

А. Наследов

SPSS 19

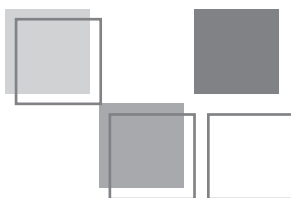
профессиональный
статистический
анализ данных

 ПИТЕР®

А. Наследов

SPSS 19

профессиональный
статистический
анализ данных



Москва • Санкт-Петербург • Нижний Новгород • Воронеж
Ростов-на-Дону • Екатеринбург • Самара • Новосибирск
Киев • Харьков • Минск

2011

ББК 32.973.233-018

УДК 004.422

НЗ1

Наследов А.

НЗ1 SPSS 19: профессиональный статистический анализ данных. — СПб.: Питер, 2011. — 400 с.: ил.

ISBN 978-5-459-00344-4

Книга представляет собой практическое руководство по анализу данных с помощью самой мощной и популярной программы статистической обработки информации — SPSS версии 19. В издании подробно описываются основы работы с пакетом SPSS, рассматривается большинство методов обработки и анализа данных, а также способов табличного и графического представления полученных результатов.

Материал книги организован таким образом, чтобы удовлетворить запросы как новичка, впервые приступающего к анализу данных на компьютере, так и опытного исследователя, желающего воспользоваться самыми современными методами. Основное содержание глав составляют пошаговые инструкции по реализации различных видов математико-статистического анализа в SPSS. Особое внимание уделяется получаемым результатам и их интерпретации. В конце книги приведены глоссарий, содержащий определения большинства статистических терминов.

Издание адресовано исследователям в области статистики, маркетинга, социологии, психологии, а также широкому кругу читателей, желающих воспользоваться программой SPSS для профессионального анализа данных.

ББК 32.973.233-018

УДК 004.422

Все права защищены. Никакая часть данной книги не может быть воспроизведена в какой бы то ни было форме без письменного разрешения владельцев авторских прав.

Информация, содержащаяся в данной книге, получена из источников, рассматриваемых издательством как надежные. Тем не менее, имея в виду возможные человеческие или технические ошибки, издательство не может гарантировать абсолютную точность и полноту приводимых сведений и не несет ответственности за возможные ошибки, связанные с использованием книги.

ISBN 978-5-459-00344-4

© ООО Издательство «Питер», 2011

Краткое содержание

Предисловие	11
Глава 1. Введение.....	14
Глава 2. Общий обзор SPSS	22
Глава 3. Создание и редактирование файлов данных.....	41
Глава 4. Управление данными	58
Глава 5. Диаграммы	91
Глава 6. Частоты	106
Глава 7. Описательные статистики	117
Глава 8. Таблицы сопряженности и критерий хи-квадрат.....	124
Глава 9. Корреляции	138
Глава 10. Средние значения	148
Глава 11. Сравнение двух средних и t-критерий	154
Глава 12. Непараметрические критерии	164
Глава 13. Однофакторный дисперсионный анализ	184
Глава 14. Многофакторный дисперсионный анализ	196
Глава 15. Многомерный дисперсионный анализ	212
Глава 16. Дисперсионный анализ с повторными измерениями	226
Глава 17. Простая линейная регрессия	239
Глава 18. Множественный регрессионный анализ	251
Глава 19. Анализ надежности	266
Глава 20. Факторный анализ	278
Глава 21. Кластерный анализ	296
Глава 22. Дискриминантный анализ	314
Глава 23. Многомерное шкалирование	332
Глава 24. Логистическая регрессия	347
Глава 25. Логлинейный анализ таблиц сопряженности	359
Глоссарий	372
Англо-русский словарь терминов	386
Литература	398

Содержание

Предисловие.....	11
Глава 1. Введение.....	14
Обработка данных на компьютере	14
Необходимые знания	15
Версии SPSS.....	16
Содержание книги.....	17
Файлы примеров	18
Структура глав и элементы описания	20
Глава 2. Общий обзор SPSS	22
Запуск программы	22
Кнопки и другие элементы управления	23
Настройка параметров программы	24
Окна программы	26
Окно редактора командного языка Syntax.....	31
Окно вывода и его редактирование.....	33
Сохранение, экспорт, перенос и печать результатов.....	38
Глава 3. Создание и редактирование файлов данных	41
Структура файла данных	41
Ввод данных	52
Редактирование данных	54
Пример файла данных	56
Глава 4. Управление данными	58
Знакомство с возможностями управления данными.....	58
Получение информации о файле.....	61
Обработка пропущенных значений	61
Преобразование данных	62
Выбор наблюдений для анализа.....	67
Перекодировка в новую переменную	70
Перекодирование существующей переменной	74
Сортировка наблюдений	77
Объединение данных разных файлов.....	78

Агрегирование данных	83
Реструктурирование данных.....	85
Глава 5. Диаграммы.....	91
Графика в программе SPSS.....	91
Настройка диаграмм	92
Команды построения диаграмм.....	93
Редактирование диаграмм	103
Выход из программы	105
Глава 6. Частоты.....	106
Пошаговые алгоритмы вычислений	107
Представление результатов	113
Завершение анализа и выход из программы.....	116
Глава 7. Описательные статистики	117
Пошаговый алгоритм вычислений	119
Представление результатов	122
Завершение анализа и выход из программы.....	123
Глава 8. Таблицы сопряженности и критерий хи-квадрат.....	124
Таблицы сопряженности	124
Критерий независимости хи-квадрат	125
Пошаговый алгоритм вычислений	126
Представление результатов	133
Терминология, используемая при выводе	136
Завершение анализа и выход из программы.....	137
Глава 9. Корреляции	138
Понятие корреляции.....	138
Дополнительные сведения.....	140
Пошаговые алгоритмы вычислений	142
Представление результатов.....	146
Завершение анализа и выход из программы.....	147
Глава 10. Средние значения	148
Пошаговый алгоритм вычислений	148
Представление результатов.....	152
Завершение анализа и выход из программы.....	153

Глава 11. Сравнение двух средних и t-критерий	154
Уровень значимости	155
Пошаговые алгоритмы вычислений	155
Представление результатов	160
Завершение анализа и выход из программы	163
 Глава 12. Непараметрические критерии	 164
Параметрические и непараметрические критерии	164
Пошаговые алгоритмы и результаты вычислений	166
Завершение анализа и выход из программы	183
 Глава 13. Однофакторный дисперсионный анализ	 184
Пошаговые алгоритмы вычислений	185
Представление результатов	191
Терминология	194
Завершение анализа и выход из программы	195
 Глава 14. Многофакторный дисперсионный анализ	 196
Файлы данных для группы методов Общая линейная модель	197
Дисперсионный анализ с двумя факторами	198
Дисперсионный анализ с тремя и более факторами	200
Влияние ковариат	201
Пошаговые алгоритмы вычислений	201
Представление результатов	207
Терминология, используемая при выводе	211
Завершение анализа и выход из программы	211
 Глава 15. Многомерный дисперсионный анализ	 212
Пошаговые алгоритмы вычислений	214
Представление результатов	220
Завершение анализа и выход из программы	225
 Глава 16. Дисперсионный анализ с повторными измерениями	 226
Пошаговые алгоритмы вычислений	228
Представление результатов	234
Завершение анализа и выход из программы	238
 Глава 17. Простая линейная регрессия	 239
Простая регрессия	239
Оценка криволинейности	241

Пошаговые алгоритмы вычислений	243
Представление результатов	247
Терминология, используемая при выводе	249
Завершение анализа и выход из программы.....	250

Глава 18. Множественный регрессионный анализ.....251

Уравнение множественной регрессии	251
Коэффициенты регрессии	252
Коэффициент детерминации и пошаговые методы.....	253
Условия получения приемлемых результатов анализа	254
Пошаговые алгоритмы вычислений	255
Представление результатов	262
Завершение анализа и выход из программы.....	265

Глава 19. Анализ надежности266

Коэффициент альфа.....	266
Надежность половинного расщепления.....	267
Пошаговые алгоритмы вычислений	267
Представление результатов	273
Завершение анализа и выход из программы.....	277

Глава 20. Факторный анализ.....278

Вычисление корреляционной матрицы	279
Извлечение факторов	279
Выбор и вращение факторов.....	280
Интерпретация факторов	282
Пошаговые алгоритмы вычислений	283
Представление результатов	290
Терминология, используемая при выводе	294
Завершение анализа и выход из программы.....	295

Глава 21. Кластерный анализ.....296

Сравнение кластерного и факторного анализов.....	297
Этапы кластерного анализа	298
Кластерный анализ матрицы различий (сходства)	300
Пошаговые алгоритмы вычислений	302
Представление результатов	309
Завершение анализа и выход из программы.....	313

Глава 22. Дискриминантный анализ314

Этапы дискриминантного анализа.....	316
Пошаговые алгоритмы вычислений	317

Представление результатов	324
Терминология, используемая при выводе	328
Завершение анализа и выход из программы.....	331
Глава 23. Многомерное шкалирование	332
Квадратная асимметричная матрица различий	333
Квадратная симметричная матрица различий	334
Модель индивидуальных различий	335
Пошаговые алгоритмы вычислений	335
Представление результатов	343
Завершение анализа и выход из программы.....	346
Глава 24. Логистическая регрессия	347
Математическое описание логистической регрессии.....	348
Пошаговые алгоритмы вычислений	349
Представление результатов	353
Терминология, используемая при выводе	356
Завершение анализа и выход из программы.....	358
Глава 25. Логлинейный анализ таблиц сопряженности	359
Понятие логлинейной модели	359
Логлинейный метод подбора модели.....	360
Пошаговые алгоритмы вычислений	362
Представление результатов	366
Завершение анализа и выход из программы.....	371
Глоссарий	372
Англо-русский словарь терминов.....	386
Литература	398

Предисловие

Внедрение компьютеров буквально во все сферы человеческой деятельности является на сегодняшний день, наверное, самым наглядным результатом научного прогресса. И как следовало ожидать, в существенной степени компьютеризация изменила исследовательский процесс. Компьютер сейчас используется для выполнения такой работы, которая раньше считалась самой скучной и утомительной: учет и организация исходных данных, вычисление различных показателей и пр. Это позволяет исследователю проводить более глубокий анализ данных, больше времени уделять интерпретации результатов и выдвижению новых предположений, то есть заниматься тем, что считается самым приятным и интересным в любом исследовании и что остается за пределами возможностей компьютера.

Эта книга является учебным пособием и руководством по применению компьютерной программы SPSS — очень мощного и широко распространенного средства профессионального компьютерного анализа данных. SPSS — это аббревиатура от Statistical Package for the Social Science (статистический пакет для социальных наук). Как следует из названия, SPSS представляет собой набор различных программ обработки данных. Эти программы облегчают процесс ввода информации, позволяют гибко менять структуру данных, использовать самые современные методы обработки (как по отдельности, так и совместно) и получать результаты в удобной и наглядной форме. Все это множество программ собрано в единую систему, обеспечивающую простой и дружелюбный диалог с исследователем и снабженную исчерпывающей справочной поддержкой. Благодаря такой дружелюбности система SPSS легко доступна для освоения даже тем, кто имеет минимальные навыки владения компьютером.

Но если SPSS настолько дружелюбная система, да еще обладающая справочной поддержкой, зачем нужна эта книга? Как показывает практика, оставшись один на один со справочной системой SPSS, даже опытный исследователь, не говоря о студентах, впадает в отчаяние, безуспешно пытаясь найти в мегабайтах документации нужную информацию. К этому следует добавить не всегда достаточное знание пользователем специфической статистической и математической терминологии. Но даже те, кто применяет SPSS для анализа данных, испытывают трудности при необходимости реорганизации исходных данных и редактировании результатов обработки (таблиц и графиков). Именно подобные трудности и стали стимулом к написанию этой книги.

В книге содержится подробное описание большинства процедур SPSS, предназначенных для анализа данных — от простейших до самых сложных, включая различные нюансы их применения. При описании любой процедуры основными подходами являлись: максимально доступное изложение ее назначения; сведение действий по ее применению к четким пошаговым инструкциям; демонстрация ее применения на простейших примерах. Однако было бы неверно думать, что эта книга предназначена только для тех, кто хочет освоить азы компьютерного анализа данных. Она, без сомнения, будет полезна и как справочник для тех, кто уже имеет опыт работы с программой SPSS. Справочному назначению книги способствуют глоссарий и англо-русский словарь терминов, приведенные в конце книги.

Содержание первых 12 глав адресовано широкой аудитории и не требует специальных знаний в области статистики, однако оставшаяся часть книги рассчитана на людей, имеющих такие знания. В начале каждой из глав обычно кратко излагается сущность статистической процедуры, которой посвящена глава. При этом делается все возможное, чтобы исключить избыточные детали и оставить лишь достаточную для анализа данных информацию. Говорить простым языком о сложных вещах — весьма нелегкое занятие, и пришлось приложить немало усилий, чтобы написать эту книгу. Насколько успешными были эти усилия, судить вам, читатель.

Программа SPSS снабжена подробной справочной системой. На нескольких тысячах страниц можно найти практически любую информацию по работе с SPSS, которая обычно необходима даже опытному пользователю. Разумеется, такой объем информации немислимо было разместить в нескольких сотнях страниц этой книги. Однако подавляющее большинство методов нашло отражение в книге, хотя временами и поверхностное, но, хочется верить, достаточное для практического применения.

Цивилизованное использование SPSS, в том числе в учебных целях, невозможно без приобретения официальной лицензии. Собственником программного обеспечения SPSS с 2010 года является SPSS Inc., an IBM Company© со штаб-квартирой в США. В российском официальном представительстве этой компании можно купить лицензионные версии SPSS. Всю необходимую информацию вы найдете по адресу <http://www.spss.ru/price/index.htm>. Кроме того, вы можете воспользоваться бесплатной пробной версией SPSS, срок действия которой 14 дней. Скачать пробную 14-дневную версию SPSS последней модификации можно по адресу <http://www.spss.ru/products/spss/index.htm>.

От издательства

Ваши замечания, предложения, вопросы отправляйте по адресу электронной почты comp@piter.com (издательство «Питер», компьютерная редакция).

Мы будем рады узнать ваше мнение!

Все файлы данных, упомянутые в книге, вы сможете найти по адресу <http://www.piter.com/download>.

Подробную информацию о наших книгах вы найдете на веб-сайте издательства <http://www.piter.com>.

1 Введение

14	Обработка данных на компьютере
15	Необходимые знания
16	Версии SPSS
17	Содержание книги
18	Файлы примеров
20	Структура глав и элементы описания

Основное содержание этой книги составляют пошаговые инструкции по применению программы SPSS для самого разностороннего анализа данных. Книга организована так, чтобы обеспечить как новичка, так и опытного пользователя исчерпывающей информацией для самостоятельной обработки данных.

В данной главе представлено краткое содержание книги, описана ее структура, приведены некоторые рекомендации по ее изучению. Если вы не знакомы с программой SPSS, рекомендуется прочитать эту главу полностью, поскольку она поможет лучше ориентироваться в материале книги.

Если вы уже имеете опыт обработки данных при помощи программы SPSS и вас интересует лишь описание конкретной статистической процедуры, достаточно найти это описание в одной из глав 6–25. Эти главы не зависят друг от друга: открыв любую из них и следуя инструкциям, вы получите результат и сможете в нем разобраться. Однако даже опытному пользователю рекомендуется все же обратить внимание на главы 2–5, которые содержат полезные рекомендации по таким важным вопросам, как управление исходными данными, использование командного языка Syntax, редактирование отчетов и таблиц результатов, создание и модификация графиков и диаграмм.

Обработка данных на компьютере

Анализ данных с применением компьютера включает выполнение ряда необходимых шагов.

1. Определение структуры исходных данных.
2. Ввод данных в компьютер в соответствии с их структурой и требованиями программы. Редактирование и преобразование данных.
3. Задание метода обработки данных в соответствии с задачами исследования.

4. Получение результата обработки данных. Его редактирование и сохранение в нужном формате.
5. Интерпретация результата обработки.

Шаги 1 (подготовительный) и 5 (заключительный) не способна выполнить ни одна компьютерная программа — их исследователь делает сам. Помощь компьютера (шаги 2–4) заключается, в конечном итоге, в переходе от длинной последовательности чисел к более компактной. На «вход» компьютера исследователь подает массив исходных данных, который недоступен осмыслению, но пригоден для компьютерной обработки (шаг 2). Затем исследователь дает программе команду на обработку данных в соответствии с поставленной задачей и структурой данных (шаг 3). На «выходе» он получает результат обработки (шаг 4) — тоже массив данных, только уже меньший, доступный осмыслению и содержательной интерпретации. При этом исчерпывающий анализ данных обычно требует многократной их обработки с применением разных методов.

В этой книге очень подробно и последовательно рассматривается все, что связано с компьютерной обработкой данных (шаги 2–4), кратко излагаются математические основы каждого метода, а интерпретация результатов дается на конкретных примерах.

Необходимые знания

Эта книга будет для вас действительно эффективным руководством по компьютерной обработке данных, если вы имеете хотя бы самые общие знания о статистике и обладаете элементарными навыками работы с компьютером.

Статистика

В начале каждой главы приводятся основные сведения, касающиеся той или иной статистической процедуры. Кроме того, в конце книги приведен глоссарий, содержащий определения большинства статистических терминов. Однако то и другое представляет собой информацию справочного характера и не рассчитано на новичков. Несмотря на то что практически любой человек с помощью этой книги может работать с пакетом SPSS и получать результаты, отсутствие базовых знаний по статистике не позволит самостоятельно решать даже простые задачи. Поэтому вы должны быть знакомы с основами статистики или, как минимум, изучать их.

От главы к главе требования к уровню знаний постепенно растут. А при описании сложных методов анализа данных (главы 13–25) даже подготовленному читателю, вероятно, потребуется обращение к специальной литературе¹.

¹ Математико-статистические основы большинства описываемых методов читатель может найти в книге: Наследов А. Д. Математические методы психологического исследования. Анализ и интерпретация данных. СПб.: Речь, 2008.

Компьютер

Минимальные требования к навыкам работы на компьютере весьма скромные: вам необходимо лишь уметь включать компьютер, а также пользоваться клавиатурой и мышью. Существенно облегчат освоение SPSS навыки владения программами Microsoft Office и особенно Excel (в основе как Excel, так и SPSS лежат электронные таблицы).

На вашем компьютере должны быть установлены операционная система Windows и программа SPSS (желательно, версии 12.0 и выше), а значок программы должен находиться на рабочем столе.

Версии SPSS

По всем параметрам SPSS является сложным и мощным статистическим пакетом. Однако, несмотря на сложность, средства взаимодействия входящих в пакет программ с пользователем весьма дружелюбны. С помощью пакета SPSS можно проводить практически любой анализ данных, а последние версии программы находят применение в самых разных научных областях и в мире бизнеса.

Основой для данной книги послужила *русифицированная* версия 19.0 пакета IBM SPSS Statistics, распространяемая с сентября 2010 года. Говоря точнее, снимки экрана, присутствующие в книге, соответствуют русифицированной версии 19.0. Однако почти весь изложенный материал может быть с успехом применен и к более ранним версиям, начиная с SPSS 9.0. Основные отличия будут иметь место в интерфейсе программ: названиях диалоговых окон, их виде и т. п. Кроме того, начиная с версии SPSS 13.0 более совершенными стали графические возможности программы. Для отечественного пользователя программы наиболее значительное новшество введено начиная с версии SPSS 12.0 — стала доступной русифицированная версия интерфейса и окон вывода результатов.

Если вы пользуетесь нерусифицированной версией программы, ориентироваться в англоязычной терминологии вам поможет англо-русский словарь терминов, приведенный в конце этой книги.

За последние два года семейство программ SPSS дважды меняло название. В 2009 году привычная аббревиатура *SPSS* сменилась на *PASW*, а в 2010 году традиционное название вернулось в виде *IBM SPSS Statistics*. Заметно расширилось и семейство программных модулей — до 22 программ, одна из которых является основной (IBM SPSS Statistics Base), а остальные могут быть добавлены для расширения возможностей SPSS в разных отношениях.

В книге речь пойдет о статистических процедурах, реализованных в трех основных модулях и наиболее часто использующихся в социальных исследованиях. Мы опишем операции, относящиеся к основному системному модулю (SPSS Base), модулю дополнительных моделей (SPSS Advanced Statistics) и модулю регрессионных моделей (SPSS Regression). Семейство программ SPSS сопрово-

ждается справочной системой, содержащей информацию обо всех процедурах, поддерживаемых этими программами. Однако для начинающего или недостаточно опытного пользователя справочная система слишком сложна, а ее объем, составляющий тысячи печатных страниц, затрудняет поиск нужной информации. В определенной степени это и послужило поводом к написанию книги, в которой в компактной форме представлено практически все, что может оказаться полезным на практике.

Содержание книги

Книга содержит описания большинства методов обработки данных. Решение задачи анализа данных, как отмечалось, происходит в три основных этапа.

1. Ввод данных в компьютер и их структурирование для дальнейшей обработки программой.
2. Определение типа анализа, который необходим для решения задачи.
3. Получение и расшифровка результатов обработки исходных данных.

В главе 2 пойдет речь об основах работы с пакетом SPSS: типах окон, использовании панелей инструментов и меню, просмотре, редактировании, сохранении и печати результатов обработки и т. д. Несмотря на то что глава в основном рассчитана на начинающих, в ней содержится немало информации, полезной и более опытным пользователям. Глава 3 посвящена первому из трех перечисленных выше этапов и включает темы создания, редактирования и форматирования файлов данных. Редактор данных SPSS является инструментом, позволяющим без труда создавать файлы данных и управлять их структурой.

В главах 4 и 5 рассматриваются такие важные вопросы, как модификация и преобразование данных, а также создание диаграмм. Глава 4 описывает различные действия с данными: создание новых переменных, сортировка, изменение структуры, объединение файлов, выбор подмножества данных для анализа. В главе 5 рассмотрены основные процедуры работы с диаграммами; заметим, однако, что большая часть материала, касающаяся диаграмм, излагается в других главах книги по мере необходимости.

Все главы, начиная с главы 6, посвящены двум оставшимся этапам решения задачи анализа данных. В них идет речь о различных методах обработки данных и о том, каким образом интерпретировать их результаты. Каждая из глав 6–25 не зависит от других; это означает, что если вы захотите выполнить какой-либо вариант анализа данных, даже не обладая начальными знаниями, то сможете, открыв соответствующую главу и следуя представленным инструкциям, получить результат. При необходимости в инструкциях приводятся ссылки на другие главы.

Как было отмечено ранее, книга охватывает основное содержимое трех модулей SPSS: Base, Advanced Statistics и Regression Models. Поскольку на практике иногда производится неполная установка пакета SPSS, некоторые из модулей могут

оказаться недоступными (кроме Base, который устанавливается всегда). Для выполнения подавляющего большинства методов анализа достаточно модуля Base. Модули Advanced Models и Regression Models позволяют дополнительно выполнять следующие методы анализа:

- ▶ одномерный (многофакторный) и многомерный дисперсионный анализ (MANOVA) и многомерный ковариационный анализ (MANCOVA) — команды ОЛМ-одномерная и ОЛМ-многомерная группы Общая линейная модель;
- ▶ многомерный дисперсионный анализ с повторными измерениями (repeated-measures MANOVA) — команда ОЛМ-повторные измерения группы Общая линейная модель;
- ▶ логистическая регрессия;
- ▶ логлинейный анализ.

Объем этой книги не позволяет вместить описание всех многочисленных процедур SPSS. Поэтому детально рассмотрены лишь наиболее востребованные методы обработки.

Файлы примеров

В большинстве глав книги при помощи разных методов анализируется один и тот же файл данных. Для последних глав были разработаны отдельные примеры, полнее демонстрирующие применение специфических возможностей пакета SPSS. Все файлы доступны на сайте издательства (<http://www.piter.com/download>). Наличие готовых исходных файлов (включая первый) полностью избавит вас от необходимости набирать данные для обучения вручную. К тому же, выполняя различные процедуры, вы сможете сравнивать получаемые результаты с «эталонными», приведенными в этой книге.

Первый файл (ex01.sav) создан таким образом, чтобы с его помощью можно было без труда продемонстрировать функционирование основных методов обработки. Он содержит информацию о 100 школьниках трех выпускных классов. О каждом из школьников имеются следующие данные:

- ▶ № (идентификационный номер);
- ▶ пол;
- ▶ класс;
- ▶ вуз (выбранный профиль вуза);
- ▶ хобби (систематическое внешкольное занятие);
- ▶ тест1;
- ▶ тест2;

- тест3;
- тест4;
- тест5;
- средняя годовая отметка за 10-й класс;
- средняя годовая отметка за 11-й класс.

Файл `ex01.sav` нам потребуется, начиная уже с вводных глав 2–5. Его содержимое частично приведено в конце главы 3. При желании вы можете вручную создать собственный файл и успешно использовать его в процессе изучения почти всей первой половины книги. Разумеется, все данные уже содержатся в готовом файле на сайте, однако целесообразно самостоятельно попрактиковаться во вводе, форматировании и других простых, но необходимых действиях с данными. Еще более полезным занятием была бы работа с собственными файлами данных.

Всего с упомянутого сайта издательства вы можете загрузить 19 файлов примеров. Из них 17 файлов содержат данные, один файл (`cross.spo`) предназначен для демонстрации результатов обработки (окна вывода), и один файл (`Syn_clust.sps`) содержит пример применения командного языка Syntax. В табл. 1.1 приведен список файлов примеров с указанием глав книги, в которых они используются. В указанных главах вы можете найти либо подробные описания соответствующих примеров, либо ссылки на места в книге, содержащие эти описания.

Таблица 1.1. Файлы примеров и главы, в которых они используются

Файлы примеров	Главы книги
<code>ex01.sav</code>	2–13
<code>cross.spo</code>	2
<code>ex01a.sav</code> , <code>ex01b.sav</code>	4
<code>ex020.sav</code> , <code>ex021.sav</code> , <code>ex022.sav</code>	14–16
<code>exam.sav</code>	17, 26
<code>help.sav</code>	18
<code>TestA.sav</code>	19
<code>TestIQ.sav</code>	20, 21
<code>mds1.sav</code> , <code>mds2.sav</code> , <code>mds3.sav</code>	23
<code>cars.sav</code>	21
<code>Syn_clust.sps</code>	21

Файлы примеров	Главы книги
class.sav	22
helpLR.sav	24
helpLLM.sav	25

В каждой из глав, начиная с главы 5 и заканчивая главой 25, имеются пошаговые процедуры, в которых шаг 3 относится к инструкциям, связанным с открытием конкретного файла данных. При этом предполагается, что эти файлы данных сохранены в отдельной папке на жестком диске вашего компьютера.

Структура глав и элементы описания

Главы 2–5, являющиеся вводными, не имеют единой структуры, поскольку едва ли возможно систематизировать начальную информацию о работе с программой. Материал остальных глав излагается в соответствии с общей структурой книги, которая выглядит следующим образом.

- ▶ Вводный раздел содержит общее описание рассматриваемой процедуры. В зависимости от сложности анализа описание занимает от 1 до 7 страниц.
- ▶ Далее идет раздел пошаговых процедур, требующихся для выполнения одной или нескольких разновидностей статистического анализа.
- ▶ Раздел результатов анализа (как правило, в сокращенном виде). В этом разделе также приводится необходимая терминология, позволяющая понять смысл результатов.
- ▶ Раздел завершения анализа и выхода из программы посвящен, очевидно, завершающим операциям редактирования результатов, их сохранения и выхода из программы.

Снимки экрана

Каждая глава содержит снимки экрана компьютера с фрагментами выводимых данных, элементами окон или целыми окнами, появляющимися на экране компьютера в процессе работы. На рис. 1.1 в качестве примера показано диалоговое окно Частоты, полученное в ходе выполнения шага 4 пошаговой процедуры из главы 6.

Иногда в книге вы можете встретить рисунок, на котором будет представлено не целое окно программы, а его фрагмент, например такой, как показанная ниже строка меню. Подобный рисунок призван напомнить вам, как выглядит упомянутый в текущем абзаце элемент интерфейса или описываемый фрагмент окна.



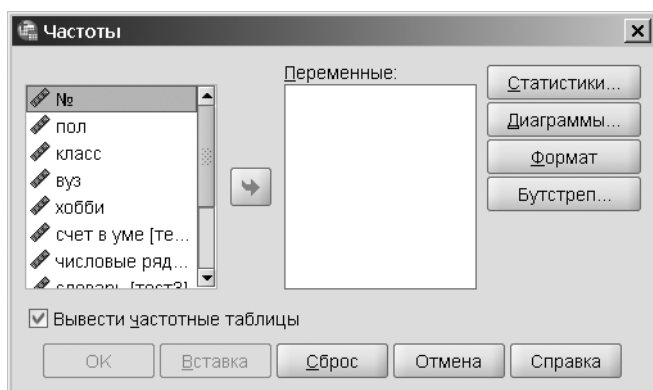


Рис. 1.1. Диалоговое окно Частоты

Помимо окон и элементов программного интерфейса в разделах «Представление результатов» приведены большие фрагменты данных, генерируемые программой. Все такие фрагменты в каждом случае будут сопровождаться подробным описанием.

Пошаговые процедуры

Основные статистические операции оформлены в виде пошаговых процедур, чтобы вы точно знали, что нужно сделать для решения поставленной задачи. Например, ниже приведен шаг 3 пошаговой процедуры из главы 4.

ШАГ 3

Откройте файл данных, с которым вы намерены работать (в нашем случае — это файл `ex01.sav`). Если он расположен в текущей папке, то выполните следующие действия.

1. Выберите в меню Файл команду Открыть ► Данные или щелкните на кнопке Открыть ► Данные панели инструментов.
2. В открывшемся диалоговом окне дважды щелкните на имени `ex01.sav` или введите его с клавиатуры и щелкните на кнопке ОК.

Как можно видеть по приведенному фрагменту пошаговой процедуры, в книге используются некоторые элементы шрифтового выделения. Например, запись Открыть ► Данные означает, что в строке меню нужно щелкнуть мышью на пункте Файл, в открывшемся меню раскрыть (щелчком мыши) подменю Открыть и уже в этом подменю выбрать (снова щелчком мыши) команду Данные. Как вы наверно успели заметить, особым шрифтом набраны имена элементов интерфейса, файлов и переменных.

2 Общий обзор SPSS

22	Запуск программы
23	Кнопки и другие элементы управления
24	Настройка параметров программы
26	Окна программы
31	Окно редактора командного языка Syntax
33	Окно вывода и его редактирование
38	Сохранение, экспорт, перенос и печать результатов

Как мы упоминали в главе 1, необходимые начальные знания пользователя ограничиваются умением включать компьютер и представлением о рабочем столе Windows. В этой главе описаны все действия, позволяющие работать с программой SPSS. Вы узнаете, как ее запускать, как переключаться между программами с помощью панели задач и т. д. В результате у вас появятся начальные навыки работы с программой SPSS.

Если вы свободно владеете компьютером и уже знакомы с SPSS, часть материала этой главы можно просмотреть «по диагонали». Тем не менее каждому читателю и даже искушенному пользователю SPSS следует обратить внимание на разделы «Окно редактора командного языка Syntax», «Окно вывода и его редактирование» и «Сохранение, экспорт, перенос и печать результатов», поскольку в них содержится важная информация, касающаяся специфики SPSS. Особое значение при работе с программой имеет окно вывода, а также возможности его редактирования и сохранения, так как это относится к результатам анализа данных. Поэтому, по мере освоения SPSS, целесообразно возвращаться к соответствующим разделам этой главы.

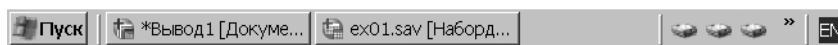
Запуск программы

Пожалуй, основное, что вам необходимо (по крайней мере, в рамках темы этой книги), — знать, как запустить программу SPSS. Для этого на большинстве компьютеров вам следует, щелкнув на кнопке Пуск (Start), последовательно переместить указатель мыши сначала на пункт Программы (Programs), а затем на пункт IBM SPSS Statistics. Когда последний пункт окажется выделенным, раскроется программная группа пакета. Щелкните мышью на команде IBM SPSS Statistics 19. После этого программа SPSS запустится.

Помимо запуска SPSS вы также должны уметь переключаться между окнами программ с помощью панели задач. Для работы с пакетом SPSS это весьма актуально, поскольку он представляет собой набор из нескольких одновременно выполняющихся программ, имеющих собственные окна. При запуске SPSS вы видите на экране единственное *окно редактора данных*, однако, как только вы начнете работу с данными, на экране также появится *окно вывода*. В зависимости от выполняемых действий число открытых окон может меняться, однако окна редактора данных и вывода, как правило, присутствуют на экране постоянно.

Иногда пакет SPSS переключается между окнами автоматически; в остальных случаях вам потребуется осуществлять переключение вручную.

Когда несколько программ выполняются одновременно, каждой из них соответствует кнопка на панели задач. Ниже показана панель задач с кнопками окна редактора данных и окна вывода.



Чтобы переключиться в окно нужной программы, достаточно подвести указатель к соответствующей кнопке и щелкнуть на ней.

Кнопки и другие элементы управления

Кнопки используются практически в каждом окне. Они выглядят по-разному. Например, кнопка со стрелкой между окнами, направленной вправо или влево (рис. 2.1), предназначена для перемещения элемента (зачастую имени переменной) из одного списка в другой. А кнопка ОК запускает процедуру обработки.

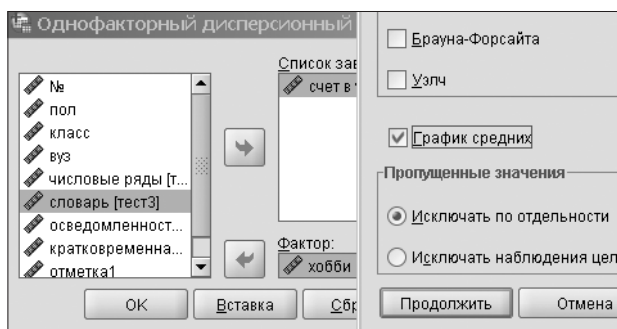


Рис. 2.1. Окно программы с кнопками, флажками и переключателями

Помимо кнопок при работе с диалоговыми окнами вы можете столкнуться с еще двумя типами элементов управления. Первый элемент — *флажок*, который может быть установлен или сброшен. Если вы устанавливаете флажок, рядом с его

названием появляется «галочка». Если «галочка» отсутствует, флажок считается сброшенным. Для изменения состояния флажка необходимо щелкнуть на месте расположения «галочки». Как правило (хотя и не всегда), флажки независимы, то есть установка или сброс одного из них никак не влияет на состояние остальных.

Область окна Пропущенные значения (см. рис. 2.1.) представляет собой группу из двух переключателей. *Переключатели* связаны друг с другом таким образом, что если один из них установлен (отмечен точкой), остальные переключатели в группе считаются сброшенными. Программные переключатели по своему смыслу аналогичны физическим переключателям: если, к примеру, на магнитофоне включен режим проигрывания, режимы перемотки и записи оказываются выключенными. Установленный переключатель обозначается точкой в соответствующем кружке.

Настройка параметров программы

После установки программы целесообразно задать минимальные параметры, облегчающие работу.

Прежде всего создайте на рабочем столе ярлык программы SPSS. Предполагается, что программа SPSS еще не запущена.

ШАГ 1

1. Выделите в главном меню Windows команду Пуск ► Программы ► IBM SPSS Statistics ► IBM SPSS Statistics 19 и щелкните по ней правой кнопкой мыши. В открывшемся меню щелкните левой кнопкой мыши на команде Копировать.
2. Переместите курсор на свободное место Рабочего стола, щелкните правой кнопкой мыши и выберите команду Вставить ярлык.

Теперь вы можете запускать программу SPSS при помощи значка на рабочем столе.

ШАГ 2

Запустите программу SPSS при помощи значка на рабочем столе, дважды щелкнув на нем левой кнопкой мыши, и в открывшемся после запуска программы диалоговом окне выбора режима работы SPSS («Что требуется выполнить?») щелкните на кнопке Отмена (Cancel).

После выполнения этого шага на экране появится окно редактора данных SPSS. В зависимости от версии интерфейс SPSS может быть русскоязычным или англоязычным. В последних версиях программы (начиная с SPSS 17) возможно задание русскоязычного интерфейса и вывода в рамках диалогового окна Параметры (Options).

ШАГ 3

В командной строке (вверху) выберите команду Правка (Edit) ► Параметры (Options).

Откроется окно Параметры (Options) (рис. 2.2), имеющее множество вкладок.

На вкладке Общие (General) задайте русский язык интерфейса в разделе Интерфейс пользователя (User interface) и русский язык вывода в разделе Вывод (Output).

На вкладке Местоположение файлов следует указать папку, к которой программа SPSS будет обращаться по умолчанию при открытии и сохранении файлов. Для файлов примеров создайте папку SPSSWORK на диске D:\ и укажите пути для нее в разделе Папки по умолчанию для диалоговых окон открытия и сохранения файла, для файлов данных и для других файлов.

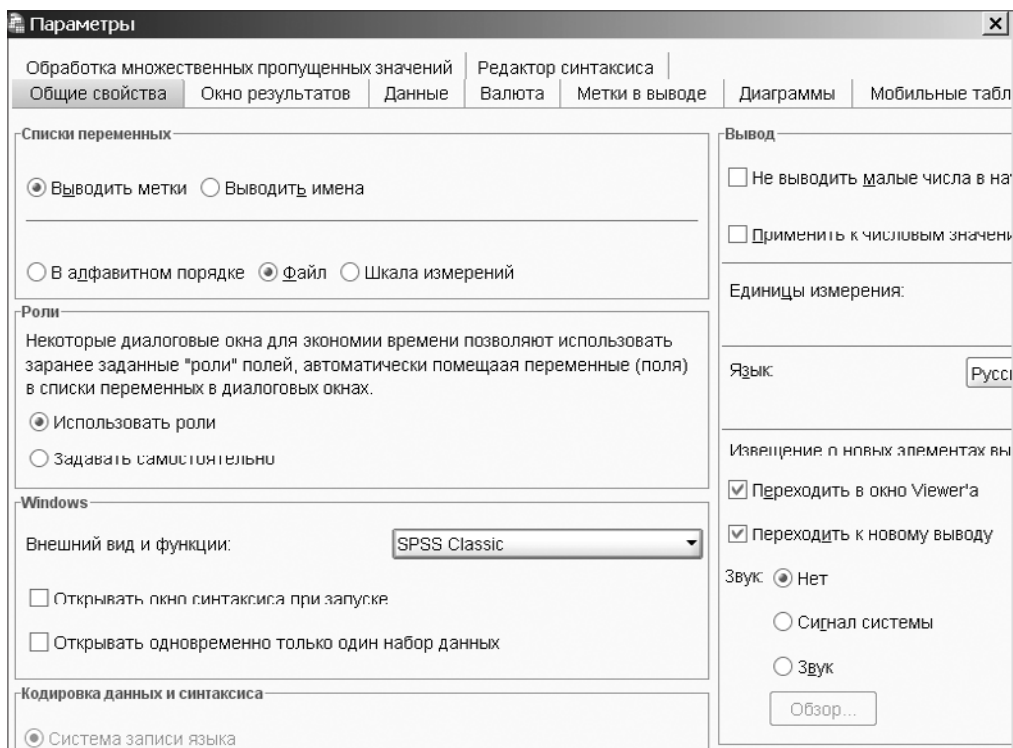


Рис. 2.2. Окно Параметры

На вкладке Окно результатов в левом нижнем углу снимите флажок Выводить команды в журнал, тем самым избавляясь от излишне частого появления окна вывода. Затем, по мере освоения программы, вы можете вернуть флажок на место — это обеспечит постоянную регистрацию и сохранение всех манипуляций с данными.

На вкладке Диаграммы в разделе Текущие для пункта Параметры цикла стиля можно установить Цикл только по узорам — тогда все графики будут черно-белыми, различаясь узором. Можно также задать стиль графиков, например, в соответствии с рекомендациями АРА (Американской психологической ассоциации). Для этого в разделе Шаблон диаграмм следует установить переключатель в положе-

ние Использовать шаблон диаграмм и указать путь к месту, где этот шаблон расположен. Если установлена версия SPSS 19, — это путь C:\Program Files\IBM\SPSS\Statistics\19\Looks\APA_Styles.sgt. Этот шаблон под именем APA_Styles.sgt также находится и в папке файлов примеров.

На вкладке Мобильные таблицы задаются параметры для таблиц результатов в окне Вывод. В правом нижнем углу измените Режим редактирования по умолчанию на Редактировать все таблицы во Viewer. Эта установка позволяет редактировать все таблицы в окне вывода, не открывая дополнительных окон редактирования таблиц. По желанию, можно выбрать предпочитаемый вид таблиц в разделе Шаблон таблиц.

Окна программы

Существует три типа окон, которые при работе с пакетом SPSS используются чаще других.

- ▶ Главное окно программы (окно редактора данных) появляется при запуске SPSS, с него начинается работа с пакетом.
- ▶ Диалоговое окно Открытие файла позволяет получить доступ к ранее созданным файлам.
- ▶ Базовые диалоговые окна (обработки данных), хотя и зависят от конкретной процедуры, имеют схожие элементы интерфейса.

Помимо этих трех типов окон особое значение имеет окно вывода. Окно вывода появляется каждый раз после окончания обработки данных. Оно содержит результаты обработки, а также краткие пояснения по их интерпретации. Мы подробно рассмотрим работу с окном вывода и печатью результатов далее в этой главе.

Дополнительные возможности управления данными и их обработки предоставляет окно редактора командного языка Syntax. Работа с этим окном — удел весьма продвинутых пользователей SPSS. Однако мы дадим лаконичное описание этого окна, так как оно позволяет выполнять действия, находящиеся за пределами стандартного кнопочного интерфейса SPSS.

А сейчас обратим внимание на окно редактора данных, или главное окно, которое появляется на экране при запуске SPSS.

Главное окно

При запуске программы SPSS на фоне неактивного главного окна появляется окно выбора режима работы SPSS (рис. 2.3). Чтобы это окно не появлялось в дальнейшем, установите флажок Не показывать это диалоговое окно в будущем. Для того чтобы перейти в главное окно SPSS (рис. 2.4), в окне выбора режима работы нажмите кнопку Отмена.

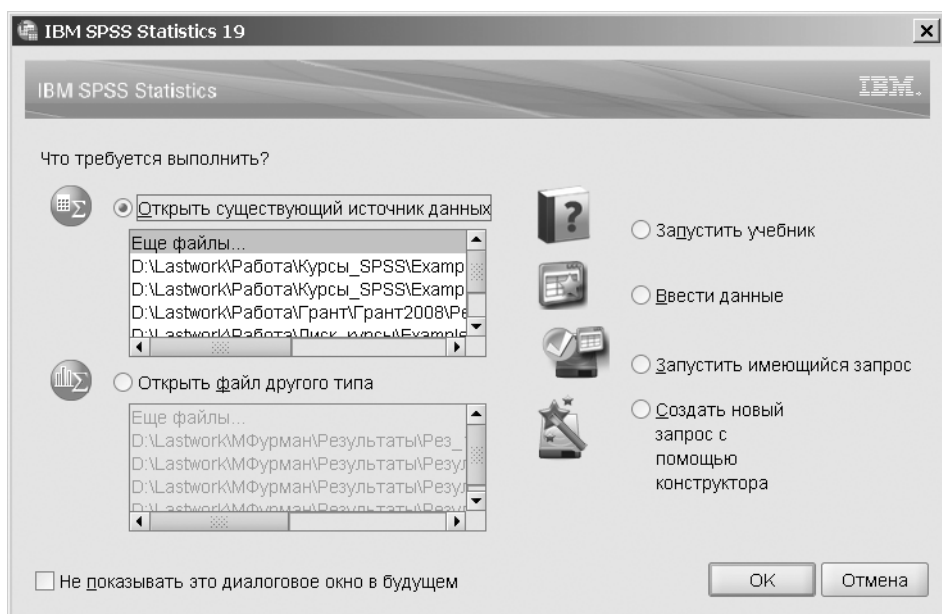


Рис. 2.3. Окно выбора режима работы

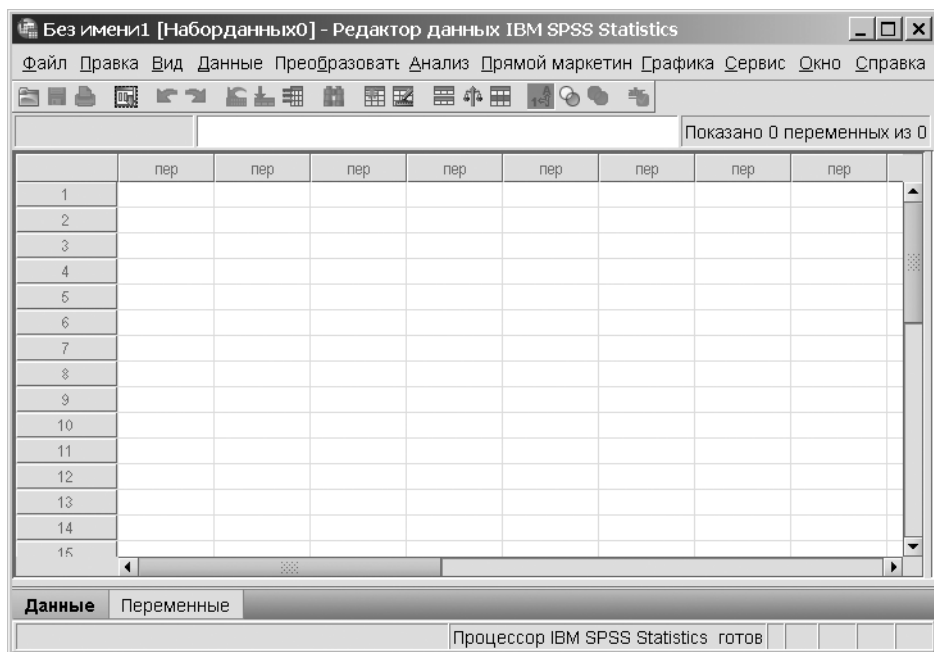


Рис. 2.4. Окно редактора данных SPSS

В верхней части окна редактора данных расположены строка меню с командами и панель инструментов. В момент запуска SPSS редактор данных пуст. Для того чтобы ввести данные, необходимо либо набрать их вручную, заполнив нужные ячейки, либо открыть существующий файл данных. Первый способ подробно описан в главе 3, а в этой главе мы займемся открытием файлов.

Панель инструментов расположена под строкой меню и содержит набор кнопок с различными значками. Щелчок на такой кнопке приводит к выполнению некоторой операции. Назначение кнопки высвечивается при подведении к ней курсора мыши. Как правило, панель инструментов позволяет выполнять без помощи меню те операции, которые требуются наиболее часто. Обратите внимание, что в различных приложениях SPSS вид панели инструментов может быть разным. Чтобы убедиться в этом, достаточно сравнить панели инструментов окна данных и окна вывода (см. далее раздел «Окно вывода и его редактирование»).

Как вы, вероятно, успели заметить, некоторые кнопки выглядят блекло. Это означает, что соответствующая команда в данный момент недоступна пользователю. Так, кнопка Печать недоступна потому, что в программе нет данных, которые можно было бы напечатать. Как только вы введете данные, кнопка Печать сразу станет доступной. Для того чтобы приобрести навыки работы с панелями инструментов, мы рекомендуем вам самостоятельно «пощелкать» на различных кнопках и понаблюдать за реакцией программы.

Строка меню содержит команды для выполнения почти всех операций, предусмотренных в программе SPSS. По мере изучения этой книги вы познакомитесь с большим количеством команд. Как правило, выполнение команды начинается с появления диалогового окна, в котором пользователю предлагается установить значения параметров. Ниже приводится краткое описание основных меню, которые будут рассмотрены в этой книге.

- ▶ **Файл.** Содержит команды, предназначенные для открытия, чтения и сохранения файлов, а также команду выхода из программы SPSS.
- ▶ **Правка.** В этом меню находятся команды редактирования, такие как команды копирования, вставки, замены, поиска и т. п.
- ▶ **Вид.** Содержит набор команд, влияющих на представление информации на экране. Наиболее часто используются команды **Метки значений** и **Шрифты**.
- ▶ **Данные.** Здесь находятся команды, предназначенные для управления вводом и представлением данных.
- ▶ **Преобразовать.** Содержит команды, модифицирующие введенные данные, а также создающие новые данные на основе существующих.
- ▶ **Анализ.** С этого меню начинаются все процедуры анализа данных.
- ▶ **Графика.** Здесь находятся команды, позволяющие создавать различные диаграммы. Иногда с помощью команд этого меню задаются параметры статистических процедур.

- ▶ Окно. С помощью этого меню можно управлять взаимным расположением и статусом открытых окон программы SPSS. Фактически, меню Окно служит средством переключения между окнами, альтернативными панели задач.
- ▶ Справка. Как следует из названия меню, оно предназначено для доступа к справочной информации. Весьма полезно при изучении этой книги было бы время от времени обращаться к справочной системе, чтобы расширять получаемые знания.

Диалоговое окно открытия файла

Диалоговое окно Открыть данные, показанное на рис. 2.5, является стандартным окном операционной системы и позволяет открывать ранее созданные файлы данных. Для того чтобы вызвать это окно, выберите в меню Файл команду Открыть ▶ Данные либо щелкните мышью на кнопке Открыть данные панели инструментов. Содержимое открытого таким образом окна будет соответствовать папке, с которой вы работали ранее, либо папке, заданной в качестве рабочего каталога (папка, открываемая программой по умолчанию).

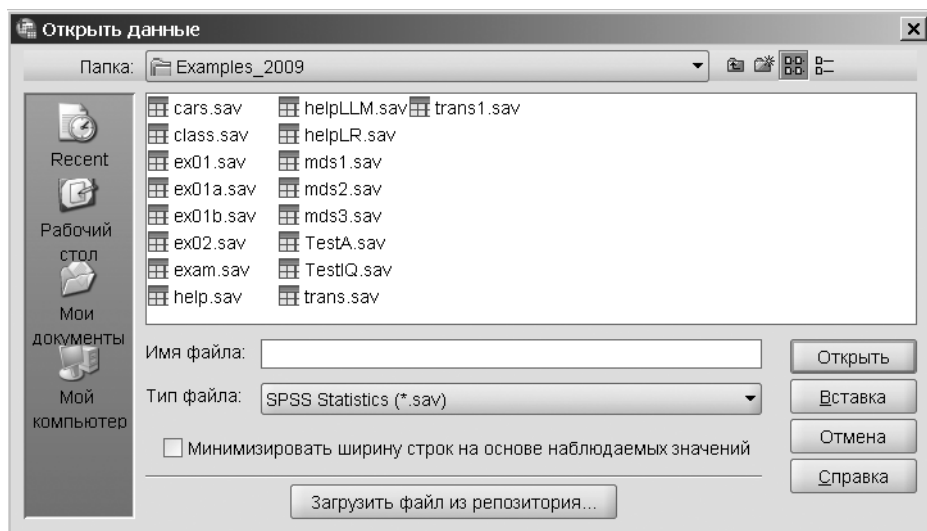


Рис. 2.5. Диалоговое окно Открыть данные

Чтобы сделать описание окна более наглядным, воспользуемся файлом данных первого из примеров с именем ex01.sav. Обратите внимание, что все файлы, созданные с помощью редактора данных SPSS, имеют расширение .sav, которое, в зависимости от настройки вашего компьютера, может отображаться или не отображаться в диалоговом окне.

Существует несколько способов открытия файлов. Если нужный вам файл находится в списке диалогового окна Открыть данные, для его открытия достаточно дважды щелкнуть на нем мышью. Вы также можете ввести имя файла с клавиатуры в поле Имя файла и щелкнуть на кнопке Открыть. При этом, если файл не находится в текущей папке, необходимо указать полный путь к нему.

Если файл не был создан в редакторе данных SPSS, перед открытием необходимо указать его тип. Для этого щелкните на стрелке раскрывающегося списка Тип файла и выберите тип, соответствующий вашему файлу. К примеру, если файл создавался программой Excel, нужно выбрать пункт Excel (*.xls). После задания типа файла вы можете воспользоваться одним из двух вышеописанных способов его открытия.

В результате открытия файла часть ячеек окна редактора данных заполнится данными, которые можно обрабатывать с помощью команд меню.

Начиная с SPSS версии 15.0 появилась возможность открывать одновременно несколько файлов данных (до этого для открытия нового или другого файла данных необходимо было завершить работу с предыдущим файлом). При этом каждому открытому файлу данных соответствует свое окно редактора данных. Например, можно сначала открыть файл ex01.sav, а затем файл ex01b.sav. Тогда окажутся открытыми два окна редактора данных: ex01.sav [Наборданных1] — Редактор данных IBM SPSS Statistics (неактивное окно) и ex01b.sav [Наборданных2] — Редактор данных IBM SPSS Statistics (активное окно). Нажимая соответствующие кнопки на панели задач, можно переключаться между этими окнами редактора данных. Возможность работать одновременно с несколькими файлами данных существенно повышает эффективность применения SPSS.

Диалоговое окно процедуры обработки

Каждая процедура обработки имеет собственное диалоговое окно. Несмотря на это, практически все диалоговые окна построены по одному и тому же принципу. Мы продемонстрируем это на примере диалогового окна процедуры Частоты, показанного на рис. 2.6 (см. также главу 6).

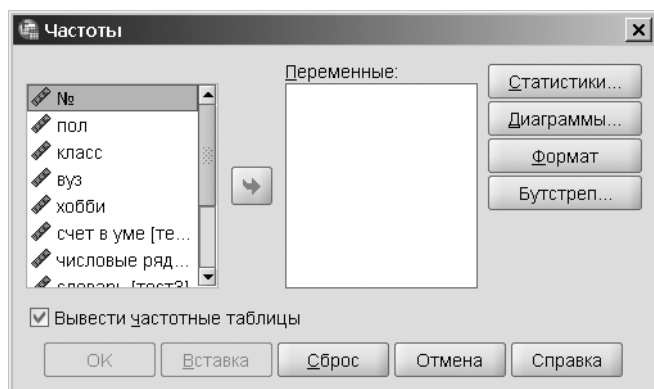


Рис. 2.6. Диалоговое окно Частоты

В левой части окна расположен список, содержащий все доступные переменные файла; справа могут находиться от одного до трех изначально пустых списков. Несмотря на то что заголовки этих списков различны, все они предназначены для указания переменных, участвующих в процедуре обработки. Между списками находится кнопка со стрелкой, позволяющая перемещать переменные из одного спи-

ска в другой. Стрелка на кнопке соответствует направлению перемещения. Если выделить пункт в левом списке, стрелка будет направлена вправо, если в правом списке — влево.

В диалоговых окнах всех процедур обработки имеются 5 кнопок.

- ▶ Кнопка ОК инициирует процедуру обработки выбранных переменных.
- ▶ Кнопка Вставка открывает окно командного языка Syntax, рассмотренное далее в этой главе.
- ▶ Кнопка Сброс возвращает параметры к значениям по умолчанию. Каждый раз при проведении обработки набор исходных данных сохраняется, что позволяет использовать его многократно с разными значениями параметров. Это удобно, если вы проводите серию статистических опытов с одними и теми же переменными. Если же вам необходимо изменить начальные условия, вы сможете очистить список переменных и установить значения параметров, принятые по умолчанию, щелкнув на кнопке Сброс.
- ▶ Кнопка Отмена позволяет немедленно закрыть окно без выполнения процедуры обработки.
- ▶ Кнопка Справка открывает доступ к *контекстной* помощи. Это означает, что, щелкнув на этой кнопке в окне Частоты, вы получите описание именно процедуры Частоты, а не общую справку по всем процедурам обработки.

Наконец, кнопки Статистики, Диаграммы, Формат и Бутстреп, находящиеся в нижней части диалогового окна, позволяют задавать различные параметры, управляющие представлением результатов процедуры и ее выполнением. Почти все диалоговые окна имеют подобные кнопки. Число кнопок и параметры, которые они позволяют задавать, меняются в зависимости от конкретной процедуры.

Окно редактора командного языка Syntax

Работа с окном редактора командного языка Syntax предоставляет более широкие возможности, чем настройка параметров команды с помощью кнопок диалогового окна. Как правило, подобные возможности доступны только опытным пользователям в специфических случаях. А для решения подавляющего большинства задач, особенно в начале освоения SPSS, достаточно стандартных средств интерфейса программы. Тем не менее, по мере освоения SPSS, мы будем обращаться к возможностям синтаксиса в главе 9 «Корреляции» и главе 21 «Кластерный анализ».

В дальнейшем, освоив обработку данных при помощи стандартных средств интерфейса SPSS, вы можете открыть для себя поистине безграничные возможности управления данными, все чаще обращаясь к командному языку Syntax. В этом вам могут помочь многочисленные советы и рекомендации, содержащиеся на сайте А. Балабанова «Raynald's SPSS Tools по-русски» (<http://www.spsstools.ru/>).

Окно редактора командного языка Syntax предназначено для ввода, редактирования и выполнения команд Syntax. Это окно имеет текстовый формат записи команд и позволяет открывать, создавать и сохранять командные файлы, которые можно сохранять и использовать в дальнейшем. Такие файлы имеют расширение *.sps, и один из них содержится среди файлов примеров к этой книге (файл Syn_clust.sps, пример к главе 21). Командный файл может быть создан или открыт тремя способами.

Самый простой способ создания командного файла — кнопка Вставка любого диалогового окна процедуры обработки (см. рис. 2.6). Эта возможность рассмотрена в главе 9 при создании нестандартной таблицы корреляций. После задания в диалоговом окне процедуры обработки всех необходимых параметров (переменных, процедур, статистик, и пр.) достаточно нажать на кнопку Вставка, чтобы открылось окно редактора командного файла. Например, если в диалоговом окне команды Частоты задать переменные пол, класс, вуз (глава 6, шаг 5), но вместо кнопки ОК, запускающую эту команду, нажать кнопку Вставка, откроется окно редактора командного файла (рис. 2.7). Выбор через командную строку этого окна команды Запуск ► Все приведет к выполнению действия, аналогичного нажатию кнопки ОК соответствующего диалогового окна процедуры обработки.

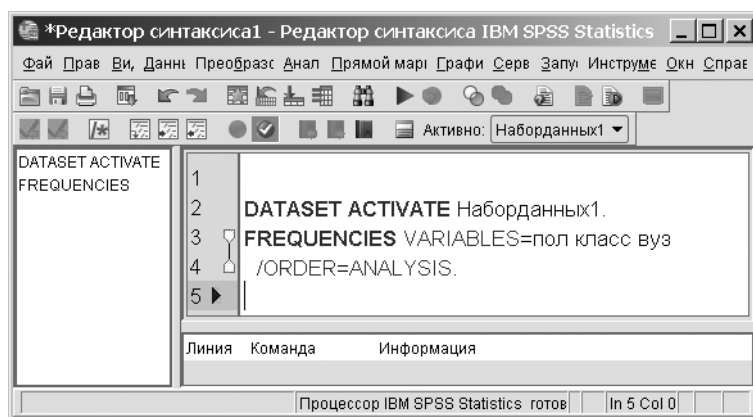


Рис. 2.7. Диалоговое окно редактора синтаксиса

Можно воспользоваться уже готовым командным файлом, выбрав в меню Файл главного окна (редактора данных) команду Открыть ► Синтаксис. Таким образом вы сможете открыть файл примера для кластерного анализа Syn_clust.sps (глава 21).

Наконец, можно самому создать новый командный файл, выбрав в меню Файл главного окна (редактора данных) команду Новый ► Синтаксис. В открывшемся (пустом) окне редактора вы можете вручную ввести необходимую синтаксическую конструкцию. А можно и вставить ее, предварительно скопировав из любого текстового редактора или с сайта «Raynald's SPSS Tools по-русски» (<http://www.spsstools.ru/>).

Для выполнения открытого командного файла достаточно выбрать через командную строку редактора командного языка Syntax команду Запуск ► Все. Перед выполнением командный файл можно отредактировать по своему усмотрению, до-

бавляя или убирая переменные, команды, процедуры и т. д. Созданный командный файл Syntax при желании можно сохранить для дальнейшего использования.

Следует отметить, что необходимость обращения к командному языку SPSS обычно возникает лишь после достаточно уверенного освоения стандартного интерфейса программы — для облегчения управления данными, выполнения длинных последовательностей команд или реализации редких методов анализа данных. Тем не менее с самого начала освоения SPSS следует делать первые шаги по изучению командного языка. Для начала достаточно при выполнении статистических процедур чаще пользоваться кнопкой Вставка, с последующим выбором команды Запуск ► Все, это позволит поближе познакомиться с языком Syntax.

Окно вывода и его редактирование

Под термином «вывод» в этой книге, как правило, понимается совокупность результатов любой процедуры обработки. Каждая глава, в которой упоминается вывод данных, обязательно снабжена указаниями о содержании окна вывода. Поскольку это окно отражает главный результат обработки данных, уделим ему особое внимание.

Пример окна вывода представлен на рис. 2.8. С подобными результатами обработки можно работать и без редактирования, сразу выводя их на печать, однако следует иметь в виду следующие моменты.

- При выводе больших объемов данных требуется большое количество бумаги, и пространство печати зачастую используется неэффективно.
- Как правило, не вся выводимая информация оказывается нужной.
- Большие таблицы, создаваемые программой SPSS, неудобны для восприятия, поэтому желательна их реорганизация.
- В отчет, который, как правило, готовится в формате Word, включается лишь часть информации, представленной в выводе.

Рассмотрим подробно возможности работы с окном вывода SPSS и его панелью инструментов. Это позволит вам в дальнейшем удалять ненужные фрагменты полученных данных, добавлять свои комментарии и реорганизовывать порядок вывода.

Для знакомства с файлом вывода используем результат обработки из главы 8 — таблицу сопряженности пол × хобби × класс. Желательно, чтобы все описываемые операции вы проделали сами, открыв файл примера cross.sprv через строку меню Файл ► Открыть ► Вывод....

Некоторые кнопки на панели инструментов окна вывода совпадают с кнопками окна редактора данных. Эти кнопки выполняют аналогичные действия; разница состоит лишь в том, что в редакторе данных обрабатываются исходные данные, а в окне вывода — результаты.

Строго говоря, окно вывода на самом деле не является окном. Как видно из рис. 2.8, результаты анализа отображены в окне Вывод — Viewer IBM SPSS Statistics — специальной программы просмотра результатов анализа. Окно Вывод разделено на две области. Правая область представляет собой то, что мы назвали окном вывода, то есть окно результатов анализа данных. В левой области отображена иерархическая структура объектов окна вывода. В сущности, весь вывод программы — заголовки, таблицы, диаграммы и т. п., на самом деле являются не чем иным, как разного рода объектами, имеющими иерархическую структуру. Таким образом, вместе с результатами анализа мы видим их внутреннее программное представление.

Сводка обработки наблюдений

	Наблюдения					
	Валидные		И пропущенные		Итого	
	N	Процент	N	Процент	N	Процент
пол * хобби * класс	100	100,0%	0	,0%	100	100,0%

Таблица сопряженности пол * хобби * класс

Частота

класс			хобби			Итого
			спорт	компьютер	искусство	
А	пол	ЖЕН	6	5	3	14
		МУЖ	11	7	1	19
	Итого		17	12	4	33
Б	пол	ЖЕН	6	6	9	21
		МУЖ	5	7	2	14
	Итого		11	13	11	35
В	пол	ЖЕН	3	8	15	26
		МУЖ	2	4	0	6
	Итого		5	12	15	32
Итого	пол	ЖЕН	15	19	27	61
		МУЖ	18	18	3	39
	Итого		33	37	30	100

Рис. 2.8. Окно вывода SPSS

Если вы внимательно посмотрите на содержимое окна Вывод, то заметите, что объект Примечания, отображенный в иерархической структуре, отсутствует в окне вывода. Это объясняется тем, что по умолчанию он является *скрытым*. Слева от объектов расположены значки, принимающие вид открытой или закрытой книги в зависимости от того, скрыты или отображены эти объекты в данный момент.

Если дважды щелкнуть на таком значке, состояние объекта изменится на противоположное.

Иерархическая структура делает удобной навигацию внутри окна вывода. Так, если вам нужно просмотреть содержимое какой-либо таблицы, достаточно щелкнуть на объекте, чтобы соответствующее изображение появилось на экране. Как нетрудно догадаться, операция удаления фрагментов вывода сводится к удалению соответствующих объектов. Для удаления объекта щелкните на его имени в левой части окна, а затем в меню Правка выберите команду Удалить. Если вы хотите изменить порядок следования объектов, выделите один или несколько объектов, которые намереваетесь переместить, а затем в меню Правка выберите команду Вырезать. После этого выделите объект, за которым вы хотите осуществить вставку, и выберите команду Правка ► Вставить после.

Уже после нескольких операций обработки одних и тех же данных навигация внутри окна вывода становится затруднительной даже при иерархической структуре объектов. Для решения проблемы используется сворачивание объектов. Как можно видеть на рис. 2.8, *слева*, около названий объектов имеется значок «-» (минус). Если вы не хотите, чтобы объекты под каким-либо из заголовков присутствовали на экране, щелкните на этом значке. После этого названия соответствующих объектов исчезнут, на схеме останется лишь заголовок, а вместо значка «-» (минус) появится значок «+» (плюс). Для того чтобы снова развернуть список объектов, достаточно щелкнуть на этом значке, и содержимое схемы будет восстановлено.

Теперь, когда вы получили элементарные знания о том, что представляет собой окно вывода программы SPSS и как с ним работать на объектном уровне, пришло время заглянуть внутрь самих объектов и научиться изменять их. Нас будет интересовать структура таблиц и диаграмм, поскольку от того, насколько удобно представлены эти элементы вывода, зависит удобочитаемость выводимых данных в целом. В этой книге мы не будем рассматривать все процедуры, изменяющие представление таблиц и диаграмм, однако знаний, которые вы получите, вполне достаточно для комфортной работы с результатами исследований. Редактирование диаграмм описывается в главе 5.

В первую очередь необходимо уяснить принцип построения таблиц. Обычно SPSS создает таблицы автоматически, самостоятельно решая, какие показатели разместить в строках, а какие — в столбцах. Конечно, не всегда удобно полагаться на выбор программы, и для этих случаев SPSS предлагает средства редактирования таблиц. Следует отметить, что средства редактирования таблиц SPSS значительно совершеннее возможностей Word.

Чтобы попрактиковаться в редактировании таблиц, обратимся снова к таблице сопряженности пол × хобби × класс (из файла примера cross.spv). Если вы успели поработать с этим файлом, закройте его без сохранения и вновь откройте, чтобы вернуться к его прежнему виду. В таблице сопряженности (из файла cross.spv), приведенной на рис. 2.9, SPSS по умолчанию размещает данные о хобби учащихся в столбцах, а о поле и классе — в строках. Несмотря на то что в подобном представлении нет ничего неправильного, можно представить себе ситуации, когда для

наглядности потребовалось бы расположить данные в таблице по-другому. При этом сами данные необходимо оставить неизменными, поскольку они являются результатом анализа.

Таблица сопряженности пол * хобби * класс						
Частота						
класс			хобби			Итого
			спорт	компьютер	искусство	
1	пол	ЖЕН	6	5	3	14
		МУЖ	11	7	1	19
	Итого		17	12	4	33
2	пол	ЖЕН	6	6	9	21
		МУЖ	5	7	2	14
	Итого		11	13	11	35
3	пол	ЖЕН	3	8	15	26
		МУЖ	2	4	0	6
	Итого		5	12	15	32

Рис. 2.9. Исходная таблица сопряженности

В столбцах рассматриваемой таблицы перечислены хобби учащихся (спорт, компьютеры, искусство), а каждая строка содержит два уровня заголовков: номер класса (1, 2, 3) и пол ученика (ЖЕН, МУЖ). Для каждого заголовка первого уровня перебраны все возможные варианты заголовков второго уровня. Данная таблица является однослойной (слой — класс), поэтому мы сразу перейдем к организации столбцов и строк.

Для того чтобы производить какие-либо действия с таблицей, сначала необходимо открыть ее для редактирования. Для этого следует навести на нее курсор и щелкнуть правой кнопкой мыши. В открывшемся меню следует выбрать раздел Изменить содержимое. Доступны два режима редактирования: В средстве просмотра — непосредственно в окне вывода; В отдельном окне — в открывшемся окне Мобильная таблица. В окне вывода удобно редактировать небольшие таблицы, а для больших таблиц лучше выбирать пункт В отдельном окне. При выборе любого из этих пунктов на экране исчезнут панели инструментов, а в строке меню появится новое меню Мобильная таблица. Кроме того, изменится набор доступных команд меню. Далее будут рассмотрены команды Повернуть внутренние метки столбцов, Повернуть наружные метки строк, Транспонировать строки и столбцы и Поля вращения меню Мобильная таблица и Формат.

Вероятно, вы уже обратили внимание на то, что ширина столбцов рассматриваемой таблицы гораздо больше, чем необходимо для отображения содержащихся в ней чисел. Это объясняется тем, что дополнительное пространство занято именами (метками) столбцов. Если изменить направление текста меток столбцов, то есть применить к таблице команду Формат ► Повернуть внутренние метки столбцов, формат таблицы станет гораздо более компактным (рис. 2.10).

Для того чтобы восстановить предыдущий формат, достаточно выбрать указанную команду еще раз. Аналогичным образом можно изменять формат строк, только в меню Формат выбирать команду Повернуть наружные метки строк.

Как уже говорилось, при создании таблицы содержимое ее строк и столбцов назначается программой автоматически. Если вам необходимо изменить расположение категорий в таблице, следует воспользоваться командой Транспонировать строки и столбцы или Поля вращения. Первая меняет местами строки и столбцы; результат ее применения к рассматриваемой таблице показан на рис. 2.11.

Таблица сопряженности пол * хобби * класс									
Частота									
			хобби			Итого			
			спорт	компьютер	искусство				
класс	1		ЖЕН	6	5	3	14		
			МУЖ	11	7	1	19		
	Итого			17	12	4	33		
2	1		ЖЕН	6	6	9	21		
			МУЖ	5	7	2	14		
	Итого			11	13	11	35		
3	1		ЖЕН	3	8	15	26		
			МУЖ	2	4	0	6		
	Итого			5	12	15	32		

Рис. 2.10. Таблица сопряженности после поворота меток столбцов

Таблица сопряженности пол * хобби * класс													
Частота													
		класс											
		1				2				3			
		пол		Итого		пол		Итого		пол		Итого	
хобби	спорт	ЖЕН	МУЖ			ЖЕН	МУЖ			ЖЕН	МУЖ		
		6	11	17		6	5	11		3	2	5	
		5	7	12		6	7	13		8	4	12	
		3	1	4		9	2	11		15	0	15	
хобби	компьютер	3	1	4		9	2	11		15	0	15	
		3	1	4		9	2	11		15	0	15	
		3	1	4		9	2	11		15	0	15	
		3	1	4		9	2	11		15	0	15	
хобби	искусство	3	1	4		9	2	11		15	0	15	
		3	1	4		9	2	11		15	0	15	
		3	1	4		9	2	11		15	0	15	
		3	1	4		9	2	11		15	0	15	
Итого		14	19	33		21	14	35		26	6	32	
		14	19	33		21	14	35		26	6	32	
		14	19	33		21	14	35		26	6	32	
		14	19	33		21	14	35		26	6	32	

Рис. 2.11. Таблица сопряженности после перестановки строк и столбцов

Вторая команда позволяет осуществить более гибкую реорганизацию таблицы. При ее выборе на экране появляется новое окно Поля вращения, показанное на рис. 2.12, в котором вы можете задать содержимое будущих строк и столбцов. Команда Поля вращения также предназначена для реорганизации слоев.

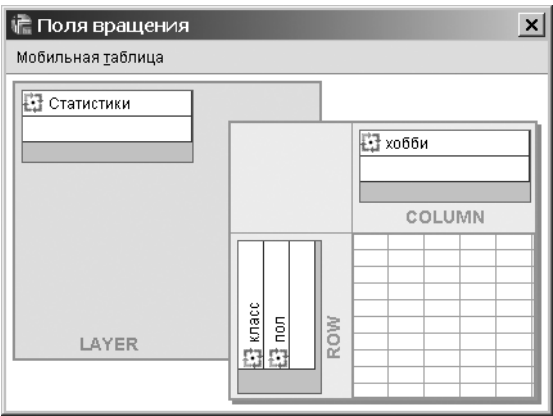


Рис. 2.12. Окно Поля вращения

Реорганизация таблицы производится с помощью осей. Под *осью строк* будем понимать столбец, содержащий метки строк (ROW), а под *осью столбцов* — строку, содержащую метки столбцов (COLUMN). В нашем примере на оси столбцов нахо-

дится информация о номере класса и поле учащихся, а на оси строк — об их хобби, что соответствует таблице на рис. 2.12.

В окне Поля вращения имеются панели Row (Строка) и Column (Столбец), соответствующие оси строк и оси столбцов. Вы можете менять положение меток, перетаскивая мышью (удерживая нажатой ее левую кнопку) цветной квадратик с названием переменной (пол, класс, хобби) в пределах одной оси или с одной оси на другую. Если вы хотите, к примеру, поменять местами класс и пол учащихся, достаточно перетащить в окне Поля вращения значок переменной пол в пустую ячейку правее переменной класс. В этом случае пол окажется заголовком первого уровня, и каждая метка переменной пол будет содержать три метки переменной класс.

Можно также реорганизовывать переменные, перемещая соответствующие квадратные значки с именами переменных между осями строк и столбцов. Так, если вам необходимо сделать пол переменной второго уровня для каждого значения переменной хобби, следует перетащить переменную пол под значок переменной хобби.

Часто возникает необходимость удаления лишнего столбца или строки таблицы. Для этого достаточно выделить заголовок щелчком левой кнопки мыши и затем открыть меню щелчком правой кнопки. В раскрывшемся меню нужно выбрать команду Выбрать ► Ячейки данных и метки. Будет выделена соответствующая область таблицы. Наведя курсор на выделенную область, правой кнопкой мыши снова откройте меню и выберите Вырезать или Очистить для удаления.

Как вы уже успели заметить, даже для такой простой таблицы, как эта, существует большое количество вариантов реорганизации. Самый простой способ научиться реорганизовывать таблицы — поэкспериментировать с ними, воспользовавшись примером cross.srv или своими результатами.

Сохранение, экспорт, перенос и печать результатов

После редактирования содержимого окна вывода в соответствии с вашими предпочтениями у вас, вероятно, возникнет желание сохранить результат для дальнейшего использования.

Сохранение результатов

В отличие от файла данных, файл вывода имеет расширение .srv. Для сохранения файла вывода в меню File (Файл) выберите команду Save (Сохранение), в появившемся диалоговом окне задайте имя файла и щелкните на кнопке Save (Сохранить)¹.

¹ В более ранних версиях файлы вывода имели расширение .spo. Для того чтобы иметь возможность открывать такие файлы вывода, созданные в более ранних версиях SPSS, необходимо дополнительно установить модуль Legacy Viewer. Его можно скачать с сайта [www.spss.com](http://support.spss.com/ProductsExt/SPSS/Documentation/Statistics/LegacyViewer/LegacyViewer_English.zip) по ссылке http://support.spss.com/ProductsExt/SPSS/Documentation/Statistics/LegacyViewer/LegacyViewer_English.zip

Помимо сохранения, программа SPSS позволяет *экспортировать* вывод в другие форматы (например, Word), *перенести* его в другой документ или *напечатать* весь файл вывода или его выделенные фрагменты. Перед этим настоятельно рекомендуется обдумать, стоит ли переносить или печатать те или иные фрагменты вывода, поскольку искать полезную информацию среди большого объема ненужных таблиц, как правило, достаточно сложно.

Чтобы *выделить* для экспорта, переноса или печати фрагмент выводимых данных, выполните в окне вывода любое из трех перечисленных ниже действий.

- ▶ Щелкните на заголовке или объекте иерархической структуры, расположенной в левой части окна вывода. При щелчке на заголовке выбранными оказываются все объекты, расположенные под этим заголовком.
- ▶ Тем же способом, описанным в предыдущем пункте, выделите один объект, а затем, удерживая клавишу Shift, выделите другой объект. В результате выбранными окажутся все объекты, расположенные между двумя указанными объектами.
- ▶ Способом, описанным в первом пункте, выделите один объект, а затем, удерживая клавишу Ctrl, поочередно выделите все объекты, предназначенные для печати. Этот способ отличается от предыдущего тем, что позволяет выбирать несмежные объекты.

Экспортирование результатов

Начиная с версии 11.5, программа SPSS предоставляет возможность сохранения файлов вывода и в более привычном для многих формате текстового редактора Word. Для этого в меню Файл выберите команду Экспортировать. На экране появится диалоговое окно Экспортировать вывод, показанное на рис. 2.13. В этом окне в разделе Документ в списке Тип необходимо выбрать пункт Word/RTF (*.doc). В разделе Имя файла ввести имя нового файла и задать место его расположения, щелкнув на кнопке Обзор. Дополнительно с помощью группы переключателей Экспортируемые объекты можно выбрать объекты для экспорта, обычно — Все видимые или Выбранные. А при помощи кнопки Изменить параметры (в разделе Документ) можно установить более тонкие настройки параметров экспортирования документа. После окончания настройки в этом диалоговом окне щелкните на кнопке ОК. Будет создан файл с расширением .doc, содержащий те объекты файла вывода, которые вы выбрали. Этот файл можно будет открывать и редактировать в программе Word.

Перенос результатов в документ Word

Наиболее популярный способ сохранения содержимого окна вывода — перенос его выделенной части в документ Word. Для этого сначала следует скопировать выделенную часть окна вывода в буфер обмена при помощи, например, правой кнопки мыши и команды Копировать.

Доступны два варианта вставки скопированного фрагмента в документ Word. При выборе команды Вставить в документе Word таблица будет вставлена как обычная таблица Word, доступная для редактирования. Недостаток такого способа за-

ключается в том, что средства редактирования таблиц SPSS заметно совершеннее возможностей Word, поэтому такой перенос может нарушить структуру таблицы. Второй вариант — перенести таблицу как рисунок. Для этого, после копирования таблицы вывода, в документе Word в разделе Правка (в командной строке) следует выбрать команду Специальная вставка..., и указать там Точечный рисунок. Перенос таблицы из SPSS в Word как рисунка гарантирует сохранность формата таблицы, но лишает возможности редактировать таблицу.

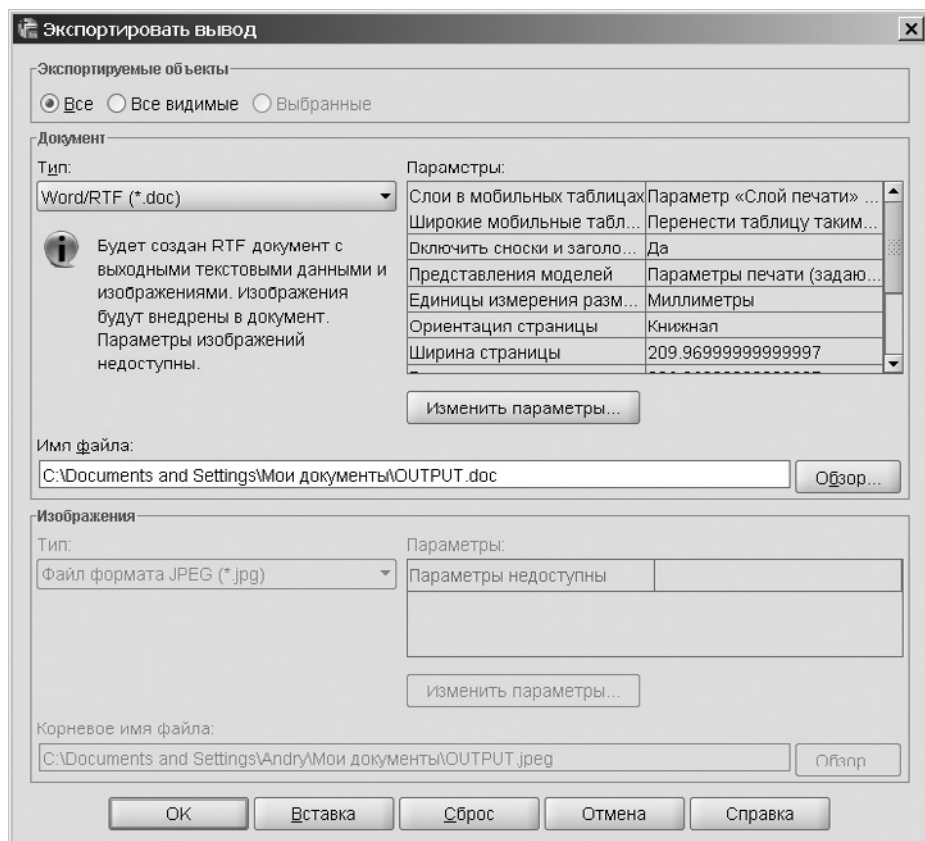


Рис. 2.13. Диалоговое окно Экспортировать вывод

Печать результатов

После выделения требуемых для печати объектов в меню Файл выберите команду Предварительный просмотр для проверки расположения результатов на бумаге. При необходимости поменяйте ориентацию листов с книжной на альбомную. Щелкните на кнопке Печатать. На экране появится системное диалоговое окно, с помощью которого можно выбрать принтер (в случае если на компьютере установлено более одного принтера), а также диапазон печати, определяющий, следует печатать все данные или только выделенные фрагменты. Когда все параметры установлены нужным образом, щелкните на кнопке ОК для завершения процесса.

3 Создание и редактирование файлов данных

41	Структура файла данных
52	Ввод данных
54	Редактирование данных
56	Пример файла данных

Эта глава поможет вам сделать первый шаг к статистическому анализу — ввести данные в программу. Как вы уже знаете из предыдущей главы, в SPSS для управления данными используется специальная программа — редактор данных. Окно редактора данных представляет собой электронную таблицу, в которой столбцы отражают переменные, а строки — объекты (случаи). На пересечении строки и столбца находится ячейка, в которой сохраняется значение переменной (столбца) для данного объекта (строки). Для работы с таблицами, не уместяющимися в пределах окна, предусмотрены вертикальные и горизонтальные полосы прокрутки, а также автоматические функции перехода к заданной переменной или объекту.

Данная глава является в книге первой, содержащей пошаговые процедуры. В них приведена последовательность действий, необходимых для ввода данных в компьютер. В качестве примера мы будем использовать файл `ex01.sav`, упоминавшийся в главе 2 и содержащий разнообразную информацию о 100 школьниках, обучающихся в трех классах. Как уже упоминалось, готовые файлы примеров можно загрузить с веб-сайта <http://www.piter.com/download>.

Структура файла данных

Перед тем как начать ввод данных, необходимо определить структуру будущего файла. Для этого вы должны ответить себе на вопрос, как будут использоваться в анализах те или иные переменные. К сожалению, многие пользователи начинают задумываться об этом гораздо позже, чем следует, и в этом кроется причина многих ошибок при проведении анализа данных: чем сложнее план исследования, тем больше шансов на то, что из-за необдуманных действий исследователя он завершится неудачей.

Итак, первое, что следует сделать, — определить последовательность действий при обработке данных. Кроме того, необходимо четко представлять себе структуру

и взаимосвязи переменных в вашем плане исследования. Ниже перечислены наиболее характерные для файлов данных ошибки и недостатки.

- ▶ Отсутствуют ключевые переменные (пол, возраст и т. п.), являющиеся основой для анализа.
- ▶ Переменная плохо отражает содержание соответствующей реальной величины (например, на сложный вопрос предусмотрено только два варианта ответа: «да» и «нет»).
- ▶ При большом количестве независимых переменных отсутствуют зависимые переменные, отражающие цель исследования (или наоборот).
- ▶ Недостаточно независимых переменных, влияющих на заданную зависимую переменную.

Эти примеры демонстрируют, что залогом успеха как исследования в целом, так и создания файла данных в частности является тщательно продуманный выбор структуры данных.

Следует отметить, что структура файла данных должна соответствовать плану исследования. С другой стороны, план исследования должен быть составлен так, чтобы его исходные данные можно было бы обработать в соответствии с задачами и гипотезами исследования. Самый оптимальный и простой путь обеспечения этих соответствий — определение структуры данных на этапе планирования исследования, еще до их сбора. Это позволит избежать большинства типичных ошибок, относящихся как к планированию исследования, так и к организации данных.

На этапе планирования исследования структура данных может быть задана в виде предварительного списка переменных с указанием их типов и диапазонов возможных значений, например так, как в табл. 3.1.

Таблица 3.1. Предварительный список переменных

№	Название	Тип	Диапазон возможных значений
1	Идентификационный номер	Номинальная	1–100
2	Пол	Номинальная	1 — жен; 2 — муж
3	Класс	Номинальная	1 — А; 2 — Б; 3 — В
4	Предполагаемый для поступления вуз	Номинальная	1 — гуманитарный; 2 — экономический; 3 — технический; 4 — естественно-научный
5	Внешкольные увлечения	Номинальная	1 — спорт; 2 — компьютер; 3 — искусство
6–10	Показатели тестов 1–5	Количественные	1–20
11	Средний балл за 10-й класс	Количественная	3–5
12	Средний балл за 11-й класс	Количественная	3–5

Каждая переменная — это имеющее значение для исследователя основание, позволяющее отличать объекты друг от друга. На предварительном этапе следует выделять два типа переменных: количественные и категориальные (номинальные). Количественная переменная позволяет различать объекты по уровню выраженности некоторого свойства, например: средний балл отметки, тестовый показатель и пр. Идентификация количественных переменных на предварительном этапе не составляет труда: обычно они соответствуют тому, что исследователь намеревается измерить. Второй тип — категориальные (номинальные) переменные. Обычно они используются как основания для деления объектов (испытуемых) на группы или категории: пол, класс, возрастная категория, уровень дохода и пр. Типичная ошибка начинающего исследователя — игнорирование возможных оснований для деления объектов на группы как самостоятельных номинальных переменных в структуре данных.

Важным свойством номинальных переменных является возможность их представления в виде набора целых чисел. Например, трем видам внешкольных увлечений (хобби) учащихся (спорт, компьютер, искусство) можно сопоставить числа 1, 2, и 3 соответственно. Числовое представление данных в компьютерных программах всегда предпочтительнее символьного, поскольку обработка чисел происходит быстрее, проще и с меньшей вероятностью ошибок. Кроме того, числовое представление легко модифицировать: вы можете переназначить числа, соответствующие созданным элементам, а также (что часто требуется на практике) без проблем включить в анализ новые элементы. Например, если в группе окажется учащийся, увлечение которого не соответствует перечисленным, будет полезно включить в переменную хобби категорию с названием другие и присвоить ей число 4. Эта операция рассмотрена подробнее в главе 4.

Порядок создания переменных также важен при вводе данных. Здесь следует придерживаться простого правила: наиболее важные и часто используемые переменные должны помещаться в начало файла, для остальных данных вопрос порядка следования не столь важен, однако рекомендуется объединять их в группы по их «физическому смыслу». Чаще всего в начало файла следует поместить категориальные переменные, которые далее предполагается использовать для деления объектов (испытуемых) на группы, например пол, семейное положение и пр. Далее можно перечислять остальные сведения, а логическое объединение переменных производить в зависимости от того, какие аспекты они отражают.

Ниже мы приведем инструкции, с помощью которых вы сможете приступить к освоению программы SPSS.

ШАГ 1

Первое, что необходимо сделать, — запустить программу SPSS. В открывшемся диалоговом окне выбора режима работы щелкните на кнопке Отмена. Вы получите доступ к окну редактора данных (рис. 3.1).

ШАГ 2

Перейдите на вкладку Переменные, щелкнув на ее ярлычке мышью (рис. 3.2).

Вкладка Данные, которая отображается сразу после запуска редактора, предназначена для ввода значений в создаваемый файл данных. Вкладка Переменные позволяет задать структуру файла данных, то есть определить имена, метки и структуры переменных. Заголовки столбцов представляют собой параметры каждой из переменных.

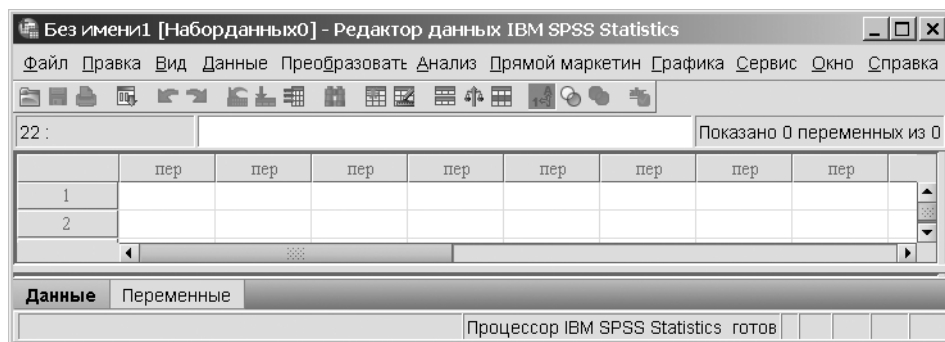


Рис. 3.1. Исходный вид окна редактора данных SPSS

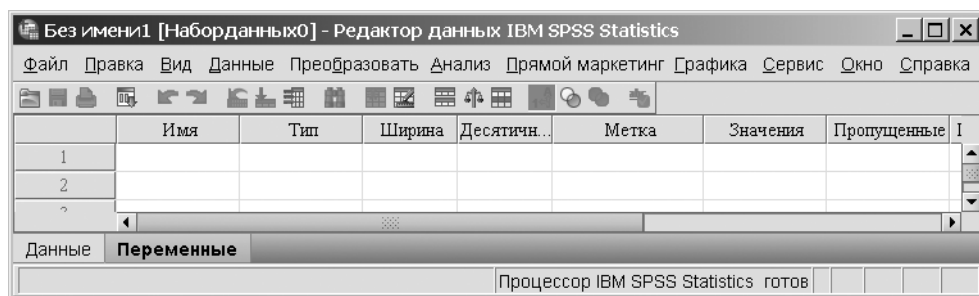


Рис. 3.2. Вкладка просмотра переменных окна редактора данных

В следующих пошаговых инструкциях мы возьмем за основу данные из файла ex01.sav. Часть содержимого этого файла приведена в конце этой главы. В процессе создания файла можно выделить три основных действия: задание имени переменной, определение ее параметров и ввод данных. Необязательно последовательно выполнять каждый из трех этапов для каждой переменной; такой порядок выбран лишь в качестве примера. На практике вам может оказаться удобнее создавать файл данных «порциями»: сначала вы полностью зададите параметры части переменных, введете их значения, затем зададите новую группу переменных и т. д. Мы же сейчас обратимся к рассмотрению параметров переменных.

Имя переменной

Параметр Имя определяет имя переменной. Чтобы задать имя первой переменной, просто введите его с клавиатуры в текущую ячейку. Имя второй переменной вводится в том же столбце под именем первой, то есть во второй строке; имя третьей переменной — в третьей строке и т. д. Для перемещения между строками пользуйтесь клавишами ↓ и ↑.

Если вы пользуетесь одной из ранних версий SPSS (до 12), возможно, вам не удастся воспользоваться кириллицей для задания имен переменных. Ничего страшного: задавайте имена и параметры переменных латинскими буквами.

ШАГ 3

Для задания имен всем переменным файла ex01.sav выполните следующие действия (должна быть открыта вкладка Переменные редактора данных).

1. Введите символ № и нажмите клавишу ↓.
2. Введите слово пол и нажмите клавишу ↓.
3. Введите слово класс и нажмите клавишу ↓.

Аналогичным образом вводятся оставшиеся имена переменных.

Как вы, наверное, уже обратили внимание, при переходе на следующую строку (то есть при окончании ввода имени переменной) оставшиеся параметры переменной автоматически заполняются значениями по умолчанию.

Имя переменной не является произвольным. Существует ряд соглашений, которым оно должно удовлетворять:

- ▶ длина имени — не более 64 символов (в ранних версиях — до 8 символов);
- ▶ в имени могут использоваться любые буквы, цифры, символы @, #, ., _, \$, однако имя всегда должно начинаться с буквы, а символ «.» (точка) не может стоять в конце имени;
- ▶ имена всех переменных должны быть разными;
- ▶ буквы верхнего и нижнего регистров символов не различаются, то есть имена ID, id, Id и iD воспринимаются программой как идентичные;
- ▶ имена переменных не должны совпадать с каким-либо из зарезервированных слов (all, ne, eq, to, le, lt, by, or, gt, and, not, ge, with).

Чтобы не запоминать все ограничения, может быть, вам будет проще пользоваться следующим простым правилом именования переменных. Имя должно быть коротким (не более 8 символов), начинаться с буквы и содержать только буквы, цифры и знак подчеркивания. И конечно, имена не должны повторяться.

Тип переменной

Параметр Тип определяет тип переменной. Если фокус ввода находится в столбце Имя какой-либо из созданных переменных, переместите его в столбец Тип, например, нажав клавишу Tab, тогда в правой части ячейки появится кнопка с многоточием. Щелчок на ней приводит к появлению на экране диалогового окна Тип переменной, представленного на рис. 3.3.

Как видите, текущим типом переменной является тип Числовая. В подавляющем большинстве случаев вам придется иметь дело именно с числовыми данными. В тех редких случаях, когда значения переменных представляют собой буквы или буквосочетания (слова), необходимо установить переключатель Текстовая. Текст-

вые данные, в отличие от числовых, могут включать буквы и другие символы, то есть нести текстовую информацию. Это могут быть ФИО респондентов, города и т. п. В частном случае строковая переменная может хранить число, однако обработка такого «числа» будет производиться так, как будто оно является текстом. *Как правило, строковые переменные не подлежат обработке. Поэтому их следует избегать, за исключением редких случаев, например, когда данная переменная содержит имена людей или названия городов.*

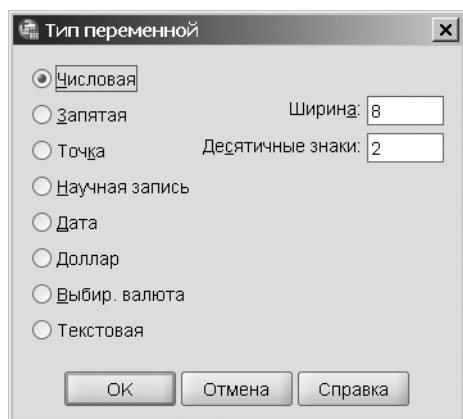


Рис. 3.3. Диалоговое окно Тип переменной

Необходимость в выборе строкового типа может возникнуть в случае, когда данные переносятся из программы Excel в SPSS путем копирования и вставки. Если значения какой-либо из переменных являются буквенными (например, «м» и «ж»), перед вставкой этой переменной необходимо изменить ее тип на строковый. В противном случае буквенные значения вставлены не будут.

Остальные 6 типов переменных, которые можно задать с помощью переключателей, присутствующих в диалоговом окне, практически не используются при обработке данных, и поэтому мы не будем на них останавливаться.

Поскольку все переменные имеют установленный по умолчанию тип Числовая, никаких дополнительных действий с ними производить не нужно.

Ширина переменной

Параметр Ширина позволяет задать максимальное количество знаков, которое может иметь значение переменной, включая дробную часть. Например, ширина переменной № (идентификатор) составляет 3 знака, поскольку числа, которые мы намерены вводить, содержат от 1 до 3 знаков. На практике заранее определить ширину переменной гораздо труднее, поскольку не всегда известно, какие данные нам понадобятся вводить в будущем. Поэтому следует задавать ширину переменной с гарантированным запасом: ее можно ограничить потом, после ввода данных.

В большинстве случаев нет необходимости менять заданную по умолчанию ширину переменной.

Зададим для всех переменных ширину 3 знака.

ШАГ 4

Для задания ширины переменных выполните следующие действия.

1. С помощью клавиш со стрелками переместите фокус ввода в ячейку столбца *Ширина*, соответствующую переменной №, и задайте значение 3. В ячейке должно остаться значение ширины, равное 3.
2. Наведите курсор мыши на ячейку столбца *Ширина*, соответствующую переменной №, нажмите правую кнопку мыши и выберите *Копировать*.
3. При помощи мыши вставьте скопированное значение для всех остальных переменных. Для этого выделите ячейки столбца *Ширина* соответствующих переменных, нажмите правую кнопку мыши и выберите *Вставить*.

Дробная часть числа

Параметр *Десятичные* предназначен для задания числа десятичных знаков после запятой в случае, если тип переменной допускает использование дробных чисел. Для строковых переменных значение в ячейке *Десятичные* автоматически устанавливается равным нулю, а для числовых переменных — равным 2. У строковых переменных значение параметра *Десятичные* недоступно для изменения.

Когда фокус ввода оказывается в ячейке столбца *Десятичные*, справа появляются две кнопки счетчика. Щелкая на этих кнопках, вы можете изменять текущее количество знаков после запятой. Изначально у всех числовых переменных оно равно 2.

В файле *ex01.sav* знаки после запятой необходимы для переменных *отметка1* и *отметка2*, причем количество этих знаков равно 2 и задано программой по умолчанию. Таким образом, задача заключается в том, чтобы установить для всех остальных числовых переменных значение параметра *Десятичные* равным 0. Отметим, что это не является необходимым, а служит для удобства отображения данных на экране или в отчете.

ШАГ 5

Для задания числа десятичных знаков в переменных выполните следующие действия.

1. С помощью клавиш со стрелками переместите фокус ввода в ячейку столбца *Десятичные*, соответствующую переменной №, и дважды щелкните на нижней кнопке счетчика. В ячейке должно остаться значение, равное 0.
2. С помощью клавиш со стрелками переместите фокус ввода в ячейку столбца *Десятичные*, соответствующую переменной *пол*, и дважды щелкните на нижней кнопке счетчика. В ячейке должно остаться значение, равное 0.
3. Аналогичные действия повторите для остальных переменных, кроме переменных *отметка1* и *отметка2*.

Метки переменных

С помощью параметра *Метка* можно создать метку переменной. Как правило, метка используется в тех случаях, когда смысл переменной недостаточно точно отражен в названии. По сути, метка — это комментарий к имени переменной. При желании вы можете отобразить метки переменных в окне вывода вместо их имен.

Ячейки столбца *Метка* представляют собой обычные текстовые поля, в которые вы можете вводить текст меток. Длина метки не должна превышать 256 символов. Ограничений на используемые символы нет. Но помните о том, что слишком длинные метки ухудшают удобочитаемость фрагментов, в которых они присутствуют. Как показывает практика, 20–30 символов вполне достаточно, чтобы описать назначение переменной. И чем короче метка, тем лучше.

Если какая-либо из надписей окажется длиннее текущей ширины столбца, последняя автоматически увеличится до необходимого размера. Если вам не обязательно видеть все надписи целиком, вы можете задать ширину столбца вручную. Для этого подведите указатель мыши к правой границе ячейки с меткой столбца — указатель примет вид двух стрелок, направленных в разные стороны. Нажмите левую кнопку мыши и, удерживая ее нажатой, перетащите правую границу столбца на новое место. Аналогичным способом можно перетаскивать границу любых столбцов и тем самым оформлять вкладку *Переменные* по своему вкусу.

ШАГ 6

Для ввода меток выполните следующие действия.

1. С помощью клавиш со стрелками переместите фокус ввода в ячейку столбца *Метка*, соответствующую переменной *тест1*, введите словосочетание *счет в уме* и нажмите клавишу *Enter*.
2. С помощью клавиш со стрелками переместите фокус ввода в ячейку столбца *Метка*, соответствующую переменной *тест2*, введите словосочетание *числовые ряды* и нажмите клавишу *Enter*.
3. Для переменной *тест3* введите слово *словарь*, для переменной *тест4* — слово *осведомленность*, для переменной *тест5* — словосочетание *кратковременная память*.

Метки значений переменных

Параметр *Значения* позволяет управлять наименованиями уровней (категорий) переменной. Под уровнем, или категорией, понимается целочисленное значение переменной, имеющее определенный смысл. Например, переменная *пол* имеет два уровня: 1 — *жен* (женский) и 2 — *муж* (мужской). Буквосочетания или слова, поставленные в соответствие уровням переменной, например *жен* и *муж*, называются *метками значений* и отражают смысл разных значений переменной для исследователя. Наличие меток значений исключительно важно для удобочитаемости результатов анализа, поскольку очень трудно запомнить смысл значений всех переменных, даже если их сравнительно немного. В крупных демографических исследованиях количество переменных может превышать 100, поэтому метки значений являются

практически незаменимым средством расшифровки результатов статистического анализа. SPSS также позволяет включать метки значений в файл данных, что очень удобно при его создании. Длина метки значения ограничивается 60 символами, однако, как правило, для пользователей достаточно 3–10 символов.

Как и в случае столбца Тип, перемещение в ячейку столбца Значения приводит к появлению кнопки с многоточием. Если щелкнуть на ней, на экране появится диалоговое окно Метки значений, в котором можно задать метки для различных значений переменной (рис. 3.4).

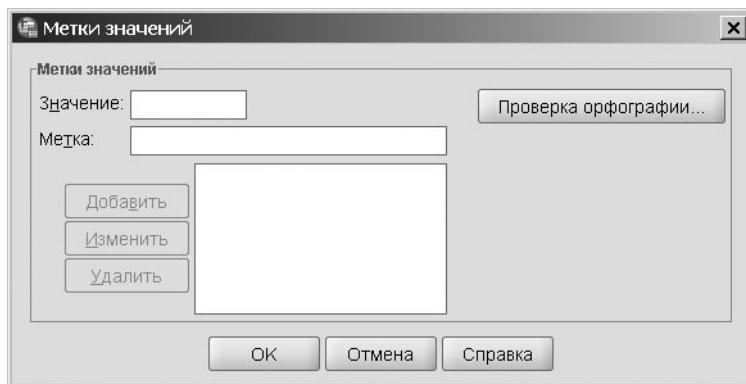


Рис. 3.4. Диалоговое окно Метки значений

ШАГ 7

Ниже в качестве примера приведены действия, позволяющие задать метки значений для переменной пол.

1. С помощью клавиш со стрелками переместите фокус ввода в ячейку столбца Значения, соответствующую переменной пол, и щелкните мышью на кнопке с многоточием.
2. В поле Значение введите число 1, нажмите клавишу Tab для перехода в поле Метка, введите туда слово жен и щелкните на кнопке Добавить.
3. В поле Значение введите число 2, нажмите клавишу Tab для перехода в поле Метка, введите туда слово муж и щелкните на кнопке Добавить.
4. Для закрытия диалогового окна Метки значений щелкните на кнопке ОК.

После задания меток значений переменных появляется еще одна удобная возможность изменения вида окна вкладки Данные. Если в меню Вид установить флажок рядом с командой Метки значений, численные значения переменных будут заменены заданными метками. Для того чтобы вернуться к численным значениям переменных, достаточно сбросить флажок рядом с командой Метки значений.

Пропуски

Параметр Пропуски используется очень редко, поскольку программа и так позволяет учитывать пропуски в данных. Необходимость в этом параметре возника-

ет, когда требуется различать причины пропусков значений. Например, пропуск в данных может быть обусловлен тем, что респондент еще не опрошен или, может быть, он отказался отвечать на данный вопрос. Так, в отношении переменной вуз (предполагаемый для поступления вуз) нам необходимо различать тех учащихся, которых мы не успели опросить, и тех, которые еще не определились. Для еще не опрошенных учащихся мы будем оставлять пустую ячейку, а не определившихся учащихся будем обозначать цифрой 9. После щелчка на кнопке с многоточием, появляющейся при перемещении фокуса ввода в ячейку столбца Пропуски, открывается диалоговое окно Пропущенные значения, в котором можно задать обозначения пропусков. Если поставить переключатель Отдельные пропущенные значения и ввести число, например, 9, это значение не будет использоваться в дальнейшем при обработке переменной, наряду с пустыми ячейками.

Столбцы

Параметр Ширина столбца, в отличие от своего соседа слева, требуется для всех переменных. С его помощью можно управлять шириной (в символах) столбцов вкладки Данные. Если вы зададите широкие столбцы, это позволит полностью видеть имена переменных, однако число одновременно видимых переменных будет невелико. Напротив, узкие столбцы позволяют наблюдать на экране много переменных сразу, но их имена могут не уместиться в столбце целиком. Например, при ширине столбцов в 3 символа на экране можно увидеть 33 переменные, а при ширине в 8 символов — всего 12. Опытные пользователи знают, что работа с данными тем удобнее, чем больше их видно одновременно. Поэтому вы часто будете сталкиваться с задачей размещения на экране как можно большего объема данных с сохранением их удобочитаемости. Пожалуй, основной принцип, которым следует руководствоваться: чем меньше знаков занимает большинство значений некой переменной, тем меньшей следует задавать ширину соответствующего столбца.

Для изменения параметра Ширина столбца используется счетчик. По умолчанию для всех переменных значение равно 8.

ШАГ 8

Для примера ниже приведены действия по изменению ширины столбцов для переменных №, пол, класс.

1. С помощью клавиш со стрелками переместите фокус ввода в ячейку столбца Ширина столбца, соответствующую переменной №, и несколько раз щелкните на нижней кнопке счетчика, пока не останется значение 4.
2. С помощью клавиш со стрелками переместите фокус ввода в ячейку столбца Ширина столбца, соответствующую переменной пол, и несколько раз щелкните на нижней кнопке счетчика, пока не останется значение 6.
3. С помощью клавиш со стрелками переместите фокус ввода в ячейку столбца Ширина столбца, соответствующую переменной класс, и три раза щелкните на нижней кнопке счетчика. В ячейке должно остаться значение 5. Подобным образом подберите ширину оставшихся столбцов, проверяя результат на вкладке Данные.

Выравнивание

Параметр Выравнивание позволяет управлять расположением данных внутри ячейки. Доступ к раскрывающемуся списку возможен, только если фокус ввода оказался в соответствующей ячейке — теперь можно выбрать один из трех типов выравнивания: По правому краю, По левому краю и По центру. По умолчанию значения числовых переменных выравниваются по правому краю, строковых — по левому. Для каждой переменной вы можете самостоятельно задать тип выравнивания.

Шкала измерения

Как и у предыдущего параметра, значение параметра Шкала выбирается в раскрывающемся списке из трех доступных: Количественная, Порядковая и Номинальная. Каждое из этих значений несет дополнительную информацию о типе данных переменной.

- Значение Количественная указывается для количественных типов данных, допускающих арифметические операции. Примером может служить возраст человека: для него имеют смысл операции вычитания, сложения и деления. В файле ex01.sav это значение параметра Шкала можно указать для переменных тест1, ..., тест5, отметка1, отметка2.
- Значение Порядковая указывается для количественных данных, отражающих измеренное качество на уровне порядка и не допускающих основных арифметических операций. Например, упорядочивание учащихся по сообразительности или кодирование ответов по степени выраженности (1 — «не нравится», 2 — «нейтрально», 3 — «нравится») — это измерения в порядковой шкале. К подобным типам можно отнести и многие дихотомические переменные (переменные, имеющие лишь два различных значения: 0 или 1, да или нет и т. п.).
- Значение Номинальная указывается для категориальных (неколичественных) типов данных, не отражающих количество измеряемого свойства. В файле ex01.sav к таким данным относятся переменные №, пол, класс, вуз, хобби.

Иногда может оказаться затруднительным сделать выбор между значениями Количественная и Порядковая. Не отчаивайтесь, если попадете в такую ситуацию: во всех статистических процедурах оба этих значения параметра Шкала сказываются на результатах анализа одинаковым образом.

Чаще всего нет необходимости менять значение параметра Количественная, заданное по умолчанию. Тем более что выбор значения Номинальная ограничивает набор возможных методов обработки данных.

Роль

Параметр Роль позволяет задавать роли, которые могут быть использованы для предварительного выбора переменных для анализа. При открытии некоторых из диалоговых окон анализа переменные, которые удовлетворяют требованиям роли, будут автоматически отображаться в списках назначения. Доступные роли:

- Входная. Переменная будет использоваться в качестве входной (например, предиктор, независимая переменная).

- ▶ Целевая. Переменная будет использоваться в качестве выходной или целевой (например, независимая переменная).
- ▶ Двойного назначения. Данная переменная будет использоваться в качестве входной и выходной.
- ▶ Не создавать. Данная переменная не имеет назначения роли.
- ▶ Разделения. Данная переменная будет использоваться, чтобы разделить данные на отдельные выборки для обучения, испытания и проверки.
- ▶ Расщепления. Включены для двустороннего совмещения с SPSS Modeler.

По умолчанию всем переменным назначается роль Входная, включая данные из форматов внешних файлов и файлов данных из более ранних версий SPSS. Назначение роли влияет только на диалоговые окна, которые поддерживают назначение роли. Это не оказывает влияния на командный синтаксис.

Задание некоторых ролей ограничивает использование переменной для анализа. Поэтому оставим для всех переменных роль, заданную по умолчанию: Входная.

Ввод данных

Ввод самих данных гораздо проще, чем задание параметров переменных, поэтому самое трудное в этой главе уже позади. В этом разделе мы опишем два способа непосредственного ввода данных в программу, а также процедуру переноса данных из программы Excel. Но сначала рассмотрим действия, необходимые для сохранения файла.

В процессе ввода рекомендуется время от времени производить сохранение файла во избежание случайной порчи или утери введенных данных. Обратите внимание на то, что перед вводом данных следует перейти на вкладку Данные, показанную ранее на рис. 3.1.

ШАГ 9

В случае если вы сохраняете файл в первый раз, потребуется задать его имя.

1. В меню Файл выберите команду Сохранить как.... На экране появится диалоговое окно Сохранить данные как.

Если в этом диалоговом окне нажать кнопку Сохранить, будет сохранено все содержимое файла данных, а если нажать кнопку Переменные..., появится возможность выбрать переменные для сохранения.

2. Введите имя файла и щелкните на кнопке Сохранить.

Перечисленные действия позволяют сохранять файлы лишь в текущей папке. Если вы создали рабочий каталог, текущей является папка C:\SPSSWORK, если нет — папка Мои документы. При необходимости вы можете сохранить файл в любой другой папке обычным способом.

ШАГ 10

Если ваш файл уже был сохранен хотя бы один раз, достаточно выбрать команду Файл ► Сохранить или щелкнуть на кнопке Сохранить панели инструментов.

Существует два метода непосредственного ввода данных в ячейки электронной таблицы: *по строкам (переменным)* и *по столбцам (объектам)*. В зависимости от конкретных данных тот или другой способ оказывается предпочтительным.

Ввод данных *по переменным* предполагает последовательное заполнение всех строк одного столбца значениями одной переменной, затем другой переменной и т. д. Вы вводите значение в левую верхнюю ячейку окна, затем нажимаете клавишу ↓ или Enter, переходя на следующую строку, заполняете еще одну ячейку и продолжаете эти действия до тех пор, пока не заполните весь первый столбец. Затем заполняете столбец, соответствующий второй переменной и т. д., пока все переменные не будут введены. Таким образом, с визуальной точки зрения ввод данных по переменным осуществляется «по вертикали».

Ввод данных *по объектам* заключается в последовательном заполнении каждой строки значениями всех переменных. Сначала вводится значение в левую верхнюю ячейку окна, затем с помощью клавиши → или Tab осуществляется переход в соседний столбец (переменную), вводится значение переменной и т. д. до тех пор, пока вся строка не будет заполнена. После этого аналогичным способом заполняются вторая, третья и все остальные строки. В отличие от предыдущего способа ввод данных по объектам осуществляется «по горизонтали».

Некоторым пользователям привычнее вводить, хранить и редактировать данные в электронной таблице программы Excel. Следует отметить, что программа SPSS лучше приспособлена для этого. Тем не менее перенос данных из Excel в SPSS не составляет труда, если в таблице данных программы Excel строки соответствуют объектам, а столбцы — переменным. Перенос можно осуществить двумя способами.

Первый способ самый простой. Он предполагает предварительную подготовку специального файла Excel. Столбцы этого файла должны соответствовать переменным, строки — объектам. В первой строке должны содержаться имена переменных, удовлетворяющие требованиям SPSS (см. выше). Достаточно открыть такой файл Excel из редактора данных SPSS, чтобы переместить содержащиеся в нем данные в SPSS. Для этого при помощи последовательности команд Файл ► Открыть ► Данные необходимо в диалоговом окне Открыть данные в списке Тип файла выбрать пункт Excel (*.xls).

Второй способ переноса данных из Excel в SPSS несколько сложнее и сводится к копированию значений переменных в таблице Excel и вставке их в таблицу редактора данных SPSS. Сначала необходимо подготовить структуру файла исходных данных в SPSS, как это было описано ранее. Прежде всего обращайтесь внимание на те переменные, которые имеют буквенные значения. Для них вместо принятого по умолчанию числового типа переменной необходимо задать строковый тип, установив в окне Тип, показанном на рис. 3.3, переключатель Текстовая. После задания структуры данных, не закрывая подготовленный файл SPSS, откройте

файл Excel, содержащий данные. Выделите и скопируйте при помощи мыши необходимый блок данных (только значения переменных) в таблице Excel (если эта процедура вызывает у вас затруднения, изучите сначала раздел «Редактирование данных»). Перейдите к файлу SPSS и выделите верхнюю левую ячейку пока еще пустой таблицы данных и в меню Правка выберите команду Вставить. Таблица SPSS будет заполнена теми данными, которые вы ранее скопировали из таблицы Excel. Сходным образом можно также пополнять данные, перенося их из Excel в SPSS, и наоборот.

Редактирование данных

Иногда у вас может возникнуть необходимость изменить данные, уже введенные в файл. Мы рассмотрим, каким образом скорректировать содержимое ячейки, создать новый объект (строку) или переменную (столбец), а также скопировать, удалить и вставить данные.

Изменение содержимого ячейки

Щелкните на ячейке, содержимое которой вы намерены отредактировать, введите новое значение, а затем перейдите в любую соседнюю ячейку с помощью клавиш Enter, Tab или клавиш со стрелками.

Вставка нового объекта

Если вам необходимо вставить новый объект (строку) между двумя соседними строками, щелкните сначала на нижней из них, а затем — на кнопке Вставить наблюдения панели инструментов. В результате будет создана пустая строка, а номера строк, находящихся ниже, увеличатся на единицу.

Вставка новой переменной

Чтобы вставить новую переменную между двумя соседними, щелкните сначала на правой из них, а затем — на кнопке Вставить переменную. Будет создан пустой столбец, а все переменные, находящиеся справа, окажутся сдвинутыми на один столбец.

Копирование и вырезание содержимого ячеек

Для того чтобы скопировать или вырезать содержимое одной или нескольких ячеек, сначала необходимо их выделить. Активная ячейка (то есть ячейка, в которой находится фокус ввода) всегда является выделенной. Чтобы выделить группу ячеек, нажмите левую кнопку мыши на угловой ячейке будущей группы и, удерживая эту кнопку, перетащите указатель к противоположной угловой ячейке группы, после чего кнопку мыши отпустите. Если вам необходимо выделить целиком строку или столбец, щелкните соответственно на нужном объекте или на имени переменной.

Когда желаемые ячейки выделены, в меню Правка выберите команду Копировать или Вырезать. В обоих случаях содержимое ячеек будет скопировано в буфер обмена, однако команда Копировать оставит исходные ячейки нетронутыми, а команда Вырезать очистит их.

Вставка ячеек

Для того чтобы вставить предварительно скопированные в буфер обмена данные, нужно переместить фокус ввода в левую верхнюю ячейку группы, в которую будет осуществляться вставка, и в меню Правка выбрать команду Вставить. Следует помнить о двух неприятностях, связанных с операцией вставки.

- ▶ Если в области, в которую планируется вставить данные, уже содержатся какие-либо другие данные, после выполнения вставки они будут утеряны.
- ▶ При копировании и вставке данных их смысл не анализируется программой, и вы можете случайно заменить значения одной переменной значениями любой другой переменной. Последствия такой вставки могут быть весьма плачевными, поскольку разрешить возникшую путаницу, как правило, весьма непросто.

Мы приводим несколько рекомендаций, которые могут быть полезными при выполнении вставки.

- ▶ Перед удалением и вставкой важных фрагментов данных сохраняйте текущий файл. В случае ошибки вы сможете открыть этот файл снова.
- ▶ По возможности создавайте новые переменные и строки для вставки данных.
- ▶ Как правило, вам придется иметь дело с копированием, вырезанием и вставкой столбцов (переменных) или строк (объектов) целиком. Однако особую аккуратность следует проявлять при вставке *блоков* столбцов или строк.
- ▶ Будьте особенно внимательны при вставке значений переменных разных типов, например, не допускайте вставки значений строковых переменных в столбцы, соответствующие числовым переменным. Это может привести к трудно идентифицируемым ошибкам в процессе статистического анализа.

Поиск данных

Очень удобным вспомогательным средством работы с данными является функция поиска. Выделив заголовок (имя) переменной, в меню Правка выберите команду Найти или щелкните на кнопке Поиск панели инструментов. На экране появится диалоговое окно Найти и заменить — Данные, с помощью которого можно найти заданное значение и (или) заменить их на другое значение (рис. 3.5).

ШАГ 11

Попробуем, к примеру, найти в файле ex01.sav ученика с номером 93.

1. На вкладке Данные щелкните на имени переменной №, затем в меню Правка выберите команду Найти или щелкните на кнопке Поиск панели инструментов.

2. В поле Поиск открывшегося диалогового окна Найти и заменить — Данные введите число 93. Щелкните на кнопке Найти далее, чтобы найти ученика с номером 93.

После завершения поиска щелкните на кнопке Отмена, чтобы закрыть диалоговое окно.

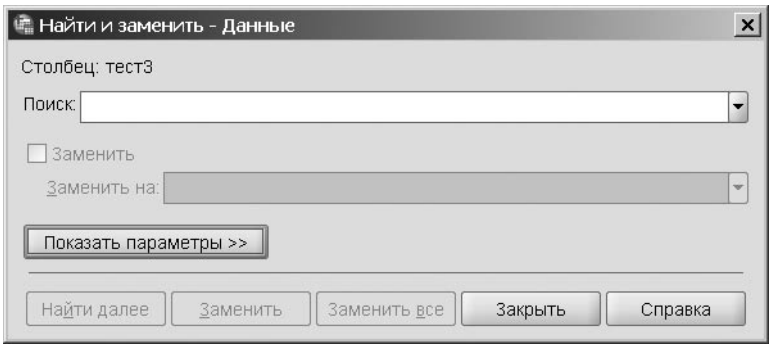


Рис. 3.5. Диалоговое окно поиска

В случае если вы обнаружите недопустимые или неверные значения какой-либо из переменных, можете применить приведенный алгоритм с добавлением флажка Заменить и указанием нужного значения в окне Заменить на: для исправления ошибки.

Пример файла данных

Файл данных ex01.sav содержит информацию об учащихся трех классов. Переменные для этого файла перечислены в табл. 3.2.

Таблица 3.2. Переменные файла данных ex01.sav

Переменная	Описание
№	Идентификационный номер ученика
пол	Пол ученика (1 — ЖЕН, 2 — МУЖ)
класс	Класс, в котором учится школьник (1 — «А», 2 — «Б», 3 — «В»)
вуз	Предполагаемый для поступления вуз: (1 — гуманитарный, 2 — экономический, 3 — технический, 4 — естественно-научный)
хобби	Внешкольное увлечение ученика (1 — спорт, 2 — компьютеры, 3 — искусство)
тест1	Показатель теста «Счет в уме»
тест2	Показатель теста «Числовые ряды»

Переменная	Описание
тест3	Показатель теста «Словарь»
тест4	Показатель теста «Осведомленность»
тест5	Показатель теста «Кратковременная память»
отметка1	Средний балл отметок за 10-й класс
отметка2	Средний балл отметок за 11-й класс

Вкладки Данные и Переменные окна редактора данных для файла ex01.sav показаны соответственно на рис. 3.6 и 3.7.

	№	пол	класс	вуз	хобби	тест1	тест2	тест3	тест4	тест5	отметка1	отметка2
1	1	МУЖ	Б	ЕСТ_Н	искус...	6	7	13	10	14	3,90	4,20
2	2	МУЖ	А	ЕСТ_Н	спорт	8	9	10	11	11	3,55	3,95
3	3	МУЖ	В	ТЕХН	компь...	10	6	10	8	9	3,75	4,65
4	4	ЖЕН	В	ГУМ	компь...	13	9	10	12	6	3,85	3,95
5	5	МУЖ	В	ТЕХН	искус...	12	8	12	18	12	4,20	3,90
6	6	ЖЕН	В	ЭКОН	искус...	12	15	17	11	11	4,25	4,25
7	7	ЖЕН	В	ЭКОН	искус...	6	7	11	16	13	4,45	4,35
8	8	ЖЕН	А	ГУМ	компь...	13	11	10	10	10	3,80	3,90
9	9	ЖЕН	В	ЕСТ_Н	искус...	9	12	14	9	15	3,90	4,00
10	10	ЖЕН	В	ЭКОН	искус...	5	9	13	13	12	4,25	3,75

Рис. 3.6. Фрагмент данных файла ex01.sav

	Имя	Тип	Ширина	Десятич...	Метка	Значения	Пропущенные	Ширина ...	Выравнивание
1	№	Числовая	3	0		Нет	Нет	4	По правом...
2	пол	Числовая	3	0		{1, ЖЕН}...	Нет	4	По правом...
3	класс	Числовая	3	0		{1, А}...	Нет	5	По правом...
4	вуз	Числовая	3	0		{1, ГУМ}...	Нет	4	По правом...
5	хобби	Числовая	3	0		{1, спорт}...	Нет	5	По правом...
6	тест1	Числовая	3	0	счет в умс	Нет	Нет	5	По правом...
7	тест2	Числовая	3	0	числовые ряды	Нет	Нет	5	По правом...
8	тест3	Числовая	3	0	словарь	Нет	Нет	5	По правом...
9	тест4	Числовая	3	0	осведомленность	Нет	Нет	5	По правом...
10	тест5	Числовая	3	0	кратковременна...	Нет	Нет	5	По правом...
11	отметка1	Числовая	3	2		Нет	Нет	8	По правом...
12	отметка2	Числовая	3	2		Нет	Нет	8	По правом...

Рис. 3.7. Переменные файла ex01.sav

4 Управление данными

58	Знакомство с возможностями управления данными
61	Получение информации о файле
61	Обработка пропущенных значений
62	Преобразование данных
67	Выбор наблюдений для анализа
70	Перекодировка в новую переменную
74	Перекодирование существующей переменной
77	Сортировка наблюдений
78	Объединение данных разных файлов
83	Агрегирование данных
85	Реструктурирование данных

Материал этой главы посвящен вопросам эффективной работы с исходными данными. Описанные здесь операции весьма полезны в большинстве случаев обработки и анализа данных, так как практически всегда существует необходимость в предварительной подготовке и преобразовании исходных данных. Поэтому изложенные рекомендации по форматированию данных помогут вам работать с программой гораздо свободнее.

Знакомство с возможностями управления данными

В процессе работы вам могут понадобиться преобразованные данные, являющиеся результатом некоторых действий над исходными данными файла. К примеру, для вашего исследования может представлять интерес среднее значение или сумма баллов по нескольким тестам для каждого учащегося, ранг каждого ученика по успеваемости и т. п. Иногда желательно упорядочить данные файла по какому-либо признаку, например по результатам выполнения какого-либо задания. Нередко возникает необходимость обработки не всех данных файла, а лишь их подмножества, выделяемого по определенным критериям (например, по полу, классу, успеваемости и пр.). Существует и обратная задача: если данные хранятся в не-

скольких небольших файлах, может возникнуть потребность в их объединении для последующего анализа.

Перечисленные проблемы указывают на то, что для регулярной аналитической работы недостаточно умения вводить данные и применять к ним статистические процедуры. Возникает задача эффективного управления. Способы решения этой задачи бывают весьма нетривиальными, и исчерпывающий рассказ о них не вполне соответствовал бы теме книги. Тем не менее представленного в этой главе материала вполне достаточно, чтобы научиться свободно манипулировать данными.

Несмотря на то что навыки управления данными приходят с опытом и требуют некоторого терпения, обязательно освойте их. Это придаст процессу исследования гибкость, простоту и легкость. Тогда выполнение статистических процедур, казавшихся сложными и громоздкими, станет для вас интуитивно понятным.

Мы рассмотрим следующие основные команды управления данными:

- ▶ команда Информация о файле данных позволяет получить сведения о переменных как открытого, так и любого внешнего файла данных SPSS: имена, метки имен и значений;
- ▶ команда Преобразовать ▶ Заменить пропущенные значения, как ясно из ее названия, работает с отсутствующими значениями переменных;
- ▶ команда Преобразовать ▶ Вычислить переменную позволяет путем вычислений создавать новые переменные на основе существующих;
- ▶ команда Преобразовать ▶ Ранжировать наблюдения позволяет создать новую переменную путем ранжирования значений существующей переменной;
- ▶ с помощью команды Данные ▶ Отобрать наблюдения можно выбрать подмножество наблюдений для дальнейшего анализа;
- ▶ команды Преобразование ▶ Перекодировать в другие переменные и Преобразование ▶ Перекодировать в те же переменные предназначены для изменения способа кодирования переменных, например уменьшения числа возможных значений;
- ▶ команда Данные ▶ Сортировать наблюдения позволяет упорядочить объекты в соответствии с назначенными критериями;
- ▶ команды подменю Слить файлы меню Данные используются для добавления в файл новых переменных или наблюдений из другого файла;
- ▶ команда Агрегировать данные меню Данные позволяет создавать такие значения переменных, каждое из которых представляет собой результат объединения группы исходных значений, например их среднее;
- ▶ команды Реструктурировать меню Данные позволяют производить сложные манипуляции со структурой файла данных, например преобразовывать набор переменных в группы значений одной переменной или, наоборот, группы значений одной переменной — в набор переменных.

Все дальнейшие главы книги, начиная с этой, построены таким образом, что они практически не связаны друг с другом. После вводной части каждой главы приве-

дено описание пошаговых процедур, которые позволяют открывать файлы с данными и получать доступ к диалоговым окнам нужных статистических операций. Затем следуют описания этих диалоговых окон, а заключительные части глав посвящены интерпретации выводимых данных и их распечатке. Как правило, действия, связанные с вызовом статистической операции, приводятся на шаге 4, а на шаге 5 рассказывается о том, как задать параметры операции и завершить анализ данных. Иногда имеется несколько альтернативных способов выполнения шагов 4 или 5, в этом случае они помечаются как 4а, 4б, 5а, 5б и т. д.

Структура данной главы несколько отличается от описанной выше, поскольку в ней идет речь не о статистических операциях, а о средствах организации данных. После выполнения трех подготовительных шагов будет детально рассмотрена работа с каждой из команд управления данными. Главу завершит обзор необходимых завершающих действий с программой. В отличие от глав, посвященных анализу данных, здесь вы не найдете заключительного раздела, относящегося к выводу результатов.

Прежде чем двигаться дальше, выполните три подготовительных шага.

ШАГ 1

Подготовьте новый или существующий файл данных. Все необходимые для этого действия описаны в главе 3.

ШАГ 2

Запустите программу SPSS при помощи значка на рабочем столе или выбрав в главном меню Windows команду Пуск ► Программы ► IBM SPSS Statistics ► IBM SPSS Statistics 19 и в открывшемся после запуска программы диалоговом окне выбора режима работы щелкните на кнопке Отмена.

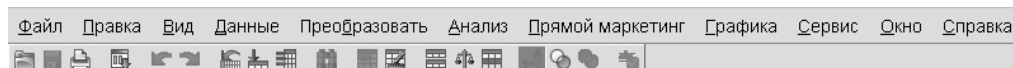
После выполнения этого шага на экране появится окно редактора данных SPSS.

ШАГ 3

Откройте файл данных, с которым вы намерены работать (в нашем случае — это файл ex01.sav). Если он расположен в текущей папке, то выполните следующие действия.

1. Выберите в меню Файл команду Открыть ► Данные или щелкните на кнопке Открыть данные панели инструментов.
2. В открывшемся диалоговом окне дважды щелкните на имени ex01.sav или введите его с клавиатуры и щелкните на кнопке ОК.

Независимо от того, открыта программа SPSS только что или какие-то процедуры уже выполнялись, в верхней части главного окна должна присутствовать строка меню (она показана ниже). Пока строка меню присутствует на экране, доступны все команды анализа данных. При этом окно с данными видеть не обязательно.



При определенных исходных настройках программы SPSS одновременно с открытием файла данных открывается окно вывода. Чтобы вернуться к главному окну и стандартной строке меню, щелкните на кнопке свертывания или восстановления текущего окна.

Получение информации о файле

Команда Информация о файле данных позволяет получить сведения о переменных файла данных SPSS. Это, пожалуй, самая простая из рассматриваемых команд. Она позволяет быстро получать список имен переменных, их меток, меток их значений не только для открытого (рабочего) файла данных, но и для любого другого (внешнего) файла данных SPSS.

При помощи шага 4 вы получите информацию об открытом (рабочем) файле ex01.sav, а выполнение шага 4а обеспечит вас сведениями о файле TestIQ.sav.

ШАГ 4

После выполнения шага 3 должно быть открыто окно редактора данных (ex01.sav — Редактор данных IBM SPSS Statistics).

1. В меню Файл выберите команду Информация о файле данных ► Рабочий файл.

В результате выполнения этой команды откроется окно вывода, содержащее две таблицы. В первой таблице перечислены все переменные файла с указанием их характеристик, а во второй — значения и метки номинативных переменных.

ШАГ 5

Находясь в окне вывода, полученного в результате выполнения шага 4, выполните следующие действия.

1. В меню Файл выберите команду Информация о файле данных ► Внешний файл.
2. В открывшемся окне выбора файла найдите файл с именем TestIQ и дважды щелкните на его имени.

После выполнения этого шага к содержимому окна вывода добавится информация о файле данных TestIQ.sav, содержащая две таблицы. В первой таблице приведены общие сведения о файле, во второй — список переменных и их метки. Таблицы со значениями и метками номинативных переменных нет, так как файл их не содержит.

После просмотра закройте окно вывода без сохранения, чтобы оно не мешало дальнейшей работе.

Обработка пропущенных значений

В процессе работы с программой SPSS вы нередко будете сталкиваться с проблемой отсутствующих данных. Обратимся к переменным из примера ex01.sav. Вполне вероятна ситуация, когда кто-либо из учеников отсутствовал при проведении тестирования или не ответил на вопрос о внешкольном увлечении, либо не определился с перспективой поступления в вуз. Подобные случаи приводят к тому, что в данных рабочего файла появляются *пропущенные значения*. Пропущенные значения не только мешают осмысливать данные, но и могут оказывать нежелательное влияние на результаты анализа. Некоторые статистические процедуры игнорируют объекты (строки), в которых содержится хотя бы одно пропущенное значение.

Если, к примеру, из 35 наблюдений 13 имеют пропущенные значения по разным переменным, то анализу будет подлежать немногим более 60 % данных файла, что, несомненно, исказит результаты.

Большинство статистических методов SPSS позволяет учитывать пропуски в данных двумя принципиально различными способами: *построчно* (listwise) и *попарно* (pairwise). При построчном учете пропусков SPSS перед выполнением операции проверяет строки (объекты) на наличие пропущенных значений и в случае обнаружения последних исключает соответствующие строки из анализа целиком. Этот способ позволяет получить наиболее корректные статистические результаты, однако потери данных при этом максимальны. При попарном учете пропусков обработка выполняется без дополнительных проверок, и в процессе вычислений не выполняются только те операции, которые требуют наличия пропущенного значения. Таким образом, в анализе участвуют все введенные данные, но результаты анализа содержат погрешности.

Мы рекомендуем вам по возможности решать проблему пропущенных значений на этапе ввода и кодирования данных, а не полагаться на то, что SPSS сделает это за вас. В любом случае, чем больше пропусков в исходных данных, тем менее точны и корректны результаты анализа.

Для номинальной переменной проблема пропущенных значений решается легко: вы можете просто ввести для нее еще одну градацию, которая соответствует пропуску в данных. Для количественной переменной (метрической или порядковой), имеющей множество возможных значений, в SPSS предусмотрены специальные процедуры заполнения пропусков: в меню Преобразовать есть команда Заменить пропущенные значения. При всем соблазне ее использовать следует помнить, что результаты обработки данных с заменой пропусков фиктивными значениями, например средними, вряд ли могут вызвать доверие. Поэтому лучше на месте пропуска честно оставлять пустую ячейку. А вопрос о построчном или попарном учете пропусков решать отдельно для каждого конкретного метода анализа данных.

В справочной системе SPSS часто используется два термина: *системные пропущенные значения* (system missing values) и *пользовательские пропущенные значения* (user missing values). Под физически пропущенными значениями понимаются значения, не введенные в компьютер. В редакторе данных пустые ячейки, не содержащие значений, помечены точкой. Логически пропущенные значения — это специальные значения переменной, отражающие невозможность адекватного кодирования некоторой ситуации. Если, например, 1, 2 и 3 — тестовые оценки испытуемого, 8 означает, что тест не завершен, а 9 фиксирует неявку испытуемого, то значения 8 и 9 относятся к логически пропущенным, поскольку их нельзя интерпретировать как результаты теста.

Преобразование данных

Вычисления

Операция Вычислить позволяет путем вычислений создавать новые переменные на основе существующих данных файла. Так, среди данных файла ex01.sav исследова-

телю может понадобиться средний балл результатов тестирования (тест1, ..., тест5) каждого учащегося. Для этого необходимо создать новую переменную, например с именем `тест_ср`.

ШАГ 4А

После выполнения шага 3 должно быть открыто окно редактора данных (`ex01.sav` - Редактор данных IBM SPSS Statistics). При активном окне редактора данных в меню Преобразовать выберите команду Вычислить переменную. На экране появится диалоговое окно Вычислить переменную, показанное на рис. 4.1.

Диалоговое окно Вычислить переменную содержит в левой части уже ставший привычным список доступных переменных. Над списком, в левом верхнем углу окна, имеется поле Вычисляемая переменная, в которое необходимо ввести имя создаваемой, или вычисляемой, переменной. В поле Числовое выражение вводится выражение, с помощью которого вычисляется новая переменная. Вы можете вводить в это поле символы с клавиатуры, через буфер обмена, а также пользоваться перечисленными ниже вспомогательными элементами интерфейса окна, позволяющими формировать выражение.

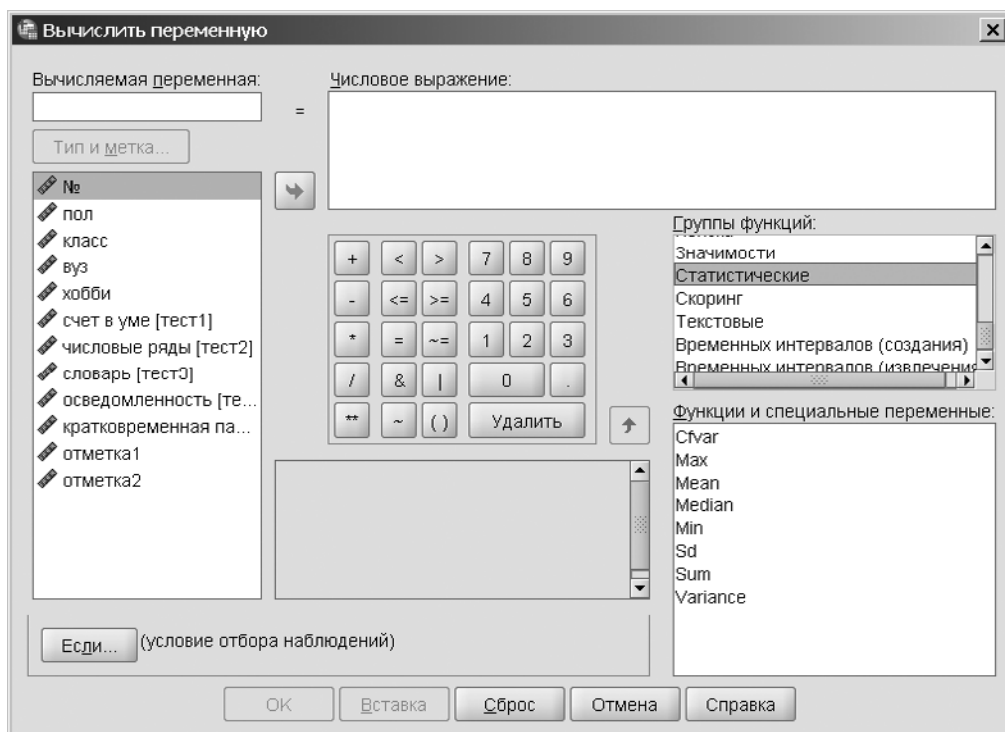


Рис. 4.1. Диалоговое окно Вычислить переменную

На панели калькулятора содержатся кнопки цифр и десятичного разделителя, а также разнообразных операций. Щелчок на каждой из этих кнопок приводит

к появлению соответствующего символа (или нескольких символов) в поле Числовое выражение. Разумеется, при составлении выражения вы можете обойтись без панели калькулятора, поскольку все представленные там символы можно ввести с клавиатуры. Кнопки основных операций перечислены в табл. 4.1.

Таблица 4.1. Кнопки основных операций, доступных при создании переменных

Арифметические операции	Операции сравнения	Логические операции
+ сложение	< меньше	& И
- вычитание	> больше	ИЛИ
* умножение	<= меньше или равно	~ НЕ
/ деление	>= больше или равно	
** возведение в степень	= равно	
() очередность операций	~= не равно	

- В списке Группы функций перечислены названия 20 групп функций, причем назначение большинства из них весьма туманно. Далее мы приводим описания нескольких групп функций, которые кажутся нам наиболее употребительными.
- В списке Функции и специальные переменные перечислены операции, входящие в состав данной группы функций. Например, на рис. 4.1 выделена группа функций Статистические и в окне ниже перечислены доступные статистические функции. Для того чтобы воспользоваться одной из них, необходимо ее выделить и при помощи стрелки, направленной вверх, перенести ее в окно Числовое выражение.
- При щелчке на кнопке Если открывается диалоговое окно выбора наблюдений. Это окно используется в том случае, когда необходимо произвести вычисления только для части наблюдений (строк). Функция кнопки Если аналогична команде Отобразить наблюдения, описанной далее в этой главе.

Ниже перечислены основные группы функций, доступные при создании переменных.

Арифметические

Арифметические и алгебраические операции. Например, ABS(числвыр) — возвращает абсолютное значение числового выражения в скобках (числвыр).

Пропущенных значений

Операции с пропущенными значениями. Например, NVALID(переменная[...]) возвращает количество действительных значений (не пропущенных) из списка переменных, имена которых заданы в квадратных скобках.

Статистические

Статистические операции. Например, `MEAN(числвыр,числвыр[...])` возвращает среднее значение числовых выражений (`числвыр,числвыр [...]`), включающих имена переменных, список которых указан в скобках.

Случайных чисел

Операции, генерирующие в качестве значений переменной случайные числа с заданными свойствами распределения. Например, `Rv.Normal(срдн,стдоткл)` возвращает массив нормально распределенных случайных чисел с заданным средним значением `срдн` и стандартным отклонением, равным `стдоткл`. Если `Случнорм` — целевая переменная, выражение `NORMAL(0,3)` создаст переменную `Случнорм`, значениями которой будут случайные числа, распределенные нормально с нулевым средним значением и стандартным отклонением, равным 3.

В следующем примере создадим новую переменную `тест_ср`. Она получается путем вычисления среднего для значений переменных `тест1`, ..., `тест5`.

ШАГ 5А

Для выполнения следующих действий диалоговое окно Вычислить переменную должно быть открыто (см. рис. 4.2) в результате выполнения шага 4.

1. В поле Вычисляемая переменная введите имя `тест_ср`.
2. В окне Группы функций найдите и выделите щелчком мыши Статистические, а в окне Функции и специальные переменные выделите `Mean` (среднее).
3. При помощи стрелки, направленной вверх, или двойным щелчком мыши на пункте `Mean` введите функцию `Mean (?.?)` в поле Числовое выражение.
4. В выражении `MEAN(?.?)` в поле Числовое выражение сотрите символы вопросов и запятую в скобках, оставив курсор между скобками.
5. Щелкните сначала на имени `тест1`, чтобы выделить его, а затем — на кнопке с направленной вправо стрелкой (можно также дважды щелкнуть на имени `тест1`), чтобы ввести его в скобки в поле Числовое выражение.
6. Щелкните на кнопке `,` (или введите запятую с клавиатуры) и дважды щелкните на переменной `тест2`.
7. Повторите те же действия для переменных `тест3`, `тест4` и `тест5`. В итоге должно получиться выражение: `MEAN(тест1,тест2,тест3,тест4,тест5)`. Если это не так, внесите изменения. Все выражение в поле Числовое выражение можно целиком вводить с клавиатуры, в том числе — и имена переменных.
8. Щелкните на кнопке `ОК`.
9. В меню `Файл` выберите команду `Сохранить`.

В результате выполнения процедуры будет создана новая переменная `тест_ср`, которая разместится в крайнем правом столбце файла данных. Если вам удобнее

держат эту переменную ближе к началу файла, можете воспользоваться операциями вырезания и вставки в нужную позицию.

После завершения процедуры программа возвращает вас в окно редактора данных, чтобы вы могли просмотреть содержимое новой переменной.

В более сложных вычислениях с привлечением разнообразных арифметических и других операций вам придется следить за приоритетами операций и при необходимости регулировать очередность вычислений с помощью скобок.

Ранжирование

Рассмотренная команда Вычислить позволяет создать новую переменную, значения которой вычисляются при помощи выражения, содержащего другие переменные. Команда Преобразовать ► Ранжировать наблюдения тоже позволяет создать новую переменную, значения которой — ранговые места наблюдений по заданной переменной. Эта процедура применяется, когда необходимо перейти от исходных значений переменной к рангам.

Команда ранжирования (перехода к рангам) выполняется при помощи диалогового окна, показанного на рис. 4.2. Для перехода к рангам достаточно задать имя переменной в поле Переменные и щелкнуть на кнопке ОК. Появится новая переменная с заданным по умолчанию именем, содержащая ранги значений исходной переменной.

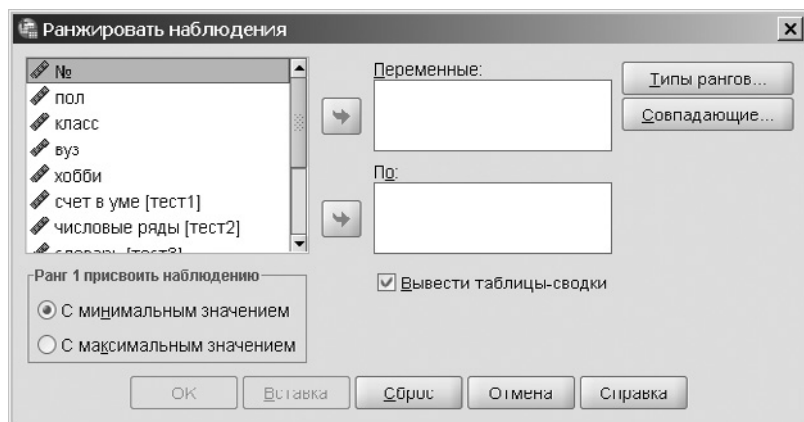


Рис. 4.2. Диалоговое окно Ранжировать наблюдения

Предположим, исследователь решил ранжировать всех учащихся по успеваемости за 11 класс. Для этого ему необходимо перейти от исходных значений переменной отметка2 к новой переменной, содержащей ранги учащихся по этой переменной.

Рассмотрим процедуру ранжирования на примере.

ШАГ 4Б

После выполнения шага 3 должно быть открыто окно редактора данных (ex01.sav - Редактор данных IBM SPSS Statistics).

Для ранжирования переменной отметка2 выполните следующие действия.

1. В меню Преобразовать выберите команду Ранжировать наблюдения. На экране появится диалоговое окно, представленное на рис. 4.2.

2. Щелкните сначала на переменной отметка2, чтобы выделить ее, а затем — на кнопке со стрелкой. Флажок Вывести таблицы-сводки можно сбросить, так как никаких результатов обработки, кроме новой переменной, получать не планируется.
3. По умолчанию в группе Ранг 1 присвоить наблюдению установлен переключатель С минимальным значением. При желании вы можете изменить порядок ранжирования на обратный, установив в группе Ранг 1 присвоить наблюдению переключатель С максимальным значением.
4. При щелчке на кнопке Типы рангов вам будет предложен дополнительный набор экстравагантных способов ранжирования, а при щелчке на кнопке Совпадающие — способов обработки одинаковых (совпадающих) рангов. По умолчанию принято обычное присвоение одинаковым значениям переменной одного и того же среднего ранга (Средний). Для ранжирования по общепринятым правилам оставьте все те параметры, которые установлены по умолчанию, и щелкните на кнопке ОК.

В результате выполнения этого шага в конце списка появится новая переменная с именем Ротметка, присвоенным по умолчанию. Это имя можно оставить или поменять на другое.

Выбор наблюдений для анализа

В этом разделе мы отойдем от привычных вводных слов и сразу перейдем к пошаговым процедурам.

ШАГ 4В

После выполнения шага 3 должно быть открыто окно редактора данных (ex01.sav - Редактор данных IBM SPSS Statistics).

В меню Данные выберите команду Отобразить наблюдения. На экране появится диалоговое окно, показанное на рис. 4.3.

Команда Отобразить наблюдения позволяет пользователю выбирать для обработки не все, а часть данных, удовлетворяющих заданным условиям. Поскольку необходимость в этом возникает довольно часто, команда Отобразить наблюдения является одной из самых востребованных при проведении исследований. Так, исследователю, использующему файл ex01.sav, могут понадобиться статистические сведения, касающиеся учащихся конкретного класса или определенного пола и т. д. Это означает, что ему нужно указать программе, какие данные следует выделить для обработки. Именно для этого предназначена команда Отобразить наблюдения. После проведения анализа выбранной подгруппы вы можете вернуться к полному набору данных, установив в окне Отобразить наблюдения переключатель Все наблюдения. Если нужно создать файл данных, содержащий только выборку, можно удалить все исключенные объекты, а затем сохранить полученный файл с помощью команды Сохранить меню Файл. Номера выбранных наблюдений, в отличие от исключенных, после выполнения команды Отобразить наблюдения остаются незачеркнутыми,

что позволяет легко находить их визуально. На рис. 4.4 показан фрагмент окна редактора данных (вкладка Данные), на котором видны выбранные (девушки) и исключенные (юноши) наблюдения.

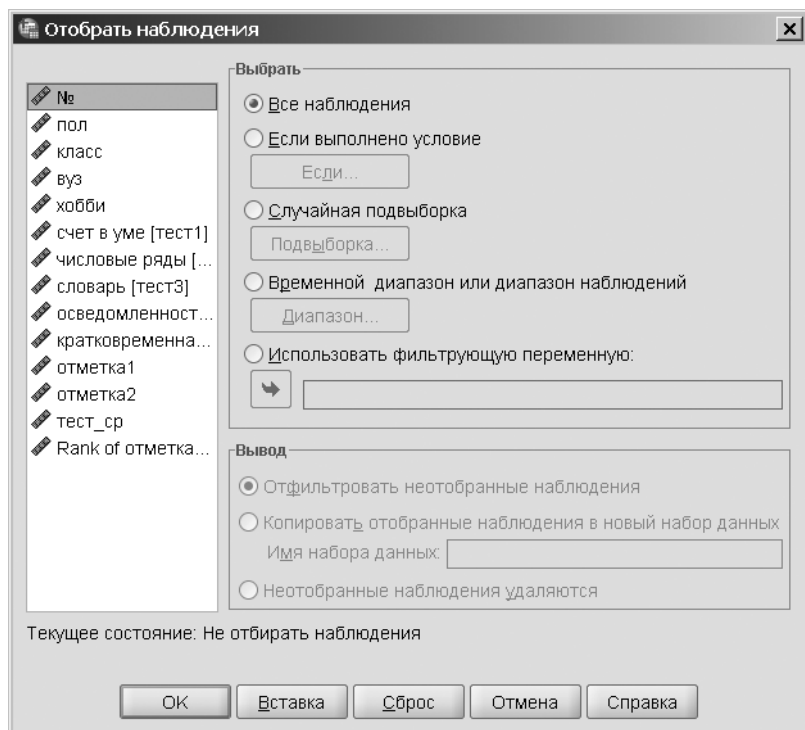


Рис. 4.3. Диалоговое окно Отобрать наблюдения

	№	пол	класс	вуз	хобби	тест1	тест2	те
1	1	МУЖ	Б	ЕСТ_Н	искусс...	6	7	
2	2	МУЖ	А	ЕСТ_Н	спорт	8	9	
3	3	МУЖ	В	ТЕХН	компь...	10	6	
4	4	ЖЕН	В	ГУМ	компь...	13	9	
5	5	МУЖ	Б	ТЕХН	искусс...	12	8	
6	6	ЖЕН	В	ЭКОН	искусс...	12	15	

Рис. 4.4. Данные после отбора наблюдений для анализа

При выборе группы наблюдений SPSS создает специальную переменную с именем `filter_$`, значение которой равно 1 для выбранных наблюдений и 0 для исключенных. Вы можете использовать эту переменную (например, для проведения сравнений между группами) или просто удалить ее.

В диалоговом окне Отобрать наблюдения имеются две группы переключателей. В группу Выбрать входит 5 переключателей, из которых в данный момент нас интересует только два.

- ▶ Переключатель Все наблюдения фактически предназначен для отмены операции, поскольку при его установке выбранными полагаются все объекты файла.
- ▶ Переключатель Если выполнено условие позволяет задать условие отбора наблюдений с помощью кнопки Если. При щелчке на этой кнопке появляется диалоговое окно, очень похожее на окно, показанное на рис. 4.2. Чтобы задать условие отбора, вы можете использовать клавиатуру, панель калькулятора, список функций, а также буфер обмена. Описание функций основных кнопок приведено в табл. 4.1. Ниже приведены примеры некоторых условий отбора.

пол = 1

Отбираются только девушки.

класс = 3

Отбираются ученики класса «В».

класс <= 3

Отбираются ученики классов «А» и «Б».

вуз >= 2 & вуз <= 3

вуз > 1 & вуз < 4

При задании любого из этих условий отбираются ученики, выбравшие экономические и технические вузы.

Чтобы ввести переменную в поле задания условия отбора, достаточно щелкнуть сначала на ней, а затем — на кнопке со стрелкой. Знаки операций можно вводить как с клавиатуры, так и с панели калькулятора. В первом из примеров мы выберем для исследований всех учащихся-девушек, а во втором — учащихся, выбирающих экономические и технические вузы.

ШАГ 5В

Сначала займемся девушками. При открытом окне Отобрать наблюдения, показанном на рис. 4.3, выполните следующие действия.

1. В группе Выбрать установите переключатель Если выполнено условие и щелкните на кнопке Если.
2. В открывшемся окне сначала щелкните на имени переменной пол, чтобы выделить ее, а затем на кнопке со стрелкой, чтобы ввести переменную пол в условие отбора. Далее на панели калькулятора щелкните на кнопке =, а затем — на кнопке 1, чтобы закончить составление условия отбора.
3. После создания условия отбора щелкните на кнопке Продолжить, чтобы закрыть первое диалоговое окно, и на кнопке ОК, чтобы закрыть второе диалоговое окно и вернуться в окно редактора данных.

После выполнения этого шага при любой обработке будут учитываться только данные для девушек. Чтобы сделать доступными все данные, достаточно в окне Отобрать наблюдения установить переключатель Все наблюдения.

ШАГ 5Г

Теперь выберем для исследований школьников, ориентированных на поступление в экономические и технические вузы. После выполнения шага 4в окно Отобрать наблюдения должно выглядеть так, как показано на рис. 4.3.

1. В группе Выбрать установите переключатель Если выполнено условие и щелкните на кнопке Если.
2. В открывшемся окне сначала щелкните на имени переменной вуз, чтобы выделить ее, а затем — на кнопке со стрелкой, чтобы ввести переменную вуз в условие отбора. Далее на панели калькулятора щелкните на кнопке \geq , введите число 2 и щелкните на кнопке &. Снова щелкните на имени переменной вуз, чтобы выделить ее, затем — на кнопке со стрелкой, чтобы вторично ввести переменную вуз в условие отбора. На панели калькулятора щелкните на кнопке \leq и введите число 3, закончив составление условия отбора.
3. После создания условия отбора щелкните на кнопке Продолжить, чтобы закрыть первое диалоговое окно, и на кнопке ОК, чтобы закрыть второе диалоговое окно и вернуться в окно редактора данных.

Перекодировка в новую переменную

Команда Преобразование ► Перекодировать в другие переменные создает новую переменную, однако ее значения определяются не как результат вычислений или ранжирования, а на основе замены множества значений существующей переменной небольшим числом категорий. Как правило, эта процедура применяется, когда необходимо разделить выборку наблюдений на подгруппы по диапазонам значений некоторой количественной переменной. При этом возможен учет значений одной или нескольких переменных.

Предположим, что на примере данных из файла ex01.sav, требуется, исходя из значений переменной отметка2 (успеваемость в 11 классе), создать две новые переменные: Гр_усп1 и Гр_усп2. Переменная Гр_усп1 должна делить всех учащихся на две группы: 1 — отметка ниже 4; 2 — отметка 4 и выше. При помощи новой переменной Гр_усп2 требуется разделить учащихся уже на 4 группы — в зависимости от успеваемости с учетом пола: выделить среди юношей и девушек тех, у кого средний балл отметки ниже 4 или 4 и выше. Таким образом, новая переменная Гр_усп2 может принимать 4 значения: 1 — девушки, отметка < 4 ; 2 — девушки, отметка 4 и выше; 3 — юноши, отметка < 4 ; 4 — юноши, отметка 4 и выше.

Команда перекодирования в другие переменные выполняется с помощью двух диалоговых окон. Одно из них показано на рис. 4.5. В нем вы можете задать входную и выходную переменные.

Как видно из рис. 4.6, окно Перекодировать в другие переменные имеет вполне обычный вид: в его левой части находится список доступных переменных, а внизу — пять стандартных кнопок (см. главу 2).

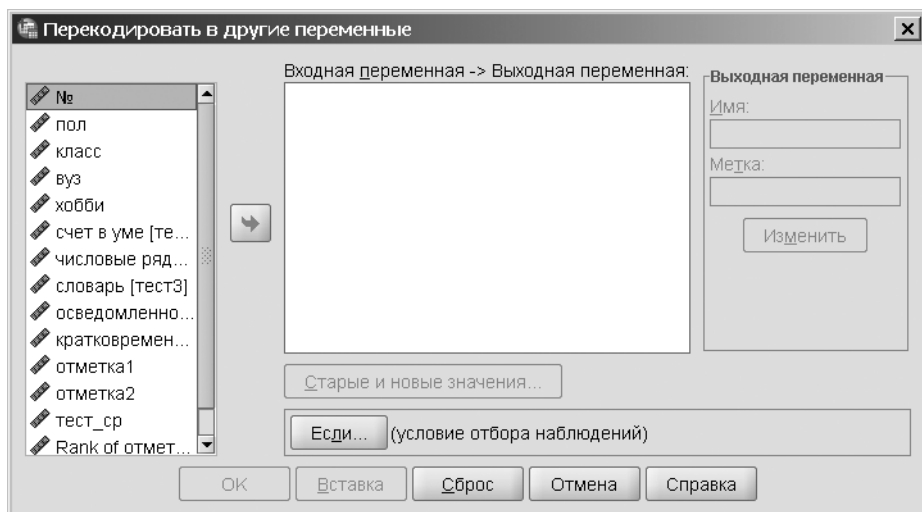


Рис. 4.5. Диалоговое окно Перекодировать в другие переменные

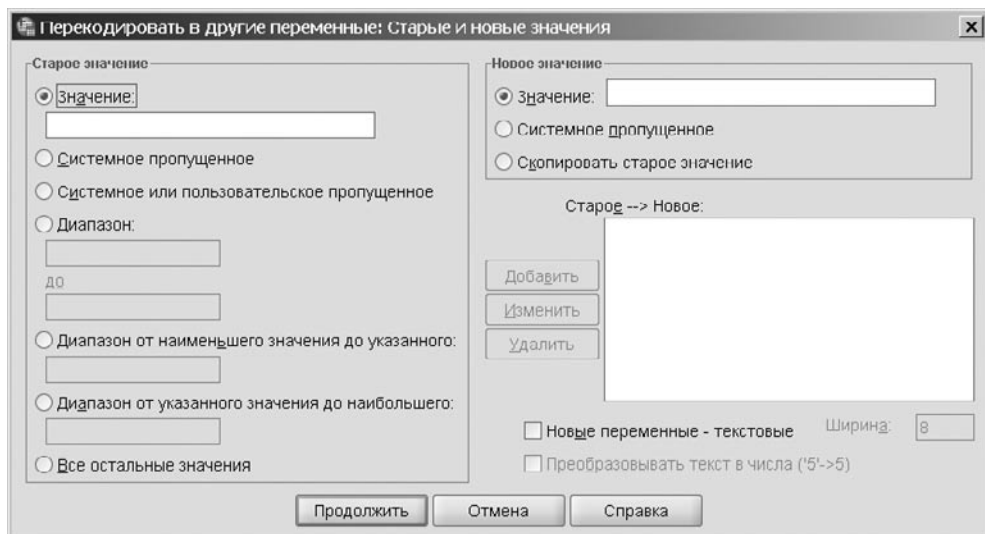


Рис. 4.6. Диалоговое окно Перекодировать в другие переменные: Старые и новые значения

Сначала создадим переменную `Гр_усп1`, разделяющую учащихся по успеваемости на две группы. Первое, что необходимо сделать в рассматриваемом нами примере, — выделить переменную `отметка2` и переместить ее в список `Входная переменная` ▶ `Выходная переменная`. Затем в поле `Имя` области `Выходная переменная`

ная следует ввести имя новой переменной (в данном случае `Гр_усп1`). Щелчок на кнопке Изменить приведет к появлению переменной `Гр_усп1` в предыдущем списке: его содержимое будет иметь вид `отметка2 > Гр_усп1`.

Щелчок на кнопке Старые и новые значения вызовет диалоговое окно, показанное на рис. 4.6. В нем можно задать значения новой переменной, которые будут соответствовать отдельным значениям или диапазонам значений старой переменной.

Семь переключателей в левой части окна (Старое значение) позволяют задавать различные способы задания заменяемых значений исходной переменной.

Теперь требуется задать соответствие между диапазонами значений переменной `отметка2` и значениями новой переменной `Гр_усп1`: 1 — отметка 4 и выше, 2 — отметка до 4. Для того чтобы настроить окно на работу с диапазоном от заданного до наибольшего значения, установите переключатель Диапазон от указанного значения до наибольшего в группе Старое значение. Введите в поле, расположенное под этим переключателем, число 4, а в поле Значение в группе Новое значение — цифру 1. После щелчка на кнопке Добавить вы увидите созданное соответствие в поле Старое > Новое: 4 through HIGHEST > 1. Всем остальным значениям следует присвоить цифру 2. Для этого достаточно в группе Старое значение установить переключатель Все остальные значения и в правом верхнем окне ввести новое значение 2. После щелчка на кнопке Добавить вы увидите последнее соответствие. Теперь щелкните сначала на кнопке Продолжить, затем, в окне Перекодировать в другие переменные — на кнопке ОК.

Новыми переменными вы можете управлять так же, как и переменными, созданными «обычным» способом: назначать метки, вручную изменять значения, менять расположение в файле данных и т. п.

Обобщим процесс создания переменной `Гр_усп1` в форме уже привычных нам пошаговых процедур.

ШАГ 4Д

После выполнения шага 3 должно быть открыто окно редактора данных (`ex01.sav` - Редактор данных IBM SPSS Statistics).

Для создания переменной `Гр_усп1` выполните следующие действия.

1. В меню Преобразовать выберите команду Перекодировать > Перекодировать в другие переменные. На экране появится диалоговое окно, показанное на рис. 4.5.
2. Щелкните сначала на переменной `отметка2`, чтобы выделить ее, а затем — на кнопке со стрелкой.
3. Для перехода в поле Имя нажмите клавишу Tab, введите имя `Гр_усп1` и щелкните на кнопке Изменить.
4. Щелкните на кнопке Старые и новые значения, чтобы открыть диалоговое окно, показанное на рис. 4.6.
5. В левой части окна установите второй снизу переключатель Диапазон от указанного значения до наибольшего. В поле под ним введите чис-

ло 4, а в поле Значение в правом верхнем углу — цифру 1 и щелкните на кнопке Добавить.

6. В группе Старое значение установите переключатель Все остальные значения и в правом верхнем окне введите цифру 2.
7. Щелкните на кнопке Продолжить, чтобы вернуться в предыдущее диалоговое окно Перекодировать в другие переменные, в котором щелкните на кнопке ОК.

В результате выполнения этого шага в окне редактора данных появится новая переменная с именем `Гр_усп1`.

Теперь создадим новую переменную `Гр_усп2`, которая позволит разделить учащихся на 4 группы по успеваемости в зависимости от пола. В данном случае новая переменная, как и ранее `Гр_усп1`, создается исходя из значений переменной `отметка2`. Но при этом необходимо учесть значения переменной `пол`. Для этого в диалоговом окне Перекодировать в другие переменные (рис. 4.5) необходимо воспользоваться кнопкой Если.... Функция этой кнопки аналогична команде Отбор наблюдений при установке переключателя Если выполнено условие. При помощи этой кнопки сначала выбираются девушки (`пол=1`) и выполняются действия, аналогичные созданию переменной `Гр_усп1`. Затем, при помощи кнопки Если... выбираются юноши (`пол=2`) и снова повторяются те же действия, но соответствующим группам присваиваются значения 3 и 4.

ШАГ 4Е

После выполнения шага 3 должно быть открыто окно редактора данных (`ex01.sav` - Редактор данных IBM SPSS Statistics).

Для создания переменной `Гр_усп2` выполните следующие действия.

1. В меню Преобразовать выберите команду Перекодировать в другие переменные. На экране появится диалоговое окно, представленное на рис. 4.5. Если вы уже успели поработать с этой процедурой, нажмите кнопку Сброс.
2. Щелкните сначала на переменной `отметка2`, чтобы выделить ее, а затем — на кнопке со стрелкой.
3. Для перехода в поле Имя нажмите клавишу Tab, введите имя `Гр_усп2` и щелкните на кнопке Изменить.
4. Щелкните на кнопке Если..., чтобы открыть диалоговое окно Перекодировать в другие переменные: Отбор наблюдений.
5. В левой части окна выделите переменную `пол`, при помощи стрелки перенесите ее в правое окно и присвойте ей 1 (`пол=1`). Нажмите кнопку Продолжить.
6. Щелкните на кнопке Старые и новые значения, чтобы открыть диалоговое окно, показанное на рис. 4.6.
7. В левой части окна установите второй снизу переключатель Диапазон от указанного значения до наибольшего. В поле под ним введите число 4, а в поле Значение в правом верхнем углу — цифру 1 и щелкните на кнопке Добавить.
8. В группе Старое значение установите переключатель Все остальные значения и в правом верхнем окне введите цифру 2. Щелкните на

кнопках Добавить и Продолжить, чтобы вернуться в предыдущее диалоговое окно Перекодировать в другие переменные.

9. Щелкните на кнопке Если..., чтобы открыть диалоговое окно Перекодировать в другие переменные: Отбор наблюдений. В правом окне меняйте значение переменной пол=2. Нажмите кнопку Продолжить и щелкните на кнопке Старые и новые значения.
10. В правой части во втором сверху окне, под надписью Старая->Новая выделите курсором строку 4 through HIGHEST (от 4 до наибольшего). В поле Значение в правом верхнем углу замените цифру 1 на цифру 3 и щелкните на кнопке Изменить.
11. Выделите строку ELSE->2. В поле Значение в правом верхнем углу замените цифру 2 на цифру 4 и щелкните на кнопке Изменить.
12. Щелкните на кнопке Продолжить, чтобы вернуться в предыдущее диалоговое окно Перекодировать в другие переменные, в котором щелкните на кнопке ОК.

В результате выполнения шага 4е в окне редактора данных появится новая переменная с именем Гр_усп2, созданная исходя из значений двух переменных: отметка2 и пол.

Сходным образом можно создавать новые переменные с учетом значений более двух исходных переменных. Кроме того, можно создавать сразу несколько новых переменных, если алгоритмы их получения из старых переменных совпадают.

Перекодирование существующей переменной

Иногда может возникнуть необходимость изменить кодирование какой-либо переменной. Это может быть обусловлено двумя причинами.

- ▶ Вы работаете одновременно с несколькими файлами данных, которые создавались разными людьми и содержат одни и те же переменные, но закодированные по-разному.
- ▶ В процессе исследования вам стало ясно, что текущее кодирование какой-либо из переменных следует изменить.

Обратимся к файлу ex01.sav. Как вы помните, в переменной пол значение 1 соответствует девушкам, а значение 2 — юношам. Вы можете столкнуться с тем, что у вас появится другой файл, в котором имеется переменная пол, но автор файла значением 1 закодировал юношей, а значением 2 — девушек. Чтобы избежать путаницы, вам придется перекодировать значения переменной пол в одном из двух файлов. Может оказаться и так, что значения, представляющие внешкольные увлечения учащихся, удобно расположить в порядке убывания распространенности

этих увлечений среди школьников, и если вы не подумали об этом сразу, то вам также придется заняться перекодированием.

Нередко при первоначальном кодировании вводится избыточное число уровней переменной. Как правило, эта избыточность проявляется на этапе проведения исследования. Типичной является следующая ситуация. Пусть имеется переменная вуз, отражающая выбираемый учащимся профиль образования. Первоначально задается 4 уровня этой переменной: 1 — гуманитарный, 2 — экономический, 3 — технический, 4 — естественно-научный. При помощи процедуры, которая будет описана далее, можно объединить значения 1 и 2 в группу гуманитарных вузов, а значения 4 и 3 — в группу физико-технических вузов и присвоить этим группам значения 1 и 2 соответственно.

Перекодирование существующей переменной выполняется с помощью команды Перекодировать в те же переменные меню Преобразовать. После выбора этой команды на экране появится диалоговое окно, показанное на рис. 4.7.

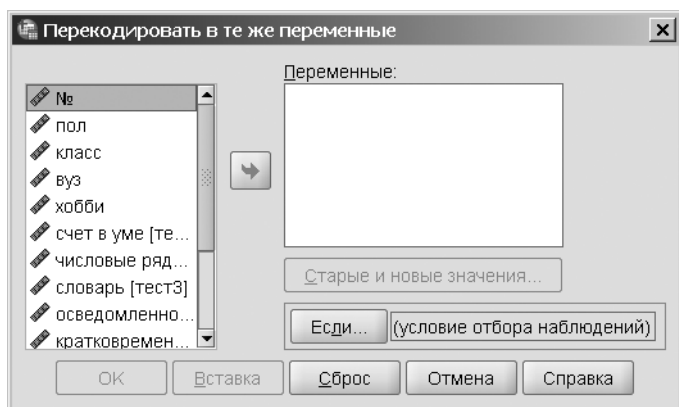


Рис. 4.7. Диалоговое окно Перекодировать в те же переменные

Операция перекодирования в те же переменные, как и предыдущая, управляется двумя диалоговыми окнами (второе показано на рис. 4.8), причем можно работать одновременно с несколькими переменными, однако для этого вам придется постоянно переключаться между диалоговыми окнами.

Первым действием, которое необходимо выполнить в окне Перекодировать в те же переменные, является заполнение списка перекодируемых переменных Переменные. Имена переменных содержатся в списке слева и выбираются при помощи кнопки со стрелкой. При щелчке на кнопке Старые и новые значения открывается диалоговое окно, показанное на рис. 4.8.

Как и в предыдущей процедуре, вы задаете одно или несколько старых и новых значений и помещаете их в списки Старое значение и Новое значение. В приведенном ниже примере мы объединим 4 градации переменной вуз в две градации.

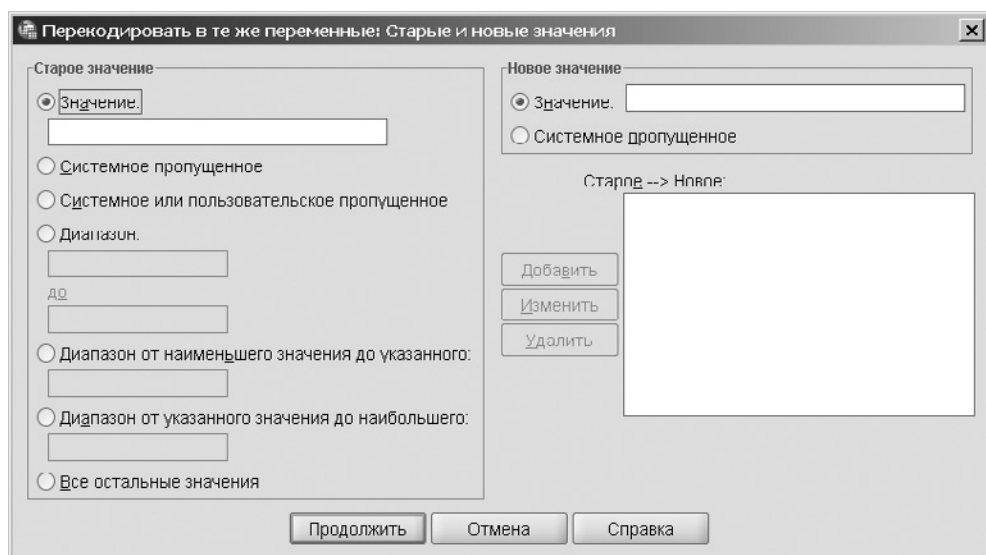


Рис. 4.8. Диалоговое окно Перекодировать в те же переменные: Старые и новые значения

ШАГ 4Ж

После выполнения шага 3 должно быть открыто окно редактора данных (ex01.sav - Редактор данных IBM SPSS Statistics).

Для перекодирования переменной вуз выполните описанные ниже действия.

1. В меню Преобразовать выберите команду Перекодировать в те же переменные. На экране появится диалоговое окно, показанное на рис. 4.7.
2. Щелкните сначала на переменной вуз, чтобы выделить ее, затем — на кнопке со стрелкой, чтобы переместить ее в список Переменные, и, наконец, — на кнопке Старые и новые значения, чтобы открыть диалоговое окно, показанное на рис. 4.8.
3. В группе Старое значение установите переключатель Диапазон, в верхнем поле введите значение 1, нажмите клавишу Tab, в нижнем поле введите значение 2, дважды нажмите клавишу Tab, чтобы перейти в поле Значение области Новое значение, введите там число 1 и щелкните на кнопке Добавить.
4. Повторите предыдущее действие для старых значений 3–4 и нового значения 2, затем щелкните на кнопке Продолжить, чтобы вернуться в исходное диалоговое окно, в котором щелкните на кнопке ОК.

Обратите внимание на то, что процедура перекодирования никак не влияет на метки значений. Другими словами, после выполнения перекодирования необходимо заново задать метки всем значениям, изменившим свой смысл. Действия с метками значений были рассмотрены в главе 3, и при необходимости вы можете обратиться к соответствующим разделам. Если вы хотите сохранить изменения, сделанные вами в файле данных, в меню Файл выберите команду Сохранить.

Сортировка наблюдений

Команда Сортировать наблюдения предназначена для реорганизации данных файла. Эта операция очень распространена, поскольку позволяет расположить информацию в том порядке, в котором это удобно исследователю в текущий момент. Так, к примеру, исследователь мог бы с помощью этой команды изменить исходный порядок следования данных, чтобы учащиеся были перечислены по классам или по убыванию успеваемости (значению переменной отметка2).

ШАГ 43

После выполнения шага 3 должно быть открыто окно редактора данных (ex01.sav - Редактор данных IBM SPSS Statistics).

В меню Данные выберите команду Сортировать наблюдения. На экране появится диалоговое окно, показанное на рис. 4.9.

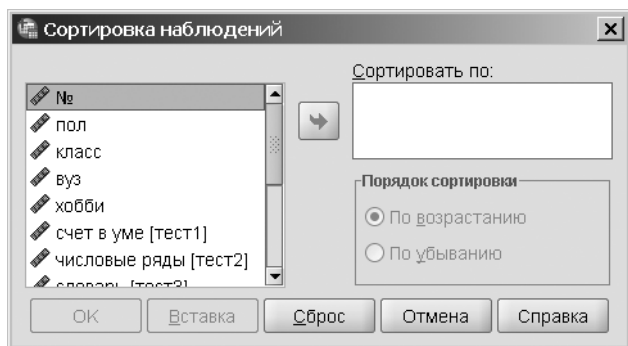


Рис. 4.9. Диалоговое окно Сортировка наблюдений

Параметры команды Сортировать наблюдения, в отличие от параметров предыдущих команд, настраиваются в одном небольшом диалоговом окне. Вы выбираете переменные в исходном списке и с помощью переключателей По возрастанию или По убыванию указываете порядок сортировки. Для текстовых переменных сортировка по возрастанию означает сортировку в алфавитном порядке.

При определении параметров сортировки можно задавать сразу несколько переменных. Это означает, что сначала данные будут отсортированы по значению первой выбранной переменной, затем наблюдения, имеющие одинаковые значения первой переменной, сортируются по значению второй переменной и т. д. Например, если в список Сортировать по ввести переменные класс и пол, то все учащиеся будут перечислены по классам, а внутри каждого класса сначала будут следовать девушки, затем — юноши.

В первом из двух приведенных ниже примеров данные файла ex01.sav упорядочиваются по убыванию значения переменной отметка2; во втором примере реализована упоминавшаяся выше сортировка учащихся по классам и полу.

ШАГ 53

При открытом диалоговом окне Сортировка наблюдений, показанном на рис. 4.9, щелкните сначала на переменной отметка2, чтобы выделить ее, а затем — на кнопке со стрелкой, чтобы переместить ее в список Сортировать по. Далее в группе Порядок сортировки установите переключатель По убыванию и щелкните на кнопке ОК.

ШАГ 51

При открытом диалоговом окне Сортировка наблюдений, показанном на рис. 4.9, выполните следующие действия.

1. Щелкните сначала на переменной класс, чтобы выделить ее, а затем — на кнопке со стрелкой, чтобы переместить ее в список Сортировать по.
2. Повторите то же действие для переменной пол и щелкните на кнопке ОК.

Обратите внимание на то, что по умолчанию выполняется сортировка по возрастанию, поэтому для такой сортировки никаких дополнительных действий по заданию способа сортировки не требуется.

Исходный порядок следования учащихся можно восстановить, назначив сортировку по переменной № (по возрастанию).

Объединение данных разных файлов

Работа с данными нескольких файлов одновременно иногда способна сбить с толку даже опытных пользователей. Нередко файлы данных создаются при помощи разного программного обеспечения и имеют разные форматы. Это порождает различные проблемы их совместного использования, из которых мы рассмотрим наиболее типичную: дополнение *рабочего* файла SPSS содержимым *внешнего* файла. При этом возможны две ситуации дополнения данных рабочего файла SPSS:

- ▶ из внешнего файла Excel;
- ▶ из внешнего файла SPSS.

Мы рассмотрим оба варианта, но сначала сформулируем общие рекомендации.

- ▶ Если вы намерены добавлять переменные, убедитесь, что порядок следования наблюдений в рабочем и внешнем файлах одинаков.
- ▶ Если вы намерены добавлять объекты, убедитесь, что порядок следования переменных в рабочем и внешнем файлах одинаков.
- ▶ Настройте форматы каждой переменной рабочего файла данных, чтобы они соответствовали данным внешнего файла (если добавляемая переменная содержит буквенные символы, либо замените буквы числами во внешнем файле, либо поменяйте тип переменной на строковую в рабочем файле).
- ▶ Перед добавлением данных создайте резервную копию рабочего файла, чтобы к ней можно было вернуться в случае неудачного переноса данных.

Следует иметь в виду, что чем больше соответствия между структурами рабочего и внешнего файлов, тем меньше вероятность ошибок при слиянии данных.

Если вы выполнили указанные рекомендации, то перенос данных из одного (внешнего) в другой (рабочий) файл не составит труда.

1. Откройте рабочий и внешний файлы с данными.
2. Во внешнем файле выделите необходимый для переноса блок данных.
3. Скопируйте его в буфер обмена.
4. Разверните свернутое окно SPSS с рабочим файлом на вкладке Данные.
5. Установите фокус ввода в левую верхнюю ячейку того места, куда вы намереваетесь поместить переносимые данные.
6. В меню Правка выберите команду Вставка.

После этого таблица редактора данных дополнится новыми данными. Проверьте корректность переноса данных. Если допущена ошибка, в меню Правка рабочего файла выберите команду Отменить. После возврата в исходное состояние проверьте свои действия и повторите попытку.

Указанная последовательность действий требует большого внимания и определенной сноровки и применима для небольших фрагментов данных. К счастью, существует более надежный способ объединения данных — *слияние* рабочего и внешнего файлов.

Слияние файлов допустимо, когда и рабочий, и внешний файлы созданы при помощи редактора данных SPSS и имеют одинаковые имена переменных (когда мы добавляем объекты) или одинаковые число и порядок следования наблюдений (когда мы добавляем переменные). Таким образом, если внешним является файл Excel, то перед слиянием необходимо на его основе создать внешний файл SPSS (см. главу 3). Структура создаваемого файла должна быть максимально согласована со структурой рабочего файла. Для этого, дополнительно к общим рекомендациям, изложенным ранее, нужно настроить форматы каждой переменной файлов так, чтобы они были одинаковыми, и проверить идентичность имен переменных файлов.

Хотя эти рекомендации не являются обязательными, все же следует иметь в виду, что чем больше соответствия между структурами файлов, тем меньше вероятность возникновения ошибок при слиянии. Особенное внимание следует уделять идентичности имен переменных, их форматам и расположению в файлах данных. Обратите внимание также на то, что при добавлении переменных одинаковый порядок следования наблюдений в файлах обязателен.

Добавление наблюдений

После завершения шага 3, описанного ранее в этой главе, у вас уже должен быть открыт файл ex01.sav.

ШАГ 4К

Для добавления наблюдений в меню Данные выберите команду Слить файлы ► Добавить наблюдения. На экране появится диалоговое окно, показанное на рис. 4.10.

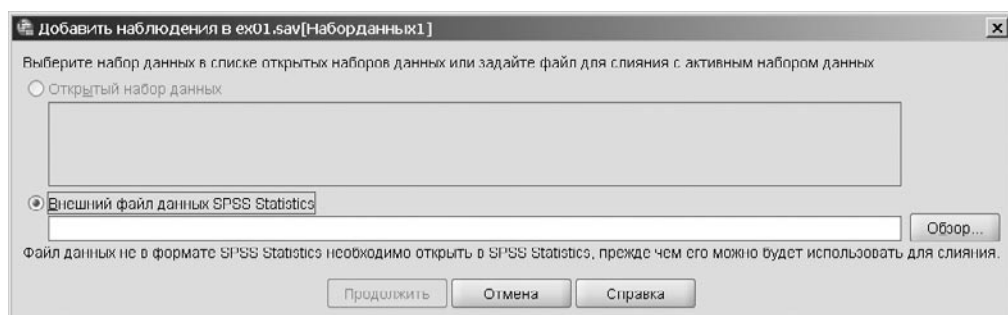


Рис. 4.10. Диалоговое окно Добавить наблюдения

С помощью диалогового окна Добавить наблюдения можно выбрать внешний файл данных, предназначенный для слияния с открытым (рабочим) файлом. Если нужный внешний файл есть в списке Открытый набор данных, то для доступа к нему достаточно щелкнуть на его имени.

Предположим, создан файл ex01a.sav, в котором содержатся данные о 10 новых учащихся с теми же переменными, что и файл ex01.sav. В данном случае ex01.sav является рабочим файлом, а ex01a.sav — внешним. Иногда внешний файл располагается на внешнем носителе. Как правило, в этом случае проще ввести полный путь к файлу с клавиатуры в формате *f:\имя_файла.sav*. После задания имени файла щелкните на кнопке Открыть и затем Продолжить.

После того как внешний файл задан, на экране появляется диалоговое окно, показанное на рис. 4.11. Как видите, в заголовке окна отображено полное имя внешнего файла.

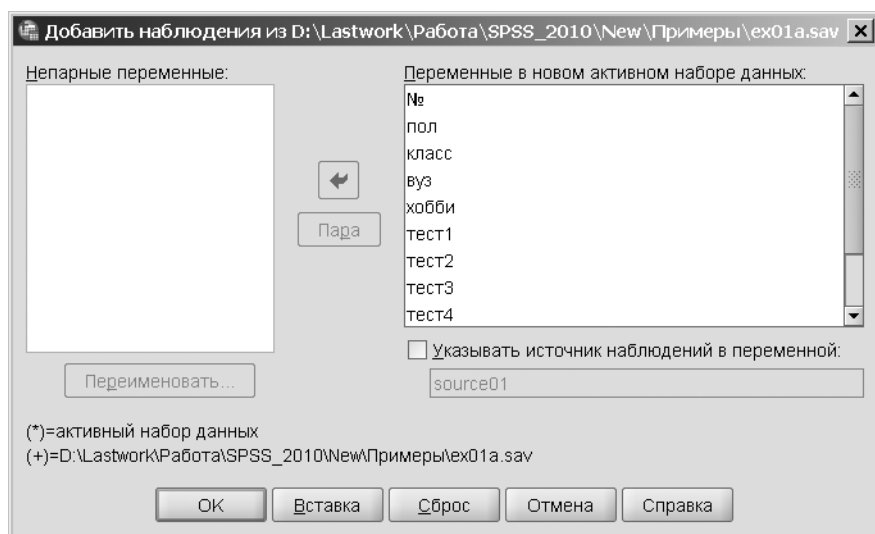


Рис. 4.11. Диалоговое окно Добавить наблюдения

Список Непарные переменные в данном случае пуст (как и следовало ожидать). Если же файлы содержат непарные переменные, все они попадают в этот список: переменные рабочего файла помечаются знаком * (звездочка), а переменные внешнего файла — знаком + (плюс). После того как вы убедитесь, что структуры рабочего и внешнего файлов совпадают, необходимо проверить на идентичность структуру каждой пары переменных. Перед щелчком на кнопке ОК вы можете при необходимости выполнить два дополнительных действия.

- ▶ Если вы не хотите, чтобы какая-либо из парных переменных присутствовала в новом файле, выделите ее щелчком в списке Переменные в новом активном наборе данных и перенесите ее в левое окно при помощи кнопки со стрелкой.
- ▶ По умолчанию все непарные переменные исключаются из нового рабочего файла. Если же вы хотите включить какую-либо из непарных переменных в рабочий файл, выделите ее в списке Непарные переменные и переместите в список Переменные в новом активном наборе данных щелчком на кнопке со стрелкой. Переменная войдет в рабочий файл, а ее отсутствующие значения будут восприниматься как пропущенные.

ШАГ 5К

Чтобы выполнить слияние файлов `ex01.sav` и `ex01a.sav`, в диалоговом окне, показанном на рис. 4.10, задайте имя файла `ex01a.sav`. Потом щелкните на кнопке Продолжить и в открывшемся диалоговом окне (см. рис. 4.11) — на кнопке ОК.

После выполнения этого шага в редакторе данных оказывается открытым новый рабочий файл с добавленными объектами, которому необходимо дать новое имя и сохранить. Содержимое старого рабочего файла и внешнего файла при этом останутся прежними.

Добавление переменных

После завершения предыдущей пошаговой процедуры вам необходимо заново открыть файл `ex01.sav`.

ШАГ 4Л

После выполнения шага 3 должно быть открыто окно редактора данных (`ex01.sav` - Редактор данных IBM SPSS Statistics).

В меню Данные окна редактора данных выберите команду Слить файлы ▶ Добавить переменные. На экране появится диалоговое окно, показанное ранее на рис. 4.10, только заголовок окна будет другим: Добавить переменные.

Слияние с добавлением переменных очень похоже на слияние с добавлением наблюдений. Начало выполнения обеих пошаговых процедур фактически одинаково: на экране появляется диалоговое окно, в котором вы указываете имя внешнего файла и щелкаете на кнопке Продолжить. При добавлении переменных после этого открывается диалоговое окно, показанное на рис. 4.12. В его заголовке отражено полное имя внешнего файла. Обратите внимание, что для демонстрации данной

процедуры мы создали еще один специальный файл ex01b.sav, который содержит переменные идентификационного номера и класса для тех же наблюдений (учащихся), что и в файле ex01.sav, а также новую переменную тест6 — результат дополнительного тестирования.

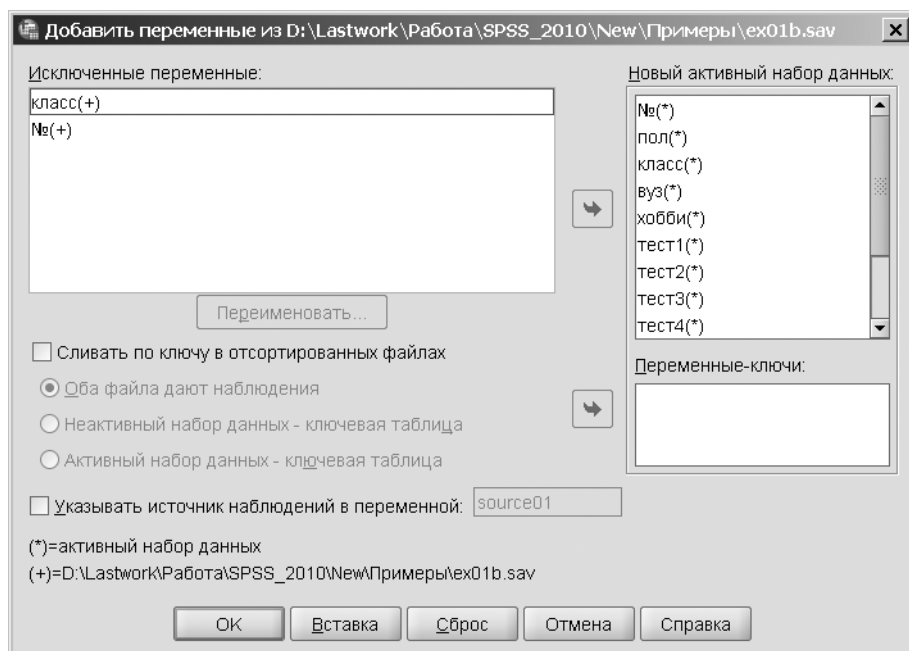


Рис. 4.12. Диалоговое окно Добавить переменные

Парные переменные попадают в список Исключенные переменные и помечаются символом + (плюс), указывающим на то, что они принадлежат внешнему файлу. Переменные, включенные в новый рабочий файл, перечисляются в списке Новый активный набор данных: переменные старого рабочего файла помечаются символом * (звездочка), а новые переменные — символом + (плюс).

Следующим и очень важным действием является задание ключевой переменной. Необходимо, чтобы порядок следования значений в ключевых переменных обоих файлов был одинаковым. В нашем примере в роли ключевой переменной рекомендуется использовать переменную № (идентификационный номер), поскольку ее значения уникальны для каждого наблюдения. Значения переменной № должны быть упорядочены, например, по возрастанию. После того как вы убедитесь, что в обоих файлах ключевые переменные отсортированы одинаково, установите флажок Сливать по ключу в отсортированных файлах, выделите ключевую переменную в списке Исключенные переменные и щелкните сначала на нижней кнопке со стрелкой, чтобы перенести ее в поле Переменные-ключи, а потом — на кнопку ОК.

Ниже приведен пример слияния рабочего файла ex01.sav с внешним файлом ex01b.sav, где в качестве ключевой переменной выступает переменная №, упорядоченная по возрастианию.

ШАГ 5Л

В диалоговом окне Добавить переменные, открытом после выполнения шага 4Л, выполните следующие действия.

1. Задайте имя файла ex01b.sav, а потом щелкните на кнопке Продолжить. На экране появится диалоговое окно, показанное на рис. 4.12.
2. Установите флажок Слить по ключу в отсортированных файлах, в списке Исключенные переменные щелкните на переменной №, чтобы выделить ее, и щелкните на нижней кнопке со стрелкой. Имя № появится в списке Переменные-ключи.
3. Щелкните на кнопке ОК.
4. На экране появится предупреждение Слияние по ключу будет невозможно, если данные не были отсортированы по ключевой переменной в порядке возрастания. Щелкните на кнопке ОК.

Если у вас возникнут проблемы при выполнении этого примера, проверьте, действительно ли переменная № упорядочена по возрастианию в обоих файлах.

После выполнения этого шага в редакторе данных оказывается открытым новый рабочий файл с добавленными объектами, которому необходимо дать новое имя и сохранить. Содержимое старого рабочего файла и внешнего файла при этом останутся прежними.

Агрегирование данных

Агрегирование данных позволяет создавать такие значения переменных, каждое из которых представляет собой результат объединения группы исходных значений, например среднее. В процессе агрегирования задается группирующая переменная (например, номер респондента), каждое значение которой неоднократно встречается в исходных данных. Затем для каждого значения группирующей переменной вычисляются новые значения агрегируемых переменных исходя из их исходных значений по заданной функции (например, среднее).

В качестве примера будем использовать файл ex021.sav (см. описание в главе 14). В нем значения переменных НАЧАЛО, СЕРЕДИНА, КОНЕЦ представлены для каждого респондента дважды (в две строки), в соответствии с переменной Отсрочка: без отсрочки (0) и с отсрочкой (1). Соответственно, номер каждого респондента (переменная N) повторяется дважды. Предположим, исследователь желает агрегировать значения переменных НАЧАЛО, СЕРЕДИНА, КОНЕЦ так, чтобы для респондента с неповторяющимся номером N каждое значение этих переменных представляло собой среднее двух исходных значений (с отсрочкой и без нее).

После выполнения шага 2 на экране должно быть окно редактора данных SPSS.

ШАГ 3М

Откройте файл данных, с которым вы намерены работать (в нашем случае — это файл `ex021.sav`). Если он расположен в текущей папке, то выполните следующие действия.

1. Выберите в меню Файл команду Открыть ► Данные или щелкните на кнопке Открыть данные панели инструментов.
2. В открывшемся диалоговом окне дважды щелкните на имени `ex021.sav` или введите его с клавиатуры и щелкните на кнопке ОК.

ШАГ 4М

1. В меню Данные выберите команду Агрегировать данные, чтобы открыть окно, показанное на рис. 4.13.

Поле Группирующая переменная предназначено для помещения переменной, повторяющимся значениям которой должна соответствовать группа агрегируемых значений. А переменные, для которых агрегируются значения, переносятся в поле Итоги для переменной(ых). Для нашего примера группирующей переменной выступает номер респондента (N), а агрегируемые переменные — НАЧАЛО, СЕРЕДИНА, КОНЕЦ. В связи с тем, что для переменных Инт и Знач значения для повторяющихся N идентичны, то результаты агрегирования для них, в случае вычисления среднего, будут равны исходным значениям. Поэтому они тоже могут быть помещены в список Итоги для переменной(ых).

Группа переключателей Сохранить управляет судьбой агрегированных переменных. В нашем случае целесообразно создание нового набора данных, который после проверки результата агрегирования может быть сохранен.

Кнопка Функция позволяет задавать другие способы агрегирования, нежели принятое по умолчанию вычисление среднего, а кнопка Имя и Метка позволяет задавать имена агрегируемым переменным, иные, чем присваиваемые по умолчанию.

На следующем шаге создадим новый набор данных с именем `ex021agg.sav`, содержащий агрегированные переменные НАЧАЛО, СЕРЕДИНА, КОНЕЦ: каждое новое их значение будет равно среднему двух исходных значений (с отсрочкой и без). Также включим в этот набор значения переменных Инт и Знач.

ШАГ 5М

1. В окне Агрегировать данные щелкните по переменной N, затем — по кнопке со стрелкой, чтобы перенести ее в поле Группирующая переменная.
2. Перенесите переменные Инт, Знач, НАЧАЛО, СЕРЕДИНА, КОНЕЦ в поле Итоги для переменной(ых). Для этого нажмите на клавиатуре клавишу `Ctrl` и, удерживая ее, щелкните последовательно на именах этих переменных, для их выделения. Затем щелкните на соответствующую кнопку со стрелкой.
3. В области Сохранить установите переключатель Создать новый набор данных... и задайте его имя `ex021agg`.
4. Для выполнения команды щелкните по кнопке ОК.

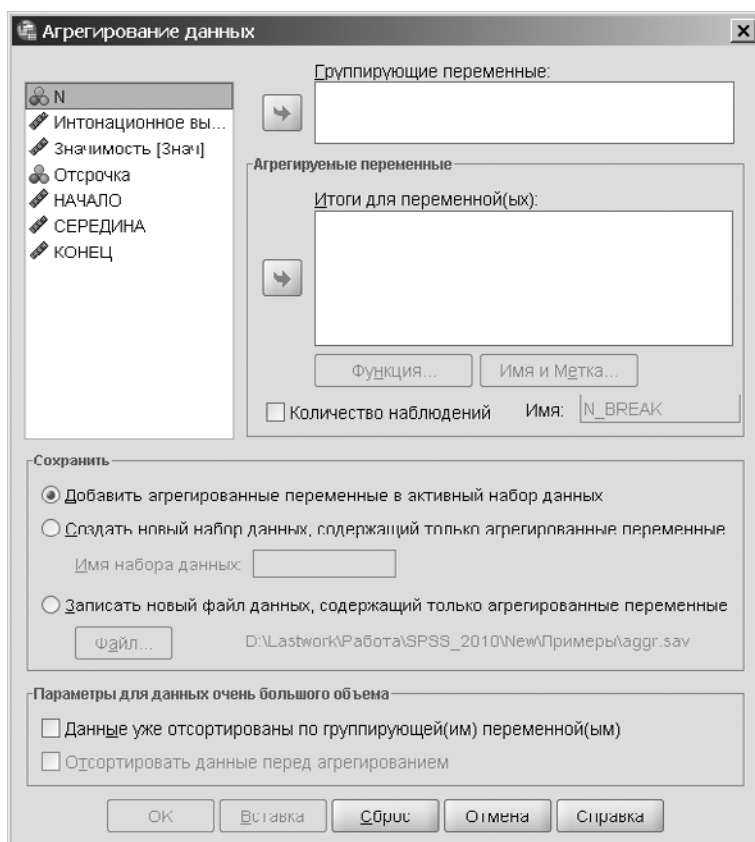


Рис. 4.13. Диалоговое окно Агрегирование данных

В результате выполнения последнего шага создан новый набор данных `ex021agg.sav`, в котором группирующая переменная `N` содержит неповторяющиеся номера, а каждое значение переменных `НАЧАЛО`, `СЕРЕДИНА` и `КОНЕЦ` представляет собой среднее для двух исходных значений соответствующих переменных (с отсрочкой и без нее). Далее этот набор данных может быть сохранен как новый файл данных.

Реструктурирование данных

Команды Реструктурировать меню Данные позволяют производить весьма сложные преобразования структуры данных, например преобразовывать набор переменных в группы значений одной переменной, или наоборот, группы значений одной переменной — в набор переменных. Чаще всего необходимость подобных преобразований возникает в случаях, когда на выборке произведено *несколько повторных измерений* одной переменной или группы переменных. Повторные измерения одной переменной (численностью P повторов) могут быть представлены в файле

как последовательность P переменных для выборки, численностью N (N строк). Те же данные могут быть представлены как P групп значений одной переменной. Тогда группы будет различать группирующая переменная, а количество строк будет $P \times N$. Подобным образом соотносятся файлы данных примеров ex020.sav, ex021.sav, ex022.sav (их описание приведено в главе 14). Далее будут приведены пошаговые инструкции сначала для преобразования 3-х групп по 2 переменные в 3 переменные (с двумя группами значений), а затем — 6 групп значений одной переменной в группу из 3-х переменных.

Преобразование группы переменных в группы значений

Для примера преобразования группы переменных в группы значений воспользуемся файлом ex020. В этом файле для $N = 20$ респондентов приведены 6 измерений продуктивности воспроизведения слов (НАЧАЛО1, ..., КОНЕЦ2): для 3-х частей ряда (начало, середина, конец), без отсрочки (индекс 1) и с отсрочкой (индекс 2). Преобразуем этот набор из 6 переменных в 3 переменные, в соответствии с частями ряда, а для различения групп в соответствии с отсрочкой, введем группирующую переменную Отсрочка.

После выполнения шага 2 на экране должно быть окно редактора данных SPSS.

ШАГ 3Н

Откройте файл данных, с которым вы намерены работать (в нашем случае — это файл ex020.sav). Если он расположен в текущей папке, то выполните следующие действия.

1. Выберите в меню Файл команду Открыть ► Данные или щелкните на кнопке Открыть данные панели инструментов.
2. В открывшемся диалоговом окне дважды щелкните на имени ex020.sav или введите его с клавиатуры и щелкните на кнопке ОК.

ШАГ 4Н

1. В меню Данные выберите команду Реструктурировать, чтобы открыть окно Конструктор реструктуризации данных, показанное на рис. 4.14. Установленный по умолчанию переключатель соответствует задаче преобразования групп переменных в группы наблюдений (значений).
2. Щелкните по кнопке Далее, чтобы перейти к окну Конструктор реструктуризации данных – Шаг 2 из 7, показанному на рис. 4.15.

Под «группой» в данном случае подразумевается набор переменных, который после реструктуризации будет представлен одной переменной. В нашем случае таких групп 3, в соответствии с частями ряда слов, по две переменные в группе (с отсрочкой и без нее). Следовательно, необходимо задать 3 группы.

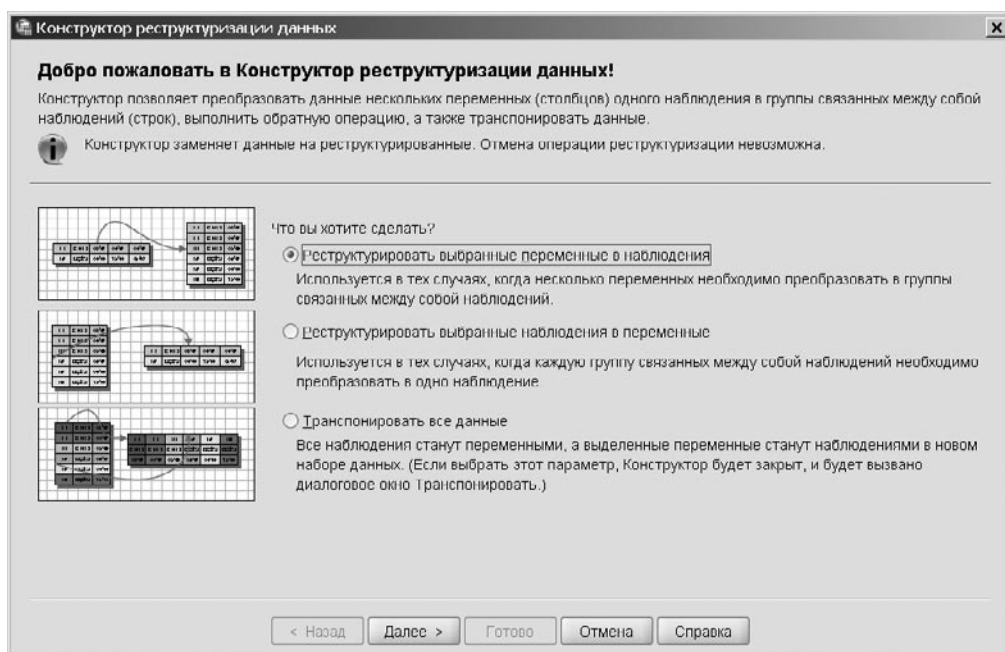


Рис. 4.14. Диалоговое окно Конструктор реструктуризации данных

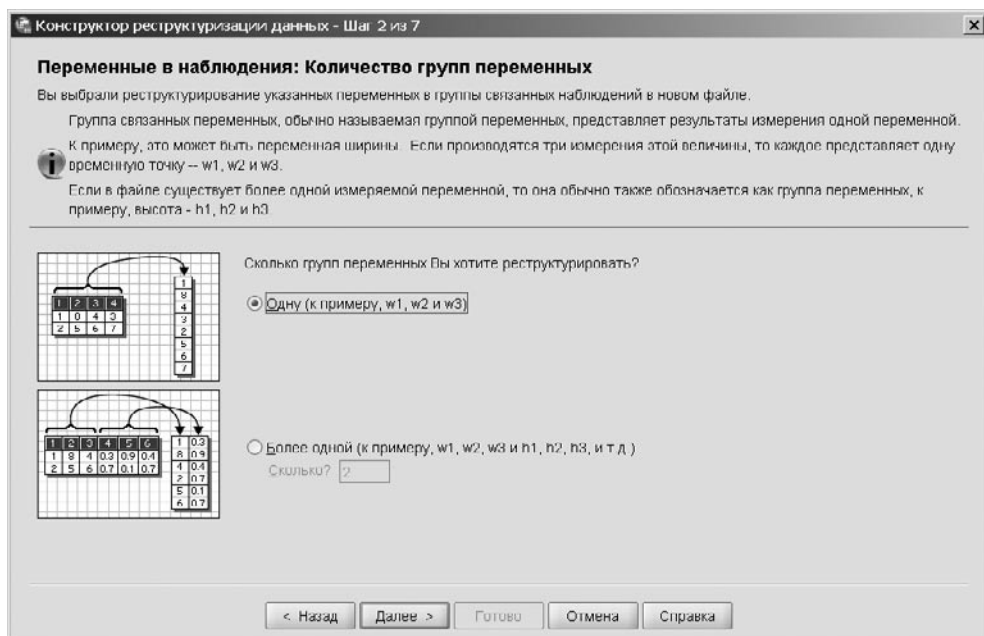


Рис. 4.15. Диалоговое окно Конструктор реструктуризации данных — Шаг 2 из 7

3. Установите переключатель Более одной и в окне Сколько? введите 3.
4. Щелкните на кнопке Далее, чтобы открыть окно Конструктор реструктуризации данных — Шаг 2 из 7, показанное на рис. 4.16.

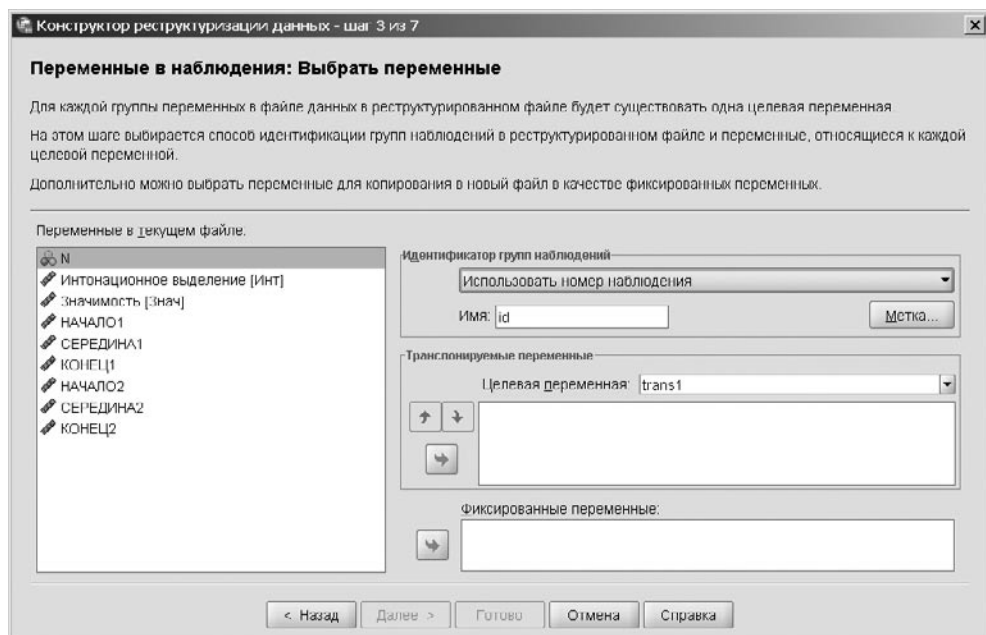


Рис. 4.16. Диалоговое окно Конструктор реструктуризации данных — Шаг 3 из 7

Идентификатор групп наблюдений — это номер группы, одинаковый для значений в одной группе. В данном случае для нумерации групп можно использовать переменную N — номер респондента, но мы оставим установку по умолчанию для контроля результата.

Группа Транспонируемые переменные предназначена для установления соответствия между исходными и итоговыми именами переменных. В итоге планируется на основе исходных 6 создать 3 переменные (НАЧАЛО, СЕРЕДИНА, КОНЕЦ). Следовательно, каждой из этих 3-х новых переменных надо сопоставить 2 исходные переменные.

5. В окне Диалоговое окно Конструктор реструктуризации данных — Шаг 3 из 7 (рис. 4.16) в окне Целевая переменная замените trans1 на Начало.
6. Удерживая нажатой на клавиатуре клавишу CTRL, щелкните сначала на переменной НАЧАЛО1, затем — на переменной НАЧАЛО2, чтобы выделить их, и при помощи кнопки со стрелкой перенесите эти переменные в список Транспонируемые переменные.
7. В окне Целевая переменная раскройте ниспадающий список Целевая переменная и выберите trans2. Замените имя trans2 на Середина.
8. Повторите пункт 6 для переменных СЕРЕДИНА1 и СЕРЕДИНА2.
9. Повторите пункты 7 и 8 для целевой переменной Конеч.

10. В окно Фиксированные переменные перенесите переменные N, Инт и Знач.
11. Щелкните на кнопке Далее, чтобы перейти к следующему окну.
12. В появляющихся далее окнах последовательно щелкайте на кнопке Далее, а в последнем — на кнопке Готово для завершения процедуры реструктурирования.

В итоге будет выполнено реструктурирование данных файла: вместо исходных 6 переменных будет создано 3 новых (НАЧАЛО, СЕРЕДИНА, КОНЕЦ), а количество строк файла удвоится. Кроме того, появится две новые переменные: id и Индекс1. Переменная id дублирует значения переменной N и нумерует группы наблюдений, а переменная Индекс1 различает два наблюдения внутри групп, в соответствии с исходными переменными, входящими в группу (1 — без отсрочки, 2 — с отсрочкой). Имя переменной Индекс1 может быть заменено на другое, в соответствии со смыслом группировки, например, в данном случае на имя Отсрочка.

Важно отметить, что результату реструктурирования приписывается имя исходного файла. Поэтому, чтобы сохранить исходный файл, результирующему файлу следует присвоить новое имя.

Преобразование групп значений в группы переменных

Для примера преобразования групп значений переменной в группу переменных воспользуемся файлом ex021.sav. В этом файле для $N = 20$ респондентов приведено 3 измерения продуктивности воспроизведения слов для 3-х частей ряда (НАЧАЛО, СЕРЕДИНА, КОНЕЦ). Для каждого респондента значения этих переменных представлены дважды, в соответствии с группирующей переменной Отсрочка (1 — без отсрочки, 2 — с отсрочкой). Преобразуем этот набор из 3-х переменных в 6 переменных, в соответствии с 3 частями ряда и отсрочкой (по 2 переменные для каждой части ряда).

После выполнения шага 2 на экране должно быть окно редактора данных SPSS.

ШАГ 30

Откройте файл данных, с которым вы намерены работать (в нашем случае — это файл ex021.sav). Если он расположен в текущей папке, выполните следующие действия.

1. Выберите в меню Файл команду Открыть ► Данные или щелкните на кнопке Открыть данные панели инструментов.
2. В открывшемся диалоговом окне дважды щелкните на имени ex021.sav или введите его с клавиатуры и щелкните на кнопке ОК.

ШАГ 40

1. В меню Данные выберите команду Реструктурировать, чтобы открыть окно Конструктор реструктуризации данных, показанное на рис. 4.14.
2. Установите переключатель в положение Реструктурировать выбранные наблюдения в переменные и щелкните на кнопке Далее, чтобы перейти к окну Конструктор реструктуризации данных — Шаг 2 из 5, показанному на рис. 4.17.

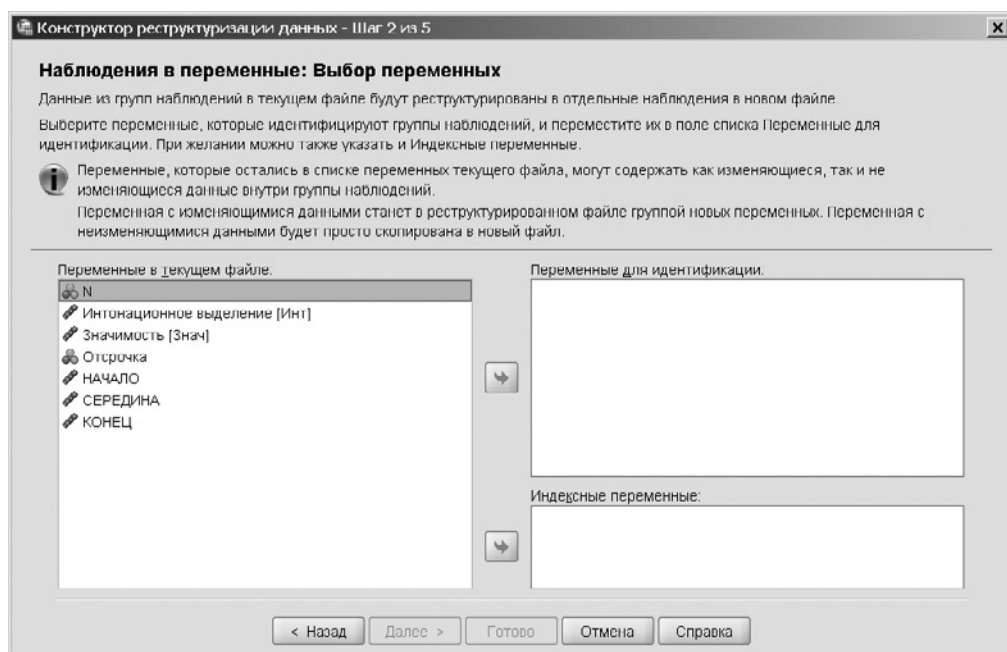


Рис. 4.17. Диалоговое окно Конструктор реструктуризации данных — Шаг 2 из 5

В окне Наблюдения в переменные: Выбор переменных (см. рис. 4.17) необходимо задать переменную для идентификации групп наблюдений. В файле ex021.sav это переменная N. Индексная переменная является группирующей внутри групп наблюдений и задает индексы для новых переменных. В нашем примере это переменная Отсрочка.

3. Выделите переменную N и перенесите ее в поле Переменные для идентификации при помощи кнопки со стрелкой.
4. Выделите переменную Отсрочка и перенесите ее в поле Индексные переменные при помощи кнопки со стрелкой.
5. Щелкните на кнопке Далее, чтобы перейти к следующему окну.
6. В появляющихся далее окнах последовательно щелкайте на кнопке Далее, а в последнем — на кнопке Готово, для завершения процедуры реструктурирования.

В итоге реструктурирования будет получен новый набор данных, содержащий вместо 3-х исходных переменных НАЧАЛО, СЕРЕДИНА, КОНЕЦ, 6 переменных (НАЧАЛО.1, НАЧАЛО.2, ... КОНЕЦ.2) в соответствии с двумя значениями группирующей переменной Отсрочка. Те переменные, значения которых дублируются внутри групп (N, Инт, Знач), войдут в новый набор данных, а индексная переменная (Отсрочка) в новом наборе будет отсутствовать. Чтобы сохранить исходный файл, результирующему файлу следует присвоить новое имя.

5 Диаграммы

91	Графика в программе IBM SPSS Statistics
92	Настройка диаграмм
93	Команды построения диаграмм
103	Редактирование диаграмм
105	Выход из программы

Программа IBM SPSS Statistics обладает обширным арсеналом мощных и эффективных средств построения диаграмм. Разумеется, в книгу такого объема, как эта, бессмысленно даже пытаться вместить описания всех этих средств. Мы лишь коснемся основных аспектов работы с графическими объектами.

Графика в программе SPSS

Как правило, диаграмма зависит от конкретной статистической процедуры, следовательно, рассматривать все тонкости построения диаграмм имеет смысл лишь в контексте обработки данных. Поэтому вся специфическая информация относительно конкретных диаграмм излагается по мере необходимости в главах книги, посвященных статистическому анализу. А здесь собраны лишь общие сведения, касающиеся диаграмм вообще, и по большей части эти сведения связаны с редактированием уже созданных диаграмм.

Ниже перечислены далеко не все, но наиболее часто используемые виды графиков.

- **Столбики (столбиковые диаграммы).** Эти графики применяются для отображения итогов для значений (категорий) переменной. Столбики соответствуют категориям переменной. Высота каждого столбика пропорциональна вычисляемому итогу. В качестве итога по умолчанию вычисляется количество наблюдений для данной категории, то есть по умолчанию строится диаграмма распределения частот. Например, с помощью столбиковой диаграммы удобно представить распределение учащихся по трем классам или по их ориентации на поступление в вузы четырех типов и т. д. В качестве итога можно задать статистику для другой переменной, например среднее. Так, задав для переменной класс в качестве итога среднее для переменной отметка, можно получить диаграмму средней отметки для разных классов. Столбиковые диаграммы делятся на простые, кластерные и состыкованные.

- ▶ Гистограммы внешне напоминают столбиковые диаграммы, однако, как правило, иллюстрируют распределение наблюдений по диапазонам значений непрерывной переменной (имеющей большое число возможных значений). С помощью гистограммы было бы удобно представить распределение учащихся по диапазонам значений успеваемости (отметки) или диапазонам тестовых значений.
- ▶ Линии (линейные диаграммы) отображают то же, что и столбики, но в виде ломаной линии.
- ▶ Круг (круговые диаграммы), как и столбиковые, зачастую применяется для иллюстрации распределения наблюдений в различных категориях.
- ▶ Ящики (коробчатые диаграммы) основаны на процентилях и являются прекрасным средством отображения различия между группами с учетом распределения данных.
- ▶ Рассеяния/точки (диаграммы рассеяния) часто используются для отображения характера взаимосвязи между переменными.

В меню Графика окна редактора данных имеется три подменю (SPSS 19): Конструктор диаграмм, Панель выбора диаграмм и Устаревшие диалоговые окна. Последнее подменю содержит список стандартных графиков, которые мы перечислили выше. Меню Конструктор диаграмм и Панель выбора диаграмм предоставляют дополнительную, очень удобную возможность построения графиков в диалоговом режиме. В этих меню представлен широкий спектр команд, позволяющих строить диаграммы в интерактивном режиме: добавлять переменные, изменять категории данных и т. п. Другими словами, работая исключительно с графиками, вы можете добиться результата и без обращения к вычислительным статистическим процедурам. Иногда такая возможность может оказаться полезной для пользователя, однако до этого необходимо освоить стандартные приемы работы с SPSS как в отношении статистики, так и в отношении графики. Поэтому описание процедур конструктора и выбора диаграмм выходит за рамки темы этой книги.

Настройка диаграмм

Установленные по умолчанию настройки для диаграмм SPSS далеко не всегда соответствуют требованиям пользователя программы. Например, вас может не устраивать установленный цвет или вообще то, что диаграммы цветные. Проблемой может оказаться и некорректное отображение русскоязычных наименований переменных и их меток, что было характерно для ранних версий SPSS. Средства редактирования диаграмм позволяют решить, по крайней мере, большую часть этих и подобных проблем для каждого графика по отдельности. Но есть более радикальный способ — изменение общих настроек диаграмм, которые станут в дальнейшем настройками программы SPSS по умолчанию.

Для решения проблем общих настроек диаграмм необходимо в строке меню редактора данных открыть меню Правка и выбрать команду Параметры. В открывшемся

диалоговом окне следует перейти на вкладку Диаграммы. На этой вкладке можно задать по своему вкусу основные свойства диаграмм, которые в дальнейшем будут применены программой по умолчанию ко всем графикам.

Рассмотрим решение наиболее часто встречающейся задачи настройки диаграмм: установка черно-белых графиков.

В программе SPSS по умолчанию установлены цветные графики: Параметры цикла стиля: цикл только по цветам. При этом однотипные (повторяющиеся) элементы, относящиеся к разным категориям, будут различаться цветом. Для установки черно-белых графиков на вкладке Диаграммы окна Параметры в раскрывающемся меню Параметры цикла стиля достаточно выбрать Цикл только по узорам. В этом случае отображаемые на диаграмме элементы будут различаться «узором», а не цветом: (сплошные и пунктирные линии, столбики со штриховкой и без нее и т. д.).

Можно также задать стиль графиков, например, в соответствии с рекомендациями APA (Американской психологической ассоциации). Для этого в разделе Шаблон диаграмм следует установить переключатель в положение Использовать шаблон диаграмм и указать путь к месту, где этот шаблон расположен. Если установлена версия SPSS 19, это путь: C:\Program Files\IBM\SPSS\Statistics\19\Looks\APA_Styles.sgt. Этот шаблон под именем APA_Styles.sgt находится также и в папке файлов примеров.

Команды построения диаграмм

Большинство статистических процедур, описанных в последующих главах книги, имеют встроенные средства построения графиков. Тем не менее программа SPSS предоставляет возможность разделения задач построения графиков и статистического анализа. Вы можете с равным успехом отредактировать как готовую диаграмму, сформированную статистической процедурой, так и диаграмму, созданную при помощи меню Диаграммы. Примеры и рекомендации, которые вы здесь найдете, позволят вам эффективно работать с диаграммами.

Основные типы команд и диалоговых окон мы рассмотрим на примере построения наиболее часто используемых столбиковых диаграмм. Будут представлены пошаговые процедуры построения столбиковых диаграмм распределения частот (простые — рис. 5.1 и кластеризованные — рис. 5.4), а также средних значений (простые — рис. 5.6 и кластеризованные — рис. 5.9).

Для построения диаграмм мы будем использовать уже знакомый вам файл ex01.sav. Однако сначала следует выполнить три подготовительных шага. Эти шаги (шаги 1–3) позволят подготовить рабочий файл данных, запустить программу IBM SPSS Statistics 19 и открыть файл (в данном случае — файл ex01.sav) Пошаговые инструкции этого процесса приведены в главе 4 (с. 60), а подробные разъяснения — в главе 2.

После завершения шага 3 на экране должно присутствовать окно редактора данных со строкой меню и загруженным файлом ex01.sav.

Сначала приведем пошаговые процедуры построения простой столбиковой диаграммы распределения частот (см. рис. 5.1).

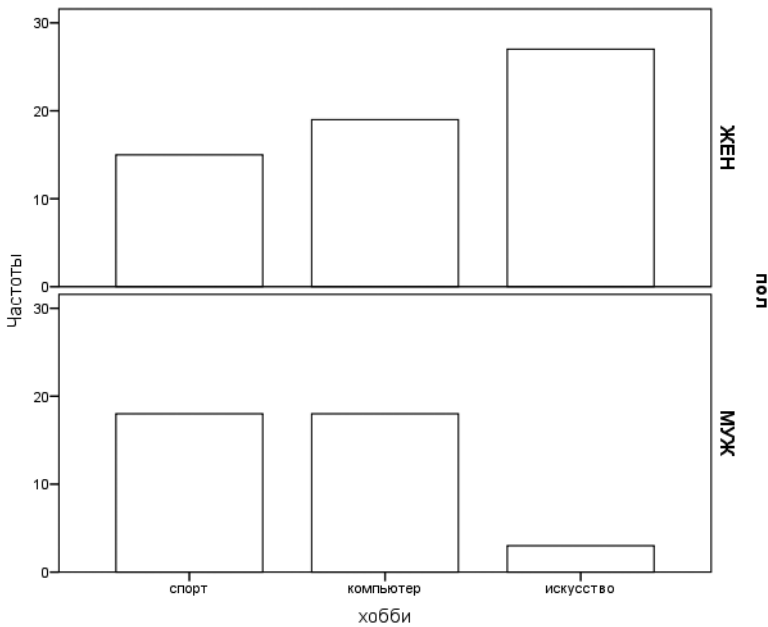


Рис. 5.1. Простая столбиковая диаграмма распределения частот

ШАГ 4

В меню Графика выберите команду Устаревшие диалоговые окна ► Столбики. На экране появится диалоговое окно, показанное на рис. 5.2

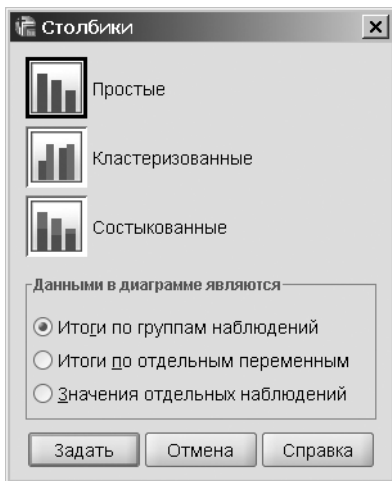


Рис. 5.2. Диалоговое окно задания типа столбиковой диаграммы

С помощью элементов интерфейса верхней части окна вы можете выбрать тип диаграммы — Простые, Кластеризованные и Состыкованные, — а в нижней части

указать, какие данные вы намерены использовать. Типы диаграмм будут меняться в зависимости от вида диалогового окна; для того чтобы указать нужный тип, щелкните на кнопке с миниатюрой диаграммы слева от ее названия. Варианты используемых данных для большинства типов диаграмм одни и те же.

- ▶ Переключатель Итоги по группам наблюдений означает, что задействуется одна переменная, а столбцы диаграммы отразят число или процент наблюдений каждой градации этой переменной. Например, для переменной класс диаграмма будет состоять из трех столбцов, соответствующих трем классам учащихся; для переменной пол — из двух столбцов, отображающих число (%) учащихся женского и мужского пола; для переменной вуз — из четырех столбцов, представляющих количество (%) учащихся в каждой из четырех групп предпочтительных вузов.
- ▶ Переключатель Итоги по отдельным переменным указывает на то, что диаграмма будет содержать несколько столбцов, каждый из которых будет соответствовать среднему значению одной из переменных. Как правило, при построении этого типа диаграмм используют переменные, связанные между собой по смыслу. В файле ex01.sav такими переменными являются переменные тест1, ..., тест5: диаграмма, построенная для них, отразит средние баллы всех учащихся для каждого из пяти тестов.
- ▶ Переключатель Значения отдельных наблюдений предназначен для данных с относительно небольшим числом объектов. Если взять данные из файла ex01.sav и построить подобную диаграмму для какой-либо переменной, например тест1, то на экране появились бы 100 столбцов, отображающих баллы каждого из учащихся по первому тесту.

ШАГ 5

После выполнения шага 4 должно быть открыто диалоговое окно, изображенное на рис. 5.2.

Оставив выделенным по умолчанию тип диаграммы Простые и переключатель в положении Итоги по группам наблюдения, щелкните на кнопке Задать, чтобы открыть диалоговое окно, фрагмент которого показан на рис. 5.3.

В верхней части окна находится группа из пяти переключателей. Чаще других используются переключатели N наблюдений и % наблюдений — каждый столбец или сегмент графика представляет число (процент) наблюдений в каждой из категорий. При выборе переключателя Другую статистику (например, среднее) становится доступным окно Переменная, находящееся ниже. Если установить этот переключатель и ввести в окно имя количественной переменной, то столбики будут отражать средние этой переменной по каждой из категорий.

Поле Категориальная ось служит для задания имени переменной, для категорий которой будут строиться столбики, в нашем примере это переменная Хобби. Поля в разделе Панель по (Строки и Столбцы) позволяют задавать имена переменных, для категорий которых будут строиться отдельные графики. В нашем примере такой переменной является Пол.

Столбики представляют

☒ N наблюдений ☐ % наблюдений

☐ Накопленное N ☐ Накопленный %

☐ Другую статистику (например, среднее)

Переменная:

Изменить статистику...

Категориальная ось:

Панель по

Строки:

☐ Вкладывать переменные (не выводить пустые строки)

Столбцы:

☐ Вкладывать переменные (не выводить пустые столбцы)

Рис. 5.3. Фрагмент диалогового окна Простые столбики: Итожащие функции по группам наблюдений

Продолжим построение столбиковой диаграммы для переменной Хобби с учетом значений переменной Пол, показанной на рис. 5.1. После выполнения шага 5 на экране должно быть окно, фрагмент которого показан на рис. 5.3.

ШАГ 6

1. В списке переменных слева щелчком выделите переменную Хобби и при помощи стрелки перенесите ее в поле Категориальная ось.
2. В списке переменных слева щелчком выделите переменную Пол и при помощи стрелки перенесите ее в поле Строки раздела Панель по.
3. Оставив установленный по умолчанию переключатель N наблюдений в разделе Столбики представляют, щелкните на кнопке ОК.

В результате выполнения шагов 4–6 появится окно вывода со столбиковой диаграммой, изображенной на рис. 5.1.

В следующем примере мы снова построим столбиковую диаграмму для переменной Хобби с учетом переменной Пол, однако для этого воспользуемся кластеризованной диаграммой и представим столбиками не N , а процент наблюдений. В результате должна получиться диаграмма, изображенная на рис. 5.4.

Выполните шаг 4. После его выполнения должно быть открыто диалоговое окно, изображенное на рис. 5.2.

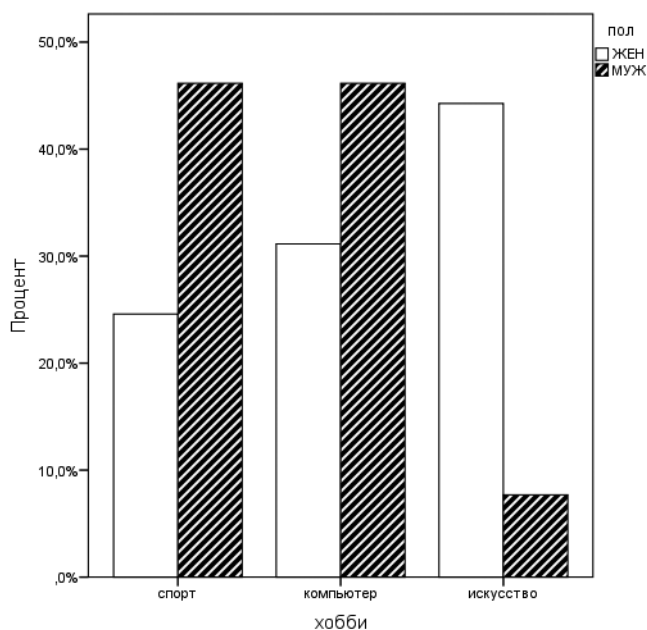


Рис. 5.4. Классифицированная столбчатая диаграмма распределения частот (в %)

ШАГ 5А

Щелчком мыши выделите тип диаграммы Классифицированные и щелкните на кнопке Задать, чтобы открыть диалоговое окно, фрагмент которого показан на рис. 5.5.

Столбцы представляют

☒ N наблюдений ☐ % наблюдений

☐ Накопленное N ☐ Накопленный %

☐ Другую статистику (например, среднее)

Переменная:

Измeнить статистику...

Категориальная ось:

Задать кластеры по:

Панель по

Рис. 5.5. Фрагмент диалогового окна Классифицированные столбцы:
Итожающие функции по группам наблюдений

Единственное отличие диалогового окна на рис. 5.5 (для классифицированных столбцов) от изображенного на рис. 5.3 (для простых столбцов) — в наличии поля

Задать кластеры по. В нем следует указать имя переменной, для значений которой будут строиться сопряженные столбики в каждой категории основной переменной. В нашем примере кластеры задаются по переменной Пол.

ШАГ 6А

1. В списке переменных слева щелчком выделите переменную Хобби и при помощи стрелки перенесите ее в поле Категориальная ось.
2. В списке переменных слева щелчком выделите переменную Пол и при помощи стрелки перенесите ее в поле Задать кластеры по.
3. В разделе Столбики представьте установите переключатель в положение % наблюдений и щелкните на кнопке ОК.

В результате выполнения последовательности шагов 4, 5а и 6а появится окно вывода с кластеризованной столбиковой диаграммой, изображенной на рис. 5.4. Обратите внимание, что за 100 % принимается количество наблюдений отдельно для муж и жен, то есть внутри каждой категории переменной Пол, по которой были заданы кластеры.

В следующих двух примерах мы рассмотрим построение и редактирование столбиковых диаграмм для средних значений (рис. 5.6 и 5.9).

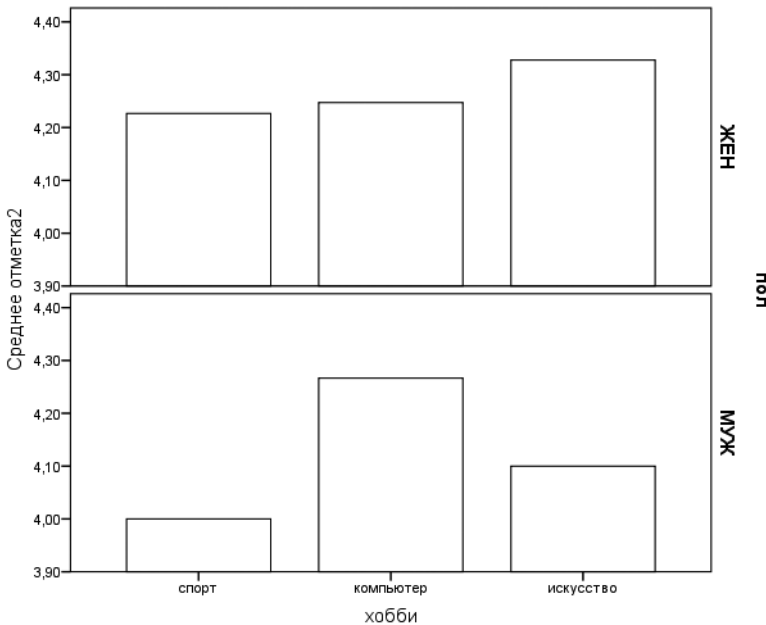


Рис. 5.6. Простая столбиковая диаграмма средних значений

Сначала приведем пошаговые инструкции построения простой столбиковой диаграммы средних значений переменной отметка2 для категорий переменной Хобби с учетом категорий переменной Пол (см. рис. 5.6).

Выполните шаг 4. После его выполнения должно быть открыто диалоговое окно, изображенное на рис. 5.2.

ШАГ 5Б

1. Щелчком мыши выделите тип диаграммы Простые и оставьте неизменным установленный переключатель Итоги по группам наблюдений. Щелкните на кнопке Задать, чтобы открыть диалоговое окно Простые столбики: Итожащие функции по группам наблюдений, фрагмент которого показан на рис. 5.3. Если вы уже поработали с этим окном, щелкните на кнопке Сброс.
2. В качестве категориальной оси задайте переменную Хобби. Для этого щелчком мыши выделите ее в списке слева и при помощи стрелки перенесите ее в поле Категориальная ось.
3. Задайте отдельные графики для категорий переменной Пол. Для этого щелчком мыши выделите ее в списке слева и при помощи стрелки перенесите в поле Строки раздела Панель по.

До этого момента шаги построения графика средних совпадают с построением графика распределения частот. Для установки средних необходимо в разделе Столбики представить установить переключатель Другую статистику (например, среднее). Тогда станет доступным окно для указания количественной переменной, по значениям которой будут вычисляться средние. При задании такой переменной станет доступной кнопка Изменить статистику, при щелчке на ней появляется окно, показанное на рис. 5.7. При помощи этого окна можно задать другую статистику вместо среднего, например медиану, если переменная порядковая (ранговая).

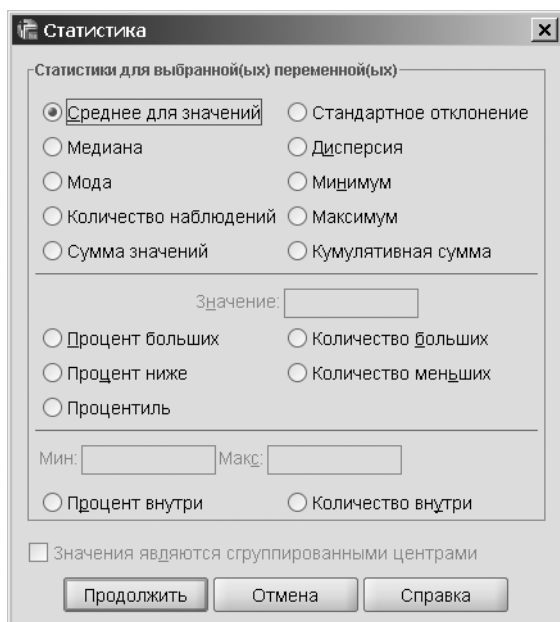


Рис. 5.7. Переключатели выбора статистики для диаграммы

ШАГ 6Б

После выполнения шага 5Б должно быть открыто окно, фрагмент которого показан на рис. 5.3. В этом окне заданы Хобби в поле Категориальная ось и Пол в поле Строки.

1. В разделе переключателей Столбики представьте установите переключатель в положение Другую статистику (например, среднее). Станет доступным поле Переменная, находящееся под установленным переключателем.
2. В списке слева выделите щелчком мыши переменную отметка2 и при помощи стрелки перенесите ее в поле Переменная.
3. Щелкните на кнопке ОК, чтобы открыть окно вывода с диаграммой.

Диаграмма, полученная в результате выполнения последнего шага 6Б, отличается от изображенной на рис. 5.6 тем, что вертикальная ось средних начинается со значения 0. В результате различия между средними визуально едва заметны. Исправить положение и сделать диаграмму средних более наглядной может изменение точки отсчета вертикальной оси, например, на значение 3,9. Следующий шаг демонстрирует некоторые возможности редактирования диаграммы.

ШАГ 7Б

После выполнения шага 6Б должно быть открыто окно вывода с диаграммой средних значений.

1. Двойным щелчком мыши на диаграмме откройте окно Редактор диаграмм, фрагмент которого показан на рис. 5.11.
2. В окне редактора диаграмм щелкните на вертикальной оси графика, чтобы выделить ее. Затем сделайте на этой оси двойной щелчок, чтобы открыть окно Свойства, показанное на рис. 5.8.
3. В диалоговом окне Свойства на вкладке Ось в строке Минимум задайте в поле значение 3,9 и щелкните на кнопке Применить, чтобы изменения вступили в силу.
4. Закройте окно Редактор диаграмм, выбрав в его командной строке Файл ► Закреть.

В результате выполнения последнего шага в окне вывода появится диаграмма средних значений, аналогичная изображенной на рис. 5.6.

В следующем примере мы построим и отредактируем кластеризованную столбиковую диаграмму средних значений для нескольких переменных. Такие диаграммы можно строить для повторных измерений или для других «родственных» переменных, представленных в едином масштабе (в одной шкале). В нашем случае такими переменными являются показатели интеллекта тест1, тест2, ..., тест5. На рис. 5.9 изображена кластеризованная диаграмма средних значений этих переменных для разных категорий переменной пол. Выполнение шагов 5в–6в реализует построение этой диаграммы.

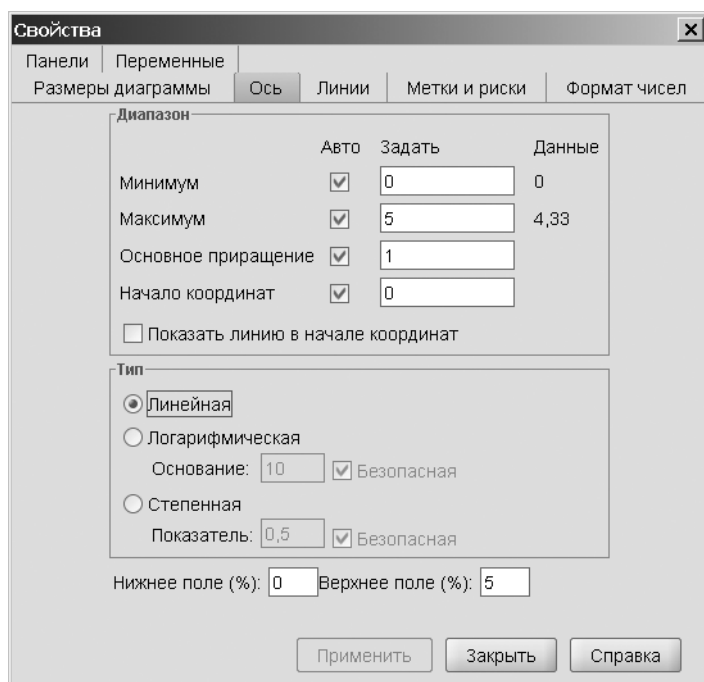


Рис. 5.8. Диалоговое окно редактора диаграмм Свойства

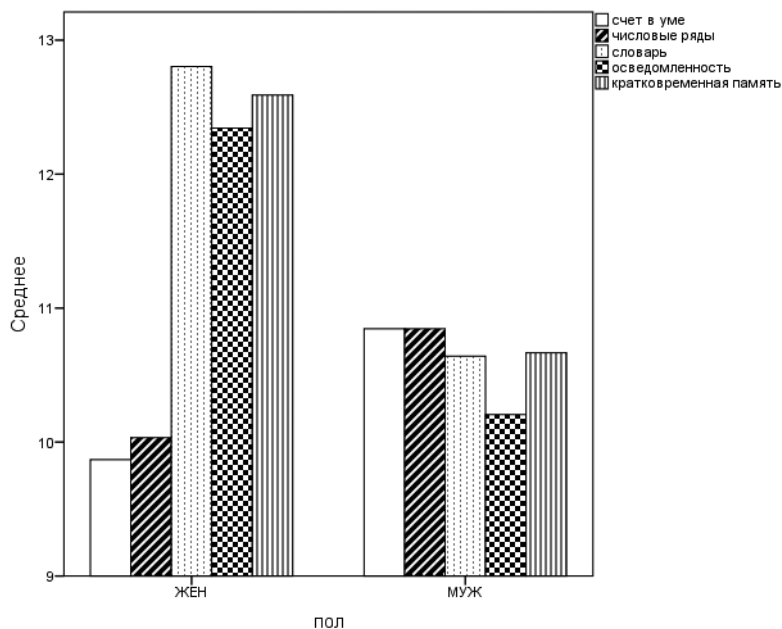


Рис. 5.9. Кластеризованная диаграмма средних значений

ШАГ 5В

Выполните шаг 4. После его выполнения должно быть открыто диалоговое окно Столбики, изображенное на рис. 5.2.

1. Щелчком мыши выделите тип диаграммы Кластеризованные и установите переключатель Итоги по отдельным переменным. Щелкните на кнопке Задать, чтобы открыть диалоговое окно Кластеризованные столбики: Итожащие функции по переменным, фрагмент которого показан на рис. 5.10.
2. В качестве категориальной оси задайте переменную Пол. Для этого щелчком мыши выделите ее в списке слева и при помощи стрелки перенесите ее в поле Категориальная ось.
3. Задайте переменные, по которым будут вычисляться средние значения. Для этого щелчком мыши выделите переменную тест1 в списке слева и при помощи стрелки перенесите ее в поле Столбики представляют. Повторите эту последовательность для переменных тест2, тест3, тест4 и тест5.
4. Щелкните на кнопке ОК, чтобы открыть окно вывода с диаграммой.

Кластеризованная диаграмма, полученная в результате выполнения шага 5в, отличается от изображенной на рис. 5.9 тем, что вертикальная ось средних начинается со значения 0. Сделать диаграмму средних более наглядной можно так же, как и в предыдущем примере, изменив точку отсчета вертикальной оси с 0 на 9.

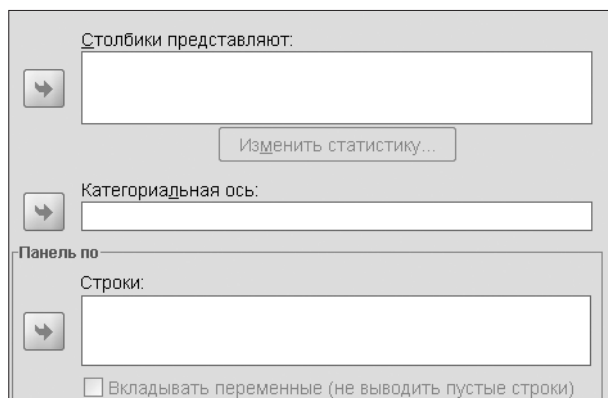


Рис. 5.10. Фрагмент диалогового окна Кластеризованные столбики: Итожащие функции по переменным

ШАГ 6В

После выполнения шага 5в должно быть открыто окно вывода с кластеризованной диаграммой средних значений.

1. Двойным щелчком мыши на диаграмме откройте окно Редактор диаграмм, фрагмент которого показан на рис. 5.11.
2. В окне редактора диаграмм щелкните на вертикальной оси графика, чтобы выделить ее. Затем сделайте на этой оси двойной щелчок, чтобы открыть окно Свойства, аналогичное показанному на рис. 5.8.

3. В диалоговом окне Свойства на вкладке Ось в строке Минимум задайте в поле значение 9 и щелкните на кнопке Применить, чтобы изменения вступили в силу.
4. Закройте окно Редактор диаграмм, выбрав в его командной строке Файл ► Заккрыть.

В результате выполнения последнего шага в окне вывода появится кластеризованная диаграмма средних значений, идентичная изображенной на рис. 5.9.

Редактирование диаграмм

После создания диаграммы можно воспользоваться большим набором команд для ее редактирования. С некоторыми из них вы познакомились при построении столбиковых диаграмм в предыдущем разделе. Для редактирования диаграммы необходимо в окне вывода дважды щелкнуть мышью на выбранном графическом объекте. При этом на экране появится окно Редактор диаграмм, содержащее строку меню с набором команд и панель инструментов, на которой находятся кнопки, соответствующие наиболее часто используемым командам (рис. 5.11).

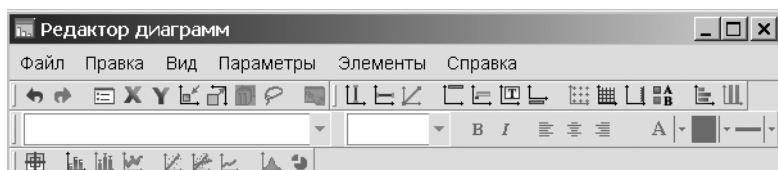


Рис. 5.11. Редактор диаграмм: строка меню и панель инструментов

Команды редактирования могут применяться к диаграмме только в том случае, если она отображена на экране в режиме редактирования. Как только вы дважды щелкнете на диаграмме, SPSS откроет новое окно с ее изображением и строкой меню в верхней части; это означает, что диаграмма доступна для редактирования.

Начиная с SPSS 13.0 графический редактор существенно изменился: теперь можно обойтись без команд строки меню и без панели управления редактора. Для редактирования любого элемента графика достаточно навести на него курсор мыши и щелкнуть левой кнопкой мыши для его выделения.

Если выделенный элемент диаграммы — текст, который вы хотите изменить, достаточно еще раз щелкнуть на нем правой кнопкой мыши. После этого текст станет доступен в режиме обычного текстового редактора.

Если выделенный элемент является графическим (линия, маркер, заливка и т. д.), то двойной щелчок на нем левой кнопкой мыши открывает диалоговое окно для его редактирования. Большим удобством стало то, что для каждого элемента диаграммы открывается диалоговое окно, предоставляющее широкие возможности редактирования именно этого элемента.

Рассмотрим пример редактирования гистограммы из главы 6 (рис. 6.7. на с. 115). Если навести курсор на столбцы гистограммы, выделить их щелчком левой кнопки мыши и еще дважды быстро щелкнуть левой кнопкой, появится диалоговое окно, фрагмент которого изображен на рис. 5.12.

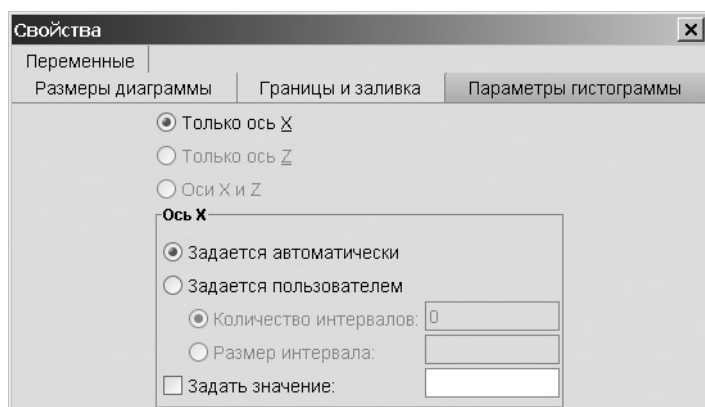


Рис. 5.12. Фрагмент диалогового окна редактирования гистограммы

Вкладки этого окна зависят от возможностей редактирования конкретного элемента. В данном случае окно открывается на вкладке Параметры гистограммы. Если в разделе Ось X переключатель установить в положение Задается пользователем, вместо параметров гистограммы по умолчанию можно задать: Размер интервала или Количество интервалов.

После внесения изменений становится доступной кнопка Применить. При нажатии на нее изменения данного элемента вступают в силу.

Щелчок правой кнопкой мыши открывает общее окно диалога, форма которого не зависит от того, в каком месте диаграммы находится курсор мыши. Это окно содержит команды, соответствующие панели инструментов редактирования диаграмм с тем набором команд, которые доступны для данного графика. Некоторые команды при этом могут быть недоступны (затенены) — в зависимости от возможностей редактирования элемента, на котором в момент щелчка находился курсор.

Заканчивая обзор средств редактирования диаграмм, отметим, что возможности графического редактора SPSS настолько широки, что описание даже части их не представляется возможным в рамках данной книги. В этом нет и необходимости, поскольку редактор обладает высокоразвитым интуитивным интерфейсом. Поэтому изложенных сведений вполне достаточно, чтобы без предварительной подготовки сразу приступить к редактированию графиков. Если по мере редактирования вы почувствовали, что лучший вариант уже был раньше, вы всегда сможете вернуться и найти его при помощи двух кнопок:

 Отменить действие пользователя.

 Вернуть действие пользователя.

После окончания редактирования графика достаточно щелкнуть на кнопке Закрыть или выбрать в командной строке Файл ► Закрыть, чтобы закрыть редактор диаграмм. В окне вывода вы увидите окончательный вариант диаграммы.

Выход из программы

Для дальнейшего использования окончательного результата все содержимое окна вывода или его фрагменты можно сохранить в файле *.spv, экспортировать в другой формат (например, Word), перенести в документ Word или вывести на печать (подробности см. в разделе «Сохранение, экспорт, перенос и печать результатов» главы 2).

Для выхода из программы выберите команду Выход в меню Файл.

6 Частоты

107	Пошаговые алгоритмы вычислений
113	Представление результатов
116	Завершение анализа и выход из программы

В этой главе рассматриваются частоты, их графическое представление (столбиковые и круговые диаграммы), гистограммы и процентиля. Частоты — одна из команд, позволяющих непосредственно работать с диаграммами, вообще не прибегая к командам меню Графика. Некоторые детали, касающиеся редактирования диаграмм, были рассмотрены в главе 5. *Столбиковые* и *круговые диаграммы* отображают количество объектов (*частоту*) в различных категориях дискретной переменной (имеющей небольшое число возможных значений). Однако команда Частоты позволяет работать не только с дискретными, но и с непрерывными переменными, которые могут принимать большое число разных значений.

Ниже даны краткие определения понятий, о которых пойдет речь в этой главе.

- *Частоты*. Команда Частоты является одной из самых простых и часто используемых команд SPSS. Действие команды сводится к подсчету количества объектов в каждой категории переменной. Это называется распределением частот по категориям переменной. Если мы анализируем переменную пол, программа подсчитает распределение численности девушек и юношей среди учащихся; если используется переменная класс, мы получим распределение численности учащихся по классам. Выводимый результат для каждой категории включает метку значения переменной, само значение переменной, частоту, процент и накопленный процент от общей частоты.
- *Столбиковые диаграммы*. Обязательное условие для применимости столбиковой диаграммы — дискретность отображаемых данных. Такими свойствами обладают переменные пол, класс, хобби и т. п. Каждая из этих переменных делит все объекты файла данных на категории. Например, переменная вуз разбивает всех учащихся на 4 группы по выбору профиля дальнейшего образования. Столбиковыми диаграммами можно также отображать и непрерывные переменные. При этом к категории относятся данные, попадающие в определенный диапазон (интервал). Примерами непрерывных переменных является средний балл отметки каждого учащегося, время финиша в гоночных состязаниях, вес пациента при медосмотре и т. п.
- *Гистограммы* используются для отображения распределения частот непрерывных переменных. Особенностью гистограмм, отличающей их от столбиковых диаграмм, является то, что каждый столбец представляет не отдельное значение, а диапазон значений переменной. Гистограммы могут использоваться и для дискретных данных в случаях, если число значений переменной слиш-

ком велико, чтобы отображать каждое из них отдельным столбцом диаграммы. Весь диапазон изменчивости переменной разделяют на интервалы, а затем по ним строят гистограмму. Именно таким методом следовало бы отображать распределение частот для переменных `отметка1` или `отметка2`.

- *Процентиль* показывает, какой процент распределения лежит ниже заданной величины. Например, если говорят, что процентиль значения 111 равен 75, это означает, что 75 % всех значений переменной меньше, чем 111, а 25 % — больше, чем 111.

Пошаговые алгоритмы вычислений

Для иллюстрации вычислений при помощи команды Частоты мы будем использовать уже знакомый вам файл `ex01.sav` (мы не приводим примеры круговых диаграмм ввиду их исключительной простоты). Число объектов в файле равно 100. Однако сначала следует выполнить три подготовительных шага. Эти шаги (шаг 1–3) позволят подготовить рабочий файл данных, запустить программу IBM SPSS Statistics 19 и открыть файл (в данном случае — файл `ex01.sav`). Пошаговые инструкции этого процесса приведены в главе 4 (с. 60), а подробные разъяснения — в главе 2.

После завершения шага 3 на экране должно присутствовать окно редактора данных со строкой меню и загруженным файлом `ex01.sav`.

ШАГ 4

В меню Анализ выберите команду Описательные статистики ► Частоты. На экране появится диалоговое окно, показанное на рис. 6.1.

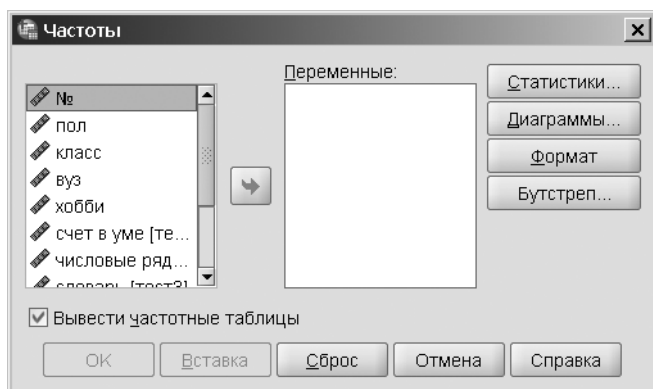


Рис. 6.1. Диалоговое окно Частоты

Частоты

Диалоговое окно Частоты типично для большинства статистических операций SPSS. В левой части окна расположен список всех доступных переменных. В нем необходимо выбрать те переменные, для которых надо вычислить распределение частот.

Для этого щелчком выделите нужную переменную, а затем щелчком на кнопке с направленной вправо стрелкой переместите ее в целевой список Переменные. Если желаемая переменная не видна в исходном списке, воспользуйтесь полосой прокрутки.

Для того чтобы удалить переменную из целевого списка, достаточно выделить ее, а затем воспользоваться кнопкой с направленной влево стрелкой, которая появляется на месте предыдущей кнопки при выделении любого пункта в списке Переменные. В этом случае переменная вновь переместится в исходный список. Чтобы полностью очистить список Переменные, щелкните на кнопке Сброс.

ШАГ 5

На этом шаге займемся вычислением частот для переменных пол, класс и вуз. При открытом диалоговом окне Частоты выполните следующие действия.

1. Щелкните сначала на переменной пол, чтобы выделить ее, а затем на кнопке со стрелкой, чтобы переместить переменную в список Переменные.
2. Повторите предыдущее действие для переменных класс и вуз.
3. Щелкните на кнопке ОК, чтобы открыть окно вывода, показанное на рис. 6.2.

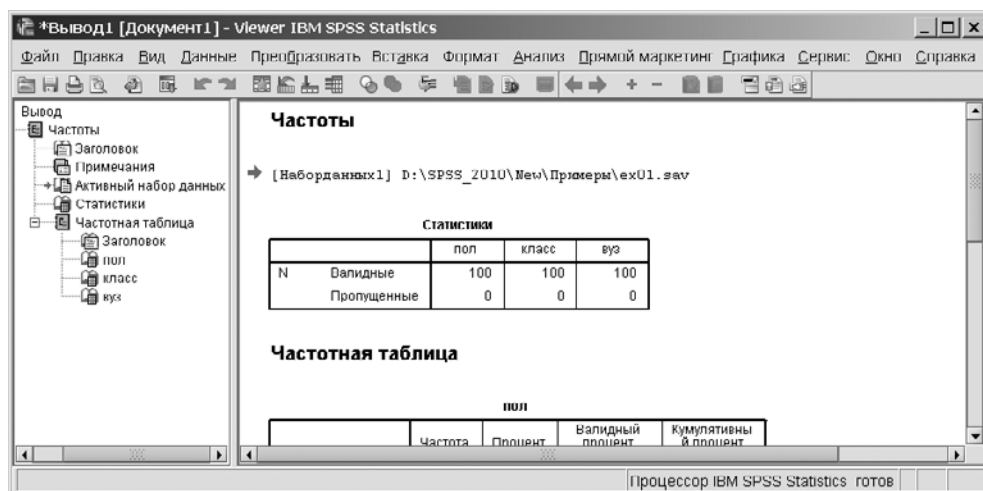


Рис. 6.2. Окно вывода программы SPSS

Переместив переменные в целевой список, вы завершаете выполнение операции щелчком на кнопке ОК. После этого программа SPSS формирует окно вывода с результатами выполнения команды. Выполнение практически всех команд статистических операций завершается именно так.

Окно с результатами работы программы имеет заголовок Вывод1 [Документ1]. Для просмотра результатов при необходимости можно воспользоваться полосой прокрутки в правой части окна. Структура выводимых результатов программы и основные приемы их редактирования описаны в разделе «Окно вывода и его редактирование»

главы 2. Поскольку в заголовке окна видна строка меню, можно снова применить команду Частоты к выбранному подмножеству данных: для этого достаточно выбрать команду Анализ ► Описательные статистики ► Частоты и в открывшемся диалоговом окне сразу щелкнуть на кнопке ОК (ранее выбранные переменные окажутся в целевом списке автоматически) или предварительно изменить список анализируемых переменных.

Столбиковые диаграммы

Для того чтобы создать столбиковую диаграмму для дискретных данных, необходимо сначала выполнить все действия шага 5, кроме последнего, а затем в диалоговом окне Частоты щелкнуть на кнопке Диаграммы. На экране появится диалоговое окно Частоты: Диаграммы, показанное на рис. 6.3.

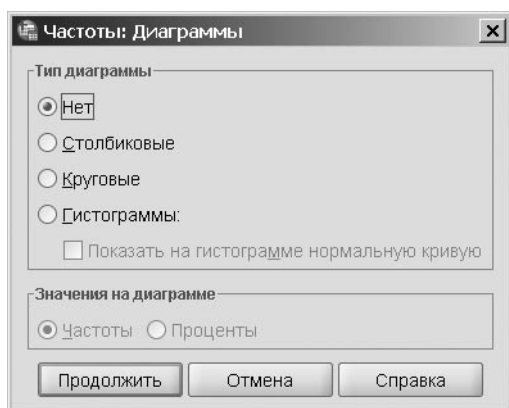


Рис. 6.3. Диалоговое окно Частоты: Диаграммы

Как видите, диалоговое окно Частоты: Диаграммы позволяет выбрать тип диаграммы с помощью переключателей. В данном примере вам понадобится установить переключатель Столбиковые. В зависимости от величины, которую вы хотите использовать для отображения частот, в группе Значения на диаграмме установите переключатель Частоты или Проценты. Щелкните на кнопке Продолжить. После этого вы вернетесь в диалоговое окно процедуры Частоты и сможете завершить операцию щелчком на кнопке ОК.

ШАГ 5А

На этом шаге мы выполним построение столбиковых диаграмм для частот тех же трех переменных пол, класс и вуз, которые использовались ранее. После выполнения шага 4 у вас должно быть открыто диалоговое окно Частоты, показанное на рис. 6.1. Если вы уже поработали с этим окном, очистите его нажатием кнопки Сброс. Для разнообразия будем заполнять целевой список двойными щелчками, а не щелчками на кнопке со стрелкой, как в предыдущем примере.

1. Дважды щелкните на переменной пол. Ее имя появится в списке Переменные. Сделайте то же самое для переменных класс и вуз и щелкните на кнопке Диаграммы, чтобы открыть диалоговое окно Частоты: Диаграммы.

2. В группе Тип диаграммы установите переключатель Столбиковые. Если вы предпочитаете вместо частот выводить на диаграмме их доли от общей частоты в процентах, в группе Значения на диаграмме установите переключатель Проценты и щелкните на кнопке Продолжить, чтобы вернуться в диалоговое окно Частоты.
3. Щелкните на кнопке ОК, чтобы открыть окно вывода.

После выполнения этого шага программа сгенерирует три диаграммы, соответствующие выбранным переменным. При помощи полосы прокрутки можно просмотреть созданные диаграммы в окне вывода. Если возникнет необходимость отредактировать какую-либо из диаграмм, дважды щелкните на ней и обратитесь к материалу главы 5 для дальнейших действий. На рис. 6.4 показано окно вывода с диаграммой частот для переменной вуз.

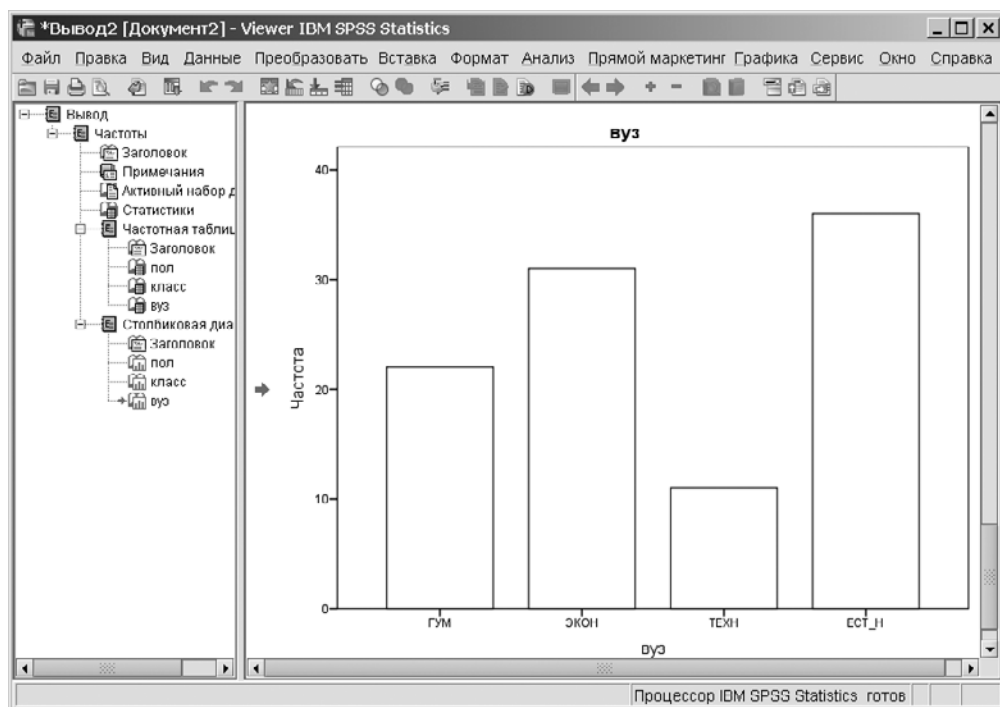


Рис. 6.4. Пример столбиковой диаграммы

Гистограммы

Процесс построения гистограмм аналогичен процессу построения столбиковых диаграмм. Единственным отличием, которое следует всегда иметь в виду, является тот факт, что гистограммы предназначены для отображения распределения непрерывных переменных. Таким образом, для переменных пол, класс и вуз гистограммы непригодны. В то же время для переменной отметка2, отражающей среднюю от-

метку для каждого учащегося, гистограмма является весьма удобным средством описания распределения частот. Ниже приведен пример построения гистограммы для переменной `отметка2`.

ШАГ 5Б

После выполнения шага 4 у вас должно быть открыто диалоговое окно Частоты, показанное на рис. 6.1. При необходимости повторите шаги 1–4. Если в целевом списке диалогового окна Частоты после работы над материалом предыдущего раздела остались переменные, щелкните на кнопке Сброс.

1. Щелкните сначала на переменной `отметка2`, чтобы выделить ее, затем на кнопке со стрелкой, чтобы переместить переменную в список Переменные, и, наконец, на кнопке Диаграммы, чтобы открыть диалоговое окно Частоты: Диаграммы.
2. В группе Тип диаграммы установите переключатель Гистограммы; если необходимо, установите флажок Показывать на гистограмме нормальную кривую и щелкните на кнопке Продолжить, чтобы вернуться в диалоговое окно Частоты.
3. В диалоговом окне Частоты сбросьте флажок Вывести частотные таблицы и щелкните на кнопке ОК.

Обратите особое внимание на необходимость сбросить флажок Вывести частотные таблицы. Этот флажок фактически отвечает за наличие в выводимых данных команды Частоты всей неграфической информации. Если оставить его установленным при использовании непрерывных переменных, размер таблицы частот будет определяться всем множеством различных значений переменной, что редко оказывается полезным для исследователя. Поэтому при анализе непрерывных данных флажок Вывести частотные таблицы лучше сбрасывать. Установленный флажок Показывать на гистограмме нормальную кривую указывает на то, что поверх гистограммы будет нанесена нормальная кривая.

Описательные статистики и процентиля

Если вернуться к диалоговому окну Частоты и щелкнуть в нем на кнопке Статистики, на экране появится диалоговое окно Частоты: Статистики, показанное на рис. 6.5. В этом окне представлены такие показатели, как процентиля и описательные статистики (подробно описательные статистики рассматриваются в главе 7). Из трех приведенных ниже примеров вы узнаете, каким образом создать гистограмму, получить описательные статистики и вычислить процентиля с заданным шагом или заданной величины. Во всех трех примерах используется переменная `отметка2`, а начинаются они с диалогового окна, полученного после выполнения шага 4 (см. рис. 6.1).

ШАГ 5В

На этом шаге мы построим гистограмму распределения частот и вычислим для него среднее значение, стандартное отклонение, асимметрию и эксцесс распределения.

1. Щелкните сначала на переменной `отметка2`, чтобы выделить ее, затем — на кнопке со стрелкой, чтобы переместить переменную в спи-

сок Переменные, и, наконец, — на кнопке Диаграммы, чтобы открыть диалоговое окно Частоты: Диаграммы.

2. В группе Тип диаграммы установите переключатель Гистограммы и щелкните на кнопке Продолжить, чтобы вернуться в диалоговое окно Частоты.
3. Щелкните на кнопке Статистики, чтобы открыть диалоговое окно Частоты: Статистики, показанное на рис. 6.5.
4. В группе Расположение установите флажок Среднее, в группе Распределение — флажки Асимметрия и Эксцесс, в группе Разброс — флажок Стандартное отклонение и щелкните на кнопке Продолжить, чтобы вернуться в диалоговое окно Частоты.
5. В диалоговом окне Частоты сбросьте флажок Вывести частотные таблицы и щелкните на кнопке ОК.

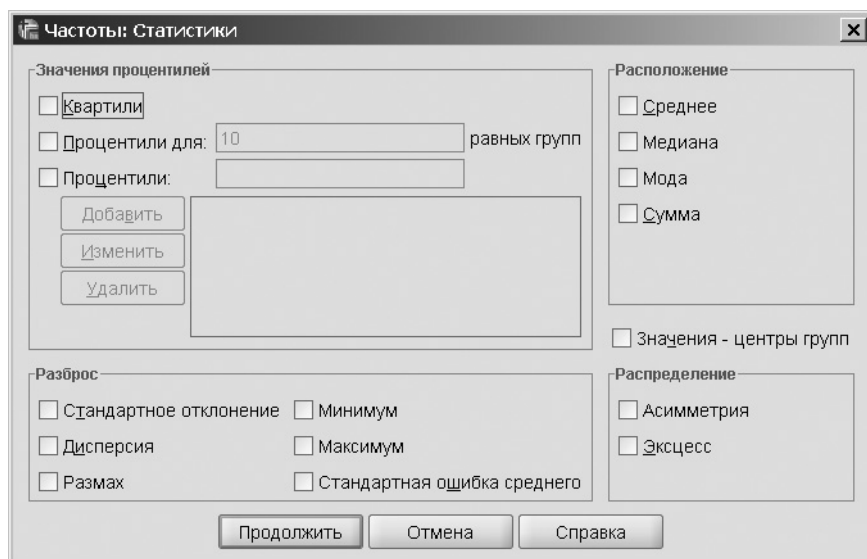


Рис. 6.5. Диалоговое окно Частоты: Статистики

ШАГ 5Г

Чтобы вычислить процентилю переменной `отметка2` с шагом 10, необходимо разделить шкалу процентилей на 10 равных промежутков с границами, равными 10, 20, 30 и т. д.

1. Щелкните сначала на переменной `отметка2`, чтобы выделить ее, затем — на кнопке со стрелкой, чтобы переместить переменную в список Переменные, и, наконец, — на кнопке Статистики, чтобы открыть диалоговое окно Частоты: Статистики, показанное на рис. 6.5.

2. В группе Значения процентилей установите флажок, в поле рядом с флажком Проценти́ли для: введите значение 10 и щелкните на кнопке Продолжить, чтобы вернуться в диалоговое окно Частоты.
3. В диалоговом окне Частоты сбросьте флажок Вывести частотные таблицы и щелкните на кнопке ОК.

ШАГ 5Д

Наконец, чтобы вычислить набор произвольных процентилей (в данном случае — 5, 70 и 95), выполните следующие действия.

1. Щелкните сначала на переменной отметка2, чтобы выделить ее, затем — на кнопке со стрелкой, чтобы переместить переменную в список Переменные, и, наконец, — на кнопке Статистики, чтобы открыть диалоговое окно Частоты: Статистики, показанное на рис. 6.5.
2. В группе Значения процентилей установите флажок Проценти́ли, в поле рядом с флажком введите значение 5 и щелкните на кнопке Добавить.
3. Введите в то же поле числа 70 и 95, по очереди добавляя их в список процентилей, и щелкните на кнопке Продолжить, чтобы вернуться в диалоговое окно Частоты.
4. В диалоговом окне Частоты сбросьте флажок Вывести частотные таблицы и щелкните на кнопке ОК.

Обратите внимание на то, что квартили (25-й, 50-й и 75-й проценти́ли) можно вычислить одним действием, установив флажок Кварти́ли в группе Значения проценти́лей диалогового окна Частоты: Статистики (см. рис. 6.5).

Представление результатов

Несмотря на то что выводимые программой SPSS результаты занимают не очень много места, как правило, их можно представить в еще более компактной форме (см. раздел «Окно вывода и его редактирование» в главе 2).

На возможные небольшие различия между данными, которые вы видите на экране своего компьютера, и иллюстрациями этой книги можно не обращать внимания.

Частоты

На рис. 6.6 приведен фрагмент окна вывода после выполнения шага 5.

Ниже дана трактовка терминов, используемых программой в окне вывода.

- ▶ Частота — число объектов, соответствующих каждой категории (градации) переменной.
- ▶ Процент — процент от общей численности (с учетом пропусков). Если бы в файле были пропущенные значения, их процент был бы указан в предпоследней строке.

- Валидный процент — процент значений для каждой категории за вычетом пропущенных значений.
- Кумулятивный процент — накопленный процент величины Валидный процент к данному значению переменной.
- Валидные — список значений переменной.
- Итого — сумма по столбцу.

пол

		Частота	Процент	Валидный процент	Кумулятивный процент
Валидные	ЖЕН	61	61,0	61,0	61,0
	МУЖ	39	39,0	39,0	100,0
	Итого	100	100,0	100,0	

класс

		Частота	Процент	Валидный процент	Кумулятивный процент
Валидные	А	33	33,0	33,0	33,0
	Б	35	35,0	35,0	68,0
	В	32	32,0	32,0	100,0
	Итого	100	100,0	100,0	

вуз

		Частота	Процент	Валидный процент	Кумулятивный процент
Валидные	ГУМ	22	22,0	22,0	22,0
	ЭКОН	31	31,0	31,0	53,0
	ТЕХН	11	11,0	11,0	64,0
	ЕСТ_Н	36	36,0	36,0	100,0
	Итого	100	100,0	100,0	

Рис. 6.6. Фрагмент окна вывода после выполнения шага 5

Гистограммы

На рис. 6.7 приведены результаты выполнения шага 5б. По горизонтальной оси отложены значения переменной с шагом, равным 0,125 (ширина каждого столбца тоже равна 0,125). Каждое указанное значение соответствует середине столбца. На вертикальной оси гистограммы отложены частоты диапазонов. Справа от гистограммы помещены вычисленные параметры: Среднее значение и Стандартное отклонение распределения, а также общее число объектов (N). Кроме того, на гистограмму нанесена нормальная кривая, поскольку при выполнении процедуры был установлен флажок Показывать на гистограмме нормальную кривую. Гистограмма без труда может быть сделана более наглядной. В частности, для данной гистограммы желательно укрупнение интервалов и, соответственно, уменьшение количества столбцов. По вопросам редактирования гистограмм обратитесь к главе 5 (см. рис. 5.12).

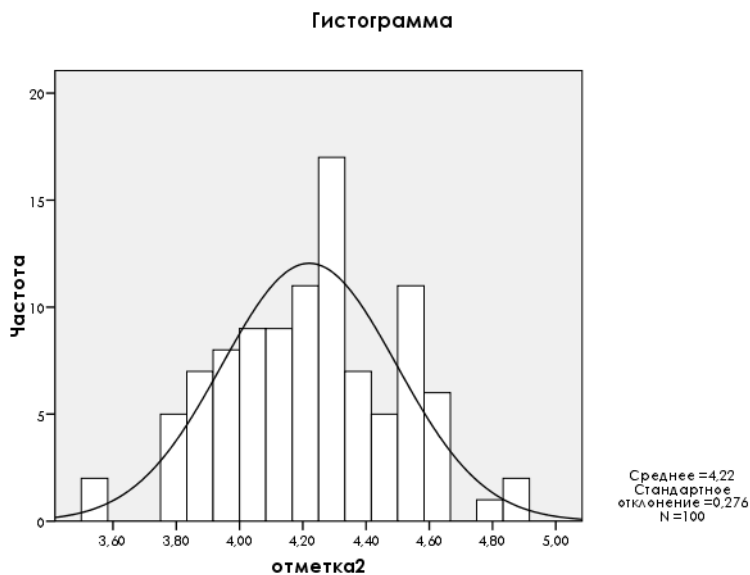


Рис. 6.7. Фрагмент окна вывода после выполнения шага 5б

Описательные статистики и процентили

На рис. 6.8 приведены результаты выполнения шагов 5г и 5д.

Статистики

отметка2		
N	Валидные	100
	Пропущенные	0
Среднее		4,2205
Стд. отклонение		,27589
Асимметрия		-,007
Стд. ошибка асимметрии		,241
Экссесс		-,357
Стд. ошибка эксцесса		,478

Статистики

отметка2		
N	Валидные	100
	Пропущенные	0
Процентили	5	3,8000
	70	4,3500
	95	4,6500

Рис. 6.8. Фрагменты окна вывода после выполнения шагов 5г и 5д

Описательные статистики рассматриваются подробнее в главе 7, поэтому мы не станем обсуждать их здесь. Обратите внимание: если в вывод результатов включены Асимметрия и Эксцесс, для них по умолчанию вычисляются стандартные ошибки (Стд. ошибка).

Что касается процентилей, то их можно трактовать следующим образом: для переменной отметка 2 5 % значений не превышают 3,8 (5 % учащихся имеют отметку не выше 3,8); 70 % значений не превышают 4,35 (30 % учащихся имеют отметку выше 4,35) и т. д.

Завершение анализа и выход из программы

Отредактируйте содержимое окна вывода в соответствии со своими предпочтениями: скройте лишнюю информацию, исправьте таблицы и пр. (см. раздел «Окно вывода и его редактирование» в главе 2). При необходимости отредактируйте диаграммы (см. главу 5).

Для дальнейшего использования окончательного результата все содержимое окна вывода или его фрагменты можно сохранить в файле *.spv, экспортировать в другой формат (например, Word), перенести в документ Word или вывести на печать (подробности см. в разделе «Сохранение, экспорт, перенос и печать результатов» главы 2).

Для выхода из программы выберите команду Выход в меню Файл.

7 Описательные статистики

119	Пошаговый алгоритм вычислений
122	Представление результатов
123	Завершение анализа и выход из программы

Описательные статистики (descriptive statistics) — это различные вычисляемые показатели, характеризующие распределение значений переменной. Эти показатели условно можно разбить на несколько групп. Первая группа — меры центральной тенденции, вокруг которых «группируются» данные: среднее значение, медиана и мода. Вторая группа характеризует изменчивость значений переменной относительно среднего: стандартное отклонение и дисперсия. Диапазон изменчивости характеризуется минимумом, максимумом и размахом. Асимметрия и эксцесс представляют меру отклонения формы распределения от нормального вида. Кроме того, существуют величины, выражающие погрешности некоторых статистик: стандартная ошибка среднего, стандартная ошибка асимметрии и стандартная ошибка эксцесса. Последние два показателя вычисляются программой вместе с асимметрией и эксцессом по умолчанию. При помощи команды *Описательные...* можно вычислить любую из указанных величин.

Меры центральной тенденции

Существует три основных меры центральной тенденции распределения.

- *Среднее значение* (mean) равно сумме всех значений распределения, деленной на их количество. Для распределения [3 5 7 5 6 8 9] среднее значение равно $(3 + 5 + 7 + 5 + 6 + 8 + 9)/7 = 6,14$.
- *Медиана* (median) определяется как значение, находящееся в середине распределения, полученного из исходного путем упорядочивания по возрастанию. Для распределения [3 5 7 5 6 8 9] медиана равна 6, поскольку значение, равное 6, находится в центре последовательности [3 5 5 6 7 8 9].
- *Мода* (mode) равна наиболее часто встречающемуся значению. В распределении [3 5 7 5 6 8 9] мода равна 5, поскольку число 5 встречается в нем дважды.

Меры изменчивости

Выделяют две величины, характеризующие изменчивость, или разброс, значений распределения относительно среднего.

- *Дисперсия* (variance) равна сумме квадратов отклонений каждого значения от среднего, деленной на $N - 1$, где N — число значений в распределении. Для

распределения [3 5 7 5 6 8 9] дисперсия равна $((3 - 6,14)^2 + (5 - 6,14)^2 + (7 - 6,14)^2 + (5 - 6,14)^2 + (6 - 6,14)^2 + (8 - 6,14)^2 + (9 - 6,14)^2)/6 = 4,1429$.

- *Стандартное отклонение* (standard deviation) равно квадратному корню из дисперсии. Для распределения [3 5 7 5 6 8 9] стандартное отклонение равно 2,0354.

Стандартное отклонение является довольно наглядной и информативной для исследователя характеристикой распределения, а дисперсия, как правило, используется как вспомогательная величина в статистических вычислениях.

Характеристики диапазона распределения

Дополнительными мерами изменчивости являются 4 простые характеристики, отражающие границы распределения и его размах.

- *Минимум* (minimum) равен наименьшему из значений распределения. Для распределения [3 5 7 5 6 8 9] минимум равен 3.
- *Максимум* (maximum) равен наибольшему из значений распределения. Для распределения [3 5 7 5 6 8 9] максимум равен 9.
- *Размах* (range) составляет разность между максимумом и минимумом распределения. В случае распределения [3 5 7 5 6 8 9] размах равен $9 - 3 = 6$.
- *Сумма* (sum) равна сумме всех значений распределения. Для распределения [3 5 7 5 6 8 9] сумма равна $3 + 5 + 7 + 5 + 6 + 8 + 9 = 43$.

Характеристики формы распределения

Для отражения близости формы распределения к нормальному виду существуют две основные характеристики.

- *Экссесс* (kurtosis) является мерой «сглаженности» («остро-» или «плосковершинности») распределения. Если значение эксцесса близко к 0, это означает, что форма распределения близка к нормальному виду. Положительный эксцесс указывает на «плосковершинное» распределение, у которого максимум вероятности выражен не столь ярко, как у нормального. Значения эксцесса, превышающие 5,0, говорят о том, что по краям распределения находится больше значений, чем вокруг среднего. Отрицательный эксцесс, напротив, характеризует «островершинное» распределение, график которого более вытянут по вертикальной оси, чем график нормального распределения. Считается, что распределение с эксцессом в диапазоне от -1 до $+1$ примерно соответствует нормальному виду. В большинстве случаев вполне допустимо считать нормальным распределение с эксцессом, по модулю не превосходящим 2.
- *Асимметрия* (skewness) показывает, в какую сторону относительно среднего сдвинуто большинство значений распределения. Нулевое значение асимметрии означает симметричность распределения относительно среднего значения, положительная асимметрия указывает на сдвиг распределения в сторону меньших значений, а отрицательная — в сторону больших значений. В боль-

шинстве случаев за нормальное принимается распределение с асимметрией, лежащей в пределах от -1 до $+1$. В исследованиях, не требующих высокой точности результатов, нормальным считают распределение с асимметрией, по модулю не превосходящей 2.

Стандартная ошибка

Стандартная ошибка (standard error) является характеристикой точности, или стабильности, величины, для которой она вычисляется. В контексте программы SPSS стандартная ошибка используется для среднего значения, асимметрии и эксцесса. Ее смысл заключается в следующем. Можно взять некоторое количество случайно выбранных значений генеральной совокупности, составить выборку и вычислить для нее среднее значение. Повторив эту операцию несколько раз, вы получите набор средних значений выборок, которые также представляют собой некоторое распределение. Стандартное отклонение этого распределения и будет являться стандартной ошибкой для среднего значения генеральной совокупности. Аналогичным способом вычисляются стандартные ошибки для асимметрии и эксцесса. Чем меньше значение стандартной ошибки, тем выше стабильность величины, для которой она вычисляется.

Пошаговый алгоритм вычислений

Для применения команды *Описательные...* мы обратимся к файлу *ex01.sav*. Число объектов, или значений каждой переменной, в этом файле равно 100, поэтому при вычислении характеристик распределения для различных переменных программа будет считать N равным 100. Сначала необходимо выполнить три подготовительных шага. Эти шаги (шаги 1–3) позволят подготовить рабочий файл данных, запустить программу IBM SPSS Statistics 19 и открыть файл (в данном случае — файл *ex01.sav*). Пошаговые инструкции этого процесса приведены в главе 4 (с. 60), а подробные разъяснения — в главе 2.

После завершения шага 3 на экране должно присутствовать окно редактора данных со строкой меню и загруженным файлом *ex01.sav*.

ШАГ 4

В меню *Анализ* выберите команду *Описательные статистики* ► *Описательные....* На экране появится диалоговое окно *Описательные статистики*, показанное на рис. 7.1.

В диалоговом окне *Описательные статистики* необходимо задать переменные, для которых будут вычислены описательные статистики. В левой части окна находится список всех доступных переменных текущего файла данных.

Для задания переменной щелчком выделите нужную переменную, а затем щелчком на кнопке с направленной вправо стрелкой переместите ее в целевой список *Переменные*. Если желаемая переменная не видна в исходном списке, воспользуйтесь полосой прокрутки.

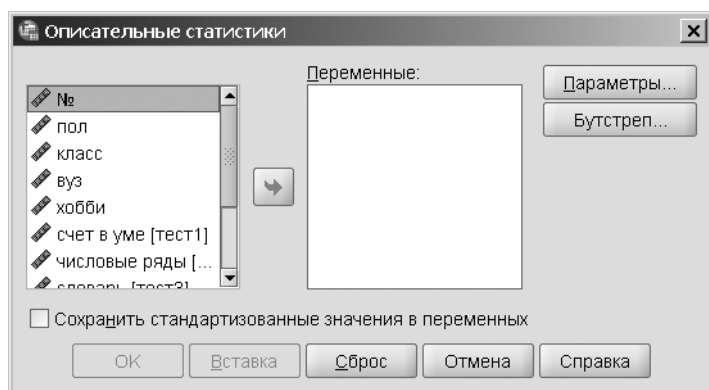


Рис. 7.1. Диалоговое окно Описательные статистики

Для того чтобы удалить переменную из целевого списка, достаточно выделить ее, а затем воспользоваться кнопкой с направленной влево стрелкой, которая появляется на месте предыдущей кнопки при выделении любого пункта в списке Переменные. В этом случае переменная вновь переместится в исходный список. Чтобы полностью очистить список Переменные, щелкните на кнопке Сброс.

Как вы могли заметить, помимо кнопок и списков диалоговое окно Описательные статистики содержит флажок Сохранить стандартизованные значения в переменных. Если этот флажок установлен, программа произведет z-преобразование (стандартизацию) всех выбранных переменных, создав таким образом новые переменные. Исходные переменные при этом останутся без изменений, а новым переменным будут присвоены старые имена, но начинающиеся с буквы z. Под стандартизованными (или z-преобразованными) значениями переменной понимается такое ее распределение, среднее значение которого равно 0, а стандартное отклонение — 1.

ШАГ 5

Для того чтобы создать таблицу описательных статистик, предлагаемую программой по умолчанию и включающую среднее значение, стандартное отклонение, максимум и минимум, выполните следующие действия.

1. Щелкните сначала на переменной отметка1, чтобы выделить ее, а затем на кнопке со стрелкой, чтобы переместить переменную в список Переменные.
2. Повторите те же действия для переменной отметка2.
3. Щелкните на кнопке ОК, чтобы открыть окно вывода.

Если вам понадобится вычислить дополнительные характеристики, не вычисляемые программой по умолчанию, перед щелчком на кнопке ОК нужно щелкнуть на кнопке Параметры, расположенной в нижнем правом углу диалогового окна Описательные статистики. Откроется диалоговое окно Описательные статистики: Параметры (рис. 7.2), в котором с помощью флажков можно задать все упоминавшиеся выше характеристики, за исключением двух: медианы и моды. Последние две характеристики доступны только через команды Частоты (глава 6) и Средние (глава 10). Кроме того, вы не увидите в диалоговом окне флажков, соответствующих стандартным ошибкам

асимметрии и эксцесса, поскольку они всегда вычисляются автоматически. Установите флажки, соответствующие нужным характеристикам, и щелкните на кнопке Продолжить. При желании можно также установить один из четырех переключателей в группе Порядок вывода. Переключатель Как в списке переменных установлен по умолчанию и означает, что в выводимых результатах переменные будут перечислены в том же порядке, в котором они представлены в файле данных. Переключатель Алфавитный указывает, что перечисление переменных будет происходить в алфавитном порядке. Переключатели По возрастанию средних и По убыванию средних позволяют отсортировать переменные в окне вывода по их вычисленным средним значениям.

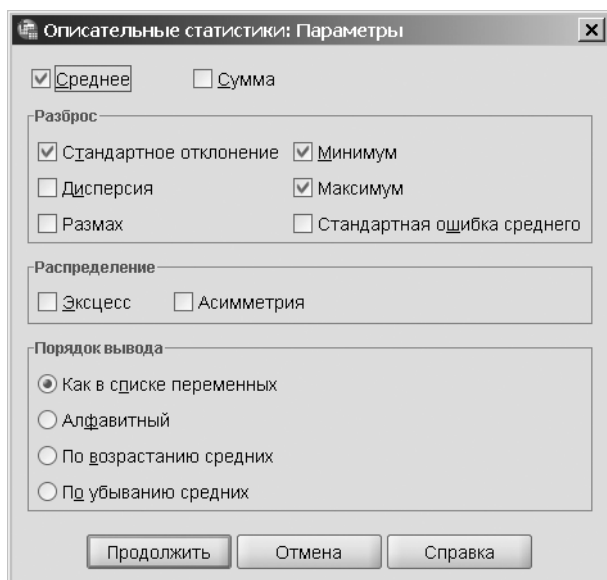


Рис. 7.2. Диалоговое окно Описательные статистики: Параметры

Далее показано, каким образом можно организовать подсчет дополнительных статистик для переменных отметка1 и отметка2.

ШАГ 5А

После выполнения шага 4 у вас должно быть открыто окно Описательные статистики, приведенное на рис. 7.1.

1. Щелкните сначала на переменной отметка1, чтобы выделить ее, а затем — на кнопке со стрелкой, чтобы переместить переменную в список Переменные.
2. Повторите те же действия для переменной отметка2 и щелкните на кнопке Параметры, чтобы открыть диалоговое окно Описательные статистики: Параметры.
3. Установите флажки для всех характеристик, которые следует вычислить, и щелкните на кнопке Продолжить, чтобы вернуться в диалоговое окно Описательные статистики.
4. Щелкните на кнопке ОК, чтобы открыть окно вывода, показанное на рис. 7.3.

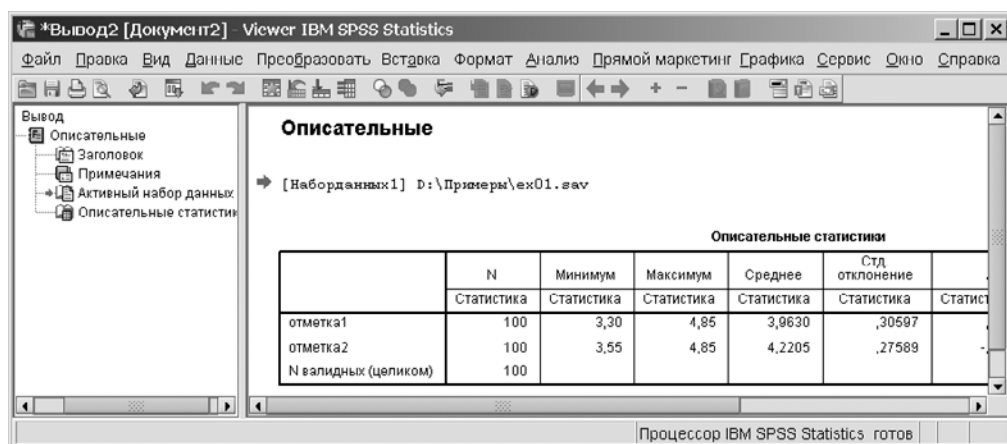


Рис. 7.3. Окно вывода SPSS

Для того чтобы полностью развернуть окно вывода, воспользуйтесь кнопкой разворачивания. Чтобы просматривать результаты внутри окна вывода, используйте вертикальную и горизонтальную полосы прокрутки. Как видите, в верхней части окна вывода имеется строка меню. Это означает, что можно выполнить и другие статистические процедуры, не обращаясь к окну редактора данных.

Представление результатов

Обратите внимание, что выводимые данные (рис. 7.3) не уместились на одной странице и целиком не видны на экране. Такая ситуация встречается достаточно часто и приходится пользоваться обеими полосами прокрутки, чтобы увидеть результаты анализа. Для быстрой навигации внутри окна вывода можно использовать иерархическую структуру в виде дерева объектов в левой части окна. Однако размер таблицы может быть таков, что при печати она не поместится на странице целиком.

Для решения этой проблемы рекомендуется воспользоваться средствами редактирования таблиц (см. раздел «Окно вывода и его редактирование» главы 2). Например, таблицу, приведенную на рис. 7.3, можно транспонировать (поменять местами строки и столбцы). Для этого необходимо войти в режим редактирования таблицы (двойным щелчком мыши по ней или через контекстное меню), а затем в появившемся меню Мобильная таблица выбрать пункт Транспонировать строки и столбцы. Иногда нужно изменить ширину того или иного столбца для лучшего отображения его содержимого. Чтобы сделанные изменения вступили в силу, следует просто выйти из режима редактирования таблиц.

На рис. 7.4 приведен один из возможных примеров улучшения вида окна вывода (см. рис. 7.3), после транспонирования таблицы в режиме ее редактирования. Как видим, таблица целиком поместилась на странице. В результат обработки включены: Число значений (N), Минимум, Максимум, Среднее значение, Стандартное от-

клонение, Асимметрия и Эксцесс. Последние две величины приведены вместе со своими стандартными ошибками.

		отметка1	отметка2	N валидных (целиком)
N	Статистика	100	100	100
Минимум	Статистика	3,30	3,55	
Максимум	Статистика	4,85	4,85	
Среднее	Статистика	3,9630	4,2205	
Стд. отклонение	Статистика	,30597	,27589	
Асимметрия	Статистика	,317	-,007	
	Стд. ошибка	,241	,241	
Эксцесс	Статистика	-,056	-,357	
	Стд. ошибка	,478	,478	

Рис. 7.4. Окно вывода после выполнения шага 5 и транспонирования таблицы

Приведенные результаты говорят о том, что в отношении рассматриваемых переменных доступны любые методы статистического анализа, так как значения асимметрии и эксцесса по модулю не превышают 1 для всех переменных. С терминологией мы уже познакомились ранее. Единственным исключением является слово «целиком» в последнем столбце таблицы. Оно указывает на то, что анализ проводился без учета тех объектов (строк), для которых пропущено хотя бы одно значение. Поскольку таких объектов в файле ex01.sav нет, число включенных в анализ объектов (100) равно числу исходных объектов.

Завершение анализа и выход из программы

Отредактируйте содержимое окна вывода в соответствии со своими предпочтениями: скройте лишнюю информацию, исправьте таблицы и пр. (см. раздел «Окно вывода и его редактирование» в главе 2).

Для дальнейшего использования окончательного результата все содержимое окна вывода или его фрагменты можно сохранить в файле *.spv, экспортировать в другой формат (например, Word), перенести в документ Word или вывести на печать (см. раздел «Сохранение, экспорт, перенос и печать результатов» в главе 2).

Для выхода из программы выберите команду Выход в меню Файл.

8 Таблицы сопряженности и критерий хи-квадрат

124	Таблицы сопряженности
125	Критерий независимости хи-квадрат
126	Пошаговый алгоритм вычислений
133	Представление результатов
136	Терминология, используемая при выводе
137	Завершение анализа и выход из программы

Таблицы сопряженности служат для описания связи двух или более номинальных (категориальных) переменных. Примерами номинальных переменных являются: пол (женский, мужской), класс (А, Б, В), местность (город, пригород, село), ответ (да, нет) и т. д. Таблицы сопряженности неприменимы к непрерывным переменным, однако последние можно разбить на интервалы. Так, возраст человека, который следует считать непрерывным из-за большого числа его возможных значений, можно разделить на интервалы от 0 до 19 лет, от 20 до 39 лет, от 40 до 59 лет и т. д. Иногда такое разделение непрерывной переменной и ее последующее представление с помощью таблиц сопряженности делает результаты более наглядными. С точки же зрения статистического анализа переход от непрерывных (количественных) переменных к номинальным нежелателен, так как при этом теряется существенная часть информации о различии объектов. Например, два человека в возрасте 39 и 40 лет попадают в соседние возрастные категории и с точки зрения анализа ничем не отличаются от пары людей в возрасте 20 и 59 лет.

Для работы с таблицами сопряженности в программе SPSS используется команда *Таблицы сопряженности*.

Таблицы сопряженности

Обратимся к файлу *ex01.sav*. С помощью команды *Частоты* мы можем узнать, что среди школьников 39 юношей и 61 девушка, что 33 из них увлекаются спортом, 37 — компьютером и 30 — искусством. Однако команда *Частоты* не позволяет ответить на вопрос, сколько девушек увлекается спортом или сколько юношей — искусством. Для этого в SPSS существует команда *Таблицы сопряженности*. Вполне логично, что для ответа на наш вопрос необходимо «сопратить», или «пересечь»,

подмножество учащихся определенного пола с подмножеством учащихся с определенным увлечением. Такое сопряжение удобно представить в виде таблицы, строки которой соответствуют полу, столбцы — увлечению. Тогда в ячейке, находящейся, например, на пересечении строки «мужской» и столбца «искусство», мы увидим количество (частоту) юношей, которые увлекаются искусством. Поскольку в файле данных существуют 2 градации пола и 3 градации внешкольных увлечений (хобби), то наша перекрестная таблица будет состоять из $2 \times 3 = 6$ ячеек. Можно составлять и более сложные таблицы сопряженности, включающие три и более переменных, но эта операция имеет смысл лишь для больших объемов данных, иначе частоты большинства ячеек будут малыми или нулевыми. Рассмотрим, что произойдет, например, если для данных файла ex01.sav создать таблицу сопряженности пол \times хобби \times класс \times вуз. Эта таблица будет содержать $2 \times 3 \times 3 \times 4 = 72$ ячейки, притом что число объектов составляет лишь 100. Большинство ячеек этой таблицы сопряженности будет иметь значения от 0 до 1–2, что делает ее малоинформативной и статистически недостоверной. При задании этих четырех номинальных переменных программа SPSS вместо «четырёхмерной» таблицы сопряженности построит 12 двухмерных таблиц размерностью 2×3 , «вложенных» в одну таблицу.

Критерий независимости хи-квадрат

Помимо частот (или *наблюдаемых* величин) SPSS может вычислять *ожидаемые* значения для каждой ячейки таблицы. Ожидаемое значение вычисляется в предположении, что две номинальные переменные независимы друг от друга. Рассмотрим простой пример. Пусть в комнате находится 100 человек, из которых 30 являются мужчинами, а 70 — женщинами. Если известно, что из этих 100 человек 10 увлекаются искусством, в случае если увлечение не зависит от пола, следует ожидать, что из 10 увлекающихся искусством 3 являются мужчинами, а 7 — женщинами. Сопоставляя эти ожидаемые частоты с наблюдаемыми частотами, мы можем судить о том, действительно ли два номинальных признака независимы. Чем больше расхождение наблюдаемых и ожидаемых частот, тем сильнее эти два признака связаны друг с другом. Целью применения критерия независимости χ^2 и является установление степени соответствия между наблюдаемыми и ожидаемыми значениями ячеек.

В основе критерия независимости лежит вычисление величины χ^2 , определяемой как сумма отношений суммы квадратов отклонений наблюдаемой величины f_0 от ожидаемой величины f_e к ожидаемой величине в каждой ячейке:

$$\chi^2 = \sum [(f_0 - f_e)^2 / f_e]$$

Как можно видеть из формулы, при больших отклонениях f_0 от f_e величина χ^2 также становится большой. Вместе с χ^2 вычисляется p -уровень значимости. При $p > 0,05$ считается, что различия между наблюдаемыми и ожидаемыми значениями незначительны. В противном случае предположение о независимости двух номинальных переменных отклоняется и делается вывод о том, что две классификации

(переменные) зависят друг от друга. Более подробные описания величин, связанных с критерием χ^2 , вы можете найти в разделе «Представление результатов» этой главы.

Иногда данные не соответствуют требованиям критерия χ^2 , например ожидаемая частота в некоторых ячейках таблицы сопряженности менее 5. Тогда можно воспользоваться одним из двух точных методов расчета значимости: методом статистических испытаний «Монте-Карло», или точными методами. Метод Точные, по определению, дает более точные результаты. Однако в некоторых случаях его расчет слишком трудоемок даже для компьютера, тогда используется метод Монте-Карло.

Зачастую величина χ^2 ошибочно воспринимается исследователями как величина *силы связи* между переменными. Однако это не так, поскольку χ^2 в значительной степени определяется числом переменных таблицы сопряженности и размером выборки. Таким образом, сравнение двух значений χ^2 , полученных при разных условиях, становится бессмысленным. По этой причине К. Пирсон предложил коэффициент «фи», получаемый извлечением квадратного корня из отношения χ^2 к размеру выборки N . Целью введения новой величины было получение наглядной интерпретации связи между переменными в виде коэффициента, лежащего в пределах от 0 до 1 и принимающего нулевое значение для независимых переменных и единичное значение для строго связанных переменных. Однако цель не была достигнута полностью: если одна из переменных таблицы сопряженности имеет более двух градаций, значение «фи» может превышать 1. Крамеру удалось исправить последний недостаток путем введения коэффициента $V = \sqrt{\chi^2 / [N(k-1)]}$, где k — наименьшее из числа градаций двух переменных. Этот коэффициент всегда принимает значения от 0 до 1 и служит характеристикой силы связи между переменными.

Пошаговый алгоритм вычислений

Для применения команды Таблицы сопряженности мы снова обратимся к файлу ex01.sav. Число объектов, или наборов значений, в этом файле равно 100, поэтому программа будет считать N равным 100. Мы создадим несколько таблиц сопряженности, а также применим критерий χ^2 для определения зависимости переменных пол и хобби, но сначала необходимо выполнить три подготовительных шага. Эти шаги (шаги 1–3) позволят подготовить рабочий файл данных, запустить программу IBM SPSS Statistics 19 и открыть файл (в данном случае — файл ex01.sav) Пошаговые инструкции этого процесса приведены в главе 4 (с. 60), а подробные разъяснения — в главе 2.

После завершения шага 3 на экране должно присутствовать окно редактора данных со строкой меню и загруженным файлом ex01.sav.

ШАГ 4

В меню Анализ выберите команду Описательные статистики ► Таблицы сопряженности. На экране появится диалоговое окно Таблицы сопряженности, показанное на рис. 8.1.

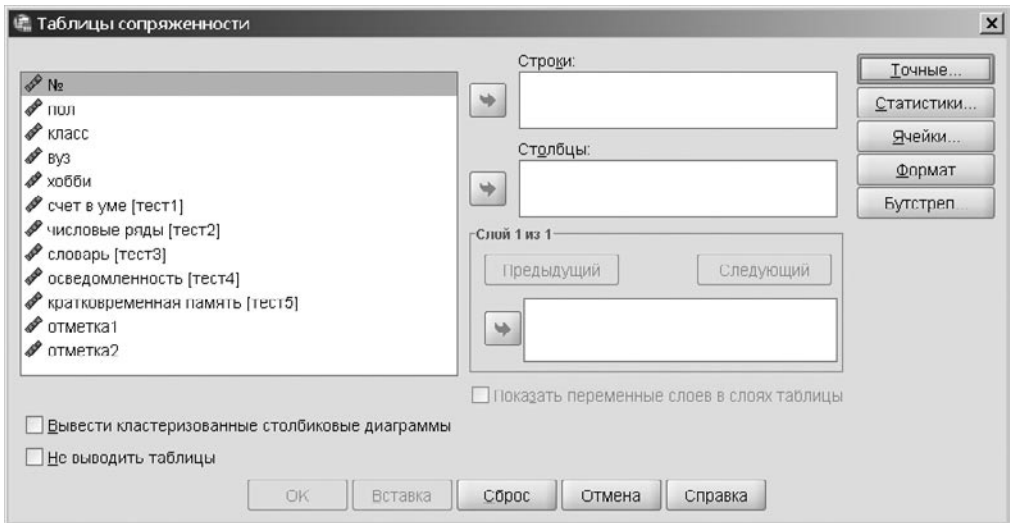


Рис. 8.1. Диалоговое окно Таблицы сопряженности

С помощью диалогового окна Таблицы сопряженности можно задать конфигурацию будущей таблицы сопряженности и столбиковую диаграмму для ее иллюстрации. Для этого требуется выбрать переменные в исходном списке и переместить их в целевые списки Строки и Столбцы, используя верхнюю и среднюю кнопки со стрелкой. Это позволит определить, значения какой переменной будут составлять строки таблицы, а какой — столбцы. Поместив в список Строки переменную пол, в список Столбцы — переменную хобби и щелкнув затем на кнопке ОК, вы создадите таблицу с двумя строками и тремя столбцами. Для создания столбиковой диаграммы следует установить флажок Вывести кластеризованные столбиковые диаграммы.

Раздел Слой 1 из 1 диалогового окна позволяет построить таблицу сопряженности для трех и более переменных. Если бы мы захотели подключить к таблице, содержащей переменные пол и хобби, еще и переменную класс, последнюю следовало бы поместить в нижний список при помощи нижней кнопки со стрелкой. В результате вы увидели бы на экране таблицу, разбитую на три части, каждая из которых представляла бы собой таблицу сопряженности пол × хобби для соответствующего значения переменной класс (1, 2 и 3). Кнопки Предыдущий и Следующий слева и справа от метки Слой 1 из 1 предназначены для случаев, когда необходимо построить таблицы сопряженности с более чем одной дополнительной переменной. Например, если бы вам понадобилось составить таблицы сопряженности с переменными пол, хобби, класс и вуз, в нижнем списке нужно было бы указать две переменные класс и вуз, а в итоге программой была бы сгенерирована общая таблица сопряженности, включающая в себя 3 таблицы пол × хобби для каждой градации переменной класс и 4 таблицы пол × хобби для каждого значения переменной вуз.

В следующем примере мы создадим таблицу сопряженности пол × хобби × класс и проиллюстрируем ее столбиковыми диаграммами.

ШАГ 5

При открытом диалоговом окне Таблицы сопряженности, представленном на рис. 8.1, выполните следующие действия.

1. Щелкните сначала на переменной пол, чтобы выделить ее, а затем — на верхней кнопке со стрелкой, чтобы переместить переменную в список Строки.
2. Щелкните сначала на переменной хобби, чтобы выделить ее, а затем — на средней кнопке со стрелкой, чтобы переместить переменную в список Столбцы.
3. Щелкните сначала на переменной класс, чтобы выделить ее, а затем на нижней кнопке со стрелкой, чтобы переместить переменную в список дополнительных переменных под меткой Слой 1 из 1.
4. Установите флажок Вывести кластеризованные столбиковые диаграммы.
5. Щелкните на кнопке ОК, чтобы открыть окно вывода.

Как правило, исследователи редко ограничиваются вычислением частот, предпочитая дополнять процедуру заданием разнообразных параметров. При щелчке на кнопке Ячейки открывается диалоговое окно Таблицы сопряженности: Вывод в ячейках, представленное на рис. 8.2. Это окно позволяет управлять информацией в ячейках. По умолчанию флажок Наблюдаемые в группе Частоты установлен, поскольку наблюдаемые частоты являются главной вычисляемой величиной в любой таблице сопряженности. Установка флажка Ожидаемые позволит вычислять в ячейках и ожидаемые частоты. Остальные параметры задаются в зависимости от целей и предпочтений конкретного исследователя.

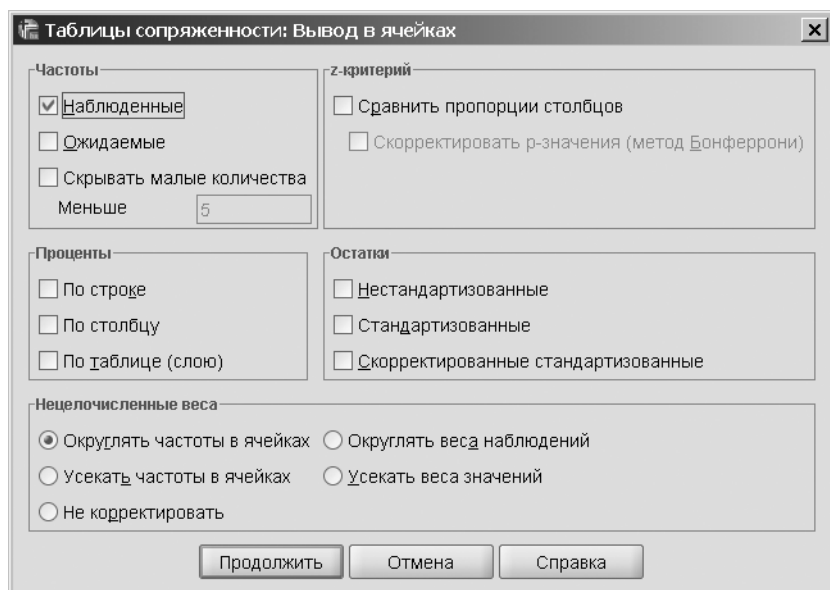


Рис. 8.2. Диалоговое окно Таблицы сопряженности: Вывод в ячейках

Ниже описано назначение некоторых флажков, управляющих содержимым ячеек в диалоговом окне Таблицы сопряженности: Вывод в ячейках.

- ▶ При установке флажка Наблюдаемые в группе Частоты отображается реальное количество объектов, соответствующее каждой ячейке.
- ▶ При установке флажка Ожидаемые в группе Частоты отображается значение ожидаемой частоты для каждой ячейки, вычисленное в предположении, что соответствующие переменные являются независимыми.
- ▶ При установке флажка По строке в группе Проценты отображается частота ячейки в процентах от суммарной частоты строки, в которой она находится.
- ▶ При установке флажка По столбцу в группе Проценты отображается частота ячейки в процентах от суммарной частоты столбца, в котором она находится.
- ▶ При установке флажка По таблице (слою) в группе Проценты отображается частота ячейки в процентах от суммарной частоты всей таблицы сопряженности.
- ▶ При установке флажка Нестандартизованные в группе Остатки отображается разность между наблюдаемой и ожидаемой частотами.
- ▶ При установке флажка Сравнить пропорции столбцов в таблице сопряженности появляется индикация столбцов, которые статистически достоверно различаются или не различаются. Дополнительный флажок Скорректировать р-значение (метод Бонферрони) вводит поправку для р-уровня, если число сравниваемых столбцов более 2.

В следующем примере мы создадим таблицу сопряженности пол \times хобби, в ячейки которой включим наблюдаемые и ожидаемые частоты, процент по строке и не-стандартизированный остаток. Также добавим индикацию статистической значимости различия столбцов.

ШАГ 5А

После выполнения шага 4 у вас должно быть открыто диалоговое окно Таблицы сопряженности, представленное на рис. 8.1. Если вы уже успели поработать с этим окном, очистите его щелчком на кнопке Сброс и выполните следующие действия.

1. Щелкните сначала на переменной пол, чтобы выделить ее, а затем — на верхней кнопке со стрелкой, чтобы переместить переменную в список Строки.
2. Щелкните сначала на переменной хобби, чтобы выделить ее, а затем — на средней кнопке со стрелкой, чтобы переместить переменную в список Столбцы.
3. Щелкните на кнопке Ячейки, чтобы открыть диалоговое окно Таблицы сопряженности: Вывод в ячейках, и установите флажки: в разделе Частоты: Ожидаемые, в разделе Проценты: По строке, в разделе Остатки: Нестандартизованные. В разделе z-критерий установите флажки Сравнить пропорции столбцов в таблице сопряженности и Скорректировать р-значение (метод Бонферрони).

Карло, или собственно точным критерием. Метод Точные, по определению, дает более точные результаты. Однако в некоторых случаях его расчет слишком трудоемок даже для компьютера, тогда используется метод Монте-Карло.

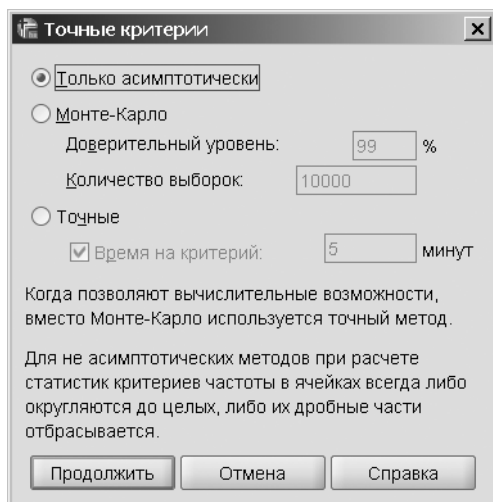


Рис. 8.4. Диалоговое окно Точные критерии

В следующем примере мы создадим таблицу сопряженности пол × хобби, в ячейки которой включим наблюдаемые частоты, проценты по строке, индикацию статистической значимости столбцов, применим критерий χ^2 и точный критерий, вычислим коэффициенты Фи и V Крамера. Анализ корреляций между переменными в данном случае не имеет смысла, так как переменные не являются количественными.

ШАГ 5Б

После выполнения шага 4 у вас должно быть открыто диалоговое окно Таблицы сопряженности, представленное на рис. 8.1. Если вы уже успели поработать с этим окном, очистите его щелчком на кнопке Сброс и выполните следующие действия.

1. Щелкните сначала на переменной пол, чтобы выделить ее, а затем — на верхней кнопке со стрелкой, чтобы переместить переменную в список Строки.
2. Щелкните сначала на переменной хобби, чтобы выделить ее, а затем — на средней кнопке со стрелкой, чтобы переместить переменную в список Столбцы.
3. Щелкните на кнопке Ячейки, чтобы открыть диалоговое окно Таблицы сопряженности: Ячейки, и установите флажки Проценты: По строке, Сравнить пропорции столбцов, Скорректировать р-значения (метод Бонферрони).
4. Щелкните на кнопке Продолжить, чтобы вернуться в диалоговое окно Таблицы сопряженности.
5. Щелкните на кнопке Статистики, чтобы открыть диалоговое окно Таблицы сопряженности: Статистики, и установите флажки Хи-квадрат и Фи и V Крамера.

6. Щелкните на кнопке Продолжить, чтобы вернуться в диалоговое окно Таблицы сопряженности.
7. Щелкните по кнопке Точные, чтобы открыть окно Точные критерии, и установите переключатель Точные. Щелкните на кнопке Продолжить, чтобы вернуться в диалоговое окно Таблицы сопряженности.
8. Установите флажок Вывести кластеризованные столбиковые диаграммы.
9. Щелкните на кнопке ОК, чтобы открыть окно вывода.

Зачастую приходится строить таблицы сопряженности и применять критерий χ^2 не для всей выборки объектов, а только для их части. Например, может возникнуть необходимость в создании таблицы пол \times хобби только для двух из трех градаций переменной хобби: спорт и искусство. Это делается следующим образом. После выбора переменных для построения таблицы сопряженности, определения набора величин, вычисляемых для каждой ячейки, и задания желаемых статистик в меню Данные выберите команду Отобрать наблюдения. На экране появится одноименное диалоговое окно (см. раздел «Выбор наблюдений для анализа» в главе 4), в котором в группе Выбрать следует установить переключатель Если выполнено условие, а затем щелкнуть на расположенной рядом кнопке Если. Откроется диалоговое окно Отобрать наблюдения: Условие. Элементы интерфейса этого окна позволяют выполнять операции, о которых шла речь в главе 4. В данном случае нас будет интересовать выбор уровней 1 и 3 переменной хобби. Для исключения уровня 2 из анализа выделите переменную хобби в исходном списке, введите ее в условие отбора, щелкнув на кнопке с направленной вправо стрелкой, на панели калькулятора щелкните на кнопке \neq (не равно) и введите в условие отбора число 2. После этого щелкните сначала на кнопке Продолжить, чтобы вернуться в диалоговое окно Отобрать наблюдения, а затем — на кнопке ОК, чтобы вернуться в диалоговое окно Таблицы сопряженности. Наконец, щелчок на кнопке ОК в окне Таблицы сопряженности приведет к построению таблицы сопряженности с двумя выбранными градациями переменной хобби и двумя градациями переменной пол.

Для того чтобы в дальнейшем анализе участвовали все наблюдения, не забудьте в диалоговом окне Отобрать наблюдения вернуть переключатель в положение Все наблюдения.

Реализуем описанный алгоритм в виде пошаговых инструкций.

ШАГ 5B

После выполнения шага 4 у вас должно быть открыто диалоговое окно Таблицы сопряженности, представленное на рис. 8.1. Если вы уже успели поработать с этим окном, очистите его щелчком на кнопке Сброс и выполните следующие действия.

1. Щелкните сначала на переменной пол, чтобы выделить ее, а затем — на верхней кнопке со стрелкой, чтобы переместить переменную в список Строки.
2. Щелкните на переменной хобби, чтобы выделить ее, а затем — на средней кнопке со стрелкой, чтобы переместить переменную в список Столбцы.

3. Щелкните на кнопке Ячейки, чтобы открыть диалоговое окно Таблицы сопряженности: Ячейки, и установите флажки По строке и Сравнить пропорции столбцов.
4. Щелкните на кнопке Продолжить, чтобы вернуться в диалоговое окно Таблицы сопряженности.
5. Щелкните на кнопке Статистики, чтобы открыть диалоговое окно Таблицы сопряженности: Статистики, и установите флажки Хи-квадрат и Фи и V Крамера.
6. Щелкните на кнопке Продолжить, чтобы вернуться в диалоговое окно Таблицы сопряженности.
7. В меню Данные выберите команду Отобразить наблюдения, чтобы открыть одноименное диалоговое окно, установите переключатель Если выполнено условие и щелкните на расположенной рядом кнопке Если, открыв этим диалоговое окно Отобразить наблюдения: Условие.
8. В списке доступных переменных щелкните сначала на переменной хобби, чтобы выделить ее, а затем — на кнопке с направленной вправо стрелкой, чтобы ввести переменную в условие отбора. Дополните условие отбора оператором $\sim =$ (не равно) и числом 2, используя панель калькулятора и клавиатуру, и щелкните на кнопке Продолжить, чтобы вернуться в диалоговое окно Отобразить наблюдения.
9. Щелкните на кнопке ОК, чтобы вернуться в диалоговое окно Таблицы сопряженности.
10. Щелкните на кнопке ОК, чтобы открыть окно вывода.
11. В меню Данные выберите команду Отобразить наблюдения, чтобы открыть одноименное диалоговое окно, установите переключатель Все наблюдения и щелкните на кнопке ОК.

После выполнения шагов 5, 5а, 5б или 5в программа автоматически активизирует окно вывода. Для просмотра результатов при необходимости можно воспользоваться вертикальной и горизонтальной полосами прокрутки. Обратите внимание на стандартную строку меню в верхней части окна вывода: ее присутствие позволяет выполнять любые статистические операции, не переключаясь обратно в окно редактора данных.

Представление результатов

На рис. 8.5 и 8.6 приведены фрагменты выводимых результатов, сгенерированные программой после выполнения шага 5б.

Первое, на что необходимо обратить внимание при интерпретации выводимых результатов, — это соотношение частот (процентов) в строках таблицы сопряженности. Беглый взгляд на первую таблицу говорит о том, что эти величины заметно различаются. Подстрочные буквы указывают, какие ячейки в строке различаются статистически значимо, а какие — нет. Буква *a* для первых двух ячеек свиде-

тельствует о том, что они не различаются статистически достоверно ($p > 0,05$), а буква b в третьей ячейке указывает, что частота в ней отличается статистически достоверно от частоты как в первой, так и во второй ячейках ($p \leq 0,05$). В соответствии с установленным флажком Скорректировать p -значения (метод Бонферрони), введена поправка, необходимая при сравнении более двух ячеек. Различие частот в ячейках свидетельствует о том, что переменные пол и хобби связаны друг с другом. Это подтверждается содержанием второй таблицы: относительно большим значением критерия χ^2 (15,405) и малым значением p -уровня ($p < 0,001$), то есть высокой статистической значимостью. Столбчатая диаграмма демонстрирует, что связь пол и хобби обусловлена, прежде всего, различием юношей и девушек в выборе искусства. Отметим, что результат применения точного критерия Точная значимость (2-сторон.), идентичен результату применения традиционного критерия χ^2 (Асимпт. значимость (2-сторон.)), в связи с тем, что не нарушены условия применения критерия χ^2 : нет ячеек с частотой менее 5. Если бы условия применения этого критерия были нарушены (наличие более 25 % ячеек с частотой менее 5), следовало бы ориентироваться на точную значимость (2-сторон.).

О величине связи переменных можно судить по симметричным мерам — значениям показателей Фи и V Крамера, которые аналогичны коэффициенту корреляции (рис. 8.5). Величина 0,392 свидетельствует об умеренной связи между переменными.

Таблица сопряженности пол * хобби

			хобби			Итого
			спорт	компьютер	искусство	
пол	ЖЕН	Частота	15 _a	19 _a	27 _b	61
		% в пол	24,6%	31,1%	44,3%	100,0%
	МУЖ	Частота	18 _a	18 _a	3 _b	39
		% в пол	46,2%	46,2%	7,7%	100,0%
Итого	Частота	33	37	30	100	
	% в пол	33,0%	37,0%	30,0%	100,0%	

Каждая подстрочная буква обозначает набор хобби категорий, для которых пропорции столбцов значимо не различаются между собой на уровне ,05.

Критерии хи-квадрат

	Значение	ст.св.	Асимпт. значимость (2-стор.)	Точная значимость (2-стор.)	Точная значимость (1-стор.)	Вероятность в точке
Хи-квадрат Пирсона	15,405 ^a	2	,000	,000		
Отношение правдоподобия	17,504	2	,000	,000		
Точный критерий Фишера	16,702			,000		
Линейно-линейная связь	12,652 ^b	1	,000	,000	,000	,000
Кол-во валидных наблюдений	100					

a. В 0 (,0%) ячейках ожидаемая частота меньше 5. Минимальная ожидаемая частота равна 11,70.
b. Стандартизованная статистика равна -3,557.

Рис. 8.5. Фрагменты окна вывода после выполнения шага 5б: таблицы

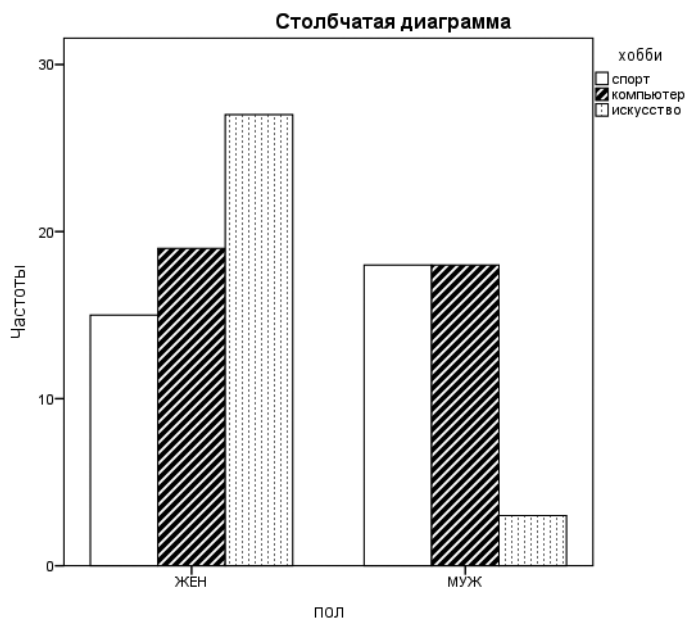
Симметричные меры

		Значение	Прибл. значимость	Точная знч.
Номинальная по номинальной	Фи	,392	,000	,000
	V Крамера	,392	,000	,000
Кол-во валидных наблюдений		100		

Рис. 8.5. Фрагменты окна вывода после выполнения шага 5б: таблицы (продолжение)

Для таблиц 2×2 при небольшом числе объектов (до 40–50) следует применять критерий χ^2 с поправкой на непрерывность. А при нарушении условий его применения (малая частота в ячейках) — Точный критерий Фишера (2-стор.).

Для таблиц сопряженности большой размерности следует помнить о проблеме малых значений частот. Малым считается значение ожидаемой частоты, меньшее 5. Предполагается, что если более 25 % ячеек таблицы сопряженности имеют малые ожидаемые значения частот, вероятность ошибки χ^2 -критерия очень высока. В этом случае следует пользоваться точным критерием.

**Рис. 8.6.** Фрагмент окна вывода после выполнения шага 5б: диаграмма.

Терминология, используемая при выводе

Трактовка терминов, используемых программой в окне вывода, дана ниже.

Хи-квадрат Пирсона и Поправка на непрерывность — два варианта критерия χ^2 . Вторым вариантом применяется для таблиц сопряженности размерностью 2×2 . При больших значениях N эти коэффициенты практически равны. Формула для хи-квадрата Пирсона выглядит следующим образом:

$$\chi^2 = \sum [(f_0 - fe)^2 / fe].$$

Точный критерий Фишера — результаты расчета точной величины значимости по методу Фишера. Используется в тех случаях, когда применение критерия χ^2 к таблице 2×2 не является корректным, например, при минимальной ожидаемой частоте в некоторых ячейках менее 5.

- Значение — для критерия χ^2 значение тем больше, чем больше зависимость между переменными (как в нашем примере). Значения близкие к 0 свидетельствуют о независимости переменных

Ст. св. — степени свободы, произведение количеств градаций переменных, уменьшенных на 1. Для данной таблицы сопряженности число степеней свободы равно $(2 - 1) \times (3 - 1) = 2$.

- Асимпт. значимость — асимптотическая значимость, вероятность случайности связи или p -уровень значимости, то есть вероятность того, что связь является случайной. Чем меньше эта величина, тем выше статистическая значимость (достоверность) связи. Величина $p \leq 0,05$ свидетельствует о статистически значимом результате, который достоин содержательной интерпретации. *Асимптотическая* значимость определяется по традиционному критерию χ^2 .
- Точная значимость — p -уровень значимости, вычисляемый точным методом; принимается во внимание, когда условия применения традиционного критерия нарушены (более 25 % ячеек таблицы сопряженности имеют частоту менее 5).

Точный критерий Фишера — вариант точного метода, который применяется для таблиц сопряженности размерностью 2×2 .

- Линейно-линейная связь — статистический критерий, определяющий степень корреляции между переменными. Чаще всего результаты этого теста являются бессмысленными, поскольку порядок уровней переменных не имеет логической интерпретации. К примеру, нет смысла говорить об упорядоченной структуре хобби. Тем не менее, если бы одной из переменных был уровень дохода, упорядоченный по возрастанию, количественная оценка корреляции несла бы вполне определенную смысловую нагрузку.

- Номинальная по номинальной — меры связи для двух номинальных переменных.
- Фи — коэффициент, являющийся мерой связи двух переменных, аналог корреляции Пирсона. Значение $\phi = 0,392$ показывает умеренную связь между двумя переменными.

V Крамера — как и коэффициент Фи, этот коэффициент является мерой связи между двумя переменными, однако отличается тем, что всегда принимает значения от 0 до 1 и более приемлем для таблиц с $df > 2$.

Завершение анализа и выход из программы

Отредактируйте содержимое окна вывода в соответствии со своими предпочтениями: скройте лишнюю информацию, исправьте таблицы и пр. (см. раздел «Окно вывода и его редактирование» в главе 2). При необходимости отредактируйте диаграммы (см. главу 5).

Для дальнейшего использования окончательного результата все содержимое окна вывода или его фрагменты можно сохранить в файле *.spv, экспортировать в другой формат (например, Word), перенести в документ Word или вывести на печать (подробности см. в разделе «Сохранение, экспорт, перенос и печать результатов» главы 2).

Для выхода из программы выберите команду Выход в меню Файл.

9 Корреляции

138	Понятие корреляции
140	Дополнительные сведения
142	Пошаговые алгоритмы вычислений
146	Представление результатов
147	Завершение анализа и выход из программы

Для вычисления корреляций между данными в программе SPSS используются команды подменю Корреляции меню Анализ. Корреляция представляет собой величину, заключенную в пределах от -1 до $+1$, и обозначается буквой r . Понятия *корреляция* и *двумерная корреляция* часто употребляются как синонимы; последнее означает «корреляция между двумя переменными» и подчеркивает, что рассматривается именно двумерное соотношение. Основной коэффициент корреляции r Пирсона предназначен для оценки связи между двумя переменными, измеренными в метрической шкале, распределение которых соответствует нормальному. Несмотря на то что величина r рассчитывается в предположении, что значения обеих переменных распределены по нормальному закону, формула для ее вычисления дает достаточно точные результаты и в случаях аномальных распределений, а также в случаях, когда одна из переменных является дискретной. Для распределений, не являющихся нормальными, предпочтительнее пользоваться ранговыми коэффициентами корреляции Спирмена или Кендалла. Команды подменю Корреляции позволяют вычислить как коэффициент Пирсона, так и коэффициенты Спирмена и Кендалла. Существуют и другие коэффициенты корреляции, применяющиеся для самых разных типов данных, однако их описание выходит за рамки темы этой книги.

Понятие корреляции

Корреляция, или *коэффициент корреляции*, — это статистический показатель вероятностной связи между двумя переменными, измеренными в количественной шкале. В отличие от функциональной связи, при которой каждому значению одной переменной соответствует строго определенное значение другой переменной, вероятностная связь характеризуется тем, что каждому значению одной переменной соответствует множество значений другой переменной. Примером вероятностной связи является связь между ростом и весом людей. Ясно, что один и тот же рост может быть у людей разного веса, как и наоборот. Величина коэффициента корреляции меняется от -1 до 1 . Крайние значения соответствуют линейной функциональной связи между двумя переменными, 0 — отсутствию связи.

Наглядное представление о связи двух переменных дает график двумерного рассеяния — соответствующая команда Рассеяние/точки имеется в меню Графика. На таком графике каждый объект представляет собой точку, координаты которой заданы значениями двух переменных. Таким образом, множество объектов представляет собой на графике множество точек. По конфигурации этого множества точек можно судить о характере связи между двумя переменными.

Строгая положительная корреляция определяется значением $r = 1$. Термин «строгая» означает, что значения одной переменной однозначно определяются значениями другой переменной, а термин «положительная» — что с возрастанием значений одной переменной значения другой переменной также возрастают.

Строгая корреляция является математической абстракцией и практически не встречается в реальных исследованиях. Примером строгой корреляции является соответствие между временем пути и пройденным расстоянием при неизменной скорости.

Положительная корреляция соответствует значениям $0 < r < 1$. Положительную корреляцию следует интерпретировать следующим образом: если значения одной переменной возрастают, то значения другой имеют *тенденцию* к возрастанию. Чем коэффициент корреляции ближе к 1, тем сильнее эта тенденция, и, наоборот, с приближением коэффициента корреляции к 0 тенденция ослабевает.

Примером значительной положительной корреляции служит зависимость между ростом и весом человека. Считается, что в этом случае коэффициент корреляции равен $r = 0,83$. Слабая положительная корреляция ($r = 0,12$) наблюдается между способностью человека к сочувствию и реальной помощью, которую он оказывает нуждающимся людям.

Отсутствие корреляции определяется значением $r = 0$. Нулевой коэффициент корреляции говорит о том, что значения переменных никак не связаны друг с другом. Примером пары величин с нулевой корреляцией является рост человека и результат его IQ-теста.

Отрицательная корреляция соответствует значениям $-1 < r < 0$. Если значения одной переменной возрастают, то значения другой имеют *тенденцию* к убыванию. Чем коэффициент корреляции ближе к -1 , тем сильнее эта тенденция, и, наоборот, с приближением коэффициента корреляции к 0 тенденция ослабевает.

Слабая отрицательная корреляция ($r = -0,13$) наблюдается между агрессивностью человека по отношению к своему другу и помощью, которую он ему оказывает. Чем агрессивней человек, тем помощь меньше, однако зависимость выражена слабо. Примером значительной отрицательной корреляции ($r = -0,73$) служит зависимость между нервной возбудимостью человека и его эмоциональной уравновешенностью. Чем выше оказывается результат его теста на возбудимость, тем более низкий результат имеет его тест на уравновешенность.

Строгая отрицательная корреляция определяется значением $r = -1$. Она, так же как и строгая положительная корреляция, является абстракцией и не находит отражения в практических исследованиях. Пример, иллюстрирующий строгую отрица-

тельную корреляцию, можно взять из школьного учебника физики: при равномерном движении расстояние равно произведению времени на скорость. При заданном расстоянии время и скорость являются обратно пропорциональными величинами: чтобы пройти путь за половину времени, необходимо идти вдвое быстрее.

Дополнительные сведения

Линейная и криволинейная корреляции

Основной коэффициент корреляции r Пирсона является мерой прямолинейной связи между переменными: его значения достигают максимума, когда точки на графике двумерного рассеяния лежат на одной прямой линии. В реальной жизни отношения между переменными часто оказываются не только вероятностными, но и непрямолинейными; монотонными или немонотонными. Если связь нелинейная, но монотонная, вместо r Пирсона следует использовать ранговые корреляции Спирмена или Кендалла.

Нередко связь между двумя переменными является не только нелинейной, но и немонотонной. В качестве примера рассмотрим такие два фактора, как нервное возбуждение перед экзаменом и успешность его сдачи. Исследования показывают, что студенты, испытывающие умеренное нервное возбуждение, имеют наилучшие результаты на экзаменах, в то время как очень спокойные или очень нервные студенты сдают экзамены значительно хуже. Если по оси абсцисс отложить степень нервного возбуждения, а по оси ординат — результаты сдачи экзаменов, график зависимости между ними примет вид, близкий к перевернутой букве U. При этом любой коэффициент корреляции, вычисленный для этих величин, окажется весьма низким. Это объясняется тем, что для немонотонных отношений нужны другие методы оценки корреляции. Частично мы коснемся этих методов в главах 15 и 16, посвященных видам регрессионного анализа.

Прежде чем оценивать корреляцию двух переменных, рекомендуется построить график зависимости между ними — график двумерного рассеяния. Если график демонстрирует монотонность связи, для вычисления корреляции можно использовать команды подменю Корреляции.

Ранговые корреляции

Как уже отмечалось, необходимость в применении ранговых корреляций возникает в двух случаях: когда распределение хотя бы одной из двух переменных не соответствует нормальному и когда связь между переменными является нелинейной (но монотонной). В этих случаях вместо корреляции r Пирсона можно выбрать ранговые корреляции: r Спирмена либо τ (читается «тау») Кендалла. Ранговыми они являются потому, что программа предварительно ранжирует переменные, между которыми они вычисляются.

Корреляцию r Спирмена программа SPSS вычисляет следующим образом. Сначала переменные переводятся в ранги, а затем к рангам применяется формула

r Пирсона. Таким образом, r Спирмена интерпретируется по аналогии с r Пирсона. Иначе дело обстоит с корреляцией τ Кендалла, которая имеет вероятностную природу.

Рассмотрим принцип вычисления τ Кендалла на примере. Предположим, оценивается связь между ростом и весом в группе людей, предварительно ранжированных по этим переменным. Тогда при сравнении любых двух человек из этой группы возможны две ситуации: однонаправленное изменение переменных («совпадение»), когда и рост, и вес одного больше, чем у другого; разнонаправленное изменение («инверсия»), когда рост у второго больше, а вес меньше, чем у первого. Перебрав все пары испытуемых, можно оценить вероятность совпадений (P) и вероятность инверсий (Q). Корреляция Кендалла — это разность вероятностей «совпадений» и «инверсий»: $\tau = P - Q$. По значению корреляции Кендалла можно всегда вычислить вероятность «совпадений» ($P = (1 + \tau)/2$) и «инверсий» ($Q = (1 - \tau)/2$). Например, если корреляция между ростом и весом $\tau = 0,5$, то вероятность «совпадений» (чем больше рост, тем больше вес) $P = 0,75$, а вероятность «инверсий» (чем больше рост, тем меньше вес) $Q = 0,25$. Таким образом, важным преимуществом корреляции τ Кендалла является ее отчетливая вероятностная интерпретация.

Значимость

Как и большинство статистических процедур, команды подменю Корреляции наряду с описательными статистиками (корреляциями в данном случае) вычисляют их уровень значимости. Напомним, что уровень значимости является мерой статистической достоверности результата вычислений, в данном случае — корреляции, и служит основанием для интерпретации. Если исследование показало, что уровень значимости корреляции не превышает 0,05 ($p \leq 0,05$), то это означает, что корреляция является случайной с вероятностью не более 5 %. Обычно это является основанием для вывода о статистической достоверности корреляции. В противном случае ($p > 0,05$) связь признается статистически недостоверной и не подлежит содержательной интерпретации.

SPSS позволяет определять два теста значимости: односторонний и двухсторонний. Обычно используется двухсторонний тест значимости. Но если вы заранее знаете направление корреляции (положительное или отрицательное) и вас интересует только одно направление, можно использовать односторонний тест значимости. Однако такая ситуация встречается редко, а если и встречается, то правомерность односторонней проверки с трудом поддается обоснованию.

Частная корреляция

Понятие *частной корреляции* связано с ковариацией. Здесь мы упоминаем частную корреляцию лишь как одну из команд подменю Корреляции. Суть частной корреляции заключается в следующем. Если две переменные коррелируют, всегда можно предположить, что эта корреляция обусловлена влиянием третьей переменной, как общей причины совместной изменчивости первых двух переменных. Для проверки этого предположения достаточно исключить влияние этой третьей

переменной и вычислить корреляцию двух переменных без учета влияния третьей переменной (при фиксированных ее значениях). Корреляция, вычисленная таким образом, и называется частной. Например, при исследовании связи между скоростью чтения и зрелостью моральных суждений у детей разного возраста наверняка будет обнаружена корреляция этих двух переменных. Ответ на вопрос, связаны ли они непосредственно, или связь обусловлена возрастом, позволяет дать частная корреляция. Если при фиксированных значениях возраста частная корреляция скорости чтения и зрелости моральных суждений приближается к нулю, можно заключить, что связь между этими переменными обусловлена возрастом.

Пошаговые алгоритмы вычислений

В пошаговых процедурах приведены несколько примеров, иллюстрирующих применение команд подменю Корреляции. В качестве файла данных используется файл ex01.sav. Сначала выполняются три подготовительных шага. Эти шаги (шаги 1–3) позволяют подготовить рабочий файл данных, запустить программу IBM SPSS Statistics 19, и открыть файл (в данном случае — файл ex01.sav). Пошаговые инструкции этого процесса приведены в главе 4 (с. 60), а подробные разъяснения — в главе 2.

После завершения шага 3 на экране должно присутствовать окно редактора данных со строкой меню и загруженным файлом ex01.sav.

ШАГ 4

В меню Анализ выберите команду Корреляции ► Парные. На экране появится диалоговое окно Парные корреляции, показанное на рис. 9.1.

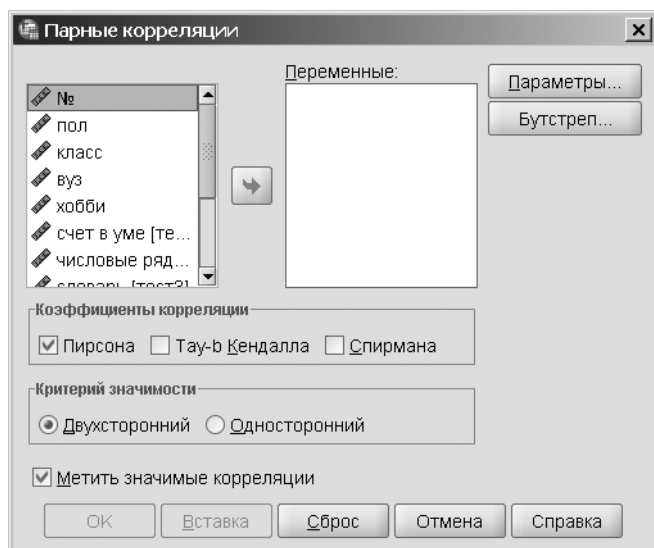


Рис. 9.1. Диалоговое окно Парные корреляции

Окно Парные корреляции позволяет настраивать параметры вычисления корреляций. В левом списке содержатся имена всех переменных, имеющих числовой тип. Строчные переменные для команды Корреляции ► Парные недоступны.

Чтобы выбрать переменные для вычисления корреляции, их требуется переместить в список Переменные при помощи кнопки со стрелкой. Если несколько нужных переменных располагаются в исходном списке друг за другом, вы можете, наведя указатель мыши на верхнюю из них и нажав кнопку мыши, переместить указатель на нижнюю переменную и отпустить кнопку мыши, тем самым выделив сразу несколько переменных.

В группе Коэффициенты корреляции по умолчанию установлен флажок Пирсона. Если требуется вычислить ранговые корреляции, то следует установить флажок Спирмена и (или) Тау-b Кендалла. Можете установить все три флажка, чтобы иметь возможность сравнивать три коэффициента корреляции для различных распределений данных.

В группе Критерий значимости по умолчанию установлен переключатель Двухсторонний. Если вы заранее уверены в направлении (знаке) корреляции, то можете установить переключатель Односторонний.

Флажок Метить значимые корреляции по умолчанию установлен. Это означает, что корреляции, вычисленные с уровнем значимости от 0,01 до 0,05, будут помечены одной звездочкой (*), а от 0 до 0,01 — двумя звездочками (**). Вне зависимости от значимости в вывод включаются коэффициенты корреляции и *p*-уровни, вычисленные с точностью до трех знаков после запятой, а также количество объектов, участвовавших в процедуре.

В примерах, приведенных в этой главе, мы будем пользоваться коэффициентом корреляции Пирсона, двусторонним критерием значимости, а также помечать звездочками значимые корреляции. Если в ваших исследованиях понадобится использовать иную конфигурацию параметров, вы сами можете легко их настроить, устанавливая нужные флажки и переключатели. «Отправной точкой» при выполнении всех примеров служит диалоговое окно Парные корреляции.

ШАГ 5

На этом шаге мы создадим корреляционную матрицу для переменных тест1, ..., тест5, с вычислением корреляций Пирсона. После выполнения шага 4 должно быть открыто диалоговое окно Парные корреляции, представленное на рис 9.1.

1. Щелкните на переменной счет в уме [тест1], чтобы выделить ее. Удерживая клавишу Shift клавиатуры, щелкните по переменной кратковременная память [тест5], для выделения группы переменных тест1, ..., тест5
2. Нажатием стрелки между окнами перенесите выделенную группу переменных в окно Переменные.
3. Щелкните на кнопке ОК, чтобы открыть окно вывода.

Кнопка Параметры позволяет задать дополнительные параметры корреляции. При щелчке на этой кнопке открывается диалоговое окно Парные корреляции: Параметры, представленное на рис. 9.2.

В группе Статистики имеется два флажка, управляющих отображением статистических величин: Средние и стандартные отклонения и Суммы перекрестных произведений отклонений и ковариации. Группа Пропущенные значения из двух переключателей позволяет выбрать способ исключения объектов, содержащих пропущенные значения. Установка переключателя Исключать наблюдения попарно означает, что если при вычислении корреляции между парой переменных для какого-нибудь объекта обнаружится отсутствующее значение, объект будет исключен из вычисления, но только для этой пары переменных. В результате может оказаться, что для разных пар переменных коэффициенты корреляции будут вычислены с разным числом объектов. При установке переключателя Исключать наблюдения целиком программа перед началом вычислительного процесса исключит из рассмотрения все объекты, содержащие хотя бы одно отсутствующее значение. В любом случае разрешение проблемы отсутствующих значений лучше провести до начала анализа.

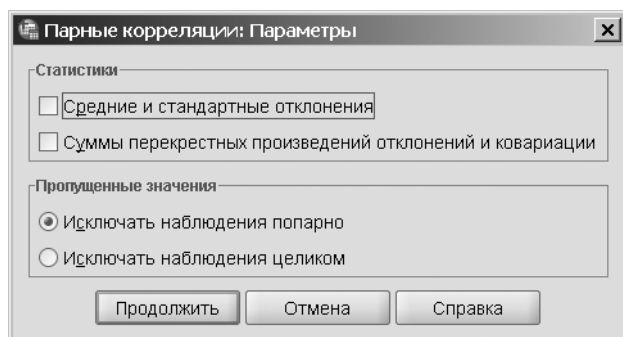


Рис. 9.2. Диалоговое окно Парные корреляции: Параметры

В приведенном примере программа генерирует квадратную корреляционную матрицу. Нередко исследователю необходимо вычислить не все корреляции, а только их часть. Например, можно представить ситуацию, когда возникает необходимость вычислить коэффициенты корреляции между одной переменной и 12 другими переменными, то есть вычислить не квадратную корреляционную матрицу. Возможности команд подменю Корреляция не позволяют этого делать. Конечно, можно сначала получить квадратную корреляционную матрицу (13×13), затем убрать лишние строки и столбцы при помощи редактора таблиц (см. главу 2, раздел «Окно вывода»). Но проще поступить по-другому — создать командный файл.

Для создания командного файла можно открыть окно *редактора синтаксиса*, выбрав команду Создать ► Синтаксис) в меню Файл. Тогда придется вручную вводить все необходимые команды. Но в нашем случае проще открыть редактор синтаксиса нажатием кнопки Вставка диалогового окна Парные корреляции (см. рис. 9.1). Если переменные для анализа заданы, при нажатии на кнопку Вставка открывается окно редактора синтаксиса, изображенное на рис. 9.3. И остается лишь внести необходимые изменения в уже существующий командный файл синтаксиса.

Чтобы выполнить введенную команду, ее нужно выделить и щелкнуть на кнопке со стрелкой Запустить выделенный фрагмент панели инструментов. Второй вариант запуска — выбрать в меню команду Запуск ► Все. Имейте в виду, что любой пропу-

щенный знак, включая завершающую точку, или неверно написанное слово, приведет к выдаче программой сообщения об ошибке.



Рис. 9.3. Окно редактора синтаксиса

ШАГ 5А

На этом шаге мы создадим корреляционную матрицу Пирсона размером 1×5 , предназначенную для оценки степени зависимости переменной от-метка2 от переменных тест1, ..., тест5. После выполнения шага 4 должно быть открыто диалоговое окно Парные корреляции, представленное на рис 9.1. Если вы успели с ним поработать, очистите его щелчком по кнопке Сброс.

1. Щелкните на переменной счет в уме [тест1], чтобы выделить ее. Удерживая клавишу Shift клавиатуры, щелкните по переменной кратковременная память [тест5] для выделения группы переменных тест1, ..., тест5
2. Нажатием стрелки между окнами перенесите выделенную группу переменных в окно Переменные.
3. Нажмите на кнопку Вставка: откроется окно редактора синтаксиса, изображенное на рис. 9.3.

4. В строке, где перечислены переменные, между именами тест5 и от-метка2 введите с клавиатуры команду WITH так, чтобы до и после нее оставались пробелы:

```
CORRELATIONS  
/VARIABLES=тест1 тест2 тест3 тест4 тест5 WITH от-метка2  
/PRINT=TWOTAIL NOSIG  
/MISSING=PAIRWISE
```

5. Выберите в командной строке команду Запуск ► Все или выделите всю команду левой кнопкой мыши и щелкните на кнопке Запустить выделенный фрагмент панели инструментов. Команда будет выполнена и откроется окно вывода.

В конструкции, приведенной здесь, можно произвольно изменять имена переменных. Допускается любое количество переменных как до, так и после ключевого

слова WITH. Переменные, перечисленные до слова WITH, составят строки новой корреляционной матрицы, а после него — столбцы. Если на экране появится сообщение об ошибке, то, скорее всего, либо неверно введено имя переменной, либо забыта точка в конце команды.

После выполнения шага 5 или 5а программа автоматически открывает окно вывода. Для просмотра результатов в нем при необходимости можно воспользоваться вертикальной и горизонтальной полосами прокрутки. Обратите внимание на стандартную строку меню в верхней части окна вывода: ее присутствие позволяет выполнять любые статистические операции, не переключаясь обратно в окно редактора данных.

Представление результатов

На рис. 9.4 приведен фрагмент выводимых данных, сгенерированных программой после выполнения шага 5.

Обратите внимание на структуру представленной на рисунке таблицы. Верхнее значение в каждой ячейке таблицы является коэффициентом корреляции между соответствующими двумя переменными, вычисленным с точностью до трех знаков после запятой. Следующее значение показывает уровень значимости этого коэффициента корреляции. Нижнее значение равно числу объектов, участвовавших в вычислении коэффициента корреляции. При наличии пропусков в анализируемых данных это число в разных ячейках может быть разным. Надписи под таблицей поясняют смысл символов * и **, используемых в качестве меток для некоторых вычисленных коэффициентов корреляции, и содержат информацию об уровне значимости.

Как легко видеть, на главной диагонали таблицы лежат единичные значения коэффициентов корреляции для всех переменных. Смысл этого результата прост: каждая переменная имеет строгую положительную корреляцию сама с собой.

Корреляции		счет в уме	числовые ряды	словарь	осведомленность	кратковременная память
счет в уме	Корреляция Пирсона	1	,436**	-,051	-,134	,017
	Знч. (2-сторон)		,000	,617	,185	,867
	N	100	100	100	100	100
числовые ряды	Корреляция Пирсона	,436**	1	-,196	-,153	,015
	Знч. (2-сторон)	,000		,050	,127	,886
	N	100	100	100	100	100
словарь	Корреляция Пирсона	-,051	-,196	1	,441**	,483**
	Знч. (2-сторон)	,617	,050		,000	,000
	N	100	100	100	100	100
осведомленность	Корреляция Пирсона	-,134	-,153	,441**	1	,475**
	Знч. (2-сторон)	,185	,127	,000		,000
	N	100	100	100	100	100
кратковременная память	Корреляция Пирсона	,017	,015	,483**	,475**	1
	Знч. (2-сторон)	,867	,886	,000	,000	
	N	100	100	100	100	100

** . Корреляция значима на уровне 0.01 (2-сторон.).

Рис. 9.4. Фрагмент окна вывода после выполнения шага 5

После выполнения шага 5а с использованием редактора синтаксиса будет получена таблица результатов, представленная на рис. 9.5.

Корреляции		отметка2
счет в уме	Корреляция Пирсона	,080
	Знч.(2-сторон)	,430
	N	100
числовые ряды	Корреляция Пирсона	,114
	Знч.(2-сторон)	,258
	N	100
словарь	Корреляция Пирсона	,040
	Знч.(2-сторон)	,692
	N	100
осведомленность	Корреляция Пирсона	,257**
	Знч.(2-сторон)	,010
	N	100
кратковременная память	Корреляция Пирсона	,294**
	Знч.(2-сторон)	,003
	N	100

** . Корреляция значима на уровне 0.01 (2-сторон).

Рис. 9.5. Фрагмент окна вывода после выполнения шага 5а

Значимая положительная корреляция в этой таблице наблюдается, в частности, между переменными кратковременная память (тест5) и отметка2 ($r = 0,294$, $p = 0,003$). Это означает, что чем лучше кратковременная память, тем выше средняя отметка за выпускной класс.

Завершение анализа и выход из программы

Отредактируйте содержимое окна вывода в соответствии со своими предпочтениями: скройте лишнюю информацию, исправьте таблицы и пр. (см. раздел «Окно вывода и его редактирование» в главе 2).

Для дальнейшего использования окончательного результата все содержимое окна вывода или его фрагменты можно сохранить в файле *.spv, экспортировать в другой формат (например, Word), перенести в документ Word или вывести на печать (подробности см. в разделе «Сохранение, экспорт, перенос и печать результатов» главы 2).

Для выхода из программы выберите команду Выход в меню Файл.

10 Средние значения

148	Пошаговый алгоритм вычислений
152	Представление результатов
153	Завершение анализа и выход из программы

С помощью команды Таблицы сопряженности, описанной в главе 8, вычисляются частоты по значениям неколичественных (номинальных) переменных. Таблицы сопряженности позволяют сравнивать частоты для разных подгрупп, которые соответствуют значениям номинативной переменной. Например, из таблицы сопряженности пол × хобби видно, что 15 девушек увлекаются спортом, а 27 — искусством и т. п. Команда Средние предназначена для сравнения подгрупп наблюдений по таким показателям количественных переменных, как средние, медианы и пр. При этом предполагается, что среди данных имеются не только количественные переменные, для которых вычисляются средние, но и номинальные переменные, разделяющие объекты на подгруппы. Команда Средние может быть рассмотрена на примере данных файла ex01.sav. Так, при помощи этой команды можно сравнить средние значения успеваемости (отметка1, отметка2) юношей и девушек (пол), учащихся разных классов (класс) и т. д. Результаты вычислений представляются в виде таблиц, похожих на таблицы сопряженности, полученные с использованием команды Таблицы сопряженности. Отличие заключается в том, что для каждой подгруппы вычисляется не только частота, но и другие статистики, например средние.

Команда Средние является одной из самых простых и часто используемых в SPSS. Для выбранных подгрупп она по умолчанию подсчитывает средние значения, стандартные отклонения и частоты. Кроме того, с помощью кнопки Параметры можно задать вывод других описательных статистик для групп наблюдений, а также результатов однофакторного дисперсионного анализа. В этой главе дисперсионный анализ упоминается без детального описания, поскольку этому вопросу целиком посвящены главы 13–16.

Пошаговый алгоритм вычислений

В пошаговых процедурах используется файл ex01.sav с входящими в него переменными класс, пол и отметка2. Сначала выполняются три подготовительных шага. Эти шаги (шаги 1–3) позволят подготовить рабочий файл данных, запустить программу IBM SPSS Statistics 19 и открыть файл (в данном случае — файл ex01.sav). Пошаговые инструкции этого процесса приведены в главе 4 (с. 60), а подробные разъяснения — в главе 2.

После завершения шага 3 на экране должно присутствовать окно редактора данных со строкой меню и загруженным файлом ex01.sav.

ШАГ 4

В меню Анализ выберите команду Сравнение средних ► Средние. На экране появится диалоговое окно Средние, показанное на рис. 10.1

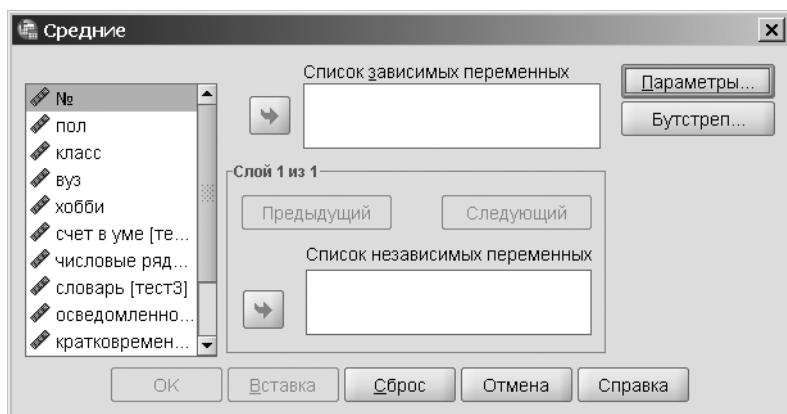


Рис. 10.1. Диалоговое окно Средние

В диалоговом окне Средние вам необходимо задать переменные, которые будут участвовать в процедуре. Список зависимых переменных в верхней части окна предназначен для количественных переменных. Например, в качестве зависимых переменных могут выступать переменные отметка1, отметка2, тест1 и т. д.

Список независимых переменных служит для задания неколичественных (номинальных) переменных, градации которых определяют сравниваемые подгруппы объектов (пол, класс, вуз и т. п.). На основе этих переменных строится таблица сопряженности. Если включить в верхний список только переменную отметка2, а в нижний список — только переменную пол, в результате выполнения команды Средние (с установками по умолчанию) мы увидим средние значения, стандартные отклонения и N отметок для юношей и девушек.

Кнопка Параметры позволяет задавать другие статистики для сравнения подвыборок наблюдений: медиану, асимметрию, эксцесс и пр.

Как правило, исследователи не ограничиваются единственной независимой переменной. Типичным является построение таблицы сопряженности с участием двух переменных. Например, таблица пол × класс позволяет получить средние значения отметок для юношей и девушек каждого из трех классов. Чтобы задать несколько переменных в списке Независимые переменные, используйте кнопки Предыдущий и Следующий справа и слева в разделе Слой 1 из 1. Как вы увидите в приведенных ниже пошаговых процедурах, все действия в окне Средние просты и понятны.

ШАГ 5

После выполнения шага 4 должно быть открыто диалоговое окно Средние, показанное на рис. 10.1. Мы вычислим средние значения отметок для учащихся каждого из трех классов.

1. Щелкните сначала на переменной отметка2, чтобы выделить ее, а затем — на верхней кнопке со стрелкой, чтобы переместить переменную в список Зависимые переменные.

- Щелкните сначала на переменной класс, чтобы выделить ее, а затем — на нижней кнопке со стрелкой, чтобы переместить переменную в список Независимые переменные.
- Щелкните на кнопке ОК, чтобы открыть окно вывода.

Можно включить в процедуру дополнительную независимую переменную, например пол, и тем самым построить таблицу сопряженности класс \times пол, содержащую средние значения и стандартные отклонения.

Количество зависимых переменных может быть любым. Программа просто строит в таблице столько столбцов, сколько переменных указано в списке Зависимые переменные.

В следующем примере мы вычислим средние значения и стандартные отклонения для переменных отметка1 и отметка2 с использованием таблицы сопряженности класс \times пол.

ШАГ 5А

После выполнения шага 4 должно быть открыто диалоговое окно Средние, показанное на рис. 10.1. Если вы уже успели поработать с этим окном, очистите его щелчком на кнопке Сброс и выполните следующие действия.

- Щелкните сначала на переменной отметка1, чтобы выделить ее, а затем — на верхней кнопке со стрелкой, чтобы переместить переменную в список Зависимые переменные.
- Повторите предыдущее действие для переменной отметка2.
- Щелкните сначала на переменной класс, чтобы выделить ее, а затем — на нижней кнопке со стрелкой, чтобы переместить переменную в список Независимые переменные.
- Щелчком на кнопке Следующий освободите список Независимые переменные — метка Слой 1 из 1 изменится на Слой 2 из 2. После этого щелкните сначала на переменной пол, чтобы выделить ее, а затем — на нижней кнопке со стрелкой, чтобы переместить переменную в список Независимые переменные.
- Щелкните на кнопке ОК, чтобы открыть окно вывода результатов.

Для того чтобы выполнить однофакторный дисперсионный анализ или включить в ячейки дополнительную информацию, можно использовать кнопку Параметры в диалоговом окне Средние. После щелчка на этой кнопке откроется диалоговое окно Средние: Параметры, представленное на рис. 10.2.

С помощью диалогового окна Средние: Параметры можно задать дополнительные параметры вывода для команды Средние. Например, помимо величин, вычисляемых по умолчанию (среднего значения, стандартного отклонения и числа наблюдений), можно указать любую совокупность показателей, перечисленных в списке Статистики. Для этого следует выделить нужный пункт списка, а затем щелчком на кнопке со стрелкой добавить его в список Статистики в ячейках.

Как уже упоминалось, команда Средние позволяет выполнять однофакторный дисперсионный анализ (ANOVA). Для этого в группе Статистики для первого слоя

нужно установить флажок Таблица дисперсионного анализа и эта. В процессе группировки зависимой переменной отметка2 по значениям независимой переменной класс программа путем однофакторного дисперсионного анализа сравнит три средних значения для значений переменной класс.

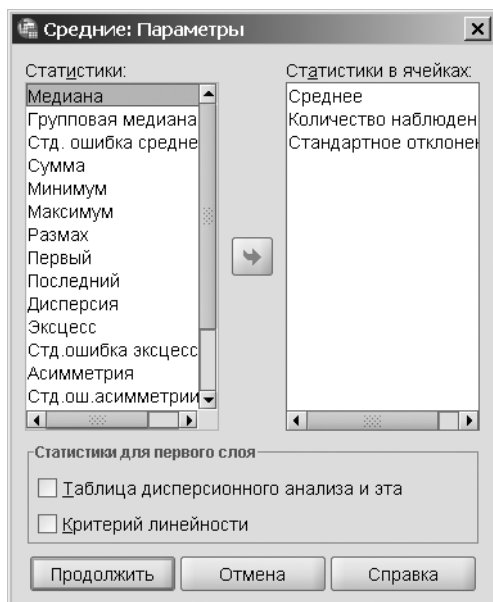


Рис. 10.2. Диалоговое окно Средние: Параметры

В следующем примере мы вычислим средние значения и медианы для каждой из шести подгрупп, образующихся при сопряжении переменных класс и пол. Кроме того, мы проведем дисперсионный анализ с зависимой переменной отметка2 и независимой переменной класс.

ШАГ 5Б

После выполнения шага 4 должно быть открыто диалоговое окно Средние, показанное на рис. 10.1. Если вы уже успели поработать с этим окном, очистите его щелчком на кнопке Сброс и выполните следующие действия.

1. Щелкните сначала на переменной отметка2, чтобы выделить ее, а затем — на верхней кнопке со стрелкой, чтобы переместить переменную в список Зависимые переменные.
2. Щелкните сначала на переменной класс, чтобы выделить ее, а затем — на нижней кнопке со стрелкой, чтобы переместить переменную в список Независимые переменные.
3. Щелчком на кнопке Следующий освободите список Независимые переменные — метка Слой 1 из 1 изменится на Слой 2 из 2. После этого щелкните сначала на переменной пол, чтобы выделить ее, а затем — на нижней кнопке со стрелкой, чтобы переместить переменную в список Независимые переменные.

4. Щелкните на кнопке Параметры, чтобы открыть диалоговое окно Средние: Параметры и установите флажок Таблица дисперсионного анализа и эта.
5. Выделите в окне слева статистику Медиана и перенесите ее в окно Статистики в ячейках, нажав на кнопку со стрелкой между этими окнами.
6. Щелкните на кнопке Продолжить, чтобы вернуться в диалоговое окно Средние.
7. Щелкните на кнопке ОК, чтобы открыть окно вывода.

После выполнения шага 5, 5а или 5б программа автоматически активизирует окно вывода результатов. Для просмотра результатов вы при необходимости можете воспользоваться вертикальной и горизонтальной полосами прокрутки. Обратите внимание на стандартную строку меню в верхней части окна вывода: ее присутствие позволяет выполнять любые статистические операции, не переключаясь обратно в окно редактора данных.

Представление результатов

На рис. 10.3 приведены фрагменты выводимых данных, сгенерированных программой после выполнения шага 5б.

В первой таблице приводятся средние значения, частоты (N) и стандартные отклонения и медианы зависимой переменной отметка2 для каждой из подгрупп. Медианы и средние значения мало отличаются друг от друга, так как распределение переменной близко к нормальному. Во второй и третьей таблицах содержатся результаты дисперсионного анализа. Обратите внимание, что дисперсионный анализ проведен только для средних значений первого слоя, который определяется градациями переменной класс. Для сравнения средних, соответствующих градациям переменной пол, следовало бы ее указать первой в списке Независимые переменные. Более подробное описание однофакторного дисперсионного анализа приводится в главе 13.

Вычисленные средние значения (4,096, 4,167 и 4,408) различаются на уровне значимости $p < 0,001$. Это свидетельствует о статистически достоверной зависимости успеваемости учащихся (отметка2) от класса, в котором они учатся. Коэффициент Эта подобен корреляции и оценивает связь между двумя переменными: количественной и номинальной. Коэффициент Эта в квадрате — мера влияния независимой переменной на дисперсию зависимой переменной. Величина 0,231 свидетельствует о том, что 23,1 % дисперсии зависимой переменной объясняются влиянием независимой переменной.

Отметим, что команда Средние не позволяет определить статистическую значимость при попарном соотношении средних значений. Так, в данном случае по результатам статистической проверки можно сделать вывод только о том, что клас-

сы различаются по успеваемости, без конкретизации, как эти классы отличаются друг от друга. Парно сравнить средние значения можно при помощи *t*-критерия Стьюдента (глава 11), а лучше — при помощи множественных сравнений в рамках однофакторного дисперсионного анализа (глава 13).

Отчет

отметка2

класс	пол	Среднее	N	Стд. Отклонение	Медиана
А	ЖЕН	4,1286	14	,21901	4,1250
	МУЖ	4,0711	19	,26579	4,0500
	Итого	4,0955	33	,24506	4,1000
Б	ЖЕН	4,2048	21	,20670	4,2500
	МУЖ	4,1107	14	,22717	4,1750
	Итого	4,1671	35	,21691	4,2000
В	ЖЕН	4,4173	26	,27711	4,5000
	МУЖ	4,3667	6	,26204	4,3500
	Итого	4,4078	32	,27093	4,4500
Итого	ЖЕН	4,2779	61	,26856	4,2500
	МУЖ	4,1308	39	,26622	4,2000
	Итого	4,2205	100	,27589	4,2000

Таблица ANOVA

		Сумма квадратов	ст.св.	Средний квадрат	F	Знч.
отметка2 * класс	Между группами	1,738	2	,869	14,544	,000
	В группах	5,797	97	,060		
	Итого	7,535	99			

Меры связи

	Эта	Эта квадрат
отметка2 * класс	,480	,231

Рис. 10.3. Фрагменты окна вывода после выполнения шага 5б

Завершение анализа и выход из программы

Отредактируйте содержимое окна вывода в соответствии со своими предпочтениями: скройте лишнюю информацию, исправьте таблицы и пр. (см. раздел «Окно вывода и его редактирование» в главе 2).

Для дальнейшего использования окончательного результата все содержимое окна вывода или его фрагменты можно сохранить в файле *.spv, экспортировать в другой формат (например, Word), перенести в документ Word или вывести на печать (подробности см. в разделе «Сохранение, экспорт, перенос и печать результатов» главы 2).

Для выхода из программы выберите команду Выход в меню Файл.

11 Сравнение двух средних и t -критерий

155	Уровень значимости
155	Пошаговые алгоритмы вычислений
160	Представление результатов
163	Завершение анализа и выход из программы

Различные варианты обработки данных с применением t -критерия позволяют сделать вывод о различии двух средних значений. Например, в случае применения t -критерия для независимых выборок проверяется достоверность различия двух выборок по количественной переменной, измеренной у представителей этих двух выборок. Для этих выборок вычисляются средние значения количественной переменной, затем по t -критерию определяется статистическая значимость различия средних. Применение t -критерия, по-видимому, — самый распространенный метод статистического вывода, так как позволяет ответить на простой вопрос: насколько существенны различия между двумя выборками по данной количественной переменной. Основное требование к данным для применения этого критерия — представление переменных, по которым сравниваются выборки, в метрической шкале измерения. Если есть сомнения в соответствии этому требованию, нужно воспользоваться непараметрическими методами, изложенными в главе 12.

SPSS позволяет применять 3 варианта t -критерия: t -критерий для независимых выборок, t -критерий для парных выборок, одновыборочный t -критерий.

- ▶ Первый из вариантов t -критерия, t -критерий для *независимых выборок*, предназначен для сравнения средних значений двух выборок. Для сравниваемых выборок должны быть определены значения одной и той же переменной. С помощью t -критерия для независимых выборок можно сравнить успеваемость студентов и студенток, степень удовлетворенности жизнью холостяков и женатых, средний рост футболистов двух команд и пр. Обязательным условием для проведения этого t -критерия является независимость выборок.
- ▶ Второй из t -критериев, t -критерий для *парных* или *зависимых выборок*, позволяет сравнить средние значения двух измерений одного признака для одной и той же выборки, например результаты первого и последнего экзаменов группы студентов или значения показателя до и после воздействия на группу. Обязательным условием применения t -критерия для зависимых выборок является наличие повторного измерения для одной выборки.

- Последний из t -критериев, *одновыборочный t -критерий*, позволяет сравнить среднее значение этой выборки с некоторой эталонной величиной. Например, отличается ли среднее значение некоторого теста для данной выборки от нормативной величины, отличается ли время, показанное бегунами во время соревнования, от 17 минут и т. д.

Уровень значимости

Результат сравнения средних значений с применением t -критерия оценивается по уровню значимости. Напомним, что *уровень значимости (p -уровень)* является мерой статистической достоверности результата вычислений, в данном случае — различий средних, и служит основанием для интерпретации. Если исследование показало, что p -уровень значимости различий не превышает 0,05, это означает, что с вероятностью не более 5 % различия являются случайными. Обычно это является основанием для вывода о статистической достоверности различий. В противном случае ($p > 0,05$) различие признается статистически недостоверным и не подлежит содержательной интерпретации.

SPSS позволяет определять два уровня значимости: односторонний и двусторонний. Обычно используется двусторонний уровень значимости. Но если вы заранее знаете направление различий и вас интересует только это направление, то можно использовать односторонний критерий значимости. Однако такая ситуация встречается редко, а если и встречается, то правомерность односторонней проверки с трудом поддается обоснованию.

Существенным различием между двумя типами уровней значимости является то, что величина двустороннего p -уровня оказывается вдвое больше, чем одностороннего. По умолчанию SPSS вычисляет двусторонний уровень значимости, но вы легко можете получить односторонний уровень значимости, разделив вычисленное программой значение на 2.

Пошаговые алгоритмы вычислений

Для применения t -критерия мы воспользуемся переменными из файла ex01.sav, но сначала необходимо выполнить три подготовительных шага. Эти шаги (шаги 1–3) позволят подготовить рабочий файл данных, запустить программу IBM SPSS Statistics 19 и открыть файл (в данном случае — файл ex01.sav). Пошаговые инструкции этого процесса приведены в главе 4, а подробные разъяснения — в главе 2.

После завершения шага 3 на экране должно присутствовать окно редактора данных со строкой меню и загруженным файлом ex01.sav.

Применение t -критерия для независимых выборок

ШАГ 4

В меню Анализ выберите команду Сравнение средних ► T -критерий для независимых выборок. На экране появится диалоговое окно T -критерий для независимых выборок, показанное на рис. 11.1.

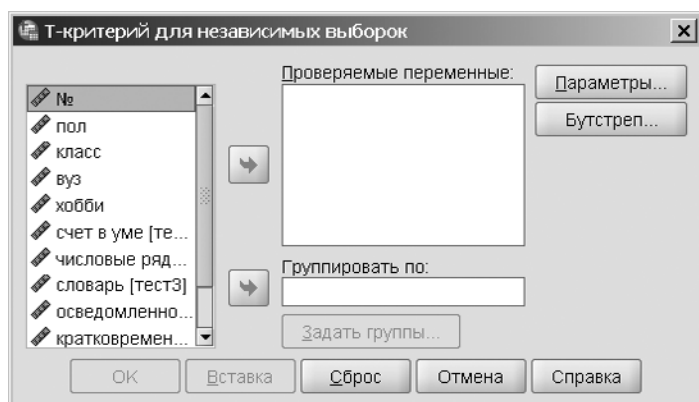


Рис. 11.1. Диалоговое окно Т-критерий для независимых выборок

В этом диалоговом окне вы можете задать параметры для применения t -критерия. Левая часть окна содержит список всех доступных переменных файла, а правая часть — список Проверяемые переменные, в который необходимо поместить имена переменных для применения t -критерия. Список может содержать любое количество переменных, а сами переменные должны быть метрического типа (переменные отметка1, отметка2, тест1 и т. п.). В поле Группировать по: указывается имя переменной, значениям (градациям) которой соответствует две независимые выборки для t -критерия. Как правило, группирующая переменная дискретна и имеет две градации. В файле ex01.sav примером двухуровневой переменной является переменная пол. Группирующая переменная может иметь более двух уровней и даже непрерывный тип, но в обоих случаях необходимо разбить множество значений переменной на две категории или задать два значения из существующих. Так, мы можем взять переменную класс в качестве группирующей переменной и указать, что в одну выборку следует включать все наблюдения, имеющие значение 1, а в другую выборку — наблюдения, имеющие значение 3. Это позволит сравнить средние значения для двух из трех классов.

Итак, после того как имя группирующей переменной задано, необходимо задать ее категории. Это делается щелчком на кнопке Задать группы. Обратите внимание на то, что данный шаг необходим даже для двухуровневых переменных. На экране появится диалоговое окно Задать группы, показанное на рис. 11.2. Если установлен переключатель Пороговое значение, в соответствующем окне следует задать значение переменной, разбивающее весь ее диапазон на два непересекающихся интервала.

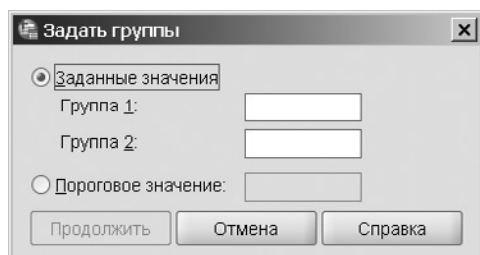


Рис. 11.2. Диалоговое окно Задать группы

ШАГ 5

На этом шаге мы применим t -критерий для сравнения юношей и девушек (значений переменной пол) по переменной отметка2. После выполнения шага 4 на экране должно быть открыто окно Т-критерий для независимых выборок (см. рис. 11.1).

1. Щелкните сначала на переменной отметка2, чтобы выделить ее, а затем — на верхней кнопке со стрелкой, чтобы переместить переменную в список Проверяемые переменные.
2. Щелкните сначала на переменной пол, чтобы выделить ее, а затем — на нижней кнопке со стрелкой, чтобы переместить переменную в поле Группировать по:.
3. Щелкните на кнопке Задать группы, чтобы открыть одноименное диалоговое окно, показанное на рис. 11.2.
4. Введите число 1 в поле Группа 1, нажмите клавишу Tab, введите число 2 в поле Группа 2 и щелкните на кнопке Продолжить, чтобы вернуться в диалоговое окно Т-критерий для независимых выборок.
5. Щелкните на кнопке ОК, чтобы открыть окно вывода.

В следующем примере в качестве группирующей будет использована переменная отметка2. Это количественная переменная — средняя отметка за 10-й класс. При помощи нее разобьем выборку на 2 части по средней отметке: до 4; от 4 и выше. В качестве проверяемой переменной используем кратковременную память (тест5).

ШАГ 5А

После выполнения шага 4 должно быть открыто диалоговое окно Т-критерий для независимых выборок, показанное на рис. 11.1. Если вы уже успели поработать с этим окном, очистите его щелчком на кнопке Сброс и выполните следующие действия.

1. Щелкните сначала на переменной кратковременная память [тест5], чтобы выделить ее, а затем — на верхней кнопке со стрелкой, чтобы переместить переменную в список Проверяемые переменные.
2. Щелкните сначала на переменной отметка2, чтобы выделить ее, а затем — на нижней кнопке со стрелкой, чтобы переместить переменную в поле Группировать по:.
3. Щелкните на кнопке Задать группы, чтобы открыть одноименное диалоговое окно, показанное на рис. 11.2.
4. Установите переключатель Пороговое значение, введите в расположенное рядом поле значение 4 и щелкните на кнопке Продолжить, чтобы вернуться в диалоговое окно Т-критерий для независимых выборок.
5. Щелкните на кнопке ОК, чтобы открыть окно вывода результатов.

Применение t -критерия для парных выборок

Теперь мы рассмотрим два примера применения t -критерия для парных (зависимых) выборок. После завершения шага 3 на экране должно присутствовать окно редактора данных со строкой меню.

ШАГ 4А

В меню Анализ выберите команду Сравнение средних ► *T*-критерий для парных выборок. На экране появится диалоговое окно *T*-критерий для парных выборок, показанное на рис. 11.3.

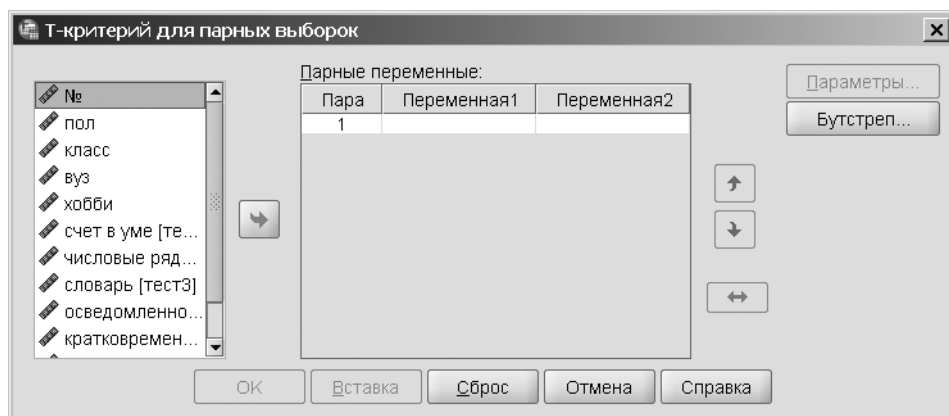


Рис. 11.3. Диалоговое окно *T*-критерий для парных выборок

Применение *t*-критерия для зависимых выборок несколько проще, чем для независимых. Процедура управляется единственным диалоговым окном и не требует указания группирующей переменной. В левой части окна находится уже знакомый нам список доступных переменных файла, в котором выбираются переменные для применения *t*-критерия. Необычным здесь является то, что нужно указывать пары переменных — отсюда название целевого списка: Парные переменные.

Сначала нужно щелкнуть на имени первой переменной, затем — на имени второй переменной, и после этого при помощи кнопки со стрелкой пара заносится в список. Количество выбираемых пар может быть любым.

ШАГ 5Б

Чтобы сравнить отметки учащихся в 10 и 11 классах (отметка1 и отметка2), выполните следующие действия.

1. Щелкните на переменной отметка1, чтобы выделить ее, затем щелчком по стрелке перенесите ее в правое окно. Ее имя появится в области Парные переменные в строке Пара: 1 рядом с меткой Переменная 1.
2. Щелкните на переменной отметка2, чтобы выделить ее, затем щелчком по стрелке перенесите ее в правое окно. Ее имя появится в области Парные переменные в строке Пара: 1 рядом с меткой Переменная 2.
3. Щелкните на кнопке ОК, чтобы открыть окно вывода.

Чтобы применить *t*-критерий одновременно для нескольких пар переменных, их все нужно переместить в список Парные переменные. В следующем примере производится сравнение результатов первого теста (тест1) с результатами остальных четырех тестов.

ШАГ 5В

После выполнения шага 4а должно быть открыто диалоговое Т-критерий для зависимых выборок, показанное на рис. 11.3. Если вы уже успели поработать с этим окном, очистите его щелчком на кнопке Сброс и выполните следующие действия.

1. Нажмите на клавиатуре клавишу Ctrl и удерживайте ее. Удерживая Ctrl, щелкните сначала на переменной счет в уме [тест1], затем — по переменной числовые ряды [тест2], чтобы они были выделены в левом списке переменных.
2. Щелкните на кнопке со стрелкой, чтобы переместить выбранную пару переменных в список Парные переменные.
3. Повторите два предыдущих действия, только вместо переменной тест2 во втором действии последовательно используйте переменные тест3, тест4 и тест5.
4. Щелкните на кнопке ОК, чтобы открыть окно вывода.

Применение *t*-критерия для одной выборки

Иногда бывает необходимо сравнить среднее значение распределения с какой-либо фиксированной величиной. Представим себе следующую ситуацию. Исследователь решил проверить, отличаются ли данные его выборки от нормативных показателей. Предположим, нормативный показатель по выбранной переменной равен 10. Для того чтобы проверить результат выборки на соответствие норме, нужно вычислить среднее значение для выборки и сравнить его с числом 10. Подобные сравнения в программе SPSS реализуются при помощи одновыборочного *t*-критерия — для одной выборки.

ШАГ 4Б

В меню Анализ выберите команду Сравнение средних ► Одновыборочный Т-критерий. На экране появится диалоговое окно Одновыборочный Т-критерий, показанное на рис. 11.4.

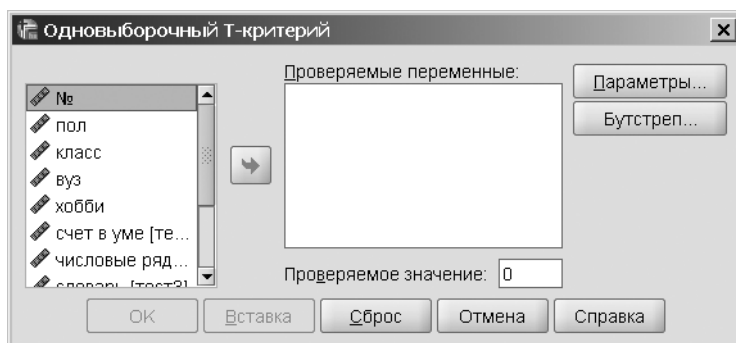


Рис. 11.4. Диалоговое окно Одновыборочный Т-критерий

Применение *t*-критерия для одной выборки является самым простым как с точки зрения вычислений, так и управления. Все, что нужно сделать, — это выбрать одну

или несколько переменных для анализа и, поместив ее (их) в список Проверяемые переменные, задать константу для сравнения в поле Проверяемое значение, после чего щелкнуть на кнопке ОК. Если в списке Проверяемые переменные будет помещено несколько переменных, то их средние значения будут сравниваться с одной и той же величиной, указанной в поле Проверяемое значение. Если нужно провести сравнения с различными константами, то придется применить t -критерий для каждой из констант.

ШАГ 5Г

На этом шаге сравниваются средние значения переменных числовые ряды [тест2] и словарь [тест3] с числом 10, как нормативным для теста.

1. Щелкните сначала на переменной тест2, чтобы выделить ее, а затем — на кнопке со стрелкой, чтобы переместить переменную в список Проверяемые переменные.
2. Повторите предыдущее действие для переменной тест3.
3. Нажмите клавишу Tab, чтобы перевести фокус ввода в поле Проверяемое значение, и введите число 10.
4. Щелкните на кнопке ОК, чтобы открыть окно вывода.

После выполнения шага 5, 5а, 5б, 5в или 5г программа автоматически перейдет в окно вывода. Для просмотра результатов при необходимости можно воспользоваться вертикальной и горизонтальной полосами прокрутки. Обратите внимание на стандартную строку меню в верхней части окна вывода: ее присутствие позволяет выполнять любые статистические операции, не переключаясь обратно в окно редактора данных.

Представление результатов

В этом разделе мы приводим результаты применения всех трех видов t -критерия. После каждого из результатов следуют краткие пояснения, раскрывающие их смысл, а в конце раздела помещен список терминов использовавшихся в окне вывода. Визуальная форма вывода немного изменена, что никак не влияет на его содержимое.

Результаты применения t -критерия для независимых выборок

На рис. 11.5 приведены фрагменты выводимых результатов, сгенерированных программой после выполнения шага 5.

Групповые статистики

пол	N	Среднее	Стд. отклонение	Стд. ошибка среднего
отметка2 ЖЕН	61	4,2779	,26856	,03439
МУЖ	39	4,1308	,26622	,04263

Рис. 11.5. Фрагменты окна вывода после выполнения шага 5

Критерий для независимых выборок								
		Критерий равенства дисперсий Ливиня		t-критерий равенства средних				
		F	Знч.	t	ст.св.	Значимость (2-сторонняя)	Разность средних	Стд. ошибка разности
отметка2	Предполагается равенство дисперсий	,060	,807	2,681	98	,009	,14710	,05488
	Равенство дисперсий не предполагается			2,686	81,647	,009	,14710	,05477

Рис. 11.5. Фрагменты окна вывода после выполнения шага 5 (продолжение)

Из результатов следует, что выборка из 39 юношей имеет средний балл 4,13, выборка из 61 девушки — средний балл 4,28. Различия статистически достоверны на высоком уровне значимости ($p = 0,009$). Критерий равенства дисперсий Ливиня указывает на то, что дисперсии двух распределений статистически значимо не различаются ($p = 0,807$), следовательно, применение t -критерия корректно.

Таблица результатов может быть велика и не помещаться на странице отчета. При желании можно удалить лишние столбцы или строки при помощи редактора таблиц. Для входа в него достаточно дважды щелкнуть на таблице левой кнопкой мыши. Более подробно редактирование таблиц рассмотрено в разделе «Окно вывода и его редактирование» главы 2.

Результаты применения t-критерия для зависимых выборок

На рис. 11.6 приведены фрагменты выводимых результатов, сгенерированных программой после выполнения шага 5б.

Как видно из результатов, для выборки объемом $N = 100$ среднее значение переменной *отметка2* (4,22) оказалось статистически значимо выше среднего значения переменной *отметка1* (3,96) с уровнем значимости $p < 0,001$. Кроме того, между переменными *отметка1* и *отметка2* существует значительная корреляция ($r = 0,434$, $p < 0,001$), свидетельствующая о том, что данные переменные действительно можно считать зависимыми выборками.

Результаты применения t-критерия для одной выборки

На рис. 11.7 приведены фрагменты выводимых результатов, сгенерированных программой после выполнения шага 5г.

Из таблиц видно, что среднее значение переменной *тест2* (числовые ряды) составляет 10,35 и статистически достоверно не отличается от 10 ($p > 0,1$). Среднее значение переменной *тест3* (словарь) равно 11,96 и статистически достоверно отличается от 10 ($p < 0,001$).

Статистики парных выборок

		Среднее	N	Стд. отклонение	Стд. ошибка среднего
Пара 1	отметка1	3,9630	100	,30597	,03060
	отметка2	4,2205	100	,27589	,02759

Корреляции парных выборок

		N	Корреляция	Знач.
Пара 1	отметка1 & отметка2	100	,434	,000

Критерий парных выборок

	Парные разности			t	ст.св.	Значимость (2-сторонняя)
	Среднее	95% доверительный интервал разности средних				
		Нижняя граница	Верхняя граница			
отметка1 - отметка2	-,25750	-,31913	-,19587	-8,290	99	,000

Рис. 11.6. Фрагменты окна вывода после выполнения шага 5б

Статистики для одновыборочного t-критерия

	N	Среднее	Стд. отклонение	Стд. ошибка среднего
числовые ряды	100	10,35	2,768	,277
словарь	100	11,96	2,857	,286

Одновыборочный t-критерий

	Проверяемое значение = 10					
	t	ст.св.	Значимость (2-сторонняя)	Разность средних	95% доверительный интервал разности средних	
					Нижняя граница	Верхняя граница
числовые ряды	1,264	99	,209	,350	-,20	,90
словарь	6,861	99	,000	1,960	1,39	2,53

Рис. 11.7. Фрагменты окна вывода после выполнения шага 5г

Трактовка терминов

Трактовка терминов, используемых программой в окне вывода результатов, дана ниже.

- ▶ Стандартная ошибка — отношение стандартного отклонения к квадратному корню из размера выборки N . Является мерой стабильности среднего значения.
- ▶ F-критерий — величина, характеризующая соотношение дисперсий двух распределений.

- ▶ Значимость — значимость, или p -уровень значимости. При сравнении дисперсии двух распределений, в зависимости от того, равны они или не равны, применяются различные виды статистических приближений. Величина $p > 0,05$ указывает на то, что дисперсии можно считать не различающимися.
- ▶ t (t -критерий) — t -критерий определяется как отношение разности средних значений к стандартному отклонению.
- ▶ ст. св. — число степеней свободы, для t -критерия с независимыми выборками при равенстве дисперсий число степеней свободы равно разности числа объектов и числа групп ($100 - 2 = 98$), а при различии дисперсий применяется более сложная формула, приводящая к дробному значению, равному 81,65. Для зависимых выборок и для одной выборки число степеней свободы для t -критерия определяется как $100 - 1 = 99$.
- ▶ Значимость (2-сторонняя) — по отношению к t -критерию двусторонняя значимость означает вероятность того, что разность между средними значениями является случайной, а по отношению к коэффициенту корреляции — вероятность того, что связь между двумя переменными является случайной.
- ▶ Стд. отклонение — стандартное отклонение. Для t -критерия с зависимыми выборками это стандартное отклонение разности между значениями повторных измерений.
- ▶ Корреляция — мера связи двух переменных, а для зависимых выборок — мера связи парных переменных. Численно определяется коэффициентом корреляции; в данном примере использовался коэффициент Пирсона. Более подробно корреляции описаны в главе 9.
- ▶ 95% доверительный интервал — в случае t -критерия термин «доверительный интервал» относится к разности между средними значениями выборок.

Завершение анализа и выход из программы

Отредактируйте содержимое окна вывода в соответствии со своими предпочтениями: скройте лишнюю информацию, исправьте таблицы и пр. (см. раздел «Окно вывода и его редактирование» в главе 2).

Для дальнейшего использования окончательного результата все содержимое окна вывода или его фрагменты можно сохранить в файле *.spv, экспортировать в другой формат (например, Word), перенести в документ Word или вывести на печать (подробности см. в разделе «Сохранение, экспорт, перенос и печать результатов» главы 2).

Для выхода из программы выберите команду Выход в меню Файл.

12 Непараметрические критерии

164	Параметрические и непараметрические критерии
166	Пошаговые алгоритмы и результаты вычислений
183	Завершение анализа и выход из программы

Параметрические и непараметрические критерии

Перед тем как ввести понятие непараметрического критерия, необходимо уточнить, что такое параметрический критерий. *Параметрический критерий* — это метод статистического вывода, который применяется в отношении *параметров* генеральной совокупности. Самым главным условием для параметрических методов является нормальность распределения переменных и, как следствие, правомерность применения таких статистик, как среднее значение и стандартное отклонение. Несмотря на то что некоторые параметрические методы позволяют анализировать данные, распределенные по другим законам (например, биномиальному или Пуассона), непараметрические методы в этом смысле гораздо функциональнее, поскольку вообще не связывают анализ с каким-либо законом распределения.

Таким образом, непараметрические методы позволяют исследовать данные без каких-либо допущений о характере распределения переменных, в том числе — при нарушении требования нормальности распределения. Так как эти методы предназначены для номинативных и ранговых переменных, в отношении которых недопустимо применение арифметических операций, они основаны на различных дополнительных вычислениях, среди которых можно отметить:

- ▶ ранжирование переменных;
- ▶ подсчет числа значений одного распределения, которые превышают значения другого распределения;
- ▶ применение весовых сравнений;
- ▶ определение степени отклонения распределения от случайного или биномиального распределения;
- ▶ проверка нормальности выборочного распределения;
- ▶ сравнения частот;
- ▶ сравнение групп путем вычисления частот значений, лежащих выше или ниже главной медианы.

Помимо всего прочего непараметрические критерии позволяют вычислять статистические показатели для одной выборки и сравнивать две выборки между собой. Несмотря на кажущуюся сложность, непараметрические методы в большинстве своем очень просты для понимания и применения.

Для использования непараметрических методов мы задействуем уже знакомый нам файл `ex01.sav` и некоторые специально подготовленные примеры. Описание переменных файла `ex01.sav` можно найти в главе 3 (число объектов $N = 100$). Ссылки на имена файлов с дополнительными примерами будут даваться по мере изложения материала.

Структура этой главы несколько отличается от других. Из-за обилия методов пошаговые алгоритмы в этой главе объединены с описанием и интерпретацией результатов. После начальных трех шагов представлены два характерных для большинства непараметрических методов диалоговых окна. Далее даны описания восьми наиболее часто используемых непараметрических методов. Помимо общего описания каждого метода показаны пример его применения в виде шагов 4 и 5, 4а и 5а, 4б и 5б и т. д., программный вывод результатов, его интерпретация и терминология. Указанные восемь методов перечислены ниже.

1. Сравнение двух независимых выборок (*критерий Манна–Уитни*) позволяет установить различия между двумя независимыми выборками по уровню выраженности порядковой переменной.
2. Сравнение двух связанных (зависимых) выборок может проводиться по двум критериям. *Критерий знаков* основан на подсчете числа отрицательных и положительных разностей между повторными измерениями; *критерий Уилкоксона* в дополнение к знакам разностей учитывает их величину.
3. *Критерий серий* определяет, является ли последовательность бинарных величин (событий) случайной или упорядоченной.
4. *Биномиальный критерий* определяет, отличается ли распределение дихотомической величины от заданного соотношения.
5. *Критерий Колмогорова–Смирнова для одной выборки* определяет отличие распределения переменной от нормального (равномерного, Пуассона и т. д.).
6. *Критерий хи-квадрат для одной выборки* определяет степень отличия наблюдаемого распределения частот по градациям переменной от ожидаемого распределения.
7. Сравнение K независимых выборок (*критерий H Крускала–Уоллеса*) позволяет установить степень различия между тремя и более независимыми выборками по уровню выраженности порядковой переменной.
8. Сравнение K связанных (зависимых) выборок (*критерий Фридмана*) позволяет установить степень различия между тремя и более зависимыми выборками по уровню выраженности порядковой переменной.

Пошаговые алгоритмы и результаты вычислений

Для выбора непараметрических методов будем использовать *традиционные диалоговые меню* (Legacy Dialogs — в англоязычном интерфейсе, в русскоязычном интерфейсе они почему-то названы «Устаревшие диалоговые окна») в группе анализа Непараметрические.

Перед применением непараметрических критериев необходимо выполнить три подготовительных шага. Эти шаги (шаги 1–3) позволят подготовить рабочий файл данных, запустить программу IBM SPSS Statistics 19 и открыть файл (в данном случае — файл ex01.sav). Пошаговые инструкции этого процесса приведены в главе 4, а подробные разъяснения — в главе 2.

Некоторые диалоговые окна непараметрических критериев

Применение непараметрических методов иногда требует определять сравниваемые группы по градациям группирующей переменной. Для этого в соответствующих диалоговых окнах предусмотрена кнопка Задать группы, при щелчке на которой открывается характерное диалоговое окно, представленное на рис. 12.1. Если группирующая переменная имеет уровни 1 и 2, то в поле Группа 1 следует ввести значение 1, в поле Группа 2 — значение 2, а затем щелкнуть на кнопке Продолжить.

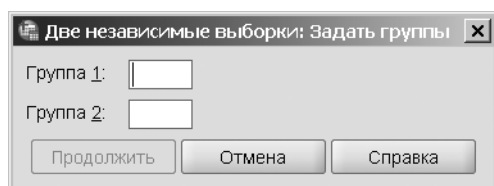


Рис. 12.1. Диалоговое окно определения групп

Еще одно новое для нас диалоговое окно появляется при щелчке на кнопке Параметры в основном диалоговом окне команды применения непараметрических методов (рис. 12.2). В группе Статистики имеются флажки Описательные и Квартили, позволяющие включить в выводимые данные соответственно описательные статистики (среднее значение, стандартное отклонение, минимум, максимум и объем выборки) и квартили.

Процедуры непараметрических методов имеют дополнительную возможность расчета *точного значения* p -уровня значимости, если диалоговое окно метода содержит кнопку Точные.... Если в диалоговом окне метода щелкнуть на кнопке Точные..., появляется окно Точные методы, позволяющее выбрать один из точных методов расчета значимости (рис. 12.3). По умолчанию переключатель в этом окне установлен в положении Только асимптотически, что соответствует определению при-

близительной значимости традиционным способом. Если данные не соответствуют требованиям применяемого метода, при помощи переключателя можно воспользоваться одним из двух точных методов расчета значимости: методом статистических испытаний «Монте-Карло» или собственно точным критерием. Метод Точные по определению дает более точные результаты. Однако в некоторых случаях его расчет слишком трудоемок даже для компьютера, тогда используется метод Монте-Карло.

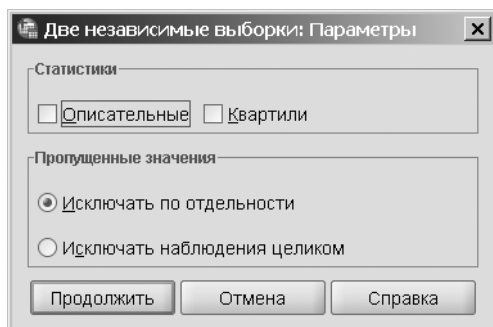


Рис. 12.2. Диалоговое окно настройки параметров непараметрических тестов

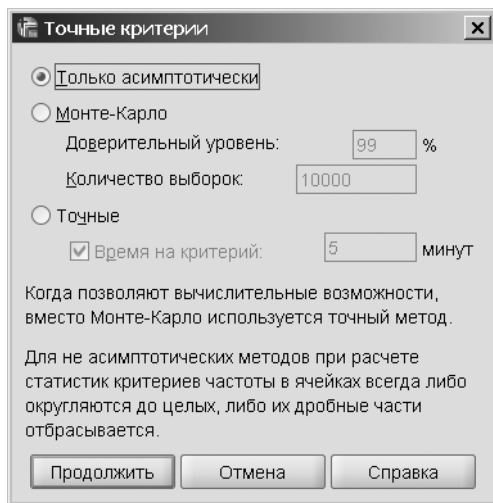


Рис. 12.3. Диалоговое окно Точные критерии

Сравнение двух независимых выборок

Критерий Манна–Уитни (Mann-Whitney), или *U-критерий*, по назначению аналогичен *t*-критерию для независимых выборок. Разница заключается в том, что *t*-критерии ориентированы на нормальные и близкие к ним распределения, а критерий Манна–Уитни — на распределения, отличные от нормальных. В частном случае критерий Манна–Уитни можно применять и для нормально распределенных данных, однако он менее чувствителен к различиям (является менее мощным),

чем t -критерий. Далее в примере мы попытаемся выяснить, различаются ли юноши и девушки по успеваемости в выпускном классе. Напомним, что пол учащихся определяется значением переменной пол, а успеваемость — переменной отметка2. При реализации метода программа сначала ранжирует все объекты без учета принадлежности к сравниваемым группам, а затем вычисляет средние ранги для каждой из двух групп. Чем выше средний ранг группы, тем выше ее успеваемость. После нахождения средних рангов определяется p -уровень. Корректность применения этого критерия сомнительна, если переменная имеет небольшое число возможных значений, например 3-балльная шкала. В этом случае следует воспользоваться точными критериями. Диалоговое окно U -критерия представлено на рис. 12.4.

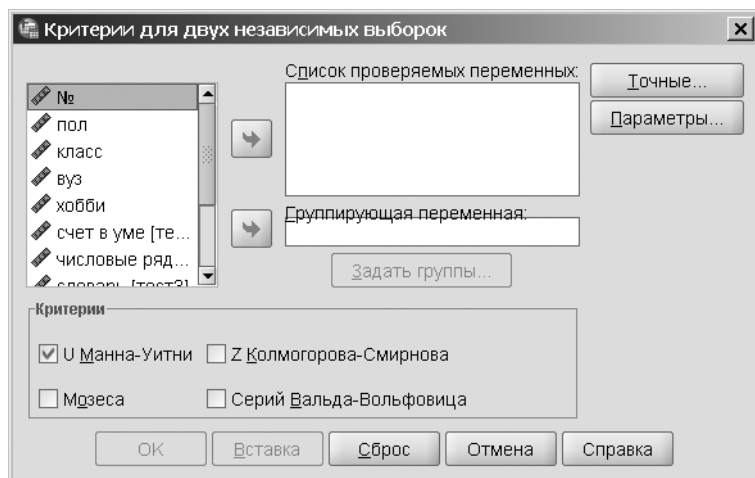


Рис. 12.4. Диалоговое окно Критерии для двух независимых выборок

После завершения шага 3 на экране должно присутствовать окно редактора данных. В этом примере мы сравним успеваемость юношей и девушек в выпускном классе.

ШАГ 4

В меню Анализ выберите команду Непараметрические критерии ► Устаревшие диалоговые окна ► Для двух независимых выборок, чтобы открыть диалоговое окно Критерии для двух независимых выборок, показанное на рис. 12.4.

ШАГ 5

Для применения метода выполните следующую последовательность действий.

1. Щелкните сначала на переменной отметка2, чтобы выделить ее, а затем — на верхней кнопке со стрелкой, чтобы переместить переменную в Список проверяемых переменных.
2. Щелкните сначала на переменной пол, чтобы выделить ее, а затем — на нижней кнопке со стрелкой, чтобы переместить переменную в поле Группирующая переменная.

3. Щелкните на кнопке Задать группы, чтобы открыть диалоговое окно, показанное на рис. 12.1.
4. В поле Группа 1 введите значение 1, нажмите клавишу Tab, чтобы переместить фокус ввода в поле Группа 2, введите значение 2 и щелкните на кнопке Продолжить, чтобы вернуться в диалоговое окно Критерий для двух независимых выборок.
5. Щелкните на кнопке ОК, чтобы открыть окно вывода.

Результаты работы программы показаны на рис. 12.5.

Ранги			
пол	N	Средний ранг	Сумма рангов
отметка2 ЖЕН	61	56,21	3429,00
МУЖ	39	41,56	1621,00
Всего	100		

Статистика критерия ^a	
	отметка2
Статистика U Манна-Уитни	841,000
Статистика W Уилкоксона	1621,000
Z	-2,469
Асимпт. знч. (двухсторонняя)	,014

a. Группирующая переменная: пол

Рис. 12.5. Фрагменты окна вывода после выполнения шага 5

Средний ранг для девушек равен 56,21, а для юношей — 41,56. Это значит, что у девушек успеваемость выше, чем у юношей. Статистика U Манна-Уитни равна 841. Значение Z является нормализованным, связанным с уровнем значимости $p = 0,014$. Поскольку величина уровня значимости (Асимпт. знч (двухсторонняя)) меньше 0,05, мы можем быть уверены в статистической достоверности вывода о том, что успеваемость девушек действительно выше успеваемости юношей.

Сравнение двух связанных (зависимых) выборок

Основные методы, которые используются для сравнения двух зависимых выборок, — это *критерий знаков* и *критерий Уилкоксона* (Wilcoxon test). Диалоговое окно Критерии для двух связанных выборок представлено на рис. 12.6.

Критерий знаков

Критерий знаков позволяет сравнить два измерения переменной на одной выборке (например, «до» и «после») по уровню ее выраженности путем сопоставления количества положительных и отрицательных разностей (сдвигов) значений. Для того чтобы продемонстрировать применение критерия, сравним результаты учащихся по второму (тест2) и четвертому (тест4) тестам. Для каждого объекта сначала определяется знак разности значений. Так, для первого объекта значение тест2 равно 7, а значение тест4 — 10. Сравнение этих значений даст отрицательный

знак ($7 - 10 = 3 < 0$). Для четвертого объекта значения переменных тест2 и тест4 составляют соответственно 9 и 6, и знак разности будет положительным ($9 - 6 > 0$). Подсчитывается число положительных, отрицательных и нулевых разностей, а затем вычисляется нормализованное z -значение и p -уровень значимости.

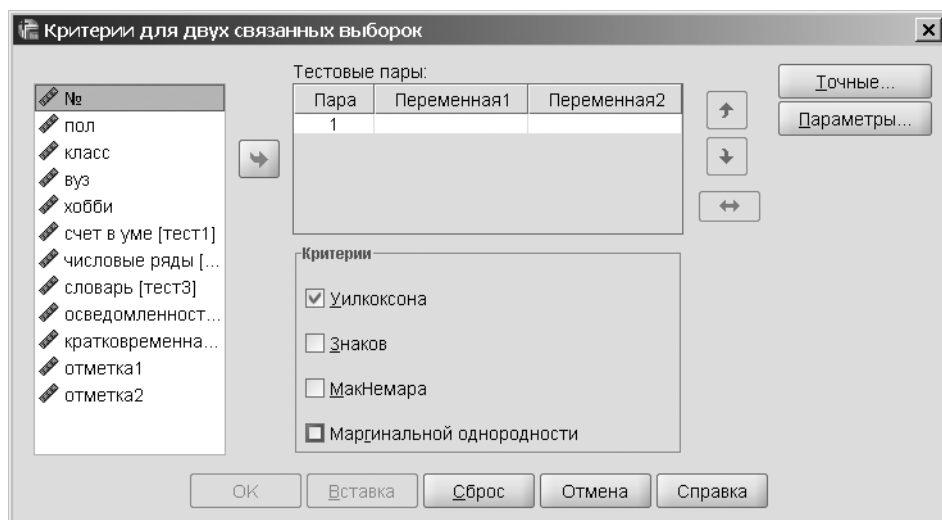


Рис. 12.6. Диалоговое окно Критерии для двух связанных выборок

После завершения шага 3 на экране должно присутствовать окно редактора данных. В этом примере мы сравним результаты учащихся по второму (тест2) и четвертому (тест4) тестам.

ШАГ 4А

В меню Анализ выберите команду Непараметрические критерии ► Устаревшие диалоговые окна ► Для двух независимых выборок, чтобы открыть диалоговое окно Критерии для двух связанных выборок, показанное на рис. 12.4.

ШАГ 5А

Для применения критерия знаков выполните следующую последовательность действий.

1. В группе Критерии установите флажок Знаков и сбросьте флажок Уилкоксона.
2. Щелкните на переменной числовые ряды [тест2], чтобы выделить ее, затем щелчком по стрелке перенесите ее в правое окно. Ее имя появится в области Парные переменные в строке Пара: 1 рядом с меткой Переменная 1.
3. Щелкните на переменной осведомленность [тест4], чтобы выделить ее, затем щелчком по стрелке перенесите ее в правое окно. Ее имя появится в области Парные переменные в строке Пара: 1 рядом с меткой Переменная 2.
4. Щелкните на кнопке ОК, чтобы открыть окно вывода.

Результаты работы программы показаны на рис. 12.7.

Частоты		N
осведомленность - числовые ряды	Отрицательные разности ^а	39
	Положительные разности ^б	57
	Связи ^с	4
	Всего	100

а. осведомленность < числовые ряды

б. осведомленность > числовые ряды

с. осведомленность = числовые ряды

Статистика критерия ^а	
	осведомл енность - числовые ряды
Z	-1,735
Асимпт. знч. (двухсторонняя)	,083

а. Критерий знаков

Рис. 12.7. Фрагменты окна вывода после выполнения шага 5а

Полученные результаты говорят о том, что в 39 случаях значения переменной тест2 оказались меньшими, чем значения переменной тест4, в 57 случаях значения переменной тест2 превысили значения переменной тест4, и 4 раза было установлено равенство значений обеих переменных. Стандартизованное значение (Z) составляет $-1,735$, а уровень значимости $p = 0,083$. Это означает, что различия между результатами тестов тест4 и тест2 статистически недостоверны. Обратите внимание: поскольку переменные тест4 и тест2 являются метрическими, к ним предпочтительней применить t -критерий для парных выборок. Он показал бы, что средние значения тест4 и тест2 различаются с уровнем значимости $p = 0,01$. Таким образом, можно на практике убедиться в том, что статистические возможности t -критерия в отношении переменных значительно выше, чем возможности критерия знаков.

Критерий Уилкоксона

Недостатком критерия знаков является то, что он никак не учитывает величину разности двух значений. Так, для него нет разницы между результатами сравнения пар (10; 0) и (6; 5): в обоих случаях знак разности будет положительным. Однако очевидно, что абсолютное значение разности также характеризует соотношение распределений. Для того чтобы учесть это, применяется критерий Уилкоксона. Этот критерий основан на подсчете абсолютных разностей между парами значений с последующим их ранжированием. Затем вычисляются средние значения рангов для положительных и отрицательных разностей (сдвигов). Уровень значимости подсчитывается на основе стандартизованного значения. Корректность применения этого критерия сомнительна, если переменная имеет небольшое число возможных значений, например 3-балльная шкала. В этом случае следует восполь-

зоваться Точными критериями. Основное диалоговое окно критерия Уилкоксона то же, что и для критерия знаков (см. рис. 12.6).

После завершения шага 3 на экране должно присутствовать окно редактора данных. В этом примере мы, как и в предыдущем, сравним результаты учащихся по второму (тест2) и четвертому (тест4) тестам. Однако на этот раз вместо критерия знаков воспользуемся критерием Уилкоксона.

ШАГ 4Б

В меню Анализ выберите команду Непараметрические критерии ► Устаревшие диалоговые окна ► Для двух независимых выборок, чтобы открыть диалоговое окно Критерии для двух связанных выборок, показанное на рис. 12.6. Если вы уже успели поработать с этим окном, щелкните на кнопке Сброс.

ШАГ 5Б

Для применения критерия Уилкоксона выполните следующую последовательность действий.

1. Щелкните на переменной числовые ряды [тест2], чтобы выделить ее, затем щелчком по стрелке перенесите ее в правое окно. Ее имя появится в области Парные переменные в строке Пара: 1 рядом с меткой Переменная 1.
2. Щелкните на переменной осведомленность [тест4], чтобы выделить ее, затем щелчком по стрелке перенесите ее в правое окно. Ее имя появится в области Парные переменные в строке Пара: 1 рядом с меткой Переменная 2.
3. Щелкните на кнопке ОК, чтобы открыть окно вывода.

Обратите внимание, что мы не устанавливали флажок Уилкоксона, поскольку он установлен по умолчанию. Результаты работы программы показаны на рис. 12.8.

Ранги		N	Средний ранг	Сумма рангов
осведомленность - числовые ряды	Отрицательные ранги	39 ^a	42,26	1648,00
	Положительные ранги	57 ^b	52,77	3008,00
	Связи	4 ^c		
	Всего	100		

a. осведомленность < числовые ряды

b. осведомленность > числовые ряды

c. осведомленность = числовые ряды

Статистика критерия^a

	осведомленность - числовые ряды
Z	-2,493 ^a
Асимпт. знч. (двухсторонняя)	,013

a. Используются отрицательные ранги.

Рис. 12.8. Фрагменты окна вывода после выполнения шага 5б

Результаты применения критерия Уилкоксона и критерия знаков очень похожи. Частота каждого из трех исходов N осталась неизменной. Информация о каждом из исходов (кроме равенства) теперь включает также среднее и суммарное значения для соответствующих рангов. Визуальный анализ исходных данных говорит о том, что значения теста 4 (осведомленность) в целом несколько превышают значения теста 2 (числовые ряды). Это демонстрирует и величина $Z = -2,493$, которая значительно превосходит по модулю соответствующее значение, полученное ранее для критерия знаков. Уровень значимости $p = 0,013$, что говорит о статистической достоверности различий. Таким образом, мы убеждаемся в том, что критерий Уилкоксона является более чувствительным к различиям (более мощным), чем критерий знаков. Тем не менее он оказывается несколько хуже t -критерия, обеспечивающего уровень значимости 0,01, что подтверждает предпочтительность последнего для анализа метрических данных.

Критерий серий

Как следует из названия, *критерий серий* применяется для анализа последовательности объектов (явлений, событий), упорядоченных во времени или в порядке возрастания (убывания) значений измеренного признака. Кроме того, критерий требует представления последовательности в виде бинарной переменной, то есть как чередования событий 0 и 1. Математическая идея критерия основана на подсчете числа серий в упорядоченной последовательности событий двух типов, например 0 и 1. *Серия* — это последовательность однотипных событий, непосредственно перед и после которой произошли события другого типа. Гипотеза о случайном распределении событий 1 среди событий 0 может быть отклонена, если количество серий либо слишком мало (однотипные события имеют тенденцию к группированию), либо слишком велико (события 0 и 1 имеют тенденцию к чередованию).

Основное диалоговое окно критерия серий представлено на рис. 12.9.

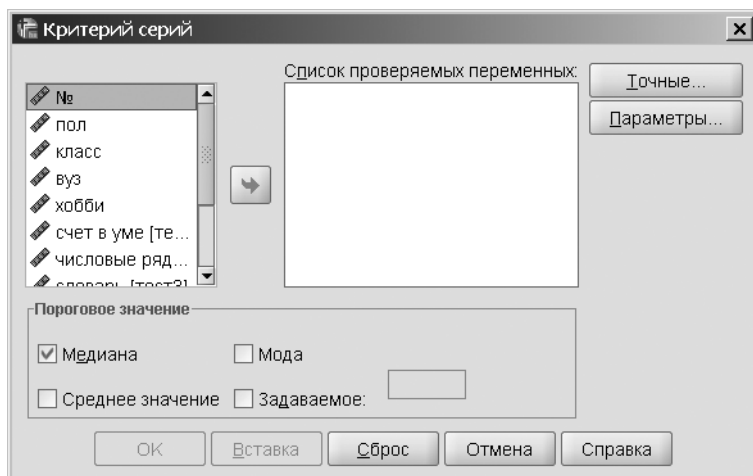


Рис. 12.9. Диалоговое окно Критерий серий

После завершения шага 3 на экране должно присутствовать окно редактора данных. Рассмотрим применение критерия серий на примере проверки гипотезы о неслучайном чередовании юношей и девушек (переменная пол) в списке испытуемых в файле ex01.sav. Для задания точки раздела значений переменной на две категории предназначены флажок и поле Задаваемое. В данном случае в одну группу надо включить значения переменной пол, равные 1, а в другую группу — значения, равные 2.

ШАГ 4А

В меню Анализ выберите команду Непараметрические критерии ► Устаревшие диалоговые окна ► Серии, чтобы открыть диалоговое окно Критерий серий, показанное на рис. 12.9.

ШАГ 5В

Для применения критерия выполните следующую последовательность действий.

1. Щелкните сначала на переменной пол, чтобы выделить ее, а затем — на кнопке со стрелкой, чтобы переместить переменную в поле Список проверяемых переменных.
2. В группе Пороговое значение установите флажок Задаваемое, введите в расположенное рядом поле значение 2 и сбросьте флажок Медиана.
3. Щелкните на кнопке ОК, чтобы открыть окно вывода.

Результаты работы программы показаны на рис. 12.10.

Критерий серий					
	Проверяемое значение ^а	Всего наблюдений	Число серий	Z	Асимпт. знч. (двухсторонняя)
пол	2,00	100	49	,089	,929

а. Заданный пользователем.

Рис. 12.10. Фрагмент окна вывода после выполнения шага 5в

Количество серий равно 49. В результаты включено значение точки деления, введенное в поле Задаваемое. Величина Z и соответствующая значимость зависят от числа серий. Число серий преобразуется к z-значению, для которого и определяется p -уровень. Большое значение p -уровня (0,929) свидетельствует о том, что чередование юношей и девушек в файле ex01.sav является случайным. Статистически значимый результат свидетельствовал бы о том, что чередование юношей и девушек в файле является неслучайным. Если при этом число серий было бы слишком велико, это свидетельствовало бы о том, что после юноши с высокой долей вероятности следует девушка (и наоборот). При малом значении числа серий можно было бы сделать вывод о том, что более вероятно группирование испытуемых в списке по половому признаку (после юноши чаще следует юноша, а после девушки — девушка).

Биномиальный критерий

Назначение *биномиального критерия* — определение вероятности того, что наблюдаемое распределение не отличается от ожидаемого (заданного) биномиального

распределения. Свойством биномиального распределения является заранее заданное соотношение вероятностей двух взаимоисключающих событий (обычно — равновероятное). Например, при многократном подбрасывании «правильной» монеты вероятности выпадения «орлов» и «решек» подчиняется биномиальному распределению. В качестве примера мы снова исследуем распределение юношей и девушек в файле ex01.sav. Зная заранее, что в файле собраны данные о 61 ученице и 39 учениках, мы сможем проверить, отличается ли статистически достоверно это распределение (наблюдаемое) от ожидаемого (теоретического) равновероятного соотношения. Основное диалоговое окно биномиального критерия представлено на рис. 12.11.

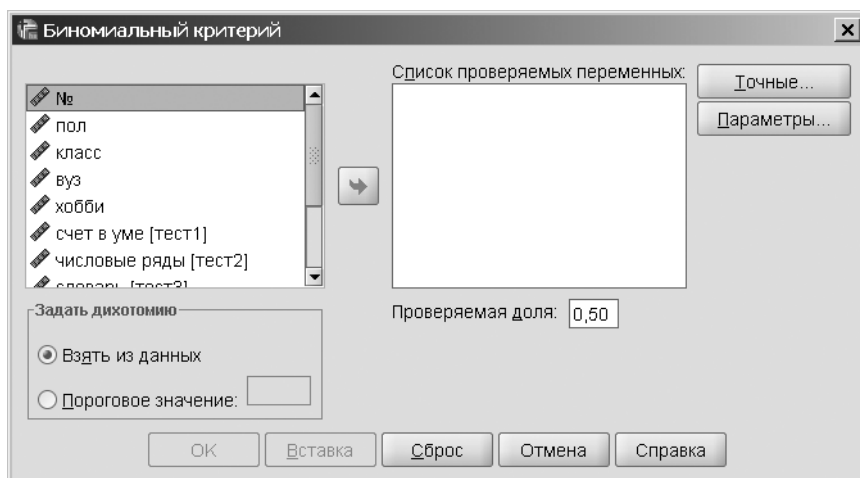


Рис. 12.11. Диалоговое окно биномиального критерия

После завершения шага 3 на экране должно присутствовать окно редактора данных. В этом примере мы попытаемся выяснить, является ли распределение мужчин и женщин в выборке биномиальным.

ШАГ 4Г

В меню Анализ выберите команду Непараметрические критерии ► Устаревшие диалоговые окна ► Биномиальный, чтобы открыть диалоговое окно Биномиальный критерий, показанное на рис. 12.11.

ШАГ 5Г

Для применения биномиального критерия выполните следующую последовательность действий.

1. Щелкните сначала на переменной пол, чтобы выделить ее, а затем — на кнопке со стрелкой, чтобы переместить переменную в Список проверяемых переменных. Обратите внимание, что ожидаемое соотношение, заданное по умолчанию (Проверяемая доля), равно 0,5.
2. Щелкните на кнопке ОК, чтобы открыть окно вывода.

Результаты работы программы показаны на рис. 12.12.

Биномиальный критерий

		Категория	N	Наблюден- ная доля	Проверяе- мая доля	Асимпт. знч. (двухсторон- ная)
пол	Группа 1	МУЖ	39	,39	,50	,035 ^a
	Группа 2	ЖЕН	61	,61		
	Всего		100	1,00		

a. Используется Z аппроксимация.

Рис. 12.12. Фрагмент окна вывода после выполнения шага 5г

Ожидаемая пропорция для биномиального теста равна 0,5 для обеих групп. Наблюдаемая пропорция для каждой из групп определяется как отношение размера группы (N) к размеру выборки (100). Как можно видеть, наблюдаемые пропорции значительно отличаются от 0,5 и составляют 0,39 для мужчин и 0,61 для женщин. Уровень значимости, равный 0,035, свидетельствует о статистически достоверном отличии исследуемого распределения от биномиального (равновероятного).

При желании можно проверять отличие наблюдаемого распределения от любого другого биномиального распределения, задавая необходимые ожидаемые пропорции в поле Проверяемая доля.

Критерий Колмогорова–Смирнова для одной выборки

Критерий Колмогорова–Смирнова для одной выборки позволяет определить, отличается ли заданное распределение от нормального (эксцесс и асимметрия распределения равны 0), равномерного (значения распределены с одинаковой плотностью, например, как у целых чисел от 1 до 1000), Пуассона (среднее значение и дисперсия равны λ ; при больших значениях λ распределение Пуассона приближается к нормальному) или экспоненциального. Суть метода заключается в сравнении эмпирического (наблюдаемого) распределения накопленных частот выборки с теоретическим (ожидаемым) распределением накопленных частот (нормальным, Пуассона и т. д.). В качестве примера мы исследуем распределение значений переменной `отметка1` в файле `ex01.sav` на соответствие нормальному распределению. Основное диалоговое окно критерия Колмогорова–Смирнова для одной выборки представлено на рис. 12.13.

После завершения шага 3 на экране должно присутствовать окно редактора данных. В этом примере мы сравним распределение переменной `отметка1` с нормальным распределением.

ШАГ 4Д

В меню Анализ выберите команду Непараметрические критерии ► Устаревшие диалоговые окна ► Одновыборочный Колмогорова–Смирнова, чтобы открыть диалоговое окно Одновыборочный критерий Колмогорова–Смирнова, показанное на рис. 12.13.

ШАГ 5Д

Для применения критерия выполните следующую последовательность действий.

1. Щелкните сначала на переменной **отметка1**, чтобы выделить ее, а затем — на кнопке со стрелкой, чтобы переместить переменную в Список тестируемых переменных.
2. Щелкните на кнопке **ОК**, чтобы открыть окно вывода.

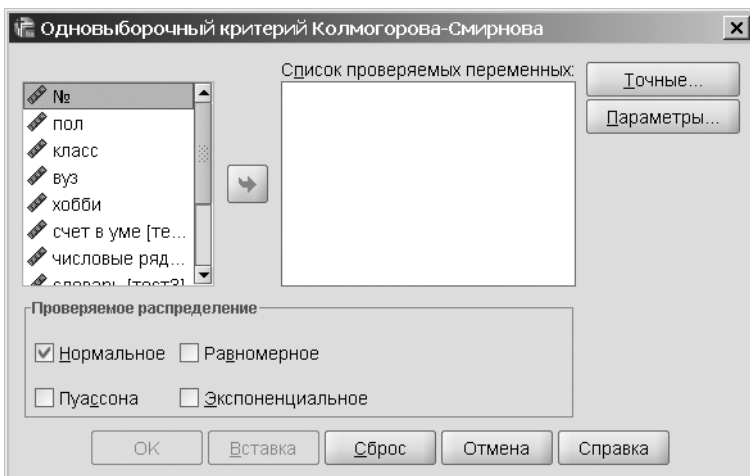


Рис. 12.13. Диалоговое окно Одновыборочный критерий Колмогорова-Смирнова

Результаты работы программы показаны на рис. 12.14.

Одновыборочный критерий Колмогорова-Смирнова		
		отметка1
N		100
Нормальные параметры	Среднее	3,9630
	Стд. отклонение	,30597
Разности экстремумов	Модуль	,072
	Положительные	,072
	Отрицательные	-,051
Статистика Z Колмогорова-Смирнова		,716
Асимпт. знч. (двухсторонняя)		,685

а. Сравнение с нормальным распределением.

б. Оценивается по данным.

Рис. 12.14. Фрагмент окна вывода после выполнения шага 5д

В строке **Разности экстремумов** приведены **Модуль**, а также **Положительные** и **Отрицательные** отклонения исследуемого распределения от теоретического (в данном случае, нормального). Строка **Статистика Z Колмогорова-Смирнова** содержит z-значение, уровень значимости которого равен 0,685 (последняя строка). Это

означает, что распределение значений переменной `отметка1` статистически не отличается от нормального ($p > 0,05$).

Критерий хи-квадрат для одной выборки

Рассматриваемый в данном разделе критерий χ^2 для одной выборки несколько отличается от рассмотренного ранее критерия χ^2 для таблиц сопряженности (см. главу 8). В данном случае в качестве ожидаемого (теоретического) распределения обычно выступает равномерное распределение объектов по градациям переменной, в отношении которой применяется критерий. Далее будет приведен пример применения критерия χ^2 к переменной `вуз` из файла `ex01.sav`. Поскольку число объектов (N) в файле `ex01.sav` равно 100, а переменная `вуз` имеет 4 градации, ожидаемые частоты для каждой градации равны $100/4 = 25$. Применение рассматриваемого критерия допускает задание не только равномерного ожидаемого распределения, но и любого другого. Например, можно проверить гипотезу о том, что соотношение учащихся, предпочитающих 4 категории специализаций, соотносятся как 20:20:30:30. Для этого в группе Ожидаемые значения следует установить переключатель Значения, а затем при помощи поля и кнопки Добавить последовательно ввести в список значения 20, 20, 30, 30. После этих действий ожидаемые частоты изменятся в соответствии с заданными пропорциями. Основное диалоговое окно критерия χ^2 для одной выборки представлено на рис. 12.15.

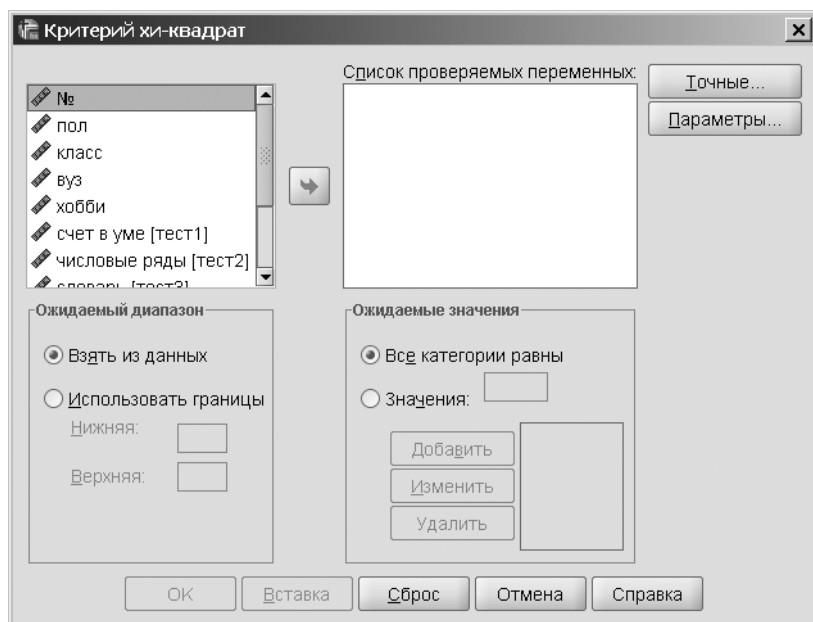


Рис. 12.15. Диалоговое окно Критерий хи-квадрат

После завершения шага 3 на экране должно присутствовать окно редактора данных. В этом примере мы применим критерий χ^2 к переменной `вуз`.

ШАГ 4Е

В меню Анализ выберите команду Непараметрические критерии ► Установившие диалоговые окна ► Хи-квадрат, чтобы открыть диалоговое окно Критерий хи-квадрат, показанное на рис. 12.15.

ШАГ 5Е

Для применения критерия выполните следующую последовательность действий.

1. Щелкните сначала на переменной вуз, чтобы выделить ее, а затем — на кнопке со стрелкой, чтобы переместить переменную в Список тестируемых переменных.
2. Щелкните на кнопке ОК, чтобы открыть окно вывода.

Результаты работы программы показаны на рис. 12.16.

вуз			
	Наблюдаемое N	Ожидаемое N	Остаток
ГУМ	22	25,0	-3,0
ЭКОН	31	25,0	6,0
ТЕХН	11	25,0	-14,0
ЕСТ_Н	36	25,0	11,0
Всего	100		

Статистики критерия	
	вуз
Хи-квадрат ^а	14,480
ст.св.	3
Асимпт. знч.	,002

а. Частоты, меньшие 5, ожидалось в 0 ячеек (0%).
Минимальная ожидаемая частота равна 25,0.

Рис. 12.16. Фрагменты окна вывода после выполнения шага 5е

Первая из таблиц демонстрирует заметные различия наблюдаемых и ожидаемых частот. Остаток — это разность между наблюдаемыми и ожидаемыми частотами. Число степеней свободы (ст.св.) определяется как число значений (градаций) переменной, уменьшенное на 1. Уровень значимости ($p = 0,002$) свидетельствует о статистически достоверном отличии наблюдаемого распределения предпочтений от равномерного распределения.

Сравнение К независимых выборок и критерий Крускала–Уоллеса

Для сравнения более двух независимых выборок по уровню выраженности переменной применяется несколько критериев: *Н-критерий Крускала–Уоллеса*, *критерий медианы*, *критерий Джонкира–Терпстра*. Из них наибольшей чувствительностью к различиям обладает *Н-критерий Крускала–Уоллеса*. Этот критерий является непараметрическим аналогом однофакторного дисперсионного анализа, отличаясь от него в двух отношениях. Во-первых, критерий Крускала–Уоллеса основан не

на сравнении средних значений и дисперсий переменных, а на сравнении средних рангов. Во-вторых, вместо вычисления F -критерия на основе сравнения средних рангов с ожидаемыми значениями вычисляется критерий хи-квадрат. Для нормальных распределений однофакторный дисперсионный анализ обеспечивает более точные результаты, чем критерий Крускала—Уоллеса, однако применение последнего рекомендуется для распределений, отличающихся от нормального.

H -критерий Крускала—Уоллеса «по идее» сходен с U -критерием Манна—Уитни. Как и последний, он оценивает степень пересечения (совпадения) нескольких рядов значений измеренного признака. Чем меньше совпадений, тем больше различаются ряды, соответствующие сравниваемым выборкам. Основная идея H -критерия Крускала—Уоллеса основана на представлении всех значений сравниваемых выборок в виде одной общей последовательности упорядоченных (ранжированных) значений с последующим вычислением среднего ранга для каждой из выборок. Если выполняется статистическая гипотеза об отсутствии различий, можно ожидать, что все средние ранги примерно равны и близки к общему среднему рангу.

Диалоговое окно непараметрических критериев для нескольких независимых выборок представлено на рис. 12.17.

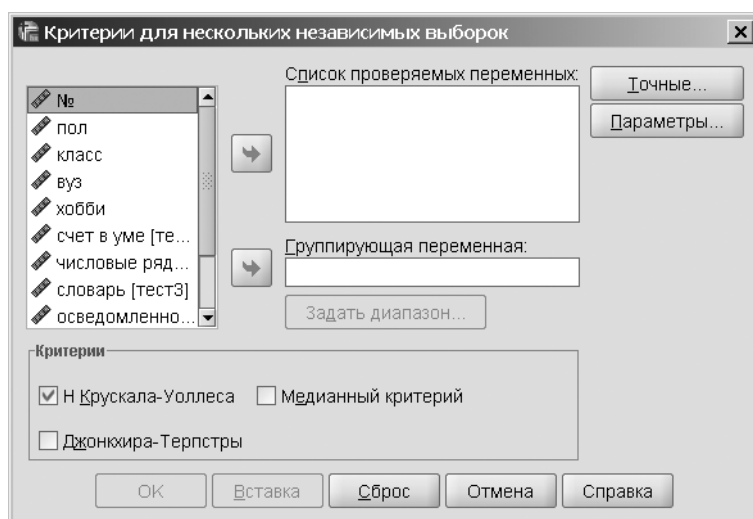


Рис. 12.17. Диалоговое окно Критерии для нескольких независимых выборок

После завершения шага 3 на экране должно присутствовать окно редактора данных. В этом примере мы проведем сравнение трех групп учащихся, отличающихся внешкольными увлечениями (переменная хобби) и успеваемостью в выпускном классе (переменная отметка2).

ШАГ 4Ж

В меню Анализ выберите команду Непараметрические критерии ► Устаревшие диалоговые окна ► Для К независимых выборок, чтобы открыть диалоговое окно Критерии для нескольких независимых выборок, показанное на рис. 12.17.

ШАГ 5Ж

Для применения критерия выполните следующую последовательность действий.

1. Щелкните сначала на переменной *отметка2*, чтобы выделить ее, а затем — на верхней кнопке со стрелкой, чтобы переместить переменную в Список проверяемых переменных.
2. Щелкните сначала на переменной *хобби*, чтобы выделить ее, а затем — на нижней кнопке со стрелкой, чтобы переместить переменную в поле Группирующая переменная.
3. Щелкните на кнопке *Задать диапазон*, чтобы открыть одноименное диалоговое окно.
4. В поле *Минимум* введите значение 1, нажмите клавишу *Tab*, чтобы переместить фокус ввода в поле *Максимум*, введите значение 3 и щелкните на кнопке *Продолжить*, чтобы вернуться в диалоговое окно *Критерии для нескольких независимых выборок*.
5. Щелкните на кнопке *ОК*, чтобы открыть окно вывода.

Обратите внимание, что флажок *Н Крускала-Уоллеса* установлен по умолчанию, поэтому мы ничего не меняли для выбора критерия.

Результаты работы программы показаны на рис. 12.18.

Ранги			
		N	Средний ранг
отметка2	хобби спорт	33	38,18
	компьютер	37	54,28
	искусство	30	59,38
	Всего	100	

Статистики критерия ^{а,б}	
	отметка2
Хи-квадрат	9,437
ст.св.	2
Асимпт. знч.	,009

а. Критерий Крускала-Уоллеса

б. Группирующая переменная: хобби

Рис. 12.18. Фрагменты окна вывода после выполнения шага 5ж

В первой таблице для каждой группы представлена ее численность и средний ранг. Во второй таблице указано значение критерия χ^2 , число степеней свободы и уровень статистической значимости. Результаты обработки показывают статистически достоверную связь внешкольных увлечений учащихся с успеваемостью в выпускном классе.

Сравнение нескольких зависимых выборок и критерий Фридмана

Критерий Фридмана является непараметрическим аналогом однофакторного дисперсионного анализа для повторных измерений. Он позволяет проверять гипотезы

о различии более двух зависимых выборок (повторных измерений) по уровню выраженности изучаемой переменной. Критерий Фридмана может быть более эффективен, чем его метрический аналог однофакторный дисперсионный анализ в случаях повторных измерений изучаемого признака на небольших выборках и при отличии распределения от нормального.

Критерий Фридмана весьма сходен с критерием Крускала—Уоллеса и основан на ранжировании ряда повторных измерений для каждого объекта выборки. Затем вычисляется сумма рангов для каждого из условий (повторных измерений). Если выполняется статистическая гипотеза об отсутствии различий между повторными измерениями, можно ожидать примерного равенства сумм рангов для этих условий. Чем больше различаются зависимые выборки по изучаемому признаку, тем больше эмпирическое значение вычисляемого значения критерия χ^2 , по которому определяется p -уровень значимости.

Диалоговое окно команды для применения критерия Фридмана представлено на рис. 12.19.

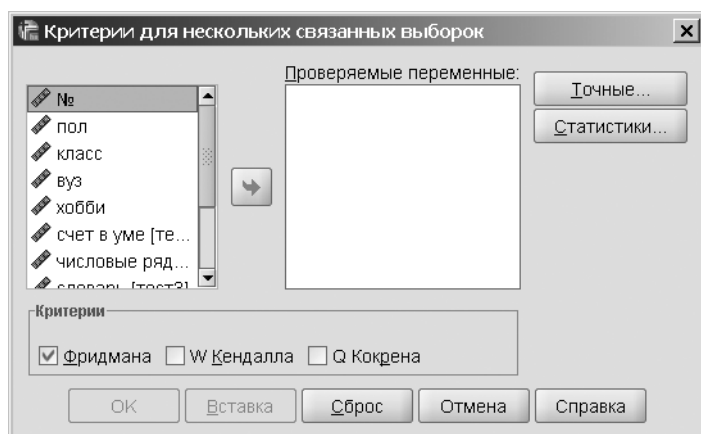


Рис. 12.19. Диалоговое окно критерия Фридмана

После завершения шага 3 на экране должно присутствовать окно редактора данных. В этом примере мы сравним результаты тестов тест1, тест2, тест3, тест4 и тест5 для всех учащихся.

ШАГ 43

В меню Анализ выберите команду Непараметрические критерии ► Устаревшие диалоговые окна ► Для К связанных выборок, чтобы открыть диалоговое окно Критерии для нескольких связанных выборок, показанное на рис. 12.19.

ШАГ 53

Для применения критерия выполните следующую последовательность действий.

1. Щелкните сначала на переменной тест1, чтобы выделить ее, а затем — на кнопке со стрелкой, чтобы переместить переменную в список Проверяемые переменные.

2. Повторите предыдущее действие для переменных тест2, тест3, тест4 и тест5.
3. Щелкните на кнопке ОК, чтобы открыть окно вывода.

Результаты работы программы показаны на рис. 12.20.

Ранги	
	Средний ранг
счет в уме	2,50
числовые ряды	2,65
словарь	3,41
осведомленность	3,08
кратковременная память	3,37

Статистики критерия ^а	
N	100
Хи-квадрат	29,696
ст.св.	4
Асимпт. знч.	,000

а. Критерий Фридмана

Рис. 12.20. Фрагменты окна вывода после выполнения шага 5з

Средние ранги определяются следующим образом: сначала для каждого наблюдения значения сравниваемых переменных ранжируются (по строке). Затем для каждой из сравниваемых переменных вычисляется средний ранг по всем объектам. Определяемый по критерию χ^2 уровень значимости Асимпт. знч. $< 0,001$. Он свидетельствует о статистически значимой разнице между пятью результатами тестирования. Различаться может любая пара переменных, и без попарного сравнения невозможно выяснить, какие именно пары вносят значимый вклад в факт статистической достоверности результата.

Завершение анализа и выход из программы

Отредактируйте содержимое окна вывода в соответствии со своими предпочтениями: скройте лишнюю информацию, исправьте таблицы и пр. (см. раздел «Окно вывода и его редактирование» главы 2).

Для дальнейшего использования окончательного результата все содержимое окна вывода или его фрагменты можно сохранить в файле *.spv, экспортировать в другой формат (например, Word), перенести в документ Word или вывести на печать (подробности см. в разделе «Сохранение, экспорт, перенос и печать результатов» главы 2).

Для выхода из программы выберите команду Выход в меню Файл.

13 Однофакторный дисперсионный анализ

185	Пошаговые алгоритмы вычислений
191	Представление результатов
194	Терминология
195	Завершение анализа и выход из программы

Однофакторный дисперсионный анализ в SPSS реализуется с помощью команды Однофакторный дисперсионный анализ. Команды подменю Общая линейная модель, описываемые в главах 14 и 15, также отчасти позволяют проводить подобный анализ, однако их возможности несколько более специфичны, чем у команды Однофакторный дисперсионный анализ.

Дисперсионный анализ (Analysis Of Variances, ANOVA — общепринятое обозначение метода) — это процедура сравнения средних значений выборок, на основании которой можно сделать вывод о соотношении средних значений генеральных совокупностей. Ближайшим и более простым аналогом ANOVA является t -критерий, применение которого было рассмотрено в главе 11. В отличие от t -критерия дисперсионный анализ предназначен для сравнения не двух, а нескольких выборок. Слово «дисперсионный» в названии указывает на то, что в процессе анализа сопоставляются компоненты дисперсии изучаемой переменной. Общая изменчивость переменной раскладывается на две составляющие: межгрупповую (факторную), обусловленную различием групп (средних значений), и внутригрупповую (ошибки), обусловленную случайными (неучтенными) причинами. Чем больше частное от деления межгрупповой и внутригрупповой изменчивости (F -отношение), тем больше различаются средние значения сравниваемых выборок и тем выше статистическая значимость этого различия.

Итак, название указывает на то, что вывод о различии средних значений делается на основе анализа компонентов дисперсии. А что означает слово «однофакторный»? Применяя команду Однофакторный дисперсионный анализ, вы увидите, что в ней можно задать единственную зависимую переменную (при этом она обязательно должна быть количественного, а точнее метрического типа) и единственную независимую переменную (всегда номинальную, имеющую несколько градаций). Различные модели дисперсионного анализа, описанные в двух следующих главах, допускают наличие нескольких независимых переменных. Многомерный дисперсионный анализ, с которым вы столкнетесь в главах 15 и 16, позволяет анализировать как множество независимых, так и множество зависимых переменных.

Итак, при однофакторном дисперсионном анализе сравниваются между собой средние значения каждой выборки и вычисляется общий уровень значимости различий. Обратите внимание, что вывод по результатам ANOVA касается общего различия всех сравниваемых средних без конкретизации того, какие именно выборки различаются, а какие нет. Для идентификации пар выборок, отличающихся друг от друга средними значениями, используются апостериорные критерии парных сравнений (Post Hoc), а для более сложных сопоставлений — метод контрастов (Contrasts).

Пошаговые алгоритмы вычислений

Для проведения однофакторного дисперсионного анализа мы будем использовать файл ex01.sav. В роли зависимой переменной выступит переменная тест1, а независимая переменная класс разделит объекты на три выборки, средние значения которых мы будем сравнивать, но сначала необходимо выполнить три подготовительных шага. Эти шаги (шаги 1–3) позволят подготовить рабочий файл данных, запустить программу IBM SPSS Statistics 19 и открыть файл (в данном случае — файл ex01.sav). Пошаговые инструкции этого процесса приведены в главе 4 (с. 60), а подробные разъяснения — в главе 2.

Однофакторный дисперсионный анализ

После завершения шага 3 на экране должно присутствовать окно редактора данных со строкой меню.

ШАГ 4

В меню Анализ выберите команду Сравнение средних ► Однофакторный дисперсионный анализ. На экране появится диалоговое окно Однофакторный дисперсионный анализ, показанное на рис. 13.1.

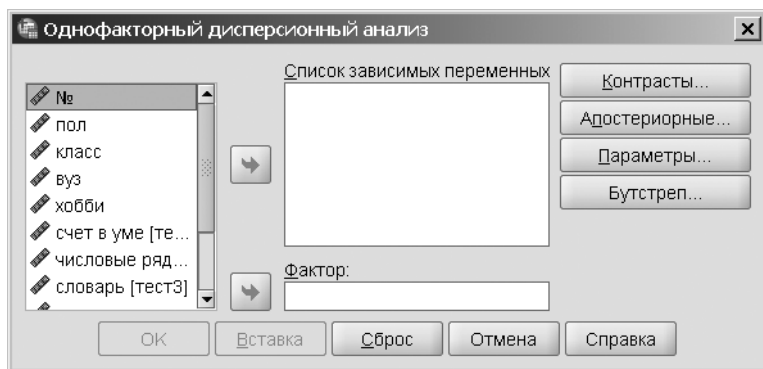


Рис. 13.1. Диалоговое окно Однофакторный дисперсионный анализ

Структура диалогового окна Однофакторный дисперсионный анализ вполне типична для большинства диалоговых окон SPSS. Слева мы видим список переменных текущего файла данных. В правой части окна расположены кнопки: Апостериорные, Контрасты и Параметры, которые мы будем использовать при обработке. Список зависимых переменных предназначен для задания одной или нескольких зависимых переменных (в нашем примере будет использоваться единственная зависимая переменная тест1). Зависимые переменные должны быть метрического типа. Если в списке указано несколько зависимых переменных, SPSS выполнит анализ для каждой из них. Под окном Список зависимых переменных находится поле Фактор, в котором нужно указать единственную независимую переменную, имеющую несколько градаций (в нашем случае — хобби). Таким образом, мы сравним результаты первого теста («счет в уме») для трех групп учащихся, различающихся внешкольными увлечениями.

На шаге 5 мы будем сравнивать между собой средние значения переменной тест1 для каждой из выборок по уровням переменной хобби.

ШАГ 5

После выполнения шага 4 должно быть открыто диалоговое окно Однофакторный дисперсионный анализ, показанное на рис. 13.1. При необходимости повторите шаг 4 и выполните следующие действия.

1. Щелкните сначала на переменной тест1, чтобы выделить ее, а затем — на верхней кнопке со стрелкой, чтобы переместить переменную в Список зависимых переменных.
2. Щелкните сначала на переменной хобби, чтобы выделить ее, а затем — на нижней кнопке со стрелкой, чтобы переместить переменную в поле Фактор.
3. Щелкните на кнопке ОК, чтобы открыть окно вывода.

Существует два дополнительных действия, которые иногда желательно выполнять в процессе анализа. Приведенная последовательность инструкций позволяет получить результаты сравнения средних значений выборок, однако ни сами средние значения, ни результаты парного сравнения выборок в выводимых данных отображены не будут. Первая проблема решается при помощи кнопки Параметры. Диалоговое окно Однофакторный дисперсионный анализ: Параметры, появляющееся после щелчка на этой кнопке, показано на рис. 13.2. Установка флажка Описательные статистики приведет к включению в выводимые данные всех средних значений, стандартных отклонений, стандартных ошибок, границ доверительных интервалов в 95 %, а также минимумов и максимумов выборок. Флажок Проверка однородности дисперсии позволяет вывести информацию о степени пригодности данных к дисперсионному анализу, а с помощью флажка График средних можно построить диаграмму, на которой будут изображены средние значения для каждой выборки.

Парные сравнения

Нередко нас могут заинтересовать результаты парных сравнений уровней независимой переменной, и для этой цели в диалоговом окне Однофакторный дисперсионный

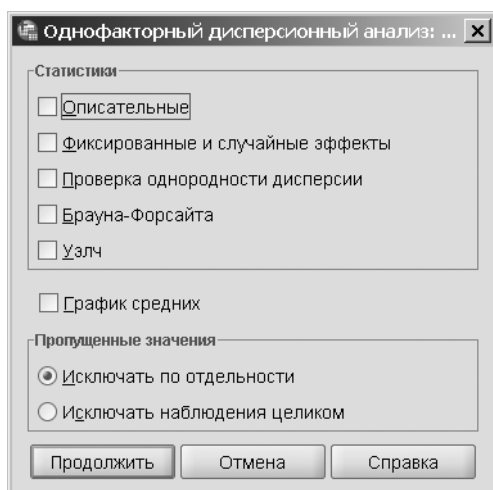


Рис. 13.2. Диалоговое окно Однофакторный дисперсионный анализ: Параметры

анализ предусмотрена специальная кнопка Апостериорные. Слово «апостериорные» означает, что эта процедура проводится после установления статистически достоверного результата однофакторного дисперсионного анализа. Если результаты дисперсионного анализа оказались статистически недостоверными, применение процедуры парных сравнений некорректно. При щелчке на кнопке Апостериорные открывается диалоговое окно Однофакторный дисперсионный анализ: Апостериорные множественные сравнения, приведенное на рис. 13.3. Это диалоговое окно с помощью флажков позволяет задать 14 критериев для выборок с одинаковой дисперсией и 4 критерия для выборок с разной дисперсией.

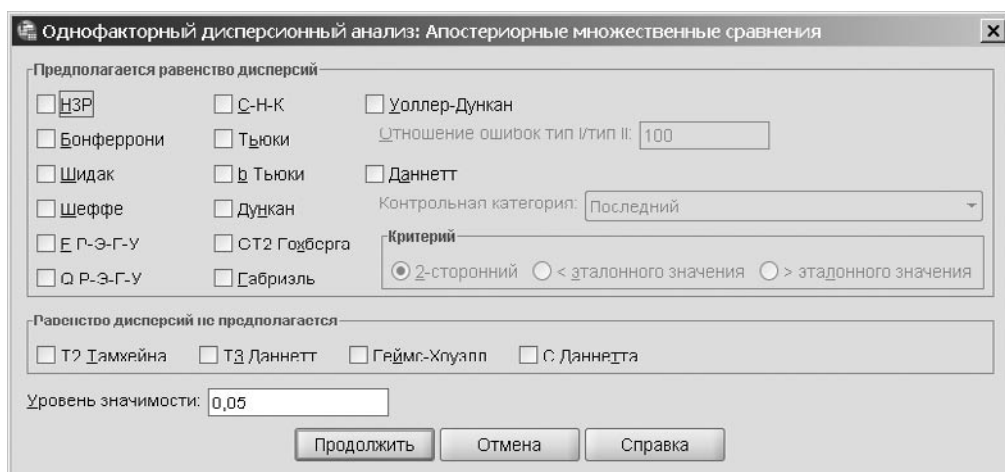


Рис. 13.3. Диалоговое окно Однофакторный дисперсионный анализ: Апостериорные множественные сравнения

Большинство из указанных тестов используются очень редко, поэтому ниже приведены описания только для нескольких наиболее популярных.

- ▶ НЗР (Наименьшая значимая разность) — этот критерий представляет собой совокупность t -критериев для всех возможных пар градаций фактора. Критерий наименьшей значимой разности является одним из самых «либеральных», поскольку наиболее подвержен ошибкам. Например, если независимая переменная имеет 5 уровней, то будет проведено 10 сравнений. При уровнях значимости каждого из сравнений, равных 0,05, существует вероятность почти в 40 % того, что хотя бы один из тестов показал значимый результат случайно.
- ▶ Бонферрони — этот критерий схож с критерием наименьшей значимой разности, однако лишен недостатка, связанного с повторными проверками: в нем уровень значимости делится на число сравнений. Таким образом, критерий Бонферрони является более «консервативным».
- ▶ Шеффе — критерий еще более «консервативный», чем критерий Бонферрони (используется F -критерий вместо t -критерия).
- ▶ Тьюки — критерий Тьюки использует статистику Стьюдента для определения различий между группами. Этот критерий часто применяется в случаях, когда исследуемый фактор имеет большое количество уровней.

Самыми консервативными из предложенных являются критерии Шеффе и Бонферрони. Часто используется также критерий Тьюки, называемый еще критерием подлинной значимости (Honestly Significant Difference, HSD). Последний задает наименьшую величину разности средних значений выборок, которую можно считать значимой. Например, если $HSD = 2,5$, а для двух выборок получены величины средних значений 3,7 и 6,3, то разность между ними, равная 2,6, согласно Тьюки, является значимой, поскольку она превышает величину HSD. При использовании критерия Тьюки программа SPSS также включает в вывод дополнительную статистическую информацию.

Необходимо отметить, что все описанные выше критерии (на самом деле, большинство критериев парных сравнений) применяются в предположении, что дисперсии всех ячеек равны. Исключение составляют 4 критерия, флажки для которых выделены в отдельную группу в нижней части диалогового окна Равенство дисперсий не предполагается, — они применяются в случаях, когда дисперсии ячеек разные.

В следующем примере мы проведем однофакторный дисперсионный анализ, а в выводимые результаты включим описательные статистики и критерий однородности дисперсии. Для парных сравнений воспользуемся критерием Шеффе. Как и в предыдущем примере, зависимой переменной будет переменная `тест1`, а независимой — переменная `хобби`.

ШАГ 5А

После выполнения шага 4 должно быть открыто диалоговое окно Однофакторный дисперсионный анализ, показанное на рис. 13.1. При необходимости повторите шаг 4 и выполните следующие действия.

1. Щелкните сначала на переменной `тест1`, чтобы выделить ее, а затем — на верхней кнопке со стрелкой, чтобы переместить переменную в Список зависимых переменных.

2. Щелкните сначала на переменной хобби, чтобы выделить ее, а затем — на нижней кнопке со стрелкой, чтобы переместить переменную в поле Фактор.
3. Щелкните на кнопке Параметры, чтобы открыть диалоговое окно Однофакторный дисперсионный анализ: Параметры, показанное на рис. 13.2.
4. Установите флажки Описательные, Проверка однородности дисперсии и График средних, а затем щелкните на кнопке Продолжить, чтобы вернуться в диалоговое окно Однофакторный дисперсионный анализ.
5. Щелкните на кнопке Апостериорные, чтобы открыть диалоговое окно Однофакторный дисперсионный анализ: Апостериорные множественные сравнения), показанное на рис. 13.3.
6. Установите флажок Шеффе и щелкните на кнопке Продолжить, чтобы вернуться в диалоговое окно Однофакторный дисперсионный анализ.
7. Щелкните на кнопке ОК, чтобы открыть окно вывода.

Контрасты

Верхняя кнопка в правой части окна Однофакторный дисперсионный анализ имеет название Контрасты. Она предназначена для вызова диалогового окна Однофакторный дисперсионный анализ: Контрасты, представленного на рис. 13.4. Это окно позволяет осуществлять различные сравнения выборок по градациям независимой переменной. Так, вы можете сравнивать одну градацию с другой, одну градацию со всеми остальными или разбить все градации на две группы и затем сравнить их между собой. Сравнение сводится к применению модифицированного варианта t -критерия. Отметим, что применение метода контрастов не требует предварительного получения статистически достоверного результата анализа ANOVA, в отличие от процедуры апостериорного множественного парного сравнения.

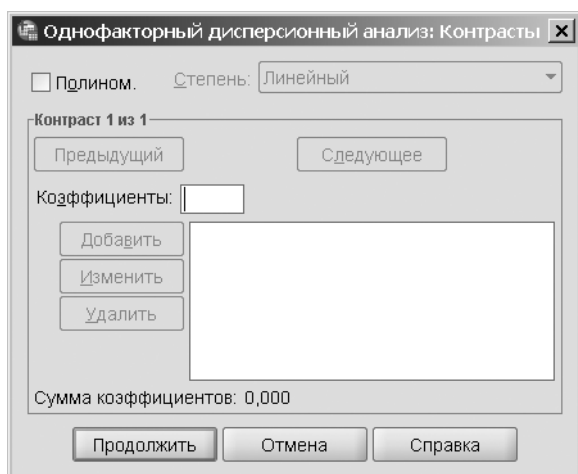


Рис. 13.4. Диалоговое окно Однофакторный дисперсионный анализ: Контрасты

В нашем примере независимая переменная хобби имеет три градации: 1 — спорт, 2 — компьютер, 3 — искусство. Для задания контраста предназначены поле и список Коэффициенты. Заполнение списка происходит следующим образом. Каждой градации фактора вы должны сопоставить число, определяющее его роль в контрасте: отрицательное число соответствует одной группе, положительное число — другой группе, а ноль означает, что градация в сравнениях не задействована. При этом абсолютные величины коэффициентов неважны, но последние должны вводиться в порядке следования градаций и в сумме давать нулевое значение. Например, если вам необходимо сравнить увлекающихся спортом с остальными учащимися, вы можете закодировать градации последовательностью -2 , 1 и 1 , а если вы хотите сравнить увлекающихся спортом только с теми, кто увлекается искусством, необходимо задать последовательность 1 , 0 и -1 . Не забывайте, что сумма всех коэффициентов обязательно должна равняться 0 .

Помещение чисел в список осуществляется путем ввода значения в поле справа от названия списка и щелчка на кнопке Добавить. Если нужно создать несколько «контрастов», то есть разбиений на группы, щелкните на кнопке Следующее справа от метки Контраст 1 из 1 и повторите описанную процедуру.

Следующий пример иллюстрирует применение контрастов: для сравнения учащихся, увлекающихся компьютером, с теми, кто имеет другие увлечения, а также для сравнения увлекающихся спортом с остальными учащимися.

ШАГ 5Б

После выполнения шага 4 должно быть открыто диалоговое окно Однофакторный дисперсионный анализ, показанное на рис. 13.1. Если вы уже успели поработать с этим окном, очистите его щелчком на кнопке Сброс и выполните следующие действия.

1. Щелкните сначала на переменной тест1, чтобы выделить ее, а затем — на верхней кнопке со стрелкой, чтобы переместить переменную в список Зависимые переменные.
2. Щелкните сначала на переменной хобби, чтобы выделить ее, а затем — на нижней кнопке со стрелкой, чтобы переместить переменную в поле Фактор.
3. Щелкните на кнопке Параметры, чтобы открыть диалоговое окно Однофакторный дисперсионный анализ: Параметры, показанное на рис. 13.2.
4. Установите флажки Описательные, Проверка однородности дисперсии и График средних, а затем щелкните на кнопке Продолжить, чтобы вернуться в диалоговое окно Однофакторный дисперсионный анализ.
5. Щелкните на кнопке Апостериорные, чтобы открыть диалоговое окно Однофакторный дисперсионный анализ: Апостериорные множественные сравнения, показанное на рис. 13.3.
6. Установите флажок НЗР и щелкните на кнопке Продолжить, чтобы вернуться в диалоговое окно Однофакторный дисперсионный анализ.
7. Щелкните на кнопке Контрасты, чтобы открыть диалоговое окно Однофакторный дисперсионный анализ: Контрасты, представленное на рис. 13.4.

8. Нажмите клавишу Tab, чтобы перевести фокус ввода в поле Коэффициенты, введите число 1 и щелкните на кнопке Добавить, задав первое число в списке.
9. Повторите предыдущее действие сначала для чисел -2 и 1, затем, щелкнув на кнопке Следующий, — для чисел -2, 1, 1, после чего щелкните на кнопке Продолжить, чтобы вернуться в диалоговое окно Однофакторный дисперсионный анализ.
10. Щелкните на кнопке ОК, чтобы открыть окно вывода.

После выполнения шага 5 программа автоматически активизирует окно вывода. Для просмотра результатов вы при необходимости можете воспользоваться вертикальной и горизонтальной полосами прокрутки. Обратите внимание на стандартную строку меню в верхней части окна вывода: ее присутствие позволяет выполнять любые статистические операции, не переключаясь обратно в окно редактора данных.

Представление результатов

В этом разделе мы приводим некоторые результаты процедур, описываемых в этой главе.

Однофакторный дисперсионный анализ

На рис. 13.5 приведен фрагмент выводимых данных, сгенерированных программой после выполнения шага 5а, относящийся к однофакторному дисперсионному анализу.

Однофакторный дисперсионный анализ

счет в уме

	Сумма квадратов	Ст. св.	Средний квадрат	F	Знач.
Между группами	82,969	2	41,484	6,556	,002
Внутри групп	613,781	97	6,328		
Итого	696,750	99			

Рис. 13.5. Фрагмент окна вывода после выполнения шага 5а

Самым важным в этой таблице является уровень значимости $p = 0,002$. Он указывает на то, что разность между средними значениями переменной тест1 для трех групп статистически достоверна.

Описательные статистики

На рис. 13.6 приведен фрагмент выводимых данных, сгенерированных программой после выполнения шага 5б, относящийся к описательным статистикам.

Описательные статистики

счет в уме								
	N	Среднее	Стд. отклонение	Стд. ошибка	95% доверительный интервал для среднего		Минимум	Максимум
					Нижняя граница	Верхняя граница		
спорт	33	9,67	2,582	,449	8,75	10,58	4	14
компьютер	37	11,43	2,102	,346	10,73	12,13	7	17
искусство	30	9,43	2,885	,527	8,36	10,51	5	15
Итого	100	10,25	2,653	,265	9,72	10,78	4	17

Рис. 13.6. Фрагмент окна вывода после выполнения шага 5б, относящийся к описательным статистикам

Апостериорные парные сравнения

На рис. 13.7 приведен фрагмент выводимых данных, сгенерированных программой после выполнения шага 5а, относящийся к парным сравнениям.

Средние значения переменной тест1 для каждой из трех выборок были перечислены в предыдущей таблице (см. рис. 13.6), здесь же даны разности между этими значениями. Знаком звездочки помечены те пары выборок, для которых разность средних значений статистически достоверна, то есть со значением уровня значимости 0,05 и меньше. Из полученных данных можно сделать вывод, что результаты теста 1 для тех, кто увлекается компьютером, статистически выше значимы, чем для тех, кто увлекается спортом и искусством. Те же, кто увлекаются спортом и искусством, по результатам теста 1 статистически достоверно не различаются.

Множественные сравнения

Зависимая переменная: счет в уме
Шеффе

(I) хобби	(J) хобби	Разность средних (I-J)	Стд. ошибка	Энч.	95% доверительный интервал	
					Нижняя граница	Верхняя граница
спорт	компьютер	-1,766*	,602	,016	-3,26	-,27
	искусство	,233	,635	,935	-1,34	1,81
компьютер	спорт	1,766*	,602	,016	,27	3,26
	искусство	1,999*	,618	,007	,46	3,54
искусство	спорт	-,233	,635	,935	-1,81	1,34
	компьютер	-1,999*	,618	,007	-3,54	-,46

*. Разность средних значима на уровне .05.

Рис. 13.7. Фрагмент окна вывода после выполнения шага 5а, относящийся к парным сравнениям

Критерий Ливиня

На рис. 13.8 приведен фрагмент выводимых результатов, сгенерированных программой после выполнения шага 5б (п. 4), относящийся к критерию однородности дисперсии Ливиня.

Критерий однородности дисперсии Ливиня со значимостью 0,161 показал, что дисперсии для каждой из групп статистически достоверно не различаются. Следова-

вательно, результаты ANOVA могут быть признаны корректными. Если бы результат применения критерия Ливиня оказался статистически достоверным, то это послужило бы основанием для сомнения в корректности применения ANOVA.

Критерий однородности дисперсий

счет в уме			
Статистика Ливиня	Ст. св. 1	Ст. св. 2	Знч.
1,858	2	97	,161

Рис. 13.8. Фрагмент окна вывода после выполнения шага 5б, относящийся к критерию Ливиня

Контрасты

На рис. 13.9 приведены фрагменты выводимых результатов, сгенерированных программой после выполнения шага 5б, относящиеся к применению метода контрастов.

Первая из таблиц, приведенных на рисунке, содержит коэффициенты, введенные при группировании уровней. Для каждого контраста было применено по два t -критерия, один из которых проводился с допущением о равных дисперсиях, а другой — с допущением о не равных дисперсиях. Первый контраст оказался статистически достоверным: те, кто увлекается компьютером, имеют статистически достоверно более высокий результат по переменной тест1. Второй контраст не достигает статистической значимости: учащиеся, увлекающиеся спортом, статистически достоверно не отличаются от других учащихся по переменной тест1.

Коэффициенты контраста

Контраст	хобби		
	спорт	компьютер	искусство
1	1	-2	1
2	-2	1	1

Критерии контраста

Контраст			Значение контраста	Стд. ошибка	t	Ст. св.	Знч. (2-сторон)
счет в уме	Предполагается равенство дисперсий	1	-3,76	1,042	-3,611	97	,000
		2	1,53	1,072	1,430	97	,156
	Равенство дисперсий не предполагается	1	-3,76	,978	-3,848	89,220	,000
		2	1,53	1,098	1,396	61,897	,168

Рис. 13.9. Фрагменты окна вывода после выполнения шага 5б, относящиеся к применению метода контрастов

Терминология

- ▶ Сумма квадратов строки Между группами означает сумму квадратов разностей между общим средним значением и средними значениями каждой группы, умноженными на весовые коэффициенты, равные числу объектов в группе, а строки Внутри групп — сумму квадратов разностей среднего значения каждой группы и каждого значения этой группы.
- ▶ Ст.св. строки Между группами означает межгрупповое число степеней свободы, равное числу групп, уменьшенному на 1, в строке Внутри групп — внутригрупповое число степеней свободы, равное разности между числом объектов и числом групп.
- ▶ Средний квадрат — отношение суммы квадратов к числу степеней свободы.
- ▶ F — F -критерий, отношение среднего квадрата между группами к среднему квадрату внутри группы.
- ▶ Знч. — статистическая значимость, вероятность того, что наблюдаемые различия случайны. Величина значимости $p = 0,002$ свидетельствует о статистически достоверных различиях.
- ▶ N — число объектов для каждой из градаций переменной хобби.
- ▶ Стд. отклонение — стандартное отклонение, мера разброса значений распределения относительно среднего.
- ▶ Стд. ошибка — стандартная ошибка среднего, отношение стандартного отклонения к квадратному корню из числа объектов.
- ▶ 95% доверительный интервал для среднего — при большом числе выборок из генеральной совокупности 95 % средних значений этих выборок попадут в интервал, определяемый указанными в таблице границами.
- ▶ Минимум — наименьшее из наблюдаемых значений для группы.
- ▶ Максимум — наибольшее из наблюдаемых значений для группы.
- ▶ Значение контраста — это значение не представляет интереса для исследователя, поскольку является всего лишь произведением разности средних с весовым коэффициентом.
- ▶ t — значение t -критерия, отношение величины различия средних к стандартной ошибке.
- ▶ Знч. (2-сторон) — двусторонний уровень значимости. Вероятность того, что отличие от нуля является случайным.

Завершение анализа и выход из программы

Отредактируйте содержимое окна вывода в соответствии со своими предпочтениями: скройте лишнюю информацию, исправьте таблицы и пр. (см. раздел «Окно вывода и его редактирование» главы 2).

Для дальнейшего использования окончательного результата все содержимое окна вывода или его фрагменты можно сохранить в файле *.spv, экспортировать в другой формат (например, Word), перенести в документ Word или вывести на печать (подробности см. в разделе «Сохранение, экспорт, перенос и печать результатов» главы 2).

Для выхода из программы выберите команду Выход в меню Файл.

14 Многофакторный дисперсионный анализ

197	Файлы данных для группы методов	Общая линейная модель
198	Дисперсионный анализ с двумя факторами	
200	Дисперсионный анализ с тремя и более факторами	
201	Влияние ковариат	
201	Пошаговые алгоритмы вычислений	
207	Представление результатов	
211	Терминология, используемая при выводе	
211	Завершение анализа и выход из программы	

В этой главе речь пойдет о дисперсионном анализе (ANOVA) с двумя и более факторами. В нем участвует единственная зависимая переменная, которая, как и при однофакторном дисперсионном анализе, должна быть метрической. Главное отличие многофакторного анализа от однофакторного заключается в том, что здесь участвуют не одна, а несколько независимых переменных, каждая из которых должна быть номинальна, то есть иметь несколько градаций, или уровней. Многофакторный дисперсионный анализ реализуется в SPSS с помощью команды **Общая линейная модель** ► **ОЛМ-одномерная**.

Дисперсионному анализу посвящены главы 13–16. Мы настоятельно рекомендуем читать эти главы по порядку, поскольку в них в развитии описывается одна и та же статистическая концепция. Представление об однофакторном анализе является основой для понимания дисперсионного анализа с двумя и более факторами, который, в свою очередь, позволяет перейти к более сложным моделям дисперсионного анализа. Отметим, что дисперсионный анализ является очень сложным разделом статистики, поэтому в этой книге мы приводим лишь самые необходимые сведения, оставляя для самостоятельного изучения все специфические вопросы.

Программа SPSS предназначена для того, чтобы сделать работу исследователя как можно более простой и комфортной. Чтобы выполнить дисперсионный анализ с одним, двумя и даже семью факторами, достаточно сделать лишь несколько щелчков на нужных кнопках. Компьютер выполнит за вас всю вычислительную работу, и вы получите результат в наглядной и удобной форме, но, к сожалению, нередко неясный по своему содержанию. Простота и удобство вычислительного процесса подчас приводят к тому, что исследователи «расслабляются» и забывают простую истину: успех исследования определяется его продуманностью. Если од-

нофакторный дисперсионный анализ не представляет сложности для понимания, то анализ с двумя и более факторами требует очень внимательного отношения и определенных навыков в интерпретации результатов. Так, для интерпретации результатов трехфакторного дисперсионного анализа, как правило, необходимы немалый опыт и умение свободно обращаться со статистическими величинами. Что же касается четырехфакторного анализа, то его интерпретация чрезвычайно сложна даже для специалиста высокой квалификации.

Многофакторный дисперсионный анализ отличается от однофакторного прежде всего тем, что появляется новая *проблема взаимодействия факторов*. Решение этой проблемы принципиально не зависит от числа факторов. С возрастанием числа факторов существенно нарастает лишь сложность интерпретации взаимодействий.

В этой главе мы выполним двух- и трехфакторный дисперсионный анализ с учетом влияния ковариаты. Мы также воспользуемся графическими средствами интерпретации средних значений и, как всегда, в конце главы проведем разбор полученных результатов.

Файлы данных для группы методов Общая линейная модель

В главах 14–16 рассматриваются сложные варианты дисперсионного анализа (ANOVA), которые реализуются тремя командами в группе методов Общая линейная модель: ОЛМ-одномерная, ОЛМ-многомерная и ОЛМ-повторные измерения. Для примеров используются три файла данных, содержащих одни и те же данные, полученные в ходе одного и того же исследования, но отличающиеся только структурой: ex020.sav, ex021.sav и ex022.sav. Каждый из этих файлов может быть реструктурирован в два других при помощи команды Реструктурировать, как это описано в разделе «Реструктурирование данных» главы 4.

Все три файла данных содержат гипотетические результаты эксперимента по изучению эффективности запоминания слов в зависимости от частоты их встречаемости и от интонации, с которой они предъявлялись (зачитывались). Ряды из 24 не связанных по смыслу слов одинаковой длины зачитывались 20 испытуемым. Сразу после предъявления испытуемых просили воспроизвести эти слова. Подсчитывалось количество правильно воспроизведенных слов из начала ряда — первых 8 слов, из середины ряда и из конца ряда — завершающих 8 слов. Через два дня испытуемым предлагалось вспомнить предъявляемые слова, и снова подсчитывалось количество правильно воспроизведенных слов из начала, середины и конца ряда.

Все три файла идентичны по следующим переменным.

N — номер испытуемого.

- Номинальная переменная Инт соответствует делению испытуемых на две группы: первой (Инт = 1) все слова читались с одинаковой интонацией; второй (Инт = 2) середина ряда интонационно выделялась.

- Количественная переменная *Знач* отражает эмоциональную значимость предъявляемого ряда слов: каждый предъявляемый ряд составлялся из слов, и ряды различались и по этой переменной.

Различия между файлами.

- Файл *ex020.sav* включает 6 переменных продуктивности воспроизведения слов из трех частей ряда без отсрочки (*Начало1*, *Середина1*, *Конец1*) и с отсрочкой в два дня (*Начало2*, *Середина2*, *Конец2*). Соответственно, число строк этого файла соответствует численности испытуемых ($N = 20$).

Файл *ex021.sav* включает 3 переменных продуктивности воспроизведения слов из трех частей ряда (*Начало*, *Середина*, *Конец*). Плюс введена номинальная переменная *Отсрочка*, «удваивающая» численность строк файла: *Отсрочка* = 1 – воспроизведение без отсрочки, *Отсрочка* = 2 – воспроизведение с отсрочкой. Таким образом, число строк файла равно 40 ($2 \times N$), а значения переменных *N*, *инт* и *знач* повторяются дважды (для двух значений переменной *Отсрочка*).

Файл *ex022.sav* включает 1 переменную продуктивности воспроизведения слов *Слова*. Для их различения по части ряда введена переменная *Ч_ряда*: 1 – начало, 2 – середина, 3 – конец. А по отсрочке воспроизведения их делит переменная *Отсрочка*: 1 – без отсрочки, 2 – с отсрочкой. Таким образом, число строк этого файла равно 120 ($N \times 2 \times 3$), а значения переменных *N*, *инт* и *Инач* повторяются 6 раз (для двух значений переменной *Отсрочка* и трех значений переменной *Ч_ряда*).

Таким образом, данные каждого из трех файлов позволяют проверить одни и те же гипотезы о влиянии интонации и значимости слов на продуктивность их воспроизведения в зависимости от части ряда и от отсрочки. Но, в зависимости от структуры данных, для этого необходимо применить три разных варианта ANOVA:

- *ex022.sav* — ОЛМ: одномерный (многофакторный ANOVA), рассматриваемый в этой главе;
- *ex021.sav* — ОЛМ: многомерный (многомерный ANOVA, глава 15);
- *ex020.sav* — ОЛМ: повторные измерения (ANOVA с повторными измерениями, глава 16).

Дисперсионный анализ с двумя факторами

Как было показано в предыдущей главе, дисперсионный анализ (ANOVA) определяет статистическую достоверность различия между выборками путем сравнения их средних значений. Чтобы получить представление о двухфакторном ANOVA, сначала немного вернемся назад и обобщим наши знания об однофакторном ANOVA. Мы сравнивали три класса (переменная *класс*) по уровню выраженности переменной *тест1*. «Однофакторность» анализа заключалась в том, что деление на

группы производилось по грациям одной независимой переменной (класс). Для однофакторного ANOVA в SPSS существует специальная упрощенная команда Однофакторный дисперсионный анализ. Команды подменю Общая линейная модель позволяют выполнять однофакторный, двухфакторный, трехфакторный и т. д. анализы, однако обращение с ними несколько сложнее.

В этой главе мы будем использовать файл ex022.sav. В качестве зависимой переменной мы возьмем переменную Слова, а роль независимых будут играть переменные Инт и Ч_ряда. Мы попытаемся определить степень влияния переменных Инт, Ч_ряда и их взаимодействия Инт \times Ч_ряда на распределение значений переменной Слова. Такая схема анализа может быть лаконично обозначена как ANOVA 2 \times 3 (Интонация \times Часть ряда). Исследование позволит получить ответы на перечисленные ниже вопросы.

- ▶ Существует ли главный эффект фактора Инт, то есть существует ли значимое различие в продуктивности воспроизведения всего ряда из 24 слов в зависимости от интонационного выделения середины ряда и какова степень этого различия?
- ▶ Существует ли главный эффект фактора Ч_ряда, то есть существует ли значимое различие в продуктивности воспроизведения трех частей ряда (начала, середины и конца) и какова степень этого различия?
- ▶ Существует ли взаимодействие переменных Инт и Ч_ряда, то есть зависит ли влияние одной из этих переменных от уровней (значений, граций) другой?

Таким образом, двухфакторный дисперсионный анализ позволяет проверить три гипотезы: две о главном эффекте и одну о взаимодействии факторов. Ответы на два первых вопроса можно было бы получить, дважды применив однофакторный дисперсионный анализ. Специфика многофакторного анализа проявляется в содержании третьего вопроса, который касается взаимодействия факторов. Взаимодействие двух факторов означает, что влияние одного из них проявляется по-разному на разных уровнях другого фактора. Весьма полезным для интерпретации взаимодействий является построение диаграмм средних значений для каждой ячейки таблицы сопряженности независимых переменных. Мы займемся построением диаграмм в конце этой главы. На диаграммах будут представлены все значения вне зависимости от степени их значимости, что позволит нам более ясно представлять, почему воздействие одних переменных оказывается значимым, а других — нет. Более детально проблема взаимодействия будет рассмотрена на примере при обсуждении результатов.

SPSS позволяет включить в окно вывода средние значения всех выборок, соответствующих всем возможным сочетаниям граций факторов (в данном случае $2 \times 3 = 6$), а также вычислить F -величины и соответствующие p -уровни. По этим характеристикам мы сможем судить о степени влияния каждой из независимых переменных на распределение зависимой переменной и о взаимодействии независимых переменных.

Дисперсионный анализ с тремя и более факторами

Как уже отмечалось, проведение дисперсионного анализа с тремя и более факторами принципиально не отличается от двухфакторного анализа. При количестве факторов более двух возрастает лишь сложность интерпретации взаимодействий факторов. Эта сложность обусловлена появлением большого числа взаимодействий, часть из которых с трудом поддается интерпретации. Рассмотрим эту проблему на примере.

Предположим, изучается влияние на переменную Слова трех факторов: Инт, Ч_ряда и Отсрочка. Такая схема анализа может быть обозначена как ANOVA $2 \times 3 \times 2$ (Интонация \times Часть ряда \times Отсрочка). Применение трехфакторного дисперсионного анализа позволило бы получить ответы на следующие вопросы.

- ▶ Существует ли главный эффект фактора Инт, то есть существует ли значимое различие в продуктивности воспроизведения всего ряда из 24 слов в зависимости от интонационного выделения середины ряда и какова степень этого различия?
- ▶ Существует ли главный эффект фактора Ч_ряда, то есть существует ли значимое различие в продуктивности воспроизведения трех частей ряда (начала, середины и конца) и какова степень этого различия?
- ▶ Существует ли главный эффект фактора Отсрочка, то есть существует ли значимое различие в продуктивности воспроизведения всего ряда в зависимости от отсрочки?
- ▶ Существует ли взаимодействие переменных Инт и Ч_ряда, то есть зависит ли влияние одной из этих переменных от уровней (значений, градаций) другой?
- ▶ Существует ли взаимодействие переменных Инт и Отсрочка, то есть зависит ли влияние одной из этих переменных от градаций другой?
- ▶ Существует ли взаимодействие переменных Ч_ряда и Отсрочка, то есть зависит ли влияние одной из этих переменных от градаций другой?
- ▶ Существует ли взаимодействие переменных Инт, Ч_ряда и Отсрочка, то есть зависит ли взаимодействие двух из этих переменных от градаций третьей?

Таким образом, трехфакторный дисперсионный анализ предполагает проверку уже семи гипотез. Из них три гипотезы касаются взаимодействия двух факторов и одна — взаимодействия трех факторов. Если добавить еще один, четвертый фактор, появится взаимодействие 4-х факторов, а число проверяемых гипотез возрастет до 15. Для 5 факторов число проверяемых гипотез составит уже 31! Общее число проверяемых гипотез равно $2^P - 1$, где P — число факторов. При увеличении числа факторов быстро возрастает не только количество проверяемых гипотез, но и сложность интерпретации взаимодействий. Если интерпретации взаимодействий двух факторов обычно не составляют труда, то взаимодействия трех факторов обязательно требуют построения графиков средних значений, а интерпретации взаимодействий более высокого порядка вряд ли вообще возможны.

Влияние ковариат

Ковариаты используются для исключения влияния количественной переменной на зависимую переменную. Ковариату проще всего представить как переменную, значительно коррелирующую с зависимой переменной и позволяющую уменьшить ее дисперсию. За счет включения в анализ ковариаты дисперсия зависимой переменной уменьшается, что позволяет сделать более очевидным влияние анализируемых факторов.

В нашем исследовании в качестве ковариаты будет использоваться переменная *Знач.* Эта переменная (эмоциональная значимость предъявляемого ряда слов) в существенной степени коррелирует с продуктивностью воспроизведения этого ряда (*Слова*). Если мы хотим проследить влияние факторов *Инт* и *Ч_ряда* на зависимую переменную *Слова* без учета влияния фактора *Знач* (исключив его влияние), последний необходимо включить в анализ в качестве ковариаты. Появление ковариаты не влияет на описательные статистики, однако может изменить сумму квадратов (как правило, в меньшую сторону) и величину *F*-критерия (как в меньшую, так и, возможно, в большую сторону).

Пошаговые алгоритмы вычислений

Для проведения многофакторного дисперсионного анализа сначала необходимо выполнить три подготовительных шага. Эти шаги (шаги 1–3) позволят подготовить рабочий файл данных, запустить программу IBM SPSS Statistics 19 и открыть файл (в данном случае — файл *ex022.sav*). Пошаговые инструкции этого процесса приведены в главе 4 (с. 60), а подробные разъяснения — в главе 2.

После завершения шага 3 на экране должно присутствовать окно редактора данных со строкой меню и загруженным файлом *ex022*.

ШАГ 4

В меню Анализ выберите команду *Общая линейная модель ► ОЛМ-одномерная*. На экране появится диалоговое окно *ОЛМ-одномерная*, показанное на рис. 14.1.

Как вы, вероятно, обратили внимание, в меню *Общая линейная модель*, помимо команды *ОЛМ-одномерная*, находятся еще три команды: *ОЛМ-многомерная*, *ОЛМ-повторные измерения* и *Компоненты дисперсии*. Команда *ОЛМ-многомерная* используется для многомерного дисперсионного и ковариационного анализов и рассмотрена в главе 15. Команда *ОЛМ-повторные измерения* может применяться как для однофакторного, так и для многофакторного анализов, ее описание приведено в главе 16. Команда *Компоненты дисперсии* обладает несколько большей гибкостью по сравнению с предыдущими, однако ее сложность не позволяет рассказать о ней в рамках данной книги.

В левой части диалогового окна *ОЛМ-одномерная* расположен список всех доступных переменных файла данных, а справа находятся поля и кнопки, с помощью

которых задаются основные установки дисперсионного анализа. Поле Зависимая переменная предназначено для указания единственной зависимой переменной анализа, поле Фиксированные факторы — для имен номинальных независимых переменных, или факторов, поле Ковариаты — для указания имен ковариат, как дополнительных количественных факторов. Кнопки Контрасты и Апостериорные действуют идентично своим аналогам из диалогового окна Однофакторный дисперсионный анализ, поэтому при необходимости вернитесь к главе 13. Кнопка Модель позволяет задать пользовательскую модель анализа — на ваш выбор.

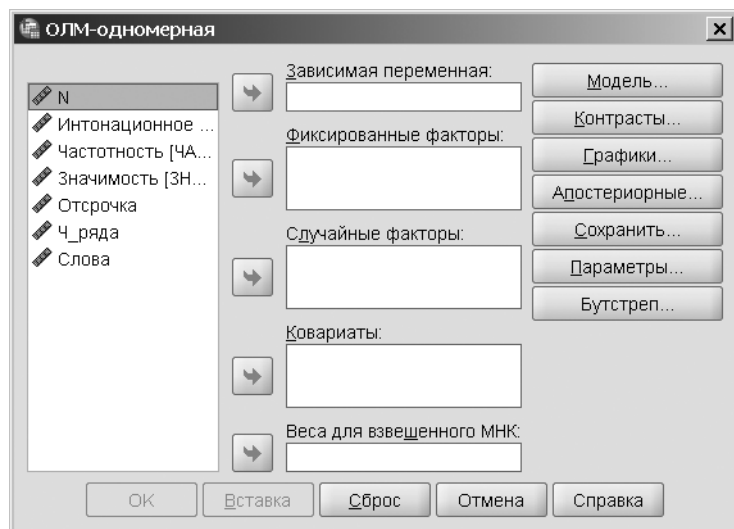


Рис. 14.1. Диалоговое окно ОЛМ-одномерная

Особое значение имеет кнопка Диаграммы, позволяющая строить графики средних значений. Для построения графиков в программе SPSS предусмотрено отдельное меню ОЛМ одномерная: Графики профилей, так как графики средних значений необходимы для интерпретации взаимодействий факторов.

Двухфакторный дисперсионный анализ

После задания зависимой переменной Слова необходимо задать независимые переменные, по очереди помещая их в поле Фиксированные факторы.

Когда имена всех переменных, участвующих в анализе, определены, можно перейти к заданию параметров выполняемых действий. Для этого щелкните на кнопке Параметры. На экране появится диалоговое окно ОЛМ: одномерная: Параметры, изображенное на рис. 14.2. Нас будет интересовать группа флажков Вывести, позволяющая задать список включаемых в вывод величин. Так, флажок Описательные статистики позволяет вывести для каждой ячейки таблицы сопряженности средние значения, стандартные отклонения и размеры выборок по всем градациям всех факторов. Нередко используется флажок Оценки размера эффекта, включающий

в выводимые данные величину эффекта (η^2) каждой из независимых переменных, а также их взаимодействий.

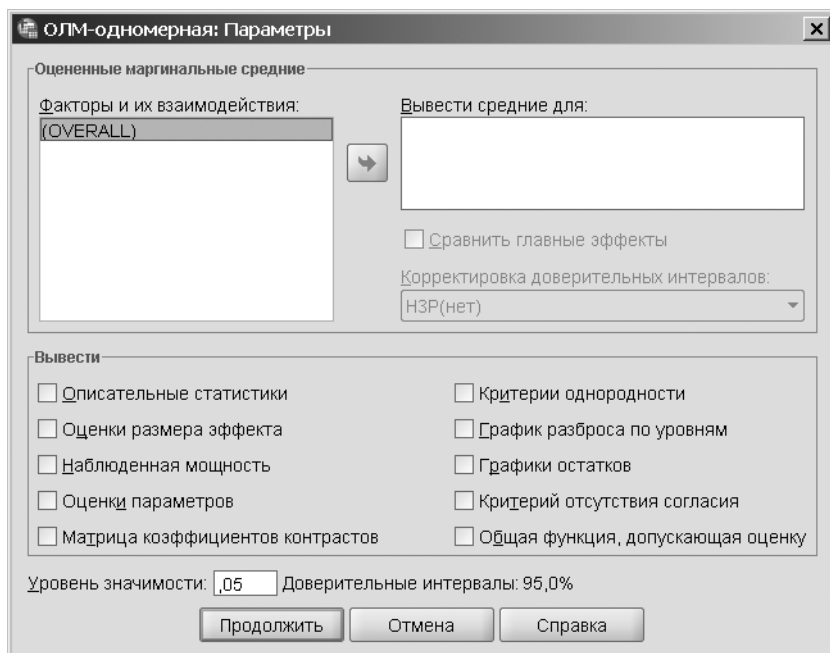


Рис. 14.2. Диалоговое окно ОЛМ: одномерная: Параметры

Следующий пример представляет собой практическую реализацию описанных действий.

ШАГ 5

После выполнения шага 4 должно быть открыто диалоговое окно ОЛМ-одномерная, показанное на рис. 14.1.

1. Щелкните сначала на переменной Слова, чтобы выделить ее, а затем — на верхней кнопке со стрелкой, чтобы переместить переменную в поле Зависимая переменная.
2. Щелкните сначала на переменной Интонационное выделение [Инт], чтобы выделить ее, а затем — на второй сверху кнопке со стрелкой, чтобы переместить переменную в список Постоянные факторы.
3. Повторите предыдущее действие для переменной Часть ряда [Ч_ряда].
4. Щелкните на кнопке Параметры, чтобы открыть диалоговое окно ОЛМ-одномерная: Параметры.
5. Установите флажки Описательные статистики, Оценка размера эффекта и Критерии однородности, а затем щелкните на кнопке Продолжить, чтобы вернуться в диалоговое окно ОЛМ-одномерная.
6. Щелкните на кнопке ОК, чтобы открыть окно вывода.

Влияние ковариаты

В следующем пошаговом алгоритме мы реализуем двухфакторный дисперсионный анализ с учетом ковариаты. Все переменные останутся теми же, что и на шаге 5, но дополнительно в качестве ковариаты включим переменную *Знач.* Это позволит сравнить результаты дисперсионного анализа с учетом и без учета влияния ковариаты.

ШАГ 5А

После выполнения шага 4 должно быть открыто диалоговое окно ОЛМ-одномерная, показанное на рис. 14.1. Если вы уже успели поработать с этим окном, щелкните на кнопке *Сброс*.

1. Щелкните сначала на переменной *Слова*, чтобы выделить ее, а затем — на верхней кнопке со стрелкой, чтобы переместить переменную в поле *Зависимая переменная*.
2. Щелкните сначала на переменной *Интонационное выделение [Инт]*, чтобы выделить ее, а затем — на второй сверху кнопке со стрелкой, чтобы переместить переменную в список *Постоянные факторы*.
3. Повторите предыдущее действие для переменной *Часть ряда [Ч_ряда]*.
4. Щелкните сначала на переменной *Знач*, чтобы выделить ее, а затем — на второй снизу кнопке со стрелкой, чтобы переместить переменную в список *Ковариаты*.
5. Щелкните на кнопке *Параметры*, чтобы открыть диалоговое окно ОЛМ-одномерная.
6. Установите флажки *Описательные статистики*, *Оценка размера эффекта* и *Критерии однородности*, а затем щелкните на кнопке *Продолжить*, чтобы вернуться в диалоговое окно ОЛМ-одномерная.
7. Щелкните на кнопке *ОК*, чтобы открыть окно вывода.

Результаты выполнения шага 5а будут отличаться от результатов выполнения шага 5 только учетом ковариаты *Знач*.

Графические средства интерпретации взаимодействий

В большинстве случаев для интерпретации взаимодействия факторов необходимо построение графиков средних значений. Для этого служит кнопка *Графики*, открывающая диалоговое окно ОЛМ-одномерная: *Графики профилей*, показанное на рис. 14.3. В этом окне можно задать параметры сразу нескольких графиков средних значений. Каждый график представляет собой ломаную линию (профиль), соединяющую точки — средние значения для групп. Горизонтальной оси соответствуют уровни одного из факторов, а величина средних значений откладывается по вертикальной оси.

Вид графиков задается тремя параметрами. В поле *Горизонтальная ось* вводится имя фактора, грациям которого будет соответствовать горизонтальная ось графика или графиков. В поле *Отдельные линии* указывается фактор, грациям которого будут соответствовать разные линии на графике. Поле *Отдельные графики*

служит для задания имени фактора, градациям которого будут соответствовать разные графики. Щелчок на кнопке **Добавить** подтверждает правильность введенных параметров и позволяет перейти к заданию графиков другого типа.

При проведении двухфакторного дисперсионного анализа, как в нашем примере, достаточно построить один график. Для этого надо сначала задать имя фактора для окна **Горизонтальная ось**. Обычно это тот фактор, который имеет больше градаций, — в нашем случае это фактор **Ч_ряда**. Затем необходимо задать имя фактора для окна **Отдельные линии**, — в нашем случае это фактор **Инт**. В результате будет построен график с двумя ломаными линиями, каждая из которых соединяет три точки. Вертикальная ось этого графика будет соответствовать величине средних значений, горизонтальная ось — переменной **Ч_ряда**, а две отдельные линии — переменной **Инт**.

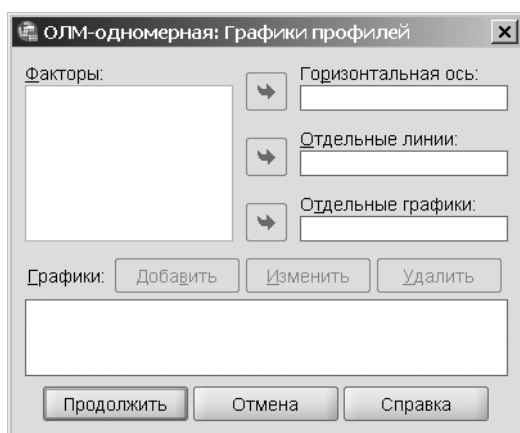


Рис. 14.3. Диалоговое окно ОЛМ-одномерная: Графики профилей

Для трехфакторного анализа подобный график позволит интерпретировать только двухфакторное взаимодействие. Для интерпретации трехфакторного взаимодействия необходимо градациям третьего фактора поставить в соответствие разные Отдельные графики. Например, если бы третьим фактором в нашем примере был фактор **Отсрочка**, то задание этого фактора как имени для разных графиков привело бы к построению двух графиков — по одному для двух условий (есть, нет), каждый из которых соответствовал бы двухфакторному взаимодействию переменных **Инт** и **Ч_ряда**.

Пошаговый алгоритм, приведенный ниже, повторяет действия предшествующего шага, но добавляет в окно вывода график средних значений, позволяющий интерпретировать взаимодействие факторов **Инт** и **Ч_ряда**. А шаг 6 демонстрирует построение альтернативного графика: столбиковой диаграммы средних значений.

ШАГ 5Б

После выполнения шага 4 должно быть открыто диалоговое окно **ОЛМ-одномерная**, показанное на рис. 14.1. Если вы уже успели поработать с этим окном, щелкните на кнопке **Сброс**.

1. Задайте переменные для анализа. Для этого щелкните на переменной **Слова**, чтобы выделить ее, а затем — на верхней кнопке со стрелкой,

чтобы переместить переменную в поле Зависимая переменная; щелкните на переменной Инт, чтобы выделить ее, а затем — на второй сверху кнопке со стрелкой, чтобы переместить переменную в список Фиксированные факторы. Повторите то же действие для переменной Ч_ряда. Щелкните на переменной Знач, чтобы выделить ее, а затем — на второй снизу кнопке со стрелкой, чтобы переместить переменную в список Ковариаты.

2. Задайте параметры результатов. Для этого щелкните на кнопке Параметры, чтобы открыть диалоговое окно ОЛМ-одномерная: Параметры, и установите флажки Описательные статистики, Критерии однородности, Оценка размера эффекта, а затем щелкните на кнопке Продолжить, чтобы вернуться в диалоговое окно ОЛМ-одномерная.
3. Щелкните на кнопке Графики, чтобы открыть диалоговое окно ОЛМ-одномерная: Графики профилей.
4. Выделите в списке Факторы переменную Ч_ряда и щелчком на верхней кнопке со стрелкой перенесите ее в поле Горизонтальная ось.
5. Выделите в списке Факторы переменную Инт и щелчком на второй сверху кнопке со стрелкой перенесите ее в поле Отдельные линии.
6. Подтвердите правильность введенных параметров щелчком на кнопке Добавить — в нижнем поле появится строка Ч_ряда*Инт, обозначающая тип графика средних.
7. Щелкните на кнопке Продолжить, чтобы вернуться в диалоговое окно ОЛМ-одномерная.
8. Щелкните на кнопке ОК, чтобы открыть окно вывода.

При желании можно построить альтернативные графики, например столбиковые, воспользовавшись меню Графика в командной строке.

ШАГ 6

На экране должно присутствовать окно редактора данных со строкой меню и загруженным файлом ex022.

1. В командной строке выберите Графика ► Устаревшие диалоговые окна ► Столбики.
2. В открывшемся окне Столбики выделите значок Кластеризованные и щелкните по кнопке Задать.
3. В открывшемся окне Кластеризованные столбики: Итожащие функции по группам наблюдений установите переключатель Другую статистику (например, среднее) и в окно Переменное перенесите из левого окна переменную Слова. В окно Категориальная ось перенесите переменную Интонация, а в окно Задать кластеры по — переменную Ч_ряда.
4. Щелкните на кнопке ОК, чтобы открыть окно вывода.

После выполнения шагов 5, 5а, 5б и 6 программа автоматически перейдет в окно вывода. Для просмотра результатов при необходимости можно воспользоваться вертикальной и горизонтальной полосами прокрутки. Обратите внимание на стандарт-

ную строку меню в верхней части окна вывода: ее присутствие позволяет выполнять любые статистические операции, не переключаясь обратно в окно редактора данных.

Представление результатов

Двухфакторный дисперсионный анализ

На рис. 14.4 приведены фрагменты выводимых данных, сгенерированных программой после выполнения шага 5.

Описательные статистики						
Зависимая переменная: Слова						
Интонационное выделение	Часть ряда	Среднее	Стд. Отклонение	N		
нет	Начало	3,50	1,573	20		
	Середина	3,10	1,021	20		
	Конец	3,75	1,585	20		
	Всего	3,45	1,419	60		
есть	Начало	2,65	1,424	20		
	Середина	4,30	1,261	20		
	Конец	2,80	,834	20		
	Всего	3,25	1,398	60		
Всего	Начало	3,08	1,542	40		
	Середина	3,70	1,285	40		
	Конец	3,28	1,339	40		
	Всего	3,35	1,406	120		

Оценка эффектов межгрупповых факторов						
Зависимая переменная: Слова						
Источник	Сумма квадратов типа III	ст.св.	Средний квадрат	F	Знач.	Частная Эта в Квадрате
Скорректированная модель	38,800 ^a	5	7,760	4,502	,001	,165
Свободный член	1346,700	1	1346,700	781,292	,000	,873
ИНТ	1,200	1	1,200	,696	,406	,006
Ч_ряда	8,150	2	4,075	2,364	,099	,040
ИНТ * Ч_ряда	29,450	2	14,725	8,543	,000	,130
Ошибка	196,500	114	1,724			
Всего	1582,000	120				
Скорректированный итог	235,300	119				

a. R квадрат = ,165 (Скорректированный R квадрат = ,128)

Рис. 14.4. Фрагменты окна вывода после выполнения шага 5

В таблице Описательные статистики содержатся характеристики всех выборок, образованных каждой из независимых переменных в отдельности (две выборки по переменной Инт и три выборки по переменной Ч_ряда), а также при пересечении градаций независимых переменных ($3 \times 2 = 6$ выборок). Таблица Оценка эффек-

тов межгрупповых факторов содержит результаты проверки трех основных гипотез двухфакторного дисперсионного анализа:

- ▶ Переменная Ч_ряда не оказывает статистически достоверное влияние на распределение зависимой переменной Слова (средние значения для начала, середины и конца ряда составили соответственно 3,08, 3,70 и 3,28, $F = 2,364$, $p = 0,099$).
- ▶ Переменная Инт не оказывает статистически значимого влияния на распределение зависимой переменной Слова (средние значения для групп «нет» и «есть» составили соответственно 3,45 и 3,25, $F = 0,696$, $p = 0,406$).
- ▶ Обнаружено статистически достоверное взаимодействие на высоком уровне статистической значимости между независимыми переменными Ч_ряда и Инт ($F = 8,543$, $p < 0,001$).

Взаимодействие независимых переменных более подробно обсуждается в последнем разделе этой главы, посвященном использованию графиков для интерпретации взаимодействий.

Влияние ковариаты

На рис. 14.5 приведен фрагмент выводимых данных, сгенерированные программой после выполнения шага 5а.

Оценка эффектов межгрупповых факторов

Зависимая переменная: Слова

Источник	Сумма квадратов типа III	ст. св.	Средний квадрат	F	Знач.	Частная Эта в Квадрате
Скорректированная модель	90,862 ^a	6	15,144	11,848	,000	,386
Свободный член	25,047	1	25,047	19,595	,000	,148
ЗНАЧ	52,062	1	52,062	40,730	,000	,265
ИНТ	5,662	1	5,662	4,429	,038	,038
Ч_ряда	8,150	2	4,075	3,188	,045	,053
ИНТ * Ч_ряда	29,450	2	14,725	11,520	,000	,169
Ошибка	144,438	113	1,278			
Всего	1582,000	120				
Скорректированный итог	235,300	119				

a. R квадрат = ,386 (Скорректированный R квадрат = ,354)

Рис. 14.5. Фрагмент окна вывода после выполнения шага 5а

Включение ковариаты ЗНАЧ не меняет описательные статистики, но вносит некоторые изменения в результаты проверки гипотез дисперсионного анализа, как видно из таблицы Оценка эффектов межгрупповых факторов.

Ковариата оказывает влияние на суммы квадратов и, как следствие, на число степеней свободы (в некоторых случаях), средние квадраты, величины F и частной эта в квадрате (η^2), а также значения p -уровней значимости. Если это влияние существенно, то, как правило, значения среднего квадрата и F -критерия увеличиваются, а соответствующие значения p -уровней уменьшаются.

Обратите внимание, что ковариата ЗНАЧ оказывает значительное влияние на разброс зависимой переменной Слова: значение η^2 составляет 0,265, то есть 26,5 % дисперсии переменной Слова обусловлено влиянием ковариаты. Дисперсия скорректированной модели представляет собой сумму всех сумм квадратов дисперсий, обусловленных влияниями независимых переменных и их взаимодействий.

Сравните между собой результаты дисперсионного анализа с ковариатой и без ковариаты. Благодаря тому, что ковариата «съедает» часть дисперсии зависимой переменной, для главных эффектов и взаимодействия факторов F -критерий увеличивается, а p -уровень уменьшается. В результате главные эффекты факторов становятся статистически достоверными. Разумеется, подобный результат проявляется не всегда: ковариата способна не только повысить, но и понизить статистическую значимость эффектов факторов и их взаимодействий — заранее это не предсказуемо.

Таким образом, двухфакторный дисперсионный анализ с зависимой переменной Слова, независимыми переменными Инт и Ч_ряда и ковариатой ЗНАЧ дал следующие результаты:

- ▶ Ковариата ЗНАЧ оказывает статистически достоверное влияние на зависимую переменную Слова ($F = 40,73, p < 0,001$).
- ▶ Переменная Инт оказывает статистическое влияние на распределение зависимой переменной Слова ($F = 4,429, p = 0,038$).
- ▶ Переменная Ч_ряда оказывает статистически значимое влияние на распределение зависимой переменной Слова ($F = 3,188, p = 0,045$).
- ▶ Обнаружено статистически достоверное взаимодействие между независимыми переменными Ч_ряда и Инт ($F = 11,52, p < 0,001$).

Отметим, что не очень удобно интерпретировать результаты по средним значениям из таблицы Описательные статистики. Для этого лучше использовать графики средних значений, полученные после выполнения шага 5б или 6.

Использование графиков для интерпретации взаимодействий

На рис. 14.6 приведен график средних значений, который генерируется программой после выполнения шага 5б.

График позволяет без труда интерпретировать взаимодействие факторов: влияние фактора Инт на переменную отметка2 проявляется по-разному на разных градациях переменной Ч_ряда: в случае наличия интонационного выделения (Инт: есть) воспроизведение слов из середины ряда существенно более продуктивно, чем в случае его отсутствия (Инт: нет).

Вообще говоря, взаимодействие факторов на графике выглядит как заметное различие формы профилей соответствующих линий на рис. 14.6 или соответствующих кластеров столбиков на рис. 14.7. Чем ближе по форме профили линий (кластеров), тем меньше взаимодействие факторов. Но интерпретации подлежат

только те графики, которые соответствуют статистически достоверному взаимодействию факторов.



Рис. 14.6. График средних значений, генерируемый после выполнения шага 5б

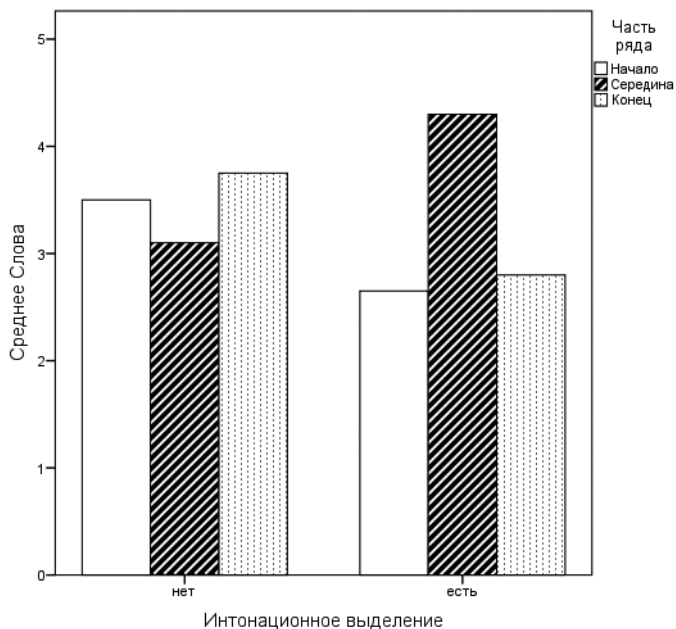


Рис. 14.7. График средних значений, генерируемый после выполнения шага 6

Если бы проводился трехфакторный анализ, то интерпретация трехфакторного взаимодействия требовала бы построения нескольких графиков. Так, при добавлении фактора Отсрочка потребовалось бы построение трех графиков двухфакторного взаимодействия: по одному для каждой градации фактора Отсрочка. Попробуйте самостоятельно провести трехфакторный дисперсионный анализ, добавив при выполнении инструкций шага 5б переменную Отсрочка в качестве третьего фактора и при задании параметров графиков средних значений в диалоговом окне указав эту переменную в поле Отдельные графики.

Терминология, используемая при выводе

Далее дана трактовка новых терминов, используемых программой в окне вывода (трактовка остальных терминов ANOVA приведена в конце главы 13).

- Частная Эта в квадрате — оценка величины эффекта (η^2), по смыслу совпадает с RSQ или R2. Под эффектом понимается вклад независимой переменной или взаимодействия переменных в разброс значений зависимой переменной (доля от 1 — полной ее дисперсии).

Завершение анализа и выход из программы

Отредактируйте содержимое окна вывода в соответствии со своими предпочтениями: скройте лишнюю информацию, исправьте таблицы и пр. (см. раздел «Окно вывода и его редактирование» главы 2). За инструкциями по редактированию графиков обратитесь к главе 5.

Для дальнейшего использования окончательного результата все содержимое окна вывода или его фрагменты можно сохранить в файле *.spv, экспортировать в другой формат (например, Word), перенести в документ Word или вывести на печать (подробности см. в разделе «Сохранение, экспорт, перенос и печать результатов» главы 2).

Для выхода из программы выберите команду Выход в меню Файл.

15 Многомерный дисперсионный анализ

214	Пошаговые алгоритмы вычислений
220	Представление результатов
225	Завершение анализа и выход из программы

В этой главе мы рассмотрим методы обработки данных, которые содержат несколько зависимых переменных: многомерный дисперсионный анализ (Multivariate Analysis Of Variances, MANOVA) и многомерный ковариационный анализ (Multivariate Analysis Of Covariance, MANCOVA). Команды многомерного анализа, входящие в подменю Общая линейная модель, относятся к наиболее сложным командам в SPSS. Помимо команды ОЛМ-многомерная, предназначенной для проведения анализов MANOVA и MANCOVA, к командам многомерного анализа относится и команда ОЛМ-повторные измерения, позволяющая провести анализ MANOVA с повторными измерениями. Первый тип многомерного анализа будет рассмотрен в этой главе, второй — в главе 16.

Ввиду значительной сложности мы упомянем лишь наиболее понятные и широко используемые параметры команды ОЛМ-многомерная. Фактически, критерии MANOVA и MANCOVA являются расширением дисперсионного анализа (ANOVA), рассмотренного нами ранее (см. главы 13–14); если вы не знакомы с дисперсионным анализом, рекомендуем, прежде чем читать дальше, обратиться к указанным главам.

Как упоминалось в предыдущих главах, t -критерий для двух выборок позволяет выяснить, существуют ли различия между двумя средними значениями для этих выборок. Эту простейшую ситуацию (единственная независимая переменная с двумя градациями и одна зависимая переменная метрического типа) можно последовательно усложнить тремя способами:

- ▶ ввести в рассмотрение независимую переменную с более чем двумя градациями — в такой ситуации применяется однофакторный дисперсионный анализ;
- ▶ ввести не одну, а несколько независимых переменных — для этого предназначен многофакторный дисперсионный анализ;
- ▶ задействовать ковариаты.

Во всех трех случаях зависимая переменная остается единственной и имеет метрический тип. Тем не менее существуют задачи, в которых требуется учитывать не одну, а несколько зависимых переменных. В этой главе мы займемся рассмотрением проблемы проведения анализа с участием более чем одной зависимой переменной; при этом мы не станем усложнять требования к независимым переменным.

В этой главе для примера используется файл данных `ex021.sav`. Подробное описание переменных файла приводится в разделе «Файлы данных для группы методов Общая линейная модель» главы 14. В нем, как и в двух других файлах (`ex020.sav` и `ex022.sav`), содержатся данные исследования влияния интонационного выделения средней части ряда из 24 предъявляемых слов на продуктивность их воспроизведения в зависимости от части ряда, эмоциональной значимости слов и отсрочки воспроизведения. В файле `ex021.sav` зависимыми переменными являются: количество воспроизведенных слов в начале ряда (НАЧАЛО); в середине ряда (СЕРЕДИНА); в конце ряда (КОНЕЦ). Независимые переменные (факторы): Отсрочка (интонационное выделение: 0 — нет, 1 — есть); Отсрочка (0 — нет, 1 — есть); Значимость (эмоциональная значимость ряда слов, количественная переменная).

Представим себе, что нам необходимо сравнить продуктивность воспроизведения при интонационном выделении и без него (переменная Отсрочка) одновременно по всем трем показателям (переменные НАЧАЛО, СЕРЕДИНА, КОНЕЦ). В подобной ситуации одним из возможных подходов является трехкратное применение t -критерия или однофакторного дисперсионного анализа (эти методы эквивалентны, поскольку $t^2 = F$). Очевидным достоинством такого решения является простота и ясность, однако нельзя не заметить и двух недостатков: во-первых, при неоднократном применении статистического критерия (в данном случае трехкратном) увеличивается вероятность ошибки, то есть вероятность случайности общего результата исследования; во-вторых, если между зависимыми переменными имеется некоторая корреляция (а в рассматриваемом случае она есть), то результат, полученный в отношении каждой из этих переменных в отдельности, не способен отразить этот важный факт.

Описанные недостатки привели к усовершенствованию как t -критерия, так и дисперсионного анализа: первый был расширен с помощью критерия Хотеллинга (Hotelling), а вместо второго стал использоваться многомерный дисперсионный анализ (MANOVA). Оба типа анализа реализуются командой ОЛМ-многомерная. Кроме того, эта команда позволяет проводить многомерный ковариационный анализ (MANCOVA), учитывающий влияние ковариат. Особенностью всех типов многомерного анализа является то, что они обрабатывают все зависимые переменные одновременно. В примерах этой главы показано, каким образом исследовать структуру изменений зависимых переменных путем применения серий одномерных F -критериев или серий множественных сравнений постфактум.

Применяя MANOVA в отношении множества зависимых переменных, следует помнить, что линейная функциональная связь между ними недопустима. Иными словами, следует избегать применения анализа MANOVA к тем зависимым переменным, корреляция между которыми близка к 1.

Как и в случае одномерного дисперсионного анализа, в MANOVA для определения значимости различий между группами используются F -критерий. Отметим, что многомерный дисперсионный анализ, как и одномерный, позволяет оценить влияние не только отдельных независимых переменных (главных эффектов), но и их взаимодействий. Поскольку мы рассматриваем наличие нескольких зависимых переменных, F -статистика носит многомерный характер и для ее формирования используется матричная алгебра.

Для многомерного анализа необходимо иметь как минимум две зависимые переменные (иначе анализ не является многомерным) и как минимум одну независимую переменную. Теоретически количество зависимых и независимых переменных не ограничено, однако на практике объем выборки диктует необходимость существенного ограничения их числа. Использование ковариат, разумеется, не является обязательным.

Пошаговые алгоритмы вычислений

Для проведения многомерного дисперсионного анализа сначала необходимо выполнить три подготовительных шага. Эти шаги (шаги 1–3) позволят подготовить рабочий файл данных, запустить программу IBM SPSS Statistics 19 и открыть файл (в данном случае — файл ex021.sav). Пошаговые инструкции этого процесса приведены в главе 4 (с. 60), а подробные разъяснения — в главе 2.

После завершения шага 3 на экране должно присутствовать окно редактора данных со строкой меню и загруженным файлом ex021.sav.

ШАГ 4

В меню Анализ выберите команду Общая линейная модель ► ОЛМ-многомерная. Откроется диалоговое окно ОЛМ-многомерная, показанное на рис. 15.1.

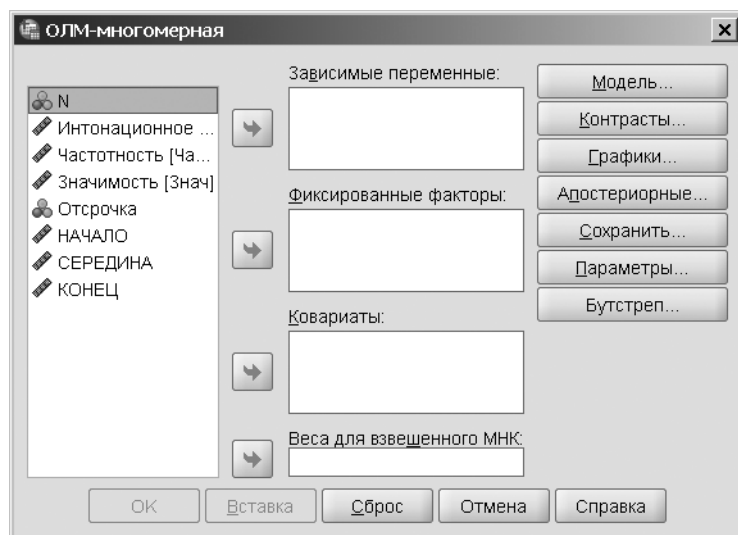


Рис. 15.1. Диалоговое окно ОЛМ-многомерная

Окно ОЛМ-многомерная позволяет задавать переменные для многомерного анализа, а также управлять диалоговыми окнами, предназначенными для определения параметров анализа. Те переменные, которые будут использоваться в качестве зависимых, необходимо поместить в список Зависимые переменные, а независимые переменные — в список Фиксированные факторы.

Как и прежде, для перемещения переменных из исходного списка в целевой используется одна из кнопок со стрелками. Мы будем использовать в наших примерах анализа три зависимые и две независимые переменные из файла ex021.sav. При желании в анализ могут быть введены одна или несколько ковариат с помощью списка Ковариаты.

В правой части диалогового окна расположены 7 кнопок, предназначенных для настройки параметров анализа. Использование кнопок Контрасты, Сохранить и Бутстреп не рассматривается, поскольку настройка соответствующих параметров требует очень высокой квалификации исследователя.

В большинстве случаев требуется многомерный анализ как для каждой из независимых переменных, так и для всех вариантов их взаимодействий (полнофакторная модель). Полнофакторная модель является моделью анализа, используемой по умолчанию. С другой стороны, иногда полезно исключить из рассмотрения часть взаимодействий или отдельные факторы. Для управления моделью анализа используется кнопка Модель. При щелчке на ней открывается диалоговое окно ОЛМ-многомерная: Модель, показанное на рис. 15.2.

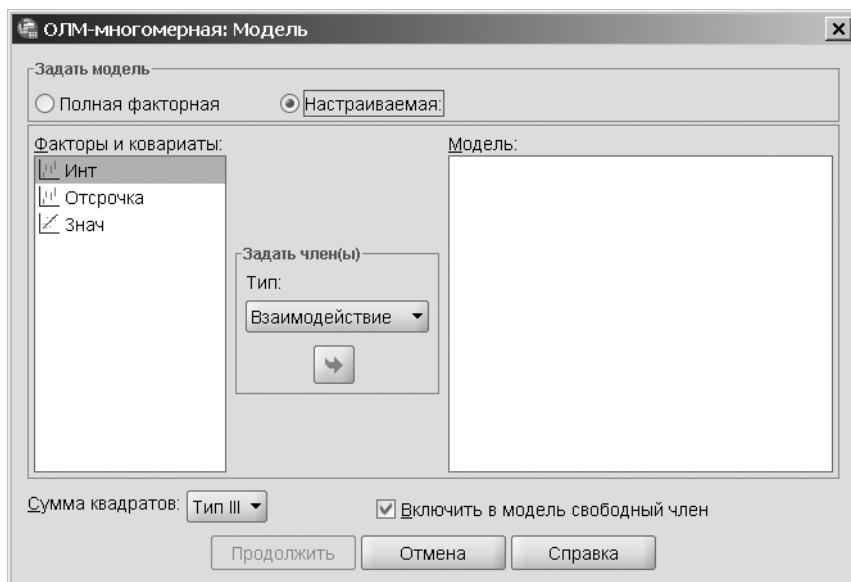


Рис. 15.2. Диалоговое окно ОЛМ-многомерная: Модель

По умолчанию в группе Задать модель установлен переключатель Полная факторная, однако вы можете установить переключатель Настраиваемая, а затем с помощью кнопки с направленной вправо стрелкой переместить в список Модель только те переменные из списка Факторы и ковариаты, которые вы хотите использовать. При этом в списке Тип (группа Задать член(ы)) для каждого сочетания факторов вы можете указать тип проверяемых гипотез (Взаимодействие, Главные эффекты, Все 2-х факторные и т. д.).

С помощью списка Сумма квадратов можно также задать вариант расчета суммы квадратов. По умолчанию выбран вариант Тип III, подходящий для большинства ситуаций. Если исходные данные содержат пропущенные значения, иногда выбирают пункт Тип IV.

При интерпретации результатов многомерного анализа зачастую удобно иметь перед глазами графическое представление средних значений зависимых переменных, определяемых различными комбинациями градаций независимых переменных (факторов). Для построения подобных графиков используется кнопка Графики, назначение которой идентично назначению одноименной кнопки в окне ОЛМ-одномерная (см. главу 14). При щелчке на ней открывается диалоговое окно ОЛМ-многомерная: Графики профилей, аналогичное изображенному на рис. 14.3 (глава 14). Данное окно позволяет определить, каким образом отображать тот или иной фактор. Для каждой зависимой переменной в выводимые данные включаются отдельные графики. Обратите внимание, что при выполнении анализа с участием ковариат вместо фактических данных отображаются приближения средних значений, учитывающие наличие ковариат. В подобных случаях может быть полезным получение изображений без использования ковариат.

Вид графиков задается тремя параметрами. В поле Горизонтальная ось вводится имя фактора, градациям которого будет соответствовать горизонтальная ось графика или графиков. В поле Отдельные линии указывается фактор, градациям которого будут соответствовать разные линии на графике. Поле Отдельные графики служит для задания имени фактора, градациям которого будут соответствовать разные графики. Щелчок на кнопке Добавить подтверждает правильность введенных параметров и позволяет перейти к заданию графиков другого типа.

Графическая интерпретация весьма полезна для оценки влияния независимых переменных и их взаимодействий. Для более детального анализа требуются множественные (парные) сравнения постфактум. Множественные сравнения проводятся в случае установления статистически достоверного влияния независимой переменной и позволяют определять, между какими именно градациями независимой переменной имеются различия. Подобные сравнения проводятся для каждой из зависимых переменных и каждого выбранного фактора. Множественные сравнения неприменимы при использовании ковариат.

Для проведения парных сравнений применяется кнопка Апостериорные. При щелчке на ней открывается ОЛМ-многомерный: Апостериорные множественные сравнения для наблюдаемых средних, представленное на рис. 15.3. Это окно идентично диалоговому окну Однофакторный дисперсионный анализ: Апостериорные множественные сравнения, которое подробно описано в главе 13. Отличается окно для многомерного анализа только наличием списков Фактор(ы) и Апостериорные критерии для.

В список Апостериорные критерии для следует включить факторы, для которых будут проводиться сравнения. Все доступные факторы перечислены в списке Фактор(ы). За описанием наиболее популярных критериев парных сравнений и результатов их применения обратитесь к главе 13.

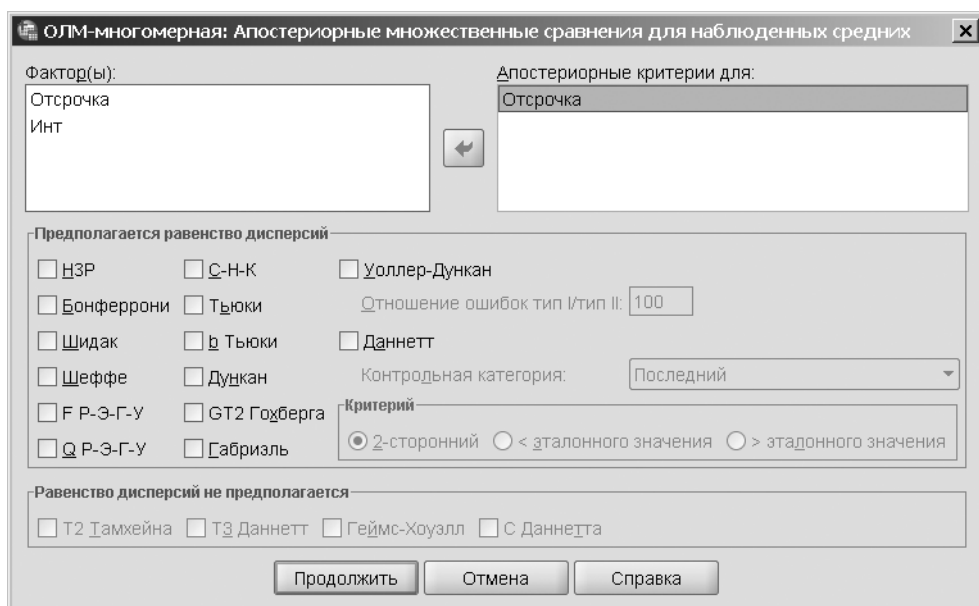


Рис. 15.3. Диалоговое окно ОЛМ-многомерная: Апостериорные множественные сравнения для наблюдаемых средних

При щелчке на кнопке Параметры открывается диалоговое окно ОЛМ-многомерная: Параметры, показанное на рис. 15.4. По умолчанию все флажки в этом окне сброшены.

Если выделить нужные пункты в списке Факторы и их взаимодействия и переместить Вывести средние для, то в выводимые данные будут включены средние значения каждой из зависимых переменных для всех уровней указанных факторов. Если анализ проводится без ковариат, включаемые средние значения будут фактическими средними значениями данных; в противном случае средние значения будут подсчитаны с учетом влияния ковариат. Если установить флажок Сравнить главные эффекты, может быть проведена серия апостериорных парных сравнений средних значений ячеек для каждого из факторов: при помощи ниспадающего меню ниже вы можете задать либо критерий Бонферрони, либо более «консервативный» критерий Шидака. Другие критерии задаются при помощи описанного выше диалогового окна ОЛМ-многомерный: Апостериорные множественные сравнения для наблюдаемых средних при условии, что анализ проводится без ковариат.

Мы опустим описания большей части флажков и упомянем лишь те, которые используются наиболее часто.

- **Описательные статистики** — вычисляются средние значения и стандартные отклонения каждой зависимой переменной для каждой ячейки.
- **Оценки размера эффекта** — вычисляется значение η^2 , характеризующее величину воздействия фактора на зависимую переменную и показывающее, какая доля общей дисперсии обусловлена данным фактором.

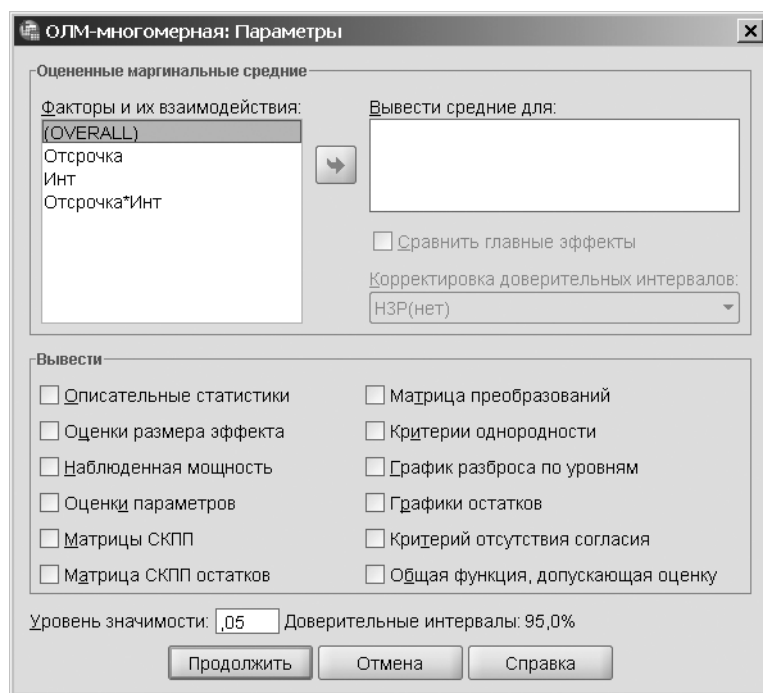


Рис. 15.4. Диалоговое окно ОЛМ-многомерная: Параметры

- ▶ Оценки параметров — вычисляются приближения и соответствующие значимости для всех факторов и ковариат модели. Это особенно удобно в случаях, когда в модель включена хотя бы одна ковариата.
- ▶ Критерии однородности — дисперсии для всех ячеек проверяются на равенство (однородность).

Поле Уровень значимости позволяет задать величину α для анализа. По умолчанию задан критический уровень значимости 0,05; если вы измените это значение, SPSS пересчитает соответствующие доверительные интервалы.

В рассматриваемом далее примере проводится многомерный дисперсионный анализ (MANOVA) с тремя зависимыми переменными НАЧАЛО, СЕРЕДИНА, КОНЕЦ и двумя независимыми переменными (факторами) Инт и Отсрочка.

ШАГ 5

После выполнения предыдущего шага у вас должно быть открыто диалоговое окно ОЛМ-многомерная, показанное на рис. 15.1. Если вы уже успели поработать с этим окном, щелкните на кнопке Сброс.

1. Щелкните на переменной НАЧАЛО, нажмите клавишу Shift и, не отпуская ее, щелкните на переменной КОНЕЦ. В результате окажутся выделенными переменные НАЧАЛО, СЕРЕДИНА, КОНЕЦ.
2. Щелкните на верхней кнопке со стрелкой, чтобы переместить выделенные переменные в список Зависимые переменные.

3. Щелкните сначала на переменной Инт, чтобы выделить ее, а затем — на второй сверху кнопке со стрелкой, чтобы переместить переменную в список Фиксированные факторы.
4. Повторите предыдущее действие для переменной Отсрочка.
5. Щелкните на кнопке ОК, чтобы открыть окно вывода.

В следующем примере мы включим в анализ ковариату Знач.

ШАГ 5А

После выполнения шага 4 у вас должно быть открыто диалоговое окно ОЛМ-многомерная, показанное на рис. 15.1. Если вы уже успели поработать с этим окном, щелкните на кнопке Сброс.

1. Щелкните на переменной НАЧАЛО, нажмите клавишу Shift и, не отпуская ее, щелкните на переменной КОНЕЦ. В результате окажутся выделенными переменные НАЧАЛО, СЕРЕДИНА, КОНЕЦ.
2. Щелкните на верхней кнопке со стрелкой, чтобы переместить выделенные переменные в список Зависимые переменные.
3. Щелкните сначала на переменной Инт, чтобы выделить ее, а затем — на второй сверху кнопке со стрелкой, чтобы переместить переменную в список Фиксированные факторы.
4. Повторите предыдущее действие для переменной Отсрочка.
5. Щелкните сначала на переменной Знач, чтобы выделить ее, а затем — на третьей сверху кнопке со стрелкой, чтобы переместить переменную в список Ковариаты.
6. Щелкните на кнопке ОК, чтобы открыть окно вывода.

В следующем примере демонстрируется использование нескольких описанных ранее параметров, генерирующих полезную статистическую и графическую информацию. Обратите внимание, что мы не включаем в данный анализ ковариату. Напоминаем, что для анализа с ковариатами парные сравнения неприменимы, поскольку все графики будут соответствовать не фактическим исходным данным, а данным, скорректированным влияниями ковариат. Кроме того, парные сравнения имеют смысл лишь для факторов, имеющих более двух уровней; по этой причине в данном примере парные сравнения не применяются. Если же у вас какой-либо фактор имеет более двух уровней, воспользуйтесь кнопкой Апостериорные. Она действует практически идентично своему аналогу из диалогового окна Однофакторный дисперсионный анализ, поэтому при необходимости обратитесь к главе 13. Отличие лишь в том, что в данном случае следует указать каждый фактор, для которого необходимы парные сравнения.

ШАГ 5Б

После выполнения шага 4 у вас должно быть открыто диалоговое окно ОЛМ-многомерная, показанное на рис. 15.1. Если вы уже успели поработать с этим окном, щелкните на кнопке Сброс.

1. Щелкните на переменной НАЧАЛО, нажмите клавишу Shift и, не отпуская ее, щелкните на переменной КОНЕЦ. В результате окажутся выделенными переменные НАЧАЛО, СЕРЕДИНА, КОНЕЦ.

2. Щелкните на верхней кнопке со стрелкой, чтобы переместить выделенные переменные в список Зависимые переменные.
3. Щелкните сначала на переменной Инт, чтобы выделить ее, а затем — на второй сверху кнопке со стрелкой, чтобы переместить переменную в список Фиксированные факторы.
4. Повторите предыдущее действие для переменной Отсрочка.
5. Щелкните на кнопке Графики, чтобы открыть диалоговое окно ОЛМ-многомерная: Графики профилей, показанное на рис. 14.3.
6. Щелкните сначала на переменной Отсрочка, чтобы выделить ее, а затем — на верхней кнопке со стрелкой, чтобы переместить переменную в поле Горизонтальная ось.
7. Щелкните сначала на переменной Инт, чтобы выделить ее, а затем — на средней кнопке со стрелкой, чтобы переместить переменную в поле Отдельные линии.
8. Щелкните сначала на кнопке Добавить, чтобы зафиксировать заданные параметры построения диаграмм, а затем — на кнопке Продолжить, чтобы вернуться в диалоговое окно ОЛМ-многомерная.
9. Щелкните на кнопке Параметры, чтобы открыть диалоговое окно ОЛМ-многомерная: Параметры, показанное на рис. 15.4. В левом окне будут перечислены факторы и их взаимодействие (Инт*Отсрочка).
10. Щелкните сначала в левом окне на строке Инт*Отсрочка, чтобы выделить ее, а затем — на кнопке со стрелкой, чтобы переместить ее в список Вывести средние для.
11. Установите флажки Описательные статистики, Оценки размера эффекта, Критерии однородности. Щелкните на кнопке Продолжить, чтобы вернуться в диалоговое окно ОЛМ-многомерная.
12. Щелкните на кнопке ОК, чтобы открыть окно вывода.

После выполнения шага 5 (5а, 5б) программа автоматически активизирует окно вывода. Для просмотра результатов при необходимости можно воспользоваться вертикальной и горизонтальной полосами прокрутки. Обратите внимание на стандартную строку меню в верхней части окна вывода: ее присутствие позволяет выполнять любые статистические операции, не переключаясь обратно в окно редактора данных.

Представление результатов

На представленных далее рисунках приведены наиболее важные фрагменты выводимых результатов, генерируемых программой при выполнении шагов 5 и 5б. Мы не приводим эти данные целиком, поскольку их объем составляет 7 страниц.

Межгрупповые факторы

В таблице, представленной на рис. 15.5, перечислены все независимые переменные со своими уровнями и объемами выборок (N).

Межгрупповые факторы			
		Метка значения	N
Интонационное выделение	1	нет	20
	2	есть	20
Отсрочка	1	нет	20
	2	есть	20

Рис. 15.5. Фрагмент окна вывода после выполнения шага 5 (межгрупповые факторы)

Описательные статистики

Таблица описательных статистик, полученная в результате выполнения шага 5 (на рис. 15.6 представлен только ее фрагмент), содержит средние значения, стандартные отклонения и объемы выборок для всех элементов модели.

Проверка ковариационных матриц на равенство

На рис. 15.7 приведены результаты проверки равенства ковариационных матриц, которые появляются, если установлен флажок Критерии однородности в окне ОЛМ-многомерная: Параметры). Эти результаты говорят о том, что дисперсионно-ковариационные матрицы не отличаются между собой (значение p -уровня превышает 0,05). Если бы результаты проверки были статистически достоверными, то мы не могли бы ручаться за достоверность многомерного анализа, одним из условий проведения которого, как вы помните, является равенство дисперсионно-ковариационных матриц.

Описательные статистики					
	Интонационное выделение	Отсрочка	Среднее	Стд. Отклонение	N
НАЧАЛО	нет	нет	4,20	1,476	10
		есть	2,80	1,398	10
		Всего	3,50	1,573	20
	есть	нет	3,30	1,494	10
		есть	2,00	1,054	10
		Всего	2,65	1,424	20
	Всего	нет	3,75	1,517	20
		есть	2,40	1,273	20
		Всего	3,08	1,542	40

Рис. 15.6. Фрагмент окна вывода после выполнения шага 5 (описательные статистики)

Критерий Бокса равенства ковариационных матриц	
М Бокса	19,003
F	,893
ст.св.1	18
ст.св.2	4579,730
Знач.	,588

Рис. 15.7. Фрагмент окна вывода после выполнения шага 5 (критерий Бокса равенства ковариационных матриц)

Многомерные критерии

Таблица, показанная на рис. 15.8, — результат выполнения шага 5 или 5б (без ковариаты). В ней содержится информация, которая является результатом проверки влияния факторов и их взаимодействий на зависимые переменные. Секция Свободный член характеризует остаточную дисперсию и не подлежит интерпретации. Как можно видеть, обнаружены статистически значимые главные эффекты факторов Инт и Отсрочка ($p < 0,05$). Взаимодействие этих факторов (Инт*Отсрочка) тоже статистически достоверно. Если в анализе участвуют ковариаты, они также отображаются в этой таблице. Выполнив шаг 5а, вы можете убедиться, что ковариата знач, применяемая в нашем примере, оказывает значимое влияние на зависимые переменные и меняет статистическую значимость влияния факторов и их взаимодействий.

Ниже дана трактовка терминов, используемых программой в окне вывода.

- Значения — в этом столбце представлены значения различных критериев для проверки влияния независимых переменных. Из всех критериев наиболее надежным считается тест Пилая (Pillai).
- F — оценка F-критерия.
- Ст. св. гипотезы — степень свободы гипотезы. Произведение числа зависимых переменных и уровней каждой независимой переменной, уменьшенных на 1. Для данного примера:

$$3 \times (2 - 1) \times (2 - 1) = 3 \times 1 \times 1 = 3.$$
- Ст. св. ошибки — величина, вычисляемая различными способами в зависимости от критерия.
- Знч. — уровень значимости для соответствующего F-критерия.

Многомерные критерии^б

Эффект		Значения	F	Ст. св. гипотезы	Ст. св. ошибки	Знч.	Частная Эта в Квадрате
Свободный член	След Пиллая	,957	253,376 ^а	3,000	34,000	,000	,957
	Лямбда Уилкса	,043	253,376 ^а	3,000	34,000	,000	,957
	След Хотеллинга	22,357	253,376 ^а	3,000	34,000	,000	,957
	Наибольший корень Роя	22,357	253,376 ^а	3,000	34,000	,000	,957
Инт	След Пиллая	,768	37,415 ^а	3,000	34,000	,000	,768
	Лямбда Уилкса	,232	37,415 ^а	3,000	34,000	,000	,768
	След Хотеллинга	3,301	37,415 ^а	3,000	34,000	,000	,768
	Наибольший корень Роя	3,301	37,415 ^а	3,000	34,000	,000	,768
Отсрочка	След Пиллая	,425	8,371 ^а	3,000	34,000	,000	,425
	Лямбда Уилкса	,575	8,371 ^а	3,000	34,000	,000	,425
	След Хотеллинга	,739	8,371 ^а	3,000	34,000	,000	,425
	Наибольший корень Роя	,739	8,371 ^а	3,000	34,000	,000	,425
Инт * Отсрочка	След Пиллая	,370	6,095 ^а	3,000	34,000	,001	,370
	Лямбда Уилкса	,622	6,095 ^а	3,000	34,000	,001	,378
	След Хотеллинга	,608	6,095 ^а	3,000	34,000	,001	,378
	Наибольший корень Роя	,608	6,095 ^а	3,000	34,000	,001	,378

а. Точная статистика

б. План: Свободный член + Инт + Отсрочка + Инт * Отсрочка

Рис. 15.8. Фрагмент окна вывода после выполнения шага 5 (многомерные критерии)

Критерий равенства дисперсий

Критерий равенства дисперсий (критерий Ливиня) позволяет проверить допущение о том, что дисперсии всех зависимых переменных одинаковы (рис. 15.9).

Критерий Ливиня проверки равенства дисперсий.

	F	ст.св.1	ст.св.2	Знч.
НАЧАЛО	,471	3	36	,704
СЕРЕДИНА	1,818	3	36	,161
КОНЕЦ	1,390	3	36	,262

Рис. 15.9. Фрагмент окна вывода после выполнения шага 5 (критерий Ливиня проверки равенства дисперсий)

Большие значения p -уровня для каждой из зависимых переменных не дают поводов для излишнего беспокойства относительно корректности дисперсионного анализа.

Одномерные критерии

В дополнение к многомерным критериям для всех зависимых переменных, к каждой из них в отдельности применяется одномерный F -критерий. Несмотря на то что многомерный критерий, как было сказано ранее, является более совершенным, чем серия одномерных критериев, последние необходимы для интерпретации статистически достоверных результатов применения многомерного критерия. Таким образом, если многомерный критерий показывает статистически значимый результат, то для детальной интерпретации используются результаты вычисления одномерных критериев.

В нашем случае статистически достоверными являются результаты применения многомерных критериев для главных эффектов и взаимодействия факторов Инт и Отсрочка. На рис. 15.10 приведен фрагмент таблицы с результатами одномерного анализа. Как видно по этому фрагменту, влияние фактора Инт статистически достоверно в отношении двух из трех зависимых переменных: СЕРЕДИНА и КОНЕЦ, а влияние фактора Отсрочка статистически достоверно в отношении каждой из трех переменных. Взаимодействие факторов (Инт*Отсрочка) статистически достоверно в отношении переменной КОНЕЦ. Для содержательной интерпретации одномерных критериев в нашем случае можно обратиться к описательным статистикам (средним), а лучше — к графикам средних значений. Если бы факторы имели более двух градаций, перед интерпретацией необходимо было бы применить метод апостериорных парных сравнений. Если анализ включает ковариаты, для каждой зависимой переменной одномерный критерий будет учитывать и их влияние.

Оценки средних значений для уровней факторов

Таблица, приведенная на рис. 15.11, получена после выполнения шага 5б. Она содержит средние значения каждой зависимой переменной для каждой градации факторов. По этим данным можно дать подробную содержательную интерпретацию статистически достоверных результатов. Так, статистически достоверное влияние фактора Инт на совокупность зависимых переменных было детализиро-

вано результатами применения одномерных критериев: фактор Отсрочка значимо влияет на переменные СЕРЕДИНА и КОНЕЦ. Судя по средним значениям для этих переменных, интонационное выделение середины ряда слов приводит к тому, что эта часть ряда запоминается лучше, а конец ряда — хуже, чем без интонационного выделения. Влияние фактора Отсрочка проявляется в том, отсроченное воспроизведение менее продуктивно для каждой из трех частей ряда.

Оценка эффектов межгрупповых факторов

Источник	Зависимая переменная	Сумма квадратов типа III	ст.св.	Средний квадрат	F	Знач.	Полная Эта в Квадрате
Скорректированная модель	НАЧАЛО	25,475 ^a	3	8,492	4,542	,008	,275
	СЕРЕДИНА	33,800 ^b	3	11,267	13,255	,000	,525
	КОНЕЦ	34,275 ^c	3	11,425	11,521	,000	,490
Свободный член	НАЧАЛО	378,225	1	378,225	202,319	,000	,849
	СЕРЕДИНА	547,600	1	547,600	644,235	,000	,947
	КОНЕЦ	429,025	1	429,025	432,630	,000	,923
Инт	НАЧАЛО	7,225	1	7,225	3,865	,057	,097
	СЕРЕДИНА	14,400	1	14,400	16,941	,000	,320
	КОНЕЦ	9,025	1	9,025	9,101	,005	,202
Отсрочка	НАЧАЛО	18,225	1	18,225	9,749	,004	,213
	СЕРЕДИНА	16,900	1	16,900	19,882	,000	,356
	КОНЕЦ	21,025	1	21,025	21,202	,000	,371
Инт * Отсрочка	НАЧАЛО	,025	1	,025	,013	,909	,000
	СЕРЕДИНА	2,500	1	2,500	2,941	,095	,076
	КОНЕЦ	4,225	1	4,225	4,261	,046	,106

Рис. 15.10. Фрагмент окна вывода после выполнения шага 5 (одномерные критерии)

Интонационное выделение * Отсрочка

Зависимая переменная	Интонационное выделение	Отсрочка	Среднее	Стд. Ошибка	95% доверительный интервал	
					Нижняя граница	Верхняя граница
НАЧАЛО	нет	нет	4,200	,432	3,323	5,077
		есть	2,800	,432	1,923	3,677
	есть	нет	3,300	,432	2,423	4,177
		есть	2,000	,432	1,123	2,877
СЕРЕДИНА	нет	нет	3,500	,292	2,909	4,091
		есть	2,700	,292	2,109	3,291
	есть	нет	5,200	,292	4,609	5,791
		есть	3,400	,292	2,809	3,991
КОНЕЦ	нет	нет	4,800	,315	4,161	5,439
		есть	2,700	,315	2,061	3,339
	есть	нет	3,200	,315	2,561	3,839
		есть	2,400	,315	1,761	3,039

Рис. 15.11. Фрагмент окна вывода после выполнения шага 5б (оценки средних значений для уровней факторов)

Статистически достоверное взаимодействие факторов легче всего интерпретировать по графикам средних, которые получены после выполнения шага 5б. На рис. 15.12 приведен один из этих графиков для переменной КОНЕЦ. Как показывает график, взаимодействие факторов проявляется в том, что спад продуктивно-

сти отсроченного воспроизведения конца ряда слов проявляется по-разному в зависимости от интонационного выделения: при интонационном выделении спад меньше, а при его отсутствии спад более выражен. Более подробно интерпретация графиков объяснялась в разделе «Представление результатов» главы 14.

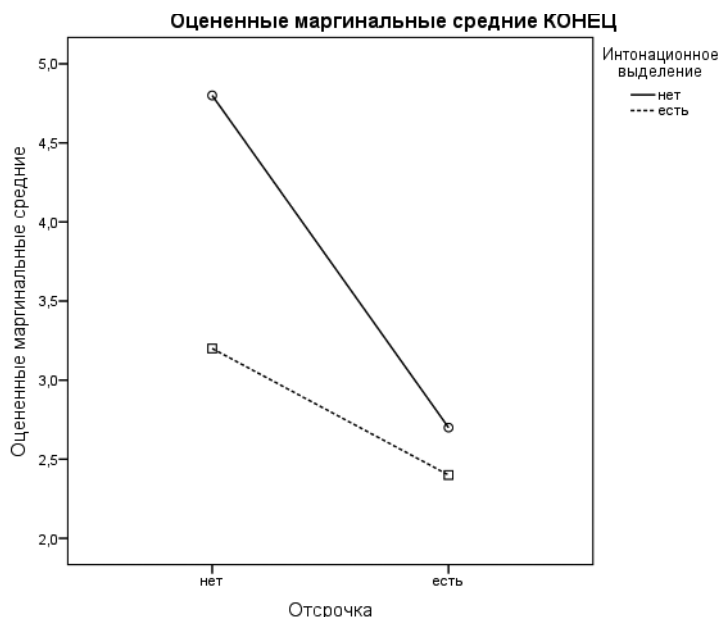


Рис. 15.12. Фрагмент окна вывода после выполнения шага 5б (график средних)

При желании можно построить альтернативные графики, например столбиковые, воспользовавшись меню в командной строке Графика ► Устаревшие диалоговые окна ► Столбики (см. шаг 6 и рис. 14.7 в главе 14).

Завершение анализа и выход из программы

Отредактируйте содержимое окна вывода в соответствии со своими предпочтениями: скройте лишнюю информацию, исправьте таблицы и пр. (см. раздел «Окно вывода и его редактирование» главы 2). За инструкциями по редактированию графиков обратитесь к главе 5.

Для дальнейшего использования окончательного результата все содержимое окна вывода или его фрагменты можно сохранить в файле *.spv, экспортировать в другой формат (например, Word), перенести в документ Word или вывести на печать (подробности см. в разделе «Сохранение, экспорт, перенос и печать результатов» главы 2).

Для выхода из программы выберите команду Выход в меню Файл.

16 Дисперсионный анализ с повторными измерениями

228	Пошаговые алгоритмы вычислений
234	Представление результатов
238	Завершение анализа и выход из программы

В предыдущей главе рассматривалась ситуация, когда необходимо было анализировать влияние на совокупность зависимых переменных нескольких независимых переменных, каждая из которых могла иметь два или более уровней. В подобных случаях проводятся межгрупповые сравнения, где группам соответствуют уровни факторов и их сочетания. Такие факторы, разным уровням которых соответствуют разные группы объектов, называются *межгрупповыми*. Однако на практике нередко используются и *внутригрупповые факторы*, разным уровням которых соответствует одна и та же группа объектов. В этом случае говорят о внутригрупповой модели, с помощью которой исследуется одна и та же группа объектов для нескольких уровней независимой переменной.

Наиболее распространенным примером применения внутригрупповой модели является исследование эффекта какого-либо воздействия. В этом случае обычно используется две выборки: на одну из них воздействие оказывается (экспериментальная выборка), а на другую — нет (контрольная выборка). Зависимые переменные, в отношении которых ожидается эффект воздействия, измеряются для обеих выборок дважды: до воздействия (предварительное тестирование) и после воздействия (итоговое тестирование).

Такая схема исследования предполагает изучение влияния на зависимые переменные двух факторов: межгруппового (два уровня: контрольная и экспериментальная выборки) и внутригруппового (два уровня: до и после воздействия). Обратите внимание, что внутригрупповому фактору в данном случае соответствует два измерения одной и той же переменной (или группы переменных). В общем случае внутригрупповой фактор может иметь и большее число уровней (повторных измерений). Отметим, что наибольший интерес для исследователя будет представлять гипотеза о взаимодействии межгруппового и внутригруппового факторов. Если она подтвердится, можно будет утверждать, что обнаружены различия между контрольной и экспериментальной группой в динамике изменения зависимой переменной.

В примерах этой главы мы будем использовать файл `ex020.sav`, содержащий те же гипотетические результаты эксперимента по изучению эффективности запомина-

ния слов, что и в файлах ex021.sav и ex022.sav. Но в файле ex021.sav (глава 14) продуктивность запоминания представлена одной (зависимой) переменной Слова, в файле ex022.sav (глава 15) — тремя, в соответствии с частями ряда (НАЧАЛО, СЕРЕДИНА, КОНЕЦ), а в файле ex020.sav — шестью переменными, в соответствии с тремя частями ряда слов и повтором их воспроизведения. Более подробное описание особенностей представления данных в соответствии с вариантами дисперсионного анализа приведено в разделе «Файлы данных для группы методов Общая линейная модель» (глава 14).

План исследования в соответствии со структурой данных в файле ex020.sav может быть представлен как комбинация двух внутригрупповых факторов (часть ряда и отсрочка), одного межгруппового фактора (интонационное выделение) и одной ковариаты (значимость ряда слов). Первый внутригрупповой фактор соответствует части ряда слов и имеет три уровня (начало, середина и конец). Второй внутригрупповой фактор — отсрочка воспроизведения, имеет два уровня (без отсрочки, с отсрочкой в два дня). Нетрудно заметить, что существует 6 комбинаций уровней этих двух факторов, каждой из которых соответствует одна из переменных (табл. 16.1).

Таблица 16.1. Сочетания комбинаций части ряда слов и отсрочки их воспроизведения

Часть ряда	Отсрочка воспроизведения	
	Нет	Два дня
Начало	начало1	начало2
Середина	середина1	середина2
Конец	конец1	конец2

Переменные, перечисленные в таблице, по сути, представляют собой 6 повторных измерений одной и той же характеристики — продуктивности воспроизведения 8 предъявленных ранее слов. Поэтому для анализа таких данных используется модель многомерного дисперсионного анализа с повторными измерениями.

Данные файла ex020.sav могут быть обработаны и иначе, в предположении наличия трех разных зависимых переменных (Начало, Середина, Конец), измеренных дважды: без отсрочки и после нее. В этом случае в рассмотрение следует ввести лишь один внутригрупповой фактор — отсрочка (2 уровня). Напомним, что при предъявлении слов для одной выборки середина ряда интонационно выделялась, для другой — нет. Следовательно, такой план исследования предполагает проведение двухфакторного дисперсионного анализа (межгрупповой фактор — «интонационное выделение», внутригрупповой фактор — «отсрочка») в отношении трех зависимых переменных (начало, середина, конец).

Модель многомерного ANOVA с повторными измерениями имеет много общего с рассмотренными ранее вариантами дисперсионного анализа без повторных измерений, поэтому материал этой главы в значительной степени опирается на мате-

риал трех предыдущих. В случае если вы пропустили главы 13–15, настоятельно рекомендуем вернуться к ним, прежде чем приступать к интерпретации результатов приводимых здесь примеров.

Следует иметь в виду, что ANOVA с повторными измерениями — самый сложный метод дисперсионного анализа. Поэтому в некоторых случаях, например при большом количестве повторов, целесообразнее реструктурировать данные так, чтобы можно было воспользоваться более простыми вариантами анализа, как это описано в разделе «Файлы данных для группы методов Общая линейная модель» (глава 14).

Пошаговые алгоритмы вычислений

При проведении многомерного дисперсионного анализа с повторными измерениями сначала необходимо выполнить три подготовительных шага.

Эти шаги (шаги 1–3) позволят подготовить рабочий файл данных, запустить программу IBM SPSS Statistics 19 и открыть файл (в данном случае — файл ex020.sav). Пошаговые инструкции этого процесса приведены в главе 4 (с. 60), а подробные разъяснения — в главе 2.

После завершения шага 3 на экране должно присутствовать окно редактора данных со строкой меню и загруженным файлом ex020.sav.

ШАГ 4

В меню Анализ выберите команду Общая линейная модель ► ОЛМ-повторные измерения. На экране появится диалоговое окно ОЛМ-повторные измерения: Задать факторы, показанное на рис. 16.1.

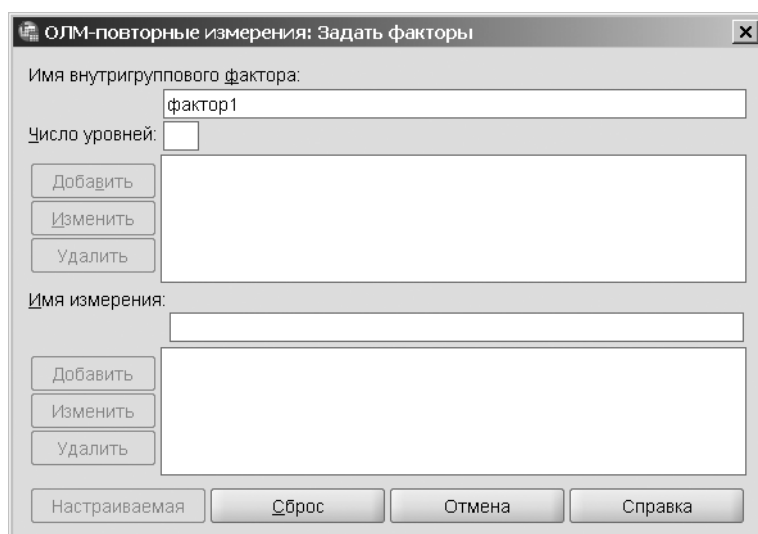


Рис. 16.1. Диалоговое окно ОЛМ-повторные измерения: Задать факторы

В этом окне вам предстоит задать имена внутригрупповых факторов. Обратите внимание, что имена, которые вы будете вводить в поле Имя внутригруппового фактора, относятся не к существующим переменным из файла данных, а определяют виртуальные переменные для команды ОЛМ-повторные измерения. В нашем примере присвоим этим переменным (внутригрупповым факторам) имена Ч_ряда (3 уровня: начало, середина, конец) и Отсрочка (2 уровня: нет, есть).

Для задания имени внутригрупповой переменной сначала необходимо ввести его в поле Имя внутригруппового фактора, затем задать число уровней переменной в поле Число уровней и щелкнуть на кнопке Добавить.

Довольно часто изучается динамика не одного показателя, а целой группы переменных, например, при оценке влияния внутригрупповых факторов (динамики) на несколько зависимых переменных. Для задания имен нескольких зависимых переменных, каждой из которых соответствует несколько повторов измерений предназначено поле Имя измерения. Так, результаты эксперимента из файла ex020.sav могут быть обработаны в предположении наличия трех *разных* зависимых переменных (запоминание слов): начало, середина и конец. Каждая из них измерена дважды: (1) непосредственно после предъявления слов и (2) с отсрочкой в два дня. В соответствии с этим планом будет произведен анализ влияния только одного внутригруппового фактора (отсрочка, 2 уровня) на три зависимых переменных. Для этого в поле Имя измерения необходимо последовательно ввести новые имена переменных, подтверждая задание каждого имени кнопкой Добавить.

После того как имена внутригрупповых факторов и переменных определены, чтобы завершить задание параметров анализа, щелкните на кнопке Настраиваемая. На экране появится диалоговое окно ОЛМ-повторные измерения, показанное на рис. 16.2.

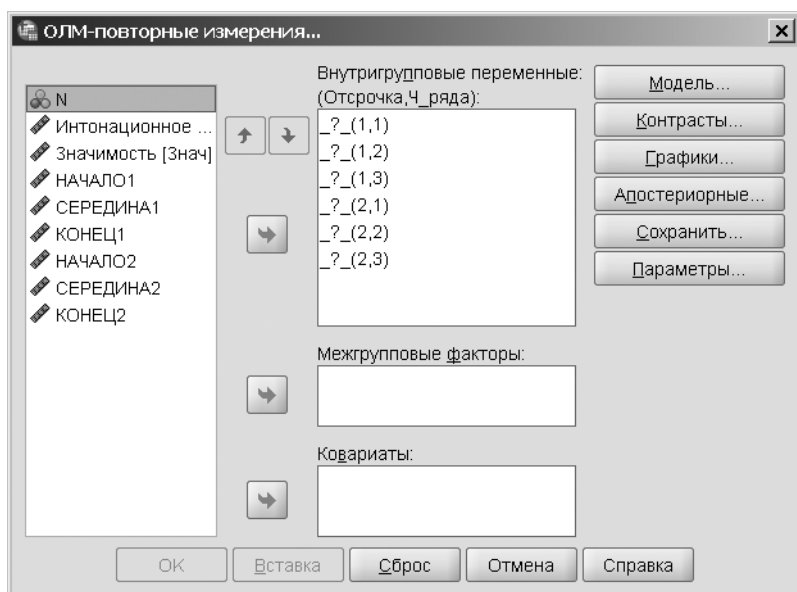


Рис. 16.2. Диалоговое окно ОЛМ-повторные измерения (два внутригрупповых фактора)

В левой части окна находится список переменных файла ex02.sav. В списке Внутригрупповые переменные вы можете указать соответствие между переменными файла и каждым уровнем внутригруппового фактора.

В нашем случае в анализ включены два внутригрупповых фактора, один из которых имеет 2, а другой — 3 уровня; таким образом, каждой из 6 комбинаций уровней необходимо сопоставить переменную из исходного списка. Пары чисел, указанные в скобках в списке Внутригрупповые переменные, определяют, к какой ячейке модели относятся переменные файла, которые далее будут обозначаться как внутригрупповые переменные (табл. 16.2).

Обратите внимание, что при движении от начала к концу списка переменных файла (первый столбец табл. 16.2) градации фактора «отсрочка» меняются реже, чем градации фактора «часть ряда». Поэтому в качестве первого фактора лучше выбирать тот, который имеет меньше уровней. Несмотря на то что эта иллюстрация относится к случаю двух факторов с небольшим числом уровней, она с легкостью может быть обобщена на любое число факторов и их уровней.

Таблица 16.2. Соответствие переменных файла и ячеек модели с двумя факторами

Переменная	Ячейка	Отсрочка	Часть ряда
начало1	(1,1)	Нет — 1	Начало — 1
середина1	(1,2)	Нет — 1	Середина — 2
конец1	(1,3)	Нет — 1	Конец — 3
начало2	(2,1)	Есть — 2	Начало — 1
середина2	(2,2)	Есть — 2	Середина — 2
конец2	(2,3)	Есть — 2	Конец — 3

Альтернативная схема анализа проще: она предполагает наличие одного внутригруппового фактора Отсрочка, но три разных зависимых переменных (начало, середина, конец). Соответствие переменных файла ячейкам модели показано в табл. 16.3.

Таблица 16.3. Соответствие переменных файла и ячеек модели с одним фактором

Переменная	Ячейка	Отсрочка
начало1	(1, начало)	Нет — 1
начало2	(2, начало)	Есть — 2
середина1	(1, середина)	Нет — 1
середина2	(2, середина)	Есть — 2
конец1	(1, конец)	Нет — 1
конец2	(2, конец)	Есть — 2

Для того чтобы ввести какую-либо переменную в список Внутригрупповые переменные, достаточно выделить ее имя в списке слева и щелкнуть на верхней кнопке с направленной вправо стрелкой.

По умолчанию SPSS сопоставляет первую введенную переменную первой ячейке модели (в нашем случае — ячейке 1,1), вторую переменную — второй ячейке (1,2) и т. д. Если же вы хотите нарушить данный порядок, воспользуйтесь кнопками с направленными вверх и вниз стрелками, чтобы переместить введенную переменную в нужное место списка ячеек.

После того как все внутригрупповые зависимые переменные анализа определены, вы можете указать все межгрупповые переменные (факторы) и ковариаты. Для этого переместите нужные переменные из исходного списка в списки Межгрупповые факторы и Ковариаты.

В правой части диалогового окна ОЛМ-повторные измерения расположены 6 кнопок: Модель, Контрасты, Графики, Апостериорные, Сохранить и Параметры. Щелчок на кнопке Параметры открывает диалоговое окно, идентичное окну ОЛМ-многомерная: Параметры, с которым мы познакомились в главе 15. Кнопка Диаграммы работает так же, как ее аналог из глав 14 и 15, но позволяет задавать диаграммы не только для межгрупповых, но и для внутригрупповых факторов. Функции кнопки Апостериорные также остались прежними; единственным отличием является то, что парные апостериорные сравнения применимы только к межгрупповым факторам. За описанием кнопок Контрасты и Сохранение обратитесь к руководству пользователя.

При щелчке на кнопке Модель открывается диалоговое окно ОЛМ-повторные измерения: Модель, представленное на рис. 16.3. В большинстве случаев исследователи предпочитают пользоваться полной факторной моделью и оставляют переключатель Полная факторная в группе Задать модель. Тем не менее иногда требуется задать набор главных эффектов и взаимодействий вручную. Для этого следует установить переключатель Настраиваемая, а затем переместить желаемые эффекты и взаимодействия из левых списков в правые. Заметьте, что вы не сможете переместить межгрупповые факторы во внутригрупповую секцию модели, и наоборот, так как это некорректно. После того как желаемая модель определена, щелкните на кнопке Продолжить, чтобы вернуться в диалоговое окно ОЛМ-повторные измерения.

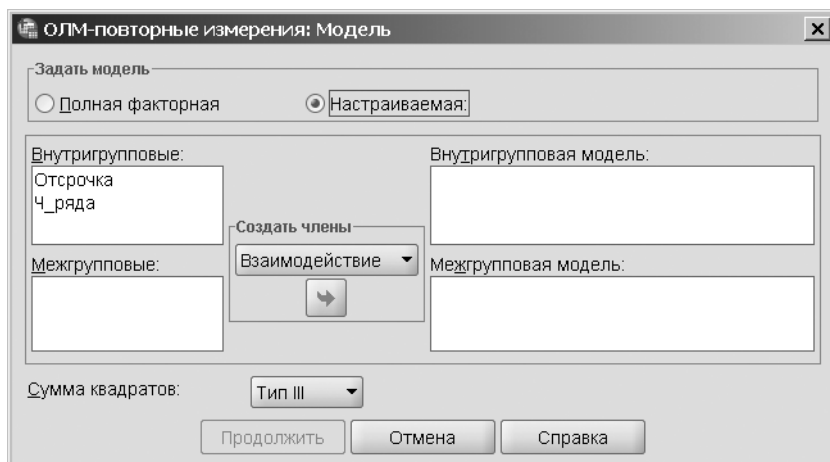


Рис. 16.3. Диалоговое окно ОЛМ-повторные измерения: Модель

Далее будут приведены два примера использования команды ОЛМ-повторные измерения. В первом примере мы проведем дисперсионный анализ с двумя внутригрупповыми факторами: Часть ряда (3 градации) и Отсрочка воспроизведения (2 градации). С каждой из 6 ячеек модели будет связана одна из переменных начало1, середина1, конец1, начало2, середина2, конец2.

ШАГ 5

После выполнения предыдущего шага должно быть открыто диалоговое окно ОЛМ-повторные измерения: Задать факторы, показанное на рис. 16.1. Если вы уже успели поработать с этим окном, щелкните на кнопке Сброс.

1. С помощью клавиши Delete или Backspace удалите имя фактор1 из поля Имя внутригруппового фактора, введите туда имя Отсрочка, в поле Число уровней введите число 2 и щелкните на кнопке Добавить.
2. В поле Имя внутригруппового фактора введите имя Ч_ряда, в поле Number of (Число уровней), введите число 3 и щелкните сначала на кнопке Добавить, а потом — на кнопке Настраиваемая, чтобы открыть диалоговое окно ОЛМ-повторные измерения, показанное на рис. 16.2.
3. Наведите указатель мыши на переменную начало1, нажмите кнопку мыши и, не отпуская ее, переместите указатель на переменную конец2 и отпустите кнопку. В результате окажутся выделенными шесть переменных: начало1, середина1, конец1, начало2, середина2, конец2.
4. Щелкните на верхней кнопке с направленной вправо стрелкой, чтобы переместить выделенные переменные в список Внутригрупповые переменные.
5. Выделите в левом окне переменную Инт и нажатием на среднюю кнопку со стрелкой перенесите ее в правое поле Межгрупповые факторы.
6. Щелкните на кнопке Параметры, чтобы открыть диалоговое окно ОЛМ-повторные измерения: Параметры.
7. Установите флажки Описательные статистики и Оценки размера эффекта, а затем щелкните на кнопке Продолжить, чтобы вернуться в диалоговое окно ОЛМ-повторные измерения.
8. Щелкните на кнопке Графики, чтобы открыть диалоговое окно ОЛМ-повторные измерения: Графики профилей, аналогичное показанному на рис. 15.3.
9. Щелкните сначала на переменной Ч_ряда, чтобы выделить ее, а затем — на верхней кнопке со стрелкой, чтобы переместить переменную в поле Горизонтальная ось.
10. Щелкните сначала на переменной Инт, чтобы выделить ее, а затем — на средней кнопке со стрелкой, чтобы переместить переменную в поле Отдельные линии.
11. Щелкните сначала на кнопке Добавить, чтобы зафиксировать заданные параметры построения диаграмм, а затем — на кнопке Продолжить, чтобы вернуться в диалоговое окно ОЛМ-повторные измерения.
12. Щелкните на кнопке ОК, чтобы открыть окно вывода.

Следующий пример реализует альтернативную схему анализа тех же данных. В нем проводится смешанный анализ влияния одного внутригруппового фактора Отсрочка и одного межгруппового фактора интонационного выделения (Инт) на три зависимые переменные (начало, середина, конец).

ШАГ 5А

После выполнения предыдущего шага должно быть открыто диалоговое окно ОЛМ-повторные измерения: Задать факторы, показанное на рис. 16.1. Если вы уже успели поработать с этим окном, щелкните на кнопке Сброс.

1. С помощью клавиши Delete или Backspace удалите имя Фактор1 из поля Имя внутригруппового фактора, введите туда имя Отсрочка, а в поле Число уровней введите число 2 и щелкните на кнопке Добавить.
2. В окно Имя измерения введите слово начало и щелкните на кнопке Добавить.
3. Повторите предыдущее действие со словами середина и конец. Щелкните на кнопке Настраиваемая, чтобы открыть диалоговое окно ОЛМ-повторные измерения, показанное на рис. 16.2.
4. Наведите указатель мыши на переменную начало1, нажатием левой кнопки мыши выделите ее, нажмите на клавиатуре клавишу Ctrl и, не отпуская ее, переместите указатель мыши на переменную начало2 и выделите ее нажатием левой кнопки мыши. Отпустите клавишу Ctrl. В результате окажутся выделенными две переменных: начало1 и начало2.
5. Щелкните на верхней кнопке с направленной вправо стрелкой, чтобы переместить выделенные переменные в список Внутригрупповые переменные. Они займут первые две позиции: (1, начало), (2, начало).
6. Повторите пункты 4 и 5 в отношении пар переменных середина1, середина2 и конец1, конец2 так, чтобы они заняли следующие пары позиций в правом поле.
7. Щелкните сначала на переменной Инт, чтобы выделить ее, а затем — на средней кнопке с направленной вправо стрелкой, чтобы переместить переменную в список Межгрупповые факторы.
8. Щелкните на кнопке Графики, чтобы открыть диалоговое окно ОЛМ-повторные измерения: Диаграммы профилей, аналогичное показанному на рис. 15.3.
9. Щелкните сначала на переменной Отсрочка, чтобы выделить ее, а затем — на верхней кнопке со стрелкой, чтобы переместить переменную в поле Горизонтальная ось.
10. Щелкните сначала на переменной Инт, чтобы выделить ее, а затем — на средней кнопке со стрелкой, чтобы переместить переменную в поле Отдельные линии.
11. Щелкните сначала на кнопке Добавить, чтобы зафиксировать заданные параметры построения диаграмм, а затем — на кнопке Продолжить, чтобы вернуться в диалоговое окно ОЛМ-повторные измерения.

12. Щелкните на кнопке Параметры, чтобы открыть диалоговое окно ОЛМ-повторные измерения: Параметры.
13. Установите флажки Описательные статистики, а затем щелкните на кнопке Продолжить, чтобы вернуться в диалоговое окно ОЛМ-повторные измерения.
14. Щелкните на кнопке ОК, чтобы открыть окно вывода.

После выполнения шага 5 программа автоматически откроет окно вывода. Для просмотра результатов при необходимости можно воспользоваться вертикальной и горизонтальной полосами прокрутки. Обратите внимание на стандартную строку меню в верхней части окна вывода: ее присутствие позволяет выполнять любые статистические операции, не переключаясь обратно в окно редактора данных.

Представление результатов

Большая часть результатов, генерируемых командой ОЛМ-повторные измерения, аналогична результатам многомерного дисперсионного анализа (глава 15), поэтому ниже будут представлены лишь те фрагменты, которые являются специфичными для внутригрупповых моделей. Также не приводятся некоторые результаты, которые являются избыточными и практически не требуют интерпретации.

Многомерные критерии

Таблица Многомерные критерии, показанная на рис. 16.4, получена после выполнения шага 5. В ней удалена часть строк, не представляющих интереса для исследователя (по вопросам редактирования окна вывода обращайтесь к соответствующему разделу главы 2).

Многомерные критерии							
Эффект		Значения	F	Ст. св. гипотезы	Ст.св. ошибки	Знач.	Частная Эта в Квадрате
Отсрочка * Инт	След Пиллая	,014	,253	1,000	18,000	,621	,014
	Лямбда Уилкса	,986	,253	1,000	18,000	,621	,014
	След Хотеллинга	,014	,253	1,000	18,000	,621	,014
	Наибольший корень Гоя	,014	,253	1,000	18,000	,621	,014
Ч_ряда * Инт	След Пиллая	,829	41,271	2,000	17,000	,000	,829
	Лямбда Уилкса	,171	41,271	2,000	17,000	,000	,829
	След Хотеллинга	4,855	41,271	2,000	17,000	,000	,829
	Наибольший корень Гоя	4,855	41,271	2,000	17,000	,000	,829
Отсрочка * Ч_ряда	След Пиллая	,022	,193	2,000	17,000	,826	,022
	Лямбда Уилкса	,978	,193	2,000	17,000	,826	,022
	След Хотеллинга	,023	,193	2,000	17,000	,026	,022
	Наибольший корень Гоя	,023	,193	2,000	17,000	,826	,022
Отсрочка * Ч_ряда * Инт	След Пиллая	,622	13,998	2,000	17,000	,000	,622
	Лямбда Уилкса	,378	13,998	2,000	17,000	,000	,622
	След Хотеллинга	1,647	13,998	2,000	17,000	,000	,622
	Наибольший корень Гоя	1,647	13,998	2,000	17,000	,000	,622

Рис. 16.4. Фрагмент окна вывода после выполнения шага 5: Многомерные критерии

Как видно из таблицы, взаимодействие факторов Часть ряда и Интонационное выделение ($\text{Ч_ряда} * \text{Инт}$) статистически достоверно. Это значит, что влияние интонационного выделения на воспроизведение слов проявляется по-разному, в зависимости от части ряда. Более детальную интерпретацию позволяет сделать график средних значений, приведенных на рис. 16.5.



Рис. 16.5. Фрагмент окна вывода после выполнения шага 5 (график средних значений)

Отметим, что тот же результат был получен с применением более простого метода, при условии реструктурирования данных (глава 14).

Проверка внутригрупповых эффектов

Таблица, показанная на рис. 16.6, получена после выполнения шага 5а.

Полученные результаты во многом дублируют результаты предшествующего применения многомерных критериев. Таблица содержит главный эффект внутригруппового фактора (Отсрочка), а также эффект его взаимодействия с межгрупповым фактором ($\text{Отсрочка} * \text{Инт}$). Все эффекты оказываются статистически достоверными. Обратите внимание на то, что данные результаты относятся к совокупности всех трех зависимых переменных. На это указывает подзаголовок таблицы «Многомерный».

Если многомерные критерии внутригрупповых эффектов дают статистически достоверные результаты, то для детальной интерпретации результатов следует обратиться к следующей таблице внутригрупповых эффектов: «Одномерные критерии», которая приведена на рис. 16.7 в отредактированном виде.

Многомерный ^{б,с}						
Внутригрупповой эффект		Значения	F	Ст. св. гипотезы	Ст.св. ошибки	Знч.
Отсрочка	След Пиллая	,861	33,101 ^а	3,000	16,000	,000
	Лямбда Уилкса	,139	33,101 ^а	3,000	16,000	,000
	След Хотеллинга	6,206	33,101 ^а	3,000	16,000	,000
	Наибольший корень Роя	6,206	33,101 ^а	3,000	16,000	,000
Отсрочка * Инт	След Пиллая	,632	9,151 ^а	3,000	16,000	,001
	Лямбда Уилкса	,368	9,151 ^а	3,000	16,000	,001
	След Хотеллинга	1,716	9,151 ^а	3,000	16,000	,001
	Наибольший корень Роя	1,716	9,151 ^а	3,000	16,000	,001

а. Точная статистика
 б. План: Свободный член + Инт
 В пределах объектов Design: Отсрочка
 с. Критерии основаны на усредненных переменных.

Рис. 16.6. Фрагмент окна вывода после выполнения шага 5а (проверка внутригрупповых эффектов: многомерные критерии)

Одномерные критерии показывают, что взаимодействие факторов Инт и Отсрочка является статистически достоверным в отношении переменных Середина и Конец. Для интерпретации взаимодействия следует обратиться к графикам средних значений (рис. 16.8).

Судя по графику для переменной середина, взаимодействие факторов Инт и Отсрочка обусловлено тем, что на графике наблюдается более крутой наклон линии для случая интонационного выделения. Иначе говоря, на продуктивность воспроизведения середины ряда слов отсрочка сильнее влияет в том случае, когда слова были интонационно выделены.

Одномерные критерии							
Источник	Измерение		Сумма квадратов типа III	ст.св.	Средний квадрат	F	Знч.
Отсрочка * Инт	Начало	Предполагаая сферичность	,025	1	,025	,086	,773
		Гринхауз-Гайссер	,025	1,000	,025	,086	,773
		Юнха-Фельдта	,025	1,000	,025	,086	,773
		Ограниченный снизу	,025	1,000	,025	,086	,773
	Середина	Предполагаая сферичность	2,500	1	2,500	8,03	,011
		Гринхауз-Гайссер	2,500	1,000	2,500	8,03	,011
		Юнха-Фельдта	2,500	1,000	2,500	8,03	,011
		Ограниченный снизу	2,500	1,000	2,500	8,03	,011
	Конец	Предполагаая сферичность	4,225	1	4,225	8,22	,010
		Гринхауз-Гайссер	4,225	1,000	4,225	8,22	,010
		Юнха-Фельдта	4,225	1,000	4,225	8,22	,010
		Ограниченный снизу	4,225	1,000	4,225	8,22	,010

Рис. 16.7. Фрагмент окна вывода после выполнения шага 5а (проверка внутригрупповых эффектов: одномерные критерии)

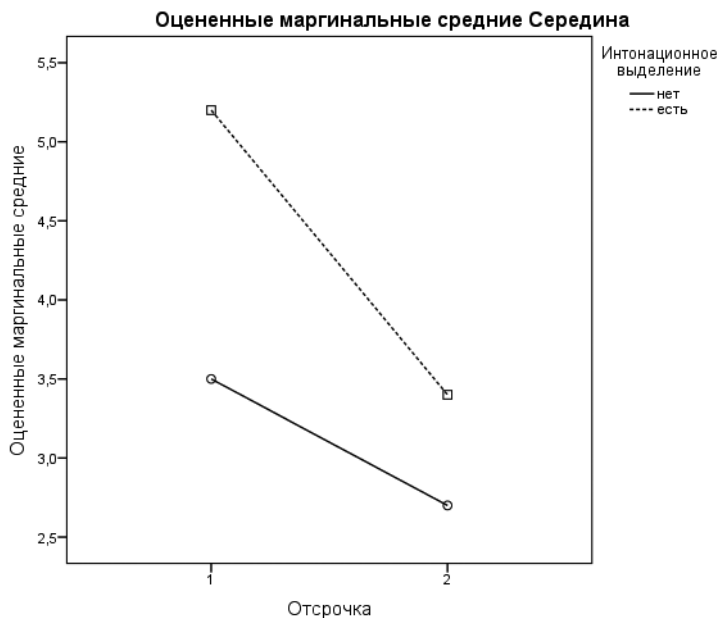


Рис. 16.8. Фрагмент окна вывода после выполнения шага 5а (график средних)

Оценка эффектов межгрупповых факторов

Таблица, показанная на рис. 16.9, получена после выполнения шага 5а.

Межгрупповые эффекты фактора инт демонстрируют статистически достоверные результаты в отношении переменных середина и конец. Это значит, что влияние интонационного выделения на эти переменные проявляется независимо от отсрочки.

Оценка эффектов межгрупповых факторов

Преобразуемая переменная: Среднее

Источник	Измерение	Сумма квадратов типа III	ст.св.	Средний квадрат	F	Знач.
Свободный член	Начало	378,225	1	378,225	109,719	,000
	Середина	547,600	1	547,600	394,272	,000
	Конец	429,025	1	429,025	291,964	,000
Инт	Начало	7,225	1	7,225	2,096	,165
	Середина	14,400	1	14,400	10,368	,005
	Конец	9,025	1	9,025	6,142	,023
Ошибка	Начало	62,050	18	3,447		
	Середина	25,000	18	1,389		
	Конец	26,450	18	1,469		

Рис. 16.9. Фрагмент окна вывода после выполнения шага 5а (оценка эффектов межгрупповых факторов)

Завершение анализа и выход из программы

Отредактируйте содержимое окна вывода в соответствии со своими предпочтениями: скройте лишнюю информацию, исправьте таблицы и пр. (см. раздел «Окно вывода и его редактирование» главы 2). За инструкциями по редактированию графиков обратитесь к главе 5.

Для дальнейшего использования окончательного результата все содержимое окна вывода или его фрагменты можно сохранить в файле *.spv, экспортировать в другой формат (например, Word), перенести в документ Word или вывести на печать (подробности см. в разделе «Сохранение, экспорт, перенос и печать результатов» главы 2).

Для выхода из программы выберите команду Выход в меню Файл.

17 Простая линейная регрессия

239	Простая регрессия
241	Оценка криволинейности
243	Пошаговые алгоритмы вычислений
247	Представление результатов
249	Терминология, используемая при выводе
250	Завершение анализа и выход из программы

Команда Линейная регрессия позволяет выполнять как простой, так и множественный регрессионный анализ. Простой регрессионный анализ обсуждается в этой главе, а множественный — в следующей. Такое разбиение одного статистического раздела было сделано нами в целях упрощения изложения материала, касающегося основ регрессионного анализа. Если вы не знакомы с множественным регрессионным анализом, эта глава послужит введением. Мы рассмотрим такие понятия, как прогнозируемые значения зависимой переменной и уравнение регрессии, покажем связь между простой регрессией и корреляцией двух переменных, рассмотрим влияние одной переменной на дисперсию другой, а также ознакомимся с оценкой криволинейности связи двух переменных.

Простая регрессия

Зачастую, имея информацию об одной характеристике объекта, мы пытаемся на основе этой информации делать выводы о другой его характеристике, связанной с исходной. Например, если мы знаем рост человека, мы могли бы попытаться оценить его возможный вес. Например, зная, что человек имеет рост 214 см, с определенной вероятностью мы можем прогнозировать, что вес этого человека превысит 91 кг. Можно привести множество пар связанных между собой величин: коэффициент интеллекта (IQ) и академическая успеваемость, число сокращений мышц ног в секунду и скорость бега, калорийность пищи и вес, симпатия к человеку и желание оказать ему помощь и т. д. Каждый день мы оцениваем связи различных факторов между собой. Именно на проведение таких оценок и нацелена простая регрессия. Этот метод не может дать абсолютно достоверного результата, однако способен ответить на вопрос о связи переменных, а также по заданному значению одной переменной рассчитать наиболее вероятное значение другой переменной.

Для иллюстрации практического применения простой регрессии воспользуемся новым файлом данных с именем `exam.sav`. Он содержит 10-балльную оценку нервной возбудимости (тревожности) 36 студентов и количество решенных ими зачетных тестовых задний (из 20 возможных). Гипотеза о линейности отношения этих двух переменных говорит о том, что чем выше нервная возбудимость студента, тем выше его результативность (например, потому, что спокойных студентов меньше волнуют их знания, а тревожные студенты проводят больше времени за подготовкой к зачету). В качестве зависимой переменной (критерия) выступает переменная `тест`, а в качестве независимой переменной (предиктора) — переменная `трев`. Мы будем пытаться прогнозировать значения переменной `тест` по известным значениям переменной `трев`, используя *уравнение регрессии*. Уравнение регрессии формируется на основе общего уравнения, связывающего фактическую успеваемость студента и нервную возбудимость:

$$\text{тест}_{\text{истина}} = \text{константа} + \text{коэффициент} \times \text{трев} + \text{остаток}$$

Здесь $\text{тест}_{\text{истина}}$ — переменная, отражающая реальный результат выполнения тестового задания, *константа* — некоторая константа, *коэффициент* — регрессионный коэффициент при значении оценки тревожности, *остаток* — остаток. В результате применения линейного регрессионного анализа константа оказалась равной 9,3114, а коэффициент регрессии 0,6751. Соответственно, уравнение для прогноза результата зачетного тестирования выглядит следующим образом:

$$\text{тест}_{\text{прогноз}} = 9,3114 + 0,6751 \times \text{трев}$$

$$\text{тест}_{\text{истина}} = \text{тест}_{\text{прогноз}} + \text{остаток}$$

Прогнозируемое значение будет отличаться от истинного значения. Разница отражает тот факт, что рассчитываемые результаты экзамена никогда не бывают абсолютно точными, и чтобы получить истинный результат, необходимо ввести в уравнение член, равный разности прогнозируемого и реального значений. Этот член и называют *остатком*:

$$\text{реальное значение} = \text{прогнозируемое значение} + \text{остаток}$$

Можно предположить, что, помимо нервной возбудимости, на результаты зачетного тестирования влияют другие факторы. Регрессионный анализ, учитывающий влияние нескольких факторов, называется *множественным* и рассматривается в следующей главе.

Ниже перечислены величины, которые вычисляются при проведении регрессионного анализа с помощью команд подменю *Линейная регрессия*.

- Коэффициент, характеризующий связь между значениями зависимой и независимой переменных, обозначается прописной буквой *R*, набранной курсивом, и представляет собой уже знакомый нам по предыдущим главам коэффициент корреляции. Использование прописной буквы вместо строчной в окне вывода результатов регрессионного анализа объясняется тем, что в общем случае речь идет о *множественных* корреляциях с участием нескольких независимых переменных. Корреляция подробно рассмотрена в главе 9.

- Для рассчитанного значения множественного коэффициента корреляции R программа определяет уровень значимости p . Как и ранее, величина $p < 0,05$ свидетельствует о значимой корреляции переменных. При $p > 0,05$ вероятность случайности результата считается слишком высокой, и в этом случае говорят, что связь между переменными слабая или не обнаружена.
- Значение R^2 является обычным квадратом величины R , однако этот коэффициент имеет собственный смысл. Коэффициент R^2 характеризует долю дисперсии одной переменной, обусловленной воздействием другой переменной. Так, для переменных трев и тест значение $R = 0,546$, а $R^2 = 0,298$. Это означает, что 29,8 % дисперсии переменной тест объясняется влиянием независимой переменной трев.
- Константу и коэффициент уравнения регрессии часто называют B -величинами. B -величины характеризуют связь между значениями переменных трев и тест.

Оценка криволинейности

Выдвинутая в предыдущем разделе гипотеза о том, что увеличение нервной возбудимости перед экзаменом всегда улучшает результат студента, очевидно, вызывает сомнения. Более разумными кажутся следующие соображения. При низкой возбудимости результаты экзамена должны быть низкими, поскольку излишнее спокойствие снижает потребность студента в подготовке. С увеличением возбудимости результат до определенного момента должен улучшаться, однако слишком возбудимые студенты вряд ли способны сконцентрироваться и показать хороший результат. Поэтому наилучших показателей должны добиваться те студенты, чей уровень возбудимости является промежуточным. При регрессионном анализе, вне зависимости от того, является он множественным или простым, соотношение между зависимой и независимой переменными считается линейным. В приведенном ранее примере мы видим значительную корреляцию между переменными трев и тест ($R = 0,546$, $p < 0,001$), однако возможная ошибка прогноза велика (только 29,8 % дисперсии переменной тест объясняется влиянием переменной трев). Можно предположить, что если изменить вид общего уравнения (например, включить в него квадрат переменной трев), прогнозируемые значения будут ближе к реальным. График на рис. 17.1 может быть получен при обращении Графика ► Устаревшие диалоговые окна Рассеяния/точки. Он наглядно демонстрирует характер отношений между переменными трев и тест.

По вертикальной оси диаграммы отложены значения переменной тест, по горизонтальной — значения переменной трев. Очевидно, что зависимость между переменными не является линейной: у нее имеется выраженный максимум, а при движении к краям диаграммы наблюдается убывание значений переменной тест. Нельзя не отметить, что только путем графической интерпретации невозможно достоверно установить характер отношений между переменными. Необходимым условием является применение статистических критериев.

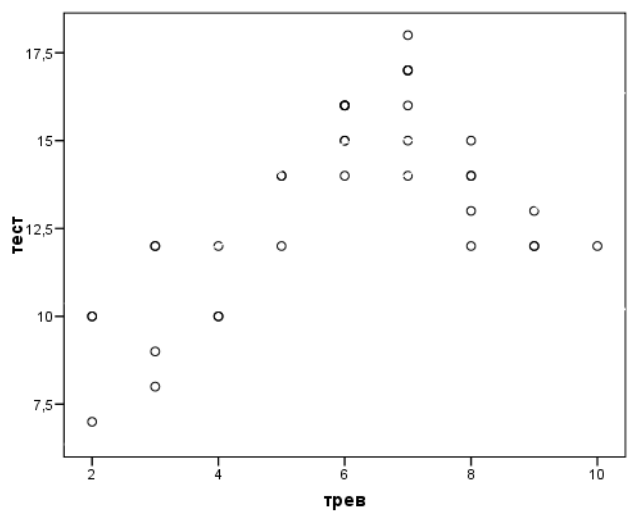


Рис. 17.1. Диаграмма рассеивания, демонстрирующая криволинейную зависимость

Для того чтобы статистически оценить криволинейность, в подменю Регрессия предусмотрена команда Подгонка кривых. В ее диалоговом окне необходимо задать зависимую переменную (тест), независимую переменную (трев) и установить флажки Линейная и Квадратичная. После щелчка на кнопке ОК появятся результаты, приведенные ниже (рис. 17.2), а также диаграмма, показанная на рис. 17.3.

Зависимая переменная: тест

Уравнение	Сводка модели					Оценки параметров		
	R квадрат	F	ст.св1	ст.св2	Знач.	Константа	b1	b2
Линейная	,298	14,466	1	34	,001	9,311	,675	
Квадратичная	,675	34,283	2	33	,000	,161	4,490	-,338

Независимой переменной является трев.

Рис. 17.2. Таблица результатов оценки криволинейности

По диаграмме можно оценить, насколько близка к линейной или квадратичной зависимость между значениями переменных. В результаты включены значения коэффициентов *B* регрессии (Константа b_0 , b_1 , b_2), поэтому не сложно составить линейное и квадратичное уравнения регрессии для прогнозируемых значений. На диаграмму помимо линейной и квадратичной зависимостей нанесен разброс данных файла. Обратите внимание на сходство диаграмм, представленных на рис. 17.1 и 17.3, а также на то, что константа и коэффициенты уравнений указаны в последних трех столбцах выводимых результатов (см. рис. 17.2).

Для линейной регрессии уравнение имеет вид:

$$\text{тест}_{\text{прогноз}} = 9,3114 + 0,6751 \times \text{трев}$$

А уравнение для квадратичной регрессии выглядит следующим образом:

$$\text{тест}_{\text{прогноз}} = 0,1615 + 4,4896 \times \text{трев} - 0,3381 \times (\text{трев})^2$$

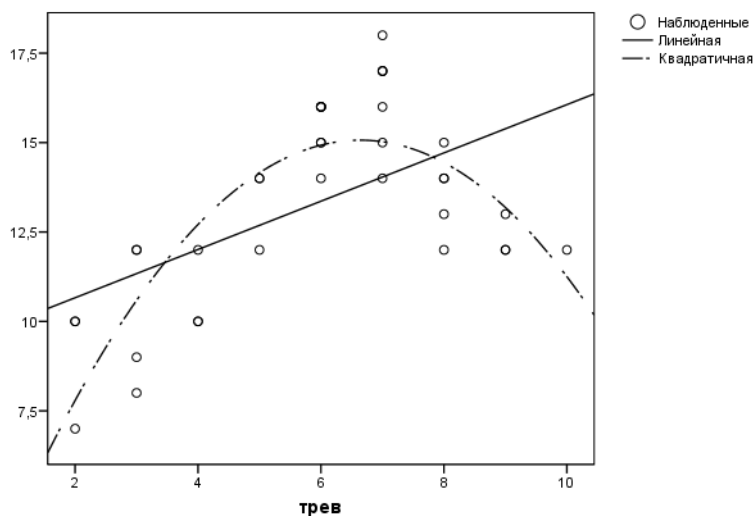


Рис. 17.3. Диаграмма, демонстрирующая линейную и криволинейную тенденции

Как видно из рис. 17.2, в случае линейной регрессии величина R^2 (столбец R квадрат в таблице выводимых результатов) равна 0,298, то есть 29,8 % дисперсии переменной тест обусловлено воздействием со стороны переменной трев. В то же время для квадратичной регрессии, которая учитывает и линейную, и криволинейную связи, $R^2 = 0,675$, то есть она обуславливает 67,5 % дисперсии переменной тест. Малый p -уровень для обоих уравнений свидетельствует об очень высокой статистической достоверности полученных результатов. Очевидно, что квадратичная регрессия описывает отношения между переменными тест и трев более адекватно, чем линейная.

Пошаговые алгоритмы вычислений

Сначала необходимо выполнить три подготовительных шага. Эти шаги (шаги 1–3) позволят подготовить рабочий файл данных, запустить программу IBM SPSS Statistics 19, и открыть файл (в данном случае — файл exam.sav). Пошаговые инструкции этого процесса приведены в главе 4 (с. 60), а подробные разъяснения — в главе 2.

После завершения шага 3 на экране должно присутствовать окно редактора данных со строкой меню и загруженным файлом exam.

Простой регрессионный анализ

ШАГ 4

В меню Анализ выберите команду Регрессия ► Линейная. На экране появится диалоговое окно Линейная регрессия, показанное на рис. 17.4.

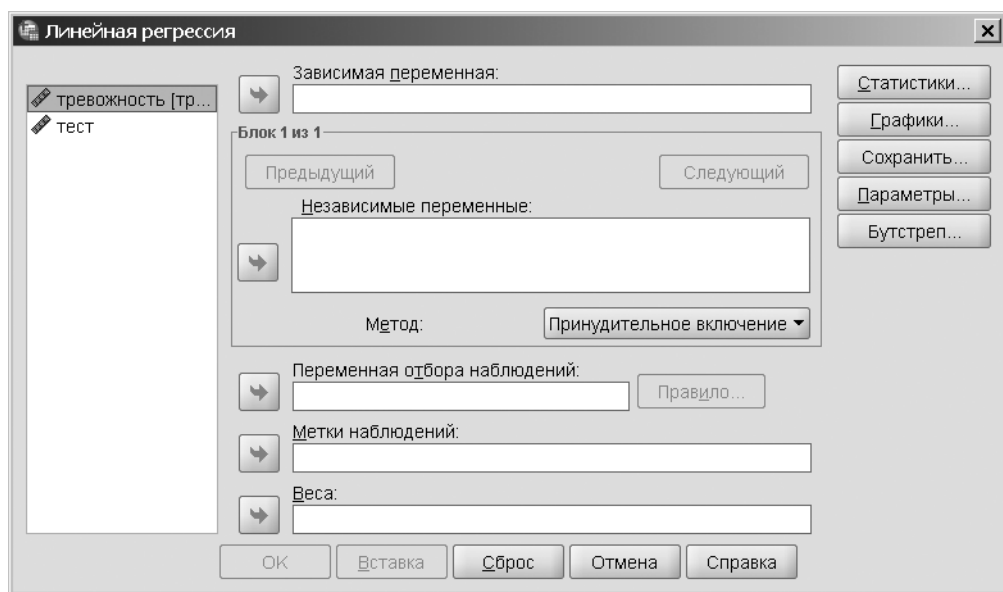


Рис. 17.4. Диалоговое окно Линейная регрессия

Диалоговое окно *Линейная регрессия* позволяет проводить разные виды регрессионного анализа. Поскольку в данном случае выполняется простой регрессионный анализ, некоторые из элементов интерфейса этого окна мы рассматривать не будем.

Поскольку в файле *exam.sav* содержится только две переменные, в списке, находящемся в левой части диалогового окна, есть лишь два имени — *трев* и *тест*. Позже нами будет создана еще одна переменная с именем *трев2*. Для того чтобы провести регрессионный анализ, достаточно выполнить несколько простых действий: поместить переменную *тест* в поле *Зависимая переменная*, переменную *трев* — в список *Независимые переменные* и щелкнуть на кнопке *ОК*. В результате программа вычислит величины R , R^2 , F , соответствующие им p -уровни, B -коэффициенты (коэффициенты и константы уравнения регрессии), а также стандартизированные β -коэффициенты, характеризующие степень зависимости между значениями исследуемых переменных.

В следующем примере мы приводим последовательность действий, необходимую для выполнения простого регрессионного анализа с участием независимой переменной *трев* и зависимой переменной *тест*.

ШАГ 5

После выполнения шага 4 должно быть открыто диалоговое окно *Линейная регрессия*, показанное на рис. 17.4.

1. Щелкните сначала на переменной *тест*, чтобы выделить ее, а затем — на верхней кнопке со стрелкой, чтобы переместить переменную в поле *Зависимая переменная*.

2. Щелкните сначала на переменной трев, чтобы выделить ее, а затем — на второй сверху кнопке со стрелкой, чтобы переместить переменную в список Независимые переменные.
3. Щелкните на кнопке ОК, чтобы открыть окно вывода.

Анализ криволинейных зависимостей

Мы рассмотрим две пошаговые процедуры, относящиеся к анализу криволинейных зависимостей. Первая позволяет получить результаты, приведенные в начале этой главы (см. таблицу на рис. 17.2 и диаграмму, показанную на рис. 17.3), вторая создает специальную «квадратичную» переменную, которую при желании можно использовать в дальнейших вычислениях. Несмотря на то что криволинейные зависимости рассматриваются в контексте простого регрессионного анализа, все сказанное в этой главе справедливо и для множественного регрессионного анализа.

После завершения шага 3 на экране должно присутствовать окно редактора данных.

ШАГ 4А

В меню Анализ выберите команду Регрессия ► Подгонка кривых. На экране появится диалоговое окно Подгонка кривых, показанное на рис. 17.5.

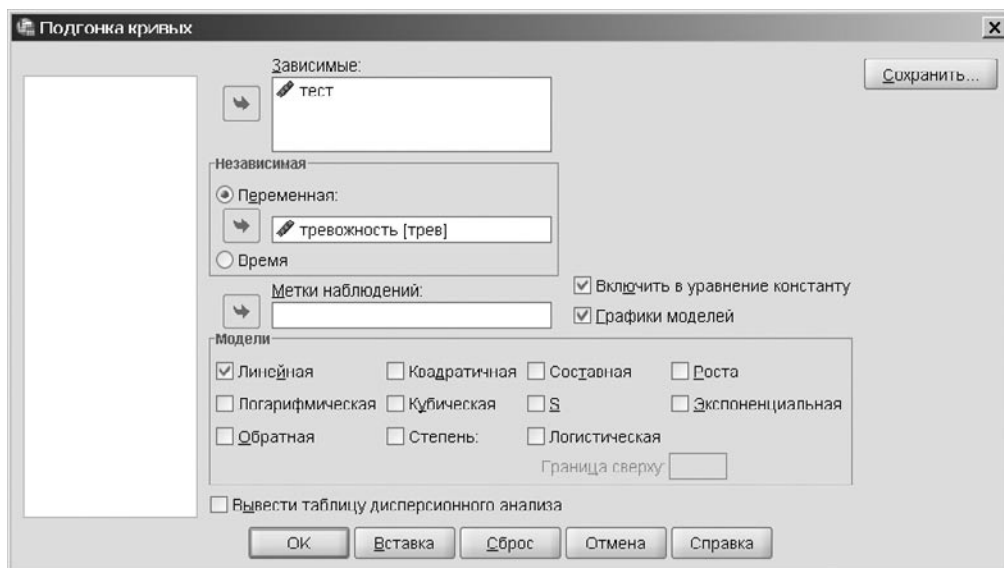


Рис. 17.5. Диалоговое окно Подгонка кривых

Первоначальные действия в окне Подгонка кривых те же, что и в окне Линейная регрессия: поместите переменную тест в список Зависимые, а переменную трев — в поле Переменная: Независимая. Обратите внимание на то, что по умолчанию в диалоговом окне установлены три флажка: флажок Включить в уравнение константу, а также флажок Графики моделей, позволяющий получить диаграмму

(см. рис. 17.3), и флажок *Линейная*, обеспечивающий проверку гипотезы о прямолинейности и включение в диаграмму прямой линии.

Как можно видеть, в группе *Модели* помимо флажка *Линейная* имеются множество других флажков, позволяющих проводить проверки на соответствие изучаемого распределения самым разным видам кривых и включать эти кривые в диаграмму. В следующем примере мы продемонстрируем проверку зависимости на квадратичность.

ШАГ 5А

Ниже перечислены действия, позволяющие провести проверку криволинейности. После выполнения шага 4а должно быть открыто диалоговое окно *Подгонка кривых*, показанное на рис. 17.5.

1. Щелкните сначала на переменной *трев*, чтобы выделить ее, а затем — на верхней кнопке со стрелкой, чтобы переместить переменную в список *Зависимые*.
2. Щелкните сначала на переменной *трев*, чтобы выделить ее, затем — на второй сверху кнопке со стрелкой, чтобы переместить переменную в поле *Независимая Переменная*; установите флажок *Квадратичная*.
3. Щелкните на кнопке *ОК*, чтобы открыть окно вывода.

При составлении квадратного уравнения регрессии SPSS вычисляет новую переменную, значения которой равны квадратам соответствующих значений переменной *трев*. Как вы, наверное, помните из школьного курса алгебры, квадратичная зависимость графически имеет вид параболы — кривой с двумя ветвями, которые могут быть направлены вверх, если коэффициент при квадратном члене положительный, или вниз, если коэффициент при квадратном члене отрицательный. Поскольку квадратное уравнение имеет не только квадратный, но и положительный линейный член, изображаемый в виде возрастающей наклонной прямой, конец левой ветви параболы расположен ниже, чем конец правой ветви.

Для того чтобы включить в анализ регрессии квадрат переменной *трев*, нам необходимо «вручную» создать переменную *трев2*, содержащую квадраты значений переменной *трев*, а затем указать ее в качестве независимой переменной. Процесс создания новых переменных описан в разделе «Преобразование данных» главы 4. Здесь мы приведем все необходимые инструкции, однако для более детального изучения вопроса вам следует обратиться к упомянутому разделу, в котором, в частности, рассказывается, что создание новой переменной начинается с диалогового окна *Вычислить переменную*.

ШАГ 5Б

На этом шаге мы создадим новую переменную *трев2* и включим ее в последующий регрессионный анализ в качестве независимой. Перед выполнением следующих инструкций у вас должно быть открыто окно редактора данных.

1. В меню *Преобразовать* окна редактора данных SPSS выберите команду *Вычислить переменную*. На экране появится одноименное диалоговое окно.
2. В поле *Вычисляемая переменная* введите имя *трев2*.

3. Щелкните сначала на переменной трев, чтобы выделить ее, а затем — на кнопке со стрелкой, чтобы ввести переменную в условие отбора, составляемое в поле Числовое выражение.
4. Следом за именем переменной в поле Числовое выражение нажмите на кнопку с изображением двух звездочек (**) или введите две звездочки с клавиатуры, затем введите цифру 2 и щелкните на кнопке ОК, чтобы вернуться в окно редактора данных.
5. В меню Анализ выберите команду Регрессия ► Линейная регрессия, чтобы открыть диалоговое окно Линейная регрессия, показанное на рис. 17.4. Если вы уже работали с этим окном, щелкните на кнопке Сброс.
6. Щелкните сначала на переменной тест, чтобы выделить ее, а затем — на верхней кнопке со стрелкой, чтобы переместить переменную в поле Зависимая переменная.
7. Щелкните сначала на переменной трев, чтобы выделить ее, а затем — на второй сверху кнопке со стрелкой, чтобы переместить переменную в список Независимые переменные.
8. Повторите предыдущее действие для переменной трев2.
9. Щелкните на кнопке ОК, чтобы открыть окно вывода.

После выполнения шага 5б программа автоматически перейдет в окно вывода. Для просмотра результатов при необходимости можно воспользоваться вертикальной и горизонтальной полосами прокрутки. Обратите внимание на стандартную строку меню в верхней части окна вывода: ее присутствие позволяет выполнять любые статистические операции, не переключаясь обратно в окно редактора данных.

Представление результатов

Простой регрессионный анализ

На рис. 17.6 приведены фрагменты выводимых данных, сгенерированные программой после выполнения шага 5.

Как можно видеть, между переменными тест и трев имеется значимая линейная связь. Это означает, что с увеличением нервной возбудимости (тревожности) студента количество выполненных им тестовых заданий на зачете также имеет тенденцию к увеличению.

Регрессионный анализ криволинейной зависимости

После выполнения шага 5а в результаты вывода войдет: сводка модели с оценками параметров (см. рис. 17.2) и диаграмма зависимостей (см. рис. 17.3).

На рис. 17.7 приведены фрагменты выводимых данных, сгенерированные программой после выполнения шага 5б.

Сводка для модели

Модель	R	R квадрат	Скорректи- рованный R квадрат	Стд. ошибка оценки
1	,546 ^a	,298	,278	2,306

а. Предикторы: (константа) трев

Дисперсионный анализ^b

Модель		Сумма квадратов	ст.св.	Средний квадрат	F	Знч.
1	Регрессия	76,901	1	76,901	14,466	,001 ^a
	Остаток	180,738	34	5,316		
	Итого	257,639	35			

а. Предикторы: (константа) трев

b. Зависимая переменная: тест

Коэффициенты^a

Модель		Нестандартизованные коэффициенты		Стандартизованные коэффициенты	t	Знч.
		B	Стд. ошибка	Бета		
1	(Константа)	9,311	1,118		8,327	,000
	трев	,675	,177		3,803	,001

а. Зависимая переменная: тест

Рис. 17.6. Фрагмент окна вывода после выполнения шага 5

Сводка для модели

Модель	R	R квадрат	Скорректированный R квадрат	Стд. ошибка оценки
1	,822 ^a	,675	,655	1,593

а. Предикторы: (константа) трев2, трев

Дисперсионный анализ^b

Модель		Сумма квадратов	ст.св.	Средний квадрат	F	Знч.
1	Регрессия	173,929	2	86,964	34,283	,000 ^a
	Остаток	83,710	33	2,537		
	Итого	257,639	35			

а. Предикторы: (константа) трев2, трев

b. Зависимая переменная: тест

Коэффициенты^a

Модель		Нестандартизованные коэффициенты		Стандартизованные коэффициенты	t	Знч.
		B	Стд. ошибка	Бета		
1	(Константа)	,161	1,669		,097	,924
	трев	4,490	,629		3,633	,000
	трев2	-,338	,055		-3,148	,000

а. Зависимая переменная: тест

Рис. 17.7. Фрагмент окна вывода после выполнения шага 5б

Термины, используемые в приведенных результатах, имеют тот же смысл, что и в предыдущем случае. Обратите внимание на значительный коэффициент корреляции зависимой переменной тест с независимыми переменными трев и трев2. Величина $R^2 = 0,675$ означает, что 67,5 % дисперсии переменной тест обусловлено влиянием со стороны переменных трев и трев2. Значения F -критерия и соответствующие значимости (для F и t) говорят о сильном воздействии на зависимую переменную как обеих независимых переменных, так и каждой переменной в отдельности. Вопросы влияния нескольких независимых переменных будут рассмотрены в следующей главе. Как вы можете видеть, значения β (Бета) для нелинейных отношений не ограничиваются диапазоном от -1 до 1 . В столбце В последней таблицы приведены коэффициенты уравнения регрессии. Сравните эти коэффициенты с коэффициентами уравнения для квадратичной регрессии, приведенного в середине главы.

Терминология, используемая при выводе

- ▶ R — поскольку в анализе участвовала единственная независимая переменная, эта величина равна коэффициенту корреляции (r) между переменными тест и трев.
- ▶ R квадрат — квадрат величины R (R^2), равный доле дисперсии переменной тест, обусловленной воздействием переменной трев.
- ▶ Скорректированный R квадрат — скорректированная величина R^2 . Величина R^2 , используемая в расчетах, на практике оказывается несколько завышенной. Скорректированная величина R^2 менее формальна и ближе к реальным результатам.
- ▶ Стд. ошибка оценки — в таблице Сводка для модели это стандартное отклонение оценок значений зависимой переменной тест.
- ▶ Регрессия — статистики, оценивающие долю дисперсии зависимой переменной, обусловленную влиянием независимых переменных.
- ▶ Остаток — статистики, оценивающие долю дисперсии зависимой переменной, не обусловленную влиянием независимых переменных.
- ▶ ст. св. — число степеней свободы, для регрессии оно равно числу независимых переменных. Для остатка — равно разности размера выборки и числа степеней свободы регрессии, уменьшенной на единицу ($36 - 1 - 1 = 34$).
- ▶ Сумма квадратов, Средний квадрат, F — показатели, относящиеся к дисперсионному анализу (глава 13).
- ▶ Знч. — величина p -уровня значимости, вероятность случайности полученного результата.

- ▶ B — коэффициент и константа линейного уравнения регрессии:
 $\text{тестпрогноз} = 9,311 + 0,675(\text{трев})$
- ▶ Стд. ошибка — стандартная ошибка, в таблице Коэффициенты это характеристика стабильности коэффициента B , равная стандартному отклонению коэффициентов B , рассчитанных для большого числа выборок из генеральной совокупности.
- ▶ Бета — стандартизованный коэффициент регрессии (β). Представляет собой коэффициент B для стандартизованных значений переменной трев. Для линейных отношений эта величина всегда лежит в диапазоне от $-1,0$ до $1,0$, а для криволинейных отношений может выходить за границы этого диапазона.
- ▶ t — отношение коэффициента B к его стандартной ошибке.

Завершение анализа и выход из программы

Отредактируйте содержимое окна вывода в соответствии со своими предпочтениями: скройте лишнюю информацию, исправьте таблицы и пр. (см. раздел «Окно вывода и его редактирование» главы 2). За инструкциями по редактированию графиков обратитесь к главе 5.

Для дальнейшего использования окончательного результата все содержимое окна вывода или его фрагменты можно сохранить в файле *.spv, экспортировать в другой формат (например, Word), перенести в документ Word или вывести на печать (подробности см. в разделе «Сохранение, экспорт, перенос и печать результатов» главы 2).

Для выхода из программы выберите команду Выход в меню Файл.

18 Множественный регрессионный анализ

251	Уравнение множественной регрессии
252	Коэффициенты регрессии
253	Коэффициент детерминации и пошаговые методы
254	Условия получения приемлемых результатов анализа
255	Пошаговые алгоритмы вычислений
262	Представление результатов
265	Завершение анализа и выход из программы

Множественная регрессия является расширением простой линейной регрессии, описанной в главе 17. С помощью простой регрессии оценивалась степень влияния одной независимой переменной (предиктора) на зависимую переменную (критерий). В отличие от простой регрессии, множественная регрессия исследует влияние двух и более предикторов на критерий.

Анализ регрессии можно свести к геометрической интерпретации. Когда вычислена простая корреляция между двумя переменными, можно построить *линию регрессии* (линию «наилучшего соответствия»). Эта линия строится на основе уравнения регрессии; ее угол определяется коэффициентом при независимой переменной, а сдвиг по вертикальной оси — константой. Далее мы продемонстрируем множественную регрессию как последовательное усложнение простого регрессионного уравнения. Специально для этой главы нами создан файл данных `help.sav`. Этот файл содержит данные психологического исследования склонности людей оказывать помощь своим знакомым. Хотя данные являются вымышленными, результаты их обработки близки к результатам одного из реальных исследований.

Уравнение множественной регрессии

Итак, в этой главе мы используем новый набор данных `help.sav`. Переменная `помощь` представляет время (в секундах), потраченное человеком на оказание помощи своему партнеру, и ее значения имеют нормальное распределение (среднее равно 30, стандартное отклонение — 10). Переменная `симпатия` отражает оценку симпатии к партнеру в баллах от 1 до 20. На примере этих двух переменных мы продемонстрируем простую регрессию. В качестве зависимой выступит перемен-

ная помощь, а в качестве независимой — переменная симпатия (предполагается, что симпатия и сочувствие заставляют человека оказывать помощь, а не наоборот). Как показал анализ, коэффициент корреляции между переменными помощь и симпатия составляет 0,416 при значимости $p = 0,004$, что говорит о значительной связи между этими переменными. Константа и коэффициент регрессии составили соответственно 14,739 и 1,547. Таким образом, уравнение регрессии имеет следующий вид:

$$\text{помощь}_{\text{прогноз}} = 14,739 + 1,547 \times (\text{симпатия}).$$

Если для некоторого испытуемого значение переменной симпатия составит 16, то на основе регрессионного уравнения мы можем прогнозировать, что переменная помощь примет следующее значение:

$$14,739 + 1,547 \times 16 = 39,5.$$

Значение 16 выше средней симпатии, в результате прогнозируемое значение помощи превышает среднее ее значение почти на одно стандартное отклонение.

Аналогичные расчеты можно выполнять и при множественном регрессионном анализе. Различие заключается лишь в том, что при множественном анализе уравнение регрессии включает более чем одну зависимую переменную.

Помимо переменной симпатия с переменной помощь коррелируют и другие переменные файла help.sav. В частности, это переменные агрессия (агрессивность человека по отношению к партнеру, измеренная в баллах от 1 до 20) и польза (самооценка собственной полезности в баллах от 1 до 20). Множественный регрессионный анализ показал следующие коэффициенты при каждой из переменных: $B(\text{симпатия}) = 1,0328$, $B(\text{агрессия}) = 1,1676$, $B(\text{польза}) = 1,2569$, константа = $-5,3147$. Уравнение регрессии для множественного анализа имеет следующий вид:

$$\text{помощь}_{\text{прогноз}} = -5,3147 + 1,0328 \times (\text{симпатия}) + 1,1676 \times (\text{агрессия}) + 1,2569 \times (\text{польза}).$$

Возьмем объект с номером 7 и рассчитаем для него прогнозируемое значение переменной помощь:

$$\text{помощь}_{\text{прогноз}} = -5,3147 + 1,0328 \times 2 + 1,1676 \times 10 + 1,2569 \times 9 = 19,74.$$

Таким образом, человек, имеющий низкий показатель симпатии и средние показатели агрессивности и самооценки полезности, должен, согласно прогнозу, оказывать незначительную помощь. Фактическое значение переменной помощь для объекта 7 составило 21, что свидетельствует о высокой точности нашего прогноза.

Коэффициенты регрессии

Положительный коэффициент при независимой переменной говорит о том, что с ее возрастанием значение зависимой переменной также возрастает. Верно и противоположное утверждение: при отрицательном коэффициенте с возрастанием независимой переменной значение зависимой переменной убывает. Тем не менее

в большинстве исследований соотношение коэффициентов не позволяет делать вывод о воздействии того или иного фактора на зависимую переменную, поскольку независимые переменные, как правило, измеряются в разных шкалах и имеют разный масштаб. Чтобы разрешить эту проблему, был введен коэффициент регрессии β , принимающий значения от -1 до 1 . Он отражает *частную корреляцию* независимой и зависимой переменных. Под частной корреляцией понимается воздействие, оказываемое на зависимую переменную со стороны одной из независимых переменных при фиксированных значениях других независимых переменных (с учетом влияния последних). Чем больше данная независимая переменная коррелирует с другими независимыми переменными, тем меньше абсолютная величина ее коэффициента β . Так, в обсуждаемом примере коэффициент β для переменной агрессия — это ее корреляция с зависимой переменной *после* исключения влияния переменных симпатия и польза. Для простой регрессии с одной независимой переменной коэффициент β равен величине парной корреляции зависимой и независимой переменных. Таким образом, коэффициент β является универсальной мерой влияния независимой переменной; его часто называют *стандартным коэффициентом регрессии*. И именно стандартные коэффициенты регрессии позволяют сравнивать независимые переменные по их влиянию на оценку зависимой переменной.

Обратимся к записанному выше уравнению регрессии. Как мы можем видеть, все три независимые переменные входят в него с положительными коэффициентами. Такой результат вполне логичен для переменных симпатия и польза, однако вызывает недоумение то, что рост агрессивности субъекта влечет за собой увеличение его стремления оказывать помощь. Причина появления такого странного результата — хороший повод для дискуссии. Каждый раз, когда вы будете попадать в подобные «непредвиденные» ситуации, проверяйте правильность кодирования и ввода данных. Помните, что программа способна лишь генерировать результаты анализа, но в отношении их практической интерпретации она беспомощна.

Коэффициент детерминации и пошаговые методы

Коэффициент R является мерой связи всей совокупности независимых переменных и зависимой переменной. Часто его называют *коэффициентом множественной корреляции*. Величина R^2 равна доле дисперсии зависимой переменной, обусловленной влиянием со стороны независимых переменных, и называется *коэффициентом детерминации*. Для регрессионного анализа с тремя независимыми переменными, речь о котором шла выше, значение R составило $0,571$, а $R^2 = 0,326$. Это означает, что $32,6\%$ дисперсии переменной помощь определяется совокупным воздействием переменных агрессия, симпатия и польза. Дополнительная информация о коэффициентах множественной корреляции и детерминации приведена в разделе «Представление результатов».

Множественный регрессионный анализ позволяет использовать любое количество предикторов, но присутствие большого числа независимых переменных не всегда удобно. Было бы предпочтительно иметь в качестве предикторов как можно больше переменных, оказывающих значимое влияние на критерий, и как можно меньше переменных, не оказывающих такого влияния. В процедуру множественной регрессии SPSS включены методы, позволяющие производить пошаговый отбор в регрессионное уравнение только значимых независимых переменных. Одним из них является метод Включение, суть которого заключается в следующем. Сначала процедура вычисляет, какая из независимых переменных имеет наибольший коэффициент корреляции с зависимой переменной, а затем составляет уравнение регрессии с участием этой переменной. Далее из числа оставшихся предикторов выбирается тот, который имеет наибольший коэффициент β , при условии, что β является значимым. Выбранный предиктор также включается в уравнение регрессии. Процесс продолжается до тех пор, пока не будут выбраны все предикторы, оказывающие значимое воздействие на зависимую переменную (имеющие статистически достоверные коэффициенты β). По умолчанию SPSS продолжает выбирать независимые переменные до тех пор, пока уровень значимости (p) коэффициентов β не превысит значения 0,05. Разумеется, при желании вы можете изменить величину порогового уровня значимости.

Условия получения приемлемых результатов анализа

Регрессия, как и корреляция, анализирует линейные зависимости. В предыдущей главе мы рассмотрели процедуру оценки криволинейных зависимостей в контексте простого регрессионного анализа. Большая часть того, что было справедливо для простой регрессии, оказывается применимым и для множественного анализа. Если теория или статистический расчет показывает, что между критерием и одним или несколькими предикторами существует криволинейная зависимость, то можно ввести в качестве дополнительного предиктора, к примеру, квадрат какой-либо из независимых переменных. О том, что для этого нужно сделать, подробно рассказано в предыдущей главе.

В заключение обзора множественной регрессии рассмотрим основные условия, выполнение которых способствует получению действительно ценных и концептуально осмысленных результатов анализа.

- ▶ Ваше исследование должно быть продумано по форме и исполнению. Анализ регрессии для не связанных по смыслу величин приводит к бесполезным результатам. К сожалению, описание тех условий, которые делают исследование «продуманным по форме и исполнению», выходит за рамки темы данной книги.
- ▶ Для того чтобы существующие корреляции были признаны значимыми, необходимо иметь достаточные размеры выборок. Трудно указать точные границы «достаточно большой выборки», однако, как правило, проблемы со значимо-

стью начинают возникать при $N < 50$. Разумеется, чем большее число переменных вы привлекаете для анализа, тем больший размер выборки требуется для получения значимых результатов.

- ▶ Ваши данные должны быть корректными и не содержать ошибок.
- ▶ Распределение значений предикторов должно быть близким к нормальному. Желательно, чтобы значения асимметрий и эксцессов по модулю не превосходили 1. Тем не менее можно получить весьма точные результаты, если это требование не выполняется строго для каждого из предикторов, и даже в случае, если в анализ входит дискретная переменная с небольшим числом значений. Нормальность распределения зависимой переменной также желательна, однако допустимы как отклонения от нормальности, так и использование дискретных переменных с малым числом значений.
- ▶ Наиболее жестким требованием является запрет на использование зависимых переменных, корреляции между которыми близки к 1 (–1). Для проверки этого требования можно использовать статистики коллинеарности.

Множественный регрессионный анализ является весьма сложной процедурой. И нескольких страниц краткого обзора явно недостаточно для его изучения и исчерпывающего понимания. Поэтому желательно, чтобы перед проведением множественного регрессионного анализа у вас уже были более глубокие знания о нем.

После беглого обзора простого и множественного регрессионного анализа мы можем приступить к его практической реализации с помощью программы SPSS, а также к интерпретации результатов.

Пошаговые алгоритмы вычислений

Файл данных, который мы будем использовать в приводимых далее примерах, называется `help.sav`. Число объектов в этом файле равно 46, то есть $N = 81$. Ниже перечислены те переменные файла, которые мы будем использовать.

- ▶ `помощь` — зависимая переменная, интерпретируемая как время (в секундах) оказания помощи партнеру (среднее — 30, стандартное отклонение — 10);
- ▶ `симпатия` — оценка своей симпатии к партнеру, нуждающемуся в помощи (по 20-балльной шкале);
- ▶ `агрессия` — оценка своей агрессивности к партнеру (по 20-балльной шкале);
- ▶ `польза` — самооценка пользы от оказываемой помощи (по 20-балльной шкале);
- ▶ `проблема` — оценка серьезности проблемы своего партнера (по 20-балльной шкале);
- ▶ `эмпатия` — оценка эмпатии (склонности к сопереживанию) как результат тестирования (по 10-балльной шкале).

Для проведения множественного регрессионного анализа сначала необходимо выполнить три подготовительных шага. Эти шаги (шаги 1–3) позволят подготовить рабочий файл данных, запустить программу IBM SPSS Statistics 19 и открыть файл (в данном случае — файл ex020.sav). Пошаговые инструкции этого процесса приведены в главе 4 (с. 60), а подробные разъяснения — в главе 2.

После завершения шага 3 на экране должно присутствовать окно редактора данных со строкой меню и загруженным файлом help.sav.

ШАГ 4

В меню Анализ выберите команду Регрессия ► Линейная. На экране появится диалоговое окно Линейная регрессия, показанное на рис. 18.1.

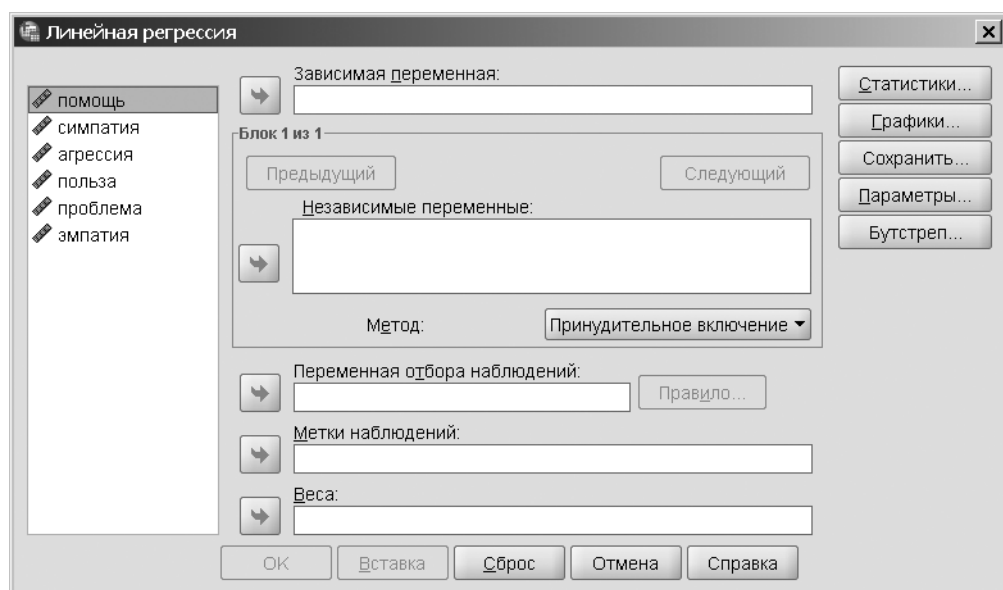


Рис. 18.1. Диалоговое окно Линейная регрессия

В верхней части диалогового окна расположено поле Зависимая переменная, предназначенное для указания единственной переменной-критерия. Ниже следует список Независимые переменные, заполняемый одной или несколькими независимыми переменными, число которых теоретически не ограничено. С помощью кнопки Следующий, расположенной справа от метки Блок 1 из 1, вы можете задать не один, а несколько наборов предикторов и тем самым провести одновременно несколько вариантов регрессионного анализа. Как только вы определите все независимые переменные и установите необходимые параметры анализа, щелчок на кнопке Следующий приведет к очистке списка Независимые переменные, который будет готов к принятию нового набора предикторов. После того как вы зададите все блоки независимых переменных и щелкните на кнопке OK, SPSS выполнит все варианты регрессионного анализа и сгенерирует результаты для каждого из них. Обратите внимание, что во всех вариантах анализа, проводимых одновременно, должна быть

общая зависимая переменная, иначе процедуру регрессии каждый раз пришлось бы запускать заново.

Важным элементом диалогового окна Линейная регрессия является раскрывающийся список Метод. Пункты этого списка определяют алгоритмы включения независимых переменных в уравнение регрессии.

- ▶ Принудительное включение — метод, применяющийся по умолчанию. Все независимые переменные включаются в уравнение независимо от степени их корреляции с переменной-критерием.
- ▶ Включение — пошаговое включение переменных с проверкой на значимость их частной корреляции с критерием. В результате в уравнение включаются все переменные, имеющие значимую частную корреляцию с переменной-критерием. Включение производится в порядке возрастания p -уровня.
- ▶ Исключение — пошаговый метод, сначала включающий в уравнение регрессии все независимые переменные, а затем поочередно удаляющий все переменные, чья корреляция с критерием имеет уровень значимости выше заданного порогового значения. Как правило, пороговым значением является $p = 0,1$.
- ▶ Шаговый отбор — комбинация пошаговых методов включения и исключения. Основной идеей является изменение доли влияния независимой переменной на критерий при появлении в уравнении других независимых переменных. Если влияние какой-либо из включенных переменных становится слишком слабым, она исключается из уравнения. Подобный метод используется при регрессионном анализе наиболее часто.
- ▶ Блочное исключение — это метод принудительного удаления переменных. Он требует предварительного задания метода Включение в качестве предыдущего блока, например Блок 1 из 1. При задании следующего блока, в данном случае Блок 2 из 2, в список Независимые переменные вы сможете ввести те независимые переменные, которые хотите исключить из уравнения регрессии. При выполнении команды вы получите результат со всеми заданными переменными, а затем — результат с удаленными переменными. Если в анализе участвуют несколько блоков, то можно задавать операцию удаления после каждого из них.

Окно и кнопка Переменная отбора наблюдений дадут возможность выбрать группирующую переменную для задания подгруппы наблюдений, в отношении которой будет проводиться анализ.

Кнопка Графики используется для графического отображения остатков. Кнопка Статистики открывает диалоговое окно Линейная регрессия: Статистики, показанное на рис. 18.2. По умолчанию в окне установлены два флажка. Флажок Оценки включает в вывод коэффициенты B , стандартные коэффициенты регрессии β , а также соответствующие стандартные ошибки, t -критерии и уровни значимости. Флажок Согласие модели генерирует значения множественного коэффициента корреляции R , величину R^2 , таблицу дисперсионного анализа, соответствующие F -величины и уровни их значимости. Таким образом, указанные флажки отвечают за генерацию основных элементов регрессионного анализа.

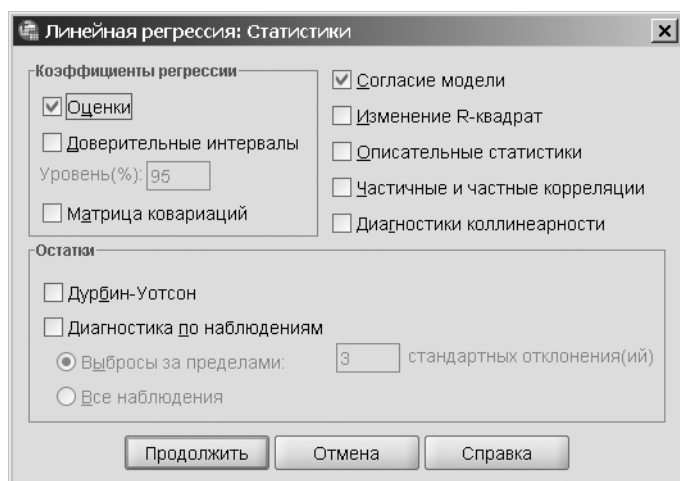


Рис. 18.2. Диалоговое окно Линейная регрессия: Статистики

Ниже пояснены наиболее важные флажки диалогового окна Линейная регрессия: Статистики.

- ▶ **Доверительные интервалы** — включает в вывод для коэффициентов B доверительный интервал в 95 %.
- ▶ **Матрица ковариаций** — генерирует таблицу, под главной диагональю которой расположены ковариации, на главной диагонали — дисперсии, а над главной диагональю — корреляции.
- ▶ **Изменение R-квадрата** — для методов Включение и Шаговый отбор указывает изменения коэффициента R^2 при введении новых переменных в уравнение регрессии.
- ▶ **Описательные статистики** — включает средние значения переменных, стандартные отклонения, а также корреляционную матрицу.
- ▶ **Диагностика коллинеарности** — устанавливает наличие коллинеарностей (корреляций, близких к 1) между переменными.

Щелчок на кнопке Сохранить в диалоговом окне Линейная регрессия приводит к открытию диалогового окна Линейная регрессия: Сохранение, показанного на рис. 18.3. В нем имеется множество флажков, названия которых, как правило, представляют загадку даже для людей, искушенных в математике. Однако некоторые весьма полезны.

Данное окно позволяет создать в файле данных новые переменные, содержащие значения, соответствующие установленным флажкам.

- ▶ В группе Предсказанные значения имеются 4 флажка. Флажок Нестандартизованные генерирует прогнозируемые значения, которые бывает полезно сравнить с фактическими значениями для оценки адекватности уравнения регрессии.

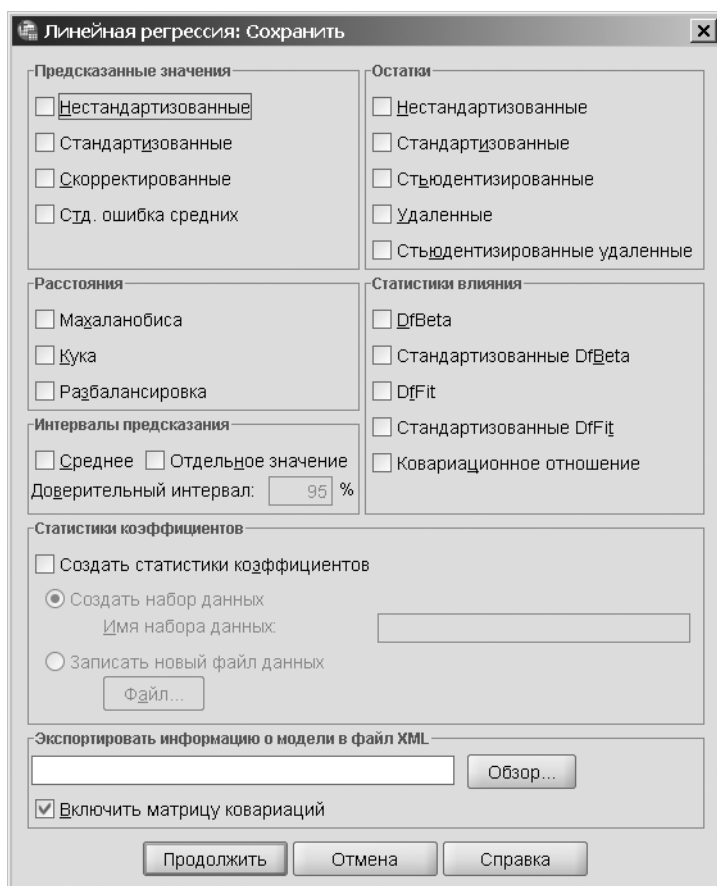


Рис. 18.3. Диалоговое окно Линейная регрессия: Сохранение

Кроме того, этот флажок позволяет получать прогнозы (оценки) зависимой переменной для тех объектов, для которых ее истинные значения неизвестны. Флажок Стандартизованные позволяет рассчитывать стандартизированные прогнозируемые значения (в z -значениях).

- В группу Остатки включены 5 флажков, позволяющих задавать сохраняемые значения остатков
- Флажки в группе Статистики влияния позволяют исключать из выборки те или иные объекты. Так, если в команде спортсменов-бегунов один пробегает дистанцию гораздо хуже или гораздо лучше других, его результаты значительно искажают статистические показатели всей команды. Иногда подобные значения («выбросы») желательно исключать из анализа. К сожалению, подробное изложение этой процедуры выходит за пределы темы данной книги.

- ▶ С помощью поля и флажков из группы Интервалы предсказания можно изменять число процентов в доверительном интервале для средних или отдельных значений (по умолчанию — 95 %).
- ▶ Флажки в группе Расстояния предоставляют три способа измерения расстояния между объектами (см. главу 22).

Наконец, рассмотрим диалоговое окно Линейная регрессия: Параметры, показанное на рис. 18.4. Это окно открывается при щелчке на кнопке Параметры в окне Линейная регрессия.

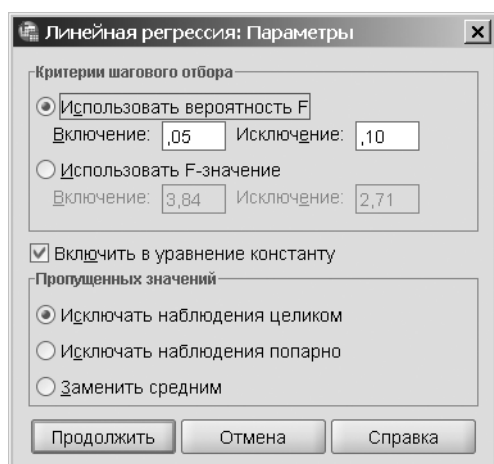


Рис. 18.4. Диалоговое окно Линейная регрессия: Параметры

Флажок Включить в уравнение константу установлен по умолчанию и без веских на то причин сбрасывать его не рекомендуется. Переключатели в группе Критерии шагового отбора позволяют управлять выполнением методов Включение, Исключение и Шаговый отбор. В поле Включение вы можете указать пороговую величину значимости для включения переменных в уравнение регрессии, а в поле Исключение — для исключения переменных; в первом по умолчанию уставлено значение 0,05, во втором — значение 0,1. Группа переключателей Пропущенных значений позволяет выбрать способ обработки отсутствующих значений (см. раздел «Обработка пропущенных значений» в главе 4).

Далее приведены два примера (шаги 5 и 5а), в первом из которых проводится множественный регрессионный анализ с участием зависимой переменной помощь и пяти предикторов симпатия, проблема, эмпатия, польза и агрессия. Для составления уравнения регрессии мы воспользуемся установленным по умолчанию методом Принудительное включение. Во втором примере мы используем некоторые из описанных выше параметров и пошаговый метод Шаговый отбор.

ШАГ 5

После выполнения шага 4 у вас должно быть открыто диалоговое окно *Линейная регрессия*, показанное на рис. 18.1.

1. Щелкните сначала на переменной *помощь*, чтобы выделить ее, а затем — на верхней кнопке со стрелкой, чтобы переместить переменную в поле *Зависимая переменная*.
2. Щелкните сначала на переменной *симпатия*, чтобы выделить ее, а затем — на второй сверху кнопке со стрелкой, чтобы переместить переменную в список *Независимые переменные*.
3. Повторите предыдущее действие для переменных *проблема*, *эмпатия*, *польза* и *агрессия*.
4. Щелкните на кнопке *ОК*, чтобы открыть окно вывода.

В результате программа сгенерирует данные, показывающие, какая из независимых переменных оказывает наибольшее влияние на зависимую переменную.

В следующем примере мы проведем регрессионный анализ с участием тех же переменных, что и в предыдущем, однако будем использовать метод *Шаговый отбор*, включим в результат статистики для коэффициентов B , описательные статистики и характеристики модели.

ШАГ 5А

После выполнения шага 4 должно быть открыто диалоговое окно *Линейная регрессия*, показанное на рис. 18.1. Если вы уже успели поработать с этим окном, очистите его щелчком на кнопке *Сброс* и выполните следующие действия.

1. Щелкните сначала на переменной *помощь*, чтобы выделить ее, а затем — на верхней кнопке со стрелкой, чтобы переместить переменную в поле *Зависимая переменная*.
2. Щелкните сначала на переменной *симпатия*, чтобы выделить ее, а затем — на второй сверху кнопке со стрелкой, чтобы переместить переменную в список *Независимые переменные*.
3. Повторите предыдущее действие для переменных *проблема*, *эмпатия*, *польза* и *агрессия*.
4. В раскрывающемся списке *Метод* выберите пункт *Шаговый отбор*.
5. Щелкните на кнопке *Статистики*, чтобы открыть диалоговое окно *Линейная регрессия: Статистики*, показанное на рис. 18.2.
6. Установите флажок *Описательные статистики* и щелкните на кнопке *Продолжить*, чтобы вернуться в диалоговое окно *Линейная регрессия*.
7. Щелкните на кнопке *Сохранить*, чтобы открыть диалоговое окно *Linear Линейная регрессия: Сохранить*, показанное на рис. 18.3.
8. Установите флажок *Нестандартизованные* и щелкните на кнопке *Продолжить*, чтобы вернуться в диалоговое окно *Линейная регрессия*.
9. Щелкните на кнопке *ОК*, чтобы открыть окно вывода.

В результате выполнения приведенных выше инструкций будут сгенерированы данные, позволяющие судить о том, какая из независимых переменных оказывает наибольшее влияние на критерий. При составлении уравнения регрессии сначала в него включаются переменные, чья частная корреляция (β) с зависимой переменной имеет уровень значимости не выше 0,05. Если затем обнаружится, что из включенных переменных какие-либо обнаруживают новый уровень значимости, превышающий значение 0,1, они исключаются из уравнения. Кроме того, в результате выполнения процедуры будет создана переменная для хранения прогнозируемых значений переменной *помощь*, рассчитанных по составленному уравнению регрессии. В окне вывода вы также сможете найти корреляционную матрицу для всех переменных и описательные статистики.

После выполнения шага 5 и шага 5а программа автоматически активизирует окно вывода. Для просмотра результатов при необходимости можно воспользоваться вертикальной и горизонтальной полосами прокрутки. Обратите внимание на стандартную строку меню в верхней части окна вывода: ее присутствие позволяет выполнять любые статистические операции, не переключаясь обратно в окно редактора данных.

Представление результатов

Множественный регрессионный анализ

Представленные на рис. 18.5 данные генерируются программой при выполнении шага 5. В таблицах представлены не все, а только окончательные результаты. Визуальная форма некоторых таблиц для удобства восприятия немного изменена без ущерба для их содержания.

В уравнение регрессии включены все пять предикторов. Коэффициент множественной корреляции R отражает связь зависимой переменной *помощь* с совокупностью независимых переменных и равен 0,598. Значение R^2 составляет 0,358 и показывает, что 35,8 % дисперсии переменной *помощь* обусловлено влиянием предикторов. Стандартные коэффициенты регрессии β отражают относительную степень влияния каждого из предикторов, но ни один из них не достигает статистической значимости ($p > 0,05$). Следовательно, вклад предикторов в оценку зависимой переменной не может быть проинтерпретирован, и результат имеет сомнительную ценность.

Изменение показателей при пошаговом добавлении переменных

Представленные на рис. 18.6 результаты генерируются программой при выполнении шага 5а. Как и в предыдущем случае, мы максимально адаптировали их для лучшего восприятия.

Сводка для модели

Модель	R	R квадрат	Скорректированный R квадрат	Стд. ошибка оценки
1	,598 ^a	,358	,277	8,627

a. Предикторы: (константа) эмпатия, проблема, польза, симпатия, агрессия

Дисперсионный анализ^b

Модель		Сумма квадратов	ст.св.	Средний квадрат	F	Знач.
1	Регрессия	1657,676	5	331,535	4,454	,003 ^a
	Остаток	2977,128	40	74,428		
	Итого	4634,804	45			

a. Предикторы: (константа) эмпатия, проблема, польза, симпатия, агрессия

b. Зависимая переменная: помощь

Коэффициенты^a

Модель		Нестандартизованные коэффициенты		Стандартизованные коэффициенты	t	Знач.
		B	Стд. ошибка	Бета		
1	(Константа)	-8,709	8,475		-1,028	,310
	симпатия	,844	,527	,227	1,601	,117
	агрессия	,832	,619	,192	1,344	,186
	польза	1,029	,626	,226	1,643	,108
	проблема	,572	,506	,162	1,131	,265
	эмпатия	,978	1,198	,116	,816	,419

a. Зависимая переменная: помощь

Рис. 18.5. Фрагменты окна вывода после выполнения шага 5

Полученные данные показывают, каким образом изменялись характеристики регрессионного анализа при поэтапном включении новых предикторов в уравнение регрессии.

С возрастанием числа предикторов значения R , R^2 и исправленная величина R^2 возрастают, в то время как значение стандартной ошибки убывает. Та же тенденция наблюдается и в отношении числа степеней свободы, суммы квадратов и среднего квадратов.

Как можно видеть, в результате применения пошагового метода из пяти предикторов в уравнение регрессии включены лишь три (модель 3): симпатия, агрессия и польза. Коэффициент множественной корреляции R отражает связь зависимой переменной помощь с совокупностью независимых переменных и равен 0,571. Значение R^2 составляет 0,326 и показывает, что 32,6 % дисперсии переменной помощь обусловлено влиянием предикторов. Стандартные коэффициенты регрессии β являются статистически достоверными, что позволяет интерпретировать относительную степень влияния каждого из предикторов; для переменной симпатия $\beta = 0,278$, а для переменных агрессия и польза соответственно $\beta = 0,276$ и $\beta = 0,269$. Каждая из независимых переменных вносит примерно одинаковый вклад в оценку зависимой переменной и коррелирует с ней положительно.

Сводка для модели^d

Модель	R	R квадрат	Скорректированный R квадрат	Стд. ошибка оценки
1	,416	,173	,154	9,335
2	,508	,258	,224	8,941
3	,571	,326	,278	8,622

Дисперсионный анализ^d

Модель	Сумма квадратов	ст. св.	Средний квадрат	F	Знач.
3 Регрессия	1512,813	3	504,271	6,784	,001 ^c
Остаток	3121,991	42	74,333		
Итого	4634,804	45			

c. Предикторы: (константа) симпатия, польза, агрессия

d. Зависимая переменная: помощь

Коэффициенты

Модель	Нестандартизованные коэффициенты		Стандартизованные коэффициенты	t	Знач.
	B	Стд. ошибка	Бета		
3 (Константа)	-5,315	8,075		-,658	,514
симпатия	1,033	,500	,278	2,065	,045
польза	1,257	,603	,276	2,083	,043
агрессия	1,168	,567	,269	2,059	,046

Рис. 18.6. Фрагменты окна вывода после выполнения шага 5а

Терминология, используемая при выводе

Ниже дана трактовка терминов, используемых программой в окне вывода.

- ▶ Вероятность F -включения — максимальный уровень значимости переменных, *вводимых* в уравнение регрессии, в данном случае равный $p = 0,050$.
- ▶ R — коэффициент множественной корреляции, отражающий связь совокупности предикторов симпатия, агрессия и польза с критерием помощь.
- ▶ R квадрат — коэффициент детерминации (R^2), равный доле дисперсии зависимой переменной помощь, обусловленной влиянием независимых переменных симпатия, агрессия и польза.
- ▶ Скорректированный R квадрат — исправленная величина R^2 . Величина R^2 , используемая в расчетах, на практике оказывается несколько завышенной. Исправленная величина R^2 ближе к реальным результатам.
- ▶ Стд. ошибка оценки — в таблице Сводка для модели стандартное отклонение ожидаемого значения переменной помощь. Как видно из приводимых данных, с добавлением каждой новой независимой переменной в уравнение регрессии эта величина уменьшается.

- ▶ Регрессия — статистика, отражающая влияние предикторов на зависимую переменную.
- ▶ Остаток — статистика, отражающая внешнее (не обусловленное предикторами) влияние на независимую переменную.
- ▶ B — нестандартизированные коэффициенты и константа уравнения регрессии, связывающего критерий и предикторы:
$$\text{помощь}_{\text{прогноз}} = -5,3147 + 1,0328 (\text{симпатия}) + 1,1676 (\text{агрессия}) + 1,2569 (\text{польза})$$
- ▶ Стд. ошибка — в таблице Коэффициенты является мерой стабильности коэффициентов B и равна стандартному отклонению их значений, рассчитанных для большого числа выборов.
- ▶ Бета — стандартизованный коэффициент регрессии (β), представляющий собой коэффициенты B для независимых переменных, представленных в z -шкале. Для линейных взаимодействий β по абсолютному значению не превосходит 1; для криволинейных взаимодействий это условие не является обязательным.
- ▶ t — отношение коэффициента B к своей стандартной ошибке.
- ▶ Бета включения — значения коэффициента β для переменных, не включенных в уравнение регрессии в предположении, что они в него включены.
- ▶ Частная корреляция — коэффициенты частной корреляции для переменных, входящих в уравнение регрессии. Наличие в этом уравнении нескольких коррелирующих переменных взаимно снижает их частную корреляцию.

Завершение анализа и выход из программы

Отредактируйте содержимое окна вывода в соответствии со своими предпочтениями: скройте лишнюю информацию, исправьте таблицы и пр. (см. раздел «Окно вывода и его редактирование» главы 2). За инструкциями по редактированию графиков обратитесь к главе 5.

Для дальнейшего использования окончательного результата все содержимое окна вывода или его фрагменты можно сохранить в файле *.spv, экспортировать в другой формат (например, Word), перенести в документ Word или вывести на печать (подробности см. в разделе «Сохранение, экспорт, перенос и печать результатов» главы 2).

Для выхода из программы выберите команду Выход в меню Файл.

19 Анализ надежности

266	Коэффициент альфа
267	Надежность половинного расщепления
267	Пошаговые алгоритмы вычислений
273	Представление результатов
277	Завершение анализа и выход из программы

Анализ надежности, рассматриваемый в этой главе, имеет отношение к надежности измерений в психологии и других социальных науках. Этот анализ еще известен как анализ заданий (пунктов теста) и применяется для отбора наиболее пригодных вопросов или заданий измерительной методики (вопросника, анкеты, теста). Обычно при разработке методики сначала составляют ее предварительный вариант, включающий избыточное количество заданий (пунктов), который апробируется на достаточно представительной выборке респондентов. Затем проводят анализ надежности, который позволяет при помощи многочисленных критериев исключить неподходящие задания.

Мы коснемся двух показателей надежности: *коэффициента Кронбаха* (Chronbach), обозначаемого буквой α (альфа), и *надежности половинного расщепления* (split-half reliability). Существуют и другие типы надежности, однако мы ограничимся лишь их упоминанием в разделе пошаговых процедур, а основное внимание сосредоточим на наиболее употребительном коэффициенте α . После описаний двух видов надежности мы приведем примеры, иллюстрирующие их применение.

Коэффициент альфа

Коэффициент α является мерой внутренней согласованности, или однородности, измерительной шкалы. Как правило, α лежит в пределах от 0 до 1, однако может принимать и отрицательные значения. Последние свидетельствуют о том, что часть элементов, или пунктов, шкалы измеряет противоположные величины. Чем ближе коэффициент α к 1, тем выше внутренняя согласованность системы заданий. Коэффициент α можно рассматривать как показатель корреляции между измеренными признаками для выборок равного объема. Если, к примеру, составить 10 тестов с сотней случайных вопросов из банка данных объемом в 1000 вопросов, коэффициент α будет характеризовать корреляцию между этими тестами. Формула для вычисления коэффициента α выглядит следующим образом:

$$\alpha = \frac{kr}{1 + (k-1)r}.$$

Здесь k — число пунктов шкалы, r — средний коэффициент корреляции между каждым пунктом и суммой остальных пунктов.

Из приведенной формулы следует, что коэффициент α возрастает как с увеличением числа пунктов шкалы, так и с усилением корреляции между пунктами.

Применение α Кронбаха основано на модели, предполагающей наличие большей дисперсии у более надежного теста: чем надежнее тест, тем большая чувствительность (различительная способность) пунктов теста.

Если в методике применяются задания дихотомического типа («да» — «нет», «правильно» — «неправильно»), формула для вычисления коэффициента α Кронбаха (Chronbach) становится идентична формуле Кюдера—Ричардсона (Kuder & Richardson).

Надежность половинного расщепления

Надежность половинного расщепления вычисляется в случае, когда число элементов достаточно велико, чтобы можно было разделить шкалу на две половины. Рассмотрим тест 16 личностных свойств Раймонда Кеттелла. Тест состоит из 187 вопросов, из которых от 10 до 14 вопросов определяют одно из 16 свойств. Разумеется, было бы бессмысленно делить на две половины все 187 вопросов, поскольку в каждой половине оказалось бы множество вопросов, относящихся к разным свойствам, а следовательно, несопоставимых друг с другом. Более разумным подходом было бы разделить пополам совокупность вопросов, измеряющих одно свойство: например, из 14 вопросов, определяющих тревожность, в случайном порядке выбрать 7 и сформировать первую группу, а из оставшихся — вторую группу. В качестве показателя надежности используется коэффициент корреляции Спирмена—Брауна между двумя образованными половинами.

Пошаговые алгоритмы вычислений

Для проведения анализа надежности мы будем использовать файл данных TestA.sav, содержащий 12 переменных для 120 объектов (респондентов). Переменные имеют имена a1, ..., a12 и соответствуют пунктам разрабатываемой семантической шкалы для оценки степени альтруизма. Каждая переменная — это оценка респондентом другого человека по наличию (1 — «да») или отсутствию (0 — «нет») у него одной из 12 характеристик. Список этих характеристик (прилагательных) вы можете найти в разделе «Представление результатов».

Чтобы приступить к анализу надежности, сначала нужно выполнить три подготовительных шага. Эти шаги (шаги 1–3) позволят подготовить рабочий файл данных, запустить программу IBM SPSS Statistics 19 и открыть файл (в данном случае — файл TestA.sav). Пошаговые инструкции этого процесса приведены в главе 4 (с. 60), а подробные разъяснения — в главе 2.

После завершения шага 3 на экране должно присутствовать окно редактора данных со строкой меню и загруженным файлом TestA.sav.

ШАГ 4

В меню Анализ выберите команду Шкалирование ► Анализ надежности. Откроется диалоговое окно Анализ надежности, показанное на рис. 19.1.

Параметры команды анализа надежности настраиваются при помощи двух диалоговых окон: основного, изображенного на рис. 19.1, и диалогового окна Анализ надежности: Статистики, показанного на рис. 19.2 и открывающегося при щелчке на кнопке Статистики.

Основное окно команды Анализ надежности почти не отличается от диалоговых окон большинства команд статистических процедур SPSS. Слева расположен список доступных переменных текущего файла данных, а внизу находятся 5 стандартных кнопок. Список Пункты предназначен для ввода имен переменных, включаемых в анализ надежности.

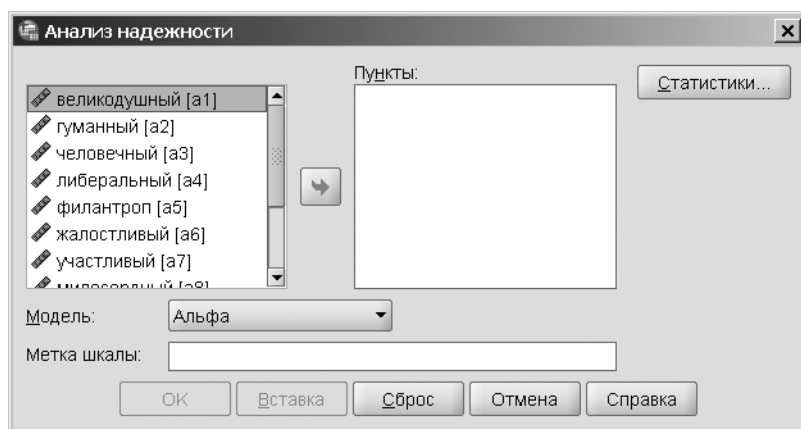


Рис. 19.1. Диалоговое окно Анализ надежности

Раскрывающийся список Модель позволяет задать модель анализа надежности, по умолчанию выбран пункт Альфа. Ниже приводятся описания остальных пунктов списка.

- Расщепления пополам — программа разбивает переменные на две группы в порядке их перечисления (число переменных в группах одинаково, если общее число переменных четно; в противном случае первая группа содержит на одну переменную больше) и сравнивает эти группы между собой.
- Гутмана — анализ основан на операции вычисления нижних границ.
- Параллельная — анализ проводится в предположении, что все переменные имеют равные дисперсии.
- Строго параллельная — анализ проводится в предположении, что все переменные имеют равные средние значения, дисперсии фактических значений и дисперсии ошибок.

Нижнее окно Метка шкалы позволяет назначить имя каждому заданному набору пунктов и варианту анализа. Если оставить это окно пустым, то по умолчанию в качестве метки шкалы будет использовано имя Шкальные: Все переменные.

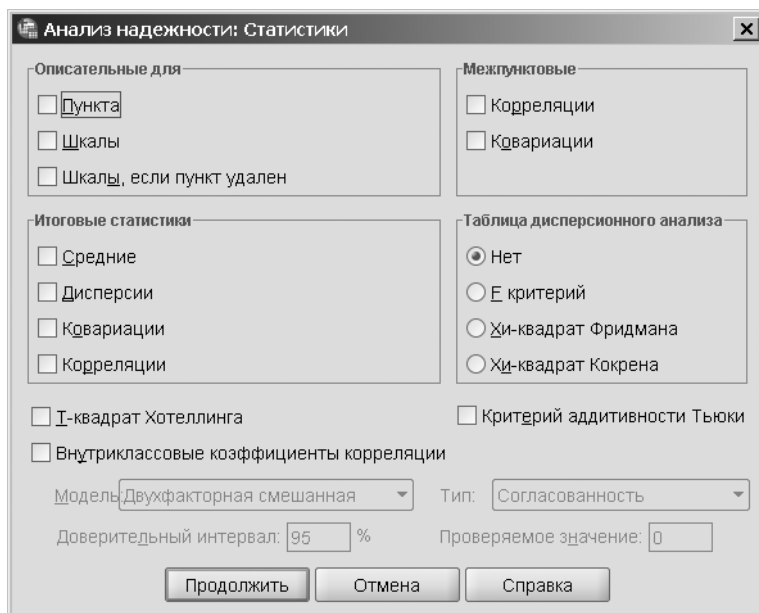


Рис. 19.2. Диалоговое окно Анализ надежности: Статистики

Как уже упоминалось, при щелчке на кнопке Статистики открывается диалоговое окно Анализ надежности: Статистики. В нем вы можете задать параметры команды Анализ надежности. Далее будут приведены описания всех флажков данного диалогового окна, за исключением тех, которые находятся в группе Таблица дисперсионного анализа. Дисперсионный анализ (ANOVA) и критерий χ^2 рассматривались нами ранее в этой книге, и, кроме того, эти виды анализа весьма редко используются в контексте анализа надежности. Остальные флажки перечислены в порядке их размещения в группах Описательные для, Итоговые статистики, Межпунктовые. Кроме того, последними даны описания флажков Т-квадрат Хотеллинга и Критерий аддитивности Тьюки.

- ▶ Пункта — вычисление средних значений и стандартных отклонений переменных, участвующих в анализе.
- ▶ Шкалы — включение в вывод среднего значения, дисперсии, стандартного отклонения и размера выборки для суммы всех переменных шкалы. Под суммой переменных понимается сумма значений всех 12 переменных (пунктов) для каждого объекта. Таким образом, обеспечивается вывод статистик для переменной, полученной в результате суммирования.
- ▶ Шкалы, если пункт удален — включение в вывод для каждой переменной значения коэффициента α , подсчитанного для всей шкалы в предположении, что

данная переменная исключена. В большинстве случаев этот флажок следует устанавливать.

- ▶ Средние — вычисление средних значений по всем наблюдениям для каждой переменной. В результат включаются все средние значения, минимальное и максимальное из этих значений, размах, отношение максимального среднего к минимальному, а также дисперсия средних значений. Несмотря на некоторую сложность, этот флажок позволяет получить важные для анализа надежности результаты.
- ▶ Дисперсии — этот флажок аналогичен флажку Средние за тем исключением, что вычисления выполняются не для средних значений, а для дисперсий.
- ▶ Ковариации в группе Итоговые статистики — вычисление ковариаций между каждой переменной и суммой всех остальных переменных. В выводимые результаты также включаются средние значения полученных ковариаций, их минимум, максимум, размах, отношение максимума к минимуму и дисперсия.
- ▶ Корреляции в группе Итоговые статистики — этот флажок отличается от флажка Ковариации лишь тем, что вычисляются не ковариации, а корреляции. Среднее корреляций для всех переменных является значением r в формуле вычисления коэффициента α .
- ▶ Корреляции в группе Межпунктовые — включение в вывод корреляционной матрицы для всех переменных, участвующих в анализе.
- ▶ Ковариации в группе Межпунктовые — создание ковариационной матрицы для всех переменных, участвующих в анализе.
- ▶ Т-квадрат Хотеллинга — применение множественного сравнения по t -критерию для проверки достоверности различий средних значений всех переменных, участвующих в анализе.
- ▶ Критерий аддитивности Тьюки — применение критерия для проверки линейности зависимости между переменными, участвующими в анализе.

Далее приведены три примера, иллюстрирующие применение критерия надежности. На шаге 5 проводится анализ надежности с параметрами, позволяющими оценить вклад каждого пункта во внутреннюю согласованность шкалы. На шаге 5а проводится анализ надежности с применением некоторых из описанных параметров. Данные, полученные в результате выполнения шага 5, показывают, что для 12 переменных надежность весьма невелика. Поэтому из списка на шаге 5а удаляются три переменные, нарушающие внутреннюю согласованность шкалы; тем самым повышается ее надежность. Наконец, на шаге 5б проводится анализ надежности половинного расщепления с 9 из 12 исходных переменных. Результаты всех трех вариантов анализа включены в раздел «Представление результатов».

В следующем примере проводится анализ надежности с включением всех 12 пунктов.

ШАГ 5

После выполнения шага 4 должно быть открыто диалоговое окно Анализ надежности, показанное на рис. 19.1. Если вы уже успели поработать с этим окном, щелкните на кнопке Сброс.

1. Перенесите группу переменных a_1, \dots, a_{12} из левого в правое поле. Для этого щелкните сначала на переменной a_1 , чтобы выделить ее. Затем нажмите клавишу Shift и, удерживая ее, щелкните на переменной a_{12} , а затем — на кнопке со стрелкой, чтобы переместить переменные в список Пункты.
2. Щелкните на кнопке Статистики, чтобы открыть диалоговое окно Анализ надежности: Статистики, показанное на рис. 19.2.
3. В группе Описательные для установите флажок Шкалы, если пункт удален. Щелкните на кнопке Продолжить, чтобы вернуться в диалоговое окно Анализ надежности.
4. Щелкните на кнопке ОК, чтобы открыть окно вывода.

При выполнении этого шага все переменные попадают в список Пункты в том порядке, в котором они были в исходном файле. В вывод вместе с именами переменных включаются метки, представляющие соответствующие вопросы теста, а также коэффициент α и объем выборки N . Кроме того, в окне вывода будут представлены результаты расчета шкалы альфа для каждого пункта в предположении, что он из шкалы удален.

Далее приведен пример анализа надежности с участием 9 переменных, для которых предыдущий анализ выявил наибольшую надежность ($a_1, a_2, a_3, a_4, a_7, a_8, a_9, a_{10}, a_{12}$). Кроме того, мы зададим некоторые дополнительные параметры и снова включим в выводимые результаты метки переменных.

ШАГ 5А

После выполнения шага 4 должно быть открыто диалоговое окно Анализ надежности, показанное на рис. 19.1. Если вы уже успели поработать с этим окном, щелкните на кнопке Сброс.

1. Щелкните сначала на переменной a_1 , чтобы выделить ее, а затем — на кнопке со стрелкой, чтобы переместить переменную в список Пункты.
2. Повторите предыдущее действие для переменных $a_2, a_3, a_4, a_7, \dots, a_{10}$ и a_{12} . Поскольку переменные a_1, \dots, a_4 в исходном списке расположены последовательно, для их одновременного выделения можно щелкнуть на переменной a_1 , нажать клавишу Shift и щелкнуть на переменной a_4 . То же касается переменных a_7, \dots, a_{10} .
3. В окне Метка шкалы введите текст, соответствующий применяемому методу и участвующим в анализе переменным: Альфа Кронбаха, пункты a_1 - a_4, a_7 - a_{10}, a_{12} .
4. Щелкните на кнопке Статистики, чтобы открыть диалоговое окно Анализ надежности: Статистики, показанное на рис. 19.2.

5. В группе Описательные для установите флажки Шкалы и Шкалы, если пункт удален, в группе Итоговые статистики — флажки Средние, Дисперсии и Корреляции, в группе Межпунктовые — флажок Корреляции. Щелкните на кнопке Продолжить, чтобы вернуться в диалоговое окно Анализ надежности.
6. Щелкните на кнопке ОК, чтобы открыть окно вывода.

В следующем примере мы выполним ту же процедуру, что и на шаге 5а, однако вместо коэффициента α будем рассчитывать надежность половинного расщепления.

ШАГ 5Б

После выполнения шага 4 должно быть открыто диалоговое окно Анализ надежности, показанное на рис. 19.1. Если вы уже успели поработать с этим окном, щелкните на кнопке Сброс.

1. Щелкните сначала на переменной a_1 , чтобы выделить ее, а затем — на кнопке со стрелкой, чтобы переместить переменную в список Пункты.
2. Повторите предыдущее действие для переменных a_2 , a_3 , a_4 , a_7 , ..., a_{10} и a_{12} . Поскольку переменные a_1 , ..., a_4 в исходном списке расположены последовательно, для их одновременного выделения можно щелкнуть на переменной a_1 , нажать клавишу Shift и щелкнуть на переменной a_4 . То же касается переменных a_7 , ..., a_{10} .
3. Раскройте список Модель и выберите пункт Расщепления пополам.
4. В поле Метка шкалы введите текст, соответствующий применяемому методу и участвующим в анализе переменным: Половинного расщепления, пункты a_1 - a_4 , a_7 - a_{10} , a_{12} .
5. Щелкните на кнопке Статистики, чтобы открыть диалоговое окно Анализ надежности: Статистики, показанное на рис. 19.2.
6. В группе Описательные для установите флажок Шкалы, в группе Итоговые статистики — флажки Средние, Дисперсии и Корреляции, в группе Межпунктовые — флажок Корреляции. Щелкните на кнопке Продолжить, чтобы вернуться в диалоговое окно Анализ надежности: Статистики.
7. Щелкните на кнопке ОК, чтобы открыть окно вывода.

После выполнения шагов 5, 5а и 5б программа автоматически переходит в окно вывода. Для просмотра результатов при необходимости можно воспользоваться вертикальной и горизонтальной полосами прокрутки. Обратите внимание на стандартную строку меню в верхней части окна вывода: ее присутствие позволяет выполнять любые статистические операции, не переключаясь обратно в окно редактора данных.

Представление результатов

В этом разделе представлены фрагменты выводимых данных, генерируемые программой при выполнении шагов 5, 5а и 5б. Каждый блок результатов озаглавлен в соответствии с введенными в поле Метка шкалы данными, например, после выполнения шага 5: Шкальные: Все переменные.

Коэффициент альфа с параметрами по умолчанию

Первая из таблиц Статистики пригодности на рис. 19.3 содержит величину Альфа Кронбаха для шкалы, содержащей все 12 переменных. Во второй таблице Статистики соотношения пункта с суммарным баллом для каждой переменной показаны статистические характеристики, позволяющие сравнить ее с совокупностью остальных переменных.

Шкальные: ВСЕ ПЕРЕМЕННЫЕ

Статистики пригодности	
Альфа Кронбаха	Количество пунктов
,696	12

Статистики соотношения пункта с суммарным баллом				
	Среднее шкалы при удалении пункта	Дисперсия шкалы при удалении пункта	Корреляция пункта с суммарным баллом	Альфа Кронбаха при удалении пункта
великодушный	5,68	6,389	,643	,629
гуманный	5,73	6,588	,546	,644
человечный	5,68	6,893	,425	,663
либеральный	5,73	7,125	,324	,678
филантроп	5,82	8,470	-,164	,745
жалостливый	5,81	7,971	,008	,723
участливый	5,76	6,924	,404	,666
милосердный	5,64	6,736	,501	,652
заботливый	5,72	6,759	,474	,655
благосклонный	5,75	6,727	,486	,653
ласковый	5,82	8,235	-,084	,735
добродушный	5,73	6,601	,540	,645

Рис. 19.3. Фрагмент окна вывода после выполнения шага 5

Ниже дана трактовка терминов, используемых программой в окне вывода и относящихся к вычислению коэффициента α с параметрами по умолчанию.

- Среднее шкалы при удалении пункта — для каждой переменной вычисляется сумма остальных 11 переменных по всем 120 объектам файла.
- Дисперсия шкалы при удалении пункта — дисперсия суммы всех переменных, с той переменной, которая удалена.

- Общая корреляция коррелированных пунктов — коэффициент корреляции переменной с суммой остальных 11 переменных.
- Альфа Кронбаха при удалении пункта — коэффициент α , полученный при удалении соответствующей переменной из шкалы.

Из рис. 19.3 видно, что для трех переменных значения в столбце Корреляция пункта с суммарным баллом оказались весьма низкими, а для переменных a5 и a11 даже отрицательными. Для увеличения коэффициента α (указанного в таблице Статистики пригодности) необходимо из рассмотрения такие переменные, величина Альфа Кронбаха при удалении которых больше, чем величина Альфа для всей шкалы. Анализ, проведенный после исключения этих трех переменных, действительно позволил найти новое, более высокое значение коэффициента α . Таким образом, удалось выявить, что наибольшее значение α принимает в случае удаления из шкалы переменных с номерами 5, 6, и 11. Обратите внимание, что подобные переменные не исключаются программой автоматически: существуют причины, по которым иногда требуется сохранять их в шкале.

Коэффициент альфа после исключения переменных и задания дополнительных параметров

Приведенные на рис. 19.4 результаты формируются после удаления из шкалы трех упомянутых переменных (шаг 5а). Первая таблица Статистики пригодности содержит величину Альфа Кронбаха для шкалы, содержащей 9 переменных. Вторая таблица содержит основные статистические показатели для шкалы: средние, дисперсии и корреляции. Таблица Статистики соотношения пункта с суммарным баллом по своему содержанию аналогична одноименной таблице, приведенной ранее на рис. 19.3, однако здесь вместо 12 исходных переменных содержатся лишь 9. Как легко видеть, значение коэффициента α увеличилось и составляет 0,825. Данное значение α больше нельзя увеличить, удалив какую-либо из 9 оставшихся переменных, так как Альфы Кронбаха при дальнейшем исключении любого из оставшихся пунктов получаются меньше, чем α , полученная по всем 9 пунктам. Таким образом, оставшиеся 9 пунктов образуют оптимальный состав шкалы.

Ниже дана трактовка новых терминов, используемых программой в окне вывода и относящихся к вычислению коэффициента α после исключения нескольких переменных и задания дополнительных параметров.

- Средние пунктов — описательная информация о средних значениях 9 пунктов шкалы. Среднее полученных средних значений равно 0,547, наименьшее из 9 средних равно 0,5 и т. д.
- Дисперсии пунктов — статистика, аналогичная средним элементов, но полученная для их дисперсий.

- Межпунктовые корреляции — описательная статистика, характеризующая корреляцию между каждой переменной шкалы и суммой остальных переменных. Для каждой переменной вычисляется коэффициент корреляции, а в окне вывода приводятся среднее значение, минимум и т. д. полученных коэффициентов. Среднее значение коэффициентов корреляции является величиной r в формуле расчета α .

Шкальные: Альфа Кронбаха, пункты a1-a4, a7-a10, a12

Статистики пригодности

Альфа Кронбаха	Альфа Кронбаха, основанная на стандартизованных пунктах	Количество пунктов
,825	,825	9

Сводка статистик пункта

	Среднее	Минимум	Максимум	Размах	Максимум / Минимум	Дисперсия
Средние пунктов	,547	,500	,617	,117	1,233	,002
Дисперсии пунктов	,248	,238	,252	,014	1,058	,000
Межпунктовые корреляции	,344	,134	,632	,499	4,733	,013

Статистики соотношения пункта с суммарным баллом

	Среднее шкалы при удалении пункта	Дисперсия шкалы при удалении пункта	Корреляция пункта с суммарным баллом	Квадрат коэффициента множественной корреляции	Альфа Кронбаха при удалении пункта
великодушный	4,34	6,462	,668	,549	,791
гуманный	4,39	6,492	,645	,487	,793
человечный	4,34	6,950	,458	,331	,815
либеральный	4,39	7,081	,397	,283	,822
участливый	4,43	6,986	,434	,259	,818
милосердный	4,31	6,837	,515	,331	,809
заботливый	4,38	6,894	,474	,383	,814
благосклонный	4,42	6,716	,546	,362	,805
добродушный	4,40	6,545	,621	,550	,796

Рис. 19.4. Фрагмент окна вывода после выполнения шага 5а

- Квадрат коэффициента множественной корреляции — представленные в этой секции величины вычислены с помощью уравнения множественной регрессии, позволяющего получать прогнозируемый коэффициент корреляции данной переменной с остальными переменными, участвующими в анализе.
- Альфа Кронбаха — коэффициент α , вычисляемый по формуле, приведенной в начале главы. Число элементов шкалы (k) в данном случае равно 9. Поскольку α зависит от числа элементов шкалы, нельзя четко определить критерий

приемлемости полученного значения α . Для большинства случаев можно придерживаться следующих рекомендаций по оценке внутренней согласованности шкалы:

- ▶ $\alpha > 0,9$ — отличная;
 - ▶ $\alpha > 0,8$ — хорошая;
 - ▶ $\alpha > 0,7$ — приемлемая;
 - ▶ $\alpha > 0,6$ — сомнительная;
 - ▶ $\alpha > 0,5$ — малопригодная;
 - ▶ $\alpha < 0,5$ — недопустимая.
- ▶ Альфа Кронбаха, основанная на стандартизованных пунктах — коэффициент α , вычисляемый в предположении, что элементы шкалы заранее стандартизованы. Как можно видеть, значения обычного и стандартизованного коэффициентов α в данном случае не отличаются.

Надежность половинного расщепления

На рис. 19.5 приведены результаты анализа надежности половинного расщепления с использованием тех же 9 переменных, что и в предыдущем случае. Как правило, такое количество переменных слишком мало для подобного анализа; тем не менее результаты вполне пригодны для интерпретации. Большая часть полученной информации по смыслу аналогична результатам, полученным для коэффициента α . Мы приводим лишь основные результаты анализа.

Обратите внимание, что значения α , вычисленные для каждой половины шкалы, ниже, чем для шкалы в целом. Это отражает тот факт, что при уменьшении числа элементов шкалы ее внутренняя согласованность снижается.

Ниже дана трактовка нескольких новых терминов, относящихся к вычислению надежности половинного расщепления.

- ▶ Корреляция между формами — приближенное значение надежности шкалы, рассчитанное в предположении, что она содержит 5 элементов.
- ▶ Коэффициент Спирмена–Брауна: Равная длина — результат гипотетического анализа надежности половинного расщепления шкалы, содержащей 10 элементов.
- ▶ Коэффициент Спирмена–Брауна: Неравная длина — результат гипотетического анализа надежности половинного расщепления шкалы, содержащей не 10, а 9 элементов, что приводит к необходимости формирования двух неравных групп.
- ▶ Коэффициент половинного расщепления Гуттмана — значение надежности, полученное путем вычисления нижних пределов.

Шкальные: Половинное расщепление, пункты а1-а4, а7-а10, а12

Статистика пригодности			
Альфа Кронбаха	Часть 1	Значение	,714
		Количество пунктов	5 ^a
	Часть 2	Значение	,695
		Количество пунктов	4 ^b
	Общее количество пунктов		9
Корреляция между формами			,696
Козффициент	Равная длина		,821
Спирмена-Брауна	Неравная длина		,822
Козффициент половинного расщепления Гуттмана			,814

а. Пункты: великодушный, гуманный, человечный, либеральный, участливый.

б. Пункты: милосердный, заботливый, благосклонный, добродушный.

Рис. 19.5. Основные статистические показатели для шкалы при анализе надежности половинного расщепления

Завершение анализа и выход из программы

Отредактируйте содержимое окна вывода в соответствии со своими предпочтениями: скройте лишнюю информацию, исправьте таблицы и пр. (см. раздел «Окно вывода и его редактирование» главы 2). За инструкциями по редактированию графиков обратитесь к главе 5.

Для дальнейшего использования окончательного результата все содержимое окна вывода или его фрагменты можно сохранить в файле *.spv, экспортировать в другой формат (например, Word), перенести в документ Word или вывести на печать (подробности см. в разделе «Сохранение, экспорт, перенос и печать результатов» главы 2).

Для выхода из программы выберите команду Выход в меню Файл.

20 Факторный анализ

279	Вычисление корреляционной матрицы
279	Извлечение факторов
280	Выбор и вращение факторов
282	Интерпретация факторов
283	Пошаговые алгоритмы вычислений
290	Представление результатов
294	Терминология, используемая при выводе
295	Завершение анализа и выход из программы

В последние 40–50 лет факторный анализ приобрел значительную популярность в самых различных исследованиях. Во многом этому способствовала разработка Раймондом Кеттеллем знаменитого 16-факторного личностного опросника (16PF). Именно при помощи факторного анализа ему удалось свести около 4500 наименований личностных особенностей к 187 вопросам, которые, в свою очередь, позволяют измерить 16 различных свойств личности. Факторный анализ дает возможность количественно определить нечто, непосредственно неизмеряемое, исходя из нескольких доступных измерению переменных. Например, характеристики «посещает развлекательные мероприятия», «много разговаривает», «охотно идет на контакт с любым незнакомым человеком» могут служить оценками качества «общительность», которое непосредственно не поддается количественному измерению. Факторный анализ позволяет установить для большого числа исходных признаков сравнительно узкий набор «свойств», характеризующих связь между группами этих признаков и называемых факторами. Процедура факторного анализа состоит из четырех основных стадий.

1. Вычисление корреляционной матрицы для всех переменных, участвующих в анализе.
2. Извлечение факторов.
3. Вращение факторов для создания упрощенной структуры.
4. Интерпретация факторов.

Первые 3 операции мы кратко рассмотрим в следующих нескольких разделах; последняя операция описывается лишь на концептуальном уровне.

Как и многие другие методы, рассматриваемые в этой книге, факторный анализ является весьма сложной процедурой. Несколько страниц краткого обзора явно недоста-

точные для его изучения и исчерпывающего понимания. Поэтому желательно, чтобы перед проведением факторного анализа у вас уже были более глубокие знания.

Вычисление корреляционной матрицы

Первая операция, которая производится при выполнении факторного анализа, — вычисление корреляционной матрицы для переменных, участвующих в анализе. Уже из этого можно сделать вывод о том, что факторный анализ основан на взаимодействии переменных. Для проведения факторного анализа совсем не обязательно специально строить корреляционную матрицу: при необходимости программа SPSS создаст ее сама на основе данных файла. Иногда не нужны даже исходные данные — достаточно иметь корреляционную матрицу, которая в этом случае вводится в командный файл SPSS. Однако подобный подход весьма сложен и в рамках этой книги не рассматривается. При необходимости вы можете обратиться к руководству пользователя программы SPSS.

Извлечение факторов

Извлечение факторов является следующим этапом факторного анализа. С математической точки зрения извлечение факторов имеет определенную аналогию с множественным регрессионным анализом. Если вы обратитесь к главе 18, то вспомните, что первым шагом множественного регрессионного анализа является выбор той независимой переменной, которая обуславливает наибольшую долю дисперсии зависимой переменной. Затем операция повторяется для оставшихся независимых переменных до тех пор, пока добавляемая доля дисперсии не перестанет быть значимой. В факторном анализе существует аналогичная процедура, и для ее описания достаточно лишь немного изменить соответствующий раздел главы 18.

Извлечение фактора начинается с подсчета суммарного разброса значений всех участвующих в анализе переменных (данная величина чем-то похожа на общую сумму квадратов). Для этого «суммарного разброса» непросто подобрать логическую интерпретацию, однако он является вполне строго определенной математической величиной. Первой задачей факторного анализа является выбор взаимодействующих переменных, чья взаимная корреляция обуславливает наибольшую долю общей дисперсии. Эти переменные образуют *первый фактор*. Затем первый фактор исключается и из оставшегося множества переменных снова выбираются те, чье взаимодействие определяет наибольшую долю оставшейся общей дисперсии. Эти переменные образуют *второй фактор*. Процедура извлечения факторов продолжается до тех пор, пока не будет исчерпана вся общая дисперсия переменных.

По умолчанию в процедуре факторного анализа каждая переменная имеет единичное значение *общности*. Этот показатель равен доле дисперсии переменной, обусловленной совокупным влиянием факторов. Общность можно сравнить с множественным коэффициентом корреляции R , принимающим значение 0 в случае, если факторы не влияют на переменную, и значение 1 в случае, если дисперсия переменной целиком определяется выделяемыми факторами. Перед началом извлечения факторов единичное значение общности установлено по умолчанию для всех переменных, участвующих в факторном анализе.

После того как процедура извлекает первый фактор, напротив его номера появляется его *собственное значение*, например рядом с числом 1 появляется значение 5,13312. Собственное значение фактора пропорционально доле общей дисперсии, определяемой данным фактором (обратите внимание, что здесь речь идет о факторе, а не о переменной, как было в случае общности). Первое собственное значение всегда является наибольшим и превышает единицу; это определяется алгоритмом работы процедуры, согласно которому факторы извлекаются в порядке убывания их влияния на дисперсию переменных. Затем вычисляется процент дисперсии, обуславливаемый данным фактором и равный отношению собственного значения фактора к числу переменных, а также соответствующий кумулятивный (накопленный) процент. С извлечением каждого нового фактора собственные значения уменьшаются, а кумулятивный процент приближается к 100. Обратите внимания, что в контексте процедуры извлечения факторов ни разу не прозвучал термин «значимость»: в отличие от регрессионного анализа извлечение факторов происходит до тех пор, пока не будет исчерпана вся дисперсия переменных, независимо от того, является ли значимым влияние фактора на дисперсию или нет.

Выбор и вращение факторов

За очень редкими исключениями для исследователя не представляют интереса все извлеченные факторы. Если факторов окажется столько же, сколько исходных переменных, факторный анализ теряет смысл, поскольку его целью является сокращение исходного набора переменных.

Итак, нужно принять решение, какие из факторов следует оставить для дальнейшего анализа. Здесь, в первую очередь, рекомендуется руководствоваться здравым смыслом и оставлять те факторы, которые имеют понятную теоретическую или логическую интерпретацию. Однако не всегда представляется возможным заранее установить назначение каждого фактора, и поэтому исследователи на первом этапе обычно используют формальные критерии. При выполнении факторного анализа с установками по умолчанию все факторы, чьи собственные значения превышают единицу, сохраняются для дальнейшего анализа. Поскольку число факторов равно числу переменных, лишь для небольшого количества факторов собственные значения оказываются больше единицы, а значит, выполнение команды с параметрами по умолчанию позволяет радикально сократить число

факторов. Существуют и другие критерии выделения факторов (например, критерий «каменистой осыпи» Р. Кеттелла); кроме того, вы можете выбирать факторы, основываясь на известных вам особенностях конкретного файла данных. В любом случае, окончательное решение о числе факторов обычно принимается после интерпретации факторов, следовательно, факторный анализ предполагает неоднократное выделение различного числа факторов. В разделе пошаговых процедур рассмотрены несколько вариантов выполнения факторного анализа, отличные от принятого по умолчанию.

Следующим шагом после выделения факторов является их вращение. Вращение требуется потому, что изначально структура факторов, будучи математически корректной, как правило, трудна для интерпретации. Целью вращения является получение простой структуры, которой соответствует большое значение нагрузки каждой переменной только по одному фактору и малое по всем остальным факторам. *Нагрузка* отражает связь между переменной и фактором, являясь подобием коэффициента корреляции. Значение нагрузки лежит в пределах от -1 до 1 . Идеальная простая структура предполагает, что каждая переменная имеет нулевые значения нагрузок для всех факторов, кроме одного, для которого нагрузка этой переменной близка к 1 (-1). Графическая интерпретация процедуры вращения представлена на рис. 20.1. До вращения (*слева*) точки, соответствующие переменным, расположены на удалении от осей факторов. После поворота осей (*справа*) переменные оказываются вблизи осей, что соответствует максимальной нагрузке каждой переменной только по одному фактору. На практике строгая ориентация переменных вдоль осей факторов обычно не достигается, однако операция поворота позволяет приблизиться к желательному результату. После выполнения вращения факторов программа SPSS включает в окно вывода похожие графические изображения, для удобства снабдив их таблицей координат.

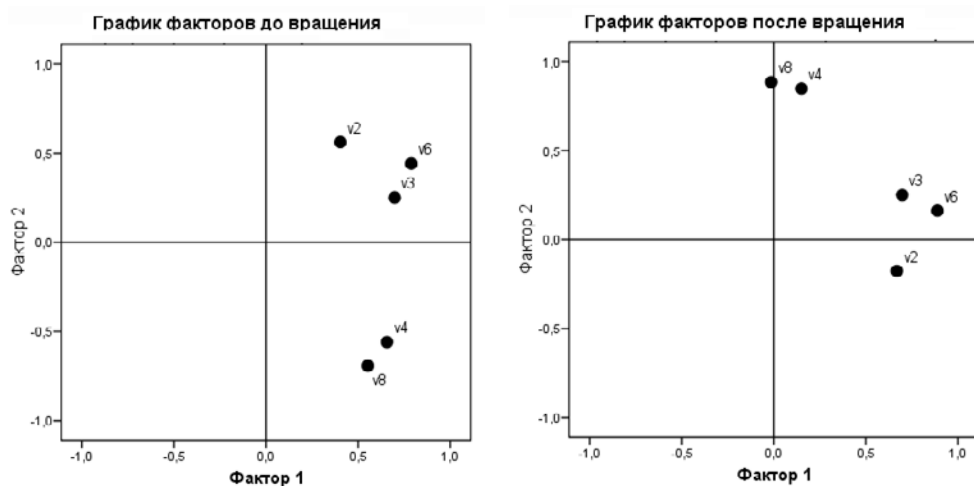


Рис. 20.1. Графическая интерпретация процедуры вращения факторов

Вращение факторов несколько не влияет на математическую строгость анализа: взаимное положение звездочек-переменных не меняется при повороте осей. Изначально процедура вращения выполнялась вручную, и исследователь самостоятельно задавал такое положение осей, при котором точки переменных максимально бы к ним приближались. Теперь SPSS позволяет выполнить несколько вариантов вращения, поворачивающих оси так, чтобы получить простую структуру, удовлетворяющую тому или иному критерию. Наиболее популярным вариантом вращения является метод Варимакс.

Вариант вращения Варимакс является ортогональным, поскольку при таком вращении оси сохраняют свое взаимное расположение под прямым углом. Иногда можно получить более предпочтительную простую структуру, если изменить угол между осями. Для этого предназначены варианты вращения Прямой облимин и Промакс. Наличие непрямого угла между осями факторов означает, что они не являются полностью независимыми друг от друга. Поскольку в реальных исследованиях факторы действительно не являются абсолютно независимыми, отклонения угла между осями от прямого при вращении вполне допустимо. Неортогональное вращение — достаточно сложная процедура и для ее применения необходимо четко представлять себе суть происходящего. В целом, аналогичную рекомендацию можно дать относительно факторного анализа вообще: не следует применять его без предварительного изучения соответствующих разделов статистики.

Интерпретация факторов

Итак, пусть в некоторой ситуации (близкой к идеальной) путем вращения мы добились того, что значение нагрузки для рассматриваемого фактора является большим (более 0,5), а для остальных факторов — малым (менее 0,2); кроме того, мы четко представляем смысл нашего фактора, то есть то, что он измеряет. Разумеется, в большинстве исследований переменные могут взаимодействовать с «ненужным» фактором, а нередко таких факторов может быть несколько. Как правило, исследователь не ограничивается только числовыми результатами факторного анализа; необходимым условием успеха факторного анализа является понимание содержательной специфики конкретных данных и взаимосвязей между ними. Вы сможете убедиться в этом при чтении раздела «Представление результатов» этой главы.

Для факторного анализа мы будем использовать данные реального тестирования интеллекта 46 школьников. Тест включал в себя 11 субтестов (переменные i_1, i_2, \dots, i_{11}), наименования которых вы найдете в разделе «Представление результатов». Предполагалось, что эти 11 субтестов позволят измерить 3 и более обобщенные интеллектуальные характеристики: математические, вербальные и невербальные (образные). Факторный анализ должен был установить соотношение субтестов и факторов.

Файл данных, который мы будем использовать в примере, называется TestIQ.sav; число объектов (N) равно 46.

В разделе пошаговых процедур приведено два варианта шага 5. Шаг 5 иллюстрирует простейший вариант факторного анализа, в котором используются значения по умолчанию для всех параметров. Шаг 5а, напротив, включает многие из действий, упомянутых ранее в этой главе.

Пошаговые алгоритмы вычислений

При проведении факторного анализа сначала выполняются три подготовительных шага. Эти шаги (шаги 1–3) позволят подготовить рабочий файл данных, запустить программу IBM SPSS Statistics 19 и открыть файл (в данном случае — файл TestIQ.sav). Пошаговые инструкции этого процесса приведены в главе 4 (с. 60), а подробные разъяснения — в главе 2.

После завершения шага 3 на экране должно присутствовать окно редактора данных со строкой меню и загруженным файлом TestIQ.sav.

ШАГ 4

В меню Анализ выберите команду Снижение размерности ► Факторный анализ. На экране появится диалоговое окно Факторный анализ, показанное на рис. 20.2.

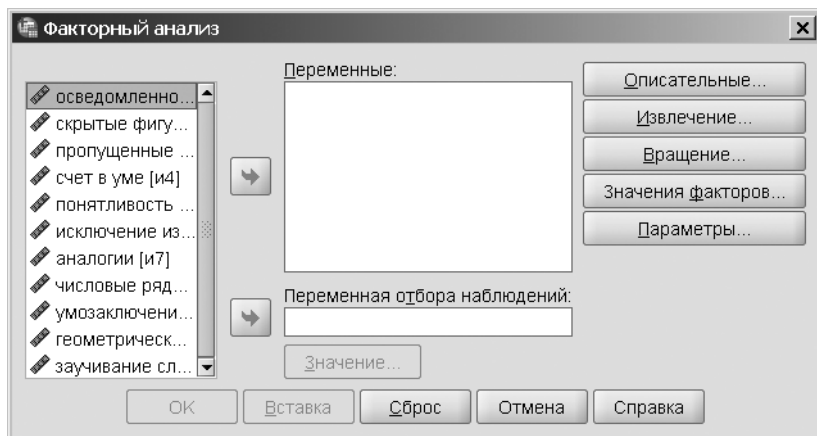


Рис. 20.2. Диалоговое окно Факторный анализ

Диалоговое окно Факторный анализ почти ничем не отличается от большинства диалоговых окон команд статистических операций SPSS: список доступных переменных находится в левой части, список Переменные, предназначенный для указания имен переменных, участвующих в анализе, — в центре, а 5 кнопок, с помощью которых задаются параметры анализа, — в правой. Несмотря на то что существует множество математических нюансов, касающихся факторного анализа, для его проведения в простейшем случае достаточно всего лишь поместить в список Переменные несколько переменных, выбрать вариант вращения и щелкнуть на кноп-

ке ОК. Далее будет приведен пример, в котором мы выберем 11 переменных и1, ..., и11 и проведем факторный анализ с параметрами по умолчанию.

Назначение поля Переменная отбора наблюдений примерно такое же, как у кнопки Если, упоминавшейся в главе 4. Оно позволяет указать имя переменной и ее уровень. Если, к примеру, поместить в поле Переменная отбора наблюдений имя пол, щелкнуть на кнопке Значение и ввести значение 1, то в факторном анализе будут участвовать только те объекты, для которых значения переменной пол равны 1.

Кнопки, расположенные в нижней части окна, позволяют управлять множеством разнообразных параметров анализа и могут использоваться независимо друг от друга в любом порядке.

Кнопка Описательные предназначена для открытия диалогового окна Факторный анализ: Описательные, показанного на рис. 20.3.

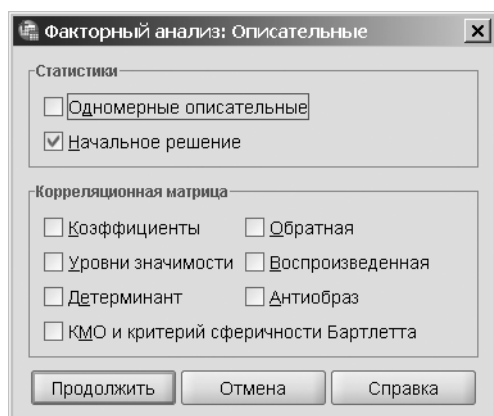


Рис. 20.3. Диалоговое окно Факторный анализ: Описательные

В этом окне находятся две группы флажков: Статистики и Корреляционная матрица. Флажок Одномерные описательные позволяет получить в таблице имена переменных, их средние значения, стандартные отклонения и метки. Флажок Начальное решение по умолчанию установлен и отвечает за включение в выводимые данные имен переменных, начальных общностей (по умолчанию равных 1,0), факторов, собственных значений, а также общего и кумулятивного процента общей дисперсии для каждого фактора. В группу Корреляционная матрица входит 7 флажков, управляющих выводом корреляционной матрицы; ниже приведены описания четырех из них, которые используются чаще других.

- ▶ Коэффициенты — включает в матрицу коэффициенты корреляции, ради которых она обычно создается.
- ▶ Уровни значимости — приводимые в отдельной таблице значения p -уровней, соответствующих вычисленным коэффициентам корреляции.

- Детерминант — детерминант корреляционной матрицы, использующийся для критериев многомерной нормальности.
- КМО и критерий сферичности Барлетта — два критерия: на многомерную нормальность (Бартлетта) и адекватность выборки (КМО определяет применимость факторного анализа к выбранным переменным). По умолчанию эти тесты не проводятся, но они обеспечивают процедуру важной начальной статистической информацией. Более подробно данный параметр рассматривается в разделе «Представление результатов».

Щелчок на кнопке Извлечение в окне Факторный анализ приводит к открытию диалогового окна Факторный анализ: Выделение факторов, представленного на рис. 20.4 и позволяющего выбрать метод извлечения, задать критерий числа факторов, сконфигурировать выводимые данные, связанные с процедурой извлечения, а также определить максимальное количество итераций, приводящих к получению результата.

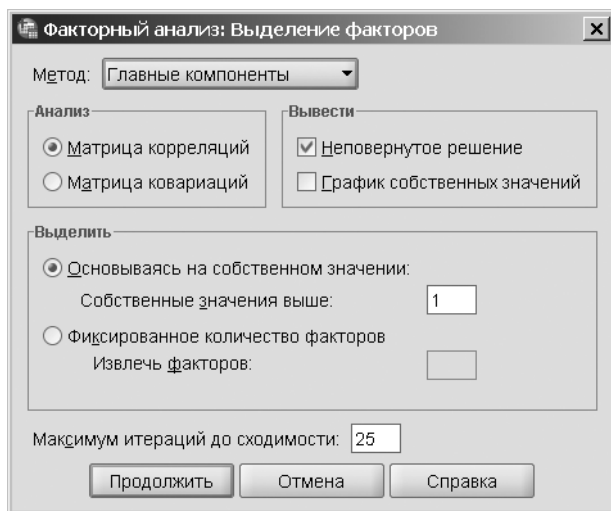


Рис. 20.4. Диалоговое окно Факторный анализ: Выделение факторов

Раскрывающийся список Метод содержит 7 пунктов. По умолчанию выбран пункт Главные компоненты; остальные пункты, соответствующие методам извлечения, перечислены ниже:

- Невзвешенный МНК;
- Обобщенный МНК;
- Максимум правдоподобия;
- Факторизация главной оси;

- Альфа-факторизация;
- Анализ образов.

Формально, именно эти 6 перечисленных методов относятся к факторному анализу. Тем не менее мы ограничимся анализом главных компонентов (АГК), как установленного по умолчанию и к тому же наиболее простого метода. Однако следует помнить, что АГК дает более грубое приближение к реальной структуре взаимосвязей, чем другие методы факторного анализа.

В группе Анализ с помощью переключателей Корреляционная матрица и Ковариационная матрица можно выбрать начальные условия анализа. Переключатели группы Выделить позволяют задать критерий числа извлекаемых факторов: если установлен переключатель Собственные значения выше, то можно воспользоваться единичным значением, установленным по умолчанию, либо задать свое значение. Переключатель Извлечь факторов позволяет извлечь указанное в поле справа количество факторов. Группа Вывести содержит два флажка: Неповернутое решение, установленный по умолчанию, и График собственных значений, часто используемый исследователями для определения числа факторов по критерию Р. Кеттелла. Если вы достаточно опытни в факторном анализе, то, вероятно, предпочтете установить оба указанных флажка. Наконец, при желании можно изменить значение в поле Максимум итераций до сходимости, по умолчанию установленное равным 25.

Щелчок на кнопке Вращение в окне Факторный анализ открывает диалоговое окно Факторный анализ: Вращение, показанное на рис. 20.5. При выполнении факторного анализа можно выбрать один из трех ортогональных методов вращения: наиболее популярные методы Варимакс, Эквимакс и Квартимакс. Переключатели Прямой облимин и Промакс позволяют выполнять неортогональное вращение факторов. Значения в полях Дельта и Каппа практически всегда рекомендуется оставлять установленными по умолчанию. Как говорилось ранее, настоятельно не рекомендуется использовать неортогональное вращение без исчерпывающих теоретических знаний в области факторного анализа.

Флажок Повернутое решение в группе Вывести по умолчанию установлен, если выбран один из методов вращения. Он позволяет включить в выводимые данные наиболее важные результаты факторного анализа. Более подробно об этом рассказано в разделе «Представление результатов». Если вам необходим визуальный анализ структуры факторов после вращения, установите флажок График(и) нагрузок. Тогда по умолчанию в выводимые данные будет включено трехмерное изображение первых трех факторов, если факторов больше двух. Вы можете перемещать оси и менять точку наблюдения изображения, но, пожалуй, неизменным в большинстве случаев будет одно: вы не сможете уловить смысла того, что видите.

Вернемся к основному диалоговому окну команды Фактор и обратимся к кнопке Значения факторов. При щелчке на ней открывается диалоговое окно, позволяющее сохранить вычисленные оценки факторов как переменные. Мы не приводим рисунок данного диалогового окна; вам достаточно знать, что для включения в выводимые данные матрицы факторных коэффициентов необходимо установить флажок

Вывести матрицу коэффициентов значений факторов. Установка флажка Сохранить как переменные позволяет сохранить в файле данных вычисленные значения факторных оценок для наблюдений в качестве новых переменных.

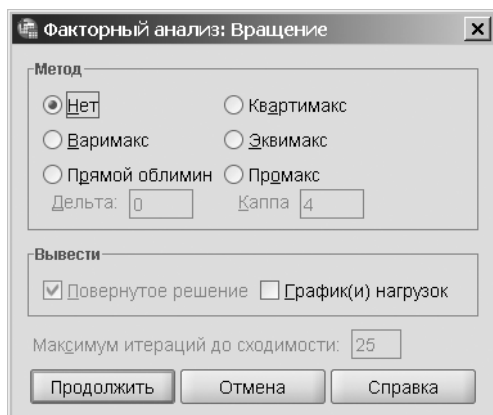


Рис. 20.5. Диалоговое окно Факторный анализ: Вращение

Кнопка Параметры в окне Факторный анализ предназначена для вызова диалогового окна Факторный анализ: Параметры, представленного на рис. 20.6.

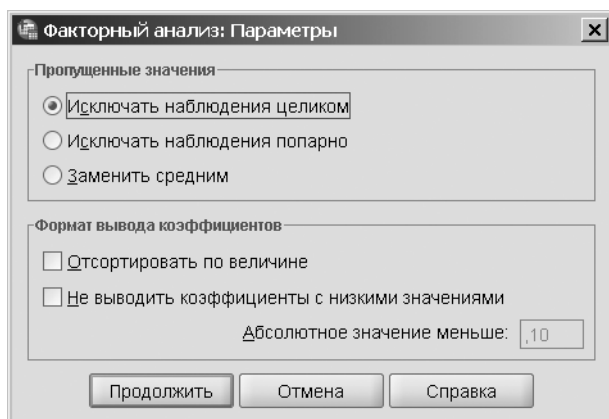


Рис. 20.6. Диалоговое окно Факторный анализ: Параметры

Элементы интерфейса, расположенные в этом окне в группе Формат вывода коэффициентов, управляют отображением матрицы факторных нагрузок после вращения. Часто полезно установить флажок Отсортировать по величине, который позволяет отсортировать переменные в соответствии с величиной их нагрузок по соответствующим факторам. Так, если 6 переменных максимально нагружают фактор 1, то эти переменные будут перечислены в порядке убывания их нагрузок в столбце с названием Фактор 1. Аналогичным образом будут выведены нагрузки для всех остальных факторов. Флажок Не выводить коэффициенты с низкими значе-

ниями позволяет исключить из таблицы все значения загрузок, меньшие заданной величины. Это делается для исключения из результатов тех загрузок, которые не имеют существенного для интерпретации значения.

Группа переключателей Пропущенные значения позволяет выбрать режим обработки отсутствующих значений при вычислении корреляций (см. главу 4). Установленный по умолчанию переключатель Исключать наблюдения целиком при вычислении корреляций исключает всю строку (наблюдение), если в ней пропущено хотя бы одно значение. Переключатель Исключать наблюдения попарно исключает только пару значений (для двух переменных), если одно из них пропущено. Это приводит к меньшим потерям наблюдений, чем при установке по умолчанию, но может вызвать искажение результатов, если пропусков много. Установка переключателя Заменить средним приводит к замене пропущенного значения переменной ее средним значением.

Далее приведены два варианта выполнения шага 5. В первом из них почти для всех параметров факторного анализа оставлены значения по умолчанию; второй вариант является более интересным и иллюстрирует применение некоторых из упоминавшихся параметров.

В следующем примере иллюстрируется применение факторного анализа к 11 переменным и1, ..., и11 файла TestIQ.sav с параметрами по умолчанию и вращением по методу Варимакс.

ШАГ 5

После выполнения предыдущего шага у вас должно быть открыто диалоговое окно Факторный анализ, показанное на рис. 20.2. Если вы уже успели поработать с этим окном, щелкните на кнопке Сброс.

1. Щелкните на переменной и1, нажмите клавишу Shift и, не отпуская ее, щелкните на переменной и11. В результате окажутся выделенными все промежуточные переменные, начиная от переменной и1 и заканчивая переменной и11.
2. Щелкните на верхней кнопке со стрелкой, чтобы переместить выделенные переменные в список Переменные.
3. Щелкните на кнопке Вращение, чтобы открыть диалоговое окно Факторный анализ: Вращение, показанное на рис. 20.5.
4. В группе Метод установите переключатель Варимакс и щелкните на кнопке Продолжить, чтобы вернуться в диалоговое окно Факторный анализ.
5. Щелкните на кнопке ОК, чтобы открыть окно вывода.

При выполнении этого шага проводится факторный анализ, включающий следующие операции.

1. Вычисление корреляционной матрицы для 11 заданных переменных.
2. Извлечение 11 факторов методом главных компонент.

3. Выбор для вращения всех факторов, чьи собственные значения не меньше 1.
4. Вращение факторов по методу Варимакс.
5. Вывод матрицы факторных нагрузок после вращения и других результатов.

В следующем примере проводится факторный анализ с участием тех же 11 переменных, что и в предыдущем случае, однако теперь задаются некоторые дополнительные параметры. Так, мы включим в вывод одномерные описательные статистики всех переменных, коэффициенты корреляции, а также применим критерии многомерной нормальности и адекватности выборки. Для извлечения факторов будет использоваться метод главных компонент, а для отображения — график собственных значений. Вращение факторов будет производиться методом Варимакс. Наконец, мы отсортируем переменные по величине их нагрузок по факторам и отобразим те нагрузки, абсолютная величина которых не менее 0,3.

ШАГ 5А

После выполнения шага 4 у вас должно быть открыто диалоговое окно Факторный анализ, показанное на рис. 20.2. Если вы уже успели поработать с этим окном, щелкните на кнопке Сброс.

1. Щелкните на переменной и1, нажмите клавишу Shift и, не отпуская ее, щелкните на переменной и11. В результате окажутся выделенными все промежуточные переменные, начиная от переменной и1 и заканчивая переменной и11.
2. Щелкните на верхней кнопке со стрелкой, чтобы переместить выделенные переменные в список Переменные.
3. Щелкните на кнопке Описательные, чтобы открыть диалоговое окно Факторный анализ: Описательные, показанное на рис. 20.3.
4. В группе Статистики установите флажок Одномерные описательные, в группе Корреляционная матрица — флажки Коэффициенты и КМО и критерий сферичности Барлетта и щелкните на кнопке Продолжить, чтобы вернуться в диалоговое окно Факторный анализ.
5. Щелкните на кнопке Извлечение, чтобы открыть диалоговое окно Факторный анализ: Выделение факторов, показанное на рис. 20.4.
6. Установите флажок График собственных значений и щелкните на кнопке Продолжить, чтобы вернуться в диалоговое окно Факторный анализ.
7. Щелкните на кнопке Вращение, чтобы открыть диалоговое окно Факторный анализ: Вращение, показанное на рис. 20.5.
8. В группе Метод установите переключатель Варимакс и щелкните на кнопке Продолжить, чтобы вернуться в диалоговое окно Факторный анализ.
9. Щелкните на кнопке Параметры, чтобы открыть диалоговое окно Факторный анализ: Параметры, показанное на рис. 20.6.

-
10. Установите флажки Отсортировать по величине и Не выводить коэф-
фициенты с низкими значениями; в окне Абсолютное значение меньше
напротив установите вместо 0,10 величину 0,3. Щелкните на кнопке
Продолжить, чтобы вернуться в диалоговое окно Факторный анализ.

11. Щелкните на кнопке ОК, чтобы открыть окно вывода.

После выполнения шага 5 программа автоматически активизирует окно вывода. Для просмотра результатов при необходимости можно воспользоваться вертикальной и горизонтальной полосами прокрутки. Обратите внимание на стандартную строку меню в верхней части окна вывода: ее присутствие позволяет выполнять любые статистические операции, не переключаясь обратно в окно редактора данных.

Представление результатов

Далее приведены данные, сгенерированные программой при выполнении шага 5а. На рис. 20.7 показан фрагмент, полученный благодаря установке флажка КМО и критерий сферичности Барлетта в окне Факторный анализ: Описательные статистики.

Мера адекватности и критерий Барлетта		
Мера выборочной адекватности Кайзера-Мейера-Олкина.		,704
Критерий сферичности Барлетта	Приблиз. хи-квадрат	146,069
	ст.св.	55
	Знач.	,000

Рис. 20.7. КМО и критерий сферичности Барлетта

Величина КМО демонстрирует приемлемую адекватность выборки для факторного анализа. Критерий сферичности Барлетта показывает статистически достоверный результат ($p < 0,05$): данные вполне приемлемы для факторного анализа.

В первой из двух таблиц, показанных на рис. 20.8, перечислены имена переменных и общности. Столбцы второй таблицы содержат характеристики выделенных факторов: их порядковые номера (с 1 по 3), суммы квадратов нагрузок, процент общей дисперсии, обусловленной фактором, и соответствующий кумулятивный (накопленный) процент (до и после вращения).

Чем больше процент дисперсии, обусловленной фактором (см. рис. 20.8), тем больший вес имеет данный фактор. А чем больше кумулятивный процент, накопленный к последнему фактору, тем более состоятельным является факторное решение. Если этот накопленный процент менее 50 %, следует либо сократить количество переменных, либо увеличить количество факторов. В данном случае накопленный процент дисперсии вполне приемлем.

Общности

	Начальные	Извлеченные
осведомленность	1,000	,572
скрытые фигуры	1,000	,618
пропущенные слова	1,000	,562
счет в уме	1,000	,687
понятливость	1,000	,573
исключение изображений	1,000	,502
анalogии	1,000	,631
числовые ряды	1,000	,628
умозаключения	1,000	,565
геометрическое сложение	1,000	,707
заучивание слов	1,000	,693

Метод выделения: Анализ главных компонент.

Полная объясненная дисперсия

Компонента	Суммы квадратов нагрузок извлечения			Суммы квадратов нагрузок вращения		
	Всего	% дисперсии	Кумуля- тивный %	Всего	% дисперсии	Кумуля- тивный %
1	3,701	33,646	33,646	2,649	24,084	24,084
2	1,796	16,332	49,978	2,089	18,989	43,073
3	1,240	11,275	61,253	2,000	18,180	61,253

Метод выделения: Анализ главных компонент.

Рис. 20.8. Имена переменных, общности и характеристики факторов

Как видите, в первом столбце первой таблицы перечислены метки, а не имена (и1, ..., и11) переменных. Если вы считаете, что метки занимают слишком много места в окне вывода, в меню командной строки Правка выберите команду Параметры, в открывшемся диалоговом окне перейдите на вкладку Метки в выводе. В разделе Метки в мобильных таблицах в раскрывающемся списке Переменные в метках вместо пункта Метки выберите пункт Имена и щелкните на кнопке ОК. После этого в таблицах окна вывода всегда будут представлены имена, а не метки переменных.

Приведенная на рис. 20.9 диаграмма называется *графиком собственных значений*, или *диаграммой каменистой осыти* (scree plot). Точками показаны соответствующие собственные значения в пространстве двух координат. Этот тип диаграммы обычно используется при определении достаточного числа факторов перед вращением. При этом руководствуются следующим правилом: оставлять нужно лишь те факторы, которым соответствуют первые точки на графике до того, как кривая станет более пологой. Отметим, что по умолчанию SPSS вращает все факторы, чьи собственные значения превышают 1; в данном примере число таких факторов равно 3, а в соответствии с упомянутым правилом нужно было бы взять не три, а четыре фактора.

Далее SPSS включает в вывод исходную структуру факторных нагрузок (до вращения). Эти данные в большинстве случаев не представляют интереса, и мы не станем приводить их в этом разделе. На рис. 20.10 показана квадратная матрица преобразований после вращения размером 3×3 .

График собственных значений

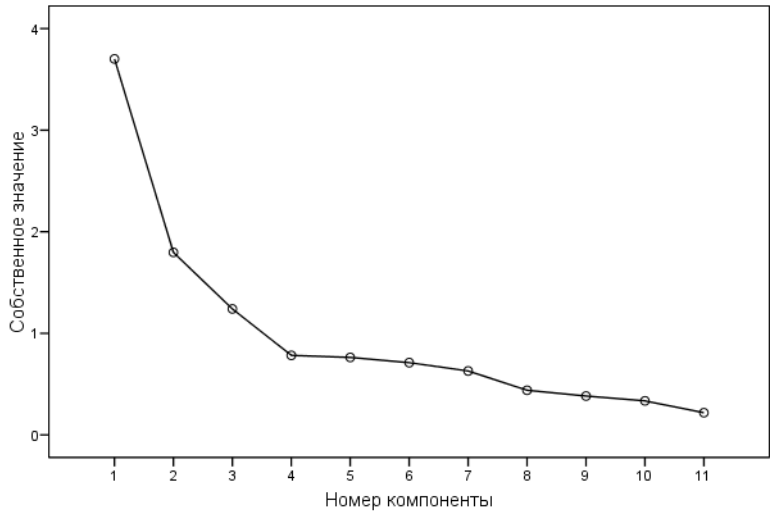


Рис. 20.9. График собственных значений

Матрица преобразования компонент

Компонента	1	2	3
1	,713	,428	,555
2	-,534	,845	,034
3	-,455	-,321	,831

Метод выделения: Анализ методом главных компонент.

Метод вращения: Варимакс с нормализацией Кайзера.

Рис. 20.10. Квадратная матрица преобразований

Если умножить матрицу преобразований на исходную матрицу факторных нагрузок (3×11), в результате получится показанная на рис. 20.11 преобразованная матрица факторных нагрузок после вращения. Именно эта матрица является главным итогом факторного анализа и подлежит содержательной интерпретации.

Обратите внимание, что благодаря установке флажка Отсортировать по величине в окне Факторный анализ: Параметры факторные нагрузки отсортированы следующим образом:

- ▶ наибольшие значения нагрузок для каждого фактора сортируются в отдельных блоках;
- ▶ внутри каждого блока нагрузки факторов упорядочены по убыванию.

Кроме того, установка флажка Не выводить коэффициенты с низкими значениями и задание Абсолютное значение меньше: 0,3 позволила вывести только существенные для интерпретации величины факторных нагрузок.

Матрица повернутых компонент^а

	Компонента		
	1	2	3
счет в уме	,804		,195
анalogии	,769	,183	
числовые ряды	,762	,217	
умозаключения	,696		,284
заучивание слов		,829	
осведомленность		,746	
пропущенные слова	,352	,630	,202
понятливость		,155	,739
скрытые фигуры	,153	,225	,737
геометрическое сложение	,421	-,224	,692
исключение изображений		,487	,506

Метод выделения: Анализ методом главных компонент.

Метод вращения: Варимакс с нормализацией Кайзера.

а. Вращение сошлось за 5 итераций.

Рис. 20.11. Матрица факторных нагрузок после вращения

Если бы описанный анализ провел реальный исследователь, он был бы удовлетворен полученными результатами. Очевидно, что первый из факторов соответствует предполагаемым математическим способностям, так как объединяет субтесты «счет в уме», «анalogии», «числовые ряды» и «умозаключения». Во второй фактор попали три субтеста, относящиеся к вербальным способностям: «заучивание слов», «осведомленность», «пропущенные слова», а в третий фактор — три субтеста, относящиеся к невербальным способностям: «скрытые фигуры», «геометрическое сложение», «исключение изображений». К «странностям» результатов можно отнести разве что распределение переменной «исключение изображений» между вторым и третьим фактором и попадание переменной «понятливость» в третий фактор. Подобные отклонения обычно требуют отдельного изучения. В частности, можно увеличить число факторов или исключить «неопределенные» переменные и повторить анализ.

Целью приведенного примера было показать, каким образом факторный анализ группирует переменные, объединяя их по факторам. Каждый фактор интерпретируется как причина совместной изменчивости (корреляции) группы переменных. Отметим напоследок, что чаще всего исследователь не ограничивается однократной факторизацией данных, а получает несколько вариантов решения с разными наборами переменных и разным числом факторов. Затем выбирается то решение, которое является наилучшим по признакам простоты структуры и концептуальной осмысленности. После получения приемлемого решения можно вычислить факторные оценки для объектов как новые переменные для дальнейшего анализа. Для этого в диалоговом окне Факторный анализ необходимо щелкнуть на кнопке Значения факторов... и в открывшемся диалоговом окне Факторный анализ: Значения факторов установить флажок Сохранить как переменные. В итоге будут созданы новые переменные (по количеству факторов), которые можно использовать в дальнейшем анализе вместо исходных переменных.

Терминология, используемая при выводе

Ниже дана трактовка терминов, используемых программой в окне вывода и относящихся к критериям КМО и Барлетта.

- ▶ КМО (мера выборочной адекватности Кайзера–Мейера–Олкина) — величина, характеризующая степень применимости факторного анализа к данной выборке:
- ▶ более 0,9 — безусловная адекватность;
- ▶ более 0,8 — высокая адекватность;
- ▶ более 0,7 — приемлемая адекватность;
- ▶ более 0,6 — удовлетворительная адекватность;
- ▶ более 0,5 — низкая адекватность;
- ▶ менее 0,5 — факторный анализ неприменим к выборке.
- ▶ Критерий сферичности Барлетта — критерий многомерной нормальности для распределения переменных. С его помощью проверяют, отличаются ли корреляции от 0. Значение p -уровня, меньшее 0,05, указывает на то, что данные вполне приемлемы для проведения факторного анализа.

Далее дана трактовка терминов, относящихся к общности и характеристикам факторов.

- ▶ Анализ главных компонент — метод извлечения, используемый программой SPSS по умолчанию.
- ▶ Начальные — по умолчанию все переменные имеют исходное значение общности, равное 1.
- ▶ Извлеченные — значения общностей после извлечения факторов.
- ▶ Компонента — номер извлеченного фактора (компоненты).
- ▶ Суммы квадратов нагрузок извлечения — величины, характеризующие информативность каждого из факторов до вращения.
- ▶ Суммы квадратов нагрузок вращения — величины, характеризующие информативность каждого из факторов после вращения.
- ▶ Всего — суммы квадратов нагрузок для каждого из факторов.
- ▶ Процент дисперсии — процент дисперсии, обусловленный фактором, равный отношению суммы квадратов нагрузок данного фактора к сумме исходных общностей (в данном случае равной 11).

- Кумулятивный процент — кумулятивный (накопленный) процент дисперсии.
- Матрица повернутых компонент — матрица факторных нагрузок после вращения, основной результат факторного анализа для содержательной интерпретации.

Завершение анализа и выход из программы

Отредактируйте содержимое окна вывода в соответствии со своими предпочтениями: скройте лишнюю информацию, исправьте таблицы и пр. (см. раздел «Окно вывода и его редактирование» главы 2). За инструкциями по редактированию графиков обратитесь к главе 5.

Для дальнейшего использования окончательного результата все содержимое окна вывода или его фрагменты можно сохранить в файле *.spv, экспортировать в другой формат (например, Word), перенести в документ Word или вывести на печать (подробности см. в разделе «Сохранение, экспорт, перенос и печать результатов» главы 2).

Для выхода из программы выберите команду Выход в меню Файл.

21 Кластерный анализ

297	Сравнение кластерного и факторного анализов
298	Этапы кластерного анализа
300	Кластерный анализ матрицы различий (сходства)
302	Пошаговые алгоритмы вычислений
309	Представление результатов
313	Завершение анализа и выход из программы

Программа SPSS реализует три метода кластерного анализа: Двухэтапный кластерный анализ (TwoStep), Кластеризация К-средними (K-means) и Иерархическая кластеризация (Hierarchical).

Двухэтапный кластерный анализ позволяет выявить группы (кластеры) объектов по заданным переменным, если эти группы действительно существуют. При этом программа автоматически определяет количество существующих кластеров (групп). Если невозможно однозначно определить количество кластеров, все объекты помещаются в один.

Кластеризация К-средними разбивает по заданным переменным все множество объектов на заданное пользователем число кластеров так, чтобы средние значения для кластеров по каждой из переменных максимально различались.

Иерархическая кластеризация, как наиболее гибкий из рассматриваемых методов, позволяет детально исследовать структуру различий между объектами и выбрать наиболее оптимальное число кластеров. В силу этого иерархический кластерный анализ применяется наиболее часто и будет нами рассмотрен более подробно.

Зачастую описание нового статистического метода удобно проводить путем его сравнения с другим методом. Именно таким образом мы и начнем рассмотрение иерархического кластерного анализа, сравнивая его с факторным анализом. При многочисленных общих чертах между указанными статистическими методами существует немало различий. Если вы еще не знакомы с понятием факторного анализа, настоятельно рекомендуем вам обратиться к началу главы 20, поскольку далее мы займемся сравнительной характеристикой факторного и кластерного анализов. Как обычно, после теоретической части последуют примеры практической реализации статистических методов средствами SPSS, оформленные в виде пошаговых процедур.

Сравнение кластерного и факторного анализов

Главное сходство между кластерным и факторным анализами заключается в том, что тот и другой предназначены для перехода от исходной совокупности множества переменных (или объектов) к существенно меньшему числу факторов (кластеров). Тем не менее реализация статистических процедур и интерпретация результатов для двух типов анализа различаются; пять основных отличий приведены ниже.

- ▶ Целью факторного анализа является замена большого числа исходных переменных меньшим числом факторов. Кластерный анализ, как правило, применяется для того, чтобы уменьшить число объектов путем их группировки. Другими словами, в процедуре кластерного анализа обычно переменные не группируются, а выступают в качестве критериев для группировки объектов. Так, в примере факторного анализа (см. главу 20) 11 субтестов интеллекта (переменных) были сведены к трем факторам, каждый из которых объединил несколько родственных исходных переменных. Кластерный анализ обычно применяется для выделения групп объектов, исходя из их сходства по измеренным признакам. Применительно к примеру с 11 субтестами интеллекта типичной задачей кластерного анализа была бы классификация учащихся (объектов) таким образом, чтобы по измеренным 11 показателям внутри каждой группы объекты были бы более похожи друг на друга, чем на объекты из других групп. Группы объектов, выделенные в результате кластерного анализа на основе заданной меры сходства между объектами, называются кластерами.
- ▶ Заявленные в предыдущем пункте различия между кластерным и факторным вариантами анализа со всей полнотой категоричности могут быть отнесены лишь к ранним версиям SPSS. Начиная с версии SPSS 10.0, программа позволяет с равным успехом проводить кластерный анализ не только объектов, но и переменных. В последнем случае кластерный анализ может выступать как более простой аналог факторного анализа. В разделе пошаговых процедур мы продемонстрируем оба варианта кластерного анализа.
- ▶ Действия, выполняемые в ходе статистических операций в каждом из вариантов анализа, принципиально различаются. В факторном анализе на каждом этапе извлечения фактора для каждой переменной подсчитывается доля дисперсии, которая обусловлена влиянием данного фактора. При кластерном анализе вычисляется расстояние между текущим объектом и всеми остальными объектами, и кластер образует та пара, для которой расстояние оказалось наименьшим. Подобным образом каждый объект группируется либо с другим объектом, либо включается в состав существующего кластера. Процесс кластеризации конечен и продолжается до тех пор, пока все объекты не будут объединены в один кластер. Разумеется, подобный результат в общем случае

не имеет смысла, и исследователь должен самостоятельно определить, в какой момент кластеризация должна быть прекращена.

- ▶ В контексте кластерного анализа особое место занимает один из его видов, называемый *иерархическим кластерным анализом*. В SPSS он реализуется с помощью команды Иерархическая кластеризация. Этот вид кластерного анализа чаще используется в биологии, экономике, социологии, политологии, нежели в психологии. Психологи обычно анализируют переменные с целью найти статистические связи между ними; эти связи, как правило, указывают на сходство между теми или иными исследуемыми факторами. Деление выборки на группы в психологических анализах редко представляет интерес; в случаях, когда это оказывается необходимым, психологи отдают предпочтение дискриминантному, а не кластерному анализу (см. главу 22).
- ▶ Поскольку кластеризация переменных оказывается весьма доступной операцией, было бы интересно сравнить ее результаты с результатами более сложного факторного анализа. Как и в случае факторного анализа, выполнение кластерного анализа и его результаты зависят от ряда параметров: способа вычисления расстояния между объектами, кластеризации индивидуальных объектов и т. д.

Этапы кластерного анализа

Кластерный анализ выполняется в несколько этапов, приводящих к конечному результату. Сначала мы рассмотрим пример, созданный специально, чтобы показать суть кластерного анализа. Отметим, что кластерный анализ неприменим к файлам данных, использовавшимся ранее, поскольку при их составлении основное внимание было уделено смыслу и связям между переменными, а содержимое объектов (то есть информация, касающаяся субъектов) практически не играло роли. Для демонстрации кластерного анализа нами был подготовлен файл `cars.sav`, содержащий гипотетические данные о 15 подержанных автомобилях разных марок, выставленных на продажу. Файл имеет структуру, подходящую для наглядной иллюстрации кластерного анализа.

Итак, выделяют несколько этапов кластерного анализа.

1. *Выбор переменных-критериев для кластеризации.* В нашем примере кластеризация будет осуществляться по следующим переменным: цена (стоимость), `t_сост` (экспертная оценка технического состояния по 10-балльной шкале), возраст (количество лет эксплуатации), пробег (пройденный километраж с начала эксплуатации).
2. *Выбор способа измерения расстояния между объектами, или кластерами* (изначально считается, что каждый объект соответствует одному кластеру). По умолчанию используется квадрат Евклидова расстояния, согласно которому расстояние между объектами равно сумме квадратов разностей между значе-

ниями одноименных переменных объектов. Предположим, что марка автомобиля А имеет показатели технического состояния и возраста 5 и 6, а марка В — соответственно 7 и 4. Тогда по этим двум переменным (координатам) расстояние между марками А и В вычисляется следующим образом: $(5 - 7)^2 + (6 - 4)^2 = 8$. При выполнении анализа сумма квадратов разностей вычисляется для всех переменных. Получаемые расстояния используются программой при формировании кластеров. Помимо Евклидова существуют и другие виды расстояний, вычисляемые по другим формулам, однако мы не будем на них останавливаться. При необходимости обратитесь к руководству пользователя SPSS.

Относительно вычисления расстояния может возникнуть следующий вопрос: будет ли адекватным результат кластерного анализа в том случае, если переменные имеют различные шкалы измерения? Так, все переменные файла cars.sav имеют самые разные шкалы. Для решения проблемы шкалирования в SPSS используется стандартизация, в частности ее простой метод — нормализация переменных, приводящая все переменные к стандартной z-шкале (среднее равно 0, стандартное отклонение — 1). При нормализации всех переменных при проведении кластерного их веса становятся одинаковыми. В случае если все исходные данные имеют одну и ту же шкалу измерения либо веса переменных по смыслу должны быть разными, стандартизацию переменных проводить не нужно.

3. *Формирование кластеров.* Существует два основных метода формирования кластеров: *метод слияния* и *метод дробления*. В первом случае исходные кластеры увеличиваются путем объединения до тех пор, пока не будет сформирован единственный кластер, содержащий все данные. Метод дробления основан на обратной операции: сначала все данные объединяются в один кластер, который затем делится на части до тех пор, пока не будет достигнут желаемый результат. По умолчанию программой SPSS используется метод слияния, и мы рассмотрим его в этой главе.

В методе слияния предусмотрены несколько способов объединения объектов. Способ, применяемый по умолчанию, называется *межгрупповым связыванием*, или *связыванием средних внутри групп*. SPSS вычисляет наименьшее среднее значение расстояния между всеми парами групп и объединяет две группы, оказавшиеся наиболее близкими. На первом шаге, когда все кластеры представляют собой одиночные объекты, данная операция сводится к обычному попарному сравнению расстояний между объектами. Термин «среднее значение» приобретает смысл лишь на втором этапе, когда сформированы кластеры, содержащие более одного объекта. Так, в нашем примере на начальном этапе имеется 15 кластеров (объектов); сначала в кластер объединяются два объекта с наименьшим расстоянием друг от друга. Затем подсчет расстояний повторяется, и в кластер объединяется еще одна пара переменных. На втором этапе вы получите либо 13 свободных объектов и один кластер, объединяющий 2 объекта, либо 11 свободных объектов и 2 кластера по 2 объекта в каждом. В конечном счете, все объекты окажутся в одном большом кластере. Существуют и другие

методы объединения объектов, однако мы не станем рассматривать их в этой книге. При необходимости обратитесь к руководству пользователя SPSS.

4. *Интерпретация результатов.* Как и в случае факторного анализа, желаемое число кластеров и оценка результатов анализа зависят от целей исследователя. Для рассматриваемого примера нам представляется наиболее предпочтительным число кластеров, равное 3. Как показывает анализ, все марки можно разделить на 3 группы: первая группа имеет высокую стоимость (среднее значение — 15 230), небольшой срок эксплуатации (4 года) и средний пробег (85 400 км). Вторая группа имеет среднюю стоимость, небольшой пробег, наибольший возраст, но хорошее техническое состояние. Третья группа содержит недорогие модели с большим пробегом и невысоким рейтингом технического состояния.

Кластерный анализ матрицы различий (сходства)

Довольно часто исходной информацией для кластерного анализа являются не данные типа «объекты-переменные», а непосредственно данные о различии (сходстве) объектов. Например, респондент может оценивать различие (сходство) объектов, предъявляемых попарно, с перебором всех пар объектов. Или данными могут быть частоты совместной встречаемости для каждой пары объектов. В этих случаях исходные данные представляют собой квадратную матрицу, как правило, симметричную относительно главной диагонали, каждый элемент которой — мера различия (или сходства) пары объектов, которым соответствует строка и столбец матрицы.

К сожалению, кнопочный интерфейс программы SPSS не позволяет обрабатывать матрицы различий методами кластерного анализа (в многомерном шкалировании это возможно). Поэтому для обработки таких данных следует обратиться к командному языку Синтаксис (Syntax).

В SPSS предусмотрен формат данных для матрицы различия (сходства). При помощи синтаксиса такая матрица создается следующим образом¹. Для создания командного файла Синтаксис необходимо открыть окно *редактора синтаксиса*, выбрав команду Открыть ► Синтаксис в меню Файл. Предположим, необходимо создать матрицу различий 8 × 8 с именами объектов a b c d e f g h. Для этого в открытом окне Редактор синтаксиса следует ввести следующий текст (не важно, прописными или строчными буквами):

```
DATA LIST FREE /ROWTYPE_(a8) VARNAME_(a8) a b c d e f g h.
BEGIN DATA
PROX      A      0 6 2 7 18 0 1 5
PROX      B      6 0 14 3 8 4 19 14
```

¹ Решение проблемы заимствовано с сайта <http://www.spsstools.ru> (А. Балабанов, Raynald's SPSS Tools по-русски).

```

PROX    C    2 14 0 30 12 4 6 8
PROX    D    7 3 30 0 1 16 3 4
PROX    E   18 8 12 1 0 19 13 14
PROX    F    0 4 4 16 19 0 5 1
PROX    G    1 19 6 3 13 5 0 16
PROX    H    5 14 8 4 14 1 16 0
END DATA.
EXECUTE.
VALUE LABELS ROWTYPE_ 'PROX' 'DISIMILARITY'.

```

Результатом выполнения этой команды будет создание матрицы различий 8×8 в окне редактора данных.

Матрица может содержать меры не различия, а сходства, например, если каждое значение матрицы — частота совместной встречаемости. Тогда последнюю строку следует заменить на следующую:

```
VALUE LABELS ROWTYPE_ 'PROX' 'SIMILARITY'.
```

Чтобы выполнить введенную команду, ее нужно выделить и щелкнуть на кнопке Запустить выделенный фрагмент панели инструментов. Второй вариант запуска — выбрать в меню команду Запуск ► Все. Имейте в виду, что любой пропущенный знак, включая завершающую точку или неверно написанное слово, приведет к выдаче программой сообщения об ошибке.

После выполнения команды в редакторе данных можно заменить полученную матрицу на свои данные, следя за тем, чтобы количество строк равнялось количеству столбцов, в каждой строке переменной ROWTYPE_ стояло значение DISIMILARITY (или SIMILARITY, если данные — меры сходства), а в строках переменной VARNAME_ присутствовало имя объекта.

Для выполнения кластерного анализа методом межгруппового связывания с выводом таблицы шагов агломерации и дендрограммы следует выполнить следующий синтаксис:

```

CLUSTER
/MATRIX IN  (*)
/METHOD BAVERAGE
/PRINT SCHEDULE
/PLOT DENDROGRAM.

```

Вся синтаксическая конструкция, создающая матрицу *различий* 8×8 (попарных различий 8 объектов) и выполняющая кластерный анализ указанным методом сохранена в папке примеров Examples в файле под именем Synt_Clust.sps¹. При желании вы можете открыть этот файл и выполнить его указанным выше способом без предварительной редакции или отредактировав. Например, вы можете заменить указание матрицы различий на задание матрицы сходства. Для этого

¹ Все файлы данных, включая и указанный файл «синтаксис», вы можете найти по адресу <http://www.piter.com/books/download/978591180318>.

достаточно заменить в строке `VALUE LABELS ROWTYPE_` слово `DISIMILARITY` на слово `SIMILARITY`.

Пошаговые алгоритмы вычислений

Структура раздела пошаговых процедур принципиально ничем не отличается от аналогичных разделов в других главах книги: шаги 1–4 являются подготовительными для выполнения анализа, шаги 6–7 требуются для печати результатов и завершения работы программы, а несколько вариантов (в данном случае два) шага 5 соответствуют разным способам выполнения статистической операции. Сначала мы продемонстрируем вариант использования кластерного анализа объектов с включением в вывод нескольких интересующих нас величин. Вторым вариантом шага 5 проиллюстрирует кластерный анализ переменных. В нем мы вновь обратимся к файлу `TestIQ.sav` для того, чтобы вы имели возможность сравнить кластерный анализ переменных с факторным анализом, описанным в предыдущей главе.

При проведении кластерного анализа сначала выполняются три подготовительных шага. Эти шаги (шаги 1–3) позволят подготовить рабочий файл данных, запустить программу IBM SPSS Statistics 19 и открыть файл (в данном случае — файл `cars.sav`). Пошаговые инструкции этого процесса приведены в главе 4 (с. 60), а подробные разъяснения — в главе 2.

После завершения шага 3 на экране должно присутствовать окно редактора данных со строкой меню и загруженным файлом `cars.sav`.

ШАГ 4

В меню Анализ выберите команду Классификация ► Иерархическая кластеризация. Откроется диалоговое окно Иерархический кластерный анализ, показанное на рис. 21.1.

Ваши дальнейшие действия в значительной степени зависят от того, какой тип кластеризации вы выберете. Для этой цели в группе Кластеризовать предусмотрены два переключателя: Наблюдения и Переменные. Вначале в списке Переменные указываются имена тех переменных, значения которых будут использоваться при кластеризации. В нашем примере в список Переменные следует поместить все переменные, кроме переменной марка, поскольку последняя представляет собой марку автомобиля. Далее следует задать способ идентификации объектов. Как правило, в роли идентификатора выступает переменная, содержащая уникальный номер объекта или его имя в виде строки. В данном случае мы будем использовать вполне подходящую для этого переменную марка. Имя идентифицирующей переменной указывается в поле Метить значениями.

Если вместо переключателя Наблюдения в группе Кластеризовать установить переключатель Переменные, в списке Переменные потребуется указать кластеризуемые переменные, а поле Метить значениями останется пустым. По умолчанию флажки Статистики и Графики в группе Вывести установлены, и в большинстве случаев нет необходимости их сбрасывать. В правой части диалогового окна расположены 4 кнопки, предназначенные для задания дополнительных параметров команды.

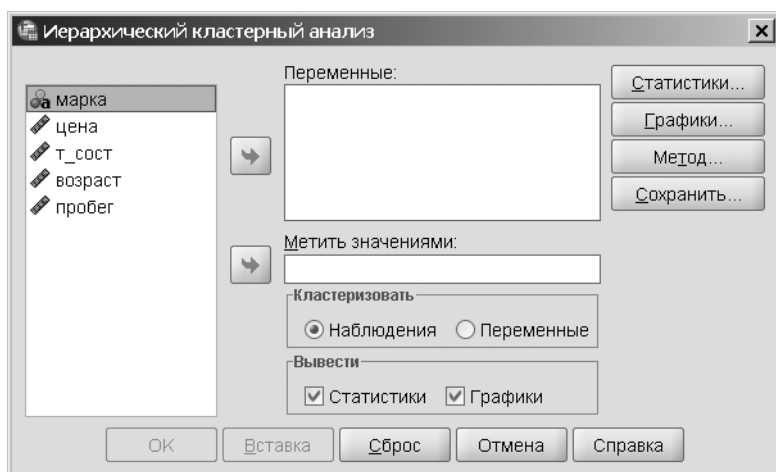


Рис. 21.1. Диалоговое окно Иерархический кластерный анализ

При щелчке на кнопке Статистики на экране появляется диалоговое окно Иерархический кластерный анализ: Статистики, показанное на рис. 21.2.

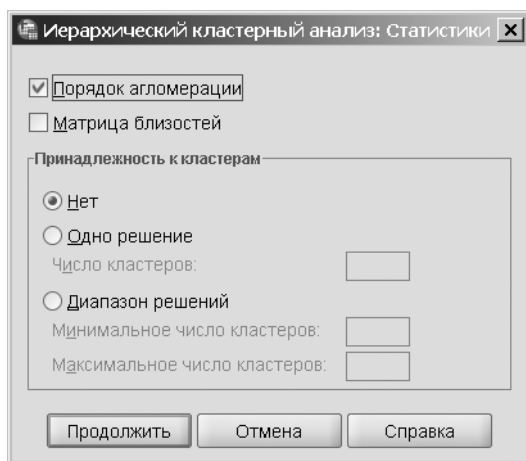


Рис. 21.2. Диалоговое окно Иерархический кластерный анализ: Статистики

Флажок **Порядок агломерации** по умолчанию установлен, обеспечивая включение в результаты кластерного анализа стандартного компонента окна вывода, который будет подробно рассмотрен в разделе «Представление результатов». Флажок **Матрица близостей** предназначен для отображения информации о расстояниях между объектами и кластерами. Использование матрицы удобно лишь для небольших файлов данных, поскольку с увеличением числа объектов размер матрицы резко возрастает, что делает ее громоздкой и неудобной для восприятия. Группа **Принадлежность к кластерам** состоит из трех переключателей, описанных ниже.

- **Нет** — в выводимые результаты включаются все кластеры. Этот вариант установлен по умолчанию.

- Одно решение — позволяет задать точное число кластеров в решении.
- Диапазон решений — обеспечивает вывод нескольких решений с разным числом кластеров. Так, если ввести в поле Минимальное число кластеров число 3, а в поле Максимальное число кластеров число 5, то в выводимые результаты будут включены все решения с 3, 4 и 5 кластерами.

Щелчок на кнопке Графики в окне Иерархический кластерный анализ открывает диалоговое окно Иерархический кластерный анализ: Графики, показанное на рис. 21.3.

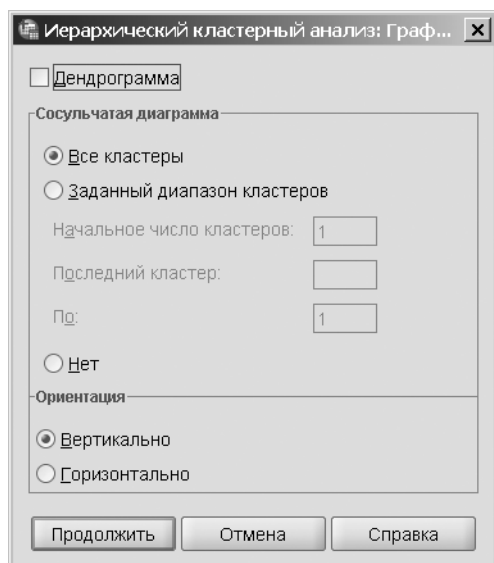


Рис. 21.3. Диалоговое окно Иерархический кластерный анализ: Графики

Флажок Дендограмма позволяет включить в выводимые результаты ту же информацию, которая содержится на «сосульчатой» диаграмме, предлагаемой по умолчанию, а также относительную величину разности между переменными или кластерами на каждом шаге процесса. Поскольку Сосульчатая диаграмма лишь дублирует информацию других разделов, а ее вид оставляет желать лучшего, мы откажемся от ее вывода. Для этого в группе Сосульчатая диаграмма достаточно установить переключатель Нет.

Щелчок на кнопке Метод в окне Иерархический кластерный анализ открывает диалоговое окно Иерархический кластерный анализ: Метод, показанное на рис. 21.4.

В этом окне для нас представляют интерес раскрывающиеся списки Метод, Мера: Интервальная и Стандартизация. Мы опишем лишь наиболее важные пункты этих списков и кратко упомянем остальные. Более детальные описания вы можете найти самостоятельно в руководстве пользователя SPSS.

В списке Метод наиболее часто используется пункт Межгрупповые связи. Суть метода заключается в том, что объединению на каждом шаге подвергаются кластеры

или объекты, расстояние между которыми минимально. Более полное описание межгруппового связывания можно найти в разделе «Представление результатов».

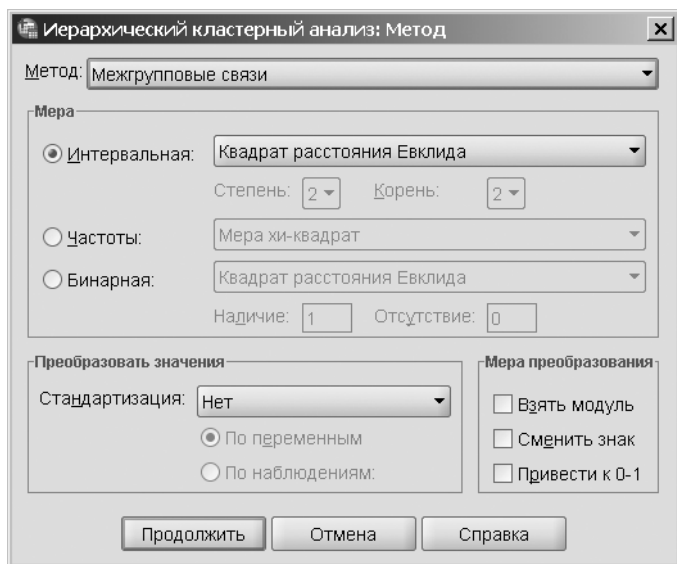


Рис. 21.4. Диалоговое окно Иерархический кластерный анализ: Метод

Помимо пункта Межгрупповые связи раскрывающийся список Метод содержит следующие пункты:

- ▶ Внутригрупповые связи
- ▶ Ближайший сосед;
- ▶ Дальний сосед;
- ▶ Центроидная кластеризация;
- ▶ Медианная кластеризация;
- ▶ Метод Варда.

В раскрывающемся списке Интервальная по умолчанию выбран пункт Квадрат расстояния Евклида. Это означает, что расстояние между объектами вычисляется как разность квадратов соответствующих переменных этих объектов, участвующих в анализе. Остальные пункты списка Интервальная:

- ▶ Косинус — метод измерения близости, основанный на косинусах векторов значений;
- ▶ Корреляция Пирсона — метод измерения близости, основанный на корреляции векторов значений;
- ▶ Чебышев — вычисление расстояния как максимума абсолютной величины разности между элементами;

- Блок — определяет меру расстояния по метрике города (см. главу 23);
- Минковского — определяет меру расстояния Минковского (см. главу 23);
- Настроенная — позволяет задавать пользовательскую меру расстояния.

Процедура стандартизации выбирается в раскрывающемся списке Стандартизация. По умолчанию выбран пункт Нет, однако в случаях, когда переменные представлены в разных шкалах (единицах измерения) стандартизация необходима, и чаще всего выбирают пункт z-значения. Оставшиеся пункты, в которых допускается варьирование среднего значения или стандартного отклонения распределения, могут привести к другим результатам; выбор какого-либо из них определяется степенью его применимости к исследуемым данным и удобством для исследователя.

В группе Преобразовать значения по умолчанию установлен переключатель По переменным, и в большинстве случаев менять его на альтернативный (По наблюдениям) нет необходимости.

В группе Мера преобразования имеются три флажка, позволяющих изменить значения переменных: Взять модуль, Сменить знак и Привести к 0–1.

Последняя из четырех функциональных кнопок окна Иерархический кластерный анализ — кнопка Сохранить — открывает диалоговое окно Иерархический кластерный анализ: Сохранить, показанное на рис. 21.5. С помощью этого окна можно создавать новые переменные, значения которых будут указывать принадлежность наблюдений кластерам.

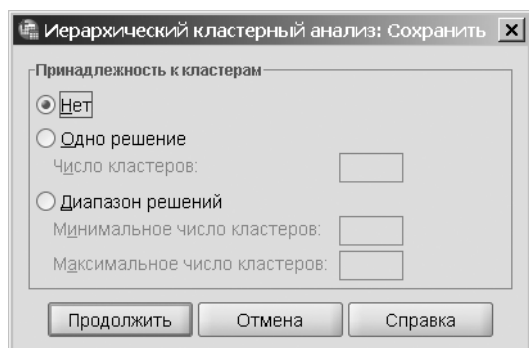


Рис. 21.5. Диалоговое окно Иерархический кластерный анализ: Сохранить

Если установлен переключатель Нет, никакого сохранения в процессе анализа не производится. Иначе при выполнении анализа будут созданы переменные, которые окажутся в конце файла. Если поставить переключатель в Одно решение и указать в поле число 3, то получим новую переменную, значение которой равно 1, 2 или 3 в зависимости от того, какому кластеру будет принадлежать соответствующий объект в решении. Если же установить переключатель Диапазон решений, в поле Минимальное число кластеров указать число 3, а в поле Максимальное число кластеров — число 5, это приведет к созданию трех новых переменных: первая будет принимать значения от 1 до 3, вторая — от 1 до 4, третья — от 1 до 5.

В следующем примере проводится кластерный анализ с несколькими дополнительными параметрами, обсуждавшимися выше. В качестве идентификатора используется переменная марка. Все остальные переменные файла задействуются для вычисления расстояния между объектами. Мы включим в выводимые результаты последовательность слияния и дендрограмму, но исключим диаграмму накопления. Значения всех переменных нормализуем для того, чтобы придать им равные веса и привести к одной шкале. В качестве расстояния между объектами зададим квадрат Евклидова расстояния, а в качестве метода кластеризации — межгрупповое связывание. Кроме того, мы создадим новую переменную, в которой сохраним решение с тремя кластерами.

ШАГ 5

После выполнения шага 4 должно быть открыто диалоговое окно Иерархический кластерный анализ, показанное на рис. 21.1. Если вы уже успели поработать с этим окном, щелкните на кнопке Сброс.

1. Щелкните сначала на переменной марка, а затем — на нижней кнопке со стрелкой, чтобы переместить переменную в поле Метить значениями.
2. Щелкните сначала на переменной цена и, нажав на клавиатуре кнопку Shift, щелкните на переменной пробег. В результате окажутся выделенными все оставшиеся в списке переменные.
3. Щелкните на верхней кнопке со стрелкой, чтобы переместить выделенные переменные в список Переменные.
4. Щелкните на кнопке Графики, чтобы открыть диалоговое окно Иерархический кластерный анализ: Графики, показанное на рис. 21.3.
5. Установите флажок Дендограмма и переключатель Нет в группе Соусльчатая диаграмма. Щелкните на кнопке Продолжить, чтобы вернуться в диалоговое окно Иерархический кластерный анализ.
6. Щелкните на кнопке Метод, чтобы открыть диалоговое окно Иерархический кластерный анализ: Метод, показанное на рис. 21.4.
7. В списке Метод оставьте выбранным пункт Межгрупповые связи, в списке Стандартизация выберите пункт z-значения и щелкните на кнопке Продолжить, чтобы вернуться в диалоговое окно Иерархический кластерный анализ.
8. Щелкните на кнопке Сохранить, чтобы открыть диалоговое окно Иерархический кластерный анализ: Сохранить, показанное на рис. 21.5.
9. Установите переключатель Одно решение, введите в расположенное рядом поле значение 3 и щелкните на кнопке Продолжить, чтобы вернуться в Иерархический кластерный анализ.
10. Щелкните на кнопке ОК, чтобы открыть окно вывода.

В следующем примере проводится кластерный анализ, в котором вместо объектов участвуют переменные. Мы используем данные файла TestIQ.sav, содержащего 11 переменных и1, ..., и11. Обратите внимание на следующее обстоятельство. Обычно при группировании переменных исследователя интересует их взаимосвязь, а не их различие (сходство), как при группировании объектов. Исключением является

случай, когда данные представляют собой оценки объектов экспертами — тогда строки соответствуют экспертам, а столбцы оцениваемым объектам. Поскольку в нашем примере интерес представляют именно взаимосвязи между переменными, и мы хотим сравнить результаты с факторным анализом, то в качестве меры близости целесообразно выбрать корреляцию. При этом корреляции надо учитывать по абсолютной величине, так как большие (по модулю) отрицательные их величины так же свидетельствуют о связи, как и большие положительные. Все это необходимо иметь в виду, если речь идет о кластеризации переменных. Большинство остальных параметров команды оставим установленными по умолчанию; даже в стандартизации в данном случае нет необходимости, так как на величину корреляции не влияют единицы измерения переменных. Добавим лишь дендрограмму в выводимые результаты и исключим оттуда сосульчатую диаграмму. Обратите внимание, поскольку для выполнения анализа необходим файл `TestIQ.sav`, а не `cars.sav`, который использовался в предыдущем примере, мы начинаем с шага 3а.

ШАГ 3А

Откройте файл данных, с которым вы намерены работать (в нашем случае — это файл `TestIQ.sav`). Если он расположен в текущей папке, то выполните следующие действия.

1. Выберите в меню Файл команду Открыть ► Данные или щелкните на кнопке Открыть файл данных панели инструментов.
2. В открывшемся диалоговом окне дважды щелкните на имени `TestIQ.sav` или введите его с клавиатуры и щелкните на кнопке ОК.

ШАГ 4А

В меню Анализ выберите команду Классификация ► Иерархическая кластеризация. Откроется диалоговое окно Иерархический кластерный анализ, показанное на рис. 21.1.

ШАГ 5А

После выполнения предыдущего шага должно быть открыто диалоговое окно Иерархический кластерный анализ. Если вы уже успели поработать с этим окном, щелкните на кнопке Сброс.

1. В группе Кластеризовать установите переключатель Переменные.
2. Нажмите кнопку мыши на переменной `и1` и, не отпуская кнопки, перетяните указатель на переменную `и11`, затем кнопку мыши отпустите. В результате окажутся выделенными все 11 переменных `и1`, `и2`, ..., `и11`.
3. Щелкните на верхней кнопке со стрелкой, чтобы переместить выделенные переменные в список Переменные.
4. Щелкните на кнопке Графики, чтобы открыть диалоговое окно Иерархический кластерный анализ: Графики, показанное на рис. 21.3.
5. Установите флажок Дендограмма и переключатель Нет в группе Сосульчатая диаграмма. Щелкните на кнопке Продолжить, чтобы вернуться в диалоговое окно Иерархический кластерный анализ.
6. Щелкните на кнопке Метод, чтобы открыть диалоговое окно Иерархический кластерный анализ: Метод, показанное на рис. 21.4.

7. В ниспадающем списке Интервальная выберите пункт Корреляция Пирсона, а в группе Мера преобразований установите флажок Взять модуль. Щелкните на кнопке Продолжить, чтобы вернуться в диалоговое окно Иерархический кластерный анализ.
8. Щелкните на кнопке ОК, чтобы открыть окно вывода.

Эта процедура реализует кластерный анализ 11-ти показателей теста интеллекта. В выводимые данные включается такая информация о переменных, как число наблюдений, число пропущенных значений и т. п. Затем выводятся шаги агломерации (объединения в кластеры) и горизонтальная дендрограмма. Рекомендуем сравнить полученные результаты с результатами факторного анализа в главе 20.

После выполнения шага 5 (5а) программа автоматически откроет окно вывода. Для просмотра результатов при необходимости можно воспользоваться вертикальной и горизонтальной полосами прокрутки. Обратите внимание на стандартную строку меню в верхней части окна вывода: ее присутствие позволяет выполнять любые статистические операции, не переключаясь обратно в окно редактора данных.

В следующем примере кластерный анализ проводится с применением языка Синтаксис.

ШАГ 3Б

Создайте файл Синтаксис, как это описано в разделе «Кластерный анализ матрицы различий (сходства)» в этой главе, или откройте файл примера Synt_Clust.sps. Если он расположен в текущей папке, то выполните следующие действия.

1. Выберите в меню Файл команду Открыть ► Синтаксис.
2. В открывшемся диалоговом окне дважды щелкните на имени Synt_Clust.sps. Откроется редактор синтаксиса с текстом команд.
3. Выберите в меню команду Запуск ► Все.

После выполнения шага 3б будет создан и открыт новый файл данных, содержащий матрицу 8×8 попарных различий 8 объектов. Кроме того, будет открыто окно вывода, содержащее результаты кластерного анализа этих 8 объектов методом средней связи (между группами). В окне редактора данных вы можете заменить представленную матрицу своими данными и выполнить кластерный анализ, отредактировав открытое окно редактора синтаксиса. Для этого надо просто удалить все строки с начала до строки CLUSTER и запустить команду Запуск ► Все.

Представление результатов

В этом разделе представлены фрагменты выводимых данных, сгенерированные программой при выполнении шагов 5 и 5а.

После выполнения шага 5 в окне вывода появится показанная на рис. 21.6 таблица Шаги агломерации. В этой таблице вторая колонка Кластер объединен с со-

держит первый (Кластер 1) и второй (Кластер 2) столбцы, которые соответствуют номерам кластеров, объединяемых на данном шаге. После объединения кластеру присваивается номер, соответствующий номеру в колонке Кластер 1. Так, на первом шаге объединяются объекты 5 и 14, и кластеру присваивается номер 5, далее этот кластер на шаге 3 объединяется с элементом 4, и новому кластеру присваивается номер 4 и т. д. Следующая колонка Коэффициент содержит значение расстояния между кластерами, которые объединяются на данном шаге. Колонка Этап первого появления кластера показывает, на каком шаге до этого появлялся первый и второй из объединяемых кластеров. Последняя колонка Следующий этап показывает, на каком шаге снова появится кластер, образованный на этом шаге.

Шаги агломерации						
Этап	Кластер объединен с		Коэффициенты	Этап первого появления кластера		Следующий этап
	Кластер 1	Кластер 2		Кластер 1	Кластер 2	
1	5	14	,439	0	0	3
2	2	12	,550	0	0	5
3	4	5	,671	0	1	6
4	7	13	,942	0	0	10
5	2	15	1,203	2	0	7
6	4	11	1,586	3	0	11
7	2	8	1,810	5	0	9
8	6	10	1,847	0	0	11
9	1	2	2,471	0	7	13
10	3	7	4,136	0	4	13
11	4	6	4,492	6	8	12
12	4	9	5,914	11	0	14
13	1	3	9,656	9	10	14
14	1	4	10,498	13	12	0

Рис. 21.6. Таблица шагов агломерации

Как видно по таблице, на первом этапе происходит объединение в кластер пары объектов, расстояние между которыми является наименьшим. На втором этапе SPSS снова подсчитывает расстояния между объектами и объединяет в кластер пару наиболее близких объектов; при этом в результате может получиться либо один кластер из трех объектов, либо два кластера из двух объектов. Процесс слияния продолжается до тех пор, пока все объекты не попадут в один кластер. Попробуем объяснить результаты, полученные на этапах 1 и 13.

- На этапе 1 происходит объединение объектов 5 и 14. Расстояние между объектами равно 0,439. Ни один из двух объектов не принадлежит какому-либо кластеру, о чем свидетельствуют нули в столбцах Cluster 1 (Кластер 1) и Cluster 2 (Кластер 2) колонки Stage Cluster First Appears (Этап первого появления кластера). Следующим этапом для данного кластера, судя по столбцу Next Stage (Следующий этап), является этап 3, на котором к кластеру присоединяется объект 4.
- На этапе 13 происходит объединение кластеров, содержащих объекты 1 и 3. Объект 1 был объединен с кластером, содержащим объект 2 на этапе 9, а объ-

ект 3 — с объектами 7 и 13 на этапе 10. Расстояние между объединяемыми на этом этапе кластерами равно 9,656. Образованный на этом этапе кластер появляется далее на следующем шаге.

По таблице шагов агломерации можно предварительно оценить число кластеров. Для этого необходимо проследить динамику увеличения расстояний по шагам кластеризации и определить шаг, на котором отмечается резкое возрастание расстояний. Оптимальному числу классов соответствует разность между числом объектов и порядковым номером шага, на котором было обнаружено резкое возрастание расстояний. Так, в нашем примере это обнаруживается при переходе от шага 12 к шагу 13. Следовательно, наиболее оптимальное количество кластеров должно быть получено на шаге 12 или 13. Оно равно численности объектов минус номер шага, то есть $15 - 12 = 3$ или $15 - 13 = 2$, то есть 3 или 2 кластера. Выбор того или иного решения зависит уже от содержательных соображений.

Дендрограмма представляет процесс кластеризации в форме древовидной структуры (рис. 21.7).

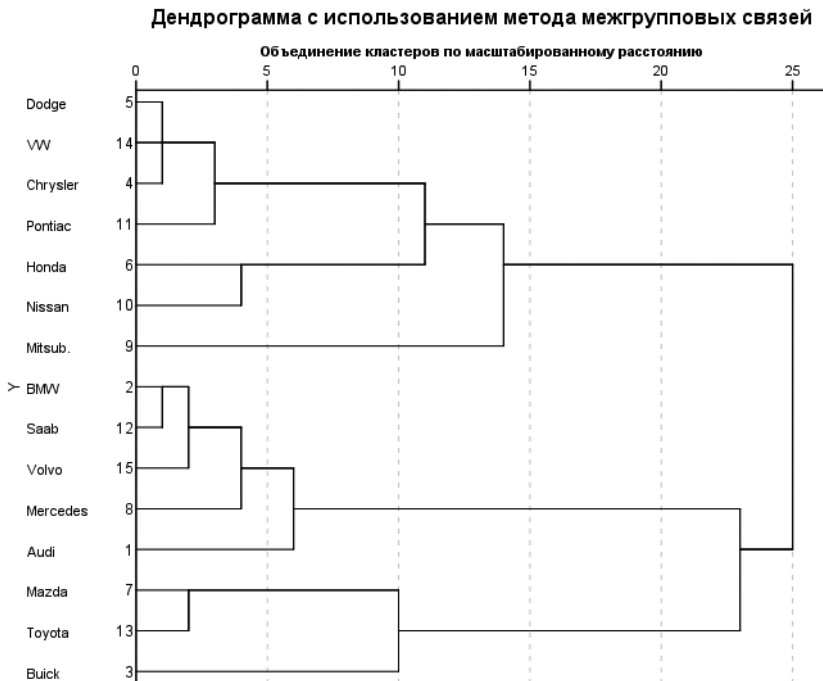


Рис. 21.7. Дендрограмма, полученная в результате выполнения шага 5

Дендрограмма не только позволяет перейти к любому объекту на любом уровне кластеризации, но и дает возможность судить о том, каково расстояние между кластерами или объектами на каждом из уровней. Числа от 0 до 25 являются

условной шкалой этих расстояний; 0 соответствует наименьшему расстоянию на первом этапе, а 25 — наибольшему расстоянию на последнем этапе.

На дендрограмме любое решение характеризуется вертикальной линией, число точек пересечения которой с деревом соответствует количеству кластеров на текущем этапе. Так, для нашего примера эту линию следует расположить на уровне 15–20: между шагами 12 и 13 кластеризации. В этом случае получается три кластера. Для того чтобы установить состав каждого кластера, необходимо вернуться к корням дерева и выяснить соответствующие номера объектов. Напомним, что в результате выполнения шага 5 в файле cars.sav появилась новая переменная, определяющая принадлежность каждого объекта к одному из трех кластеров.

В заключение на рис. 21.8 приведем дендрограмму, полученную в результате выполнения шага 5а (кластеризация 11 переменных).

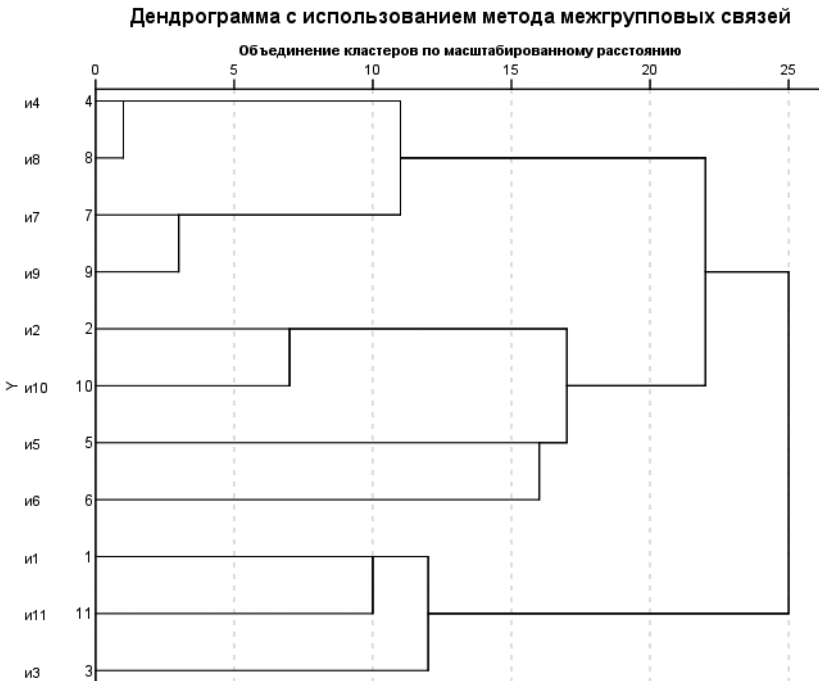


Рис. 21.8. Дендрограмма, полученная в результате выполнения шага 5а

Дендрограмма показывает, что в результате кластеризации переменные группируются в три кластера, состав которых идентичен факторам, полученным в отношении тех же данных при факторном анализе (см. главу 20).

Завершение анализа и выход из программы

Отредактируйте содержимое окна вывода в соответствии со своими предпочтениями: скройте лишнюю информацию, исправьте таблицы и пр. (см. раздел «Окно вывода и его редактирование» главы 2). За инструкциями по редактированию графиков обратитесь к главе 5.

Для дальнейшего использования окончательного результата все содержимое окна вывода или его фрагменты можно сохранить в файле *.spv, экспортировать в другой формат (например, Word), перенести в документ Word или вывести на печать (подробности см. в разделе «Сохранение, экспорт, перенос и печать результатов» главы 2).

Для выхода из программы выберите команду Выход в меню Файл.

22 Дискриминантный анализ

316	Этапы дискриминантного анализа
317	Пошаговые алгоритмы вычислений
324	Представление результатов
328	Терминология, используемая при выводе
331	Завершение анализа и выход из программы

Дискриминантный анализ позволяет предсказать принадлежность объектов к двум или более непересекающимся группам. Исходными данными для дискриминантного анализа является множество объектов, разделенных на группы так, что каждый объект может быть отнесен только к одной группе. Допускается при этом, что некоторые объекты не относятся ни к какой группе (являются «неизвестными»). Для каждого из объектов имеются данные по ряду количественных переменных. Такие переменные называются дискриминантными переменными, или предикторами. Задачами дискриминантного анализа является определение:

- ▶ решающих правил, позволяющих по значениям дискриминантных переменных (предикторов) отнести каждый объект (в том числе и «неизвестный») к одной из известных групп;
- ▶ «веса» каждой дискриминантной переменной для разделения объектов на группы.

Существует множество ситуаций, в которых было бы весьма желательно вычислить вероятность того или иного исхода в зависимости от совокупности измеряемых переменных: например, подходит ли соискатель работы на ту или иную должность, страдает психически больной человек шизофренией или психозом, вернется заключенный в тюрьму или к нормальной жизни после выхода на свободу, какие факторы влияют на увеличение риска пациента получить сердечный приступ и т. п. Во всех перечисленных ситуациях есть две общие черты: во-первых, для некоторых субъектов (не для всех) есть информация об их принадлежности к той или иной группе; во-вторых, о каждом субъекте имеется дополнительная информация для создания формулы, которая позволит спрогнозировать принадлежность субъекта к той или иной группе.

Дискриминантный анализ имеет определенное сходство с кластерным анализом; сходство заключается в том, что исследователь в обоих случаях ставит перед собой цель разделить совокупность объектов (а не переменных) на несколько более мелких (значимых) групп. Тем не менее процесс классификации в двух видах анализа принципиально различен. В кластерном анализе объекты классифицируются на основе их различий без какой-либо предварительной информации о количестве

и составе классов. В дискриминантном анализе изначально заданы количество и состав классов, и основная задача заключается в определении того, насколько точно можно предсказать принадлежность объектов к классам при помощи данного набора дискриминантных переменных (предикторов).

Дискриминантный анализ представляет собой альтернативу множественного регрессионного анализа (см. главу 18) для случая, когда зависимая переменная представляет собой не количественную, а номинальную переменную. При этом дискриминантный анализ решает, по сути, те же задачи, что и множественный регрессионный анализ: предсказание значений «зависимой» переменной (в данном случае — категорий номинального признака) и определение того, какие «независимые» переменные лучше всего подходят для такого предсказания. Дискриминантный анализ основан на составлении уравнения регрессии (см. главы 17 и 18), использующего номинальную зависимую переменную (обратите внимание на то, что она не является количественной, как в случае регрессионного анализа). Уравнение регрессии составляется на основе тех объектов, о которых известна групповая принадлежность, что позволяет максимально точно подобрать его коэффициенты. После того как уравнение регрессии получено, его можно использовать для группировки интересующих нас объектов в целях прогнозирования их принадлежности к какому-либо классу. Команда дискриминантного анализа весьма непроста и требует настройки множества параметров, описание большинства из которых лежит за рамками темы данной книги. Тем не менее при необходимости вы можете обратиться за дополнительной информацией к руководству пользователя SPSS.

Как и для большинства сложных статистических операций, параметры дискриминантного анализа в основном определяются особенностями данных, а также задачами исследователя. Как всегда, мы рассмотрим пример (на этот раз единственный) проведения дискриминантного анализа в разделе пошаговых процедур, а раздел «Представление результатов» посвятим интерпретации выводимых данных.

Для демонстрации дискриминантного анализа мы рассмотрим пример прогнозирования успешности обучения на основе предварительного тестирования. Файл `class.sav` содержит данные о 46 учащихся (объекты с 1 по 46), юношей и девушек (переменная `пол`), закончивших курс обучения, в отношении которых известны оценки успешности обучения — для этого используется переменная `оценка` (1 — низкая, 2 — высокая). Кроме того, в файл включены данные предварительного тестирования этих учащихся до начала обучения (13 переменных):

- ▶ `и1, ..., и11` — 11 показателей теста интеллекта;
- ▶ `э_и` — показатель экстраверсии по тесту Г. Айзенка (H. Eysenck);
- ▶ `н` — показатель нейротизма по тесту Г. Айзенка.

Еще для 10 претендентов на курс обучения (объекты с 47 по 56) известны лишь результаты их предварительного тестирования (13 перечисленных переменных). Значения переменной `оценка` для них, разумеется, неизвестны, и в файле данных им соответствуют пустые ячейки. В процессе дискриминантного анализа мы, в частности, попытаемся спрогнозировать успешность обучения этих 10 претендентов в предположении, что выборки закончивших обучение и претендентов идентичны.

Этапы дискриминантного анализа

Дискриминантный анализ состоит из трех основных этапов.

1. *Выбор переменных-предикторов.* Исследователь использует свои теоретические знания, практический опыт, догадки и т. п. для того, чтобы составить список переменных, которые могут повлиять на результат группировки (переменную-критерий). В рассматриваемом файле помимо переменной-критерия (оценка) содержится 13 переменных, характеризующих каждого учащегося; это позволяет нам сделать все 13 переменных предикторами и включить их в уравнение регрессии. Если бы число переменных было велико (например, несколько сотен), было бы невозможно применить дискриминантный анализ ко всем переменным одновременно. Это обусловлено как концептуальными причинами (возможная коллинеарность переменных, потеря степеней свободы и т. п.), так и практическими ограничениями (недостаточный объем оперативной памяти компьютера). Обычно на начальном этапе дискриминантного анализа для предикторов формируется корреляционная матрица. В данном контексте она имеет особый смысл, называется *общей внутригрупповой корреляционной матрицей* и содержит средние коэффициенты корреляции для двух или более корреляционных матриц (каждая для одной группы). Помимо общей внутригрупповой корреляционной матрицы можно также вычислить ковариационные матрицы для отдельных групп, для всей выборки либо общую внутригрупповую ковариационную матрицу. Нередко исследователи применяют серию t -критериев между двумя группами для каждой переменной либо однофакторный дисперсионный анализ, если число групп оказывается больше двух. Поскольку целью дискриминантного анализа является составление наилучшего уравнения регрессии, дополнительный анализ исходных данных никогда не является лишним. Так, в результате применения t -критериев для данных нашего примера были найдены значимые различия между двумя уровнями переменной оценка для 8 из 13 предикторов. Мы рассмотрим один из наиболее распространенных вариантов дискриминантного анализа, при проведении которого программа автоматически исключает несущественные для предсказания предикторы, но по критериям, которые устанавливает сам исследователь.
2. *Выбор параметров.* В этой главе будет продемонстрирован один из вариантов дискриминантного анализа. По умолчанию программа реализует метод, который основан на *принудительном включении* в регрессионное уравнение всех предикторов, указанных исследователем. В другом варианте используется метод Уилкса (Wilks), относящийся к категории *пошаговых методов* и основанный на минимизации коэффициента Уилкса (λ) после включения в уравнение регрессии каждого нового предиктора. Так же как и в случае множественного регрессионного анализа, существует критерий для включения предикторов в уравнение регрессии (по умолчанию таким критерием является $F > 3,84$) и критерий для исключения предикторов из уравнения регрессии (по умолчанию $F < 2,71$). Коэффициент λ представляет собой отношение внутригрупповой суммы квадратов к общей сумме квадратов и характеризует долю влияния

предиктора на дисперсию критерия. Со значением λ связаны величины F и p , характеризующие его значимость. Более полное описание вы можете найти в разделе «Представление результатов».

Какой же из двух методов предпочтительнее? Как показывает практика, зачастую компьютер справляется с составлением уравнения регрессии лучше, чем исследователь, задающий список предикторов вручную. Однако встречаются ситуации, когда полезней ограничить самостоятельность компьютера. Например, если провести дискриминантный анализ для наших данных с включением всех переменных, то неверно классифицированы будут 5 объектов из 46. Той же точности прогноза можно достичь всего с 7 предикторами, если выбрать пошаговый метод с *отличающимися* от принятых по умолчанию установками (как указано ниже в пошаговом алгоритме). В то же время, если использовать пошаговый метод с установками по умолчанию, оставляющий только три предиктора, количество неверно сгруппированных объектов увеличится до 9. Помимо рассмотренных программа SPSS располагает и другими методами выбора предикторов, однако их описание выходит за рамки темы данной книги, и при необходимости мы рекомендуем вам обратиться к руководству пользователя SPSS.

3. *Интерпретация результатов.* Целью дискриминантного анализа является составление уравнения регрессии с использованием выборки, для которой известны значения и предикторов, и критерия. Это уравнение позволяет по известным значениям предикторов определить неизвестные значения критерия для другой выборки. Разумеется, точность рассчитываемых значений критерия для второй выборки в общем случае не выше, чем для исходной. Так, в нашем примере регрессионное уравнение обеспечило около 90 % корректных результатов для той выборки, с помощью которой оно было создано. Соответственно, точность предсказания успешности обучения для 10 претендентов может достигать 90 % лишь в том случае, если выборка претендентов совершенно идентична тем 46 учащимся, данные для которых послужили основой для прогноза.

Пошаговые алгоритмы вычислений

Перед началом дискриминантного анализа выполняются три подготовительных шага. Эти шаги (шаги 1–3) позволят подготовить рабочий файл данных, запустить программу IBM SPSS Statistics 19 и открыть файл (в данном случае — файл class.sav). Пошаговые инструкции этого процесса приведены в главе 4 (с. 60), а подробные разъяснения — в главе 2.

После завершения шага 3 на экране должно присутствовать окно редактора данных со строкой меню и загруженным файлом class.sav.

ШАГ 4

В меню Анализ выберите команду Классификация ► Дискриминантный анализ. На экране появится диалоговое окно Дискриминантный анализ, показанное на рис. 22.1.

Помимо таких привычных элементов интерфейса, как список доступных переменных слева и 5 стандартных кнопок справа, диалоговое окно Дискриминантный анализ содержит 4 дополнительных элемента.

- ▶ Поле Группировать по предназначено для задания единственной зависимой переменной.
- ▶ В список Независимые включается любое число переменных, участвующих в дискриминантном анализе в качестве предикторов.
- ▶ Переключатели Принудительное включение и Шаговый отбор позволяют выбрать один из двух вариантов дискриминантного анализа.
- ▶ Пять кнопок управления дополнительными параметрами в правой части диалогового окна предназначены для открытия дополнительных диалоговых окон.

Назначение окна Переменная отбора наблюдений примерно такое же, как у кнопки Если, упоминавшейся в главе 4. Оно позволяет указать имя переменной и ее уровень. Если, к примеру, поместить в окно Переменная отбора имя пол, щелкнуть на кнопке Значение и ввести значение 1, в дискриминантном анализе будут участвовать только те объекты, для которых значения переменной пол равны 1.

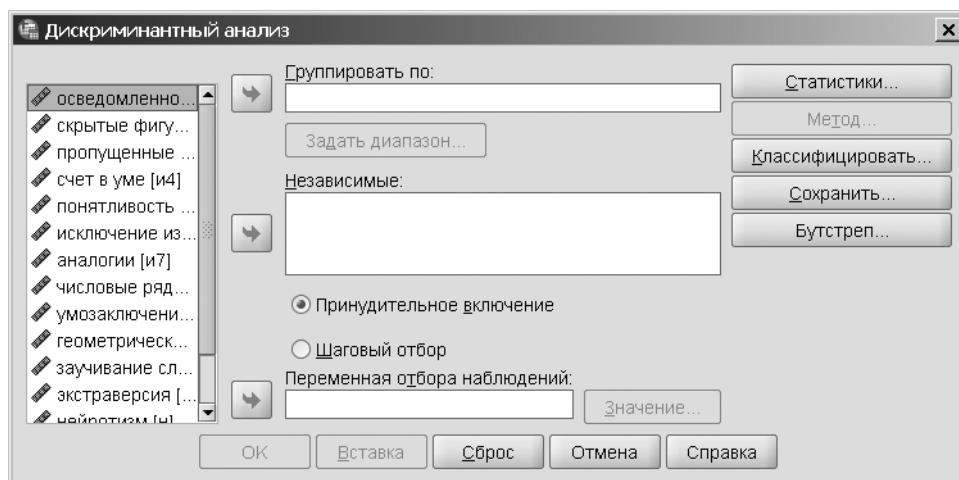


Рис. 22.1. Диалоговое окно Дискриминантный анализ

Кнопка Сохранить позволяет сохранять в качестве новых переменных следующие величины для каждого объекта (в том числе, «неизвестного»):

- ▶ прогнозируемый номер группы;
- ▶ оценки дискриминантных функций;
- ▶ вероятность принадлежности к каждой группе.

Важными элементами интерфейса диалогового окна Дискриминантный анализ являются элементы управления уровнями зависимой переменной. Хотя в нашем

примере дискриминантный анализ используется для разделения объектов на две группы, количество групп может быть и больше. После ввода имени зависимой переменной в поле Группировать по, щелчок на кнопке Задать диапазон позволяет открыть диалоговое окно Дискриминантный анализ: Задание диапазона, показанное на рис. 22.2. Поскольку переменная оценка, используемая в нашем примере как зависимая, имеет лишь два уровня (1 и 2), их следует указать в полях Минимум и Максимум, а затем щелкнуть на кнопке Продолжить. Если число уровней группирующей переменной больше двух, описанная операция позволит задать любой диапазон уровней.

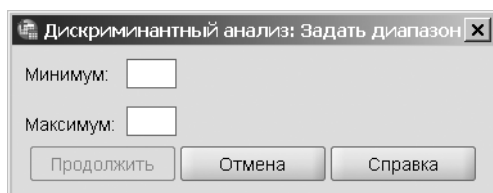


Рис. 22.2. Диалоговое окно Дискриминантный анализ: Задать диапазон

Щелчок на кнопке Статистики в окне Дискриминантный анализ открывает диалоговое окно Дискриминантный анализ: Статистики, показанное на рис. 22.3. Поскольку практически все элементы интерфейса этого окна активно используются исследователями, далее приведено описание каждого из них. Как правило, с помощью этих элементов исследователи выясняют наличие зависимостей между предикторами перед началом дискриминантного анализа.

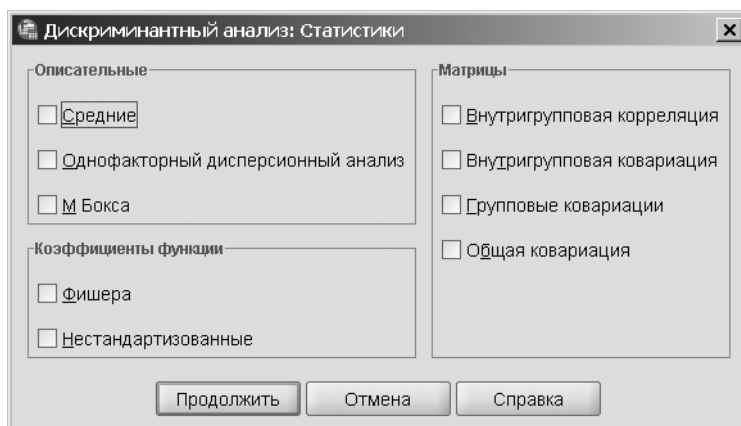


Рис. 22.3. Диалоговое окно Дискриминантный анализ: Статистики

Флажки в группе Описательные:

- Средние — средние значения и стандартные отклонения для каждой переменной каждой группы (в данном случае, двух групп, образованных уровнями переменной оценка) и всей выборки в целом;

- ▶ Однофакторный дисперсионный анализ — если число уровней зависимой переменной равно 2, то это t -критерии для сравнения между собой двух средних значений для групп по каждой переменной;
- ▶ М Бокса — критерий равенства дисперсионно-ковариационных матриц для уровней зависимой переменной.

Флажки в группе Коэффициенты функций:

- ▶ Фишера — канонические коэффициенты дискриминантных функций;
- ▶ Нестандартизованные — нестандартизованные коэффициенты дискриминантных функций, вычисленные для исходных значений предикторов.

Флажки в группе Матрицы:

- ▶ Внутригрупповая корреляция — матрица, состоящая из соответствующих средних значений корреляционных матриц для уровней зависимой переменной;
- ▶ Внутригрупповая ковариация — матрица, аналогичная предыдущей, с той разницей, что вместо корреляций используются ковариации;
- ▶ Групповые ковариации — отдельные ковариационные матрицы для каждого уровня зависимой переменной;
- ▶ Общая ковариация — ковариационная матрица для всей выборки.

Щелчок на кнопке Метод в окне Дискриминантный анализ приводит к появлению диалогового окна Дискриминантный анализ: Шаговый отбор, показанного на рис. 22.4. Обратите внимание, что кнопка Метод доступна лишь в том случае, если установлен переключатель Шаговый отбор в окне Дискриминантный анализ.

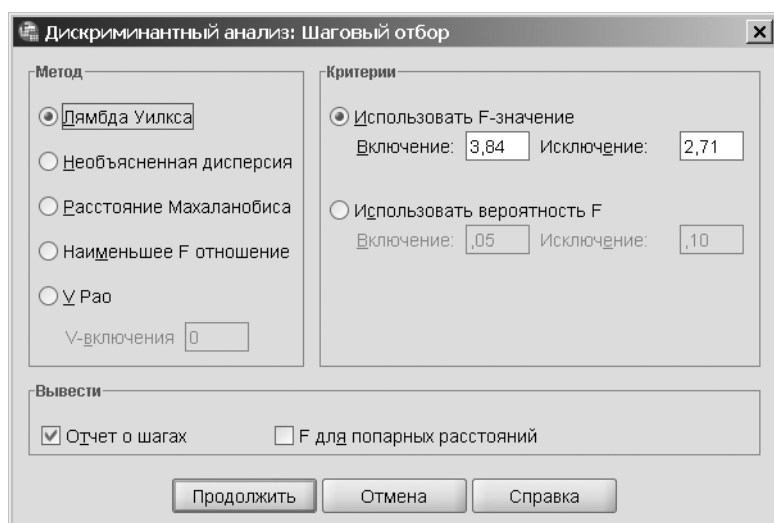


Рис. 22.4. Диалоговое окно Дискриминантный анализ: Шаговый отбор

В этом окне при помощи группы переключателей Метод можно выбрать один из пошаговых методов составления дискриминантного уравнения, в области Критерии задать критерии для включения в дискриминантное уравнение и исключения из него предикторов, а также включить в окно вывода желаемые величины при помощи флажков в группе Вывести. Из переключателей выбора метода составления уравнения для нас представляет интерес лишь переключатель Лямбда Уилкса. Соответствующий метод основан на минимизации коэффициента Уилкса (λ) после включения в уравнение регрессии каждого нового предиктора. Более подробное описание приведено в разделе «Представление результатов».

По умолчанию критериями для включения и исключения предикторов являются пороговые значения F -критерия. Как правило, значению $F = 3,84$ соответствует величина уровня значимости p , приблизительно равная 0,05, а значению $F = 2,71$ — величина p , приблизительно равная 0,1. Исследователи нередко предпочитают использовать значения по умолчанию, нежели задавать собственные.

Флажок Отчет о шагах в группе Вывести управляет пошаговым выводом основных результатов дискриминантного анализа, поэтому по умолчанию он установлен. Флажок F для попарных расстояний относится только к методу Расстояние Махаланобиса.

Щелчок на кнопке Классифицировать в окне Дискриминантный анализ открывает диалоговое окно Дискриминантный анализ: Классификация, показанное на рис. 22.5. Большая часть элементов интерфейса этого окна активно используется исследователями и описана ниже.

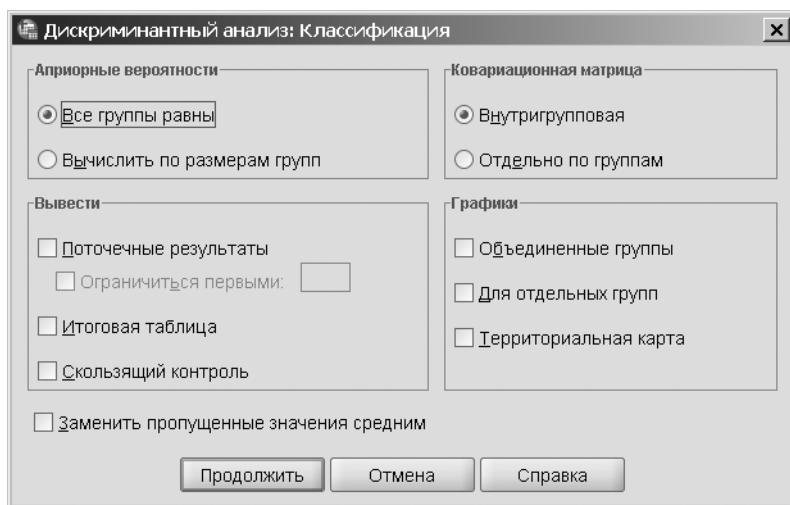


Рис. 22.5. Диалоговое окно Дискриминантный анализ: Классификация

В группе Априорные вероятности находятся два переключателя:

- Все группы равны — вероятности принадлежности объекта к каждой из групп полагаются равными;
- Вычислить по размерам групп — вероятности принадлежности объекта к каждой из групп пропорциональны размерам групп.

Флажки в группе Графики:

- ▶ Объединенные группы — в вывод включается гистограмма (если число групп равно 2) или диаграмма разброса (если число групп больше 2);
- ▶ Для отдельных групп — в вывод включаются несколько диаграмм, каждая из которых соответствует одной группе;
- ▶ Территориальная карта — этот флажок используется только в случаях, когда зависимая переменная имеет три или более уровней.

Два переключателя в группе Ковариационная матрица позволяют выбрать ковариационную матрицу:

- ▶ Внутригрупповая — этот переключатель установлен по умолчанию и классифицирует объекты при помощи общей внутригрупповой ковариационной матрицы;
- ▶ Отдельно по группам — объекты классифицируются при помощи отдельных ковариационных матриц для каждой группы.

Флажки в группе Вывести относятся к отображению результатов анализа:

- ▶ Поточечные результаты — этот флажок полезно устанавливать, если файл данных не слишком велик: в результат включается список объектов, и для каждого объекта указывается фактическая группа, прогнозируемая группа, вероятность попадания в прогнозируемую группу и значения дискриминантных функций (подобный список приведен в разделе «Представление результатов»);
- ▶ Итоговая таблица — этот флажок позволяет включить в результат число и процент корректных и некорректных классификаций для каждой группы (в разделе «Представление результатов» имеются соответствующие результаты);
- ▶ Скользящий контроль — при установке этого флажка каждый объект классифицируется с помощью функций, полученных по всем остальным наблюдениям, кроме данного. Этот метод также известен как «U-метод».

Название флажка Заменить пропущенные значения средним вполне отражает его назначение: при его установке пропущенное значение переменной заменяется ее средним (только на этапе классификации).

В этой главе приведена единственная версия шага 5. Разумеется, вы можете провести дискриминантный анализ с параметрами по умолчанию, однако значительно интереснее провести его с применением некоторых из параметров, описанных в этой главе.

В следующем примере проводится дискриминантный анализ для зависимой переменной оценка, имеющей два уровня, и 13 предикторов. Предикторы добавляются в дискриминантное уравнение пошаговым методом (Уилкса) с установками, отличающимися от предлагаемых по умолчанию: для включения предикторов в уравнение $F = 1,125$, а для исключения — значение $F = 1$. Для анализа зависимости между предикторами вычисляются все описательные статистики. Кроме того, мы

включаем в окно вывода нестандартные коэффициенты дискриминантного уравнения, результаты для каждого объекта и итоговую таблицу.

ШАГ 5

После выполнения предыдущего шага у вас должно быть открыто диалоговое окно Дискриминантный анализ, показанное на рис. 22.1. Если вы уже успели поработать с этим окном, щелкните на кнопке Сброс.

1. Щелкните сначала на переменной оценка, чтобы выделить ее, а затем — на верхней кнопке со стрелкой, чтобы переместить переменную в поле Группировать по.
2. Щелкните на кнопке Задать диапазон, чтобы открыть диалоговое окно Дискриминантный анализ: Задать диапазон, показанное на рис. 22.2.
3. В поле Минимум введите значение 1, нажмите клавишу Tab, чтобы переместить фокус ввода в поле Максимум, введите значение 2 и щелкните на кнопке Продолжит), чтобы вернуться в диалоговое окно Дискриминантный анализ.
4. Выделите переменную и1, нажмите на клавиатуре клавишу Shift, и, не отпуская ее, щелкните левой кнопкой мыши по переменной n. В результате окажутся выделенными все переменные от и1 до n.
5. Щелкните на средней кнопке со стрелкой, чтобы переместить выделенные переменные в список Независимые, установите переключатель Шаговый отбор и щелкните на кнопке Статистики, чтобы открыть диалоговое окно Дискриминантный анализ: Статистики, показанное на рис. 22.3.
6. Установите флажки Средние, Однофакторный дисперсионный анализ, М Бокса, Нестандартизованные и щелкните на кнопке Продолжить, чтобы вернуться в диалоговое окно Дискриминантный анализ.
7. Щелкните на кнопке Метод, чтобы открыть диалоговое окно Дискриминантный анализ: Шаговый отбор, показанное на рис. 22.4.
8. В окне Использовать F-значение: Включение введите значение 1,25, нажмите клавишу Tab, чтобы переместить фокус ввода в поле Исключение, введите значение 1 и щелкните на кнопке Продолжить, чтобы вернуться в диалоговое окно Дискриминантный анализ.
9. Щелкните на кнопке Классифицировать, чтобы открыть диалоговое окно Дискриминантный анализ: Классификация, показанное на рис. 22.5.
10. Установите флажки Поточечные результаты, Итоговая таблица и щелкните на кнопке Продолжить, чтобы вернуться в диалоговое окно Дискриминантный анализ.
11. Щелкните на кнопке ОК, чтобы открыть окно вывода.

После выполнения шага 5 программа автоматически откроет окно вывода. Для просмотра результатов при необходимости можно воспользоваться вертикальной и горизонтальной полосами прокрутки. Обратите внимание на стандартную строку меню в верхней части окна вывода: ее присутствие позволяет выполнять любые статистические операции, не переключаясь обратно в окно редактора данных.

Представление результатов

Далее приведены результаты, сгенерированные программой при выполнении шага 5. Как всегда, мы постарались представить данные, выведенные SPSS, в максимально удобной и понятной форме. Из весьма объемных результатов вывода выбраны наиболее важные фрагменты, иллюстрирующие суть дискриминантного анализа. Для удобства представления результатов форматы выводимых таблиц, приведенных на рисунках, изменены без ущерба для их содержания.

В начале выводимых результатов следует таблица Групповые статистики. В ней для каждой переменной показаны средние значения и стандартные отклонения на обоих уровнях (в обеих категориях) зависимой переменной, а также общие средние значения.

После таблицы Групповые статистики следует таблица Критерии равенства групповых средних, фрагмент которой приведен на рис. 22.6. В ней содержатся коэффициенты Лямбда Уилкса, F -критерии (строка F) и уровни значимости, характеризующие различия средних значений для групп по каждой из переменных (Знч.). Наиболее важная для исследователя информация относится к величинам F -критерия и уровням значимости, поскольку именно по ним можно судить, для каких переменных различие двух групп является значимым.

Критерий равенства групповых средних					
	Лямбда Уилкса	F	ст.св1	ст.св2	Знч.
осведомленность	,901	4,859	1	44	,033
скрытые фигуры	,850	7,746	1	44	,008
пропущенные слова	,931	3,271	1	44	,077
счет в уме	,806	10,563	1	44	,002

Рис. 22.6. Коэффициенты лямбда Уилкса, F -критерии и уровни значимости

Трактовка терминов приведена в конце этой главы.

На рис. 22.7 представлены фрагменты выводимых результатов, связанные с пошаговым дискриминантным анализом. В трех таблицах содержится статистическая информация о переменных, вошедших и не вошедших в дискриминантное уравнение, а также порядок включения и исключения переменных в процессе составления дискриминантного уравнения. Первая таблица иллюстрирует пошаговый процесс составления дискриминантного уравнения. В него поочередно вводятся предикторы на основе заданного критерия включения (по умолчанию критерием является $F \geq 3,84$, в нашем случае — $F \geq 1,25$), а также исключаются из уравнения те предикторы, которые удовлетворяют критерию исключения (по умолчанию таким критерием является $F \leq 2,71$, в нашем случае — $F \leq 1$). Представленные таблицы являются завершающими фрагментами очень больших таблиц, которые генерирует программа в результате выполнения семи шагов заданного пошагового метода.

Обратите внимание, что все вошедшие в уравнение переменные после шага 7 имели достаточный уровень толерантности (выше 0,1) и значения F -критерия, превышающие пороговое значение 1,125. Переменные, не попавшие в дискриминантное

Введенные/исключенные переменные

Шаг	Введенные	Лямбда Уилкса		
		Статистика	Точное значение F	
			Статистика	Знач.
1	счет в уме	,806	10,562	,002
2	заучивание слов	,708	8,849	,001
3	скрытые фигуры	,639	7,897	,000
4	умозаключения	,606	6,675	,000
5	анalogии	,570	6,043	,000
6	понятливость	,539	5,566	,000
7	экстраверсия	,514	5,125	,000

Переменные в анализе

Шаг		Толерантность	F исключения	Лямбда Уилкса
7	счет в уме	,677	6,720	,605
	заучивание слов	,932	2,073	,542
	скрытые фигуры	,889	3,012	,555
	умозаключения	,549	5,832	,593
	анalogии	,600	3,006	,555
	понятливость	,855	2,525	,549
	экстраверсия	,886	1,797	,539

Переменные, не включенные в анализ.

Шаг		Толерантность	Мин. Толерантность	F включения	Лямбда Уилкса
7	осведомленность	,768	,537	,064	,513
	пропущенные слова	,662	,509	,014	,514
	исключ. изобр.	,790	,542	,016	,514
	числовые ряды	,555	,522	,096	,513
	геометр. слож.	,519	,519	,000	,514
	нейротизм	,915	,546	1,123	,499

Рис. 22.7. Пошаговое включение и исключение переменных (фрагмент окна вывода)

уравнение, также имеют достаточную толерантность, однако значение F -критерия у них оказалось меньше 1,125. В окне вывода можно видеть, что шаги 1–7 приводят к последовательному введению в уравнение новых предикторов. Обращает на себя внимание тот факт, что в окончательный результат не вошли некоторые переменные, различия между группами для которых статистически достоверны (см. рис. 22.6). Напротив, в дискриминантное уравнение были включены переменные экстраверсия и умозаключения, различия между группами по которым статистически недостоверны. Это связано с тем, что при включении переменных в дискриминантное уравнение учитывается не только дискриминативная способность каждой переменной в отдельности, но и ее уникальный вклад в совокупности с остальными переменными.

На рис. 22.8–22.10 приведены очередные таблицы, содержащие основные показатели канонической дискриминантной функции. В таблице на рис. 22.8 в графе Функция значение 1 говорит о том, что в процессе дискриминантного анализа была получена одна дискриминантная функция. Если бы зависимая переменная имела не 2, а 3 уровня, то было бы составлено две дискриминантные функции. Чем больше значение Хи-квадрат, тем сильнее дискриминантная функция различает группы

и тем лучше она соответствует своему назначению. О ее состоятельности свидетельствует статистическая значимость Знч., заметно меньшая 0,05.

Собственные значения и Лямбда Уилкса

Функция	Собственное значение	% объясненной дисперсии	Кумулятивный %	Каноническая корреляция	Лямбда Уилкса	Хи-квадрат	ст. св.	Знч.
1	,944	100,0	100,0	,697	,514	26,926	7	,000

Рис. 22.8. Основные статистики канонической дискриминантной функции

На рис. 22.9 приведена таблица Коэффициенты канонической дискриминантной функции — список нестандартизованных коэффициентов и константа дискриминантного уравнения. Это уравнение подобно линейному уравнению множественной регрессии и применяется для предсказания. Значение функции для каждого объекта подсчитывается по этому уравнению.

Коэффициенты канонической дискриминантной функции

Функция	скрытые фигуры	счет в уме	понятливость	анalogии	умозаключения	заучивание слов	экстраверсия	(константа)
1	,162	,382	,214	,185	-,241	,157	,097	-9,865

Ненормированные коэффициенты

Рис. 22.9. Коэффициенты канонической дискриминантной функции

Структурная матрица (рис. 22.10) содержит корреляции между дискриминантной функцией и каждой из переменных. Переменные упорядочены по абсолютной величине корреляций.

Структурная матрица

Функция	счет в уме	понятливость	скрытые фигуры	анalogии	заучивание слов	геометрическое сложение	осведомленность ^а	числовые ряды ^а	пропущенные слова ^а	исключение изображений ^а	нейротизм ^а	экстраверсия	умозаключения
1	,504	,442	,432	,405	,328	,326	,290	,259	,258	,245	,181	,123	,063

Объединенные внутригрупповые корреляции между дискриминантными переменными и нормированными каноническими дискриминантными функциями.

Переменные упорядочены по абсолютной величине корреляций внутри функции.

а. Эта переменная не используется в анализе.

Рис. 22.10. Структурная матрица

На рис. 22.11 приведена таблица нормированных коэффициентов канонической дискриминантной функции. Эти коэффициенты служат для определения относительного вклада каждой переменной в значение дискриминантной функции, с учетом влияния остальных переменных. Чем больше абсолютное значение коэффициента, тем больше относительный вклад данной переменной в значение дискриминантной функции, разделяющей классы.

**Нормированные коэффициенты
канонической дискриминантной функции**

	Функция
	1
скрытые фигуры	,412
счет в уме	,676
понятливость	,387
анalogии	,501
умозаключения	-,706
заучивание слов	,338
экстраверсия	,324

Рис. 22.11. Нормированные коэффициенты дискриминантной функции

Таблица, содержащая координаты центроидов групп, приведена на рис. 22.12. Центроид представляет собой значение функции, получаемое при подстановке в дискриминантное уравнение средних значений предикторов. Обратите внимание, что центроиды равны по абсолютной величине, но имеют разные знаки. Граничным значением для двух групп является ноль.

Функции в центроидах групп

	Функция
	1
оценка успешности	
низкая	-,950
высокая	,950

Ненормированные канонические дискриминантные функции вычислены в центроидах групп.

Рис. 22.12. Координаты центроидов групп

Таблица, приведенная на рис. 22.13, содержит информацию о фактической и прогнозируемой группах для каждого объекта, вероятности его принадлежности к группе, а также значения (баллы) дискриминантной функции. В целях экономии места мы привели значения лишь для некоторых объектов. Для объектов, отмеченных двумя звездочками (**), фактическая и прогнозируемая группы не совпали. Всего таких объектов 5 из 46. В отношении последних 10 объектов, для которых принадлежность к группе не была известна, в таблице представлены результаты предсказания, полученные при помощи уравнения дискриминантной функции.

Поточечные статистики

Номер наблюдения	Фактическая группа	Наивероятнейшая группа				Дискриминантные баллы	
		Предсказанная группа	P(D>d G=g)		P(G=g D=d)		
			p	ст. св		Функция 1	
Исходные	1	1	,994	1	,861	-,958	
	2	2	,853	1	,896	1,135	
	
	6	2	1**	,507	1	,633	-,287
	
	47	несгрупп.	2	,997	1	,858	,946
	
	

** - Неправильно классифицированное наблюдение

Рис. 22.13. Фрагмент таблицы Поточечные статистики

Как показывают результаты классификации (рис. 22.14), при данном наборе дискриминантных переменных точность классификации составляет 89,13 % (41 из 46 правильных предсказаний в отношении «известных» объектов).

Результаты классификации^а

оценка успешности			Предсказанная принадлежность к группе		Итого
			низкая	высокая	
Исходные	Частота	низкая	22	1	23
		высокая	4	19	23
		Несгруппированные наблюдения	7	3	10
	%	низкая	95,7	4,3	100,0
		высокая	17,4	82,6	100,0
		Несгруппированные наблюдения	70,0	30,0	100,0

а. 89,1% исходных сгруппированных наблюдений классифицировано правильно.

Рис. 22.14. Результаты классификации

Терминология, используемая при выводе

Трактовка терминов, используемых программой в окне вывода и относящихся к средним значениям, λ Уилкса, F -критериям и уровням значимости (см. рис. 22.6).

- Среднее — средние значения для каждой категории (низкая, высокая) и для всех категорий Итого переменной оценка.
- Лябда Уилкса — отношение внутригрупповой суммы квадратов к общей сумме квадратов (λ). Данный коэффициент характеризует долю дисперсии оценок дискриминантной функции, которая не обусловлена различиями между группами, принимает значение 1 в случае, если средние значения для всех групп оказываются равными, и уменьшается с ростом разностей средних значений.

- ▶ F — значения F -критерия такие же, как при однофакторном дисперсионном анализе, и равны квадрату t -критериев двух выборок.
- ▶ Знч. — уровни значимости F -критериев, равные вероятности того, что соответствующие различия являются случайными.

Трактовка терминов, используемых программой в окне вывода и относящихся к пошаговому составлению дискриминантного уравнения (см. рис. 22.7).

- ▶ F включения — минимальное значение F , при котором предиктор включается в дискриминантное уравнение.
- ▶ F исключения — максимальное значение F , при котором предиктор исключается из дискриминантного уравнения.
- ▶ Толерантность — мера линейной зависимости между одним предиктором и всеми остальными. Если величина толерантности окажется меньше 0,001, SPSS воспримет такой результат как наличие значительной линейной зависимости и не включит соответствующий предиктор в дискриминантное уравнение.
- ▶ Знч. — значимости влияния данного предиктора на дисперсию зависимой переменной.

Трактовка терминов, используемых программой в окне вывода и относящихся к канонической дискриминантной функции, ее коэффициентам и центроидам групп (рис. 22.8–22.13).

- ▶ Коэффициенты канонической дискриминантной функции — список нестандартизованных коэффициентов и константа дискриминантного уравнения. Это уравнение подобно линейному уравнению множественной регрессии. Значение функции для каждого объекта подсчитывается по этому уравнению.
- ▶ Функция — значение 1 в этой ячейке говорит о том, что в процессе дискриминантного анализа была получена одна дискриминантная функция. Если бы зависимая переменная имела не 2, а 3 уровня, то было бы составлено две дискриминантные функции.
- ▶ Лямбда Уилкса — отношение внутригрупповой суммы квадратов к общей сумме квадратов (λ).
- ▶ Хи-квадрат — мера статистического отличия друг от друга двух уровней дискриминанта (χ^2). Чем больше данное значение, тем сильнее отличие и тем лучше дискриминантная функция соответствует своему назначению.
- ▶ Собственное значение — отношение межгрупповой суммы квадратов к внутригрупповой сумме квадратов. Чем больше данное значение, тем лучше составленная функция для дискриминантного анализа.
- ▶ % объясненной дисперсии, Кумулятивный % — дискриминантная функция всегда вычисляется для равной 100 % дисперсии зависимой переменной.

- ▶ Структурная матрица — структурная матрица содержит корреляции между значениями дискриминантной функции и каждой из переменных. Переменные упорядочены по абсолютной величине корреляций.
- ▶ Центроиды групп — средние значения дискриминантной функции для двух групп. Более точно, центроид представляет собой значение функции, получаемое при подстановке в дискриминантное уравнение средних значений предикторов.

Трактовка терминов, используемых программой в таблице Поточечные статистики (см. рис. 22.13).

- ▶ Несгруппированный — объект, для которого заранее неизвестна принадлежность к группе.
- ▶ Фактическая группа — группа, к которой принадлежит данный объект.
- ▶ Предсказанная группа — группа, вычисленная для объекта с помощью уравнения дискриминантной функции.
- ▶ $P(D > d \mid G = g)$ — вероятность принадлежности объекта к группе (G) при данной величине дискриминантной функции (D).
- ▶ $P(G = g \mid D = d)$ — вероятность наблюдаемого значения дискриминантной функции (D), если задана принадлежность объекта к группе (G).
- ▶ Наивероятнейшая группа — группа, для которой прогнозируемая вероятность включения данного объекта максимальна.
- ▶ Вторая вероятнейшая группа — группа, имеющая вторую по величине вероятность (после прогнозируемой) включения данного объекта. Поскольку число групп в данном случае равно 2, то такая группа для каждого объекта определена «заранее».
- ▶ Дискриминантные баллы — дискриминантные оценки, получаемые при подстановке значений переменных объекта в уравнение дискриминантной функции.

Завершение анализа и выход из программы

Отредактируйте содержимое окна вывода в соответствии со своими предпочтениями: скройте лишнюю информацию, исправьте таблицы и пр. (см. раздел «Окно вывода и его редактирование» главы 2). За инструкциями по редактированию графиков обратитесь к главе 5.

Для дальнейшего использования окончательного результата все содержимое окна вывода или его фрагменты можно сохранить в файле *.spv, экспортировать в другой формат (например, Word), перенести в документ Word или вывести на печать (подробности см. в разделе «Сохранение, экспорт, перенос и печать результатов» главы 2).

Для выхода из программы выберите команду Выход в меню Файл.

23 Многомерное шкалирование

333	Квадратная асимметричная матрица различий
334	Квадратная симметричная матрица различий
335	Модели индивидуальных различий
335	Пошаговые алгоритмы вычислений
343	Представление результатов
346	Завершение анализа и выход из программы

Основное достоинство многомерного шкалирования — представление больших массивов данных о различии объектов в наглядном, доступном для интерпретации графическом виде. При многомерном шкалировании матрица различий между объектами (вычисленными, например, по их экспертным оценкам) представляется в виде одно-, двух- или трехмерного графического изображения взаимного расположения этих объектов. Хотя доступно и более трех измерений, эта возможность редко применяется на практике.

Основным преимуществом многомерного шкалирования является возможность очень наглядного визуального сравнения объектов анализа. Если две точки на изображении удалены друг от друга, то между соответствующими объектами имеется значительное расхождение; и наоборот, близость точек говорит о сходстве объектов.

Многомерное шкалирование имеет много общих черт с факторным анализом (см. главу 20). Так же как и при факторном анализе, создается система координат пространства, в котором определяется расположение точек. Так же как и при факторном анализе, происходит снижение размерности и упрощение данных. Однако при факторном анализе обычно используются коэффициенты корреляции, а при многомерном шкалировании — меры различия между объектами. Наконец, в факторном анализе наибольший интерес вызывают углы между точками, представляющими данные, а в многомерном шкалировании ключевой величиной является расстояние между этими точками.

Помимо факторного анализа многомерное шкалирование имеет несколько общих черт с кластерным анализом (см. главу 21). В обоих случаях анализируется расстояние между объектами; однако при кластерном анализе типичной является количественная процедура объединения объектов в группы (кластеры), а при многомерном шкалировании качественный анализ объектов проводится визуально с помощью диаграммы.

В этой книге мы рассмотрим наиболее известную процедуру многомерного шкалирования SPSS, имеющую название ALSCAL. Фактически ALSCAL не является одной программой, а представляет собой набор процедур, каждая из которых соответствует своему типу данных. В этой главе мы проведем несколько вариантов анализа для различных типов данных, стараясь, по возможности, коротко и ясно изложить смысл сопутствующих терминов. Если вы сталкиваетесь с многомерным шкалированием впервые, то рекомендуем обратиться к дополнительной литературе.

В первом примере мы обработаем гипотетическую социограмму для группы учащихся, при этом количественные оценки их взаимоотношений будут преобразованы в соответствующее графическое изображение. Во втором примере мы рассмотрим результаты тестирования учащихся по пяти показателям и покажем различия между ними графически на плоском изображении. Наконец, третий пример будет представлять собой небольшое исследование восприятия и понимания студентами пяти многомерных методов статистического анализа.

Квадратная асимметричная матрица различий

Представим себе, что преподаватель решил создать идеальную психологическую обстановку в группе во время занятия, рассадив учащихся так, чтобы ни один из них не оказался рядом с тем, кто ему не нравится. Для этого каждому из 12 студентов было предложено оценить степень своей симпатии к своим однокурсникам по пятибалльной шкале (от 1 до 5, где 1 — максимум симпатии, а 5 — максимум антипатии). Результаты этого вымышленного опроса мы поместили в файл данных `mds1.sav`. Чтобы добиться желаемого результата, преподавателю необходимо максимально далеко рассадить негативно настроенных в отношении друг друга учащихся. Здесь весьма полезной окажется диаграмма, на которой удаленность точек будет соответствовать отношениям между учащимися. Для построения диаграммы мы воспользуемся средствами многомерного шкалирования.

Первое, что необходимо сделать для решения задачи, — создать квадратную (12×12) матрицу различий. Позже на основе этой матрицы будет построено двумерное изображение, иллюстрирующее взаимоотношения студентов. В ходе многомерного шкалирования исходная матрица 12×12 преобразуется в гораздо более простую матрицу 12×2 (где 2 — количество измерений или шкал), содержащую координаты точек для изображения. Исходную матрицу называют *квадратной асимметричной матрицей различий*. Поясним, что означают составляющие это определение термины.

- *Квадратная матрица* — это матрица, строки и столбцы которой представляют один и тот же набор объектов. В данном случае этим набором объектов является группа учащихся: каждый учащийся оценивает меру своей симпатии-антипатии к каждому из остальных. Если бы строки и столбцы матрицы

представляли не одни и те же, а разные объекты, матрица бы называлась не квадратной, а прямоугольной. Прямоугольные матрицы в контексте данной книги не рассматриваются.

- *Асимметричная матрица* — это матрица, для которой отношение двух объектов друг к другу может быть разным. Так, например, симпатия Петра к Ирине не означает, что Ирине Петр тоже симпатичен. Визуально асимметричность матрицы выражается в том, что как минимум для одной пары ячеек, симметрично расположенных относительно главной диагонали матрицы, значения различны. Если же ни одного такого различия для матрицы не установлено, матрица называется симметричной. Примером матрицы, которая всегда симметрична, является корреляционная матрица.
- *Матрица различий* — матрица, данные которой представляют меру различия. В данном случае значения матрицы отражают степень отличия отношения одного студента к другому от идеального; чем больше значение, тем больше различие. Существуют также матрицы сходств, в которых значения отражают степень некоторого сходства. Если бы преподаватель вместо матрицы различий составил матрицу сходств, то изображение, полученное в результате многомерного шкалирования, скорее внесло бы путаницу, чем помогло в решении задачи.

Квадратная симметричная матрица различий

Пусть наш гипотетический преподаватель решает другую задачу: ему нужно рассадить 12 учащихся в соответствии с результатами их тестирования по пяти показателям. Поскольку результаты тестирования не относятся к данным, характеризующим различия, необходимо сначала вычислить различия по имеющимся данным и таким образом свести задачу к предыдущей. Исходные данные для этой задачи естественно представить в виде прямоугольной матрицы 12×5 , в которой для каждого из 12 учащихся указаны результаты 5-ти тестов (файл `mds2.sav`). Затем по исходным данным строится квадратная (12×12) симметричная матрица различий между учащимися. Наконец, как и в предыдущем примере, SPSS создает матрицу координат 12×2 и визуально представляет ее в виде диаграммы.

Обратите внимание на два ключевых свойства матрицы различий: она является *квадратной* и *симметричной*. Несмотря на то, что исходная матрица является прямоугольной, то есть ее строки (объекты) соответствуют учащимся, а столбцы (переменные) — тестам, в матрице различий как строки, так и столбцы соответствуют учащимся, и, следовательно, матрица является квадратной с размером 12×12 . Кроме того, поскольку, учащийся 1 отличается от учащегося 5 по результатам тестирования так же, как учащийся 5 от учащегося 1, матрица различий является симметричной.

Модель индивидуальных различий

В обоих рассмотренных примерах создавалась единственная матрица различий, на основе которой вычислялись координаты точек-объектов для визуального представления. Однако в некоторых случаях возможно координатное представление нескольких матриц различий; примером может служить ситуация, когда несколько экспертов оценивают один и тот же набор объектов. В таких случаях говорят о многомерном шкалировании индивидуальных различий.

Предположим, преподаватель предлагает студентам сравнить между собой 5 многомерных статистических методов: множественный регрессионный анализ (mra), факторный анализ (fa), кластерный анализ (ka), дискриминантный анализ (da) и многомерное шкалирование (mds). Каждому студенту предъявляются все возможные пары из этих методов для оценки различия между ними (1 — максимально сходны, 5 — максимально различны). Затем по данным для 6 студентов составляются 6 матриц различий (5×5 каждая).

Поскольку SPSS предъявляет определенные требования к формату данных при использовании модели индивидуальных различий, мы создали специальный файл mds3.sav. Требования к формату файла, по сути, сводятся к наличию в этом файле нескольких квадратных матриц одинакового размера. Поскольку в данном примере мы имеем дело с 6 студентами, оценивающими 5 объектов (многомерных методов), в файле mds3.sav присутствуют 6 матриц размером 5×5 . Первые 5 строк файла соответствуют первой матрице, следующие 5 строк — второй матрице и т. д. Всего под матрицы отведено 30 строк.

Пошаговые алгоритмы вычислений

Для проведения многомерного шкалирования сначала необходимо выполнить три подготовительных шага. Эти шаги (шаги 1–3) позволят подготовить рабочий файл данных, запустить программу IBM SPSS Statistics 19 и открыть файл (в данном случае — файл mds1.sav). Пошаговые инструкции этого процесса приведены в главе 4 (с. 60), а подробные разъяснения — в главе 2.

После завершения шага 3 на экране должно присутствовать окно редактора данных со строкой меню и загруженным файлом mds1.sav.

ШАГ 4

В меню Анализ выберите команду Шкалирование ► Многомерное шкалирование (ALSCAL). Откроется диалоговое окно Многомерное шкалирование, показанное на рис. 23.1.

В диалоговом окне Многомерное шкалирование вы можете задать список переменных, которые будут участвовать в анализе. В SPSS существует два типа многомерного шкалирования ALSCAL: с заданной матрицей различий и с вычислением матрицы различий.

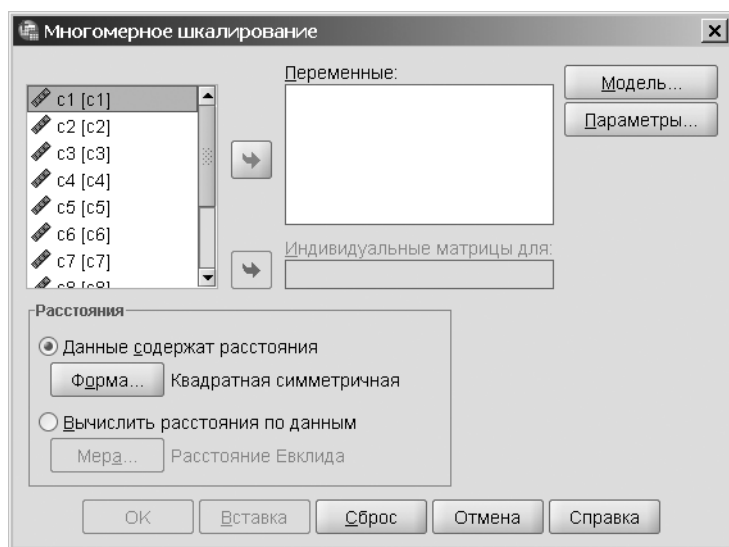


Рис. 23.1. Диалоговое окно Многомерное шкалирование

В том случае, когда матрица различий задана, вы при желании можете указать столбцы файла, которые будут участвовать в анализе. Обратите внимание на следующий нюанс: вы можете включить в процедуру многомерного шкалирования часть столбцов файла данных, однако не можете включить часть строк. Последнее достигается косвенно — при помощи описанной подробно в главе 4 команды «Отбавить наблюдения», однако вы не можете исключить «лишние» объекты непосредственно. Если используется квадратная матрица, число выбранных переменных (столбцов) должно совпадать с числом объектов файла (строк).

В случае если вы не используете матрицу различий из файла данных, а хотите, чтобы программа SPSS создала ее самостоятельно, требуется указать переменные, для которых будут вычислены расстояния между объектами. В обоих случаях выбор переменных сводится к одной и той же процедуре: вы выделяете нужную переменную (или несколько переменных) в исходном списке в левой части диалогового окна и щелчком на верхней кнопке со стрелкой помещаете ее (их) в список Переменные.

Если вы хотите, чтобы программа SPSS вычислила расстояния между объектами вашего файла, и одновременно намереваетесь использовать модель многомерного шкалирования для нескольких выборок, то можете задать имя нужной переменной в поле Отдельные матрицы для, однако эта процедура весьма трудоемка, и мы не будем на ней останавливаться.

После того как заданы переменные, которые должны быть включены в анализ, следует информировать программу SPSS о структуре вашего файла данных. Для этого предназначены два переключателя Расстояния. Если ваш файл уже содержит матрицу различий, следует оставить установленным переключатель Данные содержат расстояния. В случае если матрица с данными является квадратной и симметричной (значения над ее главной диагональю равны значениям под главной диа-

гональю), дополнительно определять тип данных не нужно. Если же вы намерены использовать асимметричную или прямоугольную матрицу, щелкните на кнопке Форма, и на экране появится диалоговое окно Многомерное шкалирование: Форма данных, показанное на рис. 23.2.

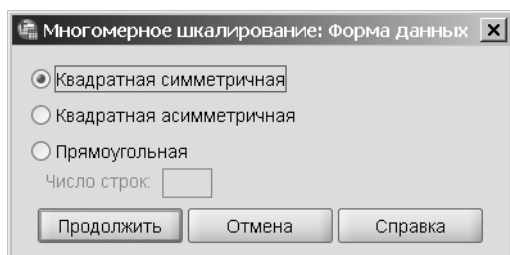


Рис. 23.2. Диалоговое окно Многомерное шкалирование: Форма данных

В этом окне вы можете выбрать один из возможных типов матрицы. Если ваша матрица является квадратной (число строк равно числу столбцов) и симметричной (значения над главной диагональю равны значениям под главной диагональю), то следует оставить установленным переключатель Квадратная симметричная. Этот тип матрицы используется в примере с многомерными методами, поскольку сравнения одного метода с другим, и наоборот, дают одинаковый результат. Если матрица является квадратной, но не симметричной, следует установить переключатель Квадратная асимметричная. Такой тип матрицы используется в примере с социограммой студентов. Наконец, если ваша матрица не является квадратной, установите переключатель Прямоугольная и введите число строк в расположенное рядом поле. Этот тип матрицы используется реже, например при шкалировании предпочтений, и в данной книге не описывается.

В случае если необходимо, чтобы программа SPSS создала матрицу различий по имеющимся данным, в окне Многомерное шкалирование следует установить переключатель Вычислить расстояния по данным. В ответ SPSS создаст матрицу различий для тех переменных, которые вы указали в списке Переменные. По умолчанию вычисляется Евклидово расстояние (по формуле Пифагора $a^2 + b^2 = c^2$), и если такой способ вычисления устраивает вас, никаких дополнительных действий с вашей стороны не требуется. Если же вы хотите применить иной способ вычисления расстояния, следует щелкнуть на кнопке Мера, открыв диалоговое окно Многомерное шкалирование: Создание меры по данным, показанное на рис. 23.3.

Как можно видеть, диалоговое окно содержит три области: Мера, Преобразовать значения и Создать матрицу расстояний. Помимо формулы для Евклидова расстояния существуют еще несколько довольно распространенных формул, например формула Минковского, которая напоминает формулу Пифагора, однако позволяет вместо квадратов использовать любую степень, например четвертую: $a^4 + b^4 = c^4$. Кроме этого, полезным способом определения расстояния является метрика города. Для того чтобы понять ее суть, представьте себе городской район, в котором вы можете передвигаться только по улицам, огибая дома. Метрика города применяется при использовании разнородных по смыслу и (или) ранговых переменных.

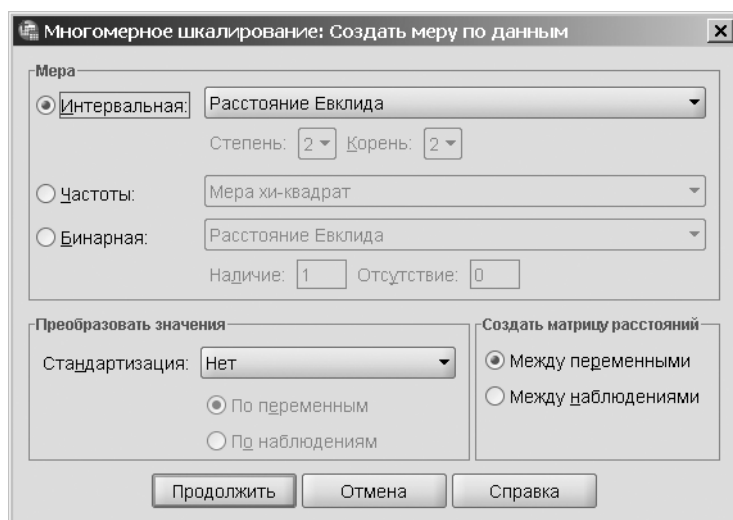


Рис. 23.3. Диалоговое окно Многомерное шкалирование: Создать меру по данным

Если все переменные рабочего файла имеют одну и ту же шкалу (например, целые числа от 1 до 5), вероятно, не придется вместо пункта Нет, выбранного в списке Стандартизация по умолчанию, выбирать что-то другое. Если же для переменных применяются различные шкалы, то может понадобиться какой-либо из вариантов стандартизации, предлагаемых SPSS. Чаще всего используется вариант z-значения. И наконец, по умолчанию SPSS вычисляет расстояния между переменными, однако при желании можно потребовать вычислить расстояние между объектами. Для этого в группе Создание матрицы расстояния необходимо вместо переключателя Между переменными установить переключатель Между наблюдениями.

После задания режима вычисления расстояний можно выбрать желаемый тип модели многомерного шкалирования. Для этого в окне Многомерное шкалирование щелкните на кнопке Модель, и на экране появится диалоговое окно Многомерное шкалирование: Модель, показанное на рис. 23.4. Наибольший интерес в этом окне представляют области Шкала измерения и Модель шкалирования. В отличие от факторного анализа (см. главу 20), при многомерном шкалировании программа не может автоматически определить количество шкал, которое следует использовать. Поскольку в подавляющем большинстве случаев требуется не более трех шкал, выбор нужного числа измерений не представляет трудностей: вы легко можете перебрать все варианты представления и остановиться на том, который наиболее удобен для зрительного восприятия. Чаще всего при многомерном шкалировании применяют модель Евклидовых расстояний; в ней вам необходимо задать либо одну матрицу различий либо несколько идентичных матриц для анализа. Однако если матрицы различий получены от разных субъектов, которых вы хотите сравнить между собой, вам придется использовать модель индивидуальных различий. Примером служит ситуация с упоминавшимися ранее многомерными методами: каждый из шести студентов сравнивает 5 методов, оценивая различия между ними.

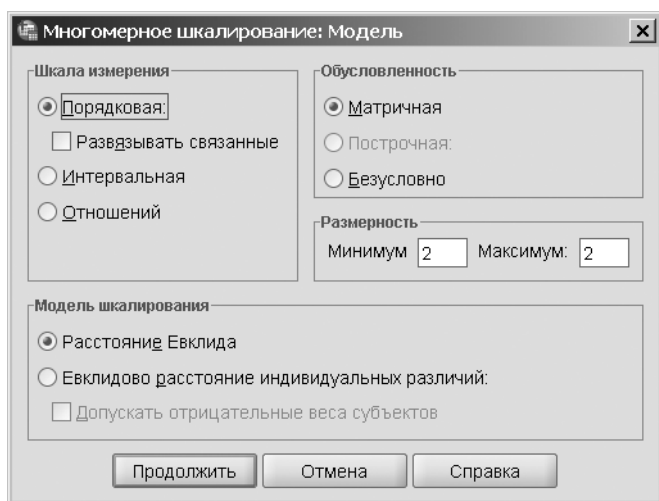


Рис. 23.4. Диалоговое окно Многомерное шкалирование: Модель

Группа переключателей **Обусловленность** позволяет выбрать, какие условия сравнения следует выполнять: если установлен переключатель **Матричная**, то все ячейки каждой матрицы сравниваются между собой (эта ситуация соответствует примеру с методами, когда сравниваются различия, оцененные одним студентом, но не сравниваются различия из разных таблиц). Установка переключателя **Построчная** означает, что сравнения производятся между ячейками одной строки. Этот переключатель следует установить в примере с социограммой, поскольку в нем каждая строка таблицы различий соответствует субъекту, оценивающему свое отношение к однокурсникам; в общем случае учащиеся могут использовать разные шкалы, однако сравнения в рамках одной строки имеют смысл. Наконец, переключатель **Безусловно** означает, что в программе предполагаются сравнения всех ячеек по всем матрицам.

Группа переключателей **Шкала измерения** позволяет указать программе, в какой шкале интерпретировать величины матрицы различий. Как правило, между переключателями **Порядковая**, **Интервальная** и **Отношений** нет существенных различий.

Последним рассматриваемым элементом интерфейса диалогового окна **Многомерное шкалирование** является кнопка **Параметры**. При щелчке на ней открывается диалоговое окно **Многомерное шкалирование: Параметры**, показанное на рис. 23.5. Наибольший интерес здесь представляет группа флажков **Вывести**. Флажок **Групповые графики** практически всегда следует устанавливать, поскольку в противном случае вам придется рисовать изображение вручную. Флажок **Матрица данных** позволяет выводить на экран исходные данные; эта возможность оказывается весьма полезной в случае, если вы используете SPSS для создания матрицы различий. Флажок **Сводка по модели и параметрам** позволяет включить в выводимые данные тип и параметры выбранной модели. Остальные элементы интерфейса диалогового окна **Многомерное шкалирование: Параметры** используются очень редко.

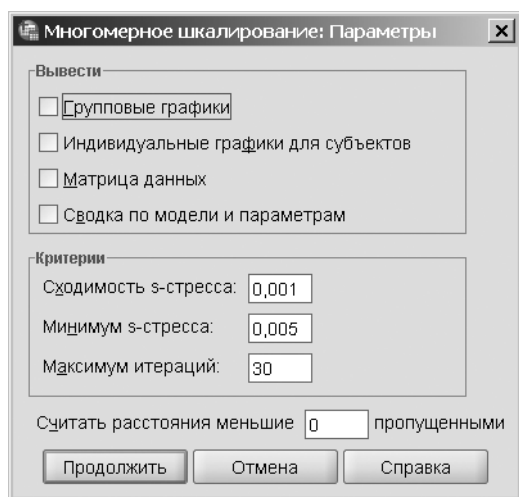


Рис. 23.5. Диалоговое окно Многомерное шкалирование: Параметры

Следующий пример иллюстрирует практическую реализацию двухмерного шкалирования для заданной матрицы различий. Подробное описание задачи приведено в разделе «Квадратная асимметричная матрица различий».

ШАГ 5

После выполнения предыдущего шага у вас должно быть открыто диалоговое окно Многомерное шкалирование, представленное на рис. 23.1. Если вы уже успели поработать с этим окном, щелкните на кнопке Сброс.

1. Щелкните на переменной s_1 , нажмите клавишу Shift и, не отпуская ее, щелкните на переменной s_{12} . В результате окажутся выделенными все промежуточные переменные, начиная от переменной s_1 и заканчивая переменной s_{12} .
2. Щелкните на верхней кнопке со стрелкой, чтобы переместить выделенные переменные в список Переменные.
3. Щелкните на кнопке Форма, чтобы открыть диалоговое окно Многомерное шкалирование: Форма данных, показанное на рис. 23.2.
4. Установите переключатель Квадратная асимметричная и щелкните на кнопке Продолжить, чтобы вернуться в диалоговое окно Многомерное шкалирование.
5. Щелкните на кнопке Модель, чтобы открыть диалоговое окно Многомерное шкалирование: Модель, показанное на рис. 23.4.
6. В группе Обусловленность установите переключатель Построчная и щелкните на кнопке Продолжить, чтобы вернуться в диалоговое окно Многомерное шкалирование.
7. Щелкните на кнопке Параметры, чтобы открыть диалоговое окно Многомерное шкалирование: Параметры, показанное на рис. 23.5.

8. Установите флажок Групповые графики и щелкните на кнопке Продолжить, чтобы вернуться в диалоговое окно Многомерное шкалирование.
9. Щелкните на кнопке ОК, чтобы открыть окно вывода.

В следующем примере демонстрируется двумерное шкалирование квадратной симметричной матрицы различий, которую SPSS создает при задании переменных из файла данных. Данные матрицы различий имеют интервальный тип. Подробное описание задачи было приведено в разделе «Квадратная симметричная матрица различий». В этом примере используется файл данных mds2.sav.

ШАГ 3А

Откройте файл данных, с которым вы намерены работать (в нашем случае — это файл mds2.sav). Если он расположен в текущей папке, то выполните следующие действия.

1. Выберите в меню Файл команду Открыть ► Данные.
2. В открывшемся диалоговом окне дважды щелкните на имени mds2.sav или введите его с клавиатуры и щелкните на кнопке ОК.

ШАГ 4А

В меню Анализ выберите команду Шкалирование ► Многомерное шкалирование (ALSCAL). Откроется диалоговое окно Многомерное шкалирование, аналогичное показанному на рис. 23.1.

ШАГ 5А

После выполнения шага 4а у вас должно быть открыто диалоговое окно Многомерное шкалирование. Если вы уже успели поработать с этим окном, щелкните на кнопке Сброс.

1. Щелкните на переменной тест1, нажмите клавишу Shift и, не отпуская ее, щелкните на переменной тест5. В результате окажутся выделенными все промежуточные переменные, начиная от переменной тест1 и заканчивая переменной тест5.
2. Щелкните на верхней кнопке со стрелкой, чтобы переместить выделенные переменные в список Переменные.
3. В группе Расстояния установите переключатель Вычислить расстояния по данным и щелкните на кнопке Мера, чтобы открыть диалоговое окно Многомерное шкалирование: Создать меру по данным, показанное на рис. 23.3.
4. В группе Создать матрицу расстояний установите переключатель Между наблюдениями и щелкните на кнопке Продолжить, чтобы вернуться в диалоговое окно Многомерное шкалирование.
5. Щелкните на кнопке Модель, чтобы открыть диалоговое окно Многомерное шкалирование: Модель, показанное на рис. 23.4.
6. В группе Шкала измерения установите переключатель Интервальная и щелкните на кнопке Продолжить, чтобы вернуться в диалоговое окно Многомерное шкалирование.

- Щелкните на кнопке Параметры, чтобы открыть диалоговое окно Многомерное шкалирование: Параметры, показанное на рис. 23.5.
- Установите флажок Групповые графики и щелкните на кнопке Продолжить, чтобы вернуться в диалоговое окно Многомерное шкалирование.
- Щелкните на кнопке ОК, чтобы открыть окно вывода.

Последним мы рассмотрим пример двумерного шкалирования с использованием нескольких квадратных симметричных матриц и модели индивидуальных различий. Подробное описание задачи было приведено в разделе «Модель индивидуальных различий» этой главы. Обратите внимание, что для выполнения анализа используется файл `mds3.sav`.

ШАГ 3Б

Откройте файл данных, с которым вы намерены работать (в нашем случае — это файл `mds3.sav`). Если он расположен в текущей папке, то выполните следующие действия.

- Выберите в меню Файл команду Открыть ► Данные.
- В открывшемся диалоговом окне дважды щелкните на имени `mds3.sav` или введите его с клавиатуры и щелкните на кнопке ОК.

ШАГ 4Б

В меню Анализ выберите команду Шкалирование ► Многомерное шкалирование (ALSCAL). Откроется диалоговое окно Многомерное шкалирование, аналогичное показанному на рис. 23.1.

ШАГ 5Б

После выполнения предыдущего шага у вас должно быть открыто диалоговое окно Многомерное шкалирование. Если вы уже успели поработать с этим окном, щелкните на кнопке Сброс.

- Щелкните на переменной `tra`, нажмите клавишу Shift и, не отпуская ее, щелкните на переменной `mds`. В результате окажутся выделенными все промежуточные переменные, начиная от переменной `tra` и заканчивая переменной `mds`.
- Щелкните на верхней кнопке со стрелкой, чтобы переместить выделенные переменные в список Переменные.
- Щелкните на кнопке Модель, чтобы открыть диалоговое окно Многомерное шкалирование: Модель, показанное на рис. 23.4.
- В группе Шкала измерения установите переключатель Отношения, а в группе Модель шкалирования — переключатель Евклидово расстояние индивидуальных различий, после чего щелкните на кнопке Продолжить, чтобы вернуться в диалоговое окно Многомерное шкалирование.
- Щелкните на кнопке Параметры, чтобы открыть диалоговое окно Многомерное шкалирование: Параметры, показанное на рис. 23.5.
- Установите флажок Групповые графики и щелкните на кнопке Продолжить, чтобы вернуться в диалоговое окно Многомерное шкалирование.
- Щелкните на кнопке ОК, чтобы открыть окно вывода.

После выполнения шага 5 программа автоматически откроет окно вывода. Для просмотра результатов при необходимости можно воспользоваться вертикальной и горизонтальной полосами прокрутки. Обратите внимание на стандартную строку меню в верхней части окна вывода: ее присутствие позволяет выполнять любые статистические операции, не переключаясь обратно в окно редактора данных.

Представление результатов

В этом разделе приведены данные, сгенерированные командой Многомерное шкалирование (ALSCAL) при выполнении шагов 5, 5а и 5б.

Итерации

Значения, записанные в столбце S-stress, характеризуют отклонение результата от идеального (точно соответствующего матрице отличий) на различных итерациях применения модели (рис. 23.6). SPSS применяет заданную модель столько раз, сколько необходимо для получения достаточно низкого значения в столбце S-stress. Если число итераций оказывается больше 30, то это, как правило, указывает на проблемы в исходных данных.

Iteration history for the 2 dimensional solution (in squared distances)

Young's S-stress formula 1 is used.

Iteration	S-stress	Improvement
1	.36850	
2	.34885	.01965
3	.33645	.01240
4	.32766	.00879
5	.32256	.00509
6	.32076	.00181
7	.32003	.00072

Iterations stopped because
S-stress improvement is less than .001000

Рис. 23.6. Фрагмент окна вывода после выполнения шага 5 (итерации)

Стрессы и квадраты коэффициентов корреляции

Для каждой строки асимметричной матрицы различий, для каждой матрицы модели индивидуальных различий, а также для всей модели при многомерном шкалировании вычисляются стресс и коэффициент R^2 (рис. 23.7). Стресс по своему смыслу схож со стрессом предыдущей модели, однако для его расчета используется другое уравнение, позволяющее упростить вычислительный процесс сравнения различий. Коэффициент R^2 (столбец RSQ) характеризует долю дисперсии в матрице различий, обусловленную данной моделью. Чем лучше модель, тем выше значение коэффициента R^2 . Если вы, к примеру, строите несколько графических изображений с разным числом шкал-координат, то величины стресса и R^2 могут служить критериями при выборе наиболее подходящей модели.

Stress and squared correlation (RSQ) in distances					
		Matrix 1			
Stimulus	Stress	RSQ	Stimulus	Stress	RSQ
1	.225	.826	2	.293	.690
3	.222	.846	4	.360	.380
5	.239	.750	6	.188	.839
7	.206	.796	8	.256	.713
9	.249	.716	10	.344	.607
11	.250	.722	12	.162	.880
Averaged (rms) over stimuli					
Stress = .256		RSQ = .730			

Рис. 23.7. Фрагмент окна вывода после выполнения шага 5 (стрессы и квадраты коэффициентов корреляции)

Координаты стимулов

Для каждого шкалируемого объекта указываются его координаты по каждой шкале (рис. 23.8). Это сделано для того, чтобы вы могли на основе этих координат построить собственное графическое изображение или использовать координаты для дальнейшего анализа. В данном случае столбец 1 соответствует координате x , а столбец 2 — координате y .

Stimulus Coordinates			
		Dimension	
Stimulus Number	Stimulus Name	1	2
1	C1	.9541	1.1290
2	C2	1.3036	.2898
3	C3	1.1314	.8383
4	C4	-.7832	.7595
5	C5	-1.6958	.2091
6	C6	-1.6310	-.0369
7	C7	-.3179	-1.5237
8	C8	.0372	-1.6126
9	C9	-1.0067	.7802
10	C10	.8591	.7622
11	C11	.8739	-.0226
12	C12	.2754	-1.5722

Рис. 23.8. Фрагмент окна вывода после выполнения шага 5 (координаты стимулов)

Субъективные веса

Если используется модель индивидуальных различий, то SPSS включает в окно вывода подсчитанные значения весов для каждого субъекта (рис. 23.9). В примере со сравнением многомерных методов каждый из 6-ти студентов по каждой из двух шкал имеет свой вес, который характеризует, какая доля дисперсии субъективных оценок приходится на соответствующую шкалу. Кроме того, в вывод включен общий вес для каждой из шкал. Величина Weirndness (предсказуемость) характеризует разброс весов для каждого субъекта; для субъекта с номером 3 она имеет максимальное значение, равное 0,51.

Subject Weights			
Subject Number	Weird- ness	Dimension	
		1	2
1	.1841	.7054	.6774
2	.1869	.7075	.6825
3	.5110	.8968	.2596
4	.1161	.7555	.6500
5	.1431	.7258	.6521
6	.2415	.8951	.4346
Overall importance of each dimension:		.6169	.3381

Рис. 23.9. Фрагмент окна вывода после выполнения шага 5б (субъективные веса)

Итоговая конфигурация объектов

Диаграмма, показанная на рис. 23.10, представляет собой итог применения модели многомерного шкалирования (в результате выполнения шага 5). Она отображает взаимоотношения 12-ти студентов таким образом, что чем больше различия между учащимися в исходной матрице, тем дальше они находятся друг от друга на диаграмме. На ней видно, что в исследуемой группе выделяются три относительно компактные подгруппы, самая крупная из которых состоит из пяти человек и располагается в правом верхнем углу диаграммы. Отношения внутри каждой из группировок характеризуются симпатией (точки расположены близко), чего не скажешь об отношениях между группами. В данном случае смысл каждой из шкал не имеет значения; главным является взаимное расположение точек.

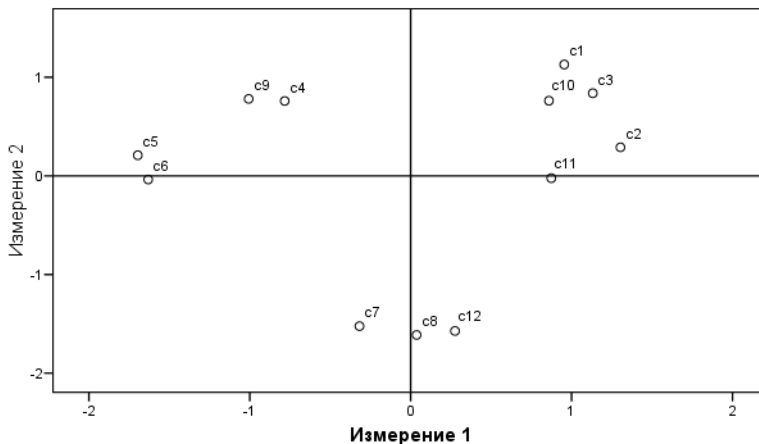


Рис. 23.10. Фрагмент окна вывода после выполнения шага 5 (конфигурация стимулов)

На диаграмме, приведенной на рис. 23.11, изображена итоговая конфигурация сравниваемых студентами статистических методов (результат выполнения шага 5б). В этом случае можно попытаться дать содержательную интерпретацию выделенных шкал как критериев, которыми пользуются студенты при сравнении методов.

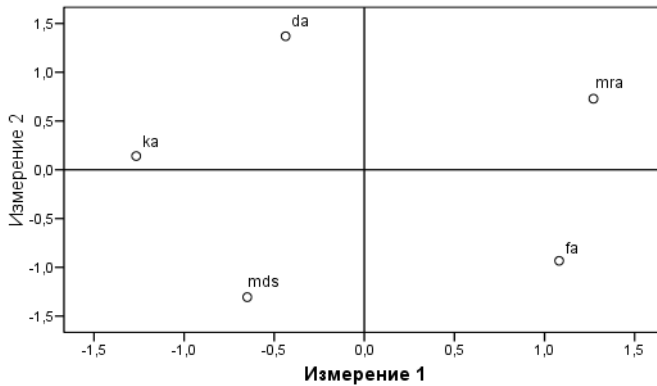


Рис. 23.11. Фрагмент окна вывода после выполнения шага 5б (итоговая диаграмма)

На правом полюсе шкалы 1 расположились множественный регрессионный (mra) и факторный анализы (fa), а на противоположном полюсе этой шкалы — кластерный анализ (ka) и многомерное шкалирование (mds). Эта шкала, таким образом, различает методы по предмету анализа: правый полюс шкалы соответствует корреляциям (факторный и регрессионный анализ), левый — различиям и расстояниям (шкалирование и кластерный анализ). На положительном полюсе шкалы 2 оказались дискриминантный (da) и множественный регрессионный анализы (mra), а на отрицательном полюсе этой шкалы — многомерное шкалирование (mds) и факторный анализ (fa). По всей видимости, эта шкала разделяет методы по их назначению: положительный полюс соответствует методам предсказания, а отрицательный — структурным методам. О значении этих шкал как субъективных критериев, которыми пользовались студенты при сравнении методов, можно судить по субъективным весам (см. рис. 23.9).

Завершение анализа и выход из программы

Отредактируйте содержимое окна вывода в соответствии со своими предпочтениями: скройте лишнюю информацию, исправьте таблицы и пр. (см. раздел «Окно вывода и его редактирование» главы 2). За инструкциями по редактированию графиков обратитесь к главе 5.

Для дальнейшего использования окончательного результата все содержимое окна вывода или его фрагменты можно сохранить в файле *.spv, экспортировать в другой формат (например, Word), перенести в документ Word или вывести на печать (подробности см. в разделе «Сохранение, экспорт, перенос и печать результатов» главы 2).

Для выхода из программы выберите команду Выход в меню Файл.

24 Логистическая регрессия

348	Математическое описание логистической регрессии
349	Пошаговые алгоритмы вычислений
353	Представление результатов
356	Терминология, используемая при выводе
358	Завершение анализа и выход из программы

Логистическая регрессия представляет собой расширение множественной регрессии, рассмотренной в главе 18, и отличается от последней тем, что в качестве зависимой переменной используется не количественная, а дихотомическая переменная, имеющая лишь два возможных значения. Как правило, эти два значения символизируют принадлежность или не принадлежность объекта какой-либо группе, ответ типа «да» или «нет» и т. п.

Поскольку логистическая и множественная регрессии основаны на сходных принципах, мы будем исходить из того, что вы уже с ними знакомы (см. главу 18). Кратко, смысл регрессионного анализа можно свести к нахождению аналитического выражения, наиболее адекватно отражающего связь между зависимой переменной и множеством независимых переменных. Главным отличием логистической регрессии от множественной является толкование уравнения регрессии. Если множественная регрессия позволяет прогнозировать количественное значение зависимой переменной (критерия) на основе известных значений независимых переменных (предикторов), то логистическая регрессия прогнозирует вероятность некоторого события, находящуюся в пределах от 0 до 1. Кроме того, при помощи индикаторной схемы кодирования допускается использование в качестве предикторов категориальных (номинативных) переменных. Более детально особенности логистической регрессии рассмотрены в разделе «Представление результатов», сейчас же достаточно знать, что категориальный предиктор может быть представлен серией бинарных переменных — по одной на каждую категорию предиктора. Этим бинарным переменным присваиваются значения 1 или 0 в зависимости от того, к какой категории относится объект.

Математически суть логистической регрессии излагается на примере переменных из файла `helpLR.sav`. Содержимое этого файла примерно то же, что и файла `help.sav`, который использовался в главе 18. При помощи файла `help.sav` прогнозировалось количественное значение переменной `помощь` в зависимости от ряда количественных предикторов (`симпатия`, `проблема`, `эмпатия`, `польза` и `агрессия`). В нашем случае мы будем прогнозировать не величину помощи, а мнение партнера о том, полезна или нет оказанная ему помощь. Как нетрудно понять, речь идет о дихотомической зависимой переменной, поскольку предполагается лишь два варианта ответа: «да» или

«нет». Помимо того что зависимая переменная польза в файле helpLR.sav представлена в дихотомическом виде, в него добавлена еще одна категориальная переменная условия. Эта переменная обозначает одно из трех условий, в которых моделировалось оказание помощи: 1 — на работе, 2 — на улице в людном месте, 3 — в лесу, при встрече «один на один». Данные для файла helpLR.sav подбирались искусственно для того, чтобы проиллюстрировать рассматриваемый тип статистического анализа.

Математическое описание логистической регрессии

С логистической регрессией связаны такие математические понятия, как вероятность, шанс и натуральный логарифм шанса. Вероятность — это ожидаемая относительная частота некоторого события. Шанс представляет собой отношение вероятности того, что событие произойдет, к вероятности того, что событие не произойдет. Так, если вероятность дождя равна 0,2, то вероятность отсутствия дождя равна 0,8, следовательно, шанс, что дождь все-таки прольется, равен $0,2/0,8 = 0,25$. Обратите внимание, что шанс, в отличие от вероятности, не ограничен максимальным единичным значением; если, к примеру, вероятность дождя составляет не 0,2, а 0,8, то получаем шанс $0,8/0,2 = 4$. Единичное значение шанса соответствует ситуации, когда вероятности появления и не появления события равны.

Ключевым параметром логистической регрессии является *логит*. Логит равен натуральному логарифму шанса. Например, логит вероятности в 20 % равен $-1,386...$

Уравнение регрессии, используемое в главе 18, имеет следующий вид:

$$\text{помощь} = B_0 + B_1 \times \text{симпатия} + B_2 \times \text{агрессия} + B_3 \times \text{польза}$$

Согласно этому уравнению величина оказываемой помощи равна сумме константы (B_0) и значений трех переменных (отражающих симпатию, агрессивность и пользу), умноженных на соответствующие коэффициенты регрессии. Несмотря на то что в логистическом анализе оценивается полезность или бесполезность помощи, а не ее величина, уравнение логистической регрессии похоже на уравнение множественной регрессии:

$$\ln \left[\frac{P_{\text{помощь}}}{1 - P_{\text{помощь}}} \right] = B_0 + B_1 x_1 + B_2 x_2 + B_3 x_3.$$

Здесь $P_{\text{помощь}}$ — вероятность оказания помощи, x_1 , x_2 и x_3 — значения переменных симпатия, агрессия и польза.

Можно избавиться от натурального логарифма в левой части и преобразовать уравнение к виду:

$$\frac{P_{\text{помощь}}}{1 - P_{\text{помощь}}} = e^{B_0} + e^{B_1 x_1} + e^{B_2 x_2} + e^{B_3 x_3}.$$

Это уравнение, в свою очередь, можно преобразовать к следующему:

$$P_{\text{помощь}} = \frac{1}{1 + e^{-B_0} \times e^{-B_1 x_1} \times e^{-B_2 x_2} \times e^{-B_3 x_3}}.$$

Вероятно, подобные формулы не дают интуитивного представления о зависимости, которую они отражают, но на первых этапах это вполне нормально, поскольку интуиция всегда приходит с опытом. Тем не менее следует проявить максимум внимания к интерпретации модели логистической регрессии, поскольку выбор модели требует ее ясного теоретического понимания.

Поскольку данный раздел статистики весьма непросто для изучения, мы рекомендуем перед проведением логистического регрессионного анализа почитать дополнительную литературу.

Пошаговые алгоритмы вычислений

Перед началом логистического регрессионного анализа необходимо выполнить три подготовительных шага. Эти шаги (шаги 1–3) позволят подготовить рабочий файл данных, запустить программу IBM SPSS Statistics 19 и открыть файл (в данном случае — файл helpLR.sav). Пошаговые инструкции этого процесса приведены в главе 4 (с. 60), а подробные разъяснения — в главе 2.

После завершения шага 3 на экране должно присутствовать окно редактора данных со строкой меню и загруженным файлом helpLR.sav.

ШАГ 4

В меню Анализ выберите команду Регрессия ► Логистическая. На экране появится диалоговое окно Логистическая регрессия, показанное на рис. 24.1.

В диалоговом окне Логистическая регрессия вводятся имена зависимой переменной и предикторов, а также задаются другие параметры анализа. Как обычно, в его левой части располагается список всех переменных файла данных. Справа от списка находится поле Зависимая переменная; в это поле нужно поместить имя бинарной переменной, значения которой должны прогнозироваться в процессе анализа. В нашем примере в качестве зависимой будет использоваться переменная помощь, имеющая два уровня: 0 (нет) означает не оказанную помощь, а 1 (есть) — оказанную. Особенностью логистического анализа является то, что в процессе его проведения любые значения двух уровней зависимой переменной заменяются нулем и единицей. Это всегда происходит автоматически, и вам не нужно предпринимать никаких дополнительных действий для перекодирования переменной.

Назначение нижнего окна Переменная отбора наблюдений примерно такое же, как у кнопки Если, упоминавшейся в главе 4. Оно позволяет указать имя переменной и ее уровень для участия в анализе только части объектов.

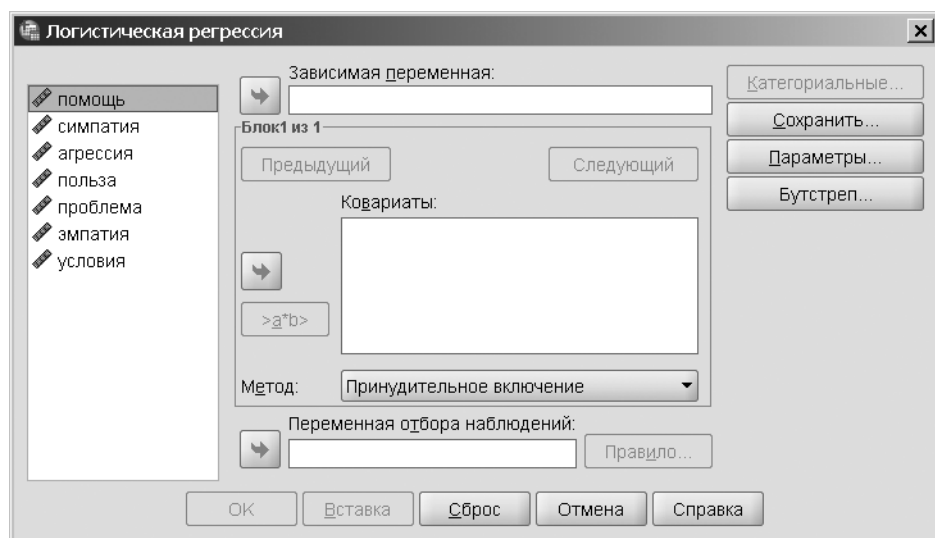


Рис. 24.1. Диалоговое окно Логистическая регрессия

В центре окна Логистическая регрессия находится список Ковариаты, предназначенный для задания предикторов, что, как обычно, выполняется щелчком на кнопке со стрелкой.

Помимо задания обычных переменных вы можете указать способ их взаимодействия; для этого следует выделить те переменные, которые участвуют во взаимодействии, а затем щелкнуть на кнопке $>a*b>$. Однако при этом следует помнить, что подобные действия ведут к не вполне очевидным последствиям, и для успешной интерпретации результатов вы должны иметь о подобных действиях ясное концептуальное представление.

Раскрывающийся список Метод позволяет выбрать способ построения уравнения регрессии. Пункт Принудительное включение, выбранный по умолчанию, предполагает включение в регрессионное уравнение всех указанных предикторов независимо от того, имеют они значимое влияние на зависимую переменную или нет. Пункт Включение: ОП (ОП — отношение правдоподобия) предполагает пошаговое включение в уравнение предикторов, оказывающих наибольшее воздействие на зависимую переменную, до последнего предиктора, чье воздействие окажется значимым. Пункт Исключение: ОП предполагает обратный принцип: сначала в уравнение регрессии включаются все выбранные предикторы, а затем происходит пошаговое исключение тех предикторов, влияние которых на зависимую переменную оказывается недостаточным согласно установленному критерию; исключение проводится до тех пор, пока в уравнении останутся только те предикторы, которые удовлетворяют заданному критерию. Описанные три метода из представленных в списке Метод используются чаще других.

В случае если к разным предикторам вы хотите применить различные методы, следует разделить последние на несколько блоков с помощью кнопок Предыдущий и Сле-

дующий, а затем для каждого блока указать собственный метод. SPSS будет последовательно обрабатывать блоки указанными методами. В нашем примере мы не станем делить предикторы на блоки, а выберем для всех предикторов метод Включение: ОП.

Если в вашем анализе используются категориальные предикторы, то после задания всех предикторов в списке Ковариаты следует воспользоваться кнопкой Категориальные. На экране появится диалоговое окно Логистическая регрессия: Задать категориальные переменные, представленное на рис. 24.2.

В списке Ковариаты вы можете видеть имена всех предикторов, указанных вами в одноименном списке диалогового окна Логистическая регрессия. В нашем примере переменная условие является не количественной, а категориальной, поэтому ее следует выделить и переместить из списка Ковариаты в список Категориальные ковариаты. Допускается применение нескольких способов обработки категориальных ковариат, каждый из которых задается при помощи списка Контраст, находящегося в правом нижнем углу диалогового окна.

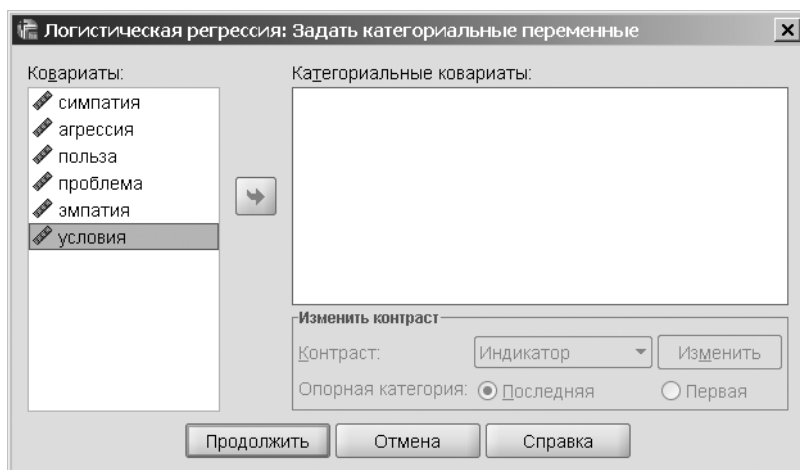


Рис. 24.2. Диалоговое окно Логистическая регрессия: Задать категориальные переменные

Кнопка Параметры в диалоговом окне Логистическая регрессия позволяет задать несколько дополнительных параметров, иногда оказывающихся весьма полезными на практике. При щелчке на ней открывается диалоговое окно Логистическая регрессия: Параметры, представленное на рис. 24.3.

Флажок Графики классификации позволяет включить в выводимые результаты диаграмму, на которой для каждого объекта вместе с вероятностью прогноза указана его классификация, полученная при помощи регрессионного уравнения. Таким образом, диаграмма наглядно иллюстрирует, насколько адекватным оказалось составленное уравнение регрессии. Флажок Корреляции оценок указывает на необходимость построения корреляционной матрицы для всех переменных, включенных в регрессионное уравнение. В области Критерии шагового отбора имеются поля Включение и Исключение. В первом указывается значение вероятности, служащее

критерием включения предиктора в уравнение регрессии, а во втором — значение вероятности для исключения из уравнения регрессии. По умолчанию в поле Включение указано значение 0,05, а в поле Исключение — значение 0,10. При желании исследователь может изменять их.

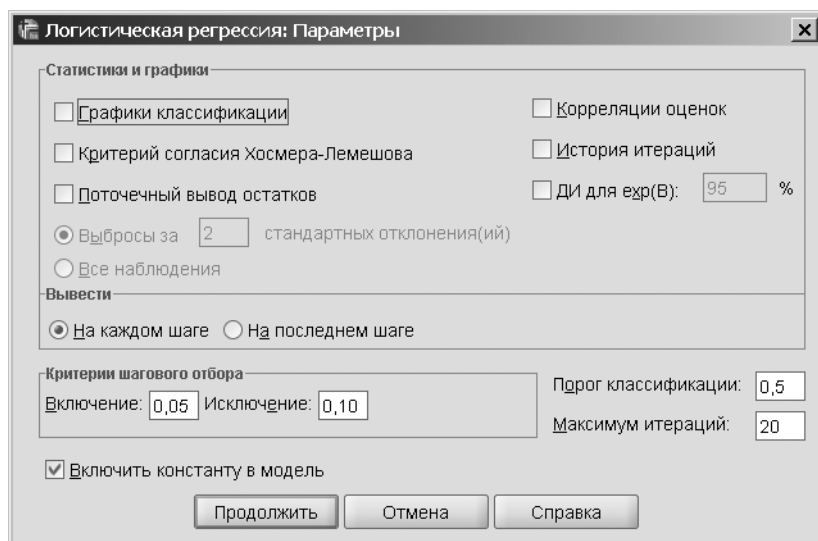


Рис. 24.3. Диалоговое окно Логистическая регрессия: Параметры

Кнопка Сохранить в диалоговом окне Логистическая регрессия позволяет сохранить вычисленные значения вероятностей и предсказанные значения зависимой переменной как новые переменные для наблюдений в файле данных. Это зачастую полезно для сравнения фактической и предсказанной группировок для отдельных объектов. В практическом плане интерес могут представлять предсказания для тех объектов, для которых фактические значения зависимой переменной неизвестны.

В следующем примере проводится логистический регрессионный анализ, в котором в качестве зависимой переменной выступает переменная помощь, а в качестве предикторов — переменные симпатия, агрессия, польза и условие.

ШАГ 5

После выполнения предыдущего шага у вас должно быть открыто диалоговое окно Логистическая регрессия, показанное на рис. 24.1. Если вы уже успели поработать с этим окном, щелкните на кнопке Сброс.

1. Щелкните сначала на переменной помощь, чтобы выделить ее, а затем — на верхней кнопке со стрелкой, чтобы переместить переменную в список Зависимая переменная.
2. Щелкните сначала на переменной симпатия, чтобы выделить ее, а затем — на нижней кнопке со стрелкой, чтобы переместить переменную в список Ковариаты.
3. Повторите предыдущее действие для переменных агрессия, польза и условие.

4. В списке Метод выберите пункт Включение: ОП и щелкните на кнопке Категориальные, чтобы открыть диалоговое окно Логистическая регрессия: Задать категориальные переменные, показанное на рис. 24.2.
5. Щелкните сначала на переменной условие, чтобы выделить ее, а затем — на кнопке со стрелкой, чтобы переместить переменную в список Категориальные ковариаты.
6. В списке Контрасты выберите пункт Отклонения и для подтверждения выбора щелкните на кнопке Изменить. Щелкните на кнопке Продолжить, чтобы вернуться в диалоговое окно Логистическая регрессия.
7. Щелкните на кнопке Параметры, чтобы открыть диалоговое окно Логистическая регрессия: Параметры, показанное на рис. 24.3.
8. Установите флажок Графики классификации, а затем щелкните на кнопке Продолжить, чтобы вернуться в диалоговое окно Логистическая регрессия.
9. Щелкните на кнопке ОК, чтобы открыть окно вывода.

После выполнения шага 5 программа автоматически откроет окно вывода. Для просмотра результатов при необходимости можно воспользоваться вертикальной и горизонтальной полосами прокрутки. Обратите внимание на стандартную строку меню в верхней части окна вывода: ее присутствие позволяет выполнять любые статистические операции, не переключаясь обратно в окно редактора данных.

Представление результатов

Ниже приведены результаты, сгенерированные программой при выполнении шага 5.

Поскольку методы Включение: ОП и Исключение: ОП являются пошаговыми, при их использовании выводится несколько таблиц, характеризующих каждый шаг и позволяющих наглядно представить себе процесс формирования уравнения регрессии. Здесь мы ограничились лишь итоговой таблицей.

Кодировки категориальных переменных

Если в вашем анализе использовались категориальные ковариаты и контрасты, то в начало выводимых результатов будет включена таблица (рис. 24.4), показывающая, каким образом исходные переменные (строки таблицы) были преобразованы в серию новых переменных (столбцы таблицы). Так, к примеру, для переменной условие, имеющей 3 уровня, были сформированы две дополнительные переменные (1 и 2), представляющие собой серию контрастов между тремя условиями.

Кодировки категориальных переменных

		Частота	Кодирование параметра	
			(1)	(2)
условия	на работе	12	1,000	,000
	на улице	17	,000	1,000
	в песу	17	-1,000	-1,000

Рис. 24.4. Кодировки категориальных переменных

Объединенные критерии для коэффициентов модели

Объединенные критерии для коэффициентов модели приведены на рис. 24.5.

Объединенные тесты для коэффициентов модели

		Хи-Квадрат	ст.св.	Знач.
Шаг 1	Шаг	15,920	2	,000
	Блок	15,920	2	,000
	Модель	15,920	2	,000
Шаг 2	Шаг	9,212	1	,002
	Блок	25,132	3	,000
	Модель	25,132	3	,000

Сводка для модели

Шаг	-2 Log Правдоподобие	R квадрат Кокса и Снелла	R квадрат Нэйджелкерка
1	47,501 ^a	,293	,391
2	38,289 ^b	,421	,563

Рис. 24.5. Объединенные критерии для коэффициентов модели

Трактовка терминов дана в конце этой главы.

Классификационная таблица

В классификационной таблице сравниваются прогнозируемые значения зависимой переменной, рассчитанные по уравнению регрессии, и фактические наблюдаемые значения (рис. 24.6). При определении прогнозируемой величины SPSS вычисляет вероятность для каждого объекта и на основании этой вероятности присваивает объекту одно из двух значений дихотомической переменной. Если вероятность оказалась менее 0,5, помощь оценивается как бесполезная (значение переменной помощь равно 0), в противном случае — как полезная (значение переменной помощь равно 1). Как показывают данные крайнего правого столбца таблицы, для 78,3 % объектов результаты прогноза оказались верными.

Таблица классификации^a

Наблюдаемое		Предсказанное		
		помощь		Процент корректных
		нет	есть	
Шаг 1	помощь нет	13	8	61,9
	есть	4	21	84,0
	Общий процент			73,9
Шаг 2	помощь нет	17	4	81,0
	есть	6	19	76,0
	Общий процент			78,3

a. Разделяющее значение = ,500

Рис. 24.6. Классификационная таблица

Переменные в уравнении

Таблица, приведенная на рис. 24.7, демонстрирует эффекты включения переменных в уравнение на каждом шаге его построения. Строка Константа для каждого шага соответствует константе B_0 регрессионного уравнения.

Переменные в уравнении						
	В	Стд.Ошибка	Вальд	ст.св.	Знч.	Ехр(В)
Шаг 1 ^а						
условия			11,413	2	,003	
условия(1)	-,279	,497	,314	1	,575	,757
условия(2)	-1,457	,495	8,659	1	,003	,233
Константа	,279	,369	,570	1	,450	1,321
Шаг 2 ^б						
агрессия	,652	,270	5,827	1	,016	1,920
условия			9,842	2	,007	
условия(1)	-,447	,559	,639	1	,424	,639
условия(2)	-1,658	,611	7,371	1	,007	,190
Константа	-6,210	2,735	5,156	1	,023	,002

а. Переменная(ые) включенная на шаге 1: условия.

б. Переменная(ые) включенная на шаге 2: агрессия.

Рис. 24.7. Переменные в уравнении регрессии

Трактовка терминов дана в конце этой главы.

Параметры модели при исключении переменных

Все переменные модели проверяются на предмет их возможного исключения, и результаты такой проверки включаются в таблицу (рис. 24.8). Уровень значимости для каждой переменной сравнивается со значимостью 0,1, указанной в поле Исключение диалогового окна Логистическая регрессия: Параметры. Если значимость переменной оказывается больше 0,1, происходит исключение переменной.

Модель при исключении члена				
Переменная	Log Правдоподобие Модели	Изменение в -2 Log Правдоподобие	ст.св.	Знч. Изменений
Шаг 1 условия	-31,711	15,920	2	,000
Шаг 2 агрессия	-23,750	9,212	1	,002
условия	-27,048	15,808	2	,000

Рис. 24.8. Результаты проверки переменных на возможность исключения из модели

Не включенные в уравнение переменные

На рис. 24.9 приведены переменные, не включенные в уравнение регрессии. Столбец Знч. характеризует значимость влияния данного предиктора на зависимую переменную без учета влияния остальных предикторов. Например, можно заключить, что переменная польза не оказывает значимого влияния на переменную мощность.

Переменные, не включенные в уравнение				Балл	ст.св.	Знач.
Шаг 1	Переменные	симпатия		5,471	1	,019
		агрессия		7,604	1	,006
		польза		3,499	1	,061
	Обобщенные статистики			12,005	3	,007
Шаг 2	Переменные	симпатия		3,622	1	,057
		польза		2,467	1	,116
	Обобщенные статистики			4,669	2	,097

Рис. 24.9. Переменные, не включенные в уравнение регрессии

Фактическая группировка и прогнозируемые вероятности

Диаграмма, показанная на рис. 24.10, генерируется в случае, если в окне Логистическая регрессия: Параметры установлен флажок Графики классификации.

В диаграмме используются первые буквы градаций зависимой переменной: е (есть — помощь оказана) и н (нет — помощь не оказана). По горизонтальной оси отложены значения прогнозируемой вероятности, вычисляемые по уравнению регрессии, а по вертикальной оси — частоты. Таким образом, каждый столбик на диаграмме соответствует определенной предсказанной вероятности, а его высота — количеству объектов, для которых предсказана данная вероятность. В случае идеальной логистической регрессии все буквы н окажутся левее букв е, а разделять их будет вероятность 0,5. Как видно из диаграммы, некоторые столбики включают в себя обе буквы, что свидетельствует об ошибках предсказания (высота в два символа соответствует одному объекту). Символам н в правой части диаграммы и символам е в левой части диаграммы соответствуют неправильные предсказания относительно оказания помощи. О количестве правильных и неправильных предсказаний позволяет судить классификационная таблица (см. рис. 24.6).

Терминология, используемая при выводе

Ниже дана трактовка терминов, используемых программой в окне вывода и относящихся к объединенным критериям для коэффициентов модели.

- ▶ Шаг — методу Включение: ОП понадобились всего два шага, чтобы включить все нужные переменные в уравнение регрессии. На самом деле, позже в модель вводятся и другие переменные, однако на данном этапе будем считать, что их всего две.
- ▶ Хи-квадрат — хи-квадрат шага, блока или модели. Это критерии статистической значимости воздействия на зависимую переменную всех предикторов заданной модели, блока или шага. На шаге 1 все три критерия χ^2 оказались равными: для модели и шага потому, что на шаге 1 они тождественны, а для

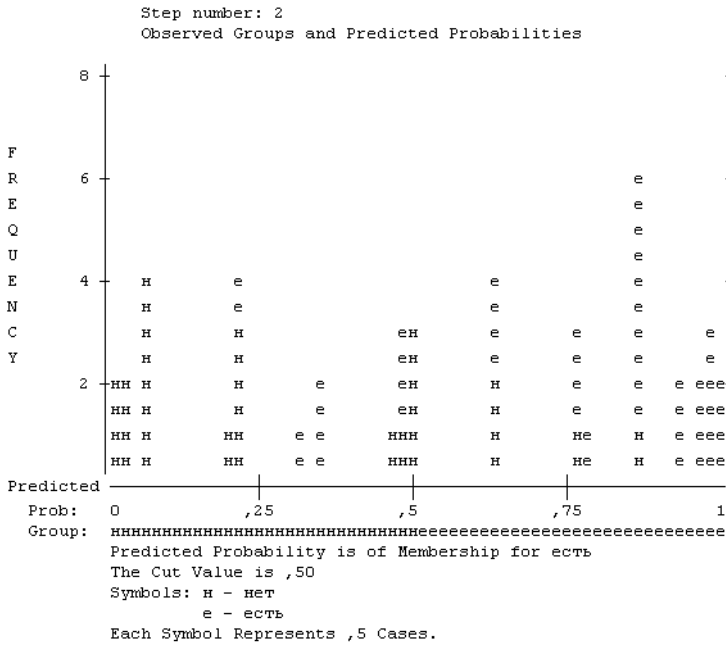


Рис. 24.10. График классификации

блока и модели потому, что модель содержит единственный блок. Большие значения критерия χ^2 говорят о том, что первая включенная переменная оказывает существенное влияние на зависимую переменную. Критерии χ^2 шага 2 свидетельствуют о том, что включение второй переменной в модель значительно улучшает последнюю; обратите внимание, что критерий χ^2 модели равен сумме критериев χ^2 шага и модели на предыдущем шаге.

- -2 Log Правдоподобие — эта величина характеризует модель и показывает, насколько хорошо она соответствует исходным данным. Чем меньше это значение, тем адекватнее сформированная модель.

R квадрат Кокса и Снелла, R квадрат Нэйджелкерка — приближения значения R^2 , показывающие долю влияния всех предикторов модели на дисперсию зависимой переменной.

Ниже дана трактовка остальных терминов, используемых программой в окне вывода и относящихся к переменным в уравнении регрессии.

- В — коэффициенты регрессионного уравнения, отражающие влияние соответствующих предикторов на зависимую переменную. Так, переменная агрессия оказывает положительное влияние на вероятность оказания помощи.
- Стд. Ошибка — стандартная ошибка, мера изменчивости коэффициентов В.
- Вальд — критерий значимости Вальда коэффициента В для соответствующего предиктора. Чем выше его значение (вместе с числом степеней свободы), тем выше значимость.

- ▶ Знч. — значимость по критерию Вальда.
- ▶ $\text{Exp}(B)$ — величина (e^B), которая может использоваться для интерпретации результатов анализа наравне с коэффициентом B (вспомните о двух формах регрессионного уравнения, в одной из которых используются коэффициенты B , а в другой — eB).

Завершение анализа и выход из программы

Отредактируйте содержимое окна вывода в соответствии со своими предпочтениями: скройте лишнюю информацию, исправьте таблицы и пр. (см. раздел «Окно вывода и его редактирование» главы 2). За инструкциями по редактированию графиков обратитесь к главе 5.

Для дальнейшего использования окончательного результата все содержимое окна вывода или его фрагменты можно сохранить в файле *.spv, экспортировать в другой формат (например, Word), перенести в документ Word или вывести на печать (подробности см. в разделе «Сохранение, экспорт, перенос и печать результатов» в главы 2).

Для выхода из программы выберите команду Выход в меню Файл.

25 Логлинейный анализ таблиц сопряженности

359	Понятие логлинейной модели
360	Логлинейный метод подбора модели
362	Пошаговые алгоритмы вычислений
366	Представление результатов
371	Завершение анализа и выход из программы

Логлинейные модели применяются для анализа таблиц сопряженности нескольких категориальных переменных. Если анализируется сопряженность двух таких переменных (двумерная таблица сопряженности), вполне достаточно применение критерия χ^2 (см. главу 8). Однако зачастую данные содержат существенно большее число категориальных переменных, и тогда визуальный анализ таблиц сопряженности становится невозможным. Например, визуально интерпретировать таблицу сопряженности пол \times класс \times вуз \times хобби практически невозможно. В подобных ситуациях на помощь исследователям приходят логлинейные модели, во многом сходные с регрессионными моделями и дисперсионным анализом.

В этой главе для иллюстрации иерархического логлинейного анализа мы используем файл `helpLLM.sav`. Этот файл содержит несколько модифицированные переменные файла `helpLR.sav`: переменная агрессия представлена в бинарном виде (1 — низкая, 2 — высокая), переменная симпатия имеет три категории (1 — слабая, 2 — средняя, 3 — сильная). Переменные помощь (2 градации) и условие (3 градации) оставлены без изменения. Таким образом, все переменные файла `helpLLM.sav` являются категориальными. В этой главе будет проводиться логлинейный анализ таблицы сопряженности агрессия \times симпатия \times условия \times помощь.

Понятие логлинейной модели

Логлинейная модель, в сущности, представляет собой множественную регрессионную модель, в которой категориальные переменные и их взаимодействия выступают в качестве предикторов, а роль зависимой переменной играет натуральный логарифм частот категорий. Именно использование логарифмической меры обуславливает линейность модели. Для взаимодействия переменных агрессия, симпатия и условия уравнение будет иметь вид:

$$\ln(\text{частота}) = \mu + \lambda A + \lambda C + \lambda Y + \lambda A \times C + \lambda A \times Y + \lambda C \times Y + \lambda A \times C \times Y$$

В этом уравнении частота — это частота текущей ячейки частотной таблицы, μ — общее среднее воздействия, эквивалентное константе во множественном регрессионном анализе, а каждое значение λ представляет собой воздействие со стороны одной или более независимых переменных. Так, λA является воздействием переменной агрессия, λU и λC — соответственно переменных условия и симпатия, $\lambda A \times C$ — взаимодействия переменных агрессия и симпатия, $\lambda A \times C \times U$ — взаимодействия переменных агрессия, симпатия и условия. Представленная здесь модель называется насыщенной, поскольку содержит все предикторы и их возможные взаимодействия. Однако насыщенная модель обычно не является оптимальной, так как редко все главные эффекты и взаимодействия оказываются значимыми. Как правило, существуют более предпочтительные альтернативы в виде ненасыщенных моделей, которые отражают лишь статистически значимые главные эффекты и взаимодействия переменных.

Подменю Логлинейный анализ содержит три команды.

- ▶ **Общий** — эта команда допускает вхождение в модель любых факторов и их взаимодействий и предполагает, что исследователь перед проведением анализа уже имеет гипотезы о составе модели.
- ▶ **Логит** — применение этой команды позволяет рассматривать дихотомические переменные как зависимые, а одну (или более) категориальную переменную как независимую. При этом зависимая дихотомическая переменная используется не для прогнозирования частот категорий, а для разделения всех категорий на две группы. Смысл понятия «логит» раскрывается в главе 24.
- ▶ **Подбор модели** — эта команда позволяет из всех возможных ненасыщенных моделей подобрать ту, которая в наибольшей степени соответствует исходным данным. Подбор осуществляется, как правило, автоматически. В результате выявляется совокупность значимых связей между категориальными переменными и вычисляются параметры μ и λ логлинейной модели.

В данной главе рассматривается последний из перечисленных методов, который основан на применении иерархической логлинейной модели.

Логлинейный метод подбора модели

Теоретически из насыщенной модели можно удалить любые элементы, получив, таким образом, произвольную ненасыщенную модель. Далее можно проверить состоятельность этой модели и в случае несоответствия ее исходным данным перейти к анализу другой ненасыщенной модели. Однако большинство исследователей отдают предпочтение иерархическим логлинейным моделям, которые позволяют упорядочить процесс подбора окончательной состоятельной модели. Основной особенностью иерархических моделей является то, что присутствие какого-либо взаимодействия переменных означает присутствие всех взаимодействий, имеющих более низкий порядок, и главных эффектов этих переменных. Например,

если в модели присутствует взаимодействие агрессия × симпатия, то в ней присутствуют главные эффекты переменных агрессия и симпатия; если в модели присутствует взаимодействие агрессия × симпатия × условия, то в ней также присутствуют взаимодействия агрессия × симпатия, агрессия × условия и симпатия × условия, и т. д.

Существуют три вспомогательных метода, которые предназначены для подбора адекватной модели. Все три метода оказываются весьма полезными и приводят к сходным результатам. Тем не менее окончательный выбор модели определяется не только статистическими результатами применения метода, но и пониманием исследователем особенностей исходных данных. Краткие описания методов приведены ниже, более подробные — в разделе «Представление результатов».

- Метод *исследования оценок параметров* предназначен для вычисления оценок параметров для насыщенной модели. Помимо обычных оценок параметров SPSS вычисляет стандартизованные оценки. Если значения последних невелики, то они не оказывают значимого влияния на модель и обычно исключаются.
- Метод *вычисления частичного критерия хи-квадрат* в дополнение к оценкам параметров модели SPSS вычисляет критерий χ^2 , характеризующий степень соответствия модели исходным данным. При помощи этого критерия проверяется, являются ли все однофакторные эффекты, а также эффекты более высоких порядков статистически значимыми (анализируются комбинации всех эффектов первого, второго и т. д. порядков). При этом отсутствие общей значимости эффектов второго порядка вовсе не означает, что все эффекты первого порядка не являются значимыми. Аналогично, из отсутствия общей значимости эффектов любого порядка не следует отсутствие значимости отдельных взаимодействий этого порядка. Вследствие этих двух особенностей в SPSS предусмотрена возможность раздельной проверки главных эффектов и эффектов взаимодействий.
- Суть метода *пошагового исключения* состоит в автоматической «подгонке» модели и сходна с методом исключения предикторов из уравнения регрессии: из насыщенной модели постепенно исключаются те элементы (переменные и их взаимодействия), которые не оказывают значимого воздействия. Данный метод построения модели относится к иерархическому логлинейному моделированию. Так, если обнаружено статистически значимое взаимодействие четырех переменных, не проверяется (на предмет исключения из модели) взаимодействие трех из этих переменных, иначе модель не являлась бы иерархической по определению. Окончательный результат «подгонки» модели наиболее приемлем, если все оставшиеся в ней элементы оказываются статистически достоверными.

Иерархические логлинейные модели относятся к очень сложному разделу статистики, поэтому мы рекомендуем перед проведением логлинейного анализа обратиться к дополнительным источникам.

Пошаговые алгоритмы вычислений

До применения логлинейного анализа для подбора модели нужно выполнить три подготовительных шага. Эти шаги (шаги 1–3) позволят подготовить рабочий файл данных, запустить программу IBM SPSS Statistics 19 и открыть файл (в данном случае — файл helpLLM.sav). Пошаговые инструкции этого процесса приведены в главе 4 (с. 60), а подробные разъяснения — в главе 2.

После завершения шага 3 на экране должно присутствовать окно редактора данных со строкой меню и загруженным файлом helpLLM.sav.

ШАГ 4

В меню Анализ выберите команду Логлинейный ► Подбор модели. На экране появится диалоговое окно Логлинейный анализ: подбор модели, показанное на рис. 25.1.

В этом диалоговом окне задаются переменные для анализа, а также выбирается метод построения модели. Слева находится список всех переменных файла данных. Для включения переменной в анализ, достаточно выделить ее и переместить в список Факторы, щелкнув на верхней кнопке со стрелкой. Данный метод применим только к категориальным переменным, диапазон каждой из которых задается кнопкой Задать диапазон. Эта кнопка открывает окно Логлинейный анализ: Задать диапазон, показанное на рис. 25.2. В нем необходимо задать минимальный и максимальный уровни каждой переменной в полях Минимум и Максимум.

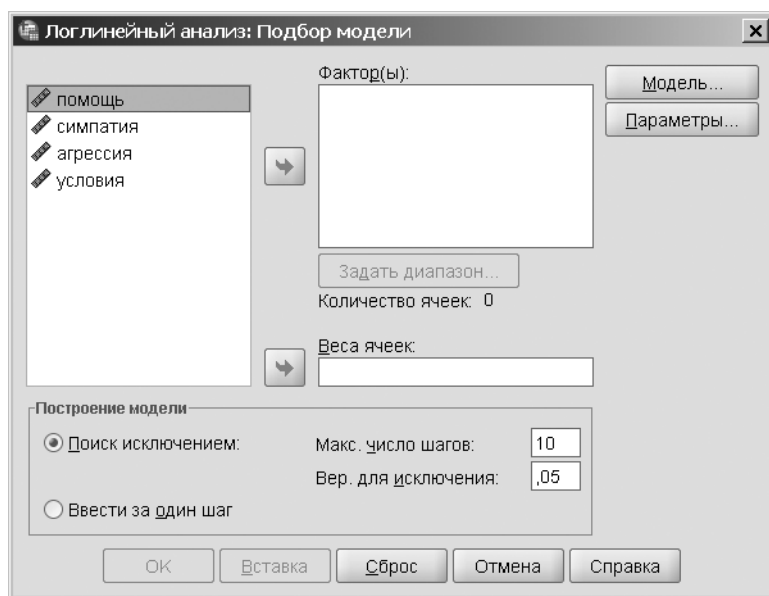


Рис. 25.1. Диалоговое окно Логлинейный анализ: Подбор модели

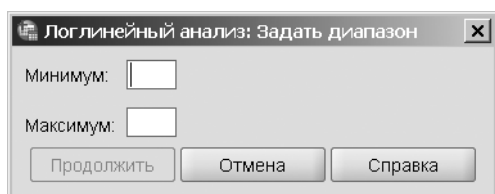


Рис. 25.2. Диалоговое окно Логлинейный анализ: Задать диапазон

Параметр Веса ячеек диалогового окна Логлинейный анализ: Подбор модели используется в том случае, если ваша модель содержит структурные нули. Понятие структурных нулей в этой книге не рассматривается; при необходимости обратитесь к руководству пользователя SPSS.

Диалоговые окна задания переменных для не рассматриваемых в этой книге команд Логлинейный ► Общий и Логлинейный ► Логит отличаются тем, что позволяют задавать в качестве факторов не только категориальные, но и количественные переменные (ковариаты). Помимо этого, при использовании логит-моделей необходимо задание одной или нескольких бинарных зависимых переменных, каждая из которых используется в анализе для деления независимых переменных на две категории.

Рассмотрим переключатели Построение модели, расположенные в нижней части диалогового окна Логлинейный анализ: Подбор модели. Положение Поиск исключением определяет построение модели методом пошагового исключения: из насыщенной модели в результате серии шагов исключаются все элементы, не оказывающие значимого влияния. Положение переключателя Ввести за один шаг позволяет получить модель со всеми элементами. Если используется метод пошагового исключения, в поле Вер. для исключения нужно указать величину значимости, которая будет служить критерием исключения элементов из модели. По умолчанию это значение $p = 0,05$, и нередко оно устраивает исследователей. Под *шагом* понимается удаление очередного элемента из модели; в поле Макс. число шагов вы можете задать максимальное количество шагов для того, чтобы процедура принудительно прекратила процесс исключения предикторов. По умолчанию максимальное число шагов равно 10.

После того как метод построения модели и переменные, входящие в ее состав, заданы, можно переходить к определению других параметров команды, используя кнопки Параметры и Модель.

При щелчке на кнопке Параметры открывается диалоговое окно Логлинейный анализ: Параметры, показанное на рис. 25.3. Оно позволяет настроить параметры генерируемых командой итоговых данных, а также задать некоторые технические аспекты ее выполнения (последние в этой книге не рассматриваются). Флажок Частоты включает в выводимые данные таблицу частот модели, а флажок Остатки — разности между фактическими и прогнозируемыми частотами, рассчитываемыми при помощи модели.

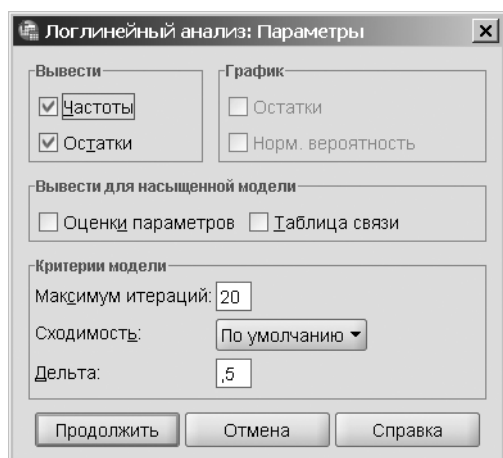


Рис. 25.3. Диалоговое окно Логлинейный анализ: Параметры

Нередко используются флажки *Оценки параметров* и *Таблица связей*. Для насыщенной модели эти флажки позволяют включить в выводимые данные оценки параметров, что может быть полезно при выборе модели, и таблицу ассоциаций, содержащую критерии χ^2 для всех главных эффектов и взаимодействий. Оба флажка при исследовании насыщенной модели рекомендуется оставлять установленными. Если же для анализа необходима ненасыщенная модель, включающая лишь некоторые взаимодействия и факторы, то, щелкнув на кнопке *Модель*, следует открыть диалоговое окно *Логлинейный анализ: Модель*, показанное на рис. 25.4.

Если в группе *Задать модель* установить переключатель *Настраиваемая*, то можно выбирать переменные в списке *Факторы*. Раскрывающийся список в центре окна позволяет выбрать один из типов их взаимодействий: *Главные эффекты*, *Взаимодействие*, *Все 2-факторные*, *Все 3-факторные*, *Все 4-факторные* и *Все 5-факторные*. Поскольку команда *Логлинейный* ► *Подбор модели* строит иерархические модели, при выборе *Взаимодействие* все взаимодействия более низких порядков включаются в модель автоматически.

Задание модели вручную требуется в тех случаях, когда есть предварительные гипотезы о составе ненасыщенной модели. Иногда модель задается для того, чтобы ограничить количество рассматриваемых взаимодействий, если взаимодействия высоких порядков вызывают трудности в интерпретации. Для настройки модели вручную следует установить переключатель *Настраиваемая*, тогда становится доступным включение в модель необходимых главных факторов и взаимодействий. Главные факторы включаются в модель следующим образом: имя соответствующего фактора выделяется в списке *Факторы* и с помощью кнопки с направленной вправо стрелкой перемещается в правый список. Если требуется включить в модель взаимодействие факторов, следует щелчками выделить в списке *Факторы* имена всех факторов, участвующих во взаимодействии, затем в раскрывающемся списке *Задать члены*: *Тип* выбрать пункт *Взаимодействия* и, наконец, как и в предыдущем случае, щелкнуть на кнопке с направленной вправо стрелкой. Чтобы построить

сразу несколько взаимодействий определенного порядка для выбранных факторов, нужно выполнить те же действия, что и для одиночного взаимодействия, но в раскрывающемся списке выбрать пункт в зависимости от желаемого результата.

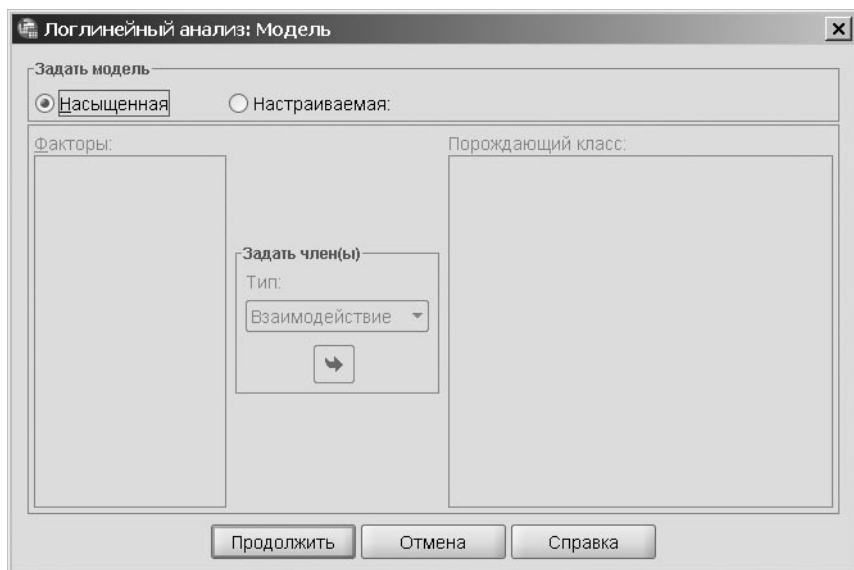


Рис. 25.4. Диалоговое окно Логлинейный анализ: Модель

В следующем примере проводится иерархический логлинейный анализ переменных агрессия, симпатия, условия и помощь, имеющих соответственно 2, 3, 3 и 2 уровня. Мы будем исследовать насыщенную модель, а в выводимые данные включим оценки параметров и критерии χ^2 для отдельных эффектов. Затем мы применим метод пошагового исключения, чтобы удалить из насыщенной модели те элементы, которые не оказывают значимого воздействия.

ШАГ 5

После выполнения предыдущего шага у вас должно быть открыто диалоговое окно Логлинейный анализ: Подбор модели, показанное на рис. 25.1. Если вы уже успели поработать с этим окном, щелкните на кнопке Сброс.

1. Щелкните сначала на переменной помощь, чтобы выделить ее, а затем — на верхней кнопке со стрелкой, чтобы переместить переменную в список Факторы.
2. Щелкните на кнопке Задать диапазон, чтобы открыть диалоговое окно Логлинейный анализ: Определение диапазона, показанное на рис. 25.2.
3. В поле Минимум введите число 0, нажмите клавишу Tab, чтобы переместить фокус ввода в поле Максимум, введите число 1 и щелкните на кнопке Продолжить, чтобы вернуться в диалоговое окно Логлинейный анализ: Подбор модели.

4. Повторите первые три действия для переменных симпатия, агрессия и условия, используя для задания значений переменных симпатия и условия числа 1 и 3, а для задания значений переменной агрессия числа 1 и 2.
5. Щелкните на кнопке Параметры, чтобы открыть диалоговое окно Логлинейный анализ: Параметры, показанное на Рис. 25.3.
6. Установите флажки Оценки параметров и Таблица связи, а затем щелкните на кнопке Продолжить, чтобы вернуться в диалоговое окно Логлинейный анализ: Подбор модели.
7. Щелкните на кнопке ОК, чтобы открыть окно вывода.

После выполнения шага 5 программа автоматически откроет окно вывода. Для просмотра результатов при необходимости можно воспользоваться вертикальной и горизонтальной полосами прокрутки. Обратите внимание на стандартную строку меню в верхней части окна вывода: ее присутствие позволяет выполнять любые статистические операции, не переключаясь обратно в окно редактора данных.

Представление результатов

В этом разделе приводятся фрагменты данных, генерируемые программой при выполнении шага 5. Напоминаем, что если выводимые данные занимают слишком много места, они представлены в окне вывода лишь частично. В этом случае в том месте окна вывода, в котором скрыты дополнительные данные, присутствует значок в виде красного треугольника. Чтобы увидеть такие результаты полностью, дважды щелкните на имеющемся фрагменте.

Проверка на равенство нулю взаимодействий порядка K и более высокого порядка

В таблице, приведенной на рис. 25.5, представлены результаты проверки эффекта от взаимодействий указанного и более высокого порядка, и отдельно — только указанного порядка. Для проверки используется два варианта критерия χ^2 . Оба критерия оценивают, являются ли ожидаемые частоты в ячейках для соответствующей модели значительно отличающимися от наблюдаемых частот или нет. Если отличие значимо, гипотеза об отсутствии взаимодействия отвергается и делается вывод о связи соответствующих переменных.

Рассмотрим первую часть таблицы: Эффекты K -го порядка и более высоких порядков. Последняя строка этой части таблицы соответствует четвертому порядку взаимодействия (значение в столбце K равно 4) и характеризует общий (суммарный) эффект от взаимодействий четырех и более переменных; большой уровень значимости и низкое значение χ^2 говорят о том, что взаимодействие всех четырех переменных не оказывает существенного влияния на модель. То же касается взаимодействия третьего и выше порядка. Противоположный вывод можно сделать

об остальных взаимодействиях: χ^2 возрастает, а значимость убывает с убыванием порядка. Первые две строки свидетельствуют о том, что среди всех взаимодействий низкого порядка имеются такие, которые оказывают на модель значимое влияние.

Эффекты К-го порядка и более высоких порядков

	К	Ст. св.	Отношение правдоподобия		Пирсона		Число итераций
			Хи-квадрат	Знач.	Хи-квадрат	Знач.	
Эффекты К-го порядка и более высоких порядков	1	35	58,726	,007	52,609	,028	0
	2	29	54,241	,003	47,940	,015	2
	3	16	21,784	,150	17,649	,345	4
	4	4	,397	,983	,212	,995	4
Эффекты К-го порядка ^а	1	6	4,486	,611	4,669	,587	0
	2	13	32,456	,002	30,291	,004	0
	3	12	21,387	,045	17,437	,134	0
	4	4	,397	,983	,212	,995	0

Число степеней свободы, использованное в данных критериях НЕ было скорректировано на наличие структурных или выборочных нулей. Критерии, основанные на данном числе степеней свободы, могут быть консервативными.

а. Результаты тестов эффектов к-го и более высоких порядков равны нулю.

б. Результаты тестов эффектов к-го порядка равны нулю.

Рис. 25.5. Проверка эффекта от взаимодействий разных порядков

Вторая часть таблицы на рис. 25.5 Эффекты К-го порядка проверяет на равенство нулю только эффекты заданного порядка. Соответственно, здесь характеризуются взаимодействия только указанного порядка. Как можно видеть в столбце Отношение правдоподобия, значимое воздействие на модель оказывают взаимодействия второго и третьего порядков.

Далее дана трактовка терминов, используемых программой в окне вывода.

- К — порядок взаимодействий. В разделе для порядка К и выше: 4 — порядок 4 и выше, 3 — порядок 3 и выше и т. д. В разделе только для порядка К: 1 — порядок 1, 2 — порядок 2 и т. д.
- Ст. св. — число степеней свободы для соответствующих взаимодействий.
- Отношение правдоподобия, Хи-квадрат — критерий χ^2 для определения вероятности того, что общий эффект от соответствующих взаимодействий равен нулю (по замыслу его авторов, более «правдоподобный», чем критерий хи-квадрат Пирсона).
- Пирсона Хи-квадрат — критерий χ^2 для определения вероятности того, что общий эффект от взаимодействий равен нулю.
- Знач. — значимость; вероятность того, что общий эффект от соответствующих взаимодействий равен нулю. Чем меньше вероятность, тем большее влияние эффект оказывает на модель.
- Число Итераций — число итераций для вычисления значений χ^2 .

Проверка частных взаимосвязей

На рис. 25.6 приведены критерии χ^2 для каждого из воздействий насыщенной модели. Воздействия с высоким значением χ^2 оказывают значимое влияние на модель; чем ниже значение χ^2 , тем ниже значимость. В данном примере главные эффекты переменных не являются статистически значимыми, а к значимым можно отнести двухфакторные взаимодействия помощь \times условие и помощь \times симпатия, а также взаимодействие трех переменных помощь \times симпатия \times агрессия. Поскольку эти частные критерии χ^2 не обязательно являются независимыми, их доля в итоговом значении χ^2 для насыщенной модели может отличаться от их доли в ненасыщенной модели.

Частные взаимосвязи

Эффект	Ст. св.	Частный Хи-квадрат	Знач.	Число итераций
помощь*симпатия*агрессия	2	6,288	,043	3
помощь*симпатия*условия	4	,091	,999	5
помощь*агрессия*условия	2	,545	,762	5
симпатия*агрессия*условия	4	2,441	,655	6
помощь*симпатия	2	7,430	,024	3
помощь*агрессия	1	1,956	,162	4
симпатия*агрессия	2	1,944	,378	4
помощь*условия	2	16,593	,000	3
симпатия*условия	4	2,277	,685	3
агрессия*условия	2	1,230	,541	4
помощь	1	,348	,555	2
симпатия	2	1,605	,448	2
агрессия	1	1,398	,237	2
условия	2	1,134	,567	2

Рис. 25.6. Проверка частных взаимосвязей

Ниже дана трактовка терминов, используемых программой в окне вывода.

- Ст. св. — число степеней свободы соответствующего эффекта.
- Частный Хи-квадрат — критерий χ^2 для отдельного эффекта.
- Знач. — статистическая значимость; вероятность того, что величина эффекта равна нулю. Чем меньше вероятность, тем большее влияние эффект оказывает на модель.
- Число итераций — число итераций для вычисления соответствующей связи.

Оценки параметров модели

Фрагмент результатов, представленный на рис. 25.7, содержит оценки параметров, представляющие собой значения λ из логлинейного уравнения: для трехфакторного взаимодействия помощь \times симпатия \times агрессия и двухфакторного взаимодействия помощь \times условия. По уровню статистической значимости первое из указанных взаимодействий приближается к статистически достоверному ($p < 0,1$), а второе статистически достоверно ($p < 0,05$). Поскольку сумма всех параметров должна

быть равна нулю, на основе полученных данных можно вычислить другие. Так, в нижней строке и правом столбце табл. 25.1 показаны значения λ , вычисленные на основе представленных выше данных для взаимодействия *помощь* × *условия*, исходя из того, что сумма значений в строках и столбцах должна быть равна 0.

Оценки параметров

Эффект	Параметр	Оценка	Станд. ошибка	Z	Знач.	95% доверительный интервал	
						Нижняя граница	Верхняя граница
помощь*симпатия*агрессия	1	-,286	,258	-1,108	,268	-,791	,220
	2	,443	,236	1,881	,060	-,019	,905
помощь*условия	1	,037	,246	,150	,881	-,446	,520
	2	,483	,239	2,026	,043	,016	,951

Рис. 25.7. Оценки параметров модели

Таблица 25.1. Пример кодирования взаимодействия

Уровни переменной «помощь»	Уровни переменной «симпатия»		
	1	2	3
0	0,0369	0,4833	-0,4798
1	-0,0369	-0,4833	0,4798

Полученные оценки можно использовать для интерпретации распределения частот. Например, из таблицы (нижняя строка) видно, что при наибольшем уровне симпатии (3) помощь оказывается чаще. Но лучше для этих целей использовать таблицы сопряженности, воспользовавшись командой *Описательные статистики* ► *Таблицы сопряженности*.

Ниже дана трактовка терминов, используемых программой в окне вывода и относящихся к оценкам параметров модели.

- Параметр — номер параметра. Нумерация параметров производится от наименьших значений независимых переменных к наибольшим (в качестве примера кодирования взаимодействий см. табл. 25.1).
- Оценка — оценка коэффициента λ в логлинейном уравнении.
- Станд. ошибка — стандартная ошибка, мера изменчивости оценки коэффициента.
- Z — z-значение коэффициента. Значения больше 1,96 (меньше -1,96) статистически достоверны (имеют уровень значимости меньше 0,05).
- Знач. — статистическая значимость оценки.
- 95% доверительный интервал — границы интервала, в который фактическое (не приближенное) значение коэффициента попадает с вероятностью 95 %.

Обратное пошаговое исключение

Результаты обратного пошагового исключения являются, пожалуй, наиболее важными среди всех остальных, поскольку дают представление о наиболее состоятельной модели (рис. 25.8). В представленном фрагменте результатов содержатся критерии χ^2 , их значимости и все взаимодействия, участвующие в процессе пошагового исключения на шагах 5–7 (шаги 1–4 для экономии места опущены). Процесс пошагового исключения начинается с насыщенной модели. Сначала SPSS вычисляет значение χ^2 для модели, включающей взаимодействия четвертого и более низких порядков (включая главные эффекты); это значение равно 0, поскольку изначально модель является насыщенной. Далее устанавливается, что при исключении взаимодействия четвертого порядка изменение величины χ^2 не значимо, и исключение производится. Затем анализируется модель, состоящая из взаимодействий третьего и более низких порядков; из модели исключаются незначимые взаимодействия. Весь процесс занимает 7 шагов, и в результате формируется иерархическая модель, в которую входят взаимодействия помощь \times симпатия \times агрессия и помощь \times условия. Оба взаимодействия статистически достоверны. Значение критерия χ^2 для всей модели составляет 13,268, число степеней свободы равно 20. Следует отметить, что критерий χ^2 для всей модели оценивает расхождение между наблюдаемыми частотами в исходной таблице сопряженности и ожидаемыми частотами, вычисленными при помощи модели.

Пошаговый отчет

Шаг ^b		Эффекты	Хи-квадрат ^a	Ст. св.	Знач.	Число итераций
5	Генерируемы классы ^c	помощь*симпатия*агрессия, помощь*условия, агрессия*условия	12,122	18	,841	
	Эффект удаления 1	помощь*симпатия*агрессия	11,939	2	,003	4
	2	помощь*условия	16,440	2	,000	2
	3	агрессия*условия	1,146	2	,564	2
6	Генерируемы классы ^c	помощь*симпатия*агрессия, помощь*условия	13,268	20	,866	
	Эффект удаления 1	помощь*симпатия*агрессия	11,944	2	,003	3
	2	помощь*условия	15,920	2	,000	2
7	Генерируемы классы ^c	помощь*симпатия*агрессия, помощь*условия	13,268	20	,866	

a. 'Эффект удаления' - это изменение Хи-квадрат в результате удаления эффекта из модели.

b. На каждом шаге удаляется эффект с наибольшим уровнем значимости отношения правдоподобия, при условии, что уровень значимости превышает ,050.

c. Выводятся статистики наилучшей модели на каждом шаге после нулевого.

Рис. 25.8. Некоторые результаты обратного пошагового исключения

Поскольку результат оказался незначимым, созданная модель адекватна исходным данным: ожидаемые в соответствии с моделью частоты вполне точно соответствуют наблюдаемым частотам. Для содержательной интерпретации модели следует обратиться к значимым компонентам модели, в данном случае, к таблицам сопряженности помощь \times симпатия \times агрессия и помощь \times условия.

Завершение анализа и выход из программы

Отредактируйте содержимое окна вывода в соответствии со своими предпочтениями: скройте лишнюю информацию, исправьте таблицы и пр. (см. раздел «Окно вывода и его редактирование» главы 2). За инструкциями по редактированию графиков обратитесь к главе 5.

Для дальнейшего использования окончательного результата все содержимое окна вывода или его фрагменты можно сохранить в файле *.spv, экспортировать в другой формат (например, Word), перенести в документ Word или вывести на печать (подробности см. в разделе «Сохранение, экспорт, перенос и печать результатов» в главе 2).

Для выхода из программы выберите команду Выход в меню Файл.

Глоссарий

В. Показатели B представляют собой набор коэффициентов и константу регрессионного уравнения. Показатель B можно рассматривать как весовой коэффициент, характеризующий влияние соответствующей независимой переменной (предиктора) на зависимую переменную (критерий). Положительное значение B указывает на то, что с возрастанием предиктора значение критерия возрастает, а отрицательное значение B — на то, что значение критерия убывает.

df. См. *Число степеней свободы*.

F-критерий. В дисперсионном анализе отношение межгруппового среднего квадрата к внутригрупповому среднему квадрату. Данная величина позволяет сравнить межгрупповую дисперсию с внутригрупповой дисперсией. В случае если первая окажется значительно выше второй, это будет означать наличие значимого различия между группами. В множественном регрессионном анализе F -критерий позволяет определить значимость множественной корреляции.

К. В иерархической логлинейной модели, порядок взаимодействия эффектов; $k = 1$ соответствует первому порядку (одна переменная), $k = 2$ — второму порядку (две переменные), и т. д.

р-уровень. См. *Значимость*.

R. Множественный коэффициент корреляции между зависимой переменной и двумя или более независимыми переменными. Значение R лежит в пределах от 0 до 1 и интерпретируется по аналогии с обычным (двухмерным) коэффициентом корреляции.

R². Квадрат коэффициента множественной корреляции (коэффициент детерминации), доля дисперсии зависимой переменной, обусловленная воздействием двух или более независимых переменных.

S-стресс. В многомерном шкалировании мера степени соответствия модели исходной матрице различий. Чем меньше это значение, тем лучше соответствие.

t-критерий. Критерий для определения статистической значимости различия двух средних.

t-критерий для зависимых выборок. Критерий, сравнивающий средние значения двух распределений для одной и той же выборки.

t-критерий для независимых выборок. Критерий, сравнивающий средние значения одной и той же переменной для двух независимых выборок.

t-критерий для одной выборки. Критерий, предназначенный для сравнения среднего значения распределения переменной с некоторой эталонной величиной.

t-критерий в регрессионном анализе. Критерий, определяющий статистическую значимость корреляций, равен отношению коэффициента B к своей стандартной ошибке.

V Крамера. Мера ассоциации между значениями двух категориальных переменных. Значение V всегда варьируется от 0 до 1 и интерпретируется по аналогии с коэффициентом корреляции (исключая отсутствия отрицательных значений). Нередко используется в контексте χ^2 -анализа; вычисление осуществляется по формуле (k — наименьшее из количеств строк и столбцов):

$$V = \sqrt{\chi^2 / [N(k-1)]}.$$

z-значения. Также называются стандартизованными значениями. После стандартизации (или z-преобразования) значений переменной среднее равно 0, стандартное отклонение равно 1. Стандартизованное значение может характеризовать направление и степень отклонения исходного значения от среднего. Для стандартизованных значений, превышающих по модулю 1,96, уровень значимости оказывается ниже 0,05.

Альфа (α). Мера внутренней согласованности измерительной шкалы, вычисляемая по формуле $\alpha = rk/[1 + (k-1)r]$, где k — число переменных в анализе, r — среднее значение корреляции между пунктами шкалы. Значение α зависит от числа переменных, поэтому нет точной интерпретации его величины; тем не менее, в большинстве случаев действует следующая оценка внутренней согласованности шкалы: $\alpha > 0,9$ — отличная; $\alpha > 0,8$ — хорошая; $\alpha > 0,7$ — приемлемая; $\alpha > 0,6$ — сомнительная; $\alpha > 0,5$ — малопригодная; $\alpha < 0,5$ — недопустимая.

Альфа, если элемент удален. В анализе надежности значение α для шкалы, получающейся путем удаления текущего ее пункта.

Априорная вероятность для каждой группы. Вероятность, характеризующая предполагаемое соотношение численности групп. Для каждой из двух групп она равна 0,5, если предполагается, что их численность одинакова.

Асимметричная матрица. Квадратная матрица, у которой хотя бы в одной паре ячеек, симметрично расположенных относительно главной диагонали, значения различны. Примером симметричной матрицы является корреляционная матрица.

Асимметрия. Мера отклонения распределения от нормального, характеризующая симметричность графика.

Асимптотические значения. Величины, на которых основано определение оценок параметров. Оценки параметров вычисляются в случаях, когда определение точных значений невозможно, в частности, в регрессионном анализе и некоторых других статистических процедурах.

Барлетта критерий сферичности. Критерий многомерной нормальности для распределения переменных. Помимо нормальности критерий проверяет, отличаются ли корреляции от 0. Значение p -уровня, меньшее 0,05, указывает на то, что данные вполне приемлемы для проведения факторного анализа

Бета (β). В регрессионном анализе β означает стандартизованный коэффициент регрессии и представляет собой B -коэффициент для нормализованных переменных. Значения β всегда лежат в интервале от -1 до $+1$ и могут сравниваться друг с другом для разных переменных.

Бета при включении. Данная величина используется во множественном регрессионном анализе для переменных, не вошедших в уравнение регрессии, и представляет собой значение коэффициента β , рассчитанного в предположении, что соответствующая переменная была включена в регрессионное уравнение.

Биномиальный критерий. Непараметрический критерий, определяющий степень близости эмпирического распределения бинарной переменной к биномиальному распределению, для которого частоты двух категорий равны.

Бонферрони критерий. Критерий для множественного сравнения средних, если в дисперсионном анализе получен значимый результат.

Вальда критерий. В модели логистической регрессии критерий значимости коэффициента B для соответствующего предиктора. Чем выше его значение (вместе с числом степеней свободы), тем выше значимость.

Вероятность. Ожидаемая относительная частота некоторого события.

Взаимодействие. Эффект совместного влияния на зависимую переменную двух и более независимых переменных, который не сводится к их раздельному влиянию. В случае двух независимых переменных проявляется в том, что эффект влияния одной из них проявляется по-разному на разных уровнях другой переменной.

Внутригрупповая сумма квадратов. Сумма квадратов отклонений наблюдаемых значений от среднего для каждой группы.

Внутригрупповой фактор. Фактор, уровням которого соответствуют повторные измерения (зависимые выборки).

Внутригрупповой эффект. Эффект внутригруппового фактора, уровням которого соответствуют повторные измерения (зависимые выборки).

Вращение. Процедура, применяемая в факторном анализе для того, чтобы получить более простую структуру факторов.

Выборка. Подмножество объектов из некоторой генеральной совокупности, выбранное для статистических выводов относительно свойств всей совокупности.

Гистограмма. Столбиковая диаграмма для отображения распределения частот по категориям (диапазонам значений) переменной. Горизонтальная ось графика соответствует значениям переменной, а вертикальная — частотам.

Главный эффект. Воздействие независимой переменной на зависимую переменную. Примеры главных эффектов можно найти в главах 14 и 15.

График собственных значений. Диаграмма, позволяющая выбрать число факторов в факторном анализе на основе критерия каменистой осыпи Р. Кеттелла.

Гуттмана критерий половинного расщепления. В анализе надежности половинного расщепления значение надежности, полученное с помощью процедуры нижних пределов.

Дендограмма. Диаграмма древовидной структуры, иллюстрирующая процесс кластеризации в кластерном анализе. Пример дендограммы приведен в главе 21

Детерминант ковариационно-дисперсионной матрицы. Величина, характеризующая степень зависимости между значениями переменных. Чем меньше значение детерминанта, тем сильнее соответствующая зависимость. Эта величина используется при вычислении М Бокса. Детерминант общей дисперсионно-ковариационной матрицы учитывает все матрицы, используемые в анализе.

Диаграмма рассеяния. График для анализа связи между двумя переменными, на котором каждый объект представляет собой точку. Положение точки задано парой значений двух переменных для данного объекта. Более подробное описание приведено в главе 9.

Диаграмма регрессии. Диаграмма разброса, включающая сдвиги точек от линии регрессии по вертикальной оси.

Дискриминантный анализ. Процедура создания формулы регрессии, на основе которой производится разбиение объектов на группы, соответствующие категориям зависимой переменной.

Дисперсионный анализ (ANOVA). Статистический анализ, устанавливающий статистическую значимость различий между средними значениями для трех или более выборок.

Дисперсии пунктов. Аналог средних значений элементов (пунктов шкалы) в анализе надежности. Поясняющий пример приведен в разделе «Представление результатов» главы 19.

Дисперсия. Характеристика выборочного распределения переменной, описывающая разброс значений вокруг среднего и вычисляемая как отношение суммы квадратов отклонений к объему выборки, уменьшенному на 1. Кроме того, дисперсия представляет собой квадрат стандартного отклонения.

Дисперсия шкалы, если элемент удален. Дисперсия суммы всех пунктов шкалы, кроме удаленного пункта.

Доверительный интервал. Диапазон, в котором находится большинство значений выборки. Например, термин «доверительный интервал в 95 %» означает интервал, в который любое случайное значение из выборки попадает с вероятностью 95 %.

Доверительный интервал в 95 %. См. *Доверительный интервал*.

Знаков критерий. Непараметрический критерий, определяющий различие двух измерений для одной выборки на основе знаков разностей пар значений.

Значимость (p -уровень). Мера случайности полученного результата, равная вероятности того, что в генеральной совокупности этот результат (различия, связь) отсутствует. Чем меньше эта вероятность (значение p -уровня), тем выше статистическая значимость результата. Результат считается статистически достоверным (значимым), если p -уровень не превышает 0,05.

Изменение R^2 . Изменение величины R^2 в результате введения новой переменной в уравнение регрессии.

Исправленная величина R^2 . Во множественном регрессионном анализе величина R^2 является точной для выборок, однако в генеральной совокупности ее значение лишь приблизительно. Исправленная величина R^2 представляет собой более точную оценку R^2 для генеральной совокупности и используется при сравнениях моделей, содержащих различное число независимых переменных.

Итерация. Стадия процесса формирования регрессионного (дискриминантного) уравнения, на которой происходит включение или исключение очередной переменной. Процесс продолжается до тех пор, пока не перестанет удовлетворяться заданный в процедуре критерий.

Кайзера–Мейера–Олкина критерий адекватности выборки. Величина, характеризующая степень применимости факторного анализа к данной выборке. Правило интерпретации этого критерия следующее: более 0,9 — безусловная адекватность; более 0,8 — высокая адекватность; более 0,7 — приемлемая адекватность; более 0,6 — удовлетворительная адекватность; более 0,5 — низкая адекватность; менее 0,5 — факторный анализ неприменим к выборке.

Канонические дискриминантные функции. Одно или более линейное дискриминантное уравнение, построенное таким образом, что классификация объектов по уровням зависимой переменной происходит наиболее точно. Более подробное описание вы можете найти в главе 22.

Канонические коэффициенты. В дискриминантном анализе канонический коэффициент представляет собой корреляцию между оценками дискриминантной функции и уровнями зависимой переменной. Более подробное описание вы можете найти в главе 22.

Категориальная (номинальная) переменная. Переменная, каждое значение которой указывает на принадлежность объекта к определенной группе (категории). Категориальная переменная не является количественной; она разделяет все объекты на непересекающиеся группы по определенному признаку (пол, хобби, класс и пр.), но не позволяет сравнивать объекты по уровню выраженности этого признака.

Квадрат евклидова расстояния. Мера, используемая по умолчанию в кластерном анализе для определения расстояния между объектами и кластерами и вы-

числяемая как сумма квадратов разностей между значениями переменных двух объектов.

Квадрат эта (η^2). Доля дисперсии зависимой переменной, обусловленная воздействием со стороны независимой переменной. Так, $\eta^2 = 0,044$ означает, что 4,4 % дисперсии зависимой переменной обусловлено данной независимой переменной.

Квадратная матрица. Матрица, строки и столбцы которой соответствуют одной и той же последовательности элементов (переменных или объектов).

Квартили. 25-й, 50-й и 75-й процентиля.

Кластерный анализ. Процедура, на основе заданного правила объединяющая объекты или переменные в группы, называемые кластерами.

Ковариата. Количественная переменная, имеющая значительную корреляцию с зависимой переменной и включаемая в анализ для более точной проверки воздействий факторов на зависимую переменную.

Количественная переменная. Значения количественной переменной (в отличие от категориальной) отражают уровень выраженности у объектов соответствующего признака в метрической или порядковой шкале.

Колмогорова–Смирнова критерий для одной выборки. Непараметрический критерий, определяющий, отличается ли данное эмпирическое распределение от теоретического распределения (нормального, равномерного, Пуассона или экспоненциального).

Контрасты. Метод контрастов — это метод множественного сравнения средних в дисперсионном анализе, который позволяет сравнивать выборки по градациям независимой переменной. Например, контрасты позволяет сравнивать одну градацию с другой, одну градацию со всеми остальными или разбить все градации на две группы и затем сравнить их между собой.

Корреляция. Мера степени и направления связи между значениями двух переменных. Понятию корреляции посвящена глава 9.

Корреляция между формами. В анализе надежности половинного расщепления приближенное значение надежности измерения в предположении, что обе половины содержат одинаковое число пунктов.

Корреляция между элементами. В анализе надежности это описательная информация о корреляциях каждого пункта с суммой всех остальных пунктов.

Коэффициент корреляции. Мера связи двух переменных, обозначаемая символом r и принимающая значения от -1 до $+1$.

Коэффициенты регрессии. B -коэффициенты, то есть множители при переменных, входящих в состав регрессионного уравнения, а также константа.

Критерий согласия. В логлинейном анализе критерий χ^2 для определения степени адекватности модели исходным данным. Чем выше его значения и чем ниже соответствующие уровни значимости, тем хуже модель соответствует данным.

Ливиния критерий. Критерий, предназначенный для проверки гипотезы о том, что все распределения зависимой переменной для сравниваемых выборок имеют одинаковые дисперсии.

Линия регрессии. Линия на графике двухмерного рассеяния, отражающая наиболее точные прогнозируемые значения («линия наилучшего соответствия»).

Логарифмический детерминант. В дискриминантном анализе натуральный логарифм определителя каждой ковариационной матрицы. Логарифмический детерминант используется для вычисления М Бокса.

Логит. Натуральный логарифм шанса. Описание понятия «логит» приводится в главе 24.

Лямбда Уилкса. Отношение внутригрупповой суммы квадратов к общей сумме квадратов, характеризующее дисперсию оценок дискриминантной функции, не обусловленную различиями между двумя группами. Единичное значение лямбда принимает в случае, если наблюдаемые средние значения групп равны; значения, близкие к нулю, означают, что внутригрупповая дисперсия мала по сравнению с общей дисперсией.

М Бокса (М). Критерий равенства дисперсионно-ковариационных матриц, основанный на близости значений определителей матриц ковариаций двух или более групп. Аналог критерия Ливина для многомерного случая.

Максимум. Наибольшее наблюдаемое значение распределения переменной.

Манна–Уитни и Уилкоксона критерий ранговых сумм. Непараметрический аналог t -критерия, определяющий различие между двумя выборками на основе рангов.

Матрица различий. Матрица, каждое значение которой соответствует различию между двумя объектами.

Матрица трансформации факторов. Результатом умножения этой матрицы и матрицы факторных нагрузок до вращения является матрица нагрузок факторов после вращения.

Медиана. Значение переменной, делящее упорядоченное множество всех значений выборки ровно пополам: у половины объектов выборки значения переменной больше, а у другой половины меньше медианы.

Медианный критерий k выборок. Непараметрический критерий для сравнения двух и более выборок по уровню выраженности признака; основан на подсчете числа элементов, лежащих выше и ниже главной медианы.

Межгрупповая сумма квадратов. Сумма квадратов разностей между главным средним значением и средними значениями групп, умноженных на весовые коэффициенты, равные числу объектов в соответствующих группах.

Метод главных компонент. Метод, применяемый SPSS по умолчанию в факторном анализе для извлечения факторов.

Метод иерархического слияния. Метод, используемый в кластерном анализе, при выполнении которого объекты объединяются в кластеры по одному на каждом шаге до тех пор, пока не будет образован единственный кластер, охватывающий все объекты. Кластерный анализ подробно описан в главе 21.

Метрическая переменная. Количественная переменная, соответствующая измерению признака в шкале интервалов или отношений. В отличие от ранговой (порядковой) переменной, при сравнении объектов позволяет судить не только о том, больше или меньше выражен признак, но и о том, насколько больше (меньше) он выражен.

Минимум. Наименьшее наблюдаемое значение распределения переменной.

Многомерные критерии значимости. В многомерном дисперсионном анализе набор критериев, позволяющих определить влияние факторов и их взаимодействий на совокупность зависимых переменных. Наиболее мощным считается критерий Пилая.

Многомерный дисперсионный анализ (ОЛМ-многомерная, MANOVA). Отличие многомерного дисперсионного анализа от одномерного (ANOVA) заключается в том, что число зависимых переменных в нем может быть теоретически любым.

Многомерный ковариационный анализ (MANCOVA). Многомерный дисперсионный анализ с включением в анализ ковариат.

Многомерный дисперсионный анализ с повторными измерениями (ОЛМ-повторные измерения). Вид дисперсионного анализа, в котором одна и та же группа объектов подвергается действию каждого уровня независимой переменной. С точки зрения вычислений этот анализ можно назвать внутригрупповым.

Многомерный критерий однородности матриц ковариаций. Критерий М Бокса определяет, являются ли ковариационные матрицы одинаковыми. Для каждого из значений вычисляется p -уровень, а также величина F или χ^2 .

Многомерное шкалирование. Метод, позволяющий на основе матрицы различий между объектами построить одно-, двух- или трехмерное изображение, иллюстрирующее удаленность этих объектов друг от друга.

Множественный регрессионный анализ. Метод, позволяющий спрогнозировать значения зависимой переменной на основе известных значений независимых переменных.

Мода. Наиболее часто повторяющееся значение распределения переменной.

Моучли критерий сферичности. Критерий многомерной нормальности. SPSS вычисляет приблизительное значение χ^2 и соответствующий уровень значимости. Если уровень значимости оказывается менее 0,05, то, вероятно, данные не являются нормально распределенными.

Наблюдаемое значение или частота. В χ^2 -анализе фактическая частота категории.

Надежность половинного расщепления. Мера надежности, для вычисления которой все пункты шкалы делятся на две эквивалентные группы, а затем на основе корреляции между двумя половинами шкалы устанавливается ее внутренняя согласованность.

Наименьшая ожидаемая частота. Наименьшая частота в ячейке таблицы сопряженности, определяемая в процессе применения критерия χ^2 .

Наименьшей значимой разности критерий (НЗР). Критерий множественного сравнения средних, представляющий собой серию t -критериев; применяется, если в дисперсионном анализе получен значимый результат.

Накопленная частота. Суммарное число объектов, имеющих значение переменной, не большее, чем указано.

Накопленный процент. Процент объектов от общего числа, имеющих значение переменной, не большее, чем указано.

Насыщенная модель. Логлинейная модель, включающая все взаимодействия и главные эффекты факторов.

Нелинейная регрессия. Процедура вычисления параметров нелинейного регрессионного уравнения.

Неортогональное вращение. Процедура, используемая в факторном анализе, допускающая результат, в котором угол между факторами отклоняется от прямого. Это иногда желательно для достижения более простой структуры.

Непараметрические критерии. Серия критериев, каждый из которых применяется без предварительных допущений относительно нормальности распределения. Непараметрические критерии основаны на ранжировании, попарных сравнениях и других средствах, не требующих нормальности распределения переменных.

Нестандартизованные коэффициенты канонической дискриминантной функции. Список коэффициентов и константа дискриминантного уравнения.

Номинальная шкала. См. *Категориальная переменная*

Нормальное распределение. Распределение частот (вероятностей), графически представляемое в виде симметричной кривой, имеющей пик в центре и асимптотически приближающееся к горизонтальной оси по краям. Идеальное нормальное распределение характеризуется нулевыми значениями асимметрии и эксцесса.

Общая внутригрупповая ковариационная матрица. Матрица, состоящая из средних значений ковариационных матриц, вычисленных для каждого уровня зависимой переменной.

Общая сумма квадратов. Сумма квадратов отклонений всех значений от среднего значения всего распределения.

Общность. В факторном анализе мера, характеризующая долю дисперсии переменной, обусловленную воздействием всех факторов.

Одномерные F-критерии. В многомерном дисперсионном анализе критерии, которые характеризуют влияние независимых переменных и их взаимодействий на каждую зависимую переменную в отдельности.

Ожидаемое значение. В перекрестной таблице при использовании критерия χ^2 значение, вычисляемое в предположении, что все переменные являются полностью независимыми друг от друга. В регрессионном анализе термин «ожидаемое значение» эквивалентен термину «прогнозируемое значение» и означает величину, получаемую для каждого объекта в результате подстановки значений переменных для него в уравнение регрессии.

Остатки и стандартизованные остатки. В логлинейных моделях остатки представляют собой разности между ожидаемыми и наблюдаемыми частотами. Чем выше значения остатков, тем менее адекватной является модель. SPSS подсчитывает исправленные величины остатков с использованием оценок стандартного отклонения. Исправленные остатки представлены в единицах нормального распределения, и, как правило, значения остатков, по модулю превышающие 1,96, являются значимыми.

Остаток. Как правило, разность между наблюдаемым и ожидаемым значениями. Эта величина относится к части дисперсии, которая не объясняется воздействием независимых переменных.

Отклонение. Расстояние и направление (отрицательное или положительное) между средним и данным значениями.

Оценка дискриминантной функции. Значение, получаемое для каждого объекта путем подстановки значений его переменных в уравнение дискриминантной функции.

Параметр. Некоторая числовая характеристика генеральной совокупности.

Параметрические критерии. Критерии, применяемые в предположении о нормальном распределении переменных в генеральной совокупности.

Переменные в уравнении. При выводе результатов пошагового регрессионного анализа SPSS включает для каждого шага статистики тех переменных, которые вошли в уравнение регрессии.

Пирсона коэффициент корреляции. Мера корреляции, идеально подходящая для двух непрерывных (метрических) переменных.

Порядковая (ранговая) шкала. См. *Ранговая переменная*.

Прямоугольная матрица. Матрица, для которой строкам и столбцам соответствуют разные последовательности элементов (объектов или переменных).

Размах. Характеристика распределения, равная разности между минимумом и максимумом распределения.

Ранговая (порядковая) переменная. Количественная переменная, отражающая измеренное качество на уровне порядка: в большей или меньшей степени оно вы-

ражено. В отличие от метрической шкалы, не позволяет судить о том, насколько больше или меньше выражено качество, поэтому не допускает применение арифметических операций.

Распределение. Статистическое понятие, обозначающее соотношение значений признака и частот (вероятностей) их встречаемости. Распределение (вероятностей, частот) может быть представлено в виде формулы для функции распределения вероятностей, графика распределения частот (гистограммы, столбиковой диаграммы), таблицы распределения частот.

Регрессионный анализ. Инструмент статистики, позволяющий прогнозировать значения зависимой переменной с помощью известных значений независимых переменных. Подробное рассмотрение регрессионного анализа проводится в главах 17 и 18.

Регрессия. В множественном регрессионном анализе этим термином обозначается статистика, отражающая влияние предикторов на зависимую переменную.

Серий критерий. Непараметрический критерий, определяющий, является ли последовательность бинарных величин (событий) случайной или упорядоченной.

Симметричная матрица. Квадратная матрица, для которой в каждой паре ячеек, расположенных симметрично относительно главной диагонали, содержатся одинаковые значения. Типичным примером симметричной матрицы является корреляционная матрица.

Скорректированная корреляция пункта и суммы. В анализе надежности корреляция между пунктом шкалы и суммой всех остальных пунктов.

Собственное значение. В факторном анализе эта величина пропорциональна доле дисперсии, обусловленной влиянием данного фактора; в дискриминантном анализе отношение межгрупповой суммы квадратов к внутригрупповой сумме квадратов. Чем больше собственное значение, тем выше точность дискриминантной функции.

Спирмена–Брауна критерий неэквивалентных форм. В анализе надежности половинного расщепления надежность, вычисленная для случая, когда «половины» имеют неравный размер.

Спирмена–Брауна критерий эквивалентных форм. Используется в анализе надежности, когда число элементов в «половинах» одинаково (вычисляется коэффициент корреляции).

Среднее шкалы при удалении пункта. В анализе надежности для каждого пункта шкалы вычисляется сумма остальных пунктов по всем объектам выборки; отношение указанной суммы к числу объектов является средним шкалы, если данный элемент удален.

Средние значения пунктов. В анализе надежности (с применением альфа Кронбаха) описательная информация, касающаяся средних значений пунктов шкалы по всем наблюдениям. Поясняющий пример приведен в разделе «Представление результатов» главы 19.

Средний квадрат. Отношение суммы квадратов к числу степеней свободы. В однофакторном дисперсионном анализе, как правило, средний квадрат вычисляется для внутригрупповой и межгрупповой сумм квадратов, а в регрессионном анализе — для регрессионной и остаточной сумм квадратов. Во всех перечисленных случаях средний квадрат используется для вычисления F -критерия.

Стандартизованный коэффициент α . В анализе надежности значение α , полученное в случае, если перед проведением анализа стандартизовать распределения всех элементов шкалы.

Стандартная ошибка. Стандартное отклонение величины, получаемое в результате ее многократного вычисления для случайных выборок. Как правило, стандартная ошибка вычисляется для среднего значения распределения.

Стандартное отклонение. Мера разброса значений распределения вокруг среднего. Стандартное отклонение определяется как квадратный корень дисперсии (суммы квадратов отклонений от среднего, деленной на $N - 1$, где N — объем выборки).

Статистики шкалы. В анализе надежности статистические характеристики суммы всех переменных по объектам.

Столбиковая диаграмма. График распределения частот по категориям (значениям) переменной. Каждый столбец на графике соответствует одному значению признака, а его высота пропорциональна частоте встречаемости этого значения. Аналогичное средство для количественных переменных, имеющих большое число возможных значений, обычно называется гистограммой.

Стресс. В многомерном шкалировании мера соответствия модели исходной матрице различий. Чем меньше значение стресса, тем лучше соответствие модели.

Сумма квадратов. Стандартная мера разброса, представляющая собой сумму квадратов отклонений всех значений величины от среднего.

Таблица распределения (частот). Таблица, устанавливающая соотношение между категориями (значениями) признака и частотами их встречаемости.

Таблица сопряженности (кросстабуляции). Обычно таблица совместного распределения частот для двух категориальных или дискретных переменных; строки соответствуют категориям (значениям) одной, а столбцы — другой переменной. Подробно таблицы сопряженности описаны в главе 8.

Толерантность. Мера линейной зависимости между одной переменной и набором других переменных. Если при дискриминантном анализе уровень толерантности составляет менее 0,001, это означает, что линейная зависимость для данной переменной настолько высока, что ее включение в дискриминантное уравнение недопустимо.

Тьюки критерий подлинной значимости. Критерий множественного сравнения, позволяющий попарно сравнивать средние значения; применяется, если дисперсионный анализ показал значимый результат.

Уилкоксона критерий. Непараметрический критерий, сходный критерием знаков, однако в отличие от последнего использующий ранги положительных и отрицательных разностей.

Фактор. В факторном анализе объединение нескольких переменных, чья взаимная корреляция исчерпывает определенную долю общей дисперсии. После процедуры вращения каждый фактор интерпретируется как некоторая общая причина взаимосвязи группы переменных.

Факторный анализ. Метод, позволяющий свести большое количество исходных переменных к значительно меньшему числу факторов, каждый из которых объединяет исходные переменные, имеющие сходный смысл.

Фи (φ). Мера связи (корреляции) двух категориальных переменных, обычно принимаемая наряду с критерием χ^2 при анализе таблиц сопряженности и вычисляемая по формуле:

$$\varphi = \sqrt{\chi^2/N}.$$

Фридмана дисперсионный анализ. Непараметрическая процедура, определяющая, различаются ли между собой три или более измерения для одной и той же выборки, на основе среднего ранга каждого измерения.

Хи-квадрат (χ^2) для модели. В анализе логистической регрессии величина, позволяющая определить, оказывают ли переменные, входящие в состав регрессионного уравнения, значимое влияние на зависимую переменную. Чем выше полученное значение, тем значительнее воздействие.

Хи-квадрат (χ^2) критерий для одной выборки. Непараметрический критерий, определяющий отличие наблюдаемого распределения переменной от ожидаемого (теоретического) распределения.

Хи-квадрат критерий (критерий χ^2). Непараметрический критерий для сравнения ожидаемых и наблюдаемых частот (как правило, для таблиц сопряженности). Критерий χ^2 может использоваться для оценки адекватности структурных и логлинейных моделей. В любом случае, χ^2 -анализ всегда отвечает на один и тот же вопрос: отличаются ли ожидаемые частоты модели от наблюдаемых. Коэффициент χ^2 Пирсона вычисляется по следующей формуле:

$$\chi^2 = \sum [(f_o - f_e)^2 / f_e].$$

Центроиды групп. В дискриминантном анализе средние значения дискриминантных функций для каждой из двух или более групп. Если число групп равно 2, центроиды будут иметь одинаковые абсолютные величины и разные знаки. Чем ближе объект в центроиду группы, тем больше вероятность, что он принадлежит к этой группе.

Частичный критерий χ^2 . Значение критерия χ^2 , характеризующее долю воздействия очередной независимой переменной на зависимую переменную.

Частота (абсолютная). Количество объектов в выборке, имеющих данное значение признака.

Частота относительная. Доля объектов в выборке, имеющих данное значение признака; равна отношению абсолютной частоты к объему выборки.

Число степеней свободы (df). Количество возможных направлений изменчивости статистического показателя, наряду с эмпирическим значением критерия служит для определения p -уровня значимости.

Шаговый отбор переменных. Процедура, включающая и исключающая переменные из дискриминантного или регрессионного уравнения в соответствии с выбранными критериями.

Шанс. Отношение вероятности того, что событие произойдет, к вероятности того, что событие не произойдет.

Шеффе критерий. Процедура, позволяющая осуществлять попарные множественные сравнения средних значений после получения статистически достоверного результата дисперсионного анализа.

Экспонента В. В логистическом регрессионном анализе величина eB используется в одной из форм регрессионного уравнения и позволяет создать одну из интерпретаций коэффициентов регрессии.

Эксцесс. Мера «сглаженности» («островершинности» или «плосковершинности») распределения. Если значение эксцесса близко к 0, это означает, что форма распределения близка к нормальному виду. Детальное описание эксцесса приведено в главе 7.

Эта (η). Мера корреляции между двумя переменными в случае, если одна из них является категориальной.

Англо-русский словарь терминов

В представленной ниже таблице перечислены основные термины, с которыми может встретиться пользователь программы SPSS (в частности, в окне вывода). Более подробную информацию по этим терминам можно найти в разделе «Представление результатов» каждой главы.

Термин	Перевод
Advanced models	Модуль дополнительных моделей
Agglomeration schedule	Шаги агломерации
Alignment	Выравнивание
Alpha	Альфа
Analysis Of Variances (ANOVA)	Дисперсионный анализ
Approximate significance	Приблизительная значимость
Asymptotic significance	Асимптотическая значимость
Bar graph	Столбиковая диаграмма
Bartlett's test of sphericity	Критерий сферичности Барлетта
Base (Statistics)	Основной системный модуль SPSS
Between-groups linkage	Межгрупповое связывание
Binomial test	Биномиальный критерий
Box plot	Коробчатая диаграмма
Box's M	М Бокса
Canonical analysis	Канонический анализ
Category (x) axes	Ось категорий (x)
Category axes label	Метка оси категорий
Category axes title	Заголовок оси категорий
Cell	Ячейка (таблицы)

Термин	Перевод
Central tendency	Центральная тенденция (мера)
Centroid	Центроид
Chi-square	Хи-квадрат
Chronbach's alpha	Альфа Кронбаха
Cluster analysis	Кластерный анализ
Cluster method	Метод кластеризации
Coefficient	Коэффициент
Commonality	Общность
Comparison	Сравнение
Complete linkage	Полная связь
Confidence interval	Доверительный интервал
Continuity correction	Поправка на непрерывность
Contrast	Контраст
Convergence	Сходимость
Correlation	Корреляция
Correlation matrix	Корреляционная матрица
Count measure	Количественный показатель (мера)
Covariance	Ковариация
Covariate	Ковариата
Cox & Snell R Square	R-квадрат Кокса и Снелла
Cramer's V	V Крамера
Criteria (criterion)	Условие
Crosstabulation (crosstab)	Таблица сопряженности
Cumulative frequencies	Кумулятивные (накопленные) частоты
Curve	Кривая (линия)
Curvilinear analysis	Криволинейный анализ

Термин	Перевод
Cut point	Точка деления
Data reduction	Сокращение данных
Degrees of Freedom (df)	Число степеней свободы
Deletion	Удаление (исключение)
Dendogram	Дендограмма
Density function	Функция плотности вероятности
Dependent sample	Зависимая выборка
Descriptive statistic	Описательная статистика
Determinant	Детерминант (определитель) матрицы
Deviation	Отклонение
Difference	Разность, различие
Dimension	Шкала
Directory	Каталог
Discriminant analysis	Дискриминантный анализ
Dispersion	Изменчивость
Dissimilarity	Различие
Distance	Расстояние
Distribution	Распределение
Distribution function	Функция распределения вероятности
Eigenvalue	Собственное значение
Enter	Включение
Epsilon corrected	С эpsilon-коррекцией
Equality of variances	Равенство дисперсий
Equal-length Spearman-Brown	Коэффициент эквивалентных форм Спирмена—Брауна
Equation	Уравнение
Error	Ошибка

Термин	Перевод
Error bar chart	Диаграмма столбцов ошибок
Eta Squared	Квадрат эта
Euclidian distance	Евклидово расстояние
Expected frequency	Ожидаемая (теоретическая) частота
Extraction method	Метод извлечения (факторов)
Factor	Фактор
Factor analysis	Факторный анализ
Factor score	Значение фактора
Factor score coefficient	Коэффициент значений факторов
Fisher's exact test	Точный критерий Фишера
Fixed factor	Фиксированный фактор
Footnote	Сноска
Frequency	Частота
Frequency table	Таблица частот
Friedman test	Критерий Фридмана
F-value	F-критерий
General linear model	Общая линейная модель
Graph	График
Grid line	Линия сетки
Group centroid	Центроид группы
Group membership	Принадлежность к группе
Grouping variable	Группирующая переменная
Guttman Split-Half	Критерий надежности половинного расщепления Гуттмана
Hierarchical cluster analysis	Иерархический кластерный анализ

Термин	Перевод
Histogram	Гистограмма
Homogeneity of variances	Гомогенность (однородность) дисперсий
Honestly Significant Difference (HSD)	Критерий подлинной значимости
Hotelling T-square	T-квадрат Хотеллинга
Icicle	Сосульчатая диаграмма
Image factoring	Факторный анализ образов
Improvement	Улучшение
Independent sample	Независимая выборка
Individual differences model	Модель индивидуальных различий
Initial Eigenvalue	Исходное собственное значение
Inner frame	Внутренняя рамка
Interaction	Взаимодействие
Inverse distribution function	Обратная функция распределения
Item	Пункт (элемент шкалы)
Item variance	Дисперсия пункта
Iteration	Итерация
Kaiser-Meyer-Olkin Measure of Sampling Adequacy	Критерий адекватности выборки Кайзера—Мейера—Олкина
Kendall's tau	Тау Кендалла
Kendall's tau-b	Тау-b Кендалла
KMO and Barlett test of sphericity	Критерии КМО и сферичности Барлетта
Kolmogorov-Smirnov test	Критерий Колмогорова-Смирнова
Kruskal-Wallis one-way analysis of variance	Однофакторный дисперсионный анализ Краскала—Уоллеса
Kurtosis	Экссесс
Legend	Легенда (условные обозначения)

Термин	Перевод
Legend label	Метка легенды
Level of measurement	Уровень (шкала) измерения
Levene's test	Критерий Ливиня
Likelihood	Правдоподобие
Line (style: dotted)	Пунктирная линия
Line (style: solid)	Сплошная линия
Line graph	Линейный график
Linear regression analysis	Линейный регрессионный анализ
Linear-by-linear association	Линейная связь между переменными
Linkage	Соединение, связь
Listwise	Построчно
Loading	Нагрузка (факторная, компонентная)
Logistic regression	Логистическая регрессия
Logit	Логит
Loglinear analysis	Логлинейный анализ
Mahalanobis distance	Расстояние Махаланобиса
Main effect	Главный эффект
Mann-Whitney U	Критерий U Манна–Уитни
Marker	Маркер
Matrix	Матрица
Mauchly's test of sphericity	Критерий сферичности Моучли
Maximum	Максимум
McNemar test	Критерий Мак-Нимара
Mean	Среднее значение

Термин	Перевод
Mean difference	Разность средних
Mean of Squares (MS)	Средний квадрат
Mean Rank	Средний ранг
Mean Square	Средний квадрат
Median	Медиана
Minimum	Минимум
Minimum expected count	Минимальная ожидаемая частота
Mode	Мода
Model selection	Подбор модели
Multidimensional scaling	Многомерное шкалирование
Multiple comparisons	Множественные сравнения (средних)
Multivariate Analysis Of Covariance (MANCOVA)	Многомерный ковариационный анализ
Multivariate Analysis Of Variances (MANOVA)	Многомерный дисперсионный анализ
Nagelkerke R Square	R-квадрат Нейджелкерка
No correlation	Отсутствие корреляции
Nonparametric test	Непараметрический критерий
Normal distribution	Нормальное распределение
Oblique rotation	Облическое (не ортогональное) вращение
Observed frequency	Наблюдаемая (эмпирическая) частота
One sample	Одна выборка
One-sample Kolmogorov-Smirnov test	Критерий Колмогорова—Смирнова для одной выборки
One-tailed significance	Односторонняя значимость
Outer frame	Внешняя рамка
Paired samples	Парные (зависимые) выборки
Pairwise	Попарно
Parametric test	Параметрический критерий

Термин	Перевод
Pareto chart	Парето-диаграмма
Partial correlation	Частная корреляция
Partial Eta Squared	Частичный квадрат Эта
Pattern	Образец
Pearson Chi-Square	Критерий хи-квадрат Пирсона
Pearson correlation	Корреляция Пирсона
Percentile	Процентиль
Perfect negative correlation	Строгая отрицательная корреляция
Perfect positive correlation	Строгая положительная корреляция
Phi	Фи
Pie chart	Круговая диаграмма
Pillai's trace	След Пиллая
Plot	Диаграмма
Principal axis factoring	Факторный анализ методом главных осей
Principle component analysis	Анализ главных компонент
Prior probability	Априорная вероятность
Probability	Вероятность
Probability of group membership	Вероятность принадлежности к группе
Proximity	Близость
Quadratic relationship	Квадратичная связь (зависимость)
Quantile	Квантиль
Quartile	Квартиль
R Square	Квадрат коэффициента корреляции (коэффициент детерминации)
R Square change	Изменение квадрата корреляции
Random factor	Случайный фактор
Range	Размах

Термин	Перевод
Rank	Ранг
Rectangular matrix	Прямоугольная матрица
Reference line	Вспомогательная линия
Regression coefficient	Коэффициент регрессии
Regression line	Линия регрессии
Regression model unit	Модуль регрессионных моделей
Related sample	Зависимая (связанная) выборка
Reliability analysis	Анализ надежности
Remove method	Метод исключения
Repeated-measures ANOVA	Анализ ANOVA с повторными измерениями
Residual	Остаток
Residual analysis	Анализ остатков
Rotation method	Метод вращения
Sample	Выборка
Saturated model	Насыщенная модель
Scale (y) axes	Ось значений (y)
Scale axes title	Заголовок оси значений
Scaling	Шкалирование
Scaling model	Модель шкалирования
Scatter plot	Диаграмма рассеяния
Scheffe test	Критерий Шеффе
Scree plot	График собственных значений («каменистой осыпи»)
Selection	Подбор
Sign	Знак
Sign test	Критерий знаков
Significance	Значимость

Термин	Перевод
Significance level	Уровень значимости
Similarity	Сходство
Skewness	Асимметрия
Solution	Решение
Source	Источник
Spearman's rho	Коэффициент корреляции Спирмена
Split-half reliability	Надежность половинного расщепления
Square asymmetric matrix	Квадратная асимметричная матрица
Square symmetric matrix	Квадратная симметричная матрица
Squared Euclidean distance	Квадрат Евклидова расстояния
Squared multiple correlation	Квадрат множественной корреляции
S-stress	s-стресс
S-stress convergence	Величина сходимости s-стресса
Standard (std.) error	Стандартная ошибка
Standard deviation (std. dev.)	Стандартное отклонение
Standard error kurtosis	Стандартная ошибка эксцесса
Standard error mean	Стандартная ошибка среднего
Standard error skewness	Стандартная ошибка асимметрии
Statistical Package for the Social Science (SPSS)	Статистический пакет для социальных наук
Statistical test	Статистический критерий
Stepwise method	Пошаговый метод
Stimulus	Стимул
Stress	Стресс
Subset	Подмножество
Subtitle	Подзаголовок
Sum	Сумма

Термин	Перевод
Sum of Squares (SS)	Сумма квадратов
System missing value	Системное пропущенное значение
Territorial map	Территориальная карта
Test	Статистический критерий
Test value	Заданное (тестовое) значение
Tick	Риска
Tied Ranks	Связанные (повторяющиеся) ранги
Title	Заголовок
Tolerance	Толерантность
Total	Итог
Tree diagram	Древовидная диаграмма
True value	Истинное значение
Tukey's test of additivity	Критерий аддитивности Тьюки
t-value	t-критерий
Two-tailed significance	Двусторонняя значимость
Unequal-length Spearman-Brown	Критерий неэквивалентных форм Спирмена—Брауна
Univariate procedure (approach)	Одномерный метод (подход)
Unstandardized coefficient	Нестандартизированный коэффициент
User missing value	Пользовательское пропущенное значение
Value	Значение
Value label	Метка значений
Variable (var)	Переменная (пер)
Variance	Дисперсия
Wald test	Критерий Вальда
Weight	Вес
Weirdness	Предсказуемость

Термин	Перевод
Wilcoxon test	Критерий Уилкоксона
Wilks' Lambda	Лямбда Уилкса
Within-groups linkage	Внутригрупповое связывание
Working directory	Рабочий каталог
Z-score	Z-значение (стандартизованное значение)

Литература

1. Айвазян С. А. и др. Прикладная статистика. Классификация и снижение размерности. М., 1989.
2. Боровиков В. Statistica. Искусство анализа данных на компьютере: Для профессионалов. СПб., 2003.
3. Гмурман В. Е. Теория вероятностей и математическая статистика. М., 1997.
4. Гублер Е. В., Генкин А. А. Применение непараметрических критериев статистики в медико-биологических исследованиях. Л., 1973.
5. Гусев А. Н. Дисперсионный анализ в экспериментальной психологии: Учебное пособие для студентов факультетов психологии. М., 2000.
6. Дэвисон М. Многомерное шкалирование: методы наглядного представления данных. М., 1988.
7. Закс Л. Статистическое оценивание. М., 1976.
8. Ибберла К. Факторный анализ. М., 1980.
9. Кендалл М., Стьюарт А. Статистические выводы и связи. М., 1973.
10. Митина О. В., Михайловская И. Б. Факторный анализ для психологов. М., 2001.
11. Наследов А. Д. Математические методы психологического исследования. Анализ и интерпретация данных. СПб., 2008.
12. Рунион Р. Справочник по непараметрической статистике. М., 1982.
13. Сидоренко Е. В. Методы математической обработки в психологии. СПб., 1996.
14. Справочник по прикладной статистике. В 2-х т. / Под ред. Э. Ллойда, У. Ледермана, Ю. Тюрина. М., 1989.
15. Суходольский Г. В. Основы математической статистики для психологов. СПб., 1998.
16. Суходольский Г. В. Математические методы психологии. СПб., 2003.
17. Терехина А. Ю. Анализ данных методами многомерного шкалирования. М., 1986.

18. Факторный, дискриминантный и кластерный анализ / Дж.-О. Ким, Ч. У. Мьюллер, У. Р. Клекка и др. М., 1989.
19. Харман Г. Современный факторный анализ. М., 1972.
20. Agresti, Alan (1996). An Introduction to Categorical Data Analysis. New York: Wiley.
21. Bishop, Y. M. M., Fienberg, S. E., & Holland, P. W. (1975). Discrete multivariate analysis. Cambridge, MA: MIT Press.

Андрей Наследов
SPSS 19: профессиональный статистический анализ данных

Заведующий редакцией
Руководитель проекта
Ведущий редактор
Литературный редактор
Художественный редактор
Корректор
Верстка

*А. Кривцов
А. Юрченко
Ю. Сергиенко
О. Некруткина
А. Татарко
В. Листова
Е. Неволainen*

ООО «Мир книг», 198206, Санкт-Петербург, Петергофское шоссе, 73, лит. А29.
Налоговая льгота — общероссийский классификатор продукции ОК 005-93, том 2; 95 3005 — литература учебная.
Подписано в печать 17.01.11. Формат 70х100/16. Усл. п. л. 33,540. Тираж 2000. Заказ 0000.
Отпечатано по технологии СtР в ОАО «Печатный двор» им. А. М. Горького.
197110, Санкт-Петербург, Чкаловский пр., 15.