. 98, no. 464, pp. 1063–1078.
127. Twisk J.W.R. Applied longitudinal data analysis for epidemiology: A practical guide. – Cambridge, UK: Cambridge University Press, 2003.
128. Van Houwelingen H.C. Cross–validated Cox regression on microarray gene expression data / H.C. Van Houwelingen, T. Bruinsma, A.A.M. Hart et al. // Statistics in medicine, 2006, vol. 25, no. 18, pp. 3201–3216.
129. Van Houwelingen H.C. Validation, calibration, revision and combination of prognostic survival models // Statistics in Medicine, 2000, vol. 19, no. 24, pp. 3401–3415.
130. Vittinghoff E. Regression methods in biostatistics: Linear, logistic, survival, and repeated measures models / E. Vittinghoff, S.C. Shiboski, D.V. Glidden et al. – New York, NY: Springer, 1995.
131. Wetterstrand W.H. Parametric models for life insurance mortality data: Gompertz's law over time // Transactions of Society of Actuaries, 1981, vo. 33, no.8, pp. 159–179.
132. Wolbers M. Prognostic models with competing risks: Methods and application to coronary risk prediction / M. Wolbers, M.T. Koller, J.C.M. Witteman et al. // Epidemiology, July 2009, vol. 20, issue 4, pp. 555–561.
133. Wu T. A method for analyzing censored survival phenotype with gene expression data / T. Wu, W. Sun, S. Yuan et al. // BMC Bioinformatics, 2008, vol. 9, pp. 417.
134. Xie J., Liu C. Adjusted Kaplan–Meier estimator and log–rank test with inverse probability of treatment weighting for survival data // Statistics in Medicine, 2005, vol. 24, issue 20, pp. 3089–3110.
135. Yousef M.H. Estimation of parameters and truncation point for the truncated Gompertz distribution // Journal of King Saud University (Administrative Sciences), 1993, vol. 5, no. 2, pp. 35–48.
136. Yu S. Privacy–preserving Cox regression for survival analysis / S. Yu, G. Fung, R. Rosales et al. // Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining table of contents, Las Vegas, Nevada, USA, 2008, pp. 1034–1042.
137. Zucker D.M. A pseudo–partial likelihood method for semiparametric survival regression with covariate errors // Journal of the American Statistical Association, December 2005, vol. 100, no. 472, pp. 1264–1277.
138. Zucker D.M., Karr A.F. Nonparametric survival analysis with time–dependent covariate effects: A penalized partial likelihood approach // The Annals of Statistics, 1990, vol. 18, no. 1, pp. 329–353.
139. Амелина Е.Л., Черняк А.В., Черняев А.Л. Муковисцидоз: определение продолжительности жизни // Пульмонология, 2001, том 11, №3, с. 61–64.
140. Арженовский С. Социально–экономические детерминанты курения в России // Квантиль, 2006, № 1, с. 81–100.
141. Башарин Г.П., Плаксина Н.Н. Применение теории конкурирующих рисков при

анализе смертности // Страховое дело, 1995, № 1.
142. Бидюк П.И., Зворыгина Т.Ф. Структурный анализ методик построения регрессионных моделей по временным рядам наблюдений // Управляющие системы и машины, 2003, № 2, с. 93–99.
143. Власов В.В. Эпидемиология. Учебное пособие для вузов. – М.: Издательский дом «ГЭОТАР–МЕД», 2004.
144. Воронков Л.Г. Предикторы 5–летней выживаемости больных и индивидуальное прогнозирование течения клинически манифестированной хронической сердечной недостаточности / Л.Г. Воронков, Г.В. Яновский, Е.В. Устименко и др. // Украинский медицинский журнал, 2003, том. 38, № 6, с. 106–109.
145. Груздев А.В. Применение анализа выживаемости Карлана-Мейера для оценки времени наступления дефолта // Банковские технологии, 2012, № 1, с. 38-38.
146. Кобзарь А.И. Прикладная математическая статистика. Для инженеров и научных работников. – М.: ФИЗМАТЛИТ, 2006.
147. Кокс Д.Р., Оукс Д. Анализ данных типа времени жизни. – М.: Финансы и статистика, 1988.
148. Мазовец О.Л. Прогностическое значение белка, связывающего жирные кислоты, у госпитализированных из–за ухудшения сердечной недостаточности больных. Результаты 6–12–месячного наблюдения / О.Л. Мазовец, И.Р. Трифонов, А.Г. Катруха и др. // Кардиология, 2008, № 1, с. 24–29.
149. Макфадден Д. Полупараметрический анализ // Квантиль, 2008, № 5, с. 29–40.
150. Один И.М. Определение параметров распределения Вейбулла методом наименьших квадратов // Надежность и контроль качества, 1975, № 7, с. 45–48.
151. Орлов А.И. Непараметрическое точечное и интервальное оценивание характеристик распределения // Заводская лаборатория. Диагностика материалов, 2004, т. 70, № 5, с. 65–70.
152. Плаксина Н.Н. Математические модели дожития и заболеваемости на основе теории конкурирующих рисков. Канд. дисс. – М.: РУДН, 1999.
153. Родригес Г. Модели выживаемости // Квантиль, 2008, № 5, с. 1–27.
154. Скрипник В.М. Анализ надежности технических систем по цензурированным выборкам / В.М. Скрипник, А.Е. Назин., Ю.Г. Приходько и др. – М.: Радио и связь, 1988.
155. Флетчер Р., Флетчер С., Вагнер Э. Клиническая эпидемиология: Основы доказательной медицины. – М.: Медиа Сфера, 2004.
156. Хан Г., Шапиро С. Статистические модели в инженерных задачах. – М.: Мир, 1969.
157. Холлендер М., Вулф Д. Непараметрические методы статистики. – М.: Финансы и статистика, 1983.
158. Цыплаков А.А. Мини–словарь англоязычных эконометрических терминов, часть 2 // Квантиль, 2008, № 5, с. 41–48.
159. Цыплаков А.А. Некоторые эконометрические методы. Метод максимального правдоподобия в эконометрии. Методическое пособие. – Новосибирск: НГУ, 1997.

Глава 18. Анализ временных рядов и прогнозирование

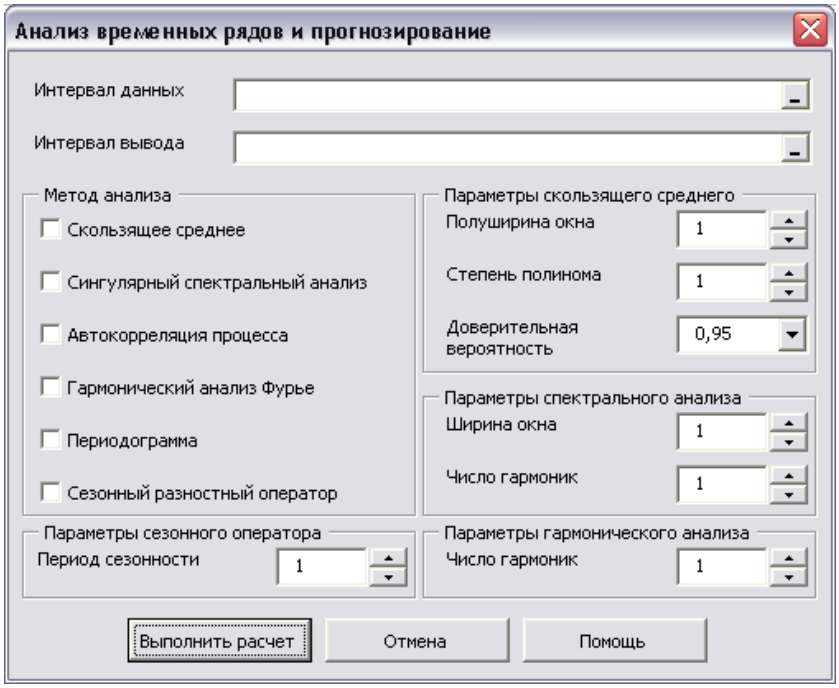
18.1. Введение

В данной главе описаны реализованные в программном обеспечении классические методы анализа временных рядов и прогнозирования.

18.2. Работа с программным обеспечением

Выберите из меню программы пункт AtteStat | Анализ временных рядов. На экране появится диалоговое окно, изображенное на рисунке:
Затем:

- Выберите или введите интервал временного ряда.
- Выберите или введите выходной интервал для выдачи результатов расчета. Начиная с первой ячейки выходного интервала (следовательно, можно указать только одну ячейку, т.к. остальные ячейки интервала игнорируются), будут выведены вычисленные показатели.



- Выберите метод анализа и относящиеся к данному методу параметры.
- Нажмите кнопку «Выполнить расчет».

При ошибках, вызванных неверными действиями пользователя при вводе исходных данных для расчета, выдаются сообщения об ошибках.

18.2.1. Сообщения об ошибках

При ошибках ввода и во время выполнения программы могут выдаваться диагностические сообщения следующих типов:

Ошибка	Комментарий
Не определен интервал переменной.	Вы не выбрали или неверно ввели интервал эмпирической выборки. Лучшим способом избежать ошибки является не ввод, а выделение интервала стандартным образом, т. е. протаскиванием курсора.
Пустая ячейка.	Проверьте исходные данные и заполните все ячейки, отмеченные Вами как входной интервал. Для избежания ошибок расчета, вызванных разногласиями, трактовать ли пустую ячейку как нуль, программа требует заполнения всех ячеек. Если в ячейке не должно быть данных по физической природе исследуемого процесса, введите в данную

	ячейку нуль.
Нечисловой тип данных.	Проверьте типы ячеек входного интервала. Тип может быть только числовым. Проще всего выделить интервал ячеек и явно определить их тип как числовой стандартными средствами.
Не определена область вывода.	Не выбран или неверно введен выходной интервал. Лучшим способом избежать ошибки является не ввод, а выделение интервала стандартным образом, т. е. протаскиванием курсора.

18.3. Теоретическое обоснование

Анализ временных рядов оперирует зависимостью случайной величины y_i, i =1,2,...,n, от контролируемой переменной t, в качестве которой обычно выступает время. В ряде моделей предполагается, что случайная величина состоит из истинного значения (тренда) с оценкой η_i, i =1,2,...,n, и нормально распределенной случайной составляющей (ошибки измерений) с нулевым средним значением. Далее, предполагается, что случайные величины y_i, i =1,2,...,n, наблюдаются через равные промежутки времени, а именно $t_i - t_{i-1} = const$, i =1,2,...,n. Данное предположение значительно упрощает все выкладки, позволяя также избавиться от ввода в программу временных отметок. Их роль играет номер отсчета. Введенные выше предположения являются общепринятыми.

Авторами выделяются основные задачи анализа временных рядов:

- исследование структуры временного ряда, в том числе описательные характеристики, выделение периодичностей, спектральный анализ,
- выделение сигнала на фоне шума,
- фильтрация и сглаживание.

Номенклатура задач не исчерпывается приведенным списком. Постоянно возникают новые прикладные задачи. Поставленные задачи решаются различными методами анализа, в том числе совокупностью представленных методов. Из методов анализа в программе реализованы следующие возможности:

- метод скользящего среднего,
- сезонный разностный оператор,
- сингулярный спектральный анализ,
- гармонический анализ Фурье,
- автокорреляционная функция,
- периодограмма.

18.3.1. Метод скользящего среднего

Метод скользящего среднего (moving average) основан на следующих соображениях. С учетом сделанных ранее предположений, определим оценку тренда в виде полинома

$$\eta_j = \sum_{i=0}^{l} x_{i+1} j^i,\ j = 1,2,...,n,$$

где n – численность временного ряда,

$x_i, i = 1,2,...,l+1,$ – коэффициенты полинома, вычисленные в точке j,

l – степень полинома.

В некоторых источниках оценка тренда называется прогнозом (прогнозируемыми значениями), однако данное наименование конфликтует с понятием прогноза, под которым подразумевается продолжение линии тренда за пределы исходного временного ряда.

Обозначим вектор коэффициентов полинома как $\bar{x}$. Методом наименьших квадратов (МНК)

найдено, что

$$\bar{x} = (A^T A)^{-1} A^T y,$$

где A – матрица размером $(2k + 1)(l + 1)$, элементы которой вычисляются по формуле

$$a_{ij} = (i - k - 1)^{j-1}, i = 1,2,...,2k + 1; j = 1,2,...,l + 1,$$

k – полуширина окна (усредняющего интервала), особенность которого для данного метода показана ниже.

Коэффициенты полинома вычисляются на основе исходного временного ряда y_i, i =1,2,...,n, причем для вычислений использован интервал данного ряда с центром в точке j, имеющий протяженность на k значений y_i, i =1,2,...,n, влево и вправо от центра интервала.

Единственное требование к выбору полуширины окна определяется, согласно МНК, тем, что число точек усредняющего интервала должно быть $l < 2k + 1$.

После всех вспомогательных вычислений в качестве значения оценки тренда в точке j берется значение x_1. Остальные компоненты вектора коэффициентов полинома применяются при расчете крайних точек тренда и соответствующих средних квадратичных отклонений, как будет показано ниже.

Алгоритм в представленной выше форме не позволяет получить оценки тренда первых k и последних k точек временного ряда. Для вычислений крайних значений используются вектора коэффициентов полинома, вычисленные, соответственно, для точек с номерами $k + 1$ и $n – k$ по формулам:

$$\eta_j = \sum_{i=0}^{l} x_{i+1}^{(k+1)} (j - k - 1)^i, j \le k,$$

$$\eta_j = \sum_{i=0}^{l} x_{i+1}^{(n-k)} (j + k - n)^i, j > n - k.$$

Значения j в показанных формулах могут быть продолжены как влево (для первой формулы), так и вправо (для второй формулы), обеспечивая потребности задачи прогнозирования.

Для вычисленного тренда программой определяются также доверительные интервалы по формуле

$$\eta_j^{\pm} = \eta_j \pm c_{11} s_j t_\beta, j = 1,2,...,n,$$

где c_{11} – элемент с индексами (1; 1) матрицы преобразования

$$C = (A^T A)^{-1},$$

$s_j, j = 1,2,...,n,$ – оценка среднего квадратичного отклонения, определяемая как

$$s_j = \sqrt{\frac{1}{2k - l} \sum_{i=-k}^{k} \left(y_j - \eta_i\right)^2}, j = 1,2,...,n,$$

причем индекс суммирования в формуле является относительным,

t_β – значение обратной функции t–распределения Стьюдента с параметрами $2k – l$ и $(1 + \beta) / 2$,

β – доверительный уровень, выраженный в долях.

Доверительный уровень выбирается из стандартной линейки и по умолчанию в программе равен 0,95. Данный параметр может быть изменен пользователем.

Для первых k и последних k точек временного ряда оценка среднего квадратичного отклонения вычисляется на основе того же принципа использования полученных коэффициентов полинома, соответственно, для точек с номерами $k + 1$ и $n – k$.

См. книги Брандта, Тюрина с соавт., Кулаичева.

18.3.2. Сезонный разностный оператор

Сезонные разностные операторы предназначены для удаления сезонных компонент. Процедура основана на формуле

$y_i = x_i - x_{i-p}, i = p+1,...,N,$

где $y_i, i = p+1,...,N,$ – элементы преобразованного временного ряда,

$x_j, j = 1,2,...,N,$ – элементы исходного временного ряда,

N – численность ряда,

p – период сезонности.

Процедура уменьшает численность ряда на величину p.

Описание см. в книге Тюрина с соавт.

18.3.3. Сингулярный спектральный анализ

Сингулярный спектральный анализ («Гусеница», singular spectrum analysis) предназначен для разделения исходного временного ряда на трендовые, сезонные и иные составляющие. Метод включает в себя ряд этапов: вложение, разложение по сингулярным числам, восстановление. Рассмотрим данные этапы подробно.

18.3.3.1. Вложение

Рассмотрим временной ряд X, состоящий из элементов x_i, i =1,2,...,N. Выберем некоторое целое число L, $1 < L < N$, которое назовем шириной окна. Затем будем двигать окно вдоль временного ряда. В результате применения данной процедуры вложения получится так называемая траекторная матрица A размером K x L, где $K = N - L + 1$.

18.3.3.2. Разложение по сингулярным числам

Каноническая формула разложения действительной прямоугольной матрицы A размером m x n (m строк, n столбцов, $m \geq n$) по сингулярным числам имеет вид

$$A = U \begin{pmatrix} \Sigma \\ 0 \end{pmatrix} V^T,$$

где U – матрица размером m x m, сформированная из m ортонормированных собственных векторов, соответствующих собственным значениям матрицы AA^T, $U^TU = I_m$,

Σ – диагональная матрица размером n x n, диагональные элементы которой представляют собой так называемые сингулярные числа – квадратные корни из λ_i, i =1,2,...,n – неотрицательных собственных значений матрицы A^TA,

0 – прямоугольная нулевая матрица размером $(m - n)n$,

V – матрица размером n x n, состоящая из n ортонормированных собственных векторов матрицы A^TA, $V^TV = VV^T = I_n$,

$I_.$ – единичная матрица соответствующего порядка.

В другой записи разложение по сингулярным числам имеет более простой вид

$A = U_n \Sigma V^T$,

где U_n – матрица размером m x n, сформированная из n ортонормированных собственных векторов, соответствующих n наибольшим из m собственным значениям матрицы AA^T, $U_n^T U_n = I_n$.

Раскрывая последнюю формулу, можно естественно получить, с учетом равенства нулю внедиагональных элементов матрицы Σ, что

$$A = \sum_{i=1}^{m} A_i,$$

где A_i, i =1,2,...,m – «элементарные» матрицы размером m x n, элементы которых определяются согласно формуле

$$a_{ij} = \sqrt{\lambda_j} U_i V_j^T, i = 1,2,..., m, j = 1,2,..., n,$$

где U_i, i =1,2,...,m – столбец матрицы U_n,
V_j, j =1,2,...,n – столбец матрицы V.

См. статьи Голуба (Golub) с соавт., Стюарта (Stewart), книги Деммеля, Голуба с соавт.

18.3.3.3. Восстановление

Существует взаимно однозначное соответствие между матрицами A_i, i =1,2,...,m, размером K x L,полученными на предыдущем этапе, и «элементарными» временными рядами каждый X_i, i =1,2,...,m, длиной N. Здесь величину m, $m \le L$, можно интерпретировать как выбираемое пользователем программы число выделяемых компонент временного ряда (гармоник). Если выбранное пользователем число гармоник превышает указанный предел, оно уменьшается до величины этого предела.
Восстановление «элементарных» временных рядов производится методом диагонального усреднения матриц A_i, i =1,2,...,m, суть которого заключается в том, что каждый элемент ряда X_i, i =1,2,...,m, будет получен как среднее арифметическое величин, стоящих на «антидиагоналях» соответствующей матрицы A_i, i =1,2,...,m.
В отличие от гармонического анализа Фурье, описываемый метод не позволяет получить разложение исходного временного ряда на «чистые» гармоники. Пользователь может в этом убедиться, проведя сравнительные расчеты разными методами. Здесь полезно привести аналогию со ступенчатым регрессионным анализом, представленном во 2 книге монографии Дрейпера с соавт. и процитировать данное там положение: «Этот метод не дает правильного МНК–решения для переменных, включенных в итоговое уравнение».

См. работы Вотарда (Vautard) с соавт., Голяндиной с соавт., Александрова с соавт.

18.3.4. Гармонический анализ Фурье

Рассматриваемый метод называют гармоническим анализом Фурье (гармоническим регрессионным анализом). Конечный ряд Фурье представляет периодическую функцию $y(t)$ в виде линейной комбинации r гармоник (гармонических векторов).
Пусть временной ряд задан в виде N отсчетов временного ряда y_n, n = 1,2,...,N, в равноотстоящих точках t_n, n = 1,2,...,N. Исходный временной ряд может быть представлен в виде конечного ряда Фурье, определяемого формулой

$$y_n = a_0 + \sum_{k=1}^{r} a_k \cos\frac{2\pi kn}{N} + \sum_{k=1}^{r} b_k \sin\frac{2\pi kn}{N}, n = 1,2,..., N,$$

где коэффициенты вычисляются по формулам

$$a_0 = \frac{1}{N}\sum_{k=1}^{N} y_k,$$

$b_0 = 0$,

$$a_m = \frac{2}{N}\sum_{k=1}^{N} y_k \cos\frac{2\pi km}{N}, m = 0,1,..., r,$$

$$b_m = \frac{2}{N}\sum_{k=1}^{N} y_k \sin\frac{2\pi km}{N}, m = 0,1,...,r,$$

где y_k, $k = 1,2,...,N$ – отсчеты временного ряда в точках t_k, $k = 1,2,...,N$.
m – номер гармоники,
N – количество наблюдений – число равных частей, на которые разделен период наблюдения,
r – количество гармоник, $r \le N / 2$.
Количество гармоник выбирается пользователем программы. Если выбранное число гармоник превышает указанный предел, оно уменьшается до величины этого предела.
Дополнительно программа выдает следующие параметры для каждой гармоники:

$A_m = \sqrt{a_m^2 + b_m^2}, m = 0,1,...,r,$ – амплитуда,

$\theta_m = \text{arctg}(-b_m / a_m) \cdot 180 / \pi$, $m = 0,1,...,r$ – фаза.

Рассмотренный метод описан в большом числе классических и современных источников. См., например, главу 18 справочника по ред. Ллойда с соавт., с. 365 монографии «Прикладной анализ случайных данных» Бендата с соавт., с. 85 книги Носача.

18.3.5. Автокорреляционная функция

Выборочная автокорреляционная функция (сериальная корреляция) строится по формуле

$$r_k = \frac{\sum_{i=1}^{N-1}(x_i - \bar{x})(x_{i+k} - \bar{x})}{\sum_{i=1}^{N}(x_i - \bar{x})^2}, k = 1,2,..., N-1,$$

где x_i, $i = 1,2,...,N$ – элементы временного ряда,
N – численность ряда,

$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$ – оценка среднего значения.

Программа производит все необходимые вычисления и выводит график автокорреляционной функции – коррелограмму. На графике показан также 95% доверительный интервал. Границы данного интервала называют корреляционной трубкой и вычисляют по формуле

$$r^{\pm} = -\frac{1}{n} \pm \frac{2}{\sqrt{n}}.$$

Подробное описание см. в книге Тюрина с соавт.

18.3.6. Периодограмма

Периодограмма временного ряда y_n, $n = 1,2,...,N$, состоит из $r = N / 2$ значений, называемых интенсивностями и вычисляемых по формуле

$$I(f_i) = \frac{N}{2}(a_i^2 + b_i^2), i = 1,2,..., r,$$

где a_i, b_i, $i = 1,2,...,r$ – коэффициенты ряда Фурье,
$f_i = i / N$, $i = 1,2,...,r$ – i–я гармоника основной частоты.

См. монографию «Прикладной анализ случайных данных» Бендата с соавт., а также с. 52 первого выпуска книги Бокса и Дженкинса.

Список использованной и рекомендуемой литературы

1. Allen M.R. Monte Carlo SSA: detecting irregular oscillations in the presence of coloured noise // Journal of Climate, December 1996, vol. 9, pp. 3373–3404.
2. Allen M.R. Optimal filtering in singular spectrum analysis // Physics Letters, October 1997, vol. 234, no. 6, pp. 419–428.
3. Bose N.K. Handbook of statistics. Vol. 10. Signal processing and its applications / Ed. by N.K. Bose, C.R. Rao. – New York, NY: Elsevier, 1993.
4. Bretthorst G.L. Bayesian spectrum analysis and parameter estimation. – Berlin: Springer–Verlag, 1988.
5. Brillinger D.R. Handbook of statistics. Vol. 3. Time series in the frequency domain / Ed. by D.R. Brillinger, P.R. Krishnaiah. – New York, NY: Elsevier, 1983.
6. Broomhead D.S., King G.P. Extracting qualitative dynamics from experimental data // Physica D: Nonlinear Phenomena, vol. 20, issues 2–3, June–July 1986, pp. 217–236.
7. Burg J.P. A new technique for time series data // Modern Spectrum Analysis / Ed. by D.G. Childers. – New York, NY: IEEE Press, 1978, pp. 42–48.
8. Cadzow J.A. Signal enhancement – A composite property mapping algorithm // IEEE Transactions of Acoustics, Speech and Signal Processing, January 1988, vol. 36, no. 1.
9. Chatfield C. The analysis of time–series: An introduction. – London, UK: Chapman & Hall / CRC, 2003.
10. Chatfield C. Time–series forecasting. – London, UK: Chapman & Hall / CRC, 2000.
11. Chen C.H. Signal processing handbook / Ed. by C.H. Chen. – New York, NY: Marcel Dekker, 1988.
12. Cornelissen G. Statistical significance without biologic signification is not enough: illustrative example / G. Cornelissen, R.B. Sothern, H.W. Wendt et al. // Chronobiologia, 1994, vol. 21, no. 3–4, pp. 315–20.
13. Cugini P. Chronobiology: principles and methods // Annali dell'Istituto Superiore di Sanita, 1993, vol. 29, no. 4, pp. 483–500.
14. DeSa R.J., Matheson I.B.C. A practical approach to interpretation of singular value decomposition results // Methods in Enzymology, 2004, vol. 384, Numerical Computer Methods, part E, pp. 1–8.
15. Doob J.L. Time series and harmonic analysis // Proceedings of the Berkeley symposium on mathematical statistics and probability, August 13–18, 1945 and January 27–29, 1946 / Ed. by J. Neyman. – Berkeley, CA: University of California Press, 1949, pp. 303–343.
16. Elkum N.B., Myles J.D. Modeling biological rhythms in failure time data // Journal of Circadian Rhythms, 2006, no. 4:14.
17. Elsner J.B., Tsonis A.A. Singular spectrum analysis. A new tool in time series analysis. – New York, NY: Plenum Press, 1996.
18. Franses P.H., van Dijk D. Nonlinear time series models in empirical finance. – Cambridge, UK: Cambridge University Press, 2000.
19. Ghil M. Advanced spectral methods for climatic time series / M. Ghil, M.R. Allen, M.D. Dettinger et al. // Reviews of Geophysics, 2002, vol. 40, no. 1, pp. 1–41.
20. Ghil M. The SSA–MTM toolkit: Applications to analysis and prediction of time series // Applications of Soft Computing, Proceedings of SPIE, Bellingham, WA / Ed. by B. Bosacchi, J.C. Bezdek, D.B. Fogel, 1997, vol. 3165, pp. 216–230.
21. Golub G., Kahan W. Calculating the singular values and pseudo–inverse of a matrix // Journal oh the Society for Industrial and Applied Mathematics, Series B: Numerical Analysis, January 1965, vol. 2, no. 2, pp. 205–224.
22. Golyandina N., Nekrutkin V., Zhigljavsky A. Analysis of time series structure: SSA and related techniques. – London, UK: Chapman & Hall / CRC, 2001.

23. Gray R.M., Davisson L.D. An introduction to statistical signal processing. – Cambridge, UK: Cambridge University Press, 2004.
24. Halberg F. Chronobiology // Annual Review of Physiology, March 1969, vol. 31, pp. 675–726.
25. Hamilton J.D. Time series analysis. – Princeton, NJ: Princeton University Press, 1994.
26. Hannan E.J. Handbook of statistics. Vol. 5. Time series in the time domain / Ed. by E.J. Hannan, P.R. Krishnaiah, M.M. Rao. – New York, NY: Elsevier, 1985.
27. Hoenen S., Schimmel M., Marques M.D. Rescuing rhythms from noise: A new method of analysis // Biological Rhythm Research, 2001, vol. 32, no. 2, pp. 271–284.
28. Hosmer D.W., Lemeshow S. Applied survival analysis: regression modeling of time to event data. – New York, NY: John Wiley & Sons, 1999.
29. Katayama T. Statistical methods in control and signal processing / Ed. by T. Katayama, S. Sugimoto. – New York, NY: Marcel Dekker, 1997.
30. Kondrashov D., M. Ghil. Spatio–temporal filling of missing points in geophysical data sets // Nonlinear Processes in Geophysics, 2006, vol. 13, pp. 151–159.
31. Kugiumtzis D., Christophersen N. State space reconstruction: Method of delays vs. singular spectrum approach // Research report, Department of Informatics, University of Oslo, 13 February 1997.
32. McLeod A.I., Yu H., Krougly Z.L. Algorithms for linear time series analysis: With R package // Journal of Statistical Software, December 2007, vol. 23, no. 5.
33. Montgomery D.C., Johnson L.A., Gardiner J.S. Forecasting and time series analysis. – New York, NY: McGraw–Hill, 1990.
34. Nelson W. Methods for cosinor–rhythmometry / W. Nelson, Y.L. Tong, J.K. Lee et al. // Chronobiologia, 1979, vol. 6, no. 4, pp. 305–323.
35. Pollock D.S.G. A handbook of time–series analysis, signal processing and dynamics. – London, UK: Academic Press, 1999.
36. Pollock D.S.G. Circulant matrices and time–series analysis // Queen Mary, University of London, Working Paper No. 422, October 2000.
37. Raynaud S. Using MSSA to determine explicity the oscillatory dynamics of weakly nonlinear climate systems / S. Raynaud, P. Yiou, R. Kleeman et al. // Nonlinear Processes in Geophysics, 2005, vol. 12, pp. 807–815.
38. Rol de Lama M.A. How to engage medical students in chronobiology: an example on autorhythmometry / M.A. Rol de Lama, J.P. Lozano, V. Ortiz et al. // Advances in Physiology Education, 2005, vol. 29, pp. 160–164.
39. Romberg T.M., Black J.L., Ledwidge T.J. Signal processing for industrial diagnostics. – Chichester, UK: John Wiley & Sons, 1996.
40. Schelter B. Handbook of time series analysis: Recent theoretical developments and applications // Ed. by B. Schelter, M. Winterhalder, J. Timmer. – New York, NY: John Wiley & Sons, 2006.
41. Schoellhamer D.H. Singular spectrum analysis for time series with missing data // Geophysical Research Letters, 2001, vol. 28, no. 16, pp. 3187–3190.
42. Stewart D.E. A new algorithm for the SVD of a long product of matrices and the stability of products // Electronic Transactions on Numerical Analysis, June 1997, vol. 5, pp. 29–47.
43. Tsay R.S. Analysis of financial time series. – New York, NY: John Wiley & Sons, 2002.
44. Varadi F. Random–lag singular cross–spectrum analysis / F. Varadi, R.K. Ulrich, L. Bertello et al. // The Astrophysical Journal, 1 January 2000, vol. 528, pp. L53–L56.
45. Varadi F. Searching for signal in noise by random–lag singular spectrum analysis / F. Varadi, J.M. Pap, R.K. Ulrich et al. // The Astrophysical Journal, 1999, vol. 526, pp.1052–1061.
46. Vautard R., Ghil M. Singular spectrum analysis in nonlinear dynamics, with applications to

paleoclimatic time series // Physica D: Nonlinear Phenomena, May 1989, vol. 35, no. 3, pp. 395–424.
47. Vautard R., Yiou P., Ghil M. Singular–spectrum analysis: A toolkit for short, noisy chaotic signals // Physica D: Nonlinear Phenomena, 15 September 1992, vol. 58, issues 1–4, pp. 95–126.
48. Yiou P. Nonlinear variability of the climate system, from singular and power spectra of quaternary records / P. Yiou, M. Ghil, J. Jouzel et al. // Climate Dynamics, 1994, vol. 9, pp. 371–389.
49. Yiou P., Sornette D., Ghil M. Data–adaptive wavelets and multi–scale singular–spectrum analysis // Physica D: Nonlinear Phenomena, 15 August 2000, vol. 142, issues 3–4, pp. 254–290.
50. Александров Ф., Голяндина Н. Автоматизация выделения трендовых и периодических составляющих временного ряда в рамках метода «Гусеница»–SSA // Exponenta Pro, 2004, № 3–4, с. 54–61.
51. Александров Ф., Голяндина Н. Выбор параметров при автоматическом выделении трендовых и периодических составляющих временного ряда в рамках подхода «Гусеница»–SSA // Труды IV Международной конференции «Идентификация систем и задачи управления» SICPRO’05. – Москва, 2005, с. 1849–1864.
52. Андерсон Т. Статистический анализ временных рядов. – М.: Мир, 1976.
53. Афанасьев В.Н., Юзбашев М.М. Анализ временных рядов и прогнозирование. – М.: Финансы и статистика, 2001.
54. Ахмед Н., Рао К.Р. Ортогональные преобразования при обработке цифровых сигналов. – М.: Связь, 1980.
55. Ашофф Ю. Биологические ритмы. Т. 1. / Под ред. Ю. Ашоффа. – М.: Мир, 1984.
56. Ашофф Ю. Биологические ритмы. Т. 2. / Под ред. Ю. Ашоффа. – М.: Мир, 1984.
57. Бассвиль М. Обнаружение изменения свойств сигналов и динамических систем / М. Бассвиль, А. Вилски, А. Банвенист и др. – М.: Мир, 1989.
58. Бендат Дж. Основы теории случайных шумов. – М.: Наука, 1965.
59. Бендат Дж., Пирсол А. Измерение и анализ случайных процессов. – М.: Мир, 1971.
60. Бендат Дж., Пирсол А. Прикладной анализ случайных данных. – М.: Мир, 1989.
61. Бендат Дж., Пирсол А. Применение корреляционного и спектрального анализа. – М.: Мир, 1983.
62. Бокс Дж., Дженкинс Г. Анализ временных рядов: Прогноз и управление. Вып. 1. – М.: Мир, 1974.
63. Бокс Дж., Дженкинс Г. Анализ временных рядов: Прогноз и управление. Вып. 2. – М.: Мир, 1974.
64. Большаков А.А., Каримов Р.Н. Методы обработки многомерных данных и временных рядов: Учебное пособие для вузов. – М.: Горячая линия – Телеком, 2007.
65. Брандт З. Анализ данных. Статистические и вычислительные методы для научных работников и инженеров. – М.: Мир, ООО «Издательство АСТ», 2003.
66. Бриллинджер Д. Временные ряды. Обработка данных и теория. – М.: Мир, 1980.
67. Гайдышев И. Анализ и обработка данных: специальный справочник. – СПб: Питер, 2001.
68. Гальберг Ф. Хронобиология / Кибернетический сборник. Новая серия. Выпуск 9. Сборник переводов / Под ред. А.А. Ляпунова, О.Б. Лупанова, с. 189–247.
69. Гласс Л., Мэки М. От часов к хаосу: Ритмы жизни. – М.: Мир, 1991.
70. Голуб Дж., Ван Лоун Ч. Матричные вычисления. – М.: Мир, 1999.
71. Гольденберг Л.М., Матюшкин Б.Д., Поляк М.Н. Цифровая обработка сигналов: Справочник. – М.: Радио и связь, 1985.

72. Голяндина Н., Некруткин В., Степанов Д. Варианты метода «Гусеница»–SSA для анализа многомерных временных рядов // Труды II Международной конференции «Идентификация систем и задачи управления» SICPRO’03. – Москва, 2003, с. 2139–2168.
73. Голяндина Н.Э. Метод «Гусеница»–SSA: Анализ временных рядов: Учебное пособие. – СПб: Издательство СПбГУ, 2004.
74. Голяндина Н.Э. Метод «Гусеница»–SSA: Прогноз временных рядов: Учебное пособие. – СПб: Издательство СПбГУ, 2004.
75. Гренджер К., Хатанака М. Спектральный анализ временных рядов в экономике. – М.: Статистика, 1972.
76. Гроп Д. Методы идентификации систем. – М.: Мир,1979.
77. Данилов Д.Л. Главные компоненты временных рядов: метод «Гусеница» / Под ред. Д.Л. Данилова, А.А. Жиглявского. – СПб: Издательство СПбГУ, 1997.
78. Деммель Дж. Вычислительная линейная алгебра. Теория и приложения. – М.: Мир, 2001.
79. Дженкинс Г., Ваттс Д. Спектральный анализ и его приложения. Выпуск 1. – М.: Мир, 1971.
80. Дженкинс Г., Ваттс Д. Спектральный анализ и его приложения. Выпуск 2. – М.: Мир, 1972.
81. Дрейпер Н., Смит Г. Прикладной Регрессионный анализ. Книга 1. – М.: Финансы и статистика, 1986.
82. Дрейпер Н., Смит Г. Прикладной Регрессионный анализ. Книга 2. – М.: Финансы и статистика, 1987.
83. Ефимов В.М., Галактионов Ю.К., Шушпанова Н.Ф. Анализ и прогноз временных рядов методом главных компонент. – Новосибирск: Наука, Сибирское отделение, 1988.
84. Зверев В.А., Стромков А.А. Выделение сигналов из помех численными методами. – Н.Новгород: ИПФ РАН, 2001.
85. Канторович Г.Г. Анализ временных рядов // Экономический журнал ВШЭ, 2002, №1, с. 85–116; 2002, №2, с. 251–273; 2002, №3, с. 379–401; 2002, №4, с. 498–523; 2003, № 1, с. 79–103.
86. Карп В.П., Катинас Г.С. Вычислительные методы анализа в хронобиологии и хрономедицине. – СПб.: Восточная корона, 1997.
87. Кендалл М., Стьюарт А. Многомерный статистический анализ и временные ряды. – М.: Наука, 1976.
88. Кендэл М. Временные ряды. – М.: Финансы и статистика, 1981.
89. Кожевникова И.А. Выявление скрытых периодичностей // Заводская лаборатория. Диагностика материалов, 2006, № 3, с. 59–65.
90. Кулаичев А.П. Компьютерный контроль процессов и анализ сигналов. – М.: Информатика и компьютеры, 1999.
91. Кулаичев А.П. Методы и средства анализа данных в среде Windows®. STADIA. – М.: Информатика и компьютеры, 1999.
92. Кулаичев А.П. Методы и средства комплексного анализа данных. – М.: ИНФРА–М, 2006.
93. Ллойд Э. Справочник по прикладной статистике. В 2–х т. Т. 2 / Под ред. Э. Ллойда, У. Ледермана, С.А. Айвазяна и др. – М.: Финансы и статистика, 1990.
94. Льюис К.Д. Методы прогнозирования экономических показателей. – М.: Финансы и статистика, 1986.
95. Макс Ж. Методы и техника обработки сигналов при физических измерениях. – М.:

Мир, 1983.
96. Марпл–мл. С.Л. Цифровой спектральный анализ и его приложения. – М.: Мир, 1990.
97. Медведев Г.А., Морозов В.А. Практикум на ЭВМ по анализу временных рядов. – Минск: Университетское, 2001.
98. Никифоров И.В. Последовательное обнаружение изменения свойств временных рядов. – М.: Наука, 1983.
99. Носач В.В. Решение задач аппроксимации с помощью персональных компьютеров. – М.: МИКАП, 1994.
100. Нуссбаумер Г. Быстрое преобразование Фурье и алгоритмы вычисления сверток. – М.: Радио и связь, 1985.
101. Отнес Р., Эноксон Л. Прикладной анализ временных рядов: Основные методы. – М.: Мир, 1982.
102. Плюта В. Сравнительный многомерный анализ в экономических исследованиях. – М.: Статистика , 1980.
103. Пойда В.Н. Спектральный анализ в дискретных ортогональных базисах. – Минск: Наука и техника, 1978.
104. Рабинер Л., Голд Б. Теория и практика цифровой обработки сигналов. – М.: Мир, 1975.
105. Сато Ю. Обработка сигналов. – М.: Издательский дом «Додэка–XXI», 2002.
106. Сергиенко А.Б. Цифровая обработка сигналов. – СПб.: Питер, 2002.
107. Серебренников М.Г., Первозванский А.А. Выявление скрытых периодичностей. – М.: Наука, 1965.
108. Степанов А., Матвеев С. Цифровая обработка зашумленной речи: методы и программные средства // Компьютер–пресс, 1999, №8.
109. Степанов Д., Голяндина Н. Варианты метода «Гусеница»–SSA для прогноза многомерных временных рядов // Труды IV Международной конференции «Идентификация систем и задачи управления» SICPRO’05. – Москва, 2005, с. 1831–1848.
110. Теребиж В.Ю. Анализ временных рядов в астрофизике. – М.: Наука, 1992.
111. Тюрин Ю.Н., Макаров А.А. Статистический анализ данных на компьютере. – М.: ИНФРА–М, 1999.
112. Уилкинсон Дж.Х. Алгебраическая проблема собственных значений. – М.: Наука, 1970.
113. Хеннан Э. Многомерные временные ряды. – М.: Мир, 1974.
114. Химмельблау Д. Анализ процессов статистическими методами. – М.: Мир, 1973.
115. Чуев Ю.В., Михайлов Ю.Б., Кузьмин И.В. Прогнозирование количественных характеристик процессов. – М.: Советское радио,1975.
116. Шугай Ю.С. Нейросетевой алгоритм прогнозирования событий в многомерных временных рядах и его применение для анализа космофизических данных / Ю.С. Шугай, С.А. Доленко, И.Г. Персианцев и др. // Труды 7–й Международной конференции «Распознавание образов и анализ изображений: новые информационные технологии» РОАИ–7–2004, Санкт–Петербург,2004.

Глава 19. Статистический контроль качества

19.1. Введение

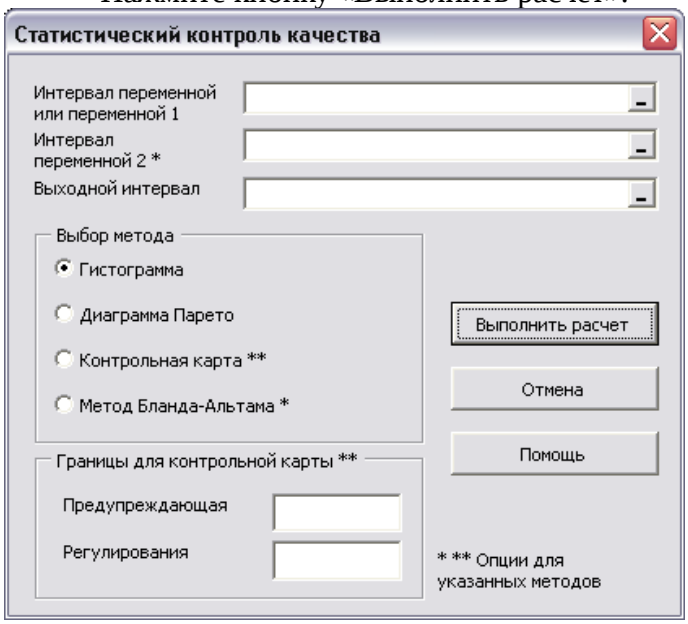
Статистические методы контроля качества контроля качества в лаборатории и на

производстве предназначены для лабораторного контроля, для контроля качества выпускаемой продукции или оказываемых услуг с целью своевременного выявления нарушений в организации производства и в технологических процессах, приводящих к снижению качества продукции или услуг ниже норм, заданных техническими условиями. Контроль качества интересен как инструмент успешного эффективного решения задач, возникающих на этапе внедрения в производство передовых методов управления.

19.2. Работа с программным обеспечением

Выберите из меню программы пункт AtteStat | Статистический контроль качества. На экране появится диалоговое окно, изображенное на рисунке:
Затем проделайте следующие шаги:

- Выберите или введите интервал переменной.
- Выберите или введите выходной интервал. Начиная с первой ячейки выходного интервала (следовательно, можно указать только одну ячейку, т.к. остальные ячейки интервала игнорируются), будут выведены результаты вычислений.
- Выберите или оставьте по умолчанию тип диаграммы.
- Для контрольной карты введите границы.
- Нажмите кнопку «Выполнить расчет».



После выполнения вычислений будут, начиная с первой ячейки выходного интервала, выведены результаты расчета, включая графическую информацию.
Программное обеспечение берет на себя верификацию исходных данных, выдавая подробную диагностику. При неверных действиях пользователя выдаются сообщения об ошибках.

19.2.1. Сообщения об ошибках

При ошибках ввода данных могут выдаваться диагностические сообщения следующих типов:

Ошибка	Комментарий
Не определена	Не выбран или неверно введен интервал данных. Лучшим способом

Ошибка	Комментарий
область данных.	избежать ошибки является не ввод, а выделение интервала ячеек стандартным образом.
Не определена область вывода.	Не выбран или неверно введен интервал ячеек, определяющих область вывода решения. Лучшим способом избежать ошибки является не ввод, а выделение интервала ячеек стандартным образом.
Не задана предупреждающая граница.	Для построения контрольной карты требуется задать предупреждающую границу.
Не задана граница регулирования	Для построения контрольной карты требуется задать границу регулирования.
Нулевая сумма вариант.	Сумма вариант выборки равна нулю, поэтому расчет показателей качества для такой конфигурации исходных данных провести нельзя.

19.3. Теоретическое обоснование

Современное управление качеством основано на широком использовании статистических методов. Статистический контроль качества, по определению У. Деминга – это применение статистических принципов и приемов на всех стадиях производства, направленное на экономичное производство изделия, максимально полезного и имеющего сбыт.
Статистическое управление качеством, по определению Дж. Мердока – это совокупность методов обнаружения неслучайных факторов, позволяющих диагностировать состояние процесса, провести его корректировку и, в конечном счете, способствующих улучшению качества продукции.
Рассмотренные методы статистического контроля качества:

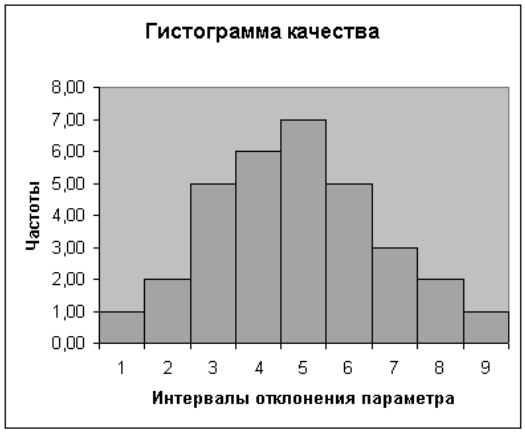
- гистограмма качества,
- диаграмма Парето,
- контрольная карта –

принадлежат к совокупности семи элементарных методов контроля качества, введенной в классических источниках по контролю качества. К другим, не рассмотренным здесь статистическим методам контроля качества, относятся точечный график и диаграмма разброса. Последние два классических метода: диаграмма Исикавы (диаграмма «причины – результат») и таблица контроля – не относятся к статистическим методам.
Анализ Бланда–Альтмана, также представленный в программе, предназначен для сравнения двух методов клинического или лабораторного контроля.
Данные методы просты, наглядны, удобны. Применение совокупности этих методов, по утверждению оригинальных источников, решает 95% всех производственных проблем. Не стоит, однако, понимать статистическое управление качеством в том смысле, что его применение даст немедленный практический эффект. Статистическое управление качеством не предназначено для решения в принципе неразрешимых проблем.
Отметим, что в статистическом управлении качеством используются разнообразные методы статистического анализа данных: описательная статистика; другие методы предварительной обработки данных; корреляционный анализ.
При анализе представленных теоретических материалов у пользователей, не являющихся специалистами в статистическом контроле качества и только начинающих знакомиться с

теорией и практикой статистического контроля качества, может возникнуть впечатление, что в силу большого объема предлагаемых материалов их изучение может быть особенно трудным. На самом деле решение возникающих задач следует начать не с изучения всевозможных материалов, а с постановки задачи, вначале словесной («нужно улучшить качество»). Затем следует перевести постановку на некий промежуточный язык («от каких параметров зависит качество») и только потом переходить к математической формулировке задачи статистического контроля качества. Предлагаемое любое программное обеспечение не заменит рассмотренного процесса постановки задачи. Оно лишь поможет в решении задачи. Не исключено, правда, что в процессе решения задачи придет более конкретное ее понимание.

19.3.1. Гистограмма качества

Гистограмма качества позволяет в наглядной форме отобразить выявленный характер разброса значений контролируемого параметра. Дополнительно на гистограмме качества могут отображаться среднее значение, стандартное отклонение и заданные границы допуска. Исходными данными для построения гистограммы качества служат количества вариант (частоты), относящиеся к каждому интервалу отклонения исследуемого параметра. На графике частоты откладываются по оси ординат. По оси абсцисс откладываются кодовые обозначения интервалов отклонения параметра (классов). Гистограмма может быть построена с помощью одноименного метода, представленного в главе «Описательная статистика».



Умение читать и анализировать гистограммы окажет неоценимую услугу специалисту не только в статистическом управлении качеством, но также и в других разделах статистического анализа данных.
Встречаются следующие основные типы гистограмм:

- Обычный тип. Гистограмма имеет симметричную колоколообразную форму. Среднее значение приходится примерно на середину размаха данных. Этот тип свидетельствует об однородности исходных данных, а в статистическом контроле качества – о нормальном протекании технологического процесса.
- Положительно (отрицательно) скошенное распределение. Форма асимметрична. Среднее значение локализуется справа (слева) от середины размаха. Такая форма встречается, когда нижняя (верхняя) граница регулируется либо теоретически, либо по значению допуска или когда левое (правое) значение недостижимо.
- Распределение с обрывом слева. Форма асимметрична. Среднее арифметическое локализуется слева (справа) от середины размаха. Эта форма встречается при 100%

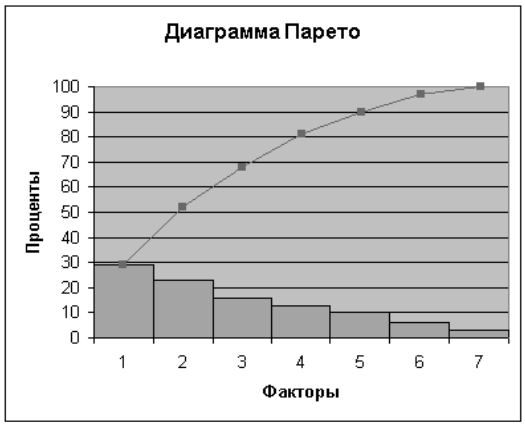
просеивании изделий из–за плохой воспроизводимости процесса.
- Плато (равномерное и прямоугольное распределения). Такая форма встречается в смеси нескольких распределений, имеющих различные средние.
- Бимодальное (двухпиковое) распределение. Такая форма встречается, когда смешиваются два распределения с далеко отстоящими друг от друга средними значениями.
- Распределение с изолированным пиком. Такая форма проявляется при наличии малых включений из другого распределения (из другого процесса), появления ошибки измерения или в случае нарушения нормальности процесса.

Рассмотрим влияние формы гистограммы качества на действия специалиста по контролю качества. В случае 1 процесс считается протекающим нормально. В случаях 2 и 3 требуется вмешательство специалистов для проверки и, если потребуется, наладки технологического процесса. Случаи 4, 5 и 6 свидетельствуют о неоднородности данных. Неоднородность может быть вызвана ошибками при сборе данных или свидетельствовать о нестабильности технологического процесса.
Как уже было отмечено, в некоторых источниках на гистограмму накладываются границы допуска (промежуток оси абсцисс между границами допуска называется полем допуска). Можно рассуждать о том, допустимо ли на одной диаграмме объединять график порядковой переменной (гистограмма) и количественной (границы допуска), а такие заблуждения распространены повсеместно. Однако так делают некоторые авторы, поэтому рассмотрим, как можно с некоторой пользой применить данную информацию. Процесс считается протекающим нормально, если почти вся гистограмма находится в границах допуска. В этом случае требуется лишь поддержание существующего состояния технологического процесса. Если гистограмма не удовлетворяет допуску, необходимо добиться смещения среднего значения ближе к центру поля допуска.

19.3.2. Диаграмма Парето

Диаграмма Парето (диаграмма распределения Парето) служит для наглядного выявления наиболее значимого фактора, влияющего на снижение качества продукции.
В разных источниках диаграмма Парето может изображаться по–разному, однако общим во всех источниках является комбинация на поле данной диаграммы двух графиков:
- графика типа гистограммы, изображающего процент брака по вине того или иного фактора,
- ломаной линии (полигоном, «кривой эффективности»), отражающей накопленные проценты.

Под браком здесь подразумевается несоответствие достигнутых показателей качества с показателями качества производственного процесса, определяемыми техническими заданиями.

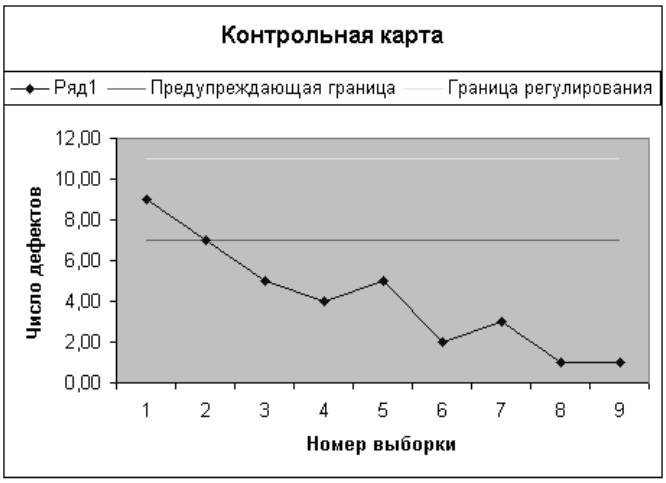


Исходными данными для построения диаграммы Парето служат количества вариант, относящиеся к каждому фактору, ответственному за снижение качества продукции, либо частоты. На графике по оси ординат откладываются частоты, выраженные в процентах. По оси абсцисс откладываются кодовые обозначения факторов. Частоты могут быть найдены так же, как указано в разделе, посвященном гистограмме качества.

19.3.3. Контрольная карта

Контрольная карта предназначена для обнаружения отклонения характеристики качества выпускаемой продукции от заданных технологических норм и допусков.
Контрольная карта представляет собой график изменения исследуемой характеристики во времени. На график дополнительно наносятся предупреждающая граница и граница регулирования. Различные типы контрольных карт (обычно их выделяют шесть) описаны в литературе. В терминологии источников рассмотренная нами контрольная карта относится к типам:

- «pn–карта», т. е. когда показатель качества представлен числом дефектных изделий в последовательности выборок фиксированного объема,
- «c–карта», т. е. когда управление качеством контролируемого производственного процесса ведется по числу дефектов в изделиях одинакового размера.



Если характеристика качества производственного процесса находится ниже предупреждающей границы, то процесс протекает нормально. Если характеристика процесса

находится между предупреждающей границей и границей регулирования, то технологический процесс функционирует, но не в соответствии с номиналом. Попадание характеристики в зону выше границы регулирования означает, что должна быть произведена коррекция технологического процесса.
За границу регулирования часто принимается величина, равная утроенному стандартному отклонению, что, как известно, означает попадание в данные границы 95% вариант в том случае, если распределение нормальное. Подробнее о нормальном распределении и его проверке см. в главе «Проверка нормальности распределения».
Исходными данными для построения контрольной карты служат количества вариант, относящиеся к каждому фактору, ответственному за снижение качества продукции. На графике по оси ординат откладываются частоты, выраженные в абсолютных величинах (в штуках). По оси абсцисс откладываются кодовые обозначения номеров выборок, отобранных в процессе производства с целью его контроля. Дополнительно в тех же единицах измерения, что и исходные данные, должны быть заданы предупреждающая граница и граница регулирования.

19.3.4. Анализ Бланда–Альтмана

Метод Бланда–Альтмана предназначен для сравнения двух методов клинического или лабораторного контроля. Метод основан на анализе графика, который представляет собой зависимость разности измерений двух методов от среднего данных измерений с указанными средним разностей и 95% доверительными интервалами этого оцениваемого среднего.
Средние значения двух измерений вычисляются по формуле

$$z_i = \frac{x_i + y_i}{2}, i = 1,2,...,n,$$

где x_i, i = 1,2,...,n – измерения 1–го метода,
y_i, i = 1,2,...,n – измерения 2–го метода,
n – число измерений каждого метода.
Разности значений двух измерений вычисляются по формуле
$d_i = x_i - y_i$, i = 1,2,...,n,
если среднее значение 1–го метода больше среднего значения 2–го метода, либо
$d_i = y_i - x_i$, i = 1,2,...,n,
если среднее значение 1–го метода меньше среднего значения 2–го метода.
При этом соответствующие средние значения вычисляются по формулам

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \text{и} \quad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i.$$

Среднее разности вычисляется по формуле

$$\bar{d} = \frac{1}{n}\sum_{i=1}^{n} d_i.$$

Вычисление двустороннего доверительного интервала оцениваемого среднего разности производится по формуле

$$I_{\bar{d}} = \left(\bar{d} - t_{(1+\beta)/2} \cdot \sqrt{\frac{DD}{n}}; \bar{d} + t_{(1+\beta)/2} \cdot \sqrt{\frac{DD}{n}} \right),$$

где DD – дисперсия разности ,
$t_{(1+\beta)/2}$ – значение обратной функции t–распределения Стьюдента с параметрами n – 1 и (1 + β) / 2,
β – доверительный уровень, выраженный в долях.

В обсуждаемом методе условились применять доверительный уровень 95%.
Дисперсия разности вычисляется как

$$DD = \frac{DX + DY}{2},$$

где DX – дисперсия 1–го метода,
DY – дисперсия 2–го метода.
При этом соответствующие дисперсии вычисляются по формулам

$$DX = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \quad \text{и} \quad DY = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2.$$

См. оригинальные статьи Альтмана (Altman) и Бланда (Bland), а также статьи Девитте (Dewitte) с соавт., Манта (Mantha) с соавт., Стокла (Stockl) с соавт.

Список использованной и рекомендуемой литературы

1. Aft L.S. Fundamentals of industrial quality control. – Boca Raton, FL: CRC Press LLC, 1997.
2. Altman D.G., Bland J.M. Measurement in medicine: The analysis of method comparison studies // The Statistician, 1983, vol. 32, pp. 307–317.
3. Barlow R.E., Irony T.Z. Foundations of statistical quality control // Current issues in statistical inference: Essays in honor of D. Basu / Ed. by M. Ghosh, P.K. Pathak. – Hayward, CA: Institute of Mathematical Statistics, 1992, pp. 99–112.
4. Bergman B., Klefsjo B. Quality from customer needs to customer satisfaction. – London, UK: McGraw–Hill, 2002.
5. Bissell D. Statistical methods for SPC and TQM. – Boca Raton, FL: CRC Press LLC, 1994.
6. Bland J.M., Altman D.G. Agreement between methods of measurement with multiple observations per individual // Journal of Biopharmaceutical Statistics, July 2007, vol. 17, issue 4, pp. 571–582.
7. Bland J.M., Altman D.G. Comparing methods of measurement: why plotting difference against standard method is misleading // The Lancet, 1995, vol. 346, pp. 1085–1087.
8. Bland J.M., Altman D.G. Statistical methods for assessing agreement between two methods of clinical measurement // The Lancet, 8 February 1986, vol. 1, no. 8476, pp. 307–310.
9. Brazdzionyte J., Macas A. Bland–Altman analysis as an alternative approach for statistical evaluation of agreement between two methods for measuring hemodynamics during acute myocardial infarction // Medicina, 2007, vol. 43, no. 3, pp. 208–214.
10. Bruynesteyn K. eciding on progression of joint damage in paired films of individual patients: smallest detectable difference or change / K. Bruynesteyn, M. Boers, P. Kostense et al. // Annals of the Rheumatic Diseases, 2005, vol. 64, pp. 179–182.
11. Burr J.T. Elementary statistical quality control. – Boca Raton, FL: CRC Press LLC, 2004.
12. Das N., Bhattacharya A. A new non–parametric control chart for controlling variability // Quality Technology of Quantitative Management, 2008, vol.5, no. 4, pp. 351–361.
13. Del Castillo E. Statistical process adjustment for quality control. – New York, NY: John Wiley & Sons, 2002.
14. Deming W.E. Elementary principles of the statistical control of quality. – Tokyo: Nippon Kagaku Gijutsu Renmei, 1952.
15. Dewitte K. Application of the Bland–Altman plot for interpretation of method–comparison studies: A critical investigation of its practice / K. Dewitte, C. Fierens, D. Stockl et al. // Clinical Chemistry, 2002, vol. 48, pp. 799–801.
16. Ghosh S., Schucany W., Smith W.B. Statistics of quality. – Boca Raton, FL: CRC Press LLC,

1996.
17. Giacalone M. On the Shewhart's operative characteristic curve // Bulletin of the International Statistical Institute, 52nd Session, Proceedings, Tome LVIII, Finland 1999. Contributed Paper Meeting 73: Quality control and statistics in industry.
18. Hanneman S.K., Kleinpell R.M. Design, analysis, and interpretation of method–comparison studies // AACN Advanced Critical Care, April/June 2008, vol. 19, no. 2, pp. 223–234.
19. Health care criteria for performance Excel®lence. – Milwaukee, WI: American Society for Quality, 2005.
20. Hilson A. Bland–Altman plot // Radiology, 2004, vol. 231 pp. 604–605.
21. Hubbard M.R. Choosing a quality control system. – Boca Raton, FL: CRC Press LLC, 1998.
22. Ishikawa K. Guide to quality control. – Tokyo: Asian Productivity Organization, 1976.
23. Ishikawa K. What is total quality control?: The Japanese Way. – London, UK: Prentice Hall, 1985.
24. Jeya Chandra M. Statistical quality control. – Boca Raton, FL: CRC Press LLC, 2001.
25. Kelley W.D., Ratliff T.A., Nenadic C. Basic statistics for laboratories: A primer for laboratory workers. – New York, NY: John Wiley & Sons, 1991.
26. Kenkel J. A Primer on quality in the analytical laboratory. – Boca Raton, FL: CRC Press LLC, 1999.
27. Krishnaiah P.R. Handbook of statistics. Vol. 7. Quality control and reliability / Ed. by P.R. Krishnaiah, C.R. Rao. – New York, NY: Elsevier, 1988.
28. Kume H. Statistical methods for quality improvement. – Tokyo: AOTS Chosakai, 1985.
29. Ledolter J., Burrill C.W. Statistical quality control: Strategies and tools for continual improvement. – New York, NY: John Wiley & Sons, 1998.
30. Mantha S. Comparing methods of clinical measurement: Reporting standards for Bland and Altman analysis / S. Mantha, M.F. Roizen, L.A. Fleisher et al. // Anesthesia & Analgesia, 2000, vol. 90, pp. 593–602.
31. Mittag H.–J., Rinne H. Statistical methods of quality assurance. – Boca Raton, FL: CRC Press LLC, 1993.
32. Montgomery D.C. Introduction to statistical quality control. – New York, NY: John Wiley & Sons, 2005.
33. Montgomery D.C., Runger G.C. Applied statistics and probability for engineers. – New York, NY: John Wiley & Sons, 2003.
34. Neupert F.–G. Statistikbasierte prozessanalysestrategie bei VW zur gezielten prozessverbesserung. – Berlin: VDI–Veranstaltung, 2003.
35. Ryan T.P. Statistical methods for quality improvement. – New York, NY: John Wiley & Sons, 2000.
36. Shirland L.E. Statistical quality control with microcomputer applications. – New York, NY: John Wiley & Sons, 1993.
37. Stamatis D.H. Six sigma and beyond: Statistical process control, volume IV. – Boca Raton, FL: CRC Press LLC, 2002.
38. Stockl D. Interpreting method comparison studies by use of the Bland–Altman plot: Reflecting the importance of sample size by incorporating confidence limits and predefined error limits in the graphic / D. Stockl, D.R. Cabaleiro, K. Van Uytfanghe et al. // Clinical Chemistry, 2004, vol. 50, pp. 2216–2218.
39. Taguchi G., Chowdhury S., Wu Y. Taguchi's quality engineering handbook. – New York, NY: John Wiley & Sons, 2004.
40. Thompson J.R., Koronacki J. Statistical process control: The Deming paradigm and beyond. – Boca Raton, FL: CRC Press LLC, 2001.
41. Vardeman S.B., Jobe J.M. Statistical quality assurance methods for engineers. – New York,

NY: John Wiley & Sons, 1998.
42. Wadsworth H.M., Stephens K.S., Godfrey A.B. Modern methods for quality control and improvement. – New York, NY: John Wiley & Sons, 1986.
43. Андреев Я.Г., Поздеева Т.Е., Шишкова Н.В. Статистический контроль качества испытаний фольги медной электролитической и катанки медной с использованием контрольных карт // Аналитика и контроль, 2004, № 4, с. 387–390.
44. Барабанова О.А. Семь инструментов контроля качества / О.А. Барабанова, В.А. Васильев, С.А. Одиноков. – М.: Издательский центр «МАТИ» РГТУ им. К.Э. Циолковского, 2003.
45. Браунли К.А. Статистические исследования в производстве. – М.: Издательство иностранной литературы, 1949.
46. Гельфанд С.Ю., Дьяконова Э.В. Статистические методы контроля качества продукции в консервной и пищеконцентратной промышленности. – М.: Легкая и пищевая промышленность, 1984.
47. Гличев А.В. Основы управления качеством продукции. – М.: ГП–Редакция журнала Стандарты и качество, 2001.
48. Гличев А.В., Круглов М.И. Управление качеством продукции. – М.: Экономика, 1979.
49. Гнеденко Б.В. Математика и контроль качества продукции. – М.: ЛКИ, 2007.
50. Гнеденко Б.В. Математическая статистика и контроль качества. – М.: Знание, 1976.
51. Гнеденко Б.В., Беляев Ю.К., Соловьев А.Д. Математические методы в теории надежности. – М.: Наука, 1965.
52. Горелов А.С. Статистическое планирование контроля качества продукции / А.С. Горелов, Е.А. Саввина, Ю.Л. Маткин и др. – Тула: ГРИФ, 2003.
53. Григорович В.Г. Информационные методы в управлении качеством / В.Г. Григорович, Н.О. Козлова, В.В. Шильдин и др. – М.: РИА Стандарты и качество, 2001.
54. Дворкин В.И. Внутрилабораторный контроль качества химического анализа и компьютерная программа «QControl» // Партнеры и конкуренты, 2000, № 4, с. 30–39.
55. Дворкин В.И. Внутрилабораторный контроль качества химического анализа при наличии контрольного материала // Журнал аналитической химии, 2001, т. 56, № 7, с. 690–702.
56. Джонсон Н., Лион Ф. Статистика и планирование эксперимента в технике и науке. Методы обработки данных. – М.: Мир, 1980.
57. Джуран Д. Все о качестве: Зарубежный опыт. Выпуск 2. Высший уровень руководства и качество. – М., 1993.
58. Диков Э. Квалиметрия // Юный техник, 1970, № 2, с. 19.
59. Дубров А.М. Последовательный анализ в статистической обработке информации. – М.: Статистика, 1976.
60. Дюк В. Обработка данных на ПК в примерах. – Спб: Питер, 1997.
61. Ефимов В.В. Статистические методы в управлении качеством: Учебное пособие. – Ульяновск: Ульяновский государственный технический университет, 2003.
62. Ефимов В.В., Барт В.В. Статистические методы в управлении качеством продукции. М.: КноРус, 2006.
63. Исикава К. Японские методы управления качеством. – М.: Экономика, 1988.
64. Катеман Т., Пийпер Ф.В. Контроль качества химического анализа. – Челябинск: Металлургия, 1989.
65. Клячкин В.Н. Статистические методы в управлении качеством: Компьютерные технологии. М.: Финансы и статистика, 2007.
66. Кобзарь А.И. Прикладная математическая статистика. Для инженеров и научных работников. – М.: ФИЗМАТЛИТ, 2006.

67. Кордонский Х.Б. Применение теории вероятностей в инженерном деле. – М.: Издательство физико–математической литературы, 1963.
68. Корнеева Т.В. Толковый словарь по метрологии, измерительной технике и управлению качеством. – М.: Русский язык, 1990.
69. Коуден Д. Статистические методы контроля качества. – М.: Издательство физико–математической литературы, 1961.
70. Круглов М.Г. Менеджмент систем качества: Учебное пособие / М.Г. Круглов, С.К. Сергеев, В.А. Такташов и др. – М.: Издательство стандартов, 1997.
71. Кулаичев А.П. Методы и средства анализа данных в среде Windows®. STADIA. – М.: Информатика и компьютеры, 1999.
72. Кулаичев А.П. Методы и средства комплексного анализа данных. – М.: ИНФРА–М, 2006.
73. Кумэ Х. Статистические методы повышения качества / Под ред. Х. Кумэ. – М.: Финансы и статистика, 1990.
74. Курицин А.Н. Секреты эффективной работы: опыт США и Японии для предпринимателей и менеджеров. – М.: Издательство стандартов,1994.
75. Лапидус В.А. Статистический контроль качества продукции на основе принципа распределения приоритетов / В.А. Лапидус, М.И. Розно, А.В. Глазунов и др. – М: Финансы и статистика, 1991.
76. Леон Р. Управление качеством. Робастное проектирование. Метод Тагути / Р. Леон, А. Шумейкер, Р. Какар и др. – М.: Сейфи, 2002.
77. Логанина В.И. Статистические методы контроля и управления качеством продукции. – Ростов–на–Дону: Феникс, 2007.
78. Макино Т. Контроль качества с помощью персональных компьютеров / Т. Макино, М. Охаси, Х. Докэ и др. – М.: Машиностроение, 1991.
79. Мердок Дж. Контрольные карты. – М: Финансы и статистика, 1986.
80. Миттаг Х.–Й., Ринне Х. Статистические методы обеспечения качества. – М.: Машиностроение. 1995.
81. Монден Я. Как работает японское предприятие / Под ред. Я. Мондена, Р. Сибакавы, С. Такаянаги и др. – М.: Экономика, 1989.
82. Мхитарян В.С. Статистические методы в управлении качеством продукции. – М.: Финансы и статистика, 1982.
83. Нив Г.Р. Пространство доктора Деминга. – М.: РИА «Стандарты и качество», 2003.
84. Никитин В.А. Управление качеством на базе стандартов ИСО 9000:2000. – СПб.: Питер, 2002.
85. Нойман Э., Хойсингтон С.Х. Качество на уровне Шесть Сигма. – Днепропетровск: Баланс–Клуб, 2004.
86. Ноулер Л. Статистические методы контроля качества продукции / Л. Ноулер, Дж. Хауэлл, Б. Голд и др. – М.: Издательство стандартов, 1989.
87. Рабинович П.М. Резервы предприятия и статистика. – М.: Статистика, 1967.
88. Розно М.И. Применение прикладных статистических методов при производстве продукции (для специалистов по управлению качеством) / М.И. Розно и др. – Нижний Новгород: СМЦ «Приоритет»,1997.
89. Сенченко И.Н. Статистические методы в спиртоводочном производстве / И.Н. Сенченко, Ю.Л. Маткин, А.С. Горелов и др. – Тула: ТулГУ, 2001.
90. Система качества. Сборник нормативно–методических документов. – М.: Издательство стандартов,1992.
91. Сиськов В.И. Статистическое измерение качества продукции. – М.: Статистика, 1966.
92. Сиськов В.И. Экономико–статистическое исследование качества продукции. – М.:

Статистика, 1971.
93. Спицнадель В.Н. Системы качества (в соответствии с международными стандартами ISO семейства 9000). – СПб.: Издательский дом «Бизнес–пресса», 2000.
94. Технический отчет ISO/TR 10017:2003. Руководство по статистическим методам применительно к ISO 9001:2000. – М.: ВНИИКИ, 2004.
95. Уилер Д., Чамберс Д. Статистическое управление процессами: Оптимизация бизнеса с использованием контрольных карт Шухарта. – М.: Альпина Паблишер, 2009.
96. Фейгенбаум А.В. Контроль качества продукции. – М.: Экономика, 1986
97. Фомин В.Н. Квалиметрия. Управление качеством. Сертификация. – М.: Ось–89, 2005.
98. Хэнсен Б. Контроль качества. – М.: Прогресс, 1968.
99. Шторм Р. Теория вероятностей. Математическая статистика. Статистический контроль качества. – М.: Мир, 1970.

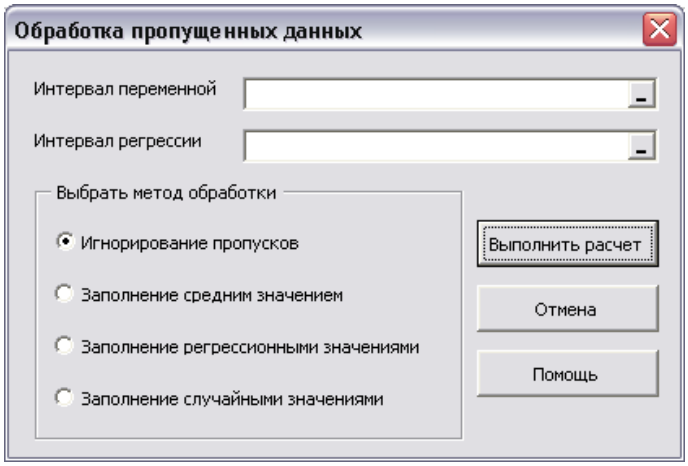
Глава 20. Обработка пропущенных данных

20.1. Введение

Программное обеспечение обеспечивает обработку пропущенных значений различными методами. Перед применением метода необходимо убедиться, что он соответствует шкале измерения исходных данных (признаков).

20.2. Работа с программным обеспечением

Выберите из меню программы пункт AtteStat | Обработка пропущенных данных. На экране появится диалоговое окно, изображенное на рисунке:



Затем проделайте следующие шаги:

- Выберите или введите интервал переменной. Исходным для расчета является эмпирический ряд в виде столбца.
- Выберите или введите интервал регрессионной переменной, записанной в виде столбца. Данный интервал необходим только, если выбран метод «Заполнение регрессионными значениями».
- Выберите метод обработки пропущенных значений. Для расчета может быть выбран только один из предлагаемых методов.
- Нажмите кнопку «Выполнить расчет».

В результате выполнения расчета пропущенные значения указанного интервала ячеек

электронной таблицы значения будут обработаны в соответствии с выбранным методом. Обработанные ячейки будут заполнены значениями, для удобства пользователя, с целью обратить его внимание, выделенными визуально шрифтом синего цвета полужирного начертания.
При ошибках, вызванных неверными действиями пользователя, выдаются сообщения об ошибках.

20.2.1. Сообщения об ошибках

При ошибках ввода и во время выполнения программы могут выдаваться диагностические сообщения следующих типов, как показано ниже:

Ошибка	Комментарий
Не определен интервал переменной.	Вы не выбрали или неверно ввели входной интервал. Лучшим способом избежать ошибки является не ввод, а выделение интервала стандартным образом, т. е. протаскиванием курсора.
Не определен интервал регрессии.	Вы не выбрали или неверно ввели интервал регрессии, требуемый для расчета методом заполнения регрессионными значениями. Лучшим способом избежать ошибки является не ввод, а выделение интервала стандартным образом, т. е. протаскиванием курсора.
Мало данных для выбранного метода.	Если выбран метод заполнения регрессионными значениями, для расчета следует выбрать или ввести входной интервал, а также интервал регрессии размером более одной числовой ячейки (иначе нельзя построить регрессию). Если выбран метод заполнения случайными значениями, необходим входной интервал, содержащий не менее двух числовых ячеек (иначе нельзя вычислить выборочную оценку дисперсии).
Нет числовых данных для расчета.	Для расчета необходимо выбрать интервал, содержащий, по меньшей мере, одну ячейку с числовым значением. Некоторые применяемые методы требуют более одной числовой ячейки.

20.3. Теоретическое обоснование

Под статистическими данными понимают любую систему данных: числовую и нечисловую информацию, извлекаемую из результатов выборочных обследований, выборки из любых генеральных совокупностей, результаты измерений и т. п. Однако в практической деятельности возникают ситуации, когда часть статистических данных по различным объективным или субъективным причинам оказывается утраченной. Существуют ряд методов, позволяющих специальным образом обработать пропущенные значения и вернуть, таким образом, утраченные данные в последующие процедуры анализа. Применение методов обработки пропущенных данных описано в литературе.
Программное обеспечение обеспечивает обработку пропущенных значений методами:

- Игнорирование пропусков.
- Заполнение средним значением.
- Заполнение регрессионными значениями.
- Заполнение случайными значениями.

Мы не рассматриваем причины возникновения пропущенных данных, однако хотелось бы думать, что они обусловлены обстоятельствами, не связанными с физикой явления. Перед

применением методов обработки пропущенных данных исследователь должен точно знать, что данные на местах пропусков обязательно должны быть, но отсутствуют, скажем, из–за невнимательности лаборанта, отказа измерительного прибора, неявки пациента на очередное обследование или по другим форс–мажорным причинам. К таким причинам не относится досрочное выбытие из эксперимента объекта, если оно вызвано условиями эксперимента. Обработку таких данных производят с помощью методов анализа цензурированных выборок. Напомним, что цензурированными называются усеченные по условиям эксперимента выборки. Например, при испытании изделий часть их может отказать, а часть не отказать в течение периода испытаний.
Подход, предлагаемый некоторыми авторами и заключающийся в совместном анализе матрицы исходных данных и матрицы пропусков, вряд ли правомерен, т.к. предполагает взаимосвязь изучаемого процесса с причиной возникновения пропусков. Мы же постулируем, что эти явления совершенно независимы.
Данное программное обеспечение поможет подкрепить теоретически интуицию исследователя.

20.3.1. Игнорирование пропусков

Самым простым и наиболее понятным способом обработки пропущенных значений является их игнорирование. На месте пропущенных значений, если не рассматривать физическую картину изучаемого явления, а именно так и поступает наука статистика, и не делать никаких дополнительных предположений, могут стоять любые значения.
Метод игнорирования пропусков в данной реализации программного обеспечения работает только с одним столбцом данных, осуществляя сдвиг ячеек вверх и заполнение, таким образом, всех пропусков. В результате обработки происходит уменьшение численности выборки на число пропущенных ячеек, к которым относятся также и все нечисловые значения.
Метод рекомендуется применять для малых выборок с малым числом пропусков, причем выборки могут принадлежать любой шкале измерения.

20.3.2. Заполнение средним значением

Заполнение пропущенных значений некоторыми допустимыми значениями является распространенным способом их восстановления. В качестве допустимого значения обычно выбирается выборочное среднее значение, вычисляемое по формуле

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i,$$

где x_i, $i = 1,2,...,n$ – присутствующие варианты выборки,
n – разность между численностью выборки и числом пропущенных значений.
Метод удобен тем, что в результате его применения важнейшая статистическая мера положения, а именно среднее значение, в выборке с заполненными пропусками не изменяется по сравнению со средним значением, вычисленным для выборки с пропущенными значениями (исходной выборки). Однако метод приводит к заниженной выборочной оценке дисперсии с заполненными значениями относительно дисперсии исходной выборки. Другим недостатком является искажение эмпирического распределения выборки независимо от типа эмпирического распределения исходной выборки.
Метод заполнения средним значением рекомендуется применять для больших выборок с малым числом пропусков. Выборка должна принадлежать количественной шкале. Данный метод представляет собой частный случай заполнения регрессионными значениями.

20.3.3. Заполнение регрессионными значениями

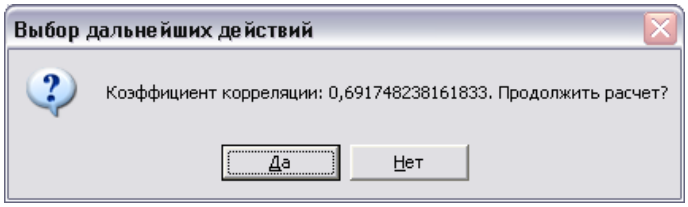
В практических наблюдениях бывают случаи, когда изменению одного признака соответствует изменение величины другого признака в среднем. Такой вид соотношений называется корреляционной зависимостью, или корреляцией. Считается, что исследование взаимной зависимости приводит к теории корреляции, тогда как изучение зависимости ведет к теории регрессии. Регрессионная модель позволяет выразить значения зависимой (регрессионной) переменной от независимой переменной без исследования функциональной (причинной) связи. Наличие статистической корреляционной зависимости не влечет зависимости причинной. Исследование причинной зависимости – предмет не статистики, а математического моделирования.
Заполнение регрессионными значениями базируется на идее заполнения пропусков, основываясь на информации о связи данной выборки с другой выборкой. При коэффициенте корреляции, по модулю близком к единице, можно предположить существующую тесную регрессионную зависимость между независимой *x* и регрессионной *y* переменными.
Коэффициент корреляции Пирсона вычисляется по формуле

$$r = \frac{Cov(x, y)}{\sqrt{Cov(x, x)Cov(y, y)}},$$

где *Cov*(.,.) – выборочная ковариация, вычисляемая по формуле

$$Cov(x, y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})\sum_{i=1}^{n}(y_i - \bar{y}),$$

x_i, $i = 1,2,...,n$ – присутствующие варианты выборки (независимая переменная),
y_i, $i = 1,2,...,n$ – присутствующие варианты выборки (зависимая переменная),
n – разность между численностью выборки и числом пропущенных значений.
После вычисления коэффициента корреляции пользователю предлагается принять решение, выполнять дальнейшие вычисления или нет, типа того, как показано на рисунке.



Если принято решение продолжить, производится вычисление линейной регрессии. В простейшем случае регрессионную зависимость можно представить полиномом 1–й степени
$y = ax + b$,
где *x* – независимая переменная,
y – зависимая (регрессионная) переменная,
a – коэффициент при первой степени *x*,
b – свободный член – коэффициент при нулевой степени *x*.
Иначе данная зависимость называется линейной, т. к. представляет собой уравнение прямой линии на плоскости. Коэффициенты полинома вычисляются по формулам

$$a = \frac{\sum_{i=1}^{n} x_i y_i - \bar{x}\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i^2 - \bar{x}\sum_{i=1}^{n} x_i},$$

$$b = \bar{y} - a\bar{x}.$$

Если установлена линейная зависимость между переменными, можно попытаться восстановить отсутствующие значения независимой переменной по регрессионной, решив обратную задачу. В данном случае формула решения обратной задачи принимает вид:

$$x = \frac{y - b}{a}.$$

Данные вычисления возможны, конечно, если в парах значений регрессионные значения известны, а независимые утрачены. Если утраченными оказываются оба значения, заполнение пропусков производится средним значением. Выборка должна принадлежать количественной шкале в случае непрерывных признаков, однако можно предположить, что представленный алгоритм может быть развит и для признаков других типов, в том числе для смешанных признаков.

20.3.4. Заполнение случайными значениями

Метод заполнения случайными значениями, вопреки бытовой трактовке наименования, на самом деле производит наиболее корректное заполнение пропусков из всех представленных методов в смысле сохранения несмещенности статистических параметров выборки. Данный метод теоретически обоснован гораздо лучше других методов, представленных данным программным обеспечением. В основе метода, построенного на квазирандомизационном подходе, лежит предположение, что меры положения (средние значения) и меры разброса (средние квадратические отклонения) в присутствующей и в пропущенной частях выборки равны. Общий подход к его реализации заключается в определении типа теоретического распределения эмпирической выборки и последующей случайной генерации отсутствующей части с тем же теоретическим распределением.

Нами решается более частная задача: предполагается, что распределение является нормальным. Если распределение относится к другому стандартному типу, то приведенные ниже выводы должны быть с учетом этого скорректированы. Метод предлагает заполнять пропущенные значения случайными значениями, имеющими нормальное распределение с параметрами, вычисленными по исходной эмпирической выборке.

Генерация выборки, распределенной по нормальному закону $N(\bar{x}, \sigma^2)$, производится по выборке, распределенной по стандартному нормальному закону $N(0,1)$, с помощью формулы

$$y_j = \bar{x} + \sigma x_j,\ j = 1,2,\ldots,m,$$

где $\bar{x}$ – выборочное среднее значение присутствующей части выборки,

σ – выборочное среднее квадратичное отклонение, квадратный корень из дисперсии присутствующей части выборки,

x_j, $j = 1,2,...,m$ – выборка, сгенерированная по стандартному нормальному закону $N(0,1)$,

m – число пропущенных значений.

Выборочное среднее определяется аналогично методу заполнения средним значением. Выборочная дисперсия вычисляется по формуле

$$\sigma^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2,$$

где x_i, $i = 1,2,...,n$ – присутствующие варианты выборки,

n – разность между численностью выборки и числом пропущенных значений.

В данной формуле в суммировании участвуют только присутствующие варианты выборки. Выборка, сгенерированная по стандартному нормальному закону, получена из выборки с равномерным распределением в интервале (0,1), путем подстановки ее вариант в нормальный интеграл (обратную функцию стандартного нормального распределения), представляющую собой решение обратной задачи

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-y^2/2} dy,$$

т. е. по известному Φ(x) определяется x.
Рассмотренный метод рекомендуется применять для больших выборок с большим числом пропусков, причем выборка должна принадлежать количественной шкале для случая непрерывных признаков.

Список использованной и рекомендуемой литературы

1. Armitage P., Berry G., Matthews J.N.S. Statistical methods in medical research. – Oxford: Blackwell Science, 2001.
2. Balder C. Detection of missing information and identification of its mechanism in a household survey / C. Balder, S. Alsina, N. Arnesi et al. // Bulletin of the International Statistical Institute, 52nd Session, Proceedings, Tome LVIII, Finland 1999. Contributed Paper Meeting 50: Missing data and non–response.
3. Royston P. Multiple imputation of missing values: Updates // The Stata Journal, 2005, vol. 5, no. 2, pp. 188–201.
4. Stoica P., Xu L., Li J. A new type of parameter estimation algorithm for missing data problems // Statistics & Probability Letters, 1 December 2005, vol. 75, issue 3, pp. 219–229.
5. Van Belle G. Biostatistics: A methodology for the health sciences // G. van Belle, L.D. Fisher, P.J. Heagerty et al. – New York, NY: John Wiley & Sons, 2003.
6. Yang Y. Multiple imputation for missing values: Concepts and new development // SUGI Proceedings, 2000.
7. Гайдышев И. Анализ и обработка данных: специальный справочник. – СПб: Питер, 2001.
8. Дюк В., Самойленко А. Data mining: учебный курс. – СПб: Питер, 2001.
9. Злоба Е., Яцкив И. Статистические методы восстановления пропущенных данных // Computer Modelling & New Technologies, 2002, vol. 6, no. 1, pp. 51–61.
10. Кулаичев А.П. Методы и средства анализа данных в среде Windows®. STADIA. – М.: Информатика и компьютеры, 1999.
11. Литтл Р.Дж.А., Рубин Д.Б. Статистический анализ данных с пропусками. – М. Финансы и статистика, 1991.
12. Мюллер П., Нойман П., Шторм Р. Таблицы по математической статистике. – М.: Финансы и статистика, 1982.
13. Сборник научных программ на Фортране. Выпуск 1. Статистика. – М.: Статистика, 1974.
14. Теннант–Смит Дж. Бейсик для статистиков. – М.: Мир, 1988.
15. Эфрон Б. Нетрадиционные методы многомерного статистического анализа. – М.: Финансы и статистика, 1988.

Глава 21. Обработка выбросов

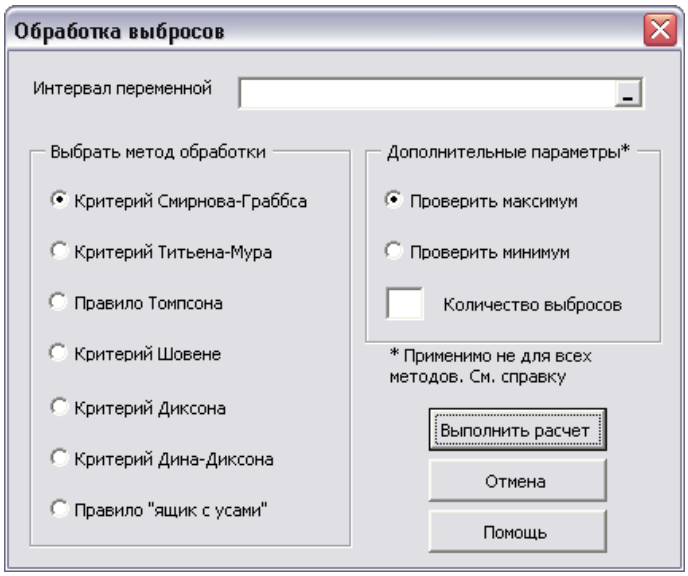
21.1. Введение

Программное обеспечение обеспечивает обработку выбросов. Выбросом называют резко (экстремально) выделяющееся наблюдение. Предположительно данное наблюдение следует исключать из анализа. Хотя возможна ситуация, когда данное значение действительно наблюдалось в эксперименте и следует найти причину его появления.
Перед применением любого метода анализа данных необходимо убедиться, что он

соответствует шкале измерения исходных данных (признаков). В случае применения представленных здесь методов анализа признаки должны принадлежать количественной шкале.

21.2. Работа с программным обеспечением

Выберите из меню программы пункт AtteStat | Обработка выбросов. На экране появится диалоговое окно, изображенное на рисунке:



Затем проделайте следующие шаги:

- Выберите или введите интервал переменной. Исходным для расчета является эмпирический ряд в виде столбца, строки или прямоугольной области электронной таблицы.
- Выберите метод обработки выбросов. Для расчета может быть выбран только один из предлагаемых методов.
- Выберите дополнительные параметры: Проверить максимум или Проверить минимум (действительно только для критериев Смирнова–Граббса и Титьена–Мура), а также укажитите количество выбросов (действительно только для критерия Титьена–Мура).
- Нажмите кнопку «Выполнить расчет».

Программа отметит все ячейки электронной таблицы, предположительно содержащие выбросы, путем установки красного цвета и жирного начертания шрифта данной ячейки. Программа не исключает выбросы, а только обращает на них внимание пользователя. Далее пользователю предлагается самому оценить, являются ли данные значения выбросами и либо игнорировать выводы программы и объяснить данные значения, либо устранить их. При ошибках, вызванных неверными действиями пользователя, выдаются сообщения об ошибках.

21.2.1. Сообщения об ошибках

При ошибках ввода и во время выполнения программы могут выдаваться диагностические сообщения следующих типов.

Ошибка	Комментарий
Не определен интервал переменной.	Вы не выбрали или неверно ввели интервал эмпирической выборки. Лучшим способом избежать ошибки является не ввод, а выделение интервала стандартным образом, т. е. протаскиванием курсора.
Пустая или нечисловая ячейка.	Для расчета необходимо выбрать интервал, содержащий только ячейки с числовым значением. Пропуски также недопустимы.
Мало данных.	Интервал исходных данных должен содержать, по меньшей мере, 3 числовых ячейки. При меньшем объеме данных представленные методы не смогут выполнить их корректную обработку.
Введите число выбросов от 1 до 10	Для метода Титьена–Мура необходимо ввести число выбросов, которые предполагается обработать. Данное число должно быть заключено в интервале от 1 до 10.
Мало данных для расчета.	Интервал исходных данных содержит слишком много значений для выбранного метода анализа. Укажите интервал объема, допустимого для данного метода, или выберите другой метод анализа.

21.3. Теоретическое обоснование

По словам профессора Клиланд (C. Cleland) из университета Колорадо, «Сила науки – в экспериментальном подходе. История развития науки свидетельствует, что в конечном итоге именно изучение аномалий приводило к потрясающим открытиям и смене существующих научных парадигм». Однако в практической деятельности иногда возникают ситуации, когда экспериментальные статистические данные по объективным или субъективным причинам оказываются засоренными резко выделяющимися, аномальными наблюдениями (выбросами). Выбросы обычно трактуются как грубые ошибки измерений, возникающие в результате просчета, неправильного чтения показаний прибора и т.п. В данном программном обеспечении решается именно такая проблема, хотя представляется, что аномальные наблюдения всё–таки должны получить объяснения. Если аномальные наблюдения не укладываются в принятую модель, логично предположить, что модель должна быть пересмотрена.

Существуют ряд методов, помимо непосредственной «ручной» проверки результатов наблюдений, позволяющих специальным образом обработать выбросы. Данные методы основаны на критериях исключения мимального (максимального) наблюдения и подробно описаны в литературе. Данное программное обеспечение обеспечивает обработку выбросов методами:

- Критерий Смирнова–Граббса.
- Критерий Титьена–Мура.
- Правило Томпсона.
- Критерий Диксона.
- Критерий Дина–Диксона.
- Критерий Шовене.
- Правило «ящик с усами».

Критерии Диксона и Дина–Диксона разработаны специально для обработки малых выборок, численностью от 3 до 30. Для локализации многочисленных выбросов в больших выборках может применяться критерий Уолша, не представленный в программе.

Критерий Кокрена, также применяющийся для обработки выбросов, представлен в главе

«Дисперсионный анализ». Для обработки выбросов в многомерных данных могут применяться многомерные методы, например, «Факторный анализ» и «Кластерный анализ». Основные предположения при разработке представленных методов заключаются в следующем:

- Исходные данные имеют нормальное распределение.
- Рассматриваем случай, когда основные статистические параметры совокупности (мера положения – среднее значение – и мера разброса – дисперсия) неизвестны и вычисляются по выборке.
- В расчетах ограничиваемся, как это принято при обработке выбросов, стандартным уровнем значимости, равным 0,05.

О выборках, загрязненных выбросами, см. справочник Родионова с соавт. Об обработке выбросов в многомерных выборках см. Афифи с соавт. В последнем случае возможен «ручной расчет» с помощью методов, представленных в программном обеспечении «Матричная и линейная алгебра».

21.3.1. Критерий Смирнова–Граббса

Критерий Смирнова–Граббса (критерий разногласий, Smirnov–Grubbs) предназначен для исключения одного выброса – резко выделяющегося максимального или минимального наблюдения из нормально распределенной выборки.

Статистика критерия Смирнова–Граббса основана, соответственно, на величине

$$T_N = \frac{\max\limits_i x_i - \bar{x}}{s} \quad \text{или} \quad T_N = \frac{\bar{x} - \min\limits_i x_i}{s},$$

где $x_1, x_2, ..., x_N$ – результаты N наблюдений,

$\bar{x}$ – выборочное среднее значение,

s – выборочное стандартное отклонение, корень квадратный из выборочной дисперсии.

Выборочное среднее значение рассчитывается по формуле

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i,$$

а выборочная дисперсия

$$s^2 = \frac{1}{N}\sum_{i=1}^{N} \left(x_i - \bar{x}\right)^2.$$

Величина статистики критерия сравнивается с критическим значением, точное распределение которого дается формулой (см. руководство NIST/SEMATECH)

$$G_N = \frac{N-1}{\sqrt{N}} \sqrt{\frac{t^2_{a/N,N-2}}{t^2_{a/N,N-2} + N - 2}},$$

где $t_{\alpha/N,N-2}$ – значение обратной функции t–распределения с параметрами α / N и $N - 2$,

α – заданный уровень значимости, обычно 0,05.

При величине статистики, большей критического значения, наблюдение исключается.

Находит применение еще один метод исключения максимального или минимального наблюдения – критерий Граббса (Груббса), связанный с представленным здесь критерием простой формулой

$$G_N = 1 - \frac{1}{N-1} T_N^2,$$

поэтому дающий точно такие же результаты своего применения.

Критерий Смирнова–Граббса, как и критерий Граббса, не годится для исключения

нескольких ($k > 1$) выбросов из–за т. н. маскирующего эффекта. Маскирующим эффектом выброса, при числе выбросов более 1, называют такое смещение параметров выборки, которое не позволяет методу обнаружить все выбросы. Поэтому выбросы уже не могут в рамках данного метода рассматриваться как нетипичные. Для исключения нескольких выбросов рекомендуется применять критерий Титьена–Мура.

См. Мюллера с соавт., Мотульски (Motulsky) с соавт.

21.3.2. Критерий Титьена–Мура

Критерий Титьена–Мура (Tietjen–Moore) является обобщением критерия Граббса на случай нескольких выбросов. Статистика критерия исключения k наибольших или наименьших аномальных значений основана, соответственно, на величине

$$L_k = \frac{\sum_{i=1}^{N-k}\left(x_i - \bar{x}_k\right)^2}{\sum_{i=1}^{N}\left(x_i - \bar{x}\right)^2} \quad \text{или} \quad \tilde{L}_k = \frac{\sum_{i=k+1}^{N}\left(x_i - \hat{x}_k\right)^2}{\sum_{i=1}^{N}\left(x_i - \bar{x}\right)^2},$$

где $x_1,x_2,...,x_N$ – упорядоченные по возрастанию результаты N наблюдений,
$\bar{x}$ – выборочное среднее значение,
Выборочное среднее значение (для всей выборки) рассчитывается по формуле

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i,$$

а средние значения после отбрасывания k наибольших или наименьших значений рассчитываются, соответственно, по формулам

$$\bar{x}_k = \frac{1}{N-k}\sum_{i=1}^{N-k} x_i \quad \text{или} \quad \hat{x}_k = \frac{1}{N-k}\sum_{i=k+1}^{N} x_i.$$

Величина статистики критерия сравнивается с табличным значением. При величине статистики, меньшей табличного значения, наблюдения исключаются.
Критерий Титьена–Мура позволяет бороться с маскирующим эффектом. Маскирующим эффектом выброса, при числе выбросов более 1, называют такое смещение параметров выборки, которое не позволяет методу обнаружить все выбросы. Поэтому выбросы уже не могут в рамках данного метода рассматриваться как нетипичные.

Описание критерия см. у Айвазяна с соавт.

21.3.3. Правило Томпсона

В правиле Томпсона (критерии Рошера) для исключения выбросов используется статистика

$$t_i = \frac{\left|x_i - \bar{x}\right|}{s},\ i = 1,2,...,N,$$

где $x_1,x_2,...,x_N$ – результаты N наблюдений,
$\bar{x}$ – выборочное среднее значение,
s – выборочное стандартное отклонение, корень квадратный из выборочной дисперсии.
Выборочное среднее значение рассчитывается по формуле

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i,$$

а выборочная дисперсия – по формуле

$$s^2 = \frac{1}{N}\sum_{i=1}^{N}\left(x_i - \bar{x}\right)^2.$$

По данным некоторых источников, при вычислении статистики может применяться несмещенная оценка дисперсии

$$s^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(x_i - \bar{x}\right)^2.$$

Величина статистики критерия сравнивается с критическим значением, точное распределение которого дается формулой

$$T = \sqrt{\frac{(N-1)t_{1-a/2,N-2}^2}{t_{1-a/2,N-2}^2 + N - 2}},$$

где $t_{\alpha/N,N-2}$ – значение обратной функции t–распределения с параметрами $1 - \alpha / 2$ и $N - 2$, α – заданный уровень значимости, обычно 0,05.
При величине статистики, большей критического значения, наблюдение исключается. Процедура повторяется для каждого наблюдения.

См. Мюллера с соавт.

21.3.4. Критерий Диксона

Критерий Диксона (критерий экстремальных значений) применяется для исключения одного максимального или минимального резко выделяющегося наблюдения в выборке численностью от 3 до 30. В критерии Диксона для исключения выбросов используются статистики

$$t_1 = \frac{z_N - z_{N-1}}{z_N - z_1},$$

$$t_2 = \frac{z_N - z_{N-1}}{z_N - z_2},$$

$$t_3 = \frac{z_N - z_{N-2}}{z_N - z_1},$$

где z_i, $i = 1,2,...,N$ – упорядоченные по возрастанию (при тестировании максимального значения) или по убыванию (при тестировании минимального значения) N наблюдений.
Если хотя бы одна из статистик превышает соответствующее ей критическое значение на заданном уровне значимости (обычно 0,05), наблюдение, соответствующее z_N, исключается. Таблицы критических значений статистик критерия получены методом компьютерного моделирования. Для расчетов таблицы очень хорошо аппроксимированы гиперболами с помощью методов главы «Регрессионный анализ».

См. результаты Мак–Бейна (McBane).

21.3.5. Критерий Дина–Диксона

Критерий Дина–Диксона (Q–критерий Дина и Диксона) применяется для исключения одного максимального или минимального резко выделяющегося наблюдения в выборке численностью от 3 до 30. В критерии для исключения выбросов используется статистика

$$Q = \frac{|z_1 - z_2|}{|z_1 - z_N|},$$

где z_i, i = 1,2,...,N – упорядоченные по убыванию (при тестировании максимального значения) или по возрастанию (при тестировании минимального значения) N наблюдений.
Если значение статистики превышает критическое значение, совпадающее с критическим значением статистики критерия Диксона t_1 на заданном уровне значимости (обычно 0,05), аномальное наблюдение, соответствующее z_1, исключается.

21.3.6. Критерий Шовене

Критерий Шовене предназначен для исключения одного максимального или минимального аномального наблюдения. Статистика критерия основана на величине

$$S_N = N\left\{1 - I\left[\frac{\max_i x_i - \bar{x}}{s}\right]\right\},$$

где x_1,x_2,...,x_N – результаты N наблюдений,
$\bar{x}$ – выборочное среднее значение,
s – выборочное стандартное отклонение, корень квадратный из выборочной дисперсии,
I[.] – нормальный интеграл.
Выборочное среднее значение рассчитывается по формуле

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i,$$

а выборочная дисперсия – по формуле

$$s^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2.$$

Нормальный интеграл определяется формулой

$$I(x) = \frac{1}{\sqrt{2\pi}}\int_{-x}^{x} e^{-t^2/2}dt$$

и практически, с учетом симметрии функции плотности нормального распределения, вычисляется через функцию стандартного нормального распределения как
$I(x) = 2[\Phi(x) - 0{,}5]$.
Величина статистики критерия сравнивается со значением 0,5. При величине статистики, меньшей 0,5, наблюдение исключается.
Некоторыми авторами рекомендуется не применять критерий второй раз с использованием пересчитанных заново (после исключения одного аномального наблюдения) значений среднего и дисперсии.

Описание см. в монографии Тейлора.

21.3.7. Правило «ящик с усами»

Правило «ящик с усами» получило название от типа соответствующего графика, используемого для наглядного представления разброса эмпирических данных с нанесенными значениями медианы и квартилей.
Порядок вычисления следующий:

- Определяются выборочные значения межквартильного размаха f и медианы μ

(подробнее о данных показателях см. главу «Описательная статистика»).

- Выборочные значения, меньшие μ – 1,5*f* и большие μ + 1,5*f*, называются мягкими (подозрительными) выбросами.
- Выборочные значения, меньшие μ – 3*f* и большие μ + 3*f*, называются экстремальными выбросами и должны быть исключены.

Критерий удобен для автоматической идентификации любого числа экстремально малых и больших значений выборки. Критерий очень популярен, однако его рекомендуется применять только в случае, если численность выборки велика.

21.3.8. Критерий Кокрена

Критерий *G* Кокрена (статистика Кокрена) используется для проверки нулевой гипотезы о равенстве дисперсий нормальных генеральных совокупностей по независимым выборкам с одинаковыми численностями. Вычисление статистики критерия производится по формуле

$$G = \frac{\max\limits_{1 \le i \le k} s_i^2}{\sum\limits_{i=1}^{k} s_i^2},$$

где $s_i^2, i = 1,2,...,k$ – выборочные дисперсии совокупностей,
k – число столбцов (выборок).
Установлено, что *P*–значение модифицированной статистики

$$G' = \frac{G(k-1)}{1-G}$$

может быть вычислено как

$$p = k \cdot F_{(n-1),(n-1)(k-1)}(G'),$$

где $F_{..}(.)$ – функция *F*–распределения,
n – численность каждой совокупности.
Согласно ГОСТ Р ИСО 5725–2–2002:

- Если значение тестовой статистики меньше (или равно) 5%–го критического значения, тестируемую позицию признают корректной.
- Если значение тестовой статистики больше 5%–го критического значения и меньше (или равно) 1%–го значения, тестируемую позицию называют квазивыбросом и отмечают одной звездочкой.
- Если значение тестовой статистики больше 1%–го критического значения, тестируемую позицию называют статистическим выбросом и отмечают двумя звездочками.

Метод реализован в главе «Дисперсионный анализ».

Описание критерия и примеры см. в монографиях Мюллера с соавт., Налимова, Siegel с соавт.

Список использованной и рекомендуемой литературы

1. Aggarwal C.C., Yu P.S. Outlier detection for high dimensional data // ACM SIGMOD Record, June 2001, vol. 30, no. 2, pp. 37–46.
2. Anscombe F.J. Rejection of outliers // Technometrics, 1960, no.2, pp.123–147.
3. Bar T. Kinetic outlier detection (KOD) in real–time PCR / T. Bar, A. Stahlberg, A. Muszta et al. // Nucleic Acids Research, 2003, vol. 31, no. 17, p. e105.
4. Barnett V., Lewis T. Outliers in statistical data. – Chichester, West Sussex, UK: John Wiley &

Sons, 1994.
5. Bates D. To reject, or not to reject… // PanVera Postings, Fall 1998.
6. Battaglia F., Orfei L. Outlier detection and estimation in nonlinear time series // Journal of Time Series Analysis, January 2005, vol. 26, no. 1, p. 107.
7. Becker C. Performance criteria for multivariate outlier identification procedures // Bulletin of the International Statistical Institute, 52nd Session, Proceedings, Tome LVIII, Finland 1999. Contributed Paper Meeting 40: Outliers.
8. Berthouex P.M., Brown L.C. Statistics for environmental engineers. – London: CRC Press, 1994.
9. Cook R.D. Detection of influential observations in linear regression // Technometrics 1977, vol. 19, pp. 15–18.
10. Cook R.D., Weisberg S. Residuals and Influence in Regression. – New York, NY: Chapman & Hall, 1982.
11. Correa J.C., Lopez V.I. A graphical method for detecting multivariate outliers // InterStat (Statistics on the Internet), February 2005.
12. Crawford K.D., Vasicek D.J., Wainwright R.L. Detecting multiple outliers in regression data using genetic algorithms // Proceedings of the 1995 ACM symposium on Applied computing, Nashville, Tennessee, USA, February 26–28, 1995, pp. 351–356.
13. Dean R.B., Dixon W.J. Simplified statistics for small numbers of observations // Journal of Analytical Chemistry, 1951, vol. 23, pp. 636–638.
14. Devore J.L. Probability and statistics. – Boston, MA: Duxbury Press, 1995.
15. Dixon W.J. Analysis of extreme values // Annals of Mathematical Statistics, 1950, vol. 21, pp. 488–506.
16. Dixon W.J. Processing data for outliers // Biometrics, 1953, vol. IX, pp. 74–89.
17. Dixon W.J. Ratios involving extreme values // Annals of Mathematical Statistics, 1951, vol. 22, pp. 68–78.
18. Evans V.P. Strategies for detecting outliers in regression analysis: An introductory primer // Advances in social science methodology, vol. 5, pp. 213–233 / Ed. by B. Thompson. – Stamford: JAI Press, 1999.
19. Ferguson T.S. On the rejection of outliers // Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics, June 20–July 30, 1960 / Ed. by J. Neyman. – Berkeley, CA: University of California Press, 1961, pp. 253–287.
20. Gibbons R.D. Statistical methods for groundwater monitoring. – New York, NY: John Wiley & Sons, 1994.
21. Gilbert R.O. Statistical methods for environmental pollution monitoring. – New York, NY: Van Nostrand Reinhold, 1987.
22. Grubbs F. Procedures for detecting outlying observations in samples // Technometrics, 1969, vol. 11, no. 1, pp. 1–21.
23. Guidance for data quality assessment. Practical methods for data analysis. EPA QA/G–9. – Washington, DC: United States Environmental Protection Agency, 2000.
24. Hamilton L.C. Regressions with graphics: A second course in applied statistics. – Monterey: Brooks/Cole, 1992.
25. Hawkins D.M. Identification of outliers. – London: Chapman & Hall, 1980.
26. Hu Y., Smeryers–Verbeke J., Massart D.L. Outlier detection in calibration // Chemometrics and Intelligent Laboratory Systems, 1990, vol. 9, pp. 31–44.
27. Huck S.W. Reading statistics and research. – New York, NY: Longman, 2000.
28. Hwang T.–Y., Hu C.–Y. On the joint distribution of Grubbs' statistics // Annals of the Institute of Statistical Mathematics, 1994, vol. 46, no. 4, pages 769–775

29. Iglewicz B., Hoaglin D.C. How to detect and handle outliers // ASQC Basic References in Quality Control, vol. 16. – Milwaukee, WI: American Society for Quality Control, 1993.
30. Iglewicz G., Hoaglin D.C. How to detect and handle outliers. – Milwaukee, WI: ASQC Quality Press, 1993.
31. Jarrell M.G. A comparison of two procedures, the Mahalanobis Distance and the Andrews–Pregibon Statistic, for identifying multivariate outliers // Research in the schools, 1994, vol. 1, 49–58.
32. Judd C.M., McClelland G.H. Data analysis: A model comparison approach. San Diego: Harcourt Brace Jovanovich, 1989.
33. Kadota K. Detecting outlying samples in microarray data: A critical assessment of the effect of outliers on sample classification / K. Kadota, D. Tominaga, Y. Akiyama et al. / Chem–Bio Informatics Journal, 2003, vol. 3, no. 1, pp. 30–45.
34. Lee J.C., Hong C.S. Identification of outlying cells in multi–way tables // InterStat (Statistics on the Internet), February 2001, No. 5.
35. Lin C.C., Chen A.P. Fuzzy discriminant analysis with outlier detection by genetic algorithm // Computers and Operations Research, May 2004, vol. 31, no. 6, pp. 877–888.
36. Liu H., Jezek K.C., O'Kelly M.E. Detecting outliers in irregularly distributed spatial data sets by locally adaptive and robust statistical analysis and GIS // International Journal of Geographical Information Science, 2001, vol. 15, no. 8, pp. 721–741.
37. Lohninger H. Teach/Me Data Analysis. – New York, NY: Springer–Verlag, 1999.
38. Lornez F.O. Teaching about influence in simple regression // Teaching Sociology, 1987, vol. 15, no. 2, pp. 173–177.
39. McBane G.C. Programs to compute distribution functions and critical values for extreme value ratios for outlier detection // Journal of Statistical Software, May 2006, vol. 16, no. 3.
40. Miller J.N. Outliers in experimental data and their treatment // Analyst, May 1993, vol. 118, pp. 455–461.
41. Miller J.N. Reaction time analysis with outlier exclusion: Bias varies with sample size // The Quarterly Journal of Experimental Psychology, 1991, vol. 43, no. 4, pp. 907–912.
42. Mitschele J. Small sample statistics // Journal of Chemical Education, June 1991, vol. 66, no. 6, pp. 470–473.
43. Motulsky H. Grubbs' test for detecting outliers // GraphPad Insight, Winter 1997, no. 14.
44. Motulsky H., Brown R. Detecting outliers when fitting data with nonlinear regression – a new method based on robust nonlinear regression and the false discovery rate // BMC Bioinformatics, March, 2006, vol. 7, no. 123.
45. Newton R.R., Rudestam K.E. Your statistical consultant: Answers to your data analysis questions. – Thousand Oaks: Sage, 1999.
46. NIST/SEMATECH e–Handbook of statistical methods (NIST Handbook 151, ver. 1/27/2005). – Gaithersburg, MD: National Institute of Standards and Technology, 2005.
47. Orr J.M., Sackett P.R., DuBois C.L.Z. Outlier detection and treatment in I/O psychology: A survey of researcher beliefs and an empirical illustration // Personnel Psychology, 1991, vol. 44, pp. 473–486.
48. Prescott P. An approximate test for outliers in linear models // Technometrics, February 1975, vol. 17, no. 1, pp. 129–132.
49. Rasmussen J.L. Evaluating outlier identification tests: Mahalanobis D Squared and Comrey D // Multivariate Behavioral Research, 1988, vol. 23, no. 2, pp. 189–202.
50. Rorabacher D.B. Statistical treatment for rejection of deviant values: Critical values of Dixon's «Q» parameter and related subrange ratios at the 95% confidence level // Journal of Analytical Chemistry, 1991, vol. 63, pp.139–146.
51. Rosado F. Maximum likelihood ratio principle in tests of discordancy for outliers // Bulletin

of the International Statistical Institute, 52nd Session, Proceedings, Tome LVIII, Finland 1999. Contributed Paper Meeting 40: Outliers.
52. Rousseeuw P.J., Leroy A.M. Robust regression and outlier detection. – New York, NY: John Wiley & Sons, 1987.
53. Sachs L. Applied statistics: A handbook of techniques. – New York, NY: Springer–Verlag, 1982.
54. Schwager S.J., Margolin B.H. Detection of multivariate outliers // The Annals of Statistics, 1982, vol. 10, pp. 943–954.
55. Siegel S., Casttelan Jr. N.J. Non–parametric statistics. – New York, NY: McGraw–Hill, 1988.
56. Sim C.H., Gan F.F., Chang T.C. Outlier labeling with boxplot procedures // Journal of the American Statistical Association, June 2005, vol. 100, no. 470, pp.642–652.
57. Simonoff J.S. The calculation of outlier detection statistics // Communications in statistics – Simulation and computation, 1984, vol. 13, pp. 275–285.
58. Solberg H.E., Lahti A. Detection of outliers in reference distributions: Performance of Horn's algorithm // Clinical Chemistry, 1 December 2005, vol. 51, no. 12, pp. 2326–2332.
59. Stefansky W. Rejecting outliers in factorial designs // Technometrics, 1972, vol. 14, pp. 469–479.
60. Stevens J.P. Outliers and influential data points in regression analysis // Psychological Bulletin, 1984, vol. 95, pp. 334–344.
61. Tabachnick B.G., Fidell L.S. Using multivariate statistics. – Needham Heights, MA: Pearson Allyn & Bacon, 2000.
62. Taylor J.K. Quality assurance of chemical measurements. – Chelsea: Lewis Publishers, 1987.
63. Tietjen G.L. The analysis and detection of outliers // Goodness–of–fit techniques (Statistics: textbooks and monographs, vol. 68) / Ed. by R.B. D'Agostino, M.A. Stephens. – New York, NY: Marcel Dekker, 1986.
64. Tietjen G.L., Moore R.H. Some Grubbs–type statistics for the detection of several outliers // Technometrics, 1972, vol. 14, no. 3, pp. 583–597.
65. Tiku M.L., Akkaya A.D. Robust estimation and hypothesis testing. – New Delhi: New Age International, 2004.
66. Van Selst M., Jolicoeur P. A solution to the effect of sample size on outlier elimination // The Quarterly Journal of Experimental Psychology, 1994, vol. 47, no. 3, pp. 631–650.
67. Viljoen H., Venter J.H. A computer intensive approach to find multivariate outliers // Bulletin of the International Statistical Institute, 52nd Session, Proceedings, Tome LVIII, Finland 1999. Contributed Paper Meeting 40: Outliers.
68. Wainer H. Robust statistics: A survey and some prescriptions // Journal of Educational Statistics, 1976, vol. 1, no. 4, pp. 285–312.
69. Weisberg S. Applied linear regression. – New York, NY: John Wiley & Sons, 1985.
70. Wilcox R.R. Fundamentals of modern statistical methods. – New York, NY: Springer, 2001.
71. Wilcox R.R. How many discoveries have been lost by ignoring modern statistical methods? // American Psychologist, 1998, vol. 53, no. 3, pp. 300–314.
72. Zimmerman D.W. A note on the influence of outliers on parametric and nonparametric tests // Journal of General Psychology, 1994, vol. 121, no. 4, pp. 391–401.
73. Zimmerman D.W. Increasing the power of nonparametric tests by detecting and downweighting outliers // Journal of Experimental Education, 1995, vol. 64, no. 1, pp. 71–78.
74. Zimmerman D.W. Invalidation of parametric and nonparametric statistical tests by concurrent violation of two assumptions // Journal of Experimental Education, 1998, vol. 67, no. 1, pp. 55–68.
75. Айвазян С.А., Енюков И.С., Мешалкин Л.Д. Прикладная статистика: Основы

моделирования и первичная обработка данных. Справочное издание. – М.: Финансы и статистика, 1983.
76. Афифи А., Эйзен С. Статистический анализ: Подход с использованием ЭВМ. – М.: Мир, 1982.
77. Большев Л.Н., Смирнов Н.В. Таблицы математической статистики. – М.: Наука, 1983.
78. ГОСТ Р ИСО 5725–2–2002. Точность (правильность и прецизионность) методов и результатов измерений. Часть 2. Основной метод определения повторяемости и воспроизводимости стандартного метода измерений. – М.: Издательство стандартов, 2002.
79. Грановский В.А., Сирая Т.Н. Методы обработки экспериментальных данных при измерениях. – Л.: Энергоатомиздат, 1990.
80. Гумбель Э. Статистика экстремальных значений. – М.: Мир, 1965.
81. Дерффель К. Статистика в аналитической химии. – М.: Мир, 1994.
82. Дэйвид Г. Порядковые статистики. – М.: Наука, 1979.
83. Лемешко Б.Ю. Робастные методы оценивания и отбраковка аномальных измерений // Заводская лаборатория, 1997, т. 63, № 5, с. 43–49.
84. Лемешко Б.Ю., Лемешко С.Б. Расширение области применения критериев типа Граббса, используемых при отбраковке аномальных измерений // Измерительная техника, 2005, № 6, с. 13–19.
85. Мюллер П., Нойман П., Шторм Р. Таблицы по математической статистике. – М.: Финансы и статистика, 1982.
86. Налимов В.В. Применение математической статистики при анализе вещества. – М.: Государственное издательство физико–математической литературы, 1960.
87. Прохоров Ю.В. Вероятность и математическая статистика. Энциклопедия / Гл. ред. Ю.В. Прохоров. – М.: Научное издательство «Большая Российская энциклопедия», 1999.
88. Родионов Д.А. Справочник по математическим методам в геологии / Д.А. Родионов, Р.И.Коган, В.А. Голубева и др. – М.: Недра, 1987.
89. Смоляк С.А., Титаренко Б.П. Устойчивые методы оценивания: статистическая обработка неоднородных совокупностей. – М.: Статистика, 1980.
90. Тейлор Дж. Введение в теорию ошибок. – М.: Мир, 1985.
91. Хальд А. Математическая статистика с техническими приложениями. – М.: Издательство иностранной литературы, 1956.
92. Хамханова Д.Н. Теоретические основы обеспечения единства экспертных измерений. – Улан–Удэ: Издательство ВСГТУ, 2006.
93. Хромов–Борисов Н.Н., Лаззаротто Г.Б., Ледур Кист Т.Б. Биометрические задачи в популяционных исследованиях // VII Всероссийский популяционный семинар «Методы популяционной биологии», 16–21 февраля 2004, г. Сыктывкар.
94. Янко Я. Математико–статистические таблицы. – М.: Госстатиздат, 1961.

Глава 22. Рандомизация и генерация случайных последовательностей

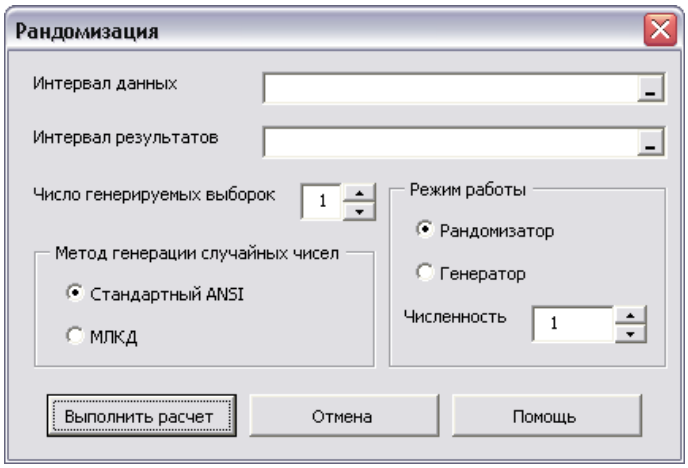
22.1. Введение

Представленное программное обеспечение позволяет выполнять рандомизацию, а также генерировать массивы случайных чисел с равномерным распределением.
Назначение рандомизации подробно описано в соответствующем разделеРандомизация.
Программное обеспечение может применяться для компьютерного моделирования и других

исследовательских целей в различных областях фундаментальной и прикладной науки. Программа явилась побочным продуктом исследовательского проекта по моделированию статистических распределений методом Монте–Карло, назначением которого были проверка стандартных статистических таблиц и вычисление мощности статистических критериев.

22.2. Работа с программным обеспечением

Выберите из меню программы пункт AtteStat | Рандомизация. На экране появится диалоговое окно, изображенное на рисунке:



Затем проделайте следующие шаги:

- Выберите или введите интервал данных. В качестве данных могут выступать ячейки, содержащие данные любых допустимых типов в любом сочетании. Это могут быть числа, номера историй болезни, закодированные фамилии пациентов и другие данные. Интервал данных необходим только в режиме работы «Рандомизатор». В режиме работы «Генератор» интервал данных вводить не обязательно.
- Выберите или введите интервал результатов. Начиная с первой интервала результатов (следовательно, можно указать только одну ячейку, т.к. остальные ячейки интервала игнорируются), будут выведены результаты вычислений.
- Введите число генерируемых выборок. Например, задано 100 пациентов, которых нужно распределить на контрольную и опытную группы. Указываем число генерируемых выборок, равное 2. Программа случайным образом распределит всех пациентов по этим двум группам. Число генерируемых выборок действительно во всех режимах работы программы.
- Выберите или оставьте по умолчанию метод генерации случайных чисел.
- Выберите или оставьте по умолчанию режим работы программы. Для режима «Генератор» обязательно введите численность генерируемой выборки.
- Нажмите кнопку «Выполнить расчет».

В ходе выполнения вычислений будут, начиная с первой ячейки интервала результатов, выведены результаты расчета. При неверных действиях пользователя выдаются сообщения об ошибках.

22.2.1. Сообщения об ошибках

При ошибках ввода данных могут выдаваться диагностические сообщения следующих типов:

Ошибка	Комментарий
Не определена область данных.	В режиме «Рандомизатор» не выбран или неверно введен интервал данных. Лучшим способом избежать ошибки является не ввод, а выделение интервала ячеек стандартным образом.
Не определена область вывода.	Не выбран или неверно введен интервал ячеек, определяющих область вывода решения. Лучшим способом избежать ошибки является не ввод, а выделение интервала ячеек стандартным образом.

22.3. Теоретическое обоснование

Под рандомизацией понимают случайное распределение объектов исследования по группам (например, контрольной и опытной). Рандомизация предполагает соблюдение двух условий: непредсказуемый (случайный) характер распределения пациентов по группам и «слепой» отбор. Предлагаемое программное обеспечение решает первую, математическую, часть задачи. Вторая часть задачи («слепой» отбор) носит в основном организационный характер. Рандомизация может осуществляться с использованием различных аппаратных, программных и аппаратно–программных генераторов равномерно распределенных случайных чисел или соответствующих таблиц. Наилучшие результаты наиболее экономичным путем в настоящее время могут быть получены с помощью специального программного обеспечения, генерирующего последовательность псевдослучайных чисел. Алгоритм рандомизации, предлагаемый программным обеспечением, начинается с генерации псевдослучайной последовательности чисел. Затем, в соответствии с данной последовательностью, осуществляется «перетасовка» всего списка (например, списка пациентов, заданных номерами историй болезни, фамилиями или любым другим удобным для пользователя способом). Для этого меняются местами идентификаторы с номерами I и J списка пациентов в соответствии с формулой

$J = Int(N \cdot R_I + 1)$,

где N – численность исходной выборки (под выборкой понимается совокупность объектов, идентифицирующих пациентов),

R_I – число псевдослучайной последовательности, стоящее на месте с номером I.

После этого производится разделение «перетасованного» списка на заданное число групп. Обычно это контрольная и опытная группы, хотя возможны различные варианты. Приведем примеры рандомизации из биомедицины и социологии.

22.3.1. Рандомизация в биомедицинских исследованиях

В настоящее время действуют «Единые требования к рукописям, представляемым в биомедицинские журналы» Всемирной ассоциации редакторов медицинских журналов (World Association of Medical Editors, WAME), объединяющей редакторов более 700 научных журналов из почти 80 стран. Официальный сайт WAME находится по ссылке http://www.wame.org). Оригинальная версия «Требований» под названием «Uniform Requirements for Manuscripts Submitted to Biomedical Journals: Writing and Editing for Biomedical Publication» общедоступна через Интернет по ссылке http://www.icmje.org на официальном сайте Международного комитета редакторов медицинских журналов (International Committee of Medical Journal Editors, ICMJE). Документ распространяется свободно. Его повсеместное распространение в информационных целях и приведение требований для авторов всех медицинских журналов в соответствие с данным документом приветствуется ICMJE. «Требования» гласят, что «сообщения о проведении рандомизированных контролируемых

исследований должны содержать информацию обо всех основных элементах исследования, включая протокол (изучаемая популяция, способы лечения или воздействия, исходы и обоснование статистического анализа), назначение лечения (методы рандомизации, способы сокрытия формирования групп лечения) и методы маскировки (обеспечения «слепого» контроля). Авторы, представляющие обзоры литературы, должны включить в них раздел, в котором описываются методы, используемые для нахождения, отбора, получения информации и синтеза данных. Эти методы также должны быть приведены в резюме». «Приведите детали процесса рандомизации. Опишите, какие методы были применены для обеспечения «слепого» контроля и насколько успешно».
Показано применение методов рандомизации в любых областях, где из некоторого списка необходимо получить тот же список, но в котором позиции будут расположены в случайном порядке, обеспечивая равноправие выбора.

О рандомизации в клинических исследованиях см. монографию Сергиенко с соавт.
В статье Орешкина на примере выборов депутатов муниципальных собраний показано, что игнорирование рандомизации является одной из возможностей фальсификации выборов. Данная возможность реализуется следующим образом. На основе основополагающих принципов человеческого восприятия, сформулированных Огилви (Ogilvy), кандидат, находящийся в первой строчке, получает несколько дополнительных процентов голосов электората, не являющегося активным, но пришедшего на выборы.

22.3.2. Генерация случайных последовательностей

Различают случайные и псевдослучайные числа. Они различаются тем, что первые из них генерируются каким–либо стохастическим устройством, а вторые генерируются арифметическим численным алгоритмом. И в том, и в другом случае последовательности чисел обладают определенными статистическими свойствами. В нашем случае программно генерируются псевдослучайные числа, равномерно распределенные в интервале [0,1].
В программном обеспечении используются 2 метода генерации псевдослучайной последовательности чисел, относящиеся к конгруэнтным генераторам:

- стандартный генератор, рекомендованный комитетом ANSI,
- мультипликативный линейный конгруэнтный датчик.

В теории чисел приятa особая запись выражений, незнакомая многим специалистам, не имеющим специального математического образования. Напомним, что она означает. Пусть a и b – целые числа. Если их разность делится на число m, то этот факт выражается записью $a = b \pmod m$.
Запись читается «a сравнимо с b по модулю m». Делитель m положительный. Он называется модулем сравнения.
Показанная формула буквально означает
$a – b = mk$,
где k – целое число.
Проще говоря, функция mod в записи $b \pmod m$ определяется как остаток от деления b на m.

Подробный обзор см. в книге Иванова с соавт. Петрович с соавт. представили датчики случайных чисел с различными законами распределения. См. также книгу Эфроса. Из других широко применяемых современных алгоритмов назовем генераторы Фибоначчи, описанные в литературе.

22.3.2.1. Стандартный генератор ANSI

Стандартный генератор ANSI, он же линейный конгруэнтный датчик, включенный в

библиотеки программ, поставляемых с различными системами программирования, дает высококачественные случайные числа, достаточные для некоторых практических применений.
Вычисление последовательности псевдослучайных чисел производится по формуле
$R_{n+1} = (a \cdot R_n + c)(\mathrm{mod}\, m)$,
где a – мультипликатор,
c – инкремент,
m – модуль.
Целые константы a, c, m выбираются определенным образом. Формула дает последовательность псевдослучайных чисел, пригодную для практических применений в некоторых областях, например, рандомизации, однако для целей компьютерного моделирования [распределений] последовательность, полученная с помощью рассматриваемого генератора, считается малопригодной.

22.3.2.2. Мультипликативный линейный конгруэнтный датчик

Если в соотношении для линейного конгруэнтного датчика положить значение инкремента $c = 0$, то оно упростится до
$R_{n+1} = (a \cdot R_n)(\mathrm{mod}\ m)$.
Датчики, основанные на этой формуле, называются мультипликативными линейными конгруэнтными датчиками (МЛКД). МЛКД в источниках называется также генератором Парка–Миллера (Park, Miller) в честь авторов, исследовавших наборы констант, входящих в формулу.
Существуют различные версии реализаций генератора, в том числе объединения нескольких МЛКД. Одной из известных реализаций такого генератора является классическая программа, составленная Лекюйе (L'Ecuyer).
Начальное значение для рекуррентной формулы может выбираться различными способами. Не лишен смысла прием, основанный на запросе программой системного времени.

Список использованной и рекомендуемой литературы

1. Borland C++. Version 3. Library reference. – Scotts Valley, CA: Borland International, 1991.
2. Brent R.P. Note on Marsaglia's xorshift random number generators // Journal of Statistical Software, August 2004, vol. 11, no. 5.
3. Julious S.A. Sample sizes for clinical trials with normal data // Statistics in Medicine, 2004, vol. 23, no. 12, pp. 1921–1986.
4. Kahaner D.K., Moler C., Nash S.G. Numerical methods and software. – Englewood Cliffs, NJ: Prentice Hall, 1989.
5. L'Ecuyer P. Random number generation // Elsevier Handbooks in Operations Research and Management Science: Simulation / Ed. by S.G. Henderson, B.L. Nelson. – Oxford, UK: Elsevier Science, 2005.
6. L'Ecuyer P. Random number generation // The Handbook of Computational Statistics / Ed. by J.E. Gentle, W. Haerdle, Y. Mori. – Heidelberg: Springer–Verlag, 2004, pp. 35–70.
7. L'Ecuyer P., Hellekalek P. Random number generators: Selection criteria and testing // Random and Quasi–Random Point Sets (Lecture Notes in Statistics, vol. 138, pp. 223–265) / Ed. by P. Hellekalek, G. Larcher. – New York, NY: Springer, 1998.
8. Marsaglia G. Random number generators // Journal of Modern Statistical Methods, May 2003, vol. 2, no. 1, pp. 2–13.
9. Marsaglia G., Tsang W.W., Wang J. Fast generation of discrete random variables // Journal of Statistical Software, July 2004, vol. 11, no. 3.

10. Ogilvy D. Ogilvy on advertising. – Toronto: John Wiley & Sons, 1983.
11. Pezeshk H. How many subjects? – A Bayesian approach to the design of a clinical trial // Iranian International Journal of Science, 2002, vol. 3, no. 1, pp. 127–133.
12. Poole W.K. An investigation of four randomization algorithms for a two arm randomized trial / W.K. Poole, R.H. Thornton, R. Perritt et al. // InterStat, May 2001, No. 1.
13. Rosenberger W.F., Lachin J.M. Randomization in clinical trials. – New York: John Wiley & Sons, 2002.
14. Van Belle G. Biostatistics: A methodology for the health sciences // G. van Belle, L.D. Fisher, P.J. Heagerty et al. – New York, NY: John Wiley & Sons, 2003.
15. Vattulainen I. A comparative study of some pseudorandom number generators / I. Vattulainen, K. Kankaala, J. Saarinen et al. // Computer Physics Communications, 1995, vol. 86, pp. 209–226.
16. Брандт З. Анализ данных. Статистические и инженерные методы для научных работников и инженеров. – М.: Мир, ООО «Издательство АСТ», 2003.
17. Власов В.В. Введение в доказательную медицину. – М.: Медиа Сфера, 2001.
18. Гайдышев И.П. Статистика в публикациях // Гений ортопедии, 2005, № 4, с. 155–161.
19. Иванов М.А., Чугунков И.В. Теория, применение и оценка качества псевдослучайных последовательностей. – М.: КУДИЦ–ОБРАЗ, 2003.
20. Каханер Д., Моулер К., Нэш С. Численные методы и программное обеспечение. – М.: Мир, 2001.
21. Козлов М.В. Мнимые повторности (pseudoreplications) в экологических исследованиях: проблема, не замеченная российскими учеными // Журнал общей биологии, 2003, т. 64, № 4, с. 292–307.
22. Котельников Г.П., Шпигель А.С. Доказательная медицина. Научно–обоснованная медицинская практика. – Самара: СамГМУ, 2000.
23. Маркова Е.В., Маслак А.А. Рандомизация и статистический вывод. – М.: Финансы и статистика, 1986.
24. Мюллер П., Нойман П., Шторм Р. Таблицы по математической статистике. – М.: Финансы и статистика, 1982.
25. Огилви Д. Огилви о рекламе. – М.: ЭКСМО, 2006.
26. Оре О. Приглашение в теорию чисел. – М.: Наука, 1980.
27. Орешкин Д. Теория вероятностей как культурный диагноз // Ежедневный журнал, 8 марта 2007.
28. Петрович М.Л., Давидович М.И. Статистическое оценивание и проверка гипотез на ЭВМ. – М.: Финансы и статистика, 1989.
29. Прохоров Ю.В. Вероятность и математическая статистика: Энциклопедия / Гл. ред. Ю.В. Прохоров. – М.: Большая Российская энциклопедия, 1999.
30. Сборник научных программ на Фортране. Руководство для программиста. Выпуск 1. Статистика. – М.: Статистика, 1974.
31. Сергиенко В.И., Бондарева И.Б. Математическая статистика в клинических исследованиях. – М.: ГЭОТАР–Медиа, 2006.
32. Теннант–Смит Дж. Бейсик для статистиков. – М.: Мир, 1988.
33. Флейс Дж. Статистические методы для изучения таблиц долей и пропорций. – М.: Финансы и статистика, 1989.
34. Эфрос А.Л. Физика и геометрия беспорядка (Библиотечка «Квант», выпуск 19). – М.: Наука, 1982.

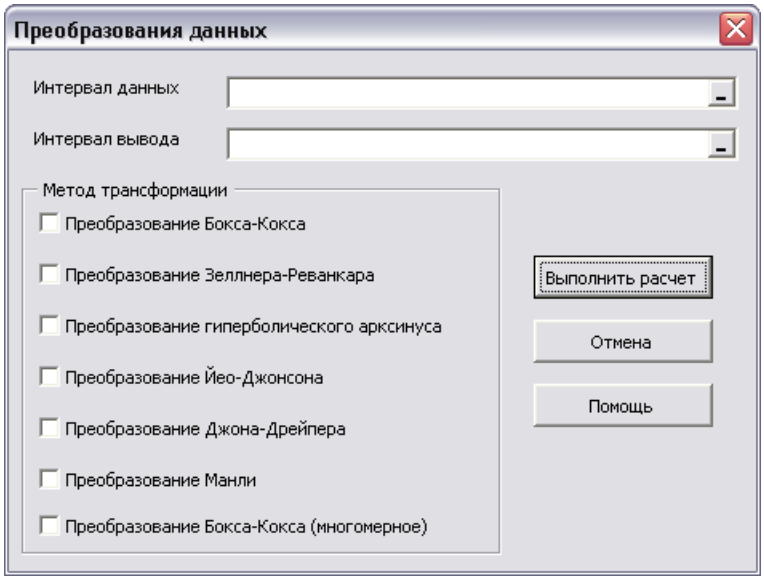
Глава 23. Преобразования данных

23.1. Введение

В данной гдаве описаны реализованные в программном обеспечении классические методы преобразования данных.

23.2. Работа с программным обеспечением

Выберите из меню программы пункт AtteStat | Преобразования данных. На экране появится диалоговое окно, изображенное на рисунке:



Затем:

- Выберите или введите интервал данных. В качестве входного интервала можно использовать любую прямоугольную область рабочего листа. При этом часть столбцов выделенной прямоугольной области может быть заполнена не до конца (см. раздел «Преобразования данных»).
- Выберите или введите выходной интервал для выдачи результатов расчета. Начиная с первой ячейки выходного интервала (следовательно, можно указать только одну ячейку, т.к. остальные ячейки интервала игнорируются), будут выведены вычисленные показатели.
- Выберите метод преобразования.
- Нажмите кнопку «Выполнить расчет».

При ошибках, вызванных неверными действиями пользователя при вводе исходных данных для расчета, выдаются сообщения об ошибках.

23.2.1. Сообщения об ошибках

При ошибках ввода и во время выполнения программы могут выдаваться диагностические сообщения следующих типов:

Ошибка	Комментарий
Не определен интервал переменной.	Вы не выбрали или неверно ввели интервал эмпирической выборки. Лучшим способом избежать ошибки является не ввод, а выделение интервала стандартным образом, т. е. протаскиванием курсора.
Не определена область вывода.	Вы не выбрали или неверно ввели выходной интервал. Лучшим способом избежать ошибки является не ввод, а выделение интервала стандартным образом, т. е. протаскиванием курсора.
Мала численность выборки.	Численность каждой выборки не может быть меньше двух. Укажите интервал матрицы исходных данных, содержащих выборки численностью от двух и более.

23.3. Теоретическое обоснование

Все преобразования данных можно разделить на 2 группы:
- универсальные,
- частные.

Универсальное преобразование – преобразование к нормальной линейной модели. В программе представлены следующие методы универсального преобразования данных:
- преобразование Бокса–Кокса,
- преобразование Зеллнера–Реванкара,
- преобразование гиперболического арксинуса,
- преобразование Йео–Джонсона,
- преобразование Джона–Дрейпера,
- преобразование Манли,
- многомерное преобразование Бокса–Кокса.

Частные преобразования подразделяются на преобразования:
- нормализующие ошибки,
- стабилизирующие дисперсии.

Частные методы преобразования обычно рассматриваются при изучении критериев обнаружения гетероскедастичности и способов ее устранения, часто в курсе эконометрики. В настоящем программном обеспечении методы не представлены. См. монографии Сокал (Sokal) с соавт., Зар (Zar).

Выбор представленной номенклатуры методов обусловлен характеристиками исходных данных, которые методы могут обрабатывать:
- классические преобразования Бокса–Кокса и Зеллнера–Реванкара – только неотрицательные варианты,
- остальные методы – любые варианты.

Программа вычисляет оптимальное значение параметра преобразования и соответствующее ему значение логарифмической функции максимального правдоподобия (ФМП), а также выводит преобразованные данные, структура которых соответствует исходным данным. ФМП характеризуют качество подгонки модели. ФМП для различных методов могут сравниваться между собой.

23.3.1. Одномерное преобразование

Классическое одномерное преобразование применяется в случаях, когда:
- Представлена одна выборка.
- Представлена совокупность выборок из одной и той же генеральной совокупности.

Данная ситуация возникает, например, в однофакторном дисперсионном анализе (см. главу «Дисперсионный анализ»). Численности каждой выборки (каждого столбца таблицы) в данном случае могут совпадать или различаться. При этом с точки зрения оптимизации вся совокупность считается одной выборкой.

Параметр преобразования находится из условия максимума ФМП, с точностью до константы,

$$L(\lambda) = \sum_{i=1}^{n} \ln J_\lambda(x_i) - \frac{n}{2} \ln \hat{\sigma}^2(\lambda),$$

где $\hat{\sigma}^2(\lambda)$ – оценка дисперсии преобразования,
λ – скалярный параметр преобразования,
n – численность выборки,
x_i, $i = 1,2,...,n$ – исходная выборка,
$y_i(\lambda)$, $i = 1,2,...,n$ – преобразованные данные,
$J_\lambda(.)$ – якобиан преобразования (определитель матрицы производных $y_i(\lambda)$ по x_i), рассчитываемый по соответствующей формуле, зависящей от применяемого метода преобразования.

Оценка дисперсии может быть рассчитана по стандартной формуле

$$\hat{\sigma}^2(\lambda) = \frac{1}{n-1}\left[\sum_{i=1}^{n} y_i^2(\lambda) - \frac{1}{n}\left(\sum_{i=1}^{n} y_i(\lambda)\right)^2\right].$$

Поиск максимума ФМП осуществляется численно одним из методов локальной оптимизации. В программе применяется дихотомический поиск (метод деления отрезка пополам).

23.3.1.1. Преобразование Бокса–Кокса

Преобразование Бокса–Кокса (Box–Cox transformation) для каждой неотрицательной варианты исходной выборки запишется как

$$y_i(\lambda) = \begin{cases} \dfrac{x_i^\lambda}{\lambda}, \lambda \neq 0, \\ \ln x_i, \lambda = 0, \end{cases}$$

где λ – параметр преобразования,
x_i, $i = 1,2,...,n$ – исходная выборка,
$y_i(\lambda)$, $i = 1,2,...,n$ – преобразованные данные,
n – численность выборки.

Якобиан преобразования рассчитывается по формуле

$$J_\lambda(x_i) = x_i^{\lambda-1}, i = 1,2,...,n.$$

См. статьи Бокса (Box) с соавт., Спитцера (Spitzer), Йенг (Yang) с соавт., Линтона (Linton) с соавт., Сакиа (Sakia), Бикел (Bickel) с соавт., монографии Армитэйдж (Armitage) с соавт., Сокал (Sokal) с соавт., работы Зарембка (Zarembka), Джонсона (Johnson R.A.). Метод упоминается в учебниках по эконометрике, например, Доугерти.

23.3.1.2. Преобразование Зеллнера–Реванкара

Преобразование Зеллнера–Реванкара (Zellner–Revankar transformation) для каждой неотрицательной варианты исходной выборки запишется как

$$y_i(\lambda) = \ln(x_i) + \lambda x_i^2, i = 1,2,...,n,$$

где λ – параметр преобразования,

x_i, i = 1,2,...,n – исходная выборка,
$y_i(\lambda)$, i = 1,2,...,n – преобразованные данные,
n – численность выборки.
Якобиан преобразования рассчитывается по формуле

$$J_\lambda(x_i) = x_i^{-1} + 2\lambda x_i, i = 1,2,...,n.$$

См. статьи Зеллнер (Zellner) с соавт., Линтон (Linton) с соавт.

23.3.1.3. Преобразование гиперболического арксинуса

Преобразование гиперболического арксинуса (Arcsinh transformation) для каждой варианты исходной выборки запишется как
$y_i(\lambda) = Ash(\lambda x_i) / \lambda$, i = 1,2,...,n,
где λ – параметр преобразования, $\lambda \neq 0$,
x_i, i = 1,2,...,n – исходная выборка,
$y_i(\lambda)$, i = 1,2,...,n – преобразованные данные,
n – численность выборки,
$Ash(.)$– функция гиперболического арксинуса.
Гиперболический арксинус может быть подсчитан по формуле

$$Ash(a) = \ln\left(a + \sqrt{1 + a^2}\right).$$

Якобиан преобразования рассчитывается по формуле

$$J_\lambda(x_i) = (1 + \lambda^2 x_i^2)^{-1/2}, i = 1,2,...,n.$$

См. статьи Линтона (Linton) с соавт., Джонсона (Johnson N.L.), Робинсона (Robinson), отчет Сперлич (Sperlich) с соавт.

23.3.1.4. Преобразование Йео–Джонсона

Преобразование Йео–Джонсона (Yeo–Johnson transformation) для каждой варианты исходной выборки запишется как

$$y_i(\lambda) = \begin{cases} [(x_i + 1)^\lambda - 1]/\lambda, \lambda \neq 0, x_i \geq 0 \\ \ln(x_i + 1), \lambda = 0, x_i \geq 0 \\ -[(1 - x_i)^{2-\lambda} - 1]/(2 - \lambda), \lambda \neq 2, x_i < 0 \\ -\ln(1 - x), \lambda = 2, x_i < 0 \end{cases}, i = 1,2,...,n,$$

где λ – параметр преобразования,
x_i, i = 1,2,...,n – исходная выборка,
$y_i(\lambda)$, i = 1,2,...,n – преобразованные данные,
n – численность выборки.
Якобиан преобразования рассчитывается по формуле

$$J_\lambda(x_i) = \begin{cases} (x_i + 1)^{\lambda - 1}, x \geq 0 \\ (1 - x_i)^{1-\lambda}, x < 0 \end{cases}, i = 1,2,...,n.$$

См. главу 7 монографии Вайсберга (Weisberg), доклад Саманта (Samanta).

23.3.1.5. Преобразование Джона–Дрейпера

Преобразование Джона–Дрейпера (John–Draper transformation) для каждой варианты исходной выборки запишется как

$$y_i(\lambda) = \begin{cases} sign(x_i)[(|x_i|+1)^{\lambda} - 1]/\lambda, \lambda \neq 0 \\ sign(x_i)\ln(|x_i|+1), \lambda = 0 \end{cases}, i = 1,2,...,n,$$

где λ – параметр преобразования,
x_i, i = 1,2,...,n – исходная выборка,
$y_i(\lambda)$, i = 1,2,...,n – преобразованные данные,
n – численность выборки.
Якобиан преобразования рассчитывается по формуле
$J_\lambda(x_i) = sign(x_i)(|x_i| + 1)^{\lambda - 1}$, i = 1,2,...,n.

См. статьи Джона (John) с соавт., фон Хиппель (von Hippel), Чен (Chen) с соавт., диссертацию Ислама (Islam) , доклад Саманта (Samanta).

23.3.1.6. Преобразование Манли

Преобразование Манли (Manly transformation) для каждой варианты исходной выборки запишется как

$$y_i(\lambda) = \begin{cases} (e^{\lambda x_i} - 1)/\lambda, \lambda \neq 0 \\ x_i, \lambda = 0 \end{cases}, i = 1,2,...,n,$$

где λ – параметр преобразования,
x_i, i = 1,2,...,n – исходная выборка,
$y_i(\lambda)$, i = 1,2,...,n – преобразованные данные,
n – численность выборки.
Якобиан преобразования рассчитывается по формуле
$J_\lambda(x_i) = e^{\lambda x_i}, i = 1,2,...,n.$

См. статью Манли (Manly), диссертацию Ислама (Islam).

23.3.2. Многомерное преобразование

Многомерное преобразование применяется в случае, если данные представлены в виде одной многомерной выборки. Предполагается, что по строкам таблицы при этом расположены варианты, по столбцам – параметры. Численности столбцов в данном случае должны совпадать.
Было бы неверным для каждого параметра многомерной выборки вместо многомерного преобразования применить одномерные маргинальные преобразования, т.к. целью многомерного преобразования является достижение требуемого эффекта (например, многомерной нормальности) всей многомерной выборки, а не только ее компонент. Соображения на эту тему см. также в главе «Проверка нормальности распределения».
В многомерном случае удобно векторные параметры обозначать соответствующими прописными латинскими и греческими литерами. В этом случае векторный параметр преобразования находится из условия максимума ФМП, с точностью до константы,

$$L(\Lambda) = \sum_{i=1}^{n} \ln J_\Lambda(X_i) - \frac{n}{2}\ln|\Sigma(\Lambda)|,$$

где $\Sigma(\Lambda)$ – оценка дисперсионно–ковариационной матрицы преобразования,
|.| – операция вычисления определителя,
Λ – векторный параметр преобразования,
n – численность многомерной выборки,
X_i, i = 1,2,...,n – исходная многомерная выборка,

$Y_i(\Lambda)$, $i = 1,2,...,n$ – преобразованные данные,
$J_\Lambda(.)$ – якобиан преобразования (определитель матрицы производных $Y_i(\Lambda)$ по X_i), рассчитываемый по соответствующей формуле, зависящей от применяемого метода преобразования.
Поиск максимума ФМП может осуществляться численно одним из методов локальной оптимизации. В программе применяется метод переменной метрики (метод Бройдена–Флетчера–Голдфарба–Шанно). Определитель эффективно рассчитывается как побочный продукт разложения Грама дисперсионно–ковариационной матрицы.

Методы оптимизации см. в книгах Дэнниса с соавт., Носача, Ортега с соавт. О разложении матриц см. монографию Уилкинсона с соавт.

23.3.2.1. Многомерное преобразование Бокса–Кокса

Многомерное преобразование Бокса–Кокса (multivariate Box–Cox transformation) для каждой неотрицательной многомерной варианты исходной выборки в поэлементном виде запишется как

$$y_{ij}(\lambda_j) = \begin{cases} \dfrac{x_{ij}^{\lambda_j}}{\lambda_j}, \lambda_j \neq 0, \\ \ln x_{ij}, \lambda_j = 0, \end{cases}$$

где λ_j, $j = 1,2,...,p$ – компоненты векторного параметра преобразования,
x_{ij}, $i = 1,2,...,n$; $j = 1,2,...,p$ – компоненты исходной многомерной выборки,
$y_{ij}(\lambda_j)$, $i = 1,2,...,n$; $j = 1,2,...,p$ – компоненты преобразованных данных,
n – численность многомерной выборки,
p – размерность многомерной выборки.
Якобиан преобразования рассчитывается по формуле

$$J_{\lambda_j}(x_{ij}) = x_{ij}^{\lambda_j - 1}, i = 1,2,...,n; j = 1,2,...,p.$$

С учетом данной явной формы якобиана логарифмическая функция максимального правдоподобия для представленного метода преобразования запишется в виде

$$L(\Lambda) = \sum_{j=1}^{p}(\lambda_j - 1)\sum_{i=1}^{n}\ln x_{ij} - \frac{n}{2}\ln|\Sigma(\Lambda)|,$$

где $\Sigma(\Lambda)$ – оценка дисперсионно–ковариационной матрицы преобразования,
Λ – векторный параметр преобразования,
$|.|$ – операция вычисления определителя.
При выборе представленного метода преобразования пользователю предоставляется напоминание о необходимости сохранить данные перед производством расчета. Это вызвано возможными вычислительными проблемами решения достаточно сложной оптимизационной задачи. Если такие проблемы возникли, аварийное снятие задачи с выполнения производится средствами операционной системы.
В отличие от одномерных методов, в результате работы данного многомерного метода выводятся компоненты вектора параметра преобразования. Расположение выведенных компонент соответствуют порядку параметров исходной многомерной выборки.

Метод представлен в статье Эндрюс (Andrews) с соавт., работе Уильямс (Williams) с соавт., препринте Рупперта (Ruppert).

Список использованной и рекомендуемой литературы

1. Andrews D.F., Gnanadesikan R., Warner J.L. Transformations of multivariate data // Biometrics, 1971, vol. 27, no. 4, pp. 825–840.
2. Armitage P., Berry G., Matthews J.N.S. Statistical methods in medical research. – Oxford, UK: Blackwell Science, 2001.
3. Bickel P.J., Doksum K.A. An analysis of transformations revisited // Journal of the American Statistical Association, 1981, vol. 76, no. 374, pp. 296–311.
4. Box G.E.P., Cox D.R. An analysis of transformations // Journal of Royal Statistical Society: Series B (Statistical Methodology), 1964, vol. 26, no. 2, pp. 211–246.
5. Carrol R.J. A robust method for testing transformation to achieve approximate normality // Journal of Royal Statistical Society: Series B (Statistical Methodology), 1980, vol. 42, no. 1, pp. 71–78.
6. Carroll R.J., Ruppert D. On prediction and the power transformation family // Biometrika, 1981, vol. 68, no. 3, pp. 609–615.
7. Chen C.–R., Wang L.–C. Likelihood inference under the general response transformation model with heteroscedastic errors // Taiwanese Journal of Mathematics, June 2003, vol. 7, no. 2, pp. 261–273.
8. Cox D.R., Small N.J.H. Testing multivariate normality // Biometrika, 1978, vol. 65, no. 2, pp. 263–272.
9. Foster A.M., Tian L., Wei L.J. Estimation for the Box–Cox transformation model without assuming parametric error distribution // Journal of the American Statistical Association, September 2001, vol. 96, pp. 1097–1101.
10. Huber W. Variance stabilization applied to microarray data calibration and to the quantification of differential expression / W. Huber, A. von Heydebreck, H. Sultmann et al. // Bioinformatics, July 2002, vol. 18, suppl. 1, pp. S96–S104.
11. Islam K. Transformed tests for homogenity of variances and means. A dissertation submitted to the graduate. – College of Bowling Green State University, August 2006.
12. John J.A., Draper N.R. An alternative family of transformations // Applied Statistics, 1980, Vol. 29, No. 2, pp. 190–197.
13. Johnson N.L. Systems of frequency curves generated by methods of translation // Biometrika, June 1949, vol. 36, no. 1/2, pp. 149–176.
14. Johnson R.A. Transformation of survival data // Survival Analysis: Proceedings of the Special Topics Meeting, 16–28 October, 1981 (Hayward, California, USA: Institute of Mathematical Statistics, 1982), pp. 118–136.
15. Lamont–Smith T. Translation to the normal distribution for radar clutter // Radar, Sonar and Navigation, IEE Proceedings, February 2000, vol. 147, no. 1, pp. 17 – 22.
16. Linton O.B. An analysis of transformations for additive nonparametric regression / O.B. Linton, R. Chen, N. Wang et al. // Journal of the American Statistical Association, 1997, vol. 92, no. 440, pp. 1512–1521.
17. Manly B.F.J. Exponential data transformations // The Statistician, March 1976, vol. 25, no. 1, pp. 37–42.
18. Olson C.L. On choosing a test statistic in multivariate analysis of variance // Psychological Bulletin, July 1976, vol. 83, no. 4, pp. 579–586.
19. Riani M. Robust multivariate transformations to normality: Constructed variables and likelihood ratio tests // Statistical Methods and Applications, September 2004, vol. 13, no. 2, pp. 179–196.
20. Robinson P.M. Best nonlinear three–stage least squares estimation of certain econometric models // Econometrica, 1991, vol. 59, no. 3, pp. 755–786.
21. Ruppert D. Multivariate transformations // International Encyclopedia of Social and

Behavioral Sciencees / Ed. by N.J. Smelser, P.B. Baltes. – St. Louis, MO: Elsevier, 2001.
22. Sakia R.M. The Box–Cox transformation technique: A review // The Statistician, 1992, vol. 41, no. 2, pp. 169–178.
23. Samanta G.P. On the new transformation–based approach to value–at–risk: Application to Indi stock market // Capital Markets 9th Capital Markets Conference Paper, Indi Institute.
24. Sokal R.R., Rohlf F.J. Biometry: The principles and practice of statistics in biological research. – New York, NY: W.H. Freeman, 1995.
25. Sperlich S., Linton O., Van Keilegom I. A computational note on estimation of a semiparametric transformation model // Working Paper 02.2008, Centre for Statistics, Georg–August Universitat Gottingen, 2008.
26. Spitzer J.J. A primer on Box–Cox estimation // The Review of Economics and Statistics, May 1982, vol. 64, no. 2, pp. 307–313.
27. Taylor J.M.G. Power transformations to symmetry // Biometrika, 1985, vol. 72, no. 1, pp. 145–152.
28. Von Hippel P.T. Normalization // Encyclopedia of Social Science Research Methods. – Thousand Oaks, CA: Sage, 2003.
29. Weisberg S. Applied linear regression. – New York, NY: John Wiley & Sons, 2005.
30. Williams J.T. Joint multipoint linkage analysis of multivariate qualitative and quantitative traits. I. Likelihood formulation and simulation results / J.T. Williams, P. Van Eerdewegh, L. Almasy et al. // The American Journal of Human Genetics, October 1999, vol. 65, no. 4, pp. 1134–1147.
31. Yang Z. A modified family of power transformations // Economics Letters, July 2006, vol. 92, no. 1, pp. 14–19.
32. Yang Z., Abeysinghe T. An explicit variance formula for the Box–Cox functional form estimator // Economics Letters, July 2002, vol. 76, no. 2, pp. 259–265.
33. Yeo I.–K., Johnson R.A. A new family of power transformations to improve normality or symmetry // Biometrika, 2000, vol. 87, no. 4, pp. 954–959.
34. Zar J.H. Biostatistical analysis. – Upper Saddle River, NJ: Prentice Hall, 1999.
35. Zarembka P. Transformation of variables in econometrics // In Frontiers in Econometrics / Ed. by P. Zarembka. – New York, NY: Academic Press, 1974.
36. Zellner A., Revankar N.S. Generalized production functions // Review of Economic Studies, 1969, vol. 37, no. 2, pp. 241–250.
37. Доугерти К. Введение в эконометрику. – М.: ИНФРА–М, 1999.
38. Дэннис Дж., мл., Шнабель Р. Численные методы безусловной оптимизации и решения нелинейных уравнений. – М.: Мир, 1999.
39. Кремер Н.Ш., Путко Б.А. Эконометрика: Учебник для вузов. – М.: ЮНИТИ–ДАНА, 2005.
40. Носач В.В. Решение задач аппроксимации с помощью персональных компьютеров. – М.: МИКАП, 1994.
41. Ортега Дж., Рейнболдт В. Итерационные методы решения нелинейных систем уравнений со многими неизвестными. – М.: Мир, 1975.
42. Уилкинсон, Райнш. Справочник алгоритмов на языке АЛГОЛ. Линейная алгебра. – М.: Машиностроение, 1976.

Глава 24. Матричная и линейная алгебра

24.1. Введение

Программное обеспечение матричной и линейной алгебры предназначено для выполнения

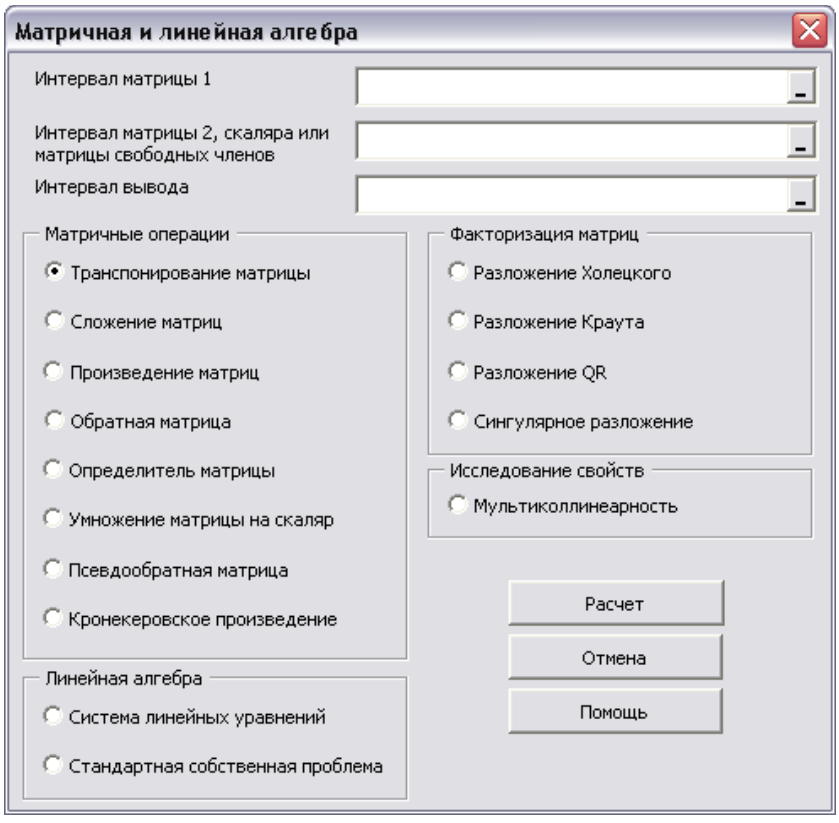
различных матричных операций. Номенклатура матричных операций насчитывает лишь несколько наиболее употребительных методов. Однако более сложные операции могут быть реализованы последовательным применением представленных элементарных операций в требуемой комбинации.
В программе также представлены несколько основных алгоритмов, составляющих предмет линейной алгебры. Методы факторизации (разложения, декомпозиции) матриц помогут при конструировании алгоритмов, примеры которых приводятся.

24.2. Работа с программным обеспечением

Для работы выберите из меню программы пункт AtteStat | Матричная и линейная алгебра. На экране появится диалоговое окно, подобное окну, изображенному на рисунке.
Затем:

- Выберите или введите интервал матрицы 1.
- Выберите или введите интервал матрицы 2 (для операций, предусматривающих работу с двумя матрицами) или скаляра (для операций, предусматривающих работу со скалярной величиной).
- Выберите или введите выходной интервал (интервал вывода). Начиная с первой ячейки выходного интервала (следовательно, можно указать только одну ячейку, т.к. остальные ячейки интервала игнорируются), будут выведены результаты вычислений.
- Выберите или оставьте по умолчанию метод расчета.
- Нажмите кнопку Расчет.

После выполнения вычислений будет, начиная с первой ячейки выходного интервала, выведено название метода и результаты расчета.
Программное обеспечение берет на себя верификацию исходных данных, выдавая подробную диагностику. При неверных действиях пользователя выдаются сообщения об ошибках.



24.2.1. Сообщения об ошибках

При ошибках ввода данных могут выдаваться диагностические сообщения следующих типов:

Ошибка	Комментарий
Не задана матрица 1.	Вы не выбрали или неверно ввели интервал матрицы 1. Лучшим способом избежать ошибки является не ввод, а выделение интервала ячеек стандартным образом.
Не задана матрица 2 или скаляр.	Вы не выбрали или неверно ввели интервал матрицы 2 или скаляра. Лучшим способом избежать ошибки является не ввод, а выделение интервала ячеек стандартным образом.
Не задан интервал вывода.	Вы не выбрали или неверно ввели интервал вывода. Лучшим способом избежать ошибки является не ввод, а выделение интервала ячеек стандартным образом.
Пустая ячейка.	В данных, указанных в качестве матрицы 1 или 2, встретилась пустая ячейка. Необходимо заполнить или устранить все пустые ячейки.
Нечисловой тип данных.	В данных, указанных в качестве матрицы 1 или 2, встретилась нечисловая ячейка. Необходимо устранить данную ошибку. Лучшим способом является выделение ячейки или группы ячеек и объявление их числовыми с помощью стандартных средств.
Не совпадают размеры матриц.	Для выбранного метода расчета размеры матриц 1 и 2 (количества строк и столбцов) должны быть совпадать. Например, для решения системы линейных уравнений требуется, чтобы количество строк в матрице системы и в матрице правых частей было одинаковым. Проверьте исходные данные.
Не соответствуют размеры матриц.	Для выбранного метода расчета размеры матриц 1 и 2 должны соответствовать друг другу. Так, для корректного выполнения операции умножения двух матриц число столбцов первого сомножителя (матрицы 1) должно равняться числу строк второго сомножителя (матрицы 2).
Матрица не квадратная.	Выбранный метод расчета может быть выполнен только для квадратной матрицы, т. е. такой матрицы, у которой количество строк равно количеству столбцов.
Число строк меньше числа столбцов.	Выбранный метод расчета может быть выполнен только для матрицы, у которой количество строк больше или равно количеству столбцов.
Ошибка вычисления.	Ошибка может быть связана с делением на нуль, потерей значимости, переполнением или выходом из области допустимых значений. В этом случае произвести расчет избранным методом не удастся. Например, данная ошибка получается, если пытаться найти обратную матрицу к матрице, состоящей из одних нулей.

24.3. Теоретическое обоснование

В раздел включены необходимые в повседневной работе исследователя алгоритмы матричных вычислений и линейной алгебры. От их качественного исполнения во многом зависит не только эффективность, но и сама работоспособность других алгоритмов.
Напомним, что матрицей называется прямоугольная таблица чисел (скаляров) размером n строк на m столбцов. При этом матрица размером n x 1 называется столбцом (вектор–столбцом), а матрица размером 1 x m называется строкой (вектор–строкой). При $m = n$ матрица называется квадратной.
Из предлагаемых алгоритмов матричных операций могут быть сконструированы почти все операции, встречающиеся на практике. Так, например, с использованием алгоритмов сложения матриц и умножения матрицы на число, в данном случае на –1, может быть получен алгоритм вычитания матриц.

24.3.1. Транспонирование матрицы

Транспонированной называется матрица, полученная из данной прямоугольной матрицы путем замены ее строк соответствующими столбцами.
Транспонирование матрицы в матричной записи записывается как:
$C = A^T$,
где A – исходная матрица,
C – транспонированная матрица.
Транспонирование матрицы ведется по формуле (в поэлементной записи):

$$c_{ji} = a_{ij}, i = 1,2,...,n; j = 1,2,...,m,$$

где $a_{..}$ – элемент исходной матрицы,
$c_{..}$ – элемент транспонированной матрицы,
n – число строк в матрице A и столбцов в матрице C,
m – число столбцов в матрице A и строк в матрице C.

24.3.2. Сложение матриц

Сложение двух матриц в матричной записи записывается как:
$C = A + B$,
где A – первое слагаемое,
B – второе слагаемое,
C – матрица – сумма двух матриц.
Сложение двух матриц ведется по формуле (в поэлементной записи):
$c_{ij} = a_{ij} + b_{ij}$, $i = 1,2,...,n$; $j = 1,2,...,m$,
где $a_{..}$ – элемент матрицы – первого слагаемого,
$b_{..}$ – элемент матрицы – второго слагаемого
$c_{..}$ – элемент матрицы – суммы,
n – число строк в каждой из матриц,
m – число столбцов в каждой из матриц.

24.3.3. Произведение матриц

Произведение двух матриц в матричной записи записывается как:
$C = AB$,
где A – матрица – первый сомножитель,
B – матрица – второй сомножитель,
C – матрица – произведение.

Произведение матриц ведется по формуле (в поэлементной записи):

$$c_{ij} = \sum_{k=1}^{m} a_{ik} b_{kj}, i = 1,2,...,n;\ j = 1,2,...,l,$$

где $a_{..}$ – элемент матрицы – первого сомножителя,
$b_{..}$ – элемент матрицы – второго сомножителя,
$c_{..}$ – элемент матрицы – произведения,
n – число строк в первом сомножителе,
m – число столбцов в первом сомножителе и строк во втором,
l – число столбцов во втором сомножителе.

24.3.4. Обратная матрица

Пусть исходная квадратная матрица A имеет отличный от нуля определитель. Тогда существует матрица A^{-1} (или просто A^{-}), такая, что
$A^{-1}A = A\,A^{-1} = I$,
где I – единичная матрица, то есть матрица, имеющая единицы на главной диагонали и нули на месте остальных элементов.
Квадратная матрица A^{-1} называется матрицей, обратной данной матрице A. Критерий обратимости матрицы дает ее определитель: если он равен нулю, матрица вырождена, если не равен, то матрица обратима.

24.3.5. Определитель матрицы

Определителем (детерминантом) матрицы называется многочлен от элементов квадратной матрицы A порядка n, каждый член которого является произведением n элементов, взятых по одному из каждого столбца и каждой строки, и снабжен определенным знаком: плюсом, если перестановка четна, и минусом, если перестановка нечетна. Число n называется порядком определителя.
По определению, детерминант может быть разложен по элементам любого столбца (строки):

$$\det A = \sum_{j=1}^{n} a_{jk} A_{jk}, k = 1,2,...,n$$

,
где k – номер столбца (строки),
$A_{jk} = (-1)^{j+k} M_{jk}$ – алгебраическое дополнение элемента a_{jk},
M_{jk} – минор элемента a_{jk}, то есть определитель порядка $n - 1$.
Минор получается при вычеркивании из исходной матрицы j–й строки и k–го столбца. При $j = k$ миноры называются главными.
Определитель является одной из важнейших характеристик квадратной матрицы, определяющей ее поведение в различных алгоритмах. Определители находят и самостоятельное применение, например, при решении систем линейных уравнений методом Крамера. Данный примитивный метод часто рассматривается в учебных курсах высшей математики, но в практических вычислениях не используется из–за низкой эффективности.
Показанная выше формула называется рекурсивным определителем детерминанта. Однако на практике, особенно с развитием средств вычислительной техники, данная формула не применяется вследствие ее низкого быстродействия и тенденции накопления ошибки вычислений. Практически вычисление детерминанта производится на основе того факта, что определитель вычисляется через продукты некоторых видов факторизации (разложения) матриц, например, разложения Холецкого и разложения Краута.

24.3.6. Умножение матрицы на скаляр

С помощью данной операции производится умножение каждого элемента матрицы на заданную скалярную величину.
Умножение матрицы на скаляр в матричной записи записывается как:
$C = kA$,
где k – скаляр – первый сомножитель,
A – исходная матрица – второй сомножитель,
C – матрица – произведение.
Умножение матрицы на скаляр ведется по формуле (в поэлементной записи):
$c_{ij} = ka_{ij}$, $i = 1,2,...,n$; $j = 1,2,...,l$,
где k – значение скаляра – первого сомножителя,
$a_{..}$ – элемент исходной матрицы – второго сомножителя,
$c_{..}$ – элемент матрицы – произведения,
n – число строк в каждой матрице,
m – число столбцов в каждой матрице.
Операция находит применение в конструировании различных алгоритмов. Так, например, с использованием алгоритмов сложения матриц и умножения матрицы на скаляр, в данном случае на –1, может быть получен алгоритм вычитания матриц.

24.3.7. Псевдообратная матрица

Для действительной прямоугольной матрицы A размера m x n, где m – число строк, n – число столбцов, вводится понятие псевдообратной матрицы A^+, то есть такой матрицы размера n x m, что имеют место свойства
$AA^+A = A$,
$A^+AA^+ = A^+$,
$(AA^+)^T = AA^+$,
$(A^+A)^T = A^+A$,
Процедура вычисления псевдообратной матрицы (называемая также обращением Мура–Пенроуза) основана на функции разложения прямоугольной матрицы по сингулярным числам, алгоритм которой требует выполнения условия $m \geq n$.

Обсуждение см. в источниках: Голуб (Golub) с соавт., Уилкинсон, Дэннис с соавт., Корн с соавт., Магнус.

24.3.8. Решение системы линейных уравнений

Методы решения систем линейных алгебраических уравнений представлены здесь методом исключения Гаусса с выбором ведущего (главного) элемента. Система линейных алгебраических уравнений в матричной записи выглядит как
$Ax = b$,
где A – матрица системы,
x – вектор подлежащих определению неизвестных,
b – столбец свободных членов.
Решение возможно, если матрица системы не вырождена.
Предполагается, что система является определенной, т. е. число уравнений равно числу неизвестных (иначе матрица A – квадратная). О решении неопределенных (число уравнений меньше числа неизвестных) и переопределенных (число уравнений больше числа неизвестных) систем см. главу «Распознавание образов с обучением».
Функция программы представляет собой расширенную процедуру по сравнению с

рассмотренной обычной системой линейных уравнений. Часто возникает необходимость (см., например, задачу вычисления обратной матрицы) решить так называемую матричную систему линейных уравнений
$AX = B$,
где вектор неизвестных и столбец свободных членов обобщаются до матриц, соответственно, X и B.
Фактически последняя формула включает в себя несколько систем линейных уравнений, объединенных общей матрицей системы уравнений A, но имеющих различные столбцы свободных членов, составляющие матрицу B. В этом случае эффективнее решать систему с помощью одной универсальной функции. Манипуляции с матрицей системы, те же самые для каждого столбца свободных членов, выгоднее производить только один раз.
Естественно, и обычные системы, когда в правой части системы стоит столбец свободных членов, решаются с помощью данной функции.

24.3.9. Стандартная проблема собственных значений

Рассмотрим стандартную проблему собственных значений
$Ax_i = \lambda_i x_i$, $i = 1,2,...,n$,
где A – квадратная матрица размером n x n,
λ_i, $i = 1,2,...,n$ – собственные значения матрицы A,
x_i, $i = 1,2,...,n$ – соответствующие собственным значениям λ_i, $i = 1,2,...,n$, вычисленные с точностью до множителя собственные вектора матрицы A.
Свойства собственных значений подробно представлены в обширной литературе и рассматриваются в вузовском курсе линейной алгебры (напомним, что линейную алгебру составляют два основных раздела: решение систем линейных уравнений и решение проблемы собственных значений).
В программе решается стандартная проблема собственных значений симметрической (симметричной относительно главной диагонали) действительной матрицы.
Симметричность матрицы гарантирует, что вычисленные собственные значения будут действительными. Программа построена таким образом, что в качестве исходных данных при вычислении используется только верхний треугольник матрицы. В связи с этим проверка симметричности не производится, а правильным может быть только верхний треугольник матрицы. Вычисляются все собственные значения и соответствующие им собственные вектора матрицы. Собственные значения выводятся в неупорядоченном виде. Расположение выводимых собственных векторов соответствует порядку собственных значений. Для решения используется метод Якоби. Программа выдает также количество затраченных для решения итераций.

См. источники: Уилкинсон, Гайдышев, Корн с соавт.

24.3.10. Обобщенная проблема собственных значений

Рассмотрим обобщенную проблему собственных значений
$Ax_i = \lambda_i Bx_i$, $i = 1,2,...,n$,
где A – симметрическая матрица размером n x n,
B – симметрическая положительно определенная матрица размером n x n,
λ_i, $i = 1,2,...,n$ – собственные значения,
x_i, $i = 1,2,...,n$ – соответствующие собственным значениям λ_i, $i = 1,2,...,n$, вычисленные с точностью до множителя собственные вектора.
Обобщенная проблема собственных значений, путем замены переменных и используя

разложение Холецкого, приводится к стандартной проблеме собственных значений.

См. источники: Уилкинсон, Гайдышев, Корн с соавт.

24.3.11. Разложение Холецкого

Разложение Холецкого (схема Холецкого, схема квадратного корня) основано на теореме Холецкого, а именно: если A – симметрическая положительно определенная матрица, то существует разложение
$A = LL^{\mathrm{T}}$,
где L – действительная невырожденная нижняя треугольная матрица.
Рассматриваемое разложение получается из разложения
$A = LDL^{\mathrm{T}}$,
где L – нижняя треугольная матрица с единичной диагональю,
D – диагональная матрица с положительными диагональными элементами,
как
$A = LD^{1/2}D^{1/2}L^{T} = \overline{L}\,\overline{L}^{T}$,
что с точностью до обозначений совпадает с показанной выше формулой.
Этот эффективный вид разложения может применяться в ряде процедур линейной алгебры, например, при решении задачи приведения обобщенной проблемы собственных значений к стандартной проблеме собственных значений и для решения систем линейных уравнений.

Различные аспекты схемы Холецкого и ее применения рассмотрены Мэйндоналдом.
Разложение может быть успешно использовано для вычисления определителя и в задаче генерации многомерного нормального распределения (Мюллер с соавт.).
См. источники: Уилкинсон с соавт., Математический энциклопедический словарь, Гилл с соавт. О вычислении разложения см. Сборник научных программ на Фортране.

24.3.12. Разложение Краута

Если A – неособая матрица, то существует разложение:
$A = LU$,
где L – нижняя треугольная матрица,
U – верхняя треугольная матрица с единичной диагональю.
Разложение не единственно. Если его записать как
$LU = (LD)(D^{-1}U)$,
где $D^{-1}U$ есть верхняя треугольная матрица с единичной диагональю, мы получим разложение Краута.
Приведем пример применения рассматриваемого разложения в задаче решения системы линейных уравнений:
$Ax = b$.
Вычислив матрицы L и U, сводим задачу к двум элементарно решаемым системам с треугольными матрицами
$Ly = b$ и $Ux = y$.
Первая из двух последних формул называется прямым исключением, а вторая – обратной подстановкой, что полностью соответствует вычислительной схеме метода Гаусса.

Эквивалентность метода исключения Гаусса и рассматриваемого типа разложения подробно обсуждается Стренгом: рассмотренное разложение представляет собой просто удобную запись метода исключения Гаусса. См. также Корна с соавт., Уилкинсона с соавт., Форсайта с

соавт. О вычислении разложения см. Сборник научных программ на Фортране.

24.3.13. Разложение QR

В некоторых приложениях применяется разложение произвольной невырожденной матрицы *A*, с помощью процесса ортогонализации Грама–Шмидта, следующего вида:
$A = QR$,
где Q – ортогональная матрица,
R – верхняя треугольная матрица.
Данное QR разложение эквивалентно построению ортонормированной системы векторов в гильбертовом пространстве, порождающей то же самое линейное многообразие, что и заданная система. Процесс Грама–Шмидта может быть истолкован как разложение невырожденной матрицы в произведение ортогональной матрицы и верхней треугольной матрицы с положительными диагональными элементами. Решение ведется по рекуррентным формулам

$$u_i = \frac{v_i}{\|v_i\|}, v_1 = e_1, v_{i+1} = e_{i+1} - \sum_{k=1}^{i} (u_k, e_{i+1}) u_k, i = 1,2,...$$

,
где u – ортонормированная система векторов,
e – заданная система векторов,
n – число векторов.
Нетрудно видеть, что вычисленные таким образом векторы n представляют собой столбцы ортогональной матрицы Q.

См. источники: Корн с соавт., Математический энциклопедический словарь, Стренг.

24.3.14. Разложение по сингулярным числам

Рассмотрим разложение действительной прямоугольной матрицы A размером m строк на n столбцов, причем $m \geq n$, вида

$$A = U \begin{pmatrix} S \\ 0 \end{pmatrix} V^T,$$

где U – матрица размером m x m , сформированная из m ортонормированных собственных векторов, соответствующих собственным значениям матрицы AA^T, $U^TU = I_m$,,
V – матрица размером n x n, состоящая из n ортонормированных собственных векторов матрицы A^TA и обладающая свойствами $V^TV = VV^T = I_n$,
S – диагональная матрица, диагональные элементы которой представляют собой так называемые сингулярные числа – квадратные корни из неотрицательных собственных значений матрицы A^TA,
0 – прямоугольная нулевая матрица размером $m – n$ строк на n столбцов,
$I_{\cdot}$ – единичная матрица соответствующего порядка.
Рассмотренное разложение называется разложением по сингулярным числам (сингулярным разложением). В другой записи, применяемой в настоящем программном обеспечении, рассмотренное разложение может быть записано как
$A = U_nSV^T$,
где U_n – матрица размером m x n, далее для простоты обозначаемая, как U, и сформированная из n ортонормированных собственных векторов, соответствующих n наибольшим из m собственных значений матрицы AA^T, и обладающая свойствами
$U^TU = V^TV = VV^T = I_n$.

Программой выводятся матрицы *U*, *S* и *V*.

См. источники: Гайдышев, Голуб (Golub) с соавт., Дэннис с соавт., Стренг, Уилкинсон с соавт.

24.3.15 Мультиколлинеарность

Вектора называются коллинеарными, если они лежат на параллельных прямых либо на одной прямой. Понятие коллинеарности, пришедшее из аналитической геометрии, тождественно линейной зависимости из линейной алгебры (линейная зависимость – также критерий мультиколлинеарности). Необходимость исследования коллинеарности возникает, например, перед применением ряда методов многомерного статистического анализа (множественная регрессия, факторный анализ) с целью исключения из рассмотрения линейно зависимых параметров. Это необходимо как для уменьшения размерности задачи, так и для снижения вычислительных сложностей.
Фаррар (Farrar) и Глаубер (Glauber) предложили совокупность статистических методов определения наличия мультиколлинеарности, известную под наименованием алгоритма Фаррара–Глаубера.
Пусть обозначено:
n – число строк в матрице исходных данных (число наблюдений),
m – число столбцов в матрице исходных данных (число векторов, в анализе данных – число параметров, в эконометрике – число объясняющих переменных),
R – корреляционная матрица (о ее вычислении см. главу «Корреляционный анализ»),
C – матрица, обратная корреляционной.
Алгоритм Фаррара–Глаубера статистически исследует проявления мультиколлинеарности, представленные далее.

24.3.15.1. Корреляция между параметрами

Вычисляется статистика критерия
$S = -[n - 1 - (2m + 5) / 6] \ln|R|$,
где |.| – определитель.
Статистика имеет распределение хи–квадрат с $m(m - 1) / 2$ степенями свободы. При значимой статистике хи–квадрат есть основание предположить наличие явления мультиколлинеарности в исследуемой системе векторов.

24.3.15.2. Коэффициенты детерминации векторов

Для каждого вектора вычисляется коэффициент детерминации
$R_k^2 = 1 - 1/c_{kk}, k = 1,2,..., m,$
где $c_{kk}, k = 1,2,..., m,$ – диагональный элемент матрицы *C*.
Также для каждого вектора вычисляется статистика

$$F_k = (c_{kk} - 1)\frac{n-m}{m-1}, k = 1,2,..., m.$$

Статистики $F_k, k = 1,2,..., m,$ подчиняются *F*–распределению со степенями свободы $n - m$ и $m - 1$ При значимых *F*–статистиках есть основание предположить коллинеарность данного вектора с некоторыми или всеми остальными векторами. Такие векторы (в задаче распознавания образов соответствующие параметрам распознавания) следует исключить из матрицы исходных данных.

24.3.15.3. Частные коэффициенты корреляции

Для каждого внедиагонального элемента корреляционной матрицы (в силу симметрии исследуется только верхняя часть) вычисляется частный коэффициент корреляции

$$r_{kj} = -\frac{c_{kj}}{\sqrt{c_{kk}c_{jj}}}, k = 1,2,...,m-1; j = k+1,...,m.$$

Также для каждого частного коэффициента корреляции вычисляется статистика

$$t_{kj} = \frac{r_{kj}\sqrt{n-m}}{\sqrt{1-r_{kj}^2}}, k = 1,2,...,m-1; j = k+1,...,m.$$

Статистики $t_{kj}, k = 1,2,...,m-1; j = k+1,...,m,$ подчиняются *t*–распределению с $n - m$ степенями свободы. При значимых *t*–статистиках есть основание предположить коллинеарность в исследуемой паре векторов *k* и *j*.

Настоящее программное обеспечение выдает результаты последовательного расчета всеми указанными методами.

Обзор методов исследования мультиколлинеарности см. в монографиях Ферстера с соавт., Айвазяна с соавт. См. также оригинальные статьи Фаррара с соавт., Карнеса (Carnes) с соавт., Уишерса (Wichers). Метод Фаррара–Глаубера представлен в книгах Лещинського с соавт., Наконечнего с соавт. См. статью Рокуэлла (Rockwell), книгу Бородича (включая вывод основных формул и рекомендательные меры по устранению мультиколлинеарности).

24.3.16. Кронекеровское произведение

Кронекеровским произведением матриц $A = (a_{ij})$ размером *m* x *n* и $B = (b_{st})$ размером *p* x *q* называется матрица $C = A \otimes B$ размером *mp* x *nq* такая, что

$$\begin{pmatrix} a_{11}B & \cdots & a_{1n}B \\ \vdots & & \vdots \\ a_{m1}B & \cdots & a_{mn}B \end{pmatrix}.$$

См. монографию Магнуса.

Список использованной и рекомендуемой литературы

1. Carnes B.A., Slade N.A. The use of regression for detecting competition with multicollinear data // Ecology, August 1988, vol. 69, no. 4, pp. 1266–1274.
2. Engineering and Scientific Subroutine Library for Linux on POWER, Version 4, Release 2. – International Business Machines Corporation, September 2003.
3. Farrar D.E., Glauber R.R. Multicollinearity in regression analysis: the problem revisited // Review of Economics and Statistics, 1967, vol. 49, pp. 92–107.
4. Galassi M. GNU scientific library reference manual / M. Galassi, J. Davies, J. Theiler et al. – Boston, MA: Network Theory, 2005.
5. Golub G., Kahan W. Calculating the singular values and pseudo–inverse of a matrix // SIAM Journal on Numerical Analysis, Series B, 1965, vol. 2, no. 2, pp. 205–223.
6. Higham N.J. Accuracy and stability of numerical algorithms. – Philadelphia, PA: Society for Industrial and Applied Mathematics, 1996.

7. IMSL C Numerical Library. User's Guide. Version 5.5. Volume 1 of 4: C Math Library (Chapters 1–7). – Visual Numerics, 2003.
8. Kahaner D.K., Moler C., Nash S.G. Numerical methods and software. – Upper Saddle River, NJ: Prentice Hall, 1989.
9. Nash J.C. Compact numerical methods for computers. Linear algebra and function minimization. – New York, NY: Adam Hilger, 1990.
10. Nash J.C., Nash M.M. Scientific computing with PCs. – Ottawa: Nash Information Services, 1993.
11. Press W.H. Numerical recipes in C. The art of scientific computing / W.H. Press, S.A. Teukolsky, W.T. Vetterling et al. – Cambridge, UK: Cambridge University Press, 2002.
12. Rockwell R.C. Assessment of multicollinearity: The Haitovsky test of the determinant // Sociological Methods & Research, 1975, vol. 3, no. 3, pp. 308–320.
13. The NAG Fortran Library Manual, Mark 21. – Oxford, UK: The Numerical Algorithms Group, 2004.
14. Wichers C.R. The detection of multicollinearity: A comment // The Review of Economics and Statistics, August 1975, vol. 57, no. 3, pp. 366–368.
15. Айвазян С.А., Мхитарян В.С. Прикладная статистика и основы эконометрики. – М.: ЮНИТИ, 1998.
16. Беллман Р. Введение в теорию матриц. – М.: Наука, 1976.
17. Бородич С.А. Вводный курс эконометрики: Учебное пособие. – Мн.: БГУ, 2000.
18. Брандт З. Анализ данных. Статистические и инженерные методы для научных работников и инженеров. – М.: Мир, ООО «Издательство АСТ», 2003.
19. Бронштейн И.Н., Семендяев К.А. Справочник по математике. – М.: Наука, 1986.
20. Вержбицкий В.М. Численные методы (линейная алгебра и нелинейные уравнения). – М.: Высшая школа, 2000.
21. Гайдышев И. Анализ и обработка данных: специальный справочник. – СПб: Питер, 2001.
22. Гантмахер Ф.Р. Теория матриц. – М.: Наука, 1988.
23. Гилл Ф., Мюррей У., Райт М. Практическая оптимизация. – М.: Мир, 1985.
24. Голуб Дж., Ван Лоун Ч. Матричные вычисления. – М.: Мир, 1999.
25. Гроссман С., Тернер Дж. Математика для биологов. – М.: Высшая школа, 1983.
26. Деммель Дж. Вычислительная линейная алгебра. Теория и приложения. – М.: Мир, 2001.
27. Дэннис Дж., мл., Шнабель Р. Численные методы безусловной оптимизации и решения нелинейных уравнений. – М.: Мир, 1999.
28. Каханер Д., Моулер К., Нэш С. Численные методы и программное обеспечение. – М.: Мир, 2001.
29. Корн Г., Корн Т. Справочник по математике для научных работников и инженеров. – М.: Наука, 1973.
30. Ланкастер П. Теория матриц. – М.: Наука, 1973.
31. Лещинський О.Л., Рязанцева В.В., Юнькова О.О. Економетрія: Навч. посіб. для студ. вищ. навч. закл. – Киев: МАУП, 2003.
32. Лоусон Ч., Хенсон Р. Численное решение задач метода наименьших квадратов. – М.: Наука, 1986.
33. Магнус Я.Р. Матричное дифференциальное исчисление с приложениями к статистике и эконометрике. – М.: Физматлит, 2002.
34. Мак–Кракен Д., Дорн У. Численные методы и программирование на ФОРТРАНе. – М.: Мир, 1977.
35. Мэйндоналд Дж. Вычислительные алгоритмы в прикладной статистике. – М.:

Финансы и статистика, 1988.
36. Мюллер П., Нойман П., Шторм Р. Таблицы по математической статистике. – М.: Финансы и статистика, 1982.
37. Наконечний С.І., Терещенко Т.О., Романюк Т.П. Економетрія: Навчальний посібник. – К.: КНЕУ, 1998.
38. Наконечний С.І., Терещенко Т.О., Романюк Т.П. Економетрія: Підручник. – К.: КНЕУ, 2004.
39. Прохоров Ю.В. Математический энциклопедический словарь / Гл. ред. Ю.В. Прохоров. – М.: Научное издательство «Большая Российская энциклопедия», 1995.
40. Сборник научных программ на Фортране. Выпуск 2. Матричная алгебра и линейная алгебра. – М.: Статистика, 1974.
41. Стренг Г. Линейная алгебра и ее применения. – М.: Мир, 1980.
42. Уилкинсон Дж. Х. Алгебраическая проблема собственных значений. – М.: Наука, 1970.
43. Уилкинсон, Райнш. Справочник алгоритмов на языке АЛГОЛ. Линейная алгебра. – М.: Машиностроение, 1976.
44. Ферстер Э., Ренц Б. Методы корреляционного и регрессионного анализа. Руководство для экономистов. – М.: Финансы и статистика, 1983.
45. Форсайт Дж., Молер К. Численное решение систем линейных алгебраических уравнений. – М.: Мир, 1969.

Глава 25. Обыкновенные дифференциальные уравнения

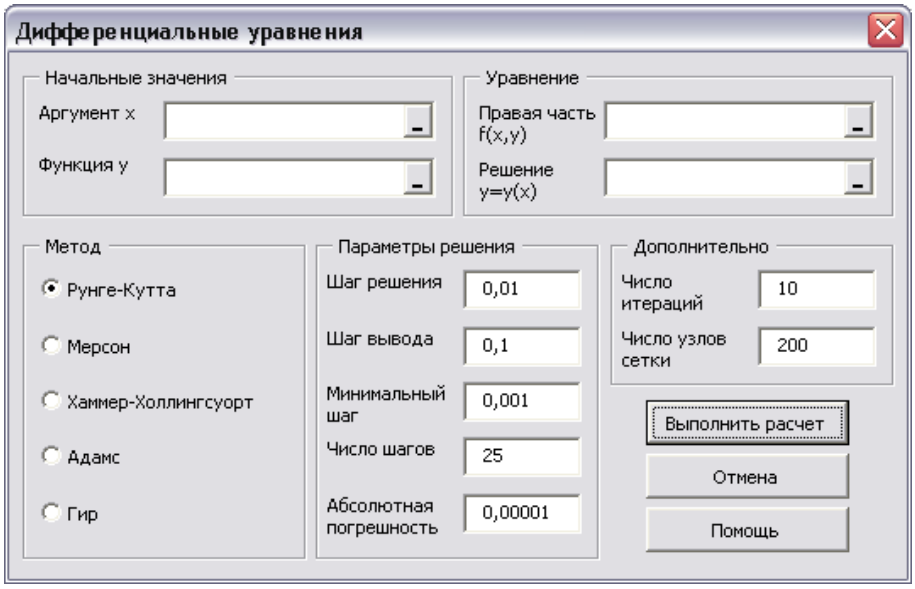
25.1. Введение

Программное обеспечение предназначено для численного решения обыкновенных дифференциальных уравнений различными методами, представленными в программе. Методы отражают многообразие подходов к решению проблемы и предназначены для решения уравнений соответствующих типов.

25.2. Работа с программным обеспечением

Выберите из меню программы пункт AtteStat | Дифференциальные уравнения. На экране появится диалоговое окно, изображенное на рисунке.
Затем:

- Выберите или введите ячейку, содержащую начальное значение аргумента.
- Выберите или введите ячейку, содержащую начальное значение функции.

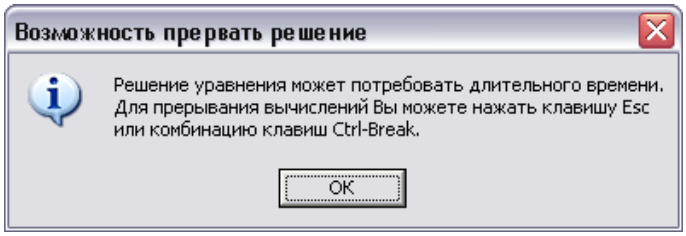


- Выберите или введите ячейку, содержащую формулу правой части уравнения. Контроль допустимости введенной правой части и ее соответствие правилам записи формул осуществляется стандартным образом самими электронными таблицами.
- Выберите или введите выходной интервал. Начиная с первой ячейки выходного интервала (следовательно, можно указать только одну ячейку, т.к. остальные ячейки интервала игнорируются), будут выведены результаты вычислений.
- Выберите или оставьте по умолчанию метод интегрирования дифференциального уравнения.
- Выберите или оставьте по умолчанию параметры решения. Обратите внимания, что абсолютная погрешность действительна только для методов с автоматическим контролем точности.
- Выберите или оставьте по умолчанию дополнительные параметры решения. Обратите внимание, что число итераций действительно только для методов, имеющих в схемах алгоритмов итерационные фрагменты, а число узлов сетки действительно только для многошаговых методов.
- Нажмите кнопку «Выполнить расчет».

После выполнения вычислений будет, начиная с первой ячейки выходного интервала, выведено название метода и результаты расчета.

За выбор адекватного исходным данным метода расчета несет ответственность пользователь. Программное обеспечение берет на себя верификацию исходных данных, выдавая подробную диагностику. При неверных действиях пользователя выдаются сообщения об ошибках. Обратим внимание, что не все сочетания параметров допустимы. Если встречается недопустимое сочетание параметров, программное обеспечение, как правило, само принимает решение о выполнении расчета.

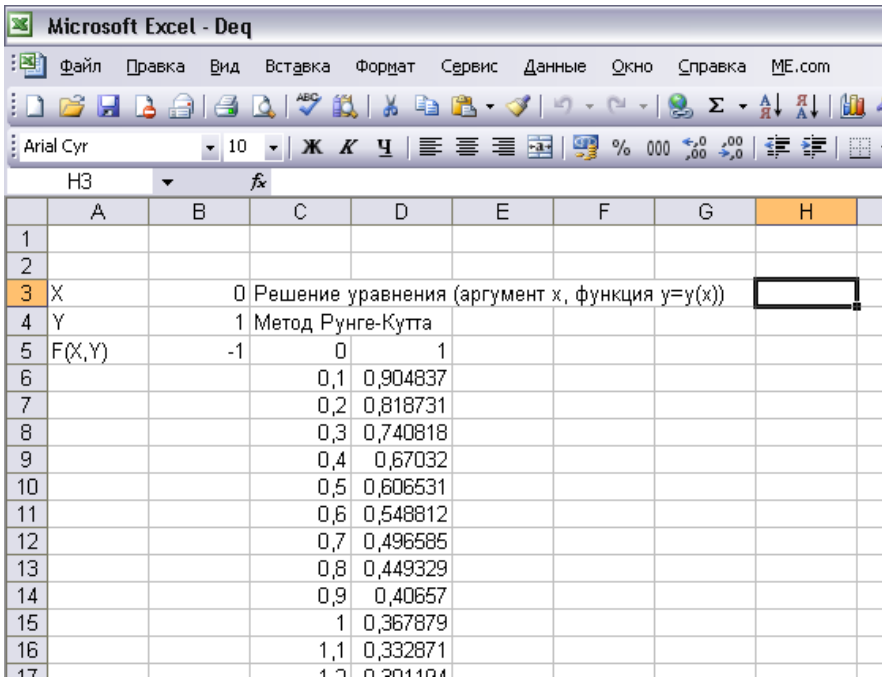
Время решения даже сравнительно простых по структуре дифференциальных уравнений различными методами может быть длительным и сильно зависеть от производительности компьютерной системы, поэтому в программу заложена возможность прерывания решения по желанию пользователя до нормального окончания с заданными параметрами. О данной возможности пользователю сообщается в специальном информационном окне, показанном на рисунке, перед любым производством самого решения.



Для начала решения следует нажать кнопку OK.

25.2.1. Пример применения

Следующий рисунок показывает, каким образом вводится формула правой части. Так, в ячейку **B3** введено начальное значение аргумента X, в ячейку **B4** введено начальное значение функции Y. В ячейку **B5** введено выражение правой части решенного в примере дифференциального уравнения $Y'=-Y$, а именно, =–B4. Легко заметить, что в ячейку формулы правой части в качестве значений X и Y вводятся адреса соответствующих ячеек. Пример ввода правой части дифференциального уравнения показан на рисунке.



	A	B	C	D	E	F	G	H
1								
2								
3	X	0	Решение уравнения (аргумент x, функция y=y(x))					
4	Y	1	Метод Рунге-Кутта					
5	F(X,Y)	-1	0	1				
6			0,1	0,904837				
7			0,2	0,818731				
8			0,3	0,740818				
9			0,4	0,67032				
10			0,5	0,606531				
11			0,6	0,548812				
12			0,7	0,496585				
13			0,8	0,449329				
14			0,9	0,40657				
15			1	0,367879				
16			1,1	0,332871				

25.2.2. Сообщения об ошибках

При ошибках ввода данных и во время выполнения программы могут выдаваться диагностические сообщения следующих типов:

Ошибка	Комментарий
Не определена ячейка.	Вы не выбрали или неверно ввели интервалы ячеек, определяющих начальные значения или начало области вывода результатов расчета. Лучшим способом избежать ошибки является не ввод, а выделение интервала ячеек стандартным образом.

Ошибка	Комментарий
Неверно задана формула.	Проверьте содержимое ячейки рабочего листа, заданное Вами как формула. Запись формулы должна соответствовать принятым правилам.
Недопустимая комбинация значений шага.	Пользователь задал минимальный шаг, превышающий шаг решения, или шаг вывода, меньший шага решения. Допустимые сочетания параметров установлены в полях формы по умолчанию.
Не задан параметр.	Не задан один из требуемых параметров решения. Допустимые значения параметров установлены в полях формы по умолчанию. Пользователю, не полностью уверенному в своих действиях, следует оставить параметры, установленные программой по умолчанию.
Ошибка вычисления.	Данные ошибки обычно являются отражением специфических свойств решаемых уравнений. В процессе решения дифференциального уравнения нужно быть готовым к тому, что иногда решение получить не удается.

25.3. Теоретическое обоснование

Порядком дифференциального уравнения называют наибольший порядок входящих в него производных. Дифференциальное уравнение первого порядка называется также обыкновенным дифференциальным уравнением и может быть записано как
$F(x,y,y') = 0$,
где y – скалярная или s–компонентная векторная функция, подлежащая определению, в механике называемая вектором обобщенных координат,
x – независимая переменная, при исследовании динамики протяженных во времени процессов в ее качестве обычно берется время t,
штрих означает полную производную по независимой переменной x и может быть заменен стандартным обозначением d / dx.
Если в явном виде независимая переменная в уравнения не входит, система уравнений называется автономной. Если система уравнений не автономна, то можно выбрать в качестве времени новую независимую переменную τ, а затем, добавив к системе еще одно уравнение
$d\tau / dt = 1$,
превратить исходную систему уравнений в систему автономную, хотя явной выгоды от данного процесса можно и не ощутить. Выгода может быть получена при вычислениях, если составлять все алгоритмы только для автономных систем уравнений. Это несколько упростит интерфейс составленных функций. Можно также заранее предусмотреть автоматическое преобразование неавтономной системы уравнений к автономной системе уравнений по рассмотренному выше простому алгоритму.
Степенью дифференциального уравнения называют высший показатель степени, в которой старшая производная входит в уравнение после того, как уравнение приведено к рациональному виду. Если в уравнение искомая функция и все ее производные входят в первой степени, уравнение называется линейным, в противном случае уравнение будет нелинейным.
Если исходное уравнение разрешимо относительно производной, то получается уравнение
$y' = f(x,y)$,
где $f(x,y)$ – функция правых частей (правая часть).
Общее решение дифференциального уравнения определяется с точностью до постоянных C,

определяющих семейство кривых
$y = y(x) + C$.
При задании начального условия и в случае выполнения перечисленных требований решение существует и единственно. Задача Коши формулируется как аналитическое или численное решение системы s обыкновенных дифференциальных уравнений, в общем случае нелинейных (с нелинейной функцией правых частей), на интервале $x \in [a,b]$ с начальными условиями
$y(x_0) = y_0$,
где $x_0 = a$.
Иногда, чтобы отличить одно дифференциальное уравнение от системы дифференциальных уравнений, первое называют скалярным дифференциальным уравнением. Уравнение высшего порядка может быть сведено к системе уравнений первого порядка простым изменением обозначений и введением новых переменных. Размерность системы – это число уравнений, содержащих производные порядка не выше первого. Все выводы, сделанные относительно одного дифференциального уравнения, естественно обобщаются на систему, поэтому в теоретических выкладках не делается разницы, является y скалярной или s–компонентной векторной функцией. Все показанные формулы могут быть использованы как для исследования одного уравнения, так и для исследования системы уравнений. В последнем случае можно просто считать, что величины, входящие в формулы, представляют собой не скаляры, а векторы.
Везде приводятся функции решения только одного уравнения, однако нетрудно обобщить их на решение системы уравнений. Для этого потребуется модернизация приведенных функций. Чтобы научить их решать системы уравнений, следует заменить векторами (массивами) функции и входящие в формулы скалярные параметры. Кроме того, там, где производится проверка точности, нужно в цикле производить проверку каждого элемента вектора (или их совокупность), ответственного за точность решения.
Единственная особенность системы уравнений относительно одного уравнения заключается в том, что принятие решения в случае недостаточной точности следует делать сразу же, как только один из элементов вектора даст для этого повод. Принятие же решения в случае чрезмерной точности делать только после того, как проверки всех элементов покажут, что точность чрезмерна (см. второй вариант метода Мерсона). Для систем, соответственно, должны быть изменены также краевые – начальные и граничные – условия задачи.
В заключение раздела отметим один научный факт, который будет использован в дальнейших рассуждениях. Система нелинейных дифференциальных уравнений может быть представлена в виде линеаризованной системы, которая получена с использованием только первых двух членов разложения решения в ряд Тейлора
$y' = J(x)y + (**)$,
где J – [локально] постоянная матрица Якоби системы – матрица частных производных правых частей $\partial f / \partial y$,
(**) – нелинейные члены.

25.3.1. Математическое моделирование

Математическими формулами можно описать, но нельзя объяснить физическую картину явлений. Наличие формул, описывающих наблюдаемое явление, не компенсирует отсутствия точных знаний или хотя бы гипотез о причине явления.
Реальные объекты или явления представляют собой сущности с бесконечно богатым содержанием. Более того, они самостоятельно или в совокупности с другими объектами или явлениями способны порождать новые сущности. В математической модели переменные и

константы моделируют реальные сущности, но не могут самостоятельно их порождать, представляя собой только символы на бумаге или на экране компьютера. При попытках более детального исследования или получении новых экспериментальных данных приходится создавать новые математические модели или совершенствовать старые. Поэтому главное в математическом моделировании – выделение свойств исследуемых объектов или явлений, существенных для решаемой в данный момент конкретной практической задачи. «Дать точное описание наблюдавшихся явлений природы, выхватить из многообразия деталей и мелочей главные, характерные черты, в резкой и краткой форме сформулировать все, что видел глаз и охватила мысль – это настолько сложная и важная задача, что перед ней бледнеют все трудности лабораторного исследования или теоретического анализа в кабинетах ученых» (А.Е. Ферсман).

Пытаясь описать всевозможные характеристики явления, до которых исследователь только смог добраться, он рискует запутаться в частностях и не описать их вовсе. Только опыт и здравый смысл помогут отличить влияние существенной части влияния на исследуемый процесс от совокупного влияния малозначительных черт. «Искусство быть мудрым состоит в умении знать, на что не следует обращать внимания» (У. Джеймс). Истина не может быть сложной для понимания – в противном случае не может быть даже установлено, что есть истина. Разумный уровень абстрактности позволит добиться успеха при исследовании даже очень сложных явлений.

Сказанное в предыдущем абзаце не означает, что исследование математических моделей позволяет только описать и наглядно отобразить явления, как это делают методы описательной статистики. Наоборот, исследование корректно составленных моделей позволяет выявить совершенно новые, часто неподдающиеся прямому исследованию в эксперименте, характеристики явлений, дополняя и даже иногда частично заменяя экспериментальные методы исследований, и породить новые неожиданные идеи.

В математическом моделировании, когда модель записана в виде системы обыкновенных дифференциальных уравнений, искомая функция *y* часто называется выходом модели в противоположность тому, что ее экспериментальный «аналог» называется выходом эксперимента. Выход модели в математическом моделировании может также представлять собой ту или иную, в том числе нелинейную, комбинацию всех или некоторых из *s* компонент функции *y* либо только наблюдаемую часть компонент векторной функции *y*, если по условиям эксперимента наблюдение всех компонент затруднительно.

Дадим несколько определений, которые могут пригодиться при математическом моделировании и исследовании математических моделей, в том числе на предельных (критических) режимах.

Детерминированная система – такая система, для которой существует правило в виде дифференциальных или разностных уравнений, определяющее ее будущее поведение, исходя из заданных начальных условий. Противоположностью детерминированной системы является система вероятностная (статистическая).

Выбор адекватной математической модели (детерминированной или вероятностной) может быть сделан на основе информационного анализа изучаемой физической системы.

Потенциальной называется сила, работа которой зависит только от начального и конечного положения точки ее приложения и не зависит от вида траектории и закона движения точки.

Гамильтонов подход к описанию динамики физических систем основан на системе обыкновенных дифференциальных уравнений

$$\dot{q}_i = \partial H / \partial p_i, \dot{p}_i = \partial H / \partial q_i, i = 1,2,...,n,$$

где точка означает полную производную по времени,

$q_i, p_i, i = 1,2,...,n,$ – соответственно, обобщенные координаты и обобщенные импульсы, их

совокупность называется каноническими переменными,
n – число степеней свободы (число независимых обобщенных координат),
$H = H(q,p)$– функция Гамильтона (гамильтониан), характеризующая физическое состояние системы.

Задача решается с начальными условиями при $t = t_0$
$p_i(t_0) = p_i^0, q_i(t_0) = q_i^0$.
Если система автономна и действующие силы потенциальны, гамильтониан является полной энергией системы, выраженной через канонические переменные.
Решение представленной выше системы можно представить как движение точки в $2n$ – мерном пространстве с координатами q и p. Такое пространство называется фазовым, его точки – фазовыми точками, а траектории движения фазовых точек – фазовыми траекториями (фазовыми кривыми). Совокупность фазовых кривых называют потоком.
Консервативная система – система, для которой имеет место закон сохранения энергии. Другие законы сохранения (например, количества движения), могут не соблюдаться. Для консервативной системы элемент фазового пространства изменяет форму, но сохраняет объем. Примером консервативной системы является Солнечная система.
Для Гамильтоновых систем траектории в фазовом пространстве не пересекаются. Гамильтоновы системы консервативны. Большинство изучаемых систем не являются Гамильтоновыми.
Диссипативная система – система, полная механическая энергия которой (кинетическая плюс потенциальная) при движении убывает (рассеивается), переходя в другие формы энергии, например, в энергию теплового хаотического движения молекул. К диссипативным системам относится большинство изучаемых систем. Примерами диссипативных систем являются механические системы с трением, движение вязких жидкостей. Для диссипативной системы объем элемента фазового пространства сокращается (фазовый элемент сжимается) с течением времени. Сокращение фазового объема приводит к тому, что при $t \to \infty$, где t – параметр (время), все решения диссипативной системы будут стягиваться к некоторому подмножеству фазового пространства, называемому аттрактором.
Неконсервативная система может не быть диссипативной, если в ней рассеяние энергии компенсируется притоком энергии извне, хотя многие авторы не делают таких тонких отличий. Так, в одном литературном источнике утверждается, что в диссипативной системе переход к хаосу возможен якобы только при внешнем возбуждении (подводе энергии в открытую систему извне).

25.3.2. Основные предположения

Введем предположения относительно свойств исследуемой системы дифференциальных уравнений, позволяющие надеяться, что применение рассмотренных методов даст приемлемый результат.
Основное предположение относительно функции $f(x,y)$ состоит в том, что она удовлетворяет условию Липшица
$\|f(x,y_1) - f(x,y_2)\| \le L\|y_1 - y_2\|$,
где L – константа Липшица,
для всех значений $x \in [a,b]$ и, в общем случае, всех соответствующих компонент векторов y_1 и y_2.
Константа Липшица играет важную роль в теории численных методов, в частности, при исследовании их численной устойчивости, и может быть вычислена как

$L = \| \partial f / \partial y \|$,
в чем нетрудно заметить норму матрицы Якоби.
При выводе методов решения, основанных на разложении в ряд Тейлора, предполагается надлежащая дифференцируемость *y*. При этом говорят, что функция *y* должна обладать той степенью гладкости, которая требуется в конкретном частном случае. Проблема может возникнуть, если искомые функции имеют разрыв. Кроме того, разрыв могут иметь правые части дифференциальных уравнений. Такая ситуация может быть обусловлена конструктивными особенностями моделируемой физической системы (примеры: работа многокамерного амортизатора опоры шасси летательного аппарата; гидравлический удар в системе упругих трубопроводов).
Напомним, что точкой разрыва (особой точкой) функции называется такая точка, в которой функция не является непрерывной. Точка разрыва будет 1 рода, если существуют пределы $\lim_{x \to x_0 + 0} f(x) = f(x_0 + 0)$ и $\lim_{x \to x_0 - 0} f(x) = f(x_0 - 0)$.
Величина $f(x_0 + 0) - f(x_0 - 0)$ называется скачком функции.
Точка разрыва будет 2 рода, если функция определена в окрестности особой точки за исключением, быть может, самой точки. Также один из рассмотренных выше пределов не существует.
Разрыв 1 рода в правых частях дифференциальных уравнений обычно преодолевается численным алгоритмом без какой–либо модернизации алгоритма.

25.3.3. Устойчивость

Применительно к дифференциальным уравнениям различают численную устойчивость методов решения и устойчивость системы дифференциальных уравнений. Таким образом, различают устойчивость метода и устойчивость уравнений как математической модели некоторого физического явления, причем в последнем случае неустойчивость уравнений прямо означает неустойчивость моделируемого физического явления. При сходстве методов анализа устойчивости в том и в другом смысле совершенно различаются объекты, к которым применяется данный анализ.
Анализу устойчивости методов посвящен значительный объем источников по численным методам решения дифференциальных уравнений. Более того, описание вновь введенного метода принято сопровождать анализом его устойчивости. Анализ устойчивости метода сводится к изучению глобальной ошибки решения, складывающейся из ошибки усечения и ошибки распространения. На основе этого анализа для различных методов, если говорить о численных методах решения обыкновенных дифференциальных уравнений, выводятся рекомендации по выбору минимально допустимого шага интегрирования либо, в общем смысле, интервала устойчивости.
Другое определение устойчивого алгоритма проще: устойчивым называется алгоритм, в котором не накапливаются ошибки округления.

25.3.3.1. Жесткие задачи

Устойчивое дифференциальное уравнение называется жестким, если оно имеет частное решение в виде убывающей экспоненты, постоянная времени которого очень мала по сравнению с длиной интервала, на котором разыскивается решение. Математически о высокой жесткости системы свидетельствует большое значение константы Липшица.
Есть и другой показатель. Согласно Лэмберту, задача Коши для устойчивой системы называется жесткой, если локальный коэффициент жесткости задачи

$$S(x) = \max_{i=1,2,\ldots,n} \mathrm{Re}(-\lambda_i) / \min_{i=1,2,\ldots,n} \mathrm{Re}(-\lambda_i) >> 1,$$

где $\lambda_i, i = 1,2,\ldots,n$ – собственные значения матрицы Якоби системы.

Жесткой считается система уже при значении $S(x) = 10$, хотя на практике могут встречаться значения до величин $S(x) = 10^6$.

Жесткие системы трудны для решения, причем это проблема чисто вычислительная, не отражающая каких–то особых свойств моделируемой физической системы. Жесткие задачи решаются с помощью специально разработанных методов, некоторые из которых представлены в предлагаемом программном обеспечении. Не лишена некоторого смысла идея путем введения специальных коэффициентов «растянуть» решение по времени. Решение можно сделать более пологим, приняв за единицу времени, например, не 1 секунду, а 0,001 секунды. Вероятно, в этом случае можно было бы применять для решения обычные методы.

25.3.3.2. Устойчивость решения

Устойчивость решения системы дифференциальных уравнений соответствует устойчивости движения моделируемой физической системы. Описываемое системой дифференциальных уравнений движение (под которым можно понимать любой изменяющийся процесс) называется устойчивым асимптотически, если достаточно малым возмущениям будет соответствовать наперед заданная малость возмущенного движения. Для устойчивого движения имеет место также стремление амплитуды этого возмущенного движения к нулю при неограниченном росте времени.

Практически неустойчивое поведение физической системы ведет к ее разрушению, к выходу из интервала допустимых эксплуатационных параметров либо (в случае физической системы, описываемой системой нелинейных дифференциальных уравнений) к хаосу (беспорядку, нерегулярности) того или иного типа. «... маленькая ошибка в начале может стать большой в конце» (Фома Аквинский). Критерием перехода регулярной организованной структуры к хаосу может служить устойчивость структур по отношению к малым возмущениям. Если такая устойчивость отсутствует, детерминированное описание структур теряет смысл, и необходимо использовать статистические методы.

Часто для анализа поведения объекта или явления, описываемого системой дифференциальных уравнений, совсем не обязательно решать эту систему. Если исследователя, например, интересуют значения некоторых параметров системы или начальных условий задачи, определяющих, будут ли в системе затухать возмущения, или, наоборот, малые возмущения будут приводить к нарастанию амплитуды возмущенного движения, значит, исследователю нужно исследовать устойчивость системы (интегрировать систему дифференциальных уравнений нет необходимости). «Начальные возмущения для реализации такого движения всегда найдутся» (Р. Лампер).

25.3.4. Численное решение дифференциальных уравнений

Если решается уравнение порядка выше первого, количество начальных условий будет равно порядку уравнения, т.к. начальные условия накладываются как на функцию *y*, так и на все ее производные. Сказанное естественно обобщается и на систему уравнений.

Среди численных методов наиболее употребительны разностные методы, требующие для своего функционирования знания значений функции в нескольких последовательных точках (многошаговые методы), и одношаговые методы (например, метод Рунге–Кутта), требующие знания значения функции только в одной предыдущей точке. И у одношаговых, и у

многошаговых методов есть свои достоинства и недостатки, перечисленные при их описании.
Все рассмотренные численные методы являются дискретными, т. е. с их помощью ищется последовательность значений искомой функции *y* в *N* заданных точках
$y_n \approx y(x_n),$
где $x_{n+1} = x_n + h_n, n = 0,1,..., N-1, x_0 = a, x_N = b$ – заданное множество точек интегрирования, $h_n > 0$ – шаг сетки.
Чаще всего рассматривается случай $h_n = h$, где h – постоянная величина.
Обратим внимание, что следует отличать шаг сетки решения от шага вывода решения, т. е. шага, с которым программное решение должно выдаваться пользователю. Хотя эти величины могут совпадать, в общем случае они различны, причем вторая обычно намного (возможно, на порядки) превосходит первую. Удобно, чтобы данные величины были кратными для гарантии от получения чрезмерно малого значения шага сетки в конце интервала. Это справедливо для всех методов, особенно для методов с автоматическим выбором длины шага, при вызове которых как начальный шаг решения, так и минимальный шаг должны быть кратны шагу вывода. Впрочем, за счет некоторого усложнения алгоритмов представленного программного обеспечения удовлетворение этого требования не обязательно.
Наиболее распространенные схемы решения обыкновенных дифференциальных уравнений без потери общности рассмотрим на примере решения одного дифференциального уравнения.

25.3.4.1. Одношаговые методы

Решение обыкновенного дифференциального уравнения в точке x_{n+1}, если оно известно в точке x_n, на основе разложения в ряд Тейлора ищется по формуле
$y(x_{n+1}) = y(x_n) + h\Delta(x_n, y_n, h),$
где
$$\Delta(x_n, y_n, h) = \sum_{k=1}^{\infty} \frac{h^{k-1}}{k!} y^{(k)}(x).$$
Если бесконечный ряд оборвать на некотором числе членов и точное значение $y(x_n)$ заменить приближенным значением y_n, получится формула,
$y_{n+1} = y_n + h\varphi(x_n, y_n, h),$
где
$$\varphi(x, y, h) = \sum_{k=1}^{p} \frac{h^{k-1}}{k!} f^{(k-1)}(x_n, y_n),$$
где *p* – степень метода решения.
Считается (возможно, что это утверждение не является однозначно верным), что формулы более высоких степеней дают не только более высокую точность, но и уменьшают количество вычислительной работы, т.к. допускают большие величины шага. При *p* = 1 имеем известный метод Эйлера.
С целью исключения производных из формулы для φ при построении методов со степенями *p* > 1 Рунге, Хойн и Кутта предложили процесс «подгонки» рядов Тейлора выражениями, в общем случае имеющими вид

$$\varphi(x,y,h)=\sum_{r=1}^{m}c_r k_r,$$

где m – количество этапов решения.
В названии конкретного метода, таким образом, фигурирует как число этапов, означающее количество вычислений правой части, так и степень метода, означающее число удерживаемых членов разложения и характеризующего теоретическую точность метода. Для степени $p > 4$ в источниках показано, что количество этапов $m > p$.
В источниках все одношаговые методы часто традиционно называют методами Рунге–Кутта.

25.3.4.1.1. Явные схемы

Пусть коэффициенты, входящие в предыдущую формулу, вычисляются как

$$k_1 = f(x,y), k_r = f(x+\alpha_r h, y+h\sum_{s=1}^{r-1}\beta_{r,s}k_s), r = 2,3,...,m,$$

где $\alpha_{\cdot},\beta_{\cdots}$ – константы.
Данная формула определяет общий вид так называемых явных схем одношаговых методов. Подробный вывод громоздок и здесь не приводится.
Основное применение явные схемы находят при решении нежестких систем. В программе представлены метод Рунге–Кутта и методы Мерсона.

25.3.4.1.2. Неявные схемы

Для неявных схем отличие от явных схем в вычислении коэффициентов, входящих в общую формулу одношаговых методов, следующее

$$k_r = f(x+\alpha_r h, y+h\sum_{s=1}^{m}\beta_{r,s}k_s), r = 1,3,\ldots,m,$$

где $\alpha_{\cdot},\beta_{\cdots}$ – константы.
Неявные методы ориентированы на решение жестких систем. Они дают хорошие результаты и для нежестких систем, но по времени счета превосходят их.
В программе представлен метод Хаммера–Холлингсуорта.

25.3.4.1.3. Метод Рунге–Кутта

Расчетные формулы для схемы классического метода Рунге–Кутта имеют вид

$$y_{n+1} = y_n + (k_1 + 2k_2 + 2k_3 + k_4)/6,$$

где

$$k_1 = hf(x_n, y_n),$$

$$k_2 = hf(x_n + h/2, y_n + k_1/2),$$

$$k_3 = hf(x_n + h/2, y_n + k_2/2),$$

$$k_4 = hf(x_n + h, y_n + k_3).$$

Наилучший прием проверки точности решения состоит в том, чтобы после окончания решения задачи методом Рунге–Кутта уменьшить шаг решения h в 2 раза и провести решение с новым шагом. Если незначительность разницы между двумя решениями устраивает исследователя, можно удовлетвориться решением с первоначальным шагом. В противном случае алгоритм уменьшения шага решения следует повторить.
Классический метод Рунге–Кутта может рассматриваться, как обобщение на дифференциальные уравнения квадратурной формулы Симпсона, предназначенной для

численного вычисления определенных интегралов.

25.3.4.1.4. Методы Мерсона

Рассматриваемые далее формулы получены на основе идеи, что если схема решения дифференциального уравнения наряду с приращением функции будет содержать некоторое приближение более высокого порядка, последнее можно использовать для управления погрешностью решения и длиной шага интегрирования.
Созданная Мерсоном оригинальная модификация метода Рунге–Кутта предоставляет возможность автоматического выбора шага *h* для достижения заданной точности решения.
Метод Мерсона (Кутты–Мерсона, Рунге–Кутта–Мерсона), являющийся пятиэтапным методом порядка 4, дает оценку погрешности решения в общем случае (т. е. для нелинейных уравнений) сверху, что позволяет управлять шагом сетки решения, добиваясь точности, не хуже заданной. Формулы первого варианта метода Мерсона могут быть представлены в виде:

$$y_{n+1} = y_n + y_1,$$

где

$$y_1 = k_1/6 + 2k_4/3 + k_5/6,$$

$$k_1 = hf(x_n, y_n),$$

$$k_2 = hf(x_n + h/3, y_n + k_1/3),$$

$$k_3 = hf(x_n + h/3, y_n + k_1/6 + k_2/6),$$

$$k_4 = hf(x_n + h/2, y_n + k_1/8 + 3k_3/8),$$

$$k_5 = hf(x_n + h, y_n + k_1/2 - 3k_3/2 + 2k_4).$$

Кроме того, подсчитывается

$$y_2 = k_1/2 - 3k_3/2 + 2k_4,$$

после чего производится вычисление величины

$$\varepsilon' = |y_1 - y_2|,$$

которая сравнивается с заданной абсолютной погрешностью ε. Если $\varepsilon' > \varepsilon$, шаг интегрирования уменьшается вдвое и вычисление повторяется с предыдущей точки.
Недостатком подхода является то, что не предусмотрено увеличение шага интегрирования при получении «слишком малой» погрешности. Это иногда ведет к более медленной работе алгоритма, чем могло бы быть, исходя из заданной погрешности.
Формулы второго варианта метода Мерсона, предусматривающие увеличение длины шага интегрирования в случае «слишком малой» вычисленной погрешности, имеют вид:

$$y_{n+1} = y_n + (k_1 + 4k_4 + k_5)/2,$$

где

$$k_1 = hf(x_n, y_n)/3,$$

$$k_2 = hf(x_n + h/3, y_n + k_1)/3,$$

$$k_3 = hf(x_n + h/3, y_n + k_1/2 + k_2/2)/3,$$

$$k_4 = hf(x_n + h/2, y_n + 3k_1/8 + 9k_3/8)/3,$$

$$k_5 = hf(x_n + h, y_n + 3k_1/2 - 9k_3/2 + 6k_4)/3.$$

Шаг решения выбирается автоматически в зависимости от оценки абсолютной погрешности решения, данной формулой

$$\varepsilon' = (k_1 - 9k_3/2 + 4k_4 - k_5/2)/5.$$

Возможны три случая:

- Если $\varepsilon' > \varepsilon$, точность неудовлетворительная, шаг уменьшается в 2 раза, и вычисление повторяется с новым, уменьшенным, шагом.
- Если $\varepsilon' \le \varepsilon / 32$, точность чрезмерно высока, шаг должен быть увеличен в 2 раза, и вычисление продолжается с новым, увеличенным, шагом.
- Если $\varepsilon / 32 < \varepsilon' \le \varepsilon$, точность в заданных пределах, шаг выбран (настроен) верно, и вычисление продолжается с текущим «правильным» шагом.

В настоящее время популярен другой метод решения с контролем погрешности – шестиэтапный метод 4 порядка Фельберга.

25.3.4.1.5. Метод Хаммера–Холлингсуорта

Рассмотрим двухэтапный неявный метод 4 порядка Хаммера–Холлингсуорта (Хаммера–Холлингсворта):

$$y_{n+1} = y_n + (k_1 + k_2)/2,$$

где

$$k_1 = hf(x_n + (1/2 - \sqrt{3}/6)h, y_n + k_1/4 + (1/4 - \sqrt{3}/6)k_2),$$

$$k_2 = hf(x_n + (1/2 + \sqrt{3}/6)h, y_n + (1/4 + \sqrt{3}/6)k_1 + k_2/4).$$

Две последних формулы определяют итерационный процесс, для нежестких уравнений сходящийся очень быстро.

Для нежестких систем представленный неявный алгоритм не эффективнее классического метода Рунге–Кутта. Однако в случае жестких систем, по данным литературы, явные одношаговые методы не работают вообще, а с помощью неявных одношаговых методов решение получить иногда удается.

Другие часто применяемые неявные методы: метод Кунцмана–Бутчера порядка 6 и 8, а также комбинированные алгоритмы Бутчера (Батчера) порядка 4 и 6.

25.3.4.2. Многошаговые методы

Многошаговые методы, в отличие от рассмотренных выше одношаговых методов, оперируют значениями функции в нескольких предыдущих точках. Общая формула линейного *k*–шагового метода имеет вид

$$y_{n+1} = \sum_{i=1}^{k} \alpha_i y_{n+1-i} + h \sum_{i=0}^{k} \beta_i f_{n+1-i},$$

где $\alpha_.$ – коэффициенты перед значениями функции,
$\beta_.$ – коэффициенты перед значениями правых частей.

С помощью многошагового метода нельзя начать решения, т.к. для вычисления y_{n+1} должны быть известны все или некоторые (в зависимости от метода) величины

$$y_n, y_{n-1}, \ldots, y_{n-k+1}, f_n, f_{n-1}, f_{n-k+1}.$$

Бывают явные и неявные многошаговые методы. Некоторые многошаговые методы включают в себя два шага: явный, называемый предиктором, и неявный (или несколько неявных), называемый корректором.

Многошаговые методы показали свою эффективность при решении жестких систем.

При $\beta_0 = 0$ метод называется явным. В программе представлен классический метод Адамса.

Если $\beta_0 \neq 0$, метод будет называться неявным. В программе представлен метод Гира 4 порядка.

25.3.4.2.1. Метод Адамса

Формула метода Адамса 4 порядка имеет вид

$y_{n+1} = y_n + h(55 f_n - 59 f_{n-1} + 37 f_{n-2} - 9 f_{n-3}) / 24.$

В варианте предиктор–корректор данная формула дополняется выражением

$y_{n+1} = y_n + h(9 f_{n+1} + 19 f_n - 5 f_{n-1} + f_{n-2}) / 24.$

25.3.4.2.2. Методы Гира

Для наиболее простых частных случаев рассматриваемый метод введен Кертисом и Хиршфельдером (Гиршфельдером). Общий случай рассмотрел Гир. Метод Гира 4 порядка относится к неявным многошаговым методам, использующим идею так называемого дифференцирования назад. Для рассматриваемого случая формула имеет вид

$25/12 y_{n+1} - 4 y_n + 3 y_{n-1} - 4/3 y_{n-2} + 1/4 y_{n-3} = h f_{n+1}.$

Методы Гира применяются для интегрирования жестких дифференциальных уравнений. Для нежестких систем получены также хорошие результаты.

Список использованной и рекомендуемой литературы

1. Bartoszewski Z., Jackiewicz Z. Toward a two–step Runge–Kutta code for nonstiff differential systems // Applicationes Mathematicae, 2001, vol. 28, no. 3, pp. 353–365.
2. Goyal M. Computer–based numerical & statistical techniques. – Hingman, MA: Infinity Science Press, 2007.
3. Murray J.D. Mathematical biology. I. An introduction. – New York, NY: Springer–Verlag, 2002.
4. Press W.H. Numerical recipes: The art of scientific computing / W.H. Press, S.A. Teukolsky, W.T. Vetterling et al. – New York, NY: Cambridge University Press, 2007.
5. Аганбегян А.Г. Математика в социологии: моделирование и обработка информации / Под ред. А.Г. Аганбегяна, Ф.М. Бородкина. – М.: Мир, 1977.
6. Бабушка И., Витасек Э., Прагер М. Численные процессы решения дифференциальных уравнений. – М.: Мир, 1969.
7. Балантер Б.И., Ханин М.А., Чернавский Д.С. Введение в математическое моделирование патологических процессов. – М.: Медицина, 1980.
8. Васильков Ю.В., Василькова Н.Н. Компьютерные технологии вычислений в математическом моделировании. – М.: Финансы и статистика, 2002.
9. Вольтерра В. Математическая теория борьбы за существование. – М.: Наука, 1976.
10. Выгодский М.Я. Справочник по высшей математике. – М.: Большая медведица, 2001.
11. Гантмахер Ф.Р. Теория матриц. – М.: Наука, 1988.
12. Губин В.Б. Об одном варианте принципа бритвы Оккама // Философские науки, 1998, № 2, с. 136–150.
13. Иванов В.В. Методы вычислений на ЭВМ: Справочное пособие. – Киев: Наукова думка, 1986.
14. Кадыров Х.К., Антомонов Ю.Г. Синтез математических моделей биологических и медицинских систем. – Киев: Наукова думка, 1974.
15. Каханер Д., Моулер К., Нэш С. Численные методы и программное обеспечение. – М.: Мир, 2001.
16. Кемени Дж., Снелл Дж. Кибернетическое моделирование. Некоторые приложения. – М.: Советское радио, 1972.
17. Краснощеков П.С., Петров А.А. Принципы построения моделей. – М.: ФАЗИС: ВЦ РАН, 2000.

18. Лампер Р.Е. Введение в теорию флаттера. – М.: Машиностроение, 1990.
19. Ланс Дж.Н. Численные методы для быстродействующих вычислительных машин. – М.: Издательство иностранной литературы, 1962.
20. Лоскутов А.Ю., Михайлов А.С. Введение в синергетику: Учебное руководство. – М.: Наука, 1990.
21. Мак–Кракен Д., Дорн У. Численные методы и программирование на ФОРТРАНе. – М.: Мир, 1977.
22. Марри Дж. Нелинейные дифференциальные уравнения в биологии. Лекции о моделях. – М.: Мир, 1983.
23. Мудров А.Е. Численные методы для ПЭВМ на языках Бейсик, Фортран и Паскаль. – Томск: МП «РАСКО», 1991.
24. Олейник О.А. Роль теории дифференциальных уравнений в современной математике и ее приложениях // Соросовский образовательный журнал, 1996, т. 2, № 4, с. 114–121.
25. Орлов А.И. Устойчивость в социально–экономических моделях. – М.: Наука, 1979.
26. Трикоми Ф. Дифференциальные уравнения. – М.: Издательство иностранной литературы, 1962.
27. Тутубалин В.Н. Математическое моделирование в экологии: Историко–методологический анализ / В.Н. Тутубалин, Ю.М. Барабашева, А.А. Григорян и др. – М.: Языки русской культуры, 1999.
28. Хайрер Э., Ваннер Г. Решение обыкновенных дифференциальных уравнений. Жесткие и алгебро–дифференциальные задачи. – М.: Мир, 1999.
29. Хайрер Э., Нерсетт С., Ваннер Г. Решение обыкновенных дифференциальных уравнений. Нежесткие задачи. – М.: Мир, 1990.
30. Хаммуд Г.М. Трехмерное семейство 7–шаговых методов Рунге–Кутта порядка 6 // Вычислительные методы и программирование, 2001, т. 2, с. 159–166.
31. Хиценко В.Е. Самоорганизация: элементы теории и социальные приложения. – М.: КомКнига, 2005.
32. Холл Дж. Современные численные методы решения обыкновенных дифференциальных уравнений / Под ред. Дж. Холла и Дж. Уатта. – М.: Мир, 1979.
33. Шубин М.А. Математический анализ для решения физических задач. – М.: МЦНМО, 2003.
34. Шуп Т. Решение инженерных задач на ЭВМ: Практическое руководство. – М.: Мир, 1982.
35. Шустер Г. Детерминированный хаос: Введение. – М.: Мир, 1988.
36. Эндрюс Дж. Математические моделирование / Под ред. Дж. Эндрюса и Р. Мак–Лоуна. – М.: Мир, 1979.
37. Эрроусмит Д., Плейс. Обыкновенные дифференциальные уравнения. Качественная теория с приложениями. – М.: Мир, 1986.

Глава 26. Многочлены

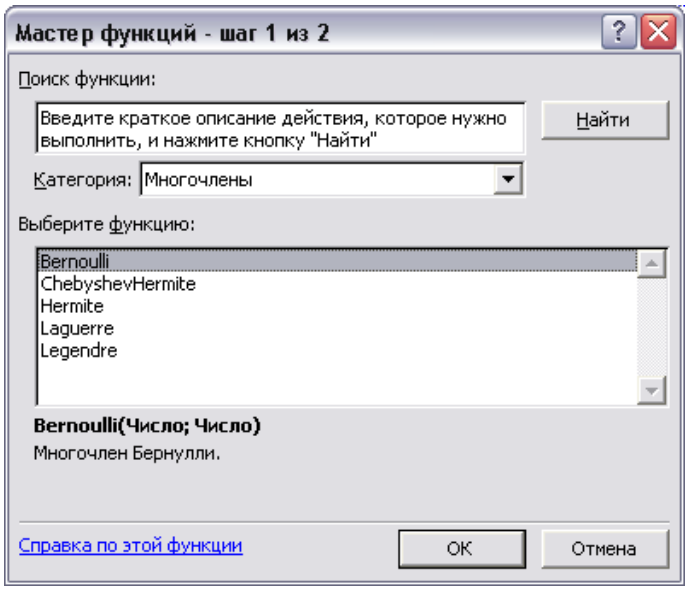
26.1. Введение

В программном обеспечении представлены функции вычисления полиномов (многочленов). Полиномом называют выражение, состоящее из нескольких частей одного типа. Многочлены возникают в различных приложениях, например, при решении дифференциальных уравнений, интерполировании и т. д. Возможные приложения

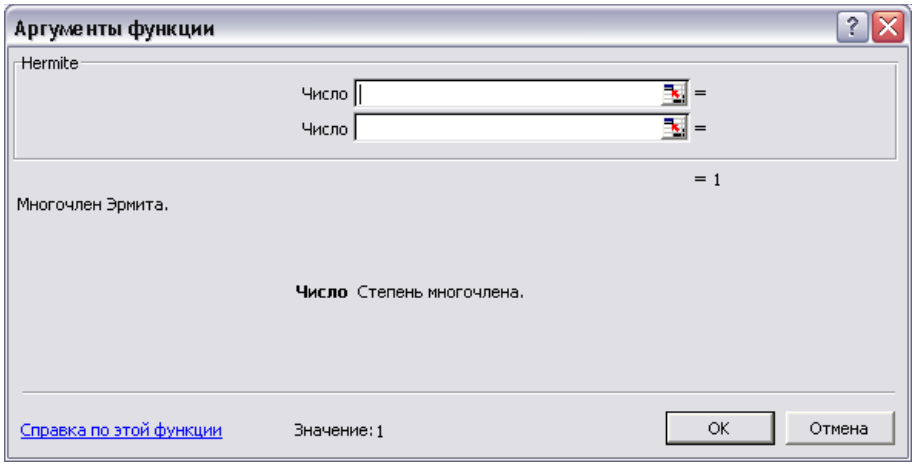
многочленов к статистическому анализу данных и математическому моделированию рассмотрены при описании конкретных примеров.

26.2. Работа с программным обеспечением

Особенностью является отсутствие какой–либо интерфейсной части, кроме возможности вызова содержащихся в разделе функций непосредственно с рабочего листа электронных таблиц. Функции вычисления полиномов (многочленов), содержащиеся в программном обеспечении, будут доступны для вызова в категории «Многочлены». Работа с этими функциями ничем не отличается от работы со стандартными функциями, как показано на снимке, сделанном с экрана во время работы Мастера функций, который запускается выбором из меню пункта Вставка | Функция.



На шаге 1 из 2 работы Мастера функций следует выбрать категорию Многочлены и нужную функцию, после чего нажать кнопку OK для перехода к следующему шагу. На экране появится окно, похожее на следующее изображение.



На шаге 2 из 2 следует ввести или выбрать все необходимые параметры в соответствующих полях окна Мастера функций. После ввода или выбора нажать кнопку OK. Будет выполнен

требуемый расчет, а результат помещен в ячейку, в которую была введена функция. Некоторым недостатком данного варианта является то, что при передаче рабочей книги (файла), содержащей вызовы функций программного обеспечения, адресат должен иметь на компьютере установленную копию данного программного обеспечения.

26.3. Теоретическое обоснование

Мы рассмотрим вычисление некоторых многочленов из класса многочленов Аппеля, содержащего такие важные системы многочленов, как многочлены Бернулли, Лагерра, Эрмита, а также вычисление многочленов Чебышева и Лежандра. Многочлены Чебышева первого и второго рода и многочлены Лежандра являются частными случаями многочленов Якоби, являющихся, в свою очередь, специальным случаем гипергеометрической функции. Связь многочленов Якоби и функции B–распределения обсуждается в литературе.
Многие из рассмотренных многочленов обладают свойством ортогональности, которое заключается в том, что

$$\sum_{s=1}^{n} P_i\ (x_s) P_j\ (x_s) = 0, i \neq j.$$

Некоторые многочлены связаны с аналитическим решением специальных типов дифференциальных уравнений. Для асимптотических разложений статистических распределений применяются многочлены Эрмита и Лагерра.

26.3.1. Многочлены Бернулли

Многочлены Бернулли определяются формулой

$$B_n(x) = \sum_{s=0}^{n} C_n^s B_s x^{n-s}, n = 0,1,2,...,$$

где B_s – числа Бернулли.
Так как все нечетные числа Бернулли, кроме числа с номером 1, равны нулю, для обозначения чисел и многочленов Бернулли применяют также обозначение $B_{2m}(x), m=0,1,2,...$.
Рекуррентные соотношения для определения чисел Бернулли имеют вид:
$B_0(x) = 1$,

$$B_n(x) = x^n - \frac{1}{n+1} \sum_{s=0}^{n-1} C_{n+1}^s B_s(x), n = 1,2,...,$$

где можно было бы продолжить и составить рекурсию для числа сочетаний, чего мы пока делать не будем.

26.3.2. Многочлены Лагерра

Обобщенные (присоединенные) многочлены Лагерра являются решениями дифференциального уравнения
$xy'' + (\lambda + 1 - x)y' + ny = 0$
и определяются формулой

$$L_n^{(\lambda)}(x) = (-1)^n x^{-\lambda} e^x \frac{d^n}{dx^n}(x^{\lambda+n} e^{-x}), n = 0,1,2,...,$$

где $\lambda > -1$,
$0 \le x < \infty$.
Рекуррентные соотношения для вычисления обобщенных многочленов Лагерра выглядят как
$L_0^{(\lambda)}(x) = 1$,

$L_1^{(\lambda)}(x) = x - \lambda - 1,$
$L_n^{(\lambda)}(x) = (x - \lambda - 2n + 1)L_{n-1}^{(\lambda)}(x) - (n-1)(\lambda + n - 1)L_{n-2}^{(\lambda)}(x),\ n = 2,3,....$

Связь многочленов Лагерра с интегралом вероятностей χ^2 и с Г–распределением обсуждается в литературе.

26.3.3. Многочлены Эрмита

Многочлены Эрмита являются решениями дифференциального уравнения
$y'' - 2xy' + 2ny = 0$, $n = 0,1,2,...,$
и определяются формулой

$$H_n(x) = (-1)^n e^{x^2} \frac{d^n e^{-x^2}}{dx^n}.$$

Рекуррентные соотношения для вычисления многочленов Эрмита выглядят как
$H_0(x) = 1,$
$H_1(x) = 2x,$
$H_n(x) = 2xH_{n-1}(x) - 2(n-1)H_{n-2}(x),\ n = 2,3,...$.
Рекуррентные соотношения для вычисления многочленов Эрмита можно записать по–другому. Тогда их называют многочленами Чебышева–Эрмита и вычисляют как
$H_0(x) = 1,$
$H_1(x) = -x,$
$H_n(x) = -xH_{n-1}(x) - (n-1)H_{n-2}(x),\ n = 2,3,...$.
В такой форме многочлены применяются для вычисления производных плотности нормального распределения по формуле

$$\varphi^{(n)}(x) = \frac{d^n \varphi(x)}{dx^n} = H_n(x)\varphi_n(x).$$

26.3.4. Многочлены Чебышева

Многочлены Чебышева первого рода являются решениями дифференциального уравнения
$(1 - x^2)y'' - xy' + n^2 y = 0$
и определяются формулой

$$T_n(x) = \cos(n \arccos(x)) = \frac{2^n n!}{(2n)!}\sqrt{1-x^2}\,\frac{d^n}{dx^n}\left[(1-x^2)^{n-1/2}\right], n = 0,1,2,...,$$

где $-1 \le x \le 1$.
Рекуррентные соотношения выглядят гораздо проще
$T_0(x) = 1,$
$T_1(x) = x,$
$T_n(x) = 2xT_{n-1}(x) - T_{n-2}(x),\ n = 2,3,...$.
Многочлены Чебышева второго рода являются решениями дифференциального уравнения
$(1-x^2)y'' - 3xy' + n(n+2)y = 0$
и определяются формулой

$$U_n(x) = \frac{\sin[(n+1)\arccos(x)]}{\sqrt{1-x^2}} = \frac{2^n (n+1)!}{(2n+1)!}\frac{1}{\sqrt{1-x^2}}\frac{d^n}{dx^n}\left[(1-x^2)^{n+1/2}\right],\ n = 0,1,2,...,$$

где $-1 \le x \le 1$.
Рекуррентные соотношения выглядят как
$U_0(x) = 1,$

$U_1(x) = 2x$,
$U_n(x) = 2xU_{n-1}(x) - U_{n-2}(x)$, $n = 2,3,\ldots$.
Многочлены Чебышева образуют систему, ортогональную на отрезке $-1 \le x \le 1$. В литературе дается формула связи многочленов Чебышева первого и второго рода.

26.3.5. Многочлены Лежандра

Рассмотрим дифференциальное уравнение

$$(1-x^2)y'' - 2xy' + \left[n(n+1) - \frac{\mu^2}{1-x^2} \right] y = 0$$

,

где n и μ – произвольные числа.
Если $n = 0,1,2,\ldots$, а $\mu = 0$, то ограниченные на отрезке $-1 \le x \le 1$ решения уравнения называются многочленами Лежандра (сферическими многочленами)

$$P_n(x) = \frac{1}{n!2^n} \frac{d^n}{dx^n} (x^2 - 1)^n$$

и определяются по рекуррентным формулам
$P_0(x) = 1$,
$P_1(x) = x$,
$P_n(x) = [(1 - 2n)xP_{n-1}(x) + (n - 1)P_{n-2}(x)] / n$, $n = 2,3,\ldots$.
Благодаря тому, что многочлены Лежандра ортогональны на отрезке $-1 \le x \le 1$, они образуют полную систему функций и могут быть использованы для разложения в ряд произвольной функции, интегрируемой на отрезке $-1 \le x \le 1$.

Список использованной и рекомендуемой литературы

1. Galassi M. GNU Scientific Library Reference Manual / M. Galassi, J. Davies, J. Theiler et al. – Network Theory, 2005.
2. Ronveaux A. Classical orthogonal polynomials: dependence of parameters / A. Ronveaux, A. Zarzo, I. Areac et al. // Journal of Computational and Applied Mathematics, 2000, vol. 121, pp. 95–112.
3. Большев Л.Н., Смирнов Н.В. Таблицы математической статистики. – М.: Наука, 1983.
4. Васильев Н., Зелевинский А. Многочлены Чебышева и рекуррентные соотношения // Квант, 1982, № 1, с. 12–19.
5. Винберг Э.Б. Симметрия многочленов. – М.: МЦНМО, 2001.
6. Данилов Ю.А. Многочлены Чебышева. – Минск: Вышэйшая школа, 1994.
7. Керов С.В. Равновесие и ортогональные полиномы // Алгебра и анализ, 2000, т. 12, вып. 6, с. 224–237.
8. Корн Г., Корн Т. Справочник по математике для научных работников и инженеров. – М.: Наука, 1973.
9. Малоземов В.Н., Певный А.Б. Рекуррентные вычисления. – Л.: Издательство Ленинградского университета, 1976.
10. Прасолов В.В. Многочлены. – М.: МЦНМО, 2003.
11. Прохоров С.А., Дегтярева О.А. Аппроксимация плотности вероятности ортогональными функциями Лагерра и получение аналитических выражений для характеристических функций по параметрам модели // Электронный научный журнал «Исследовано в России», 2005, т. 8, с. 1184–1189.
12. Прохоров Ю.В. Вероятность и математическая статистика: Энциклопедия / Гл. ред. Ю.В. Прохоров. – М.: Научное издательство «Большая Российская энциклопедия», 1999.

13. Прохоров Ю.В. Математический энциклопедический словарь / Гл. ред. Ю.В. Прохоров. – М.: Научное издательство «Большая Российская энциклопедия», 1995.
14. Хемминг Р.В. Численные методы для научных работников и инженеров. – М.: Наука, 1972.