И.Н. Дубина

МАТЕМАТИЧЕСКИЕ ОСНОВЫ ЭМПИРИЧЕСКИХ СОЦИАЛЬНО-ЭКОНОМИЧЕСКИХ ИССЛЕДОВАНИЙ

Учебное пособие

Издательство Алтайского университета

Барнаул 2006

УДК 330.105

ББК 65в631я73

Д 792

Рецензенты:

доктор экономических наук, профессор, зав. кафедрой информационных систем в экономике, декан экономического факультета Алтайского государственного университета **О.П. Мамченко;**

доктор социологических наук, профессор кафедры социальной антропологии и социальной работы Саратовского государственного технического университета **В.К. Антонова**

Д 792 Дубина И.Н.

Математические основы эмпирических социально-экономических исследований : учебное пособие. – Барнаул : Изд-во Алт. ун-та, 2006. – 263 с.

ISBN 5-7904-0553-3

В пособии рассмотрены основы теории социально-экономических измерений, представлены подходы к проектированию измерительных шкал, проверке и обоснованию надежности результатов измерения, применению математико-статистических методов обработки и анализа эмпирических данных и проверки исследовательских гипотез. Представленный в книге материал служит для формирования, развития и закрепления навыков использования специальных методов и прикладного программного обеспечения для математически корректного анализа результатов количественных исследований в экономических и социальных областях.

Рекомендуется в качестве учебного пособия для студентов, магистрантов и аспирантов экономических и социологических специальностей вузов. Книга может быть полезна также студентам и аспирантам гуманитарных специальностей, широкому кругу исследователей и практиков, использующих в своей работе методы теории измерений и статистического анализа.

ISBN 5-7904-0553-3

Издание учебного пособия осуществлено по гранту, выделенному в рамках Программы инициативных грантов для выпускников обменных образовательных программ, финансируемой Бюро по культуре и образованию Государственного департамента США и администрируемой Американскими Советами по международному образованию.

Support for this publication was provided by the Alumni Local Initiative Grants Program, which is funded by the Bureau of Educational and Cultural Affairs of the United States Department of State (ECA), under authority of the Fulbright-Hays Act of 1961 as amended, and administered by American Councils for International Education: ACTR/ACCELS. The opinions expressed herein are the author's own and do not necessarily express the views of either ECA or American Councils.

Подготовка материалов учебного пособия частично осуществлялась в рамках проекта «Научное сообщество», финансируемого Фондом Форда и администрируемого образовательно-исследовательским центром Con-text (Томск), а также при поддержке со стороны экономического факультета Алтайского государственного университета.

При подготовке 2 и 4 глав книги частично использованы результаты исследований, выполненных при финансовой поддержке Российского гуманитарного научного фонда (проект №06-02-60201 а/Т).

Оглавление

Введение

Термины и понятия из областей статистики и теории измерений являются составляющей частью профессионального тезауруса многих социологических и экономических дисциплин. Получить и обработать необходимую информацию, используемую в теории и практике социально-экономических наук, можно лишь на основе *грамотного* применения статистических подходов и методов. Хотя статистика является обязательной дисциплиной для социологических и экономических специальностей в вузе, студенты, изучающие этот курс, имеют часто лишь самое общее представление о возможностях статистического анализа. В частности, в рамках базовых курсов по математике и статистике, как правило, не рассматриваются (или рассматриваются на самом вводном уровне) специальные математико-статистические методы и программное обеспечение для проектирования экспериментальных исследований и обработки данных. Для сравнения, многие американские и европейские университеты предлагают специальные курсы по основам теории измерений в гуманитарной и социальной сферах, методам проектирования измерительных шкал, количественным методам исследования бизнеса и социально-экономических процессов.

У студентов российской высшей школы часто отсутствуют необходимые знания и навыки социально-экономических измерений, проектирования исследований и статистической обработки данных. Результаты, представляемые в некоторых дипломных и даже диссертационных работах, а также публикациях, часто являются не только малообоснованными, но даже полностью лишенными смысла с математической точки зрения.

Методы и инструменты, используемые российскими молодыми исследователями, часто не соответствуют современным научным стандартам, и это обстоятельство является одной из основных причин того, что они плохо «вписываются» в мировой научный дискурс, а работы российских ученых почти не представлены в международных нетехнических научных журналах. Учебное пособие «Математические основы эмпирических социально-экономических исследований» направлено на частичное исправление указанных пробелов в образовании и подготовке специалистов высшей квалификации.

Среди исследователей, к сожалению, распространено убеждение, что статистические методы нужны только для обработки результатов наблюдений. В данном учебном пособии акцент сделан на применении математико-статических методов на этапах планирования эксперимента и оценки качества полученных исследовательских результатов. Именно в этих

областях у многих студентов и молодых исследователей, даже имеющих солидную математико-статистическую подготовку, отсутствуют необходимые знания.

Заметную часть данного учебного пособия составили материалы дистанционно-очного учебного курса «Математические основы проектирования и анализа результатов эмпирических социально-экономических и гуманитарных исследований», который автор разрабатывал и проводил в рамках проекта «Научное сообщество», финансируемого Фондом Форда и администрируемого образовательно-исследовательским центром Con-text, в 2006 г. Опыт проведения данного курса свидетельствует о том, что знания методов обработки статистической информации являются весьма востребованными в общественных, экономических и гуманитарных исследовательских областях. Это, в частности, демонстрируют результаты конкурсного отбора участников учебного курса. В период с 15 февраля по 15 марта 2006 г. на конкурс для участия в курсе было направлено 227 заявок из 90 вузов и организаций из 56 городов России, Казахстана и США. Заявки на участие в курсе направлялись не только студентами, магистрантами и аспирантами (как это ожидалось исходя из указанной в проекте целевой группы), но и уже состоявшимися и квалифицированными исследователями – кандидатами и докторами наук, доцентами, профессорами, старшими научными сотрудниками. После прохождения курса в отзывах его участников отмечалась полезность курса для выполнения квалификационных работ и конкретных исследований, проведения лекций и семинаров. Количество заявок, статус заявителей, положительные отзывы участников и экспертов можно рассматривать в качестве индикаторов необходимости специального учебного курса по данной тематике. Это обстоятельство и послужило причиной решения автора расширить и углубить содержание материалов курса и оформить их в виде отдельного учебного пособия.

Учебное пособие направлено на формирование основ статистической грамотности в тех областях, которые, как правило, остаются за пределами базовых вузовских курсов по математике и статистике, а также большинства учебников и учебных пособий. В пособии рассматриваются темы, слабо или совсем не отраженные в учебной литературе. Особое внимание уделяется математическим основам проектирования количественных социально-экономических измерений: подробно рассматриваются виды измерительных шкал, методы калибровки и конвертирования данных из неметрических в метрические шкалы (в том числе с использованием модели Г. Раша), подходы и методы для оценки точности и надежности социально-экономических измерений (с использованием коэффициента альфа Кронбаха, коэффициентов Гутмана и других критериев и показателей). Также более обстоятельно, чем в большинстве учебников и учебных пособий, рассматриваются вопросы статистической проверки исследовательских гипотез, разбираются особенности использования

статистических критериев и методов для такой проверки и получения корректных (с математической точки зрения) результатов.

Данное учебное пособие не охватывает всего многообразия существующих методов, применяемых в социально-экономических исследованиях. За рамками учебного пособия мы оставляем материал, подробно рассматриваемый в доступной учебной литературе, в частности корреляционный и регрессионный анализ. Но при этом предполагается, что читатели уже имеют представление об основах корреляционного и регрессионного анализа, а также базовые знания из области теории вероятностей и математической статистически (например, о видах распределений, статистических показателях вариации и др.).

Большое внимание уделяется работе с программными продуктами для обработки и анализа исследовательских данных WINSTEPS, STATISTICA, SPSS. Помимо специальных статистических пакетов, в книге приводится описание возможности использования основных статистических функций Excel, которые могут быть применены для решения относительно простых задач статистического анализа.

В книге приводится много простых примеров и решений для иллюстрации «внутренних» механизмов работы тех или иных процедур для анализа исследовательских результатов. Это необходимо, поскольку работа с современными статистическими пакетами, реализующими эти процедуры, и, что самое важное, корректная интерпретация полученных результатов требуют ясного понимания теоретических основ применяемых процедур. В частности, необходимо четко понимать, что означают те или иные показатели, как они связаны друг с другом, на каких принципах основано их получение, с какими предпосылками и ограничениями они связаны. После рассмотрения всех этих моментов с иллюстрацией на сравнительно простых примерах приводится описание того, как аналогичную, но более сложную задачу можно решить с помощью того или иного программного продукта.

Таким образом, эта книга ориентирована на читателей, уже имеющих начальные знания в области математической статистики, хотя для освоения материала, представленного в пособии, не требуется основательной математической подготовки. Также предполагается, что читатели книги имеют базовые навыки работы на компьютере с общепользовательскими и прикладными программами, указываются литература и ресурсы, в которых разбирается материал, не вошедший в данное учебное пособие, но используемый в упражнениях и изложении учебного материала. Автор старался не перегружать изложение материала абстрактными формулами, теориями и выводами в целях большей доступности, однако в то же время избегать тривиализации статистических подходов и выводов.

Учебный материал в книге подается структурированно, ко многим задачам и упражнениям даны комментарии, поэтому материал можно изучать самостоятельно. Поскольку современные технологии сделали бессмысленным проведение рутинных и затратных по времени процедур по статистической обработке данных «вручную», в качестве практических задач предлагаются сравнительно небольшие массивы данных. Многие упражнения, приведенные в пособии, основаны на реальном исследовательском материале.

Автор выражает надежду, что эта книга поможет читателям повысить уровень исследовательской квалификации, получить знания и навыки по проектированию измерительных шкал для проведения социально-экономических исследований, оценке точности и надежности социально-экономических измерений, использованию математико-статистических методов анализа результатов исследования и проверки исследовательских гипотез.

Индикативное задание

Умение хорошо понимать статистические материалы (исследования, отчеты и т.п.) и разбираться в них необходимо для любого статистически грамотного исследователя. В определенной степени эта книга направлена также и на развитие данного качества, поэтому читателю сразу же предлагается первое задание.

Возможно, у читателя возникло некоторое недоумение от того, что непосредственно за введением предлагается учебное задание. Это действительно не совсем привычная логика построения учебного пособия, ведь учебник, как правило, начинается с изложения материала по некоторой теме, после чего и следуют задания. Но «классические» традиции построения учебного пособия здесь почти не нарушены. Дело в том, что это задание является ознакомительным и в определенном смысле индикативным.

Читателю предлагается найти любую статью, связанную с использованием математико-статистического аппарата для обработки данных эмпирических исследований в социальной или экономической сфере (желательно в области профильной специальности и (или) исследовательских интересов читателя). Такую статью можно найти, например, в журналах «Экономическая социология», «Экономика и математические методы», «Социология: методология, методы, математические модели», «Социологические исследования» и других специальных периодических изданиях. В разделе «Библиографический список» этой книги указаны адреса интернет-сайтов, на которых представлены содержание выпусков этих журналов, аннотации и полнотекстовые варианты многих статей. Исследования с использованием статистических методов и моделей также содержатся на сайтах «Статистические методы», «Образовательный портал Экономика, Социология, Менеджмент: Методы сбора и анализа социологических данных» и других сайтах, адреса которых также указаны в разделе «Библиографический список».

После выбора статьи и ознакомления с текстом попытайтесь ответить на приведенные ниже вопросы и выполнить простые задания. Возможно, в тексте нет прямых ответов на вопросы, которые Вам предлагаются. В таком случае попытайтесь восстановить логику исследователя по косвенным признакам и *предположить* возможные ответы.

Для выполнения этого задания не обязательно предварительно обращаться к каким-либо учебным материалам. Не стоит беспокоиться, если содержание статьи Вам не очень ясно или не понятны вопросы, на которые Вам предлагается ответить после прочтения статьи. Основная цель этого задания – самоопределение исходного уровня понимания академического текста, описывающего использование подходов статического анализа. Поэтому данное задание и названо индикативным. Автор советует читателю сохранить

статью, а также записать и сохранить ответы на предлагаемые вопросы, а после освоения всех разделов книги вернуться к этой же статье и вновь ответить на вопросы данного задания. Так Вы сможете сравнить ваши ответы с ответами, которые Вы давали до ознакомления с материалами, представленными в этой книге. Изменились ли ваши ответы и ваше понимание статьи после работы с книгой?

Впрочем, данное задание имеет не только индикативный характер. Разбираться с чужим текстом гораздо сложнее, чем со своим, тем более восстанавливать логику авторов статьи и находить ответы на предложенные вопросы. Поэтому автор книги советует читателю при ответе на вопросы пытаться аргументировать ответы, а не ограничиваться простым «да-нет-не знаю». Неважно, «правильным» будет обоснование или нет, главное на этом этапе – попытка понимания и осмысления подходов и методов для решения задач, рассматриваемых в статье.

Вопросы и задания по статье

- Основано ли исследование, представленное в статье, на выборке из генеральной совокупности?

  Если да, то

  а) каков размер выборки?

  б) какова репрезентативность выборки?

  в) каким образом формировалась выборка?

  г) можно ли обобщить полученные результаты на генеральную совокупность?
- Какие переменные использовались? Какие из них являются зависимыми, а какие независимыми?
- Какие измерительные шкалы применяются в данном исследовании?
- Какие методы и подходы к обработке данных используются авторами?
- Какие статистические гипотезы проверяются и какими методами?
- Укажите несколько (до 10) терминов, встречающихся в статье и непонятных для Вас (если такие есть):

- Оцените, насколько понятна Вам терминология, используемая в статье

  полностью понятна **0**
  в основном понятна **0**
  отчасти понятна **0**
  плохо понятна **0**
  совсем не понятна **0**

- Оцените, насколько Вам понятны используемые в статье математические символы и формулы

  полностью понятны **0**
  в основном понятны **0**
  отчасти понятны **0**
  плохо понятны **0**
  совсем не понятны **0**

- Насколько Вам понятны основные результаты и выводы исследования, представленного в статье?

  полностью понятны **0**
  в основном понятны **0**
  отчасти понятны **0**
  плохо понятны **0**
  совсем не понятны **0**

- Насколько Вы знакомы со статистическими подходами и методами, которые описаны и/или используются в статье?

  я их постоянно использую **0**
  хорошо знаком (а) **0**
  отчасти знаком (а) **0**
  плохо знаком (а) **0**
  совсем не знаком (а) **0**

Часть 1. Проектирование и оценка качества социально-экономических измерений

Если что-либо существует, оно может быть измерено.
Э. Торндайк

Точно так же, как все искусства стремятся к музыке,

так и все науки стремятся к измерениям.
Дж. Сантаяна

Не все то, что может быть сосчитано, может быть измерено,

и не все то, что может быть измерено, может быть сосчитано.
А. Эйнштейн

Цель расчетов – не числа, а понимание.
Р. Хэмминг

Глава 1. Измерение и типы измерительных шкал

1.1. Число, измерение и измерительная шкала

Высказывания, вынесенные в качестве эпиграфа к первой части этой книги, в определенной степени отражают проблему понимания измерения и его роли в научном исследовании. Измерение, понимаемое прежде всего как численное выражение некоторого свойства, традиционно рассматривалось в качестве необходимого и важнейшего компонента научного исследования. Более того, измеряемость часто выступала критерием научности вообще: если мы не можем выделить свойства объектов, измерить их признаки и установить на этой основе причинно-следственные связи, мы не можем говорить о научности и объективности исследования. Причем этот принцип распространялся не только на естественно-научные, но и на социальные и гуманитарные области исследования. Иллюстрацией к такому подходу могут служить знаменитые высказывания Э. Торндайка и Дж. Сантаяны, приведенные в качестве эпиграфа. В то же время в современной науке сформировалось осознание сложности измерительных процедур и несводимости их к простому численному представлению свойств исследуемых объектов. Это понимание выразил А. Эйнштейн фразой, также приведенной в эпиграфе.

В широком смысле слова любое исследование – это процесс по сбору и анализу информации для формулировки и проверки идей и выводов. Статистические подходы применимы для исследований, в которых используются *данные*, т.е. информация, представимая в числовом виде. При всей тривиальности этих положений следует понимать, что недостаточно просто собрать данные и никакие данные не говорят сами за себя.

Более того, число обладает некоторой «магической силой» для исследователя: число как бы «само по себе» *a priori* означает и даже подчеркивает объективность исследования, а иногда и направляет порядок действия исследователя. Ведь всем знакомы элементарные арифметические операции, и у начинающих исследователей часто возникает соблазн получить некий результат, «что-нибудь сложив, разделив или умножив».

Иллюстрацией к сказанному является история (наверняка, придуманная), которую автор книги услышал на лекциях профессора Школы бизнеса Университета Дж. Вашингтона (США) Э. Формана. За этой забавной историей, возможно, стоит и реальный случай. Некий начинающий маркетолог, оценивая привлекательность населенных пунктов для строительства горнолыжных курортов в регионе, использовал 3 критерия для расчета «интегрального показателя привлекательности». В результате «исследования» был выявлен «лучший» пункт с максимальным рейтингом:

Население, чел.	3012
Высота над уровнем моря, м	1753
Дата основания, год	1925
Итого:	**6690**

Комментировать эти результаты нет необходимости – их абсурдность очевидна. Но подобные анекдоты, как это ни печально, встречаются и в реальной исследовательской практике. Профессор Форман приводит в своей книге (Forman, 2001) следующий поучительный случай. В 1990 г. в США был опубликован рейтинг мест отдыха пенсионеров, в котором по 7 критериям был ранжирован 151 город, и в результате суммирования получен некий итоговый показатель. На основе примененной методики Лас-Вегас оказался на 105 месте. В 1994 г. была использована иная процедура сбора и обработки данных. В результате Лас-Вегас оказался первым в рейтинге. Безусловно, причина столь радикальных перемен – не в изменении условий отдыха в Лас-Вегасе, а в изменении методологии измерений!

Основное искусство[1] статистического анализа заключается в получении и интерпретации данных. Получая данные, исследователь приписывает изучаемому объекту или событию числа по некоторому правилу, т.е. он так или иначе использует ту или иную процедуру измерения. Объектами исследования могут быть города, страны, машины, детали, люди, явления и т.д. Предметом изучения являются свойства этих объектов или событий.

[1] Здесь слово «искусство» употреблено в смысле греч. *techne*, т.е. скорее как *технология*, умение применять определенные правила.

Например, в качестве свойства человека (группы людей, города, страны) может выступать богатство. Исследователь фиксирует выраженность интересующих его свойств изучаемых объектов с помощью обнаружения некоторых признаков (например количества денег). Степень выраженности свойств может (но не обязательно) характеризоваться числом. Например, богатство можно определить как наличие или отсутствие денег (что может кодироваться, например, 0 и 1), а также их количество.

Один из распространенных, хотя и не бесспорных, подходов к пониманию измерения можно сформулировать так: *Измерение* – это кодирование и соотнесение степени выраженности признаков эмпирических объектов или событий с помощью чисел в соответствии с определенными правилами (*измерительными шкалами*). Очень важно понимать, что эти правила измерения определяют как результат измерения, так и исследовательский вывод (сравните результаты в примере о Лас-Вегасе).

Измерительная шкала определяет тип получаемых данных и множество операций, которые можно с этими данными осуществлять. *Измерительная шкала* – это математическая характеристика переменной, определяемая процессом измерения и типом измеряемой переменной, являющаяся основным критерием для выбора соответствующего математико-статистического метода. *Используемая измерительная шкала является основным элементом при обосновании возможных методов обработки данных и интерпретации результатов.* Прежде чем использовать ту или иную математическую операцию (пусть даже из области элементарной арифметики, не говоря уже о сложных статистических методах), очень важно понимать, позволяет ли выбранная шкала осуществлять такую операцию, проведено ли измерение таким способом, для которого возможно *корректное* использование данной операции.

Вопрос о типе используемой шкалы особо важен в социально-экономических и гуманитарных исследованиях, поскольку многие измерения в этих сферах являются косвенными. Непосредственно мы фиксируем некоторые признаки (время решения задачи, ответы на определенные вопросы, оценки респондентов и т.д.) и на этой основе делаем выводы о некотором скрытом (*латентном*) свойстве, недоступном прямому измерению (например, коммуникативность, креативность, организационный климат для творчества и инноваций и т.п.).

Определение шкалы измерений является первым этапом любого эмпирического исследования. К сожалению, многие исследователи проходят этот этап «автоматически», совсем не задумываясь об этом или руководствуясь лишь «здравым смыслом». Именно поэтому первая и вторая главы этой книги посвящены измерительным шкалам.

1.2. Типы измерительных шкал

Выделяют три основных атрибута измерительных шкал, наличие или отсутствие которых определяет принадлежность шкалы к той или иной категории.

Упорядоченность шкалы означает, что одна позиция шкалы, определяемая числом и соответствующая выраженности измеряемого свойства, больше, меньше или равна другой позиции.

Интервальность шкалы означает, что интервалы между позициями шкалы равны между собой.

Нулевая точка (или точка отсчета) шкалы означает, что набор чисел, соответствующих выраженности измеряемого признака, имеет точку отсчета, обозначаемую за 0, которая соответствует полному отсутствию измеряемого свойства.

Наиболее общей классификацией измерительных шкал является разделение на *неметрические шкалы* (в которых единица измерения отсутствует) и *метрические шкалы* (в которых может быть установлена единица измерения, например, килограмм, градус, рубль и т.д.). Метрические шкалы являются более *мощными*, т.е. они лучше дифференцируют измеряемые объекты (отражают больше информации о различии измеряемых объектов).

К неметрическим шкалам относят номинативные и порядковые шкалы.

В *номинативной шкале* (*nominal scale*) отсутствуют все главные атрибуты измерительных шкал, а именно упорядоченность, интервальность, нулевая точка. Для обозначения такой шкалы также используются термины *шкала наименований* и *номинальная шкала*.

Номинативная шкала используется для классификации или идентификации объектов (группировки по классам, каждому из которых приписывается число). Объекты группируются по классам таким образом, чтобы внутри класса они были идентичны по измеряемому свойству. Это самая простая шкала из тех, что могут рассматриваться как измерительные, хотя фактически эта шкала не ассоциируется с измерением и не связана с понятиями «величина» и «число». Она используется только с целью отличить один объект от другого.

Примером свойства, измеряемого в такой шкале, может являться пол. Эта переменная имеет две категории (наименьшее количество категорий, которое может иметь измеряемая переменная). Это свойство может быть измерено отнесением человека (объекта) в соответствующую категорию при подсчете количества мужчин и женщин. Категория «мужчины» может, например, кодироваться цифрой «1», а категория «женщины» – цифрой «2». В результате можно сделать заключение о том, к какой категории отнесено больше

объектов, а к какой меньше. Однако на основании этих данных нельзя произвести логические или математические операции над самими категориями, т.е. для такой шкалы нельзя, например, сказать, что 1 меньше, чем 2, или что 1 + 2 = 3. Другими примерами свойств, измеряемых в номинативной шкале, являются семейное положение, национальность, религиозная принадлежность, место рождения.

Категории переменных, измеряемых в номинативной шкале, должны быть взаимоисключающими, чтобы не было неопределенности относительно классификации исследуемых объектов. Эти категории должны также быть исчерпывающими, т.е. каждому возможному значению должна однозначно соответствовать определенная категория (хотя бы категория *другие* или *прочие*). Категории данных в номинативной шкале должны включать сопоставимые экземпляры.

Категории измеряемой переменной фактически не являются численными, и их можно сравнивать друг с другом только на основе количества наблюдений, относящихся к ним. При использовании номинативных шкал цифры, кодирующие соответствующие категории, являются лишь метками (именами) и не несут в себе численного значения. Как мы показали на примере, к данным, представленным в такой шкале, арифметические операции не применимы: их нельзя складывать, вычитать, умножать или делить. Однако разработаны специальные процедуры обработки и анализа данных в номинативных шкалах (например критерий χ^2 для статистической проверки гипотез). Они основаны на определении частоты (подсчете количества появлений переменной в различных категориях).

В *порядковой шкале* (*ordinal scale*) присутствует упорядоченность, но отсутствуют атрибуты интервальности и нулевой точки. Для обозначения такой шкалы также используются термины *ранговая шкала* и *шкала рангов*. Порядковая шкала является наиболее распространенной в социальных и гуманитарных исследованиях.

Результатом измерений в порядковой шкале является упорядочение объектов. Шкала ранжирует объекты, приписывает им числа в зависимости от выраженности измеряемого свойства по некоторому признаку (в порядке убывания или возрастания). В отличие от номинативной шкалы здесь можно не просто определить, что один объект отличен от другого, но и то, что по определенному признаку один объект больше или меньше другого. Следовательно, шкала показывает, больше или меньше выражено свойство (измеряемая величина), но не насколько больше или насколько меньше оно выражено, а тем более – во сколько раз больше или меньше.

Приведем несколько примеров измерений в порядковой шкале.

Пример 1. Социально-экономический статус определяется в следующих категориях: 1) «верхний класс»; 2) «средний класс»; 3) «низший класс». Эти категории ранжируются,

например, по признаку уровня дохода. Понятно, что в данной шкале значение 1 больше, чем 2, но непонятно, насколько. То же самое можно сказать о должностном статусе, например: специалист, ведущий специалист, главный специалист, заместитель начальника отдела, начальник отдела и т.д.

Пример 2. Четверым гонщикам присвоены «места» в соответствии с тем, кто раньше достиг финиша:

Гонщик	Место
A	1
B	2
C	3
D	4

Основываясь только на этих данных, можно судить лишь о том, кто раньше приехал, а кто позже. Нельзя сказать, *насколько* один из гонщиков приехал раньше, чем другой. Более того, совершенно неочевидно, что разрыв между гонщиками, например, A и B меньше, чем B и D, хотя 2 – 1 = 1 и 4 – 2 = 2. Гонщик A мог, например, достичь финиша на 20 мин раньше B, от которого D отстал всего на 1 мин.

Пример 3. Выступления спортсменов-фигуристов, как правило, оцениваются по шкале от 1 до 10. При этом оценки, например, от 3 до 5 встречаются гораздо реже, чем оценки от 7 до 9, а наиболее значимыми оценками являются оценки от 9 до 10. Очевидно, что интервал между 9,9 и 10,0 для судей и для спортсменов гораздо больше, чем интервал между 4,9 и 5,0.

Пример 4. В школе и в вузе используется 5-балльная система оценки знаний (1, 2, 3, 4, 5). Можем ли мы ответить, *насколько* 5 «лучше», чем 4? Настолько же, насколько 4 «лучше», чем 3? Возможно, для одного студента разница между 5 и 4 невелика, а разница между 4 и 3 значительна. Возможно, эта разница изменяется от курса к курсу. Если считать, что оценки учеников или студентов отражают их знания, то в случае применения к этой шкале действий арифметики получается, что знания отличника равны сумме знаний двоечника и троечника (5 = 2 + 3), а между знаниями отличника и троечника такая же разница, как между знаниями «ударника» и двоечника (5 – 3 = 4 – 2).

Таким образом, мы не можем сказать, что в такой шкале сохраняется равенство интервалов между пунктами. Сложение, как и большинство других математических операций, предполагает, что интервалы между значениями равны. Если интервалы не равны, то 2 + 2 может означать 5 или 55. Поэтому широко распространенная процедура усреднения оценок (баллов), т.е. вычисления среднего арифметического – это операция, лишенная смысла с математической точки зрения.

Таким образом, возможности статистического анализа для данных в порядковых шкалах ограничены. Строго говоря, упорядочение – это единственная математическая операция, применимая к порядковой шкале. Использование многих операций с порядковыми

данными (например вычисление среднего арифметического) математически некорректно, но широко распространено на практике. Например, в рейтинге мест отдыха американских пенсионеров (см. 1.1) усреднение баллов порядковой шкалы представляло собой математически бессмысленную процедуру.

В то же время шкала может вполне корректно использоваться в экспериментальных исследованиях, но для этого необходимо применять адекватные методы обработки данных, например, конвертирование порядковых шкал в метрические шкалы или использование непараметрических методов для обработки данных. (Подробно методы обработки результатов измерений в порядковой шкале будут рассмотрены в следующих главах данной книги).

В *интервальной шкале* (*interval scale*) присутствуют упорядоченность и интервальность, но нет нулевой точки. В этой шкале исследуемому объекту присваивается число единиц измерения, пропорциональное выраженности измеряемого свойства. Соответствующие интервалы разных участков шкалы имеют одно и то же значение, поэтому измерения в интервальной шкале допускают не только классификацию и ранжирование, но и точное определение различий между категориями.

Примеры интервальной шкалы: фиксация времени и даты, температурные шкалы Цельсия и Фаренгейта. Например, для известных шкал измерения времени можно сказать, что интервал между 3 и 6 ч равен интервалу между 4 и 7 ч. Но можно ли сказать, что 6 ч в два раза больше, чем 3 ч? Очевидно, нет. Аналогично, если сегодня уличный термометр фиксирует температуру -10^{o}C, а вчера было -20^{o}C, мы можем сказать, что сегодня теплее на 10 градусов. Но можно ли сказать, что сегодня теплее в 2 раза? Это кажется нелепым, не так ли? По крайней мере, так никто не говорит. А можно ли сказать, что температура в 300^{o}C отличается от температуры в 200^{o}C в два раза сильнее, чем температура в 150^{o}C отличается от температуры в 100^{o}C? Это кажется логичным.

Подобные ответы на такие вопросы связаны с тем, что интервальные шкалы (например шкалы Цельсия и Фаренгейта) не имеют нулевой точки отсчета. Точнее, выбор нулевой точки в интервальной шкале условен (произволен). Имея данные, представленные в интервальной шкале, мы можем судить о том, *насколько* больше или *насколько* меньше выражено измеряемое свойство, но не о том, *во сколько раз* больше или меньше. Для интервальных шкал характерна произвольность выбора нулевой точки, т.е. 0 не соответствует полному отсутствию измеряемого свойства, поэтому измерения не соответствуют абсолютному количеству измеряемого свойства. Например, нулевое значение температуры по шкале Цельсия не соответствует полному отсутствию тепла. Для

интервальных шкал мы можем корректно использовать большинство математических операций, поэтому если, например, имеются данные в порядковой шкале, то с точки зрения математической корректности целесообразно их конвертировать в интервальную шкалу, используя специальные процедуры, которые будут рассмотрены далее. Но на практике иногда возникают сложности как при преобразовании исходных данных в метрические шкалы, так и при интерпретации и анализе полученных результатов.

В *относительной шкале* (*ratio scale*) присутствуют все атрибуты измерительных шкал: упорядоченность, интервальность, нулевая точка. Для обозначения такой шкалы также используются термины *шкала отношений* и *абсолютная шкала*. Последний термин подчеркивает абсолютный характер нулевой точки.

Относительная шкала позволяет оценивать, во сколько раз свойство одного объекта больше или меньше аналогичного свойства другого объекта, принимаемого за эталон, единицу. Эта шкала характеризуется всеми атрибутами интервальной шкалы и, кроме того, имеет фиксированную нулевую точку (0), которая не является условной, так как она соответствует полному отсутствию измеряемого свойства. Например, переменная *количество сотрудников* имеет *фактическое* начало отсчета, так как нулевое значение соответствует отсутствию сотрудников вообще. Аналогично, нулевое значение может соответствовать отсутствию образования, дохода, детей, количеству лет в браке и т.п. Другими примерами измерения в относительной шкале являются определение физических характеристик (веса, длины, высоты, площади, температуры по температурной шкале Кельвина), запись возраста респондента, денежных потоков, численности населения, количества сотрудников, временных периодов (например времени выполнения задания) и т.п.

В силу фиксированности нулевой точки при сравнении исследуемых объектов можно судить не только о том, насколько больше или меньше выражено свойство, но и во сколько раз больше или меньше оно выражено. Для относительной шкалы возможно корректное использование любых математико-статистических операций и процедур обработки данных. Впрочем, стоит отметить, что существование нулевой точки редко является критичным для статистического анализа. Например, для выполнения большинства известных параметрических методов (работающих с метрическими шкалами) для проверки статистических гипотез достаточно, чтобы данные были представлены в интервальной шкале.

Рассмотренная классификация измерительных шкал не является единственной. Некоторые специалисты (например Дж. Хили) выделяют лишь три категории измерительных шкал, не проводя существенных различий между интервальной и относительной шкалами и

объединяя эти шкалы в одну. Авторы некоторых учебных и специальных изданий, напротив, выделяют большее количество шкал. Например, профессор А.И. Орлов выделяет шесть типов шкал (шкалу наименований, порядковую шкалу, шкалу интервалов, шкалу отношений, шкалу разностей, абсолютную шкалу).

В качестве примеров конкретных шкал, используемых, например, в маркетинговых исследованиях, можно назвать следующие.

Шкала Лайкерта (Likert scale) используется при определении представлений, мнений и отношений к товарам, торговым маркам, магазинам и т.п. Шкала состоит из набора утверждений, по которым респондент выражает свое отношение («полностью согласен», «согласен», «затрудняюсь ответить», «не согласен», «полностью не согласен»).

Шкала Осгуда (Osgood scale) используется при определении различий в отношении к товарам, магазинам и иному путем указания респондентом мнения по поводу различных утверждений, касающихся предмета опроса в категориях «сложный – простой», «удобный – неудобный», «дорогой – дешевый». Шкала может содержать 5 или 7 позиций по мере возрастания (убывания) признака.

Ценностная шкала Рокича (Rokeach value scale) используется для оценки поведения потребителей на основе измерения ценностей. Респондентам предлагается ранжировать по степени важности список из наборов задач, целей, а также наиболее предпочтительных для них форм поведения в отношении приобретения ими товаров.

Ценностная шкала Шварца (Schwartz value scale) используется для исследования поведения потребителей на основе опросного листа, состоящего из десяти типов ценностей, выбор которых рассматривается в качестве руководящих жизненных принципов потребителей, побуждающих их к покупке.

1.3. Базовые процедуры обработки данных и проблема выбора измерительной шкалы

Базовыми и наиболее распространенными процедурами обработки данных практически любого экспериментального исследования являются процедуры получения агрегированных и комплексных оценок. В эксперименте, основанном на выборочном исследовании, измеряются показатели отдельных элементов выборки. Набор таких измерений по каждому показателю может быть значительным (десятки, сотни, тысячи). Для анализа показателя по всей выборке необходимо *агрегировать* измерения, т.е. получить *агрегированные* (коллективные, групповые, обобщенные) оценки. Получение агрегированных оценок должно выполняться корректно. Для величин, измеренных в метрических шкалах (интервальной и

относительной), агрегирование, как правило, осуществляется путем вычисления среднего арифметического по группе. Для метрических шкал эта простая процедура вполне корректна. Наиболее распространенная ошибка (с точки зрения математической корректности) – использование арифметических действий для агрегирования данных в неметрических шкалах (чаще всего порядковых), например вычисление среднего балла успеваемости по отдельному предмету по студенческой группе. Более подробно этот вопрос разбирается на примере шкалы Лайкерта в следующей главе.

На практике также довольно часто встречаются ситуации, когда необходимо синтезировать отдельные показатели в некоторый интегрированный (*комплексный*) показатель. Например, показатель инновационности компании интегрируется на основе нескольких показателей, характеризующих ее организационный климат. Другим, более распространенным, примером является усреднение баллов успеваемости студента (или студенческой группы) по нескольким дисциплинам.

Очень часто, как и в случае с агрегированными оценками, исследователи используют арифметические действия для интегрирования показателей, не задумываясь о содержательной и математической строгости используемых процедур. Как уже отмечалось, для неметрических шкал использование операций сложения, вычитания, умножения или деления для получения агрегированных или комплексных оценок с математической точки зрения некорректно и бессмысленно.

Таким образом, при выборе методов статистической обработки полученных данных в первую очередь следует обращать внимание на тип используемой измерительной шкалы. Идеальным с точки зрения корректности использования математико-статистического аппарата является выбор только тех методов, которые могут применяться для обработки данных, представленных в соответствующих шкалах. Однако, как уже обсуждалось и еще неоднократно будет показано ниже, многие математические операции в реальной практике используются для данных, измеренных в таких шкалах, для которых эти операции в принципе не применимы. В частности, многие базовые показатели, например среднее арифметическое, стандартное отклонение или коэффициент корреляции, являются настолько привычными для представления результатов, что исследователь редко задумывается о корректности использования этих показателей для данных разных типов.

Дифференциация номинативной шкалы от порядковой, а тем более от метрической шкалы, обычно не вызывает сложностей. Гораздо сложнее определить различие между другими шкалами (особенно порядковой и интервальной). Весьма распространенной является ошибка, когда исследователь интерпретирует данные так, как если бы они были измерены по шкале, допускающей более сложные операции, чем эта шкала в

действительности позволяет. Наиболее часто встречается ситуация, когда данные порядковой шкалы интерпретируются как измерения в интервальной и даже относительной шкале, поскольку статистические процедуры для этих шкал являются более информативными. (Читателям будут предложены соответствующие примеры и задания для анализа). Это, безусловно, создает очень серьезные трудности для анализа и интерпретации результатов исследования с точки зрения обоснования их корректности, поэтому исследователь должен проявлять осторожность при интерпретации и оценке результатов.

В следующей главе мы рассмотрим подходы к обоснованию выбора шкал и конвертации данных из неметрических в метрические шкалы. В частности, для более основательного ознакомления с вариантами и нюансами использования различных шкал и преобразования данных мы рассмотрим шкалу Лайкерта, которая используется почти повсеместно в психологических и социально-экономических исследованиях, и шкалу Раша (часто встречаемую в образовательно-психологической практике и имеющую серьезные перспективы для социально-экономических и гуманитарных исследований).

Глава 2. Оценки и измерения

В предыдущей главе мы предложили достаточно широкое определение измерения как операции соотнесения признаков эмпирических объектов с помощью чисел в соответствии с определенными правилами, которые определяются измерительными шкалами. Это определение не является единственным и бесспорным. Существует разделяемый многими исследователями подход, согласно которому измерение следует отличать от оценки – другого способа количественной характеристики величин. Различие измерения и оценки основано на следующих характеристиках: а) используемая измерительная модель (включающая в себя измерительную шкалу) и б) получаемые результаты. *Оценка* дает приближенное значение величины или параметра, причем в отличие от измерения нет однозначного соответствия между величиной и ее количественным выражением в определенных единицах. Считается, что по исходным данным, представленным в неметрических шкалах, получаются *оценки* свойств исследуемого объекта, тогда как *измерения* могут быть получены при использовании либо метрических шкал, либо определенных процедур преобразования исходных данных в метрические шкалы.

Принято считать, что измерение, в отличие от оценки, характеризуется научной обоснованностью, корректностью использования математико-статистического аппарата, большей точностью, надежностью и объективностью, стандартизованностью и универсальностью, меньшей чувствительностью к смене наборов исходных данных и т.д. Хотя четкое и однозначное различие между оценками и измерениями провести достаточно сложно, в данной главе с целью демонстрации такого различия подробно рассмотрим два подхода к измерению: «измерение» на основе шкалы Лайкерта и измерение с использованием модели Раша. Можно считать, что первый подход продуцирует оценки, а второй – именно измерения.

2.1. Шкала Лайкерта

Шкала Лайкерта (Likert scale) впервые была предложена в 1932 г. американским специалистом в областях организационной психологии и управления Ренсисом Лайкертом (Rensis Likert)[2]. Эта шкала была разработана им во время обучения в аспирантуре в Колумбийском университете (Columbia University) при подготовке докторской диссертации.

[2] Вариантом произношения фамилии является «Ликерт».

С тех пор эта шкала широко применяется для измерения отношений респондентов к предмету исследования.

Метод предполагает, что респондент выражает свое согласие или несогласие с каждым суждением (из предложенного набора) по шкале оценок (из 3, 4, 5, 6, 7 или более пунктов), а его место на итоговой шкале определяется суммой оценок каждого отдельного суждения. Лайкерт обосновал предположение о самом существовании измеряемой однородной латентной переменной и показал, что его шкала приемлема для ее измерения. Он также дал обоснование того, что предложенная шкала является порядковой.

Хотя «классическая» шкала Лайкерта включает 5 пунктов, сходные шкалы с другим количеством позиций шкалы (3, 4, 6, 7…) также называют шкалами Лайкерта; чаще всего используется 5- или 7-позиционная шкала. Шкала может содержать или не содержать «нейтральную точку» (при нечетном числе позиций). Частным случаем шкалы Лайкерта является дихотомическая шкала (2 позиции).

Оценки по шкале Лайкерта могут быть выражены в категориях согласия, частоты, важности, качества и т.п., например:

(5) Полностью согласен – (4) Согласен – (3) Нейтральное отношение – (2) Не согласен – (1) Полностью не согласен;

(5) Всегда – (4) Очень часто – (3) Иногда – (2) Редко – (1) Никогда;

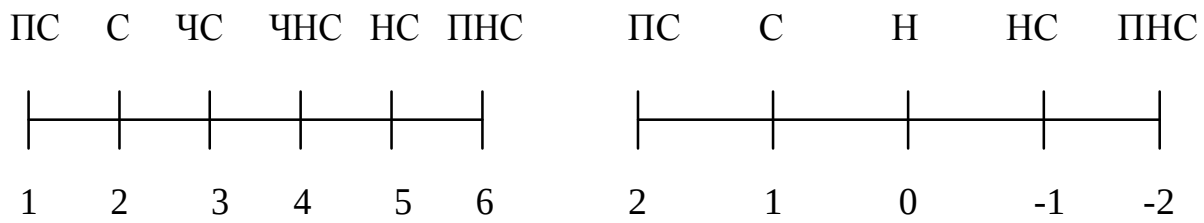
(4) Очень важно – (3) Важно – (2) В некоторой степени важно – (1) Неважно;

(5) Очень хорошо – (4) Хорошо – (3) Приемлемо – (2) Плохо – (1) Очень плохо;

В качестве примера конкретного использования 4-позиционной шкалы Лайкерта в реальном исследовании приведем фрагмент опросного листа для оценки организационного климата для творчества и инноваций KEYS Questionnaire (Amabile, Burnside and Gryskiewicz, 1999).

		NEVER N	SOMETIMES S	OFTEN O	ALWAYS A
34.	People are quite concerned about negative criticism of their work in this organization.	N	S	O	A
35.	People are recognized for creative work in this organization.	N	S	O	A
36.	The tasks in my work call out the best in me.	N	S	O	A
37.	My supervisor plans poorly.	N	S	O	A
38.	The organization has an urgent need for successful completion of the work I am now doing.	N	S	O	A
39.	People in this organization feel pressure to produce anything acceptable, even if quality is lacking.	N	S	O	A
40.	There is an open atmosphere in this organization.	N	S	O	A
41.	There is a good blend of skills in my work group.	N	S	O	A
42.	Ideas are judged fairly in this organization.	N	S	O	A
43.	Top management does not want to take risks in this organization.	N	S	O	A
44.	In my daily work environment, I feel a sense of control over my own work and my own ideas.	N	S	O	A
45.	Failure is acceptable in this organization, if the effort on the project was good.	N	S	O	A
46.	The budget for my project(s) is generally adequate.	N	S	O	A
47.	My area of this organization is creative.	N	S	O	A
48.	My area of this organization is productive.	N	S	O	A
49.	People are encouraged to solve problems creatively in this organization.	N	S	O	A
50.	People are rewarded for creative work in this organization.	N	S	O	A

Позиции шкалы Лайкерта обычно кодируются числами от 1 до 5 (для пятипозиционной шкалы), от 1 до 6 (для шестипозиционной шкалы) и т.д., хотя кодирование может включать отрицательные и нулевое значение (например для нейтральной позиции), что, впрочем, используется нечасто.



ПС – полностью согласен; С – согласен; ЧС – частично согласен;
ЧНС – частично не согласен; НС – не согласен; ПНС – полностью не согласен;
Н – нейтральное отношение

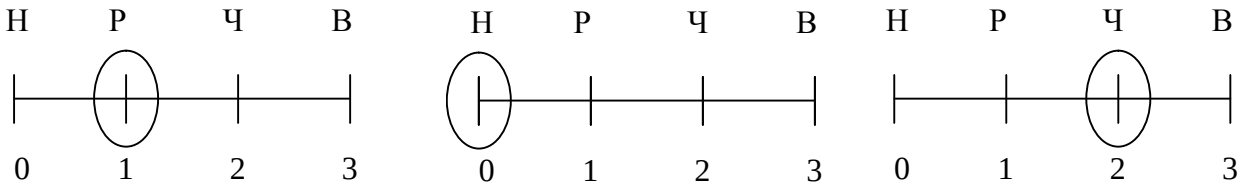
Направление кодирования зависит от формулировки вопроса. Для позитивно сформулированных утверждений (например «Творческие идеи в данной организации поощряются») позитивно выраженное отношение (например «согласен») кодируется, как правило, большим числом (например 5), а негативно выраженное («не согласен») – меньшим. При негативной формулировке (например «Мой руководитель игнорирует мои предложения») позитивно выраженное отношение может кодироваться меньшим числом, а негативно выраженное – большим.

В случае, если анкета содержит как прямые, так и обратные вопросы или утверждения (например «Мой руководитель поддерживает мои инициативы» и «Мой руководитель не поддерживает мои инициативы»), отклики на обратные утверждения перекодируются по формуле

$$r = (min - s) + max,$$

где r – отклик на прямые утверждения (вопросы); s – отклик на обратные утверждения (вопросы); min, max – минимальное и максимальное значения шкалы.

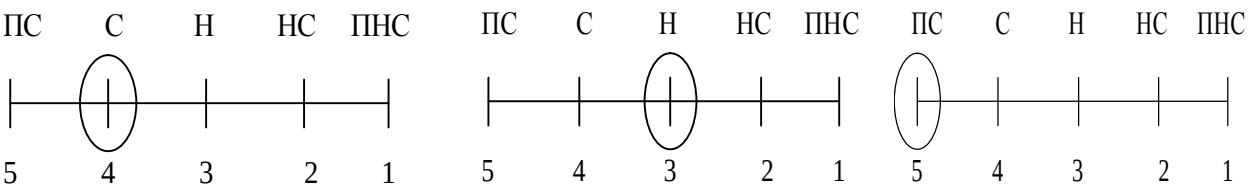
Рассмотрим пример агрегирования измерений в шкале Лайкерта (без нейтральной позиции). Предположим, что три респондента отвечают на утверждение «Мой руководитель поддерживает мои инициативы», и отклики респондентов распределились следующим образом



Н – Никогда Р – Редко Ч – Часто В – Всегда

Если использовать традиционный подход для агрегирования оценок, то мы получим 1 + 0 + 2 = 3. Разделив полученное значение на количество респондентов, получим значение 1. Что можно сказать в этом случае о поддержке руководителем организации / отдела творческих инициатив сотрудников? Если изменить кодировку позиций шкалы, например, сдвинув на 1 (от 1 до 4), то мы получим 2 + 1 + 3 = 6; 6 / 3 = 2. Агрегированная оценка в численном виде изменилась, но средний «словесный» результат («редко»), казалось бы, остается тем же. Аналогичный результат мы получим при сдвиге кодировки шкалы на 2, 3, 4 и т.д. Является ли эта «инвариантность» результата к подобной перекодировке свидетельством корректности используемого подхода к получению агрегированной оценки?

Прежде чем ответить на этот вопрос, рассмотрим еще один пример. Возьмем другую шкалу Лайкерта (с другим количеством позиций и с наличием нейтральной позиции). Будем считать, что три других сотрудника отвечают на утверждение «Мой руководитель поддерживает мои инициативы». Предположим, что отклики респондентов распределились следующим образом:



ПС – полностью согласен; С – согласен; Н – нейтральное отношение;
НС – не согласен; ПНС – полностью не согласен

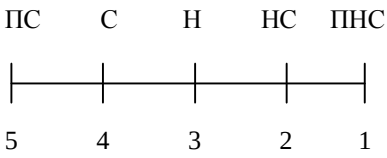
Агрегированная оценка составит (4 + 3 + 5) / 3 = 4. Средний «словесный» результат (в данном случае «согласен») сохранится при любом линейном преобразовании численных значений пунктов шкалы. На первый взгляд никаких изъянов в таком подходе к агрегированию нет. Выше мы уже обсуждали, что агрегирование оценок на основе вычисления среднего арифметического корректно лишь для метрических шкал. Может ли шкала Лайкерта рассматриваться как метрическая? Другими словами, могут ли измерения в шкале Лайкерта, по крайней мере, интерпретироваться как интервальные?

В шкале Лайкерта очень наглядно проявляется дилемма между простотой измерительных процедур и математической строгостью. В литературе встречаются и аргументируются различные точки зрения на возможности использования шкалы Лайкерта. Эта шкала, как показал еще ее автор, является порядковой (неметрической) шкалой, однако многие исследователи считают возможным использовать ее как интервальную шкалу и применять процедуры для обработки данных, приемлемые для метрических шкал, так как, по их мнению, это не приводит к серьезным ошибкам в результатах. Основными аргументами в пользу такого утверждения являются следующие *пред*-суждения:

- интервалы между позициями шкалы Лайкерта в отдельном пункте (вопросе или утверждении) анкеты приблизительно одинаковы;
- относительные интервалы между позициями шкалы Лайкерта изменяются незначительно при переходе от вопроса к вопросу (от утверждения к утверждению).

Если эти два условия можно считать выполненными, то к шкале Лайкерта могут быть применены процедуры, приемлемые для интервальных шкал.

Давайте посмотрим, что означает предпосылка о равенстве интервалов. Пусть мы имеем следующую шкалу:

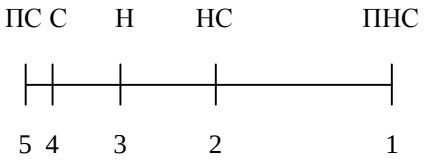


ПС – полностью согласен; С – согласен; Н – нейтральное отношение;
НС – не согласен; ПНС – полностью не согласен

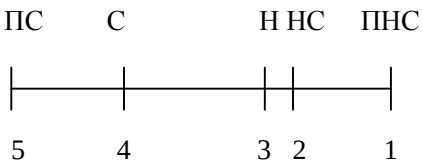
Равенство межпозиционных интервалов должно означать следующее:

$$ПС - С = С - Н = Н - НС = НС - ПНС = 5 - 4 = 4 - 3 = 3 - 2 = 2 - 1 = 1.$$

Пусть мы имеем в качестве пункта анкеты следующее утверждение: «Члены моей рабочей группы открыты новым идеям». Для респондента А, который в целом согласен с этим утверждением, интервалы между позициями шкалы могут быть такими:



В этом случае респондент А с большей вероятностью выберет одну из позиций – ПС или С, нежели Н, а тем более ПНС, т.е. в этом случае говорить о равенстве интервалов (даже приблизительном) не приходится. Предположим, что некий респондент В скорее не согласен с предложенным утверждением. Для него интервалы между позициями шкалы могут быть следующими:

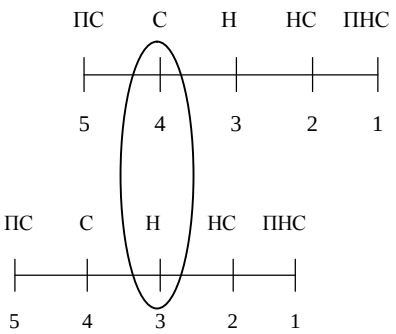


Интервал между Н и НС для респондента В много меньше, чем интервалы между другими позициями. Очевидно, что вычисление арифметического среднего для данного случая некорректно.

Аналогично складывается ситуация и с получением комплексных оценок в шкале Лайкерта. Предположим, что измеряемым показателем является «Поддержка творческой инициативы сотрудников руководством компании». Анкета включает два пункта: «Мой руководитель открыт для новых идей» и «Мой руководитель поддерживает мои новые предложения». Очевидно, что эти пункты имеют разный «вес». Поэтому отклик «согласен» при ответе на первый пункт может не соответствовать позиции «согласен» при отклике на второй пункт (Одно дело – просто воспринимать новые идеи, другое – их поддерживать). Возможно, «вес» позиции «согласен» (4) в первом пункте соответствует нейтральной позиции (3) во втором пункте.

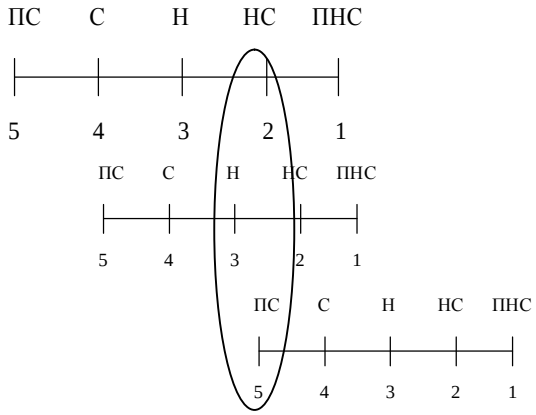
- Мой руководитель открыт для новых идей

- Мой руководитель поддерживает мои новые предложения



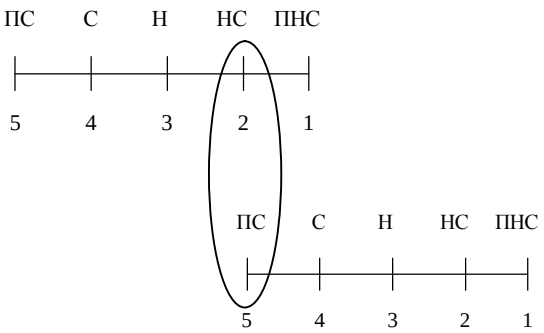
Еще более рельефно неэквивалентность значений позиций шкалы при переходе от пункта к пункту при получении комплексных оценок в шкале Лайкерта демонстрирует пример, приведенный в книге Т. Бонда и К. Фокс (Bond & Fox, 2001). Измеряемым показателем является «Озабоченность избыточным весом». В анкете 3 пункта: «Я регулярно принимаю рвотное для контроля веса», «Я придерживаюсь строжайшей диеты и всегда считаю каждую калорию и каждый грамм», «Я регулярно делаю гимнастику для того, чтобы «сжечь» лишние калории». Очевидно, что одни и те же значения позиций шкалы (как числовые, так и «словесные») в каждом пункте различны. Например, значение *2* в первом пункте может соответствовать значению *3* во втором и значению *5* в третьем.

- Я регулярно принимаю рвотное для контроля веса
- Я придерживаюсь строжайшей диеты и всегда считаю каждую калорию и каждый грамм
- Я регулярно делаю гимнастику для того, чтобы «сжечь» лишние калории



Приведем еще один пример из этой же книги. Оценивается отношение школьников к работе за компьютером. В анкете используется несколько пунктов, в том числе: «Я настолько боюсь компьютера, что стараюсь избегать пользоваться им» и «Я опасаюсь, что сделаю ошибку при работе за компьютером». Школьник, выбирающий «полностью согласен» по второму пункту, может выбрать «не согласен» по первому, и комплексная оценка, получаемая традиционным способом, будет равна 7, хотя очевидно, что «вклад» пунктов в общую оценку не равноценен. Школьник, выбравший «полностью согласен» по первому пункту и «не согласен» по второму, получит те же 7 баллов, что кажется нелогичным. Более реалистичную оценку можно было бы получить, «выровняв» «вес» этих пунктов, например, следующим образом: *2* в первом пункте равно *5* во втором.

- Я настолько боюсь компьютера, что стараюсь избегать пользоваться им
- Я опасаюсь, что сделаю ошибку при работе за компьютером



Таким образом, в реальности достаточно сложно обеспечить выполнение двух основных условий (равенство межпозиционных интервалов шкалы и эквивалентность позиций шкалы при переходе от пункта к пункту) для того, чтобы шкалу Лайкерта можно было бы считать интервальной. Даже для приблизительного выполнения этих условий «веса» пунктов анкеты должны быть приблизительно равны, а отношения респондентов примерно одинаковы.

2.2. Возможности преобразования данных из неметрических шкал в метрические

Метрические шкалы, как мы знаем, являются более «мощными», т.е. они отражают больше информации о различии исследуемых объектов по измеряемому свойству. Поэтому если у исследователя есть такая возможность, то надо использовать более мощную шкалу. При планировании экспериментального исследования целесообразно выбирать такие способы сбора данных, которые позволяют применить метрические шкалы. Однако очень часто такого выбора нет, и приходится использовать доступную измерительную шкалу. Часто исследователи не задумываются над тем, какую шкалу они применяют, однако, мы неоднократно обращали на это внимание, определение шкалы, в которой проводится или будет проводиться измерение, – ключевой момент анализа данных, так как выбор последующих методов для обработки измерений зависит от применяемой измерительной шкалы.

При применении статистических методов требуется обосновать, что выбранная шкала позволяет эти методы использовать, поэтому исследователь стремится найти свидетельства того, что его шкала может рассматриваться как метрическая.

Почти повсеместно используемые в исследовательской практике подходы к получению агрегированных и комплексных оценок по шкале Лайкерта путем вычисления арифметического среднего работают с практической точки зрения, но лишены смысла с точки зрения математической, поскольку эта шкала не является интервальной. С другой стороны, для получения корректных результатов исследования необходимо обеспечить корректность получаемых измерений и применяемых процедур обработки. В то же время шкала Лайкерта является простым и удобным инструментом для сбора данных. Возможен ли некий компромисс между практическим удобством и математической строгостью? Возможно ли значимое обоснование данных в шкале Лайкерта как данных интервальной

шкалы? Возможно ли некое преобразование данных из шкалы Лайкерта в метрическую шкалу?

Одним из самых «сильных» аргументов в пользу метрического характера шкалы является распределение измерений по нормальному закону[3]. Исходным принципом здесь является предположение о том, что измеряемое свойство распределено в генеральной совокупности в соответствии с нормальным законом. Следовательно, измерение данного свойства также должно обеспечивать нормальное распределение. Если это так, то шкала может считаться по крайней мере интервальной (Наследов, 2006). Если это не так, то измеряемое свойство отражается в не метрической шкале, а, например, в порядковой. Таким образом, основной подход к разработке и анализу измерительной шкалы заключается в обеспечении и проверке нормального распределения полученных измерений. Если выборочное распределение не отличается от нормального, то, по мнению, многих специалистов, можно считать, что измеряемое свойство отражается в интервальной шкале. Однако такой подход к обоснованию метрического характера измерений принимается не всеми исследователями. Например, А.И. Орлов в одной из своих работ показывает, что в большинстве случаев распределения результатов измерения существенно отличаются от нормальных.

Причиной отклонения формы выборочного распределения признака от нормального вида чаще всего является особенность процедуры измерения: используемая шкала может обладать неравномерной чувствительностью к измеряемому свойству в разных частях диапазона его изменчивости. Для проверки нормальности используются различные процедуры, позволяющие выяснить, отличается ли от нормального выборочное распределение измеренной переменной. Необходимость такого сравнения особенно актуальна, когда есть сомнения, в какой шкале представлен признак – порядковой или интервальной. Это важно, поскольку от этого зависит возможность использования методов для дальнейшего анализа. А метрические шкалы позволяют использовать гораздо более широкий диапазон методов, чем порядковые, поэтому и исследовательские выводы могут быть более глубокими и информативными. При значительном отклонении эмпирического распределения от нормального следует отказаться от предположения о том, что признак измерен в метрической шкале. Однако остается вопрос о *степени значительности* отклонения распределений измерения от нормального закона.

Существуют и используются несколько способов проверки нормальности:

[3] При дальнейшем изложении мы предполагаем, что читатели знакомы с нормальным законом распределения. При необходимости пополнить свои знания в этой области читатели могут, обратившись к любому вузовскому учебнику по теории вероятностей и математической статистике.

- визуально-графический способ (строятся графики частот встречаемости признака);
- оценка отклонений асимметрии и эксцесса распределения от нулевых значений;
- критерий Колмогорова-Смирнова и другие статистические критерии.

Процедуры проверки нормальности распределения включены в статистические пакеты SPSS, STATISTICA и др. Удобный инструментарий для оценки вида распределения включает в себя программный продукт @Risk, предлагаемый компанией Palisade, специализирующейся на разработке программного обеспечения для принятия решений в условиях риска и неопределенности (см., например, www.palisade.com и Winston & Albright, 2001).

Если исходные данные (*raw data*) не проходят проверку на соответствие нормальному распределению, можно их нормализовать, т.е. привести к нормальному виду. Наиболее простым и распространенным нормализирующим преобразованием исходных данных является приведение данных к так называемому стандартному *z*-распределению со средним, равным 0, и стандартным отклонением, равным 1:

$$z = \frac{r - m}{s},$$

где r – исходные данные; m – среднее значение по признаку (по исходным данным); s – стандартное отклонение по признаку (по исходным данным).

Допустим, мы имеем по некоторому измеряемому признаку 10 измерений, полученных, например, по шкале Лайкерта.

№	1	2	3	4	5	6	7	8	9	10
r	3	4	4	2	3	4	5	2	3	4

Среднее значение по имеющимся исходным данным составляет 3,4; стандартное отклонение 0,966. Этого достаточно для нормализации данных:

№	1	2	3	4	5	6	7	8	9	10
r	-0,41	0,62	0,62	-1,45	-0,41	0,62	1,66	-1,45	-0,41	0,62

Среднее значение по нормализованным данным равно 0; стандартное отклонение – 1. В пакете SPSS при вычислении базовых статистик (*Analyze, Descriptive Statistics, Descriptive*) можно выбрать опцию перевода исходных данных в нормализованную *z*-шкалу (*Save standardized values as variables*). Более подробное описание возможностей пакета SPSS для обработки и анализа исследовательских данных приводится в третьей и последующих главах книги.

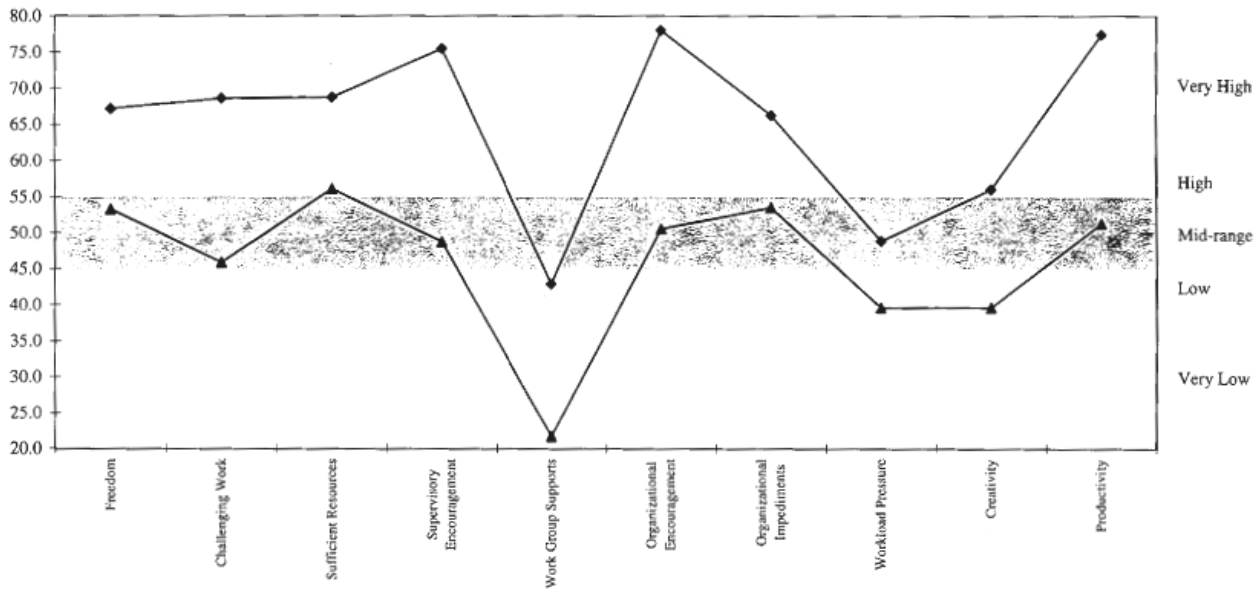
В психологических исследованиях широко используется *стандартная T-шкала*, связанная с *z*-шкалой простым преобразованием:

$$T = 10z + 50.$$

T-распределение имеет среднее, равное 50, и стандартное отклонение, равное 10. Например, для исходных данных, представленных в таблице выше, значения в *T*-шкалу пересчитываются следующим образом:

№	1	2	3	4	5	6	7	8	9	10
T	45,9	56,2	56,2	35,5	45,9	56,2	66,6	35,5	45,9	56,2

В качестве примера практического использования *T*-шкалы на рисунке приведены результаты оценки организационного климата (профили) двух компаний по 10 показателям (Amabile, Burnside and Gryskiewicz, 1999).



Описанные подходы к нормализации данных имеют принципиальный недостаток с точки зрения математической строгости. Они изначально используют процедуры, неприменимые для порядковых шкал (определение среднего значения и стандартного отклонения) по причинам, которые подробно обсуждались выше. Однако исследователи «забывают» о такой «мелочи», и обычно считается, что нормализованные данные, полученные в порядковых шкалах, можно рассматривать как интервальные со всеми вытекающими последствиями по применению методов обработки данных. В качестве упражнения читателю будет предложено для небольшой выборки осуществить подобные преобразования исходных данных и сравнить результаты.

Еще одним распространенным подходом к преобразованию порядковых шкал в интервальные является замена интервалов на их средние значения: (позиция *i* – позиция *I* –

1) / 2. Предположим, что имеются данные о среднемесячном доходе, представленные в порядковой шкале с позициями:

1 – до 2000 руб.;

2 – от 2000 до 4000 руб.;

3 – от 4000 до 8000 руб.;

4 – от 8000 до 12000 руб.;

5 – свыше 12000 руб.

В таком случае мы можем перейти, например, к шкале, которая имеет свойства относительной: 1000, 3000, 6000, 10000, 15000. Очевидно, что подобный подход имеет существенные ограничения. Он применим лишь в случаях, когда содержательный смысл измеряемых величин (денежные величины, временные периоды, размеры, количество определенных товаров и т.п.) позволяет осуществить такое преобразование, т.е. когда изначально имеются признаки, которые могли бы быть измерены в метрических шкалах, но по каким-то причинам (например этическим или в целях упрощения анкеты и процесса обработки данных) они были представлены в порядковой шкале. Тем не менее подобный подход позволяет корректно использовать те методы обработки данных, которые неприменимы к порядковым шкалам. Однако точность такого преобразования и полученных результатов существенно зависит от выбора позиций порядковой шкалы.

2.3. Модель Раша

Процедуру более корректного преобразования данных из шкалы Лайкерта в интервальную шкалу можно осуществить на основе *модели Раша* (*Rasch model*), предложенной датским математиком Г. Рашем еще в конце 1950-х гг., хотя понимание и признание среди исследователей и специалистов эта модель получила лишь в 1970–1980-х гг. Первоначально модель Раша была разработана для оценки способностей (знаний, умений и т.п.), позднее ее также начали использовать для оценки отношений (*attitudes*) и восприятий (*perceptions*). В настоящее время модель Раша наиболее активно применяется в образовательно-педагогической практике (в инструментах тестирования), но имеет серьезные перспективы для социально-экономических и гуманитарных исследований.

Разработка этой модели была связана с обнаружившимися проблемами представления данных в порядковых шкалах. Во-первых, эти шкалы не обеспечивают линейности (пропорциональности) в представлении данных, а во-вторых, получаемые данные субъективны и ситуативны в том смысле, что они зависят от конкретных респондентов и конкретных вопросов. Изменение набора респондентов и/или набора вопросов в анкете

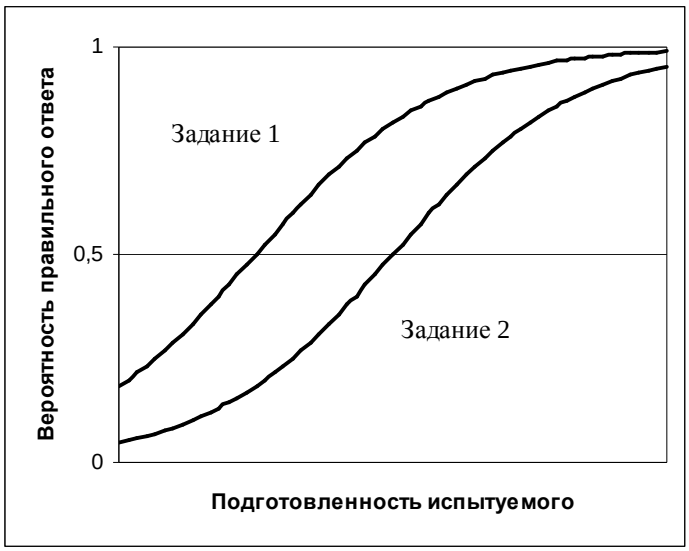
может привести к другим результатам. Г. Раш предложил сравнительно простую модель для «объективного» измерения, которая снимает эти две проблемы.

В контексте измерения способностей ключевая идея модели Раша может быть сформулирована следующим образом: вероятность правильного ответа на вопрос или задание теста (P) является функцией способностей (знаний) тестируемого и зависит от способностей (знаний) тестируемого (B) и сложности вопроса (D):

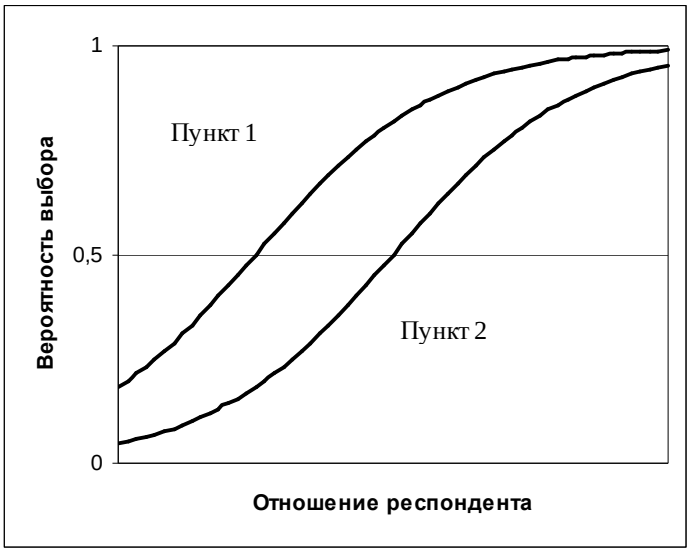
$$P_i(B) = \frac{e^{B-D}}{1+e^{B-D}}.$$

В контексте измерения отношений (оценок) эта идея может быть интерпретирована следующим образом. Вероятность выбора (позитивного отклика) определенной позиции пункта анкеты (*endorsement*) зависит от отношения (*attitude*) респондента к пункту (вопросу или утверждению) анкеты и от степени соответствия (*agreeability*) этого пункта и позиции действительному положению вещей.

Эта базовая идея может быть проиллюстрирована с помощью *характеристической кривой* заданий теста (или пунктов анкеты). Пусть в тесте Задание 2 сложнее Задания 1. Тогда при той же подготовленности испытуемого (или при тех же самых знаниях) вероятность правильного ответа на Задание 2 будет меньше, чем вероятность правильного ответа на Задание 1.



Аналогично в случае измерения отношений, если в анкете Пункт 2 объективно сложнее для выбора респондента, чем Пункт 1, то при одном и том же отношении респондента к предмету исследования вероятность выбора Пункта 2 будет меньше, чем вероятность выбора Пункта 1.



Единицей шкалы Раша является «логит» (*d*):

$$d = \ln(\frac{P}{1-P}).$$

Здесь P – вероятность правильного ответа на вопрос или (в контексте измерения отношений) вероятность выбора пункта (позиции) анкеты. Чем больше вероятность правильного ответа на вопрос испытуемым, тем большее значение по шкале Раша он имеет. На практике при получении данных в шкале Раша вместо вероятностей используют частоты откликов, т.е. чем чаще тестируемый отвечает правильно, тем выше его показатель в шкале Раша.

Если подставить в эту формулу значения вероятности успешного (правильного) ответа на вопрос, как она определяется в модели Раша, то получим $\boldsymbol{d = B - D}$. Этот результат показывает, что измерения в шкале Раша зависят только от способности испытуемого и сложности вопроса, т.е. мы получаем измерение, *независимое от выборки испытуемых (person free measurement) и набора вопросов теста (item free calibration)*. Это одно из важнейших, наряду с интервальностью, свойств шкалы Раша. Независимость измерения от теста и респондентов (в пределах ошибки измерения) позволяет специалистам в области теории измерений называть шкалу Раша «объективной» измерительной шкалой и сравнивать со шкалами, используемыми в точных и естественных науках, например в физике.

Для определения сложности пункта шкалы используется следующая формула:

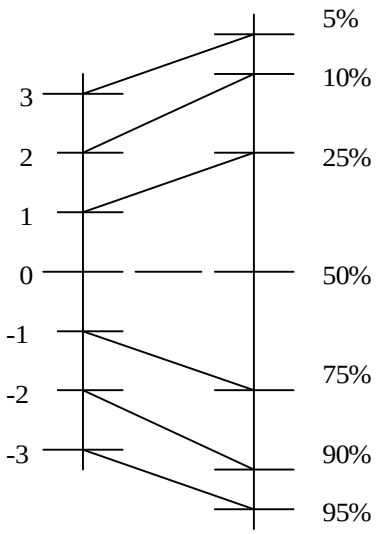
$$d = \ln(\frac{1-P}{P}).$$

Следовательно, чем проще пункт анкеты, тем меньшее значение имеет этот пункт по шкале Раша.

Некоторую сложность может вызывать интерпретация данных в шкале Раша. Значение 0 по шкале Раша отражает ситуацию, когда знания (способности) тестируемого

соответствуют сложности задачи (вопроса), а вероятность правильного ответа на вопрос составляет 50%. При уменьшении сложности вопроса на 1 логит вероятность правильного ответа возрастает приблизительно до 75%. При увеличении сложности вопроса на 1 логит вероятность правильного ответа уменьшается примерно до 25%.

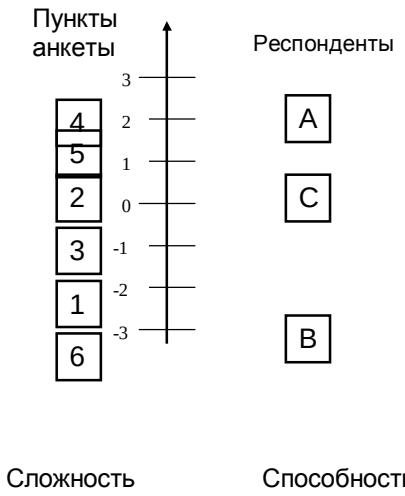
Иначе значение 0 можно интерпретировать следующим образом: 50% тестируемых правильно отвечают на вопрос, 50% – неправильно. Значение –2,94 соответствует «простому» вопросу, на который 95% тестируемых отвечают правильно, и лишь 5% – неправильно. Значение +2,94, напротив, соответствует «сложному» вопросу, на который лишь 5% тестируемых отвечают правильно (см. рис.).



Несколько иначе можно интерпретировать значения шкалы Раша при измерении отношений. Например, при оценке творческого климата в организации значение 0 соответствует «нейтральному» климату (50% положительных и 50% отрицательных откликов), значение 2,94 соответствует 95% положительных и 5% отрицательных откликов и т.д. Иначе значение 2,94 может быть интерпретировано как показатель организационного климата, в котором 95% положительных факторов для творчества и инноваций и лишь 5% отрицательных факторов. Значение –2,94 соответствует резко отрицательному («холодному») климату для творчества и инноваций. Такая интерпретация в контексте приведенного примера может показаться несколько «вольной» по сравнению с «классической» интерпретацией значений шкалы Раша, поскольку в «классической» интерпретации более часто встречаемый признак (с большим количеством баллов) считается более «легким» и имеет меньшее значение. Однако, как мы еще увидим, результат интерпретации во многом зависит от формулировки вопросов или утверждений (в «негативной» или «позитивной» форме), а также от прямого или обратного кодирования позиций в исходной шкале. Впрочем, на данный момент это не столь принципиально для

рассмотрения модели Раша в качестве способа конвертирования данных из порядковой шкалы в интервальную.

Таким образом, модель Раша оценивает «сложность» каждого пункта анкеты и «способность» тестируемого в логитах (единицах интервальной шкалы Раша). Специальные программные продукты (например WINSTEPS) позволяют представлять сложность пунктов и способности респондентов на *диаграмме Раша* (см. рис.).



Для приведенного на рисунке примера пункт 4 «сложнее» пункта 5 на 1 логит и «сложнее» пункта 3 на 3 логита. Респондент А «способнее» респондента В на 5 логитов, а респондента С – на 2 логита. Слова «сложность» и «способность» могут наполняться различными смыслами в зависимости от контекста измерения.

Модель Раша обладает следующими преимуществами в сравнении с порядковыми шкалами, в частности шкалой Лайкерта.

- Модель обеспечивает калибровку пунктов анкеты в соответствии с их «весом» («сложностью»).
- «Вес» пунктов анкеты может быть сравнительно просто вычислен на основе частотности откликов.
- Единица измерения шкалы Раша, представляющая собой результат преобразования исходных данных, имеет стандартизованный характер, что дает возможность объективно сравнить результаты, полученные по разным инструментам и выборкам.
- Данные представляются в интервальной шкале.
- Модель обеспечивает статистику соответствия (*fit statistics*) каждого пункта анкеты и откликов каждого респондента (см. главу 4).

- Пункты анкеты, которые каким-либо респондентом оставлены без ответа, не представляют трудностей при использовании модели Раша, т.е. пропуски заданий (пунктов анкеты) не критичны для обработки данных.
- Измерения по модели Раша не зависят от выборки (***sample free***) и набора пунктов анкеты (***item free***), т.е. теоретически они не зависят от распределения индивидуальных откликов и набора вопросов. Это позволяет получать объективированные результаты, что, во-первых, положительно отражается на качестве исследовательских результатов, а во-вторых, следует принципу объективности как одного из критериев научного исследования. По этой причине измерения, основанные на модели Раша, называют «*объективными измерениями*» (***objective measurement***) и даже «*фундаментальными измерениями*» (***fundamental measurement***).

Разработаны специальные программные продукты для конвертирования исходных данных в шкалу Раша и выполнения всех необходимых расчетов (например WINSTEPS, RUMM, BiLog, MultiLog). Процедуры, реализованные в этих программах, достаточно сложны и громоздки. Чтобы проиллюстрировать общий подход к реализации подобных процедур, рассмотрим простейший алгоритм преобразования исходных данных в шкалу Раша. Для простоты начнем с дихотомической шкалы, в которой измерение представлено двумя возможными состояниями (например 1 и 0). Примером такой шкалы может являться некий тест на знание учебного предмета, в котором испытуемому требуется выбрать правильный ответ из двух вариантов по каждому вопросу. Правильный ответ кодируется 1, неправильный – 0.

Предположим, что в тест включено 7 заданий (вопросов): 1,2, … 7. Допустим, что на вопросы (задания) ответили 7 испытуемых, которых обозначим латинскими буквами A, B, … G. Исходные данные представлены в следующем виде.

	Задания теста						
Испытуемые	**1**	**2**	**3**	**4**	**5**	**6**	**7**
A	1	0	1	1	0	1	0
B	0	1	0	1	1	0	0
C	0	1	1	1	1	1	0
D	0	1	1	1	1	1	1
E	1	0	1	0	1	1	0
F	0	0	1	1	0	1	1
G	1	1	1	1	1	0	0

Для конвертирования исходных данных в шкалу Раша необходимо определить общую сумму баллов (S_1), набранных испытуемыми, и общую сумму баллов по каждому пункту теста (S_2). Далее определяется максимальное значение, которое *мог бы* набрать испытуемый (*Max1*): *Max1* = 1 * 7 = 7. В контексте тестового исследования (измерения способностей, знаний, подготовленности и т.д.) для дихотомической шкалы это значение соответствует количеству вопросов, на каждый из которых может быть дан правильный ответ (7). Отношение S_1 к *Max1* может быть интерпретировано (при большом количестве пунктов теста) как частота (или вероятность) правильного ответа (P_1). Это значение определяется для каждого испытуемого. В итоге для каждого испытуемого определяется его балл по шкале Раша $D_1 = ln(P_1 / (1 - P_1))$. Например, при проверке знаний или подготовленности тестируемого более высокие значения D_1 соответствуют лучшей подготовленности и наоборот.

Аналогично вычисляется максимальное количество баллов, которое может быть приписано каждому пункту анкеты (*Max2*). В данном случае количество пунктов равно 7, максимальное значение по каждому пункту равно 1, поэтому *Max2* = 1 * 7 = 7. Далее для каждого пункта вычисляется отношение $P_2 = S_2 / Max2$ и значение $D_2 = ln((1 - P_2) / P_2)$. Обратите внимание на то, что в последней формуле числитель и знаменатель поменялись местами (что эквивалентно смене знака перед логарифмом). В классическом варианте модели Раша, которая разрабатывалась в контексте именно тестовых исследований, это было сделано для более «естественной» интерпретации результатов. В таком случае более высокие значения D2 соответствуют большей сложности тестового задания (вопроса) и наоборот. Следовательно, чем более сложным является вопрос, тем меньше вероятность правильного ответа на него и, следовательно, тем меньшее количество испытуемых отвечает на него правильно.

	Задания теста									
Испытуемые	**1**	**2**	**3**	**4**	**5**	**6**	**7**	S_1	P_1	D_1
A	1	0	1	1	0	1	0	4	0,57	0,29
B	0	1	0	1	1	0	0	3	0,43	-0,29
C	0	1	1	1	1	1	0	5	0,71	0,92
D	0	1	1	1	1	1	1	6	0,86	1,79
E	1	0	1	0	1	1	0	4	0,57	0,29
F	0	0	1	1	0	1	1	4	0,57	0,29
G	1	1	1	1	1	0	0	5	0,71	0,92
S_2	3	4	6	6	5	5	2			
P_2	0,43	0,57	0,86	0,86	0,71	0,71	0,29			

D_2	0,29	-0,29	-1,79	-1,79	-0,92	-0,92	0,92			

Дихотомические шкалы используются также в социологических, маркетинговых и иных опросах, например:

«Я планирую приобрести ноутбук (… название …) в течение ближайших 6 мес.»

- **0** Да - **0** Нет

Если в таком случае положительный отклик респондента («да») будет кодироваться как 1, а отрицательный («нет») как 0, то большие значения для респондента по шкале Раша будут соответствовать его более положительному (благоприятному) отношению к исследуемому явлению (объекту). Поэтому рассмотренный подход к обработке данных в дихотомической шкале может быть перенесен и в сферу социально-экономических исследований. В то же время в контексте измерения отношений логичнее получать более высокие значения по шкале Раша для пунктов, по которым большим количеством респондентов выражено благоприятное отношение к предмету исследования. В такой ситуации более логичным представляется использовать для оценки пунктов (индикаторов) «прямую» формулу перевода суммарных откликов в логиты, а именно $D_2 = ln(P_2 / (1 – P_2)$.

Такой подход к конвертированию данных может быть расширен и на многопозиционные порядковые шкалы, в частности, на шкалу Лайкерта. В отличие от дихотомической шкалы эта шкала имеет не один пороговый переход «правильно-неправильно», а несколько, например переходы между позициями «всегда» и «часто», «часто» и «иногда», «иногда» и «никогда». Переход от одной позиции к другой может рассматриваться как более или менее «сложный» для респондента, что отражает степень соответствия (*agreeability*) позиции действительному положению вещей.

Например, первый пороговый переход моделируется как вероятность выбора 1 («всегда») вместо 0 («часто») по каждому пункту анкеты:

$$P_{ni1}(x = 1/ B_n, D_i, F_1) = \frac{e^{(B-[D_i+F_1])}}{1+e^{(B-[D_i+F_1])}}.$$

где P_{ni1} – вероятность выбора респондентом n позиции «всегда» (1) вместо «часто» (0) по i-му пункту анкеты; F_1 – сложность первого порогового перехода; D_i – сложность пункта анкеты.

Рассмотрим пример. Предположим, что в анкету включено 7 пунктов (1,2, … 7), представляющих собой утверждение, по которому респондент выражает степень своего согласия/несогласия. Для оценки откликов используется 5-позиционная шкала Лайкерта с

позициями «Полностью не согласен» (ПНС), «Не согласен» (НС), «Нейтральное отношение» (Н), «Согласен» (С), «Полностью согласен» (ПС). Позиции шкалы кодируются от 1 до 5. Пусть на вопросы анкеты ответили 40 респондентов. Исходные данные (частоты откликов по каждому пункту) представлены в таблице.

Пункты анкеты	**ПНС (1)**	**НС (2)**	**Н (3)**	**С (4)**	**ПС (5)**	***S***	***P***	***D***	D_1
1	5	8	11	10	6	124	0,62	-0,49	0,49
2	9	7	10	9	5	114	0,57	-0,28	0,28
3	9	8	9	10	4	112	0,56	-0,24	0,24
4	7	9	9	8	7	119	0,595	-0,38	0,38
5	3	4	11	12	10	142	0,71	-0,90	0,90
6	5	9	9	9	8	126	0,63	-0,53	0,53
7	12	13	8	4	3	93	0,465	0,14	-0,14

Из таблицы видно, что при ответе на первый пункт анкеты 5 чел. выбрали ПНС, 8 чел. – НС, 11 – Н, 10 – С и 6 чел. выбрали ПС и т.д.

В наиболее простом варианте алгоритм перевода этих данных в шкалу Раша следующий. Рассчитывается балл соответствующего пункта (*item counts*) как сумма произведений количества респондентов, выбравших соответствующую категорию, и «балл» этой категории (*S*). Например, для первого пункта: 5 * 1 + 8 * 2 + 11 * 3 + 10 * 4 + 6 * 5 = 124. Максимальный балл для каждого пункта составляет 40 * 5 = 200. Тогда «вероятность» выбора первого пункта составит $P = 124 / 200 = 0{,}62$. Значение этого пункта в шкале Раша (в классическом варианте) определяется как $D = ln((1 - P) / P) = -0{,}49$. Аналогично вычисляются агрегированные оценки всех пунктов анкеты.

Здесь мы вновь сталкиваемся с вопросом об интерпретации результатов. В данном случае пункты, оцененные большим баллом, имеют меньшее значение, что полностью соответствует классической («тестовой») интерпретации измерений на основе модели Раша. В ситуациях измерения латентных переменных через фиксацию откликов респондентов, если большей степени измеряемого свойства соответствует большее значение в порядковой шкале, то более удобным и понятным является использование формулы $D = ln(P / (1 - P))$ для преобразования данных в шкалу Раша. Понятно, что результаты измерения будут одинаковыми по абсолютной величине и отличаются лишь знаком (значения D_1 в таблице). Все свойства шкалы (интервальность, независимость от выборки и т.д.) сохраняются.

Например, мы измеряем творческий климат в компании. При позитивно сформулированных пунктах анкеты (например «У меня достаточно времени для реализации

своих идей в компании») и при соответствующей кодировке позиций шкалы, мы, пользуясь последней формулой, получим данные, большее значение которых соответствует большей выраженности свойства. При использовании «классической» формулы мы получили бы, что организационный климат, в меньшей степени способствующий генерации и реализации новых идей, характеризовался бы большими значениями в шкале Раша.

Еще раз отметим, что при использовании описанного метода конвертирования данных пункты анкеты, на которые респонденты не ответили по каким-то причинам, не являются препятствием для использования модели Раша. Те пункты, по которым не получено отклика от респондентов, не учитываются при подсчете максимально возможного балла для каждого пункта. Например, если по первому пункту один из респондентов не выразил своего мнения, то значение максимального балла для этого пункта составляет не 40 * 5 = 200, а 39 * 5 = 195. В остальном процедура конвертирования данных не отличается от описанной выше.

В данном примере мы определяли лишь агрегированные оценки для пунктов анкеты и не рассчитывали значения для каждого респондента и для каждой конкретной позиции каждого пункта. Более сложный подход к преобразованию данных в шкалу Раша позволяет учесть и эти показатели, что необходимо, в частности, для анализа надежности откликов респондентов и пунктов шкалы (реализовано в WINSTEPS).

На основе модели Раша были разработаны другие измерительные модели, включающие в себя большее количество параметров. Например, известна двухпараметрическая модель, в которой параметр a соответствует кривизне характеристической кривой, т.е. дифференцирующей способности тестового задания:

$$P_i(B) = \frac{1}{1 + e^{-a(B-D)}}.$$

Трехпараметрическая модель, предложенная А. Бирнбаумом, включает в себя еще один дополнительный параметр (часто называемый «*параметром угадывания*»), определяющий точку пересечения характеристической кривой с осью ординат.

$$P_i(B) = c_j + (1 - c_j)\frac{1}{1 + e^{-a(B-D)}}.$$

О преимуществах и недостатках двух- и трехпараметрических моделей в сравнении с классической моделью Раша до сих пор спорят специалисты в области теории измерений.

2.4. Реализация модели Раша в программе WINSTEPS

Рассмотрим процедуру обработки исходных данных на основе модели Раша с помощью программы WINSTEPS. Демонстрационная версия этой программы доступна на сайте www.winsteps.com. В программе WINSTEPS используются более сложные алгоритмы, чем те, что мы иллюстрировали выше на простых примерах. Эти алгоритмы основаны на итерационных процедурах для подбора и оценки параметров модели Раша.

Начнем с примера теста, в котором данные были представлены в дихотомической шкале (7 заданий, 7 испытуемых). Исходный файл для работы в WINSTEPS имеет достаточно сложный формат, поэтому мы покажем, как этот файл может быть сформирован на основе данных, предварительно представленных в программе SPSS, с помощью конвертера, включенного в состав WINSTEPS. Исходные данные могут быть перенесены в SPSS из Excel или Word обычным копированием (более подробно работа с данными в SPSS рассматривается в следующих главах). Структура данных должна быть следующей: столбцы массива должны соответствовать испытуемым, а строки – заданиям. Предположим, что исходные данные перенесены в SPSS и переменные, соответствующие заданиям, обозначены как item1, item2, …, item7. Файл сохранен под именем Rasch-binary.sav. После запуска WINSTEPS необходимо на вопрос *Would you like help setting up your analysis?* ответить *No*. Затем из пункта меню SPSS (или SPSS/SAS) выбрать опцию *Select SPSS file and variables*. В диалоговом окне выбрать сохраненный ранее файл (Rasch-binary.sav). В открывшемся окне WordPad необходимо скопировать имена переменных, которые расположены ниже строки `;Variable Format Label` (в конце файла), и вставить их после строки `! Item Response Variables.`

В результате редактируемый файл в WordPad содержит следующий фрагмент.

```
;      Number of Cases: 7
;      Number of SPSS Variables: 7

! Person Label Variables. (Do not delete this line)

! Item Response Variables. (Do not delete this line)

ITEM1     ; F4.0
ITEM2     ; F4.0
ITEM3     ; F4.0
ITEM4     ; F4.0
ITEM5     ; F4.0
ITEM6     ; F4.0
ITEM7     ; F4.0

!Other SPSS Variables (ignored)
;Variable   Format    Label
ITEM1     ; F4.0
ITEM2     ; F4.0
ITEM3     ; F4.0
```

После сохранения файла, не закрывая его, нужно перейти в WINSTEPS и из пункта меню SPSS (SPSS/SAS) выбрать опцию *Construct WINSTEPS file from SPSS file*. В появившемся окне *Permanent Winsteps control file* задать имя файла с расширением .txt, например Rasch-binary.txt. В открывшемся окне WordPad необходимо проверить значения параметров `NI` (количество заданий `NI=7`), `ITEM1` (номер столбца, с которого начинаются числовые данные после строки `END NAMES: ITEM1=1`), `CODES` (коды переменной в исходных данных, `CODES = "01"`). В итоге подготовленный файл с исходными данными содержит следующий фрагмент.

```
ITEM1=1 ; Starting column of item responses
NI=7          ; Number of items
; SPSS Cases processed = 7
; datum: 0          count: 18
; datum: 1          count: 31
XWIDE = 1          ; this matches the biggest data value observed
CODES = "01" ; matches the data
NAME1 = 9          ; Starting column for person label in data record
; Person Label variables: in label: in line
NAMLEN = 0          ; Length of person label
&END ; Item labels follow:
ITEM1     ; Item 1
ITEM2     ; Item 2
ITEM3     ; Item 3
ITEM4     ; Item 4
ITEM5     ; Item 5
ITEM6     ; Item 6
ITEM7     ; Item 7
END NAMES
1011010 1
0101100 2
0111110 3
0111111 4
1010110 5
0011011 6
1111100 7
```

После сохранения файла WordPad можно закрыть. Далее необходимо перейти в WINSTEPS, нажать *Enter*, в меню *File* выбрать *Open File* и открыть сохраненный файл (Rasch-binary.txt). В командной строке программы WINSTEPS на запрос *Report output file name* указать имя файла, который будет содержать отчет о рабочем сеансе в WINSTEPS, например output.txt (если такой отчет не нужен, нажать *Enter*). На запрос *Extra specifications* нажать *Enter*. Ввод исходных данных в WINSTEPS завершен. На экран и в файл отчета выводится общая статистика по обработке исходных данных (ее описание мы опускаем).

Результаты конвертирования исходных данных в измерения Раша по всем испытуемым приводятся в отчете `PERSON STATISTICS` (меню *Output Tables, 17. PERSON: measure*).

```
PERSON STATISTICS:  MEASURE ORDER

+-------------------------------------------------------------------------+
|ENTRY    RAW                    MODEL|   INFIT  |  OUTFIT  |PTMEA|       |
|NUMBER  SCORE  COUNT  MEASURE  S.E. |MNSQ  ZSTD|MNSQ  ZSTD|CORR.| PERSON|
```

```
|-----------------------------------+---------+---------+-----+-------|
|     4      6      7    2.13    1.14|1.00    .2| .59    .0|  .40| 4     |
|     3      5      7    1.10     .92| .51  -1.2| .40   -.9|  .84| 3     |
|     7      5      7    1.10     .92| .99    .1| .92    .1|  .43| 7     |
|     1      4      7     .34     .85| .97    .0| .86   -.2|  .50| 1     |
|     5      4      7     .34     .85|1.22    .7|1.33    .8|  .24| 5     |
|     6      4      7     .34     .85|1.17    .6|1.18    .5|  .31| 6     |
|     2      3      7    -.36     .85|1.15    .5|1.04    .3|  .33| 2     |
|-----------------------------------+---------+---------+-----+-------|
| MEAN     4.4    7.0     .71     .91|1.00    .1| .90    .1|     |       |
| S.D.      .9     .0     .74     .10| .22    .6| .30    .5|     |       |
+-------------------------------------------------------------------------+
```

В столбце ENTRY NUMBER указаны номера испытуемых. Результаты измерений (столбец MEASURE) приведены по убыванию (в контексте тестирования это означает, что подготовленность (знания) испытуемого 4 по всем заданиям теста на 1,03 логита лучше, чем, например, испытуемого 3, и на 2,49 логита лучше, чем испытуемого 2). В столбце RAW SCORE приведены суммарные баллы испытуемых по исходным данным, в столбце COUNT – количество заданий, по которым проводился расчет, в столбце MODEL S.E. – ошибка измерения на основе модели Раша, в столбце PERSON указываются имена (метки) испытуемых (в данном случае имена испытуемых не задавались, поэтому повторяется их номер). Содержание других столбцов отчета будет рассмотрено в главе 4. В строках MEAN и S.D. указаны средние значения и стандартные отклонения по соответствующим значениям в столбцах.

Результаты измерений по заданиям теста приводятся в отчете `ITEM STATISTICS` (меню *Output Tables, 13. ITEM: measure*).

```
ITEM STATISTICS:  MEASURE ORDER

+-------------------------------------------------------------------------+
|ENTRY    RAW                    MODEL|   INFIT   |  OUTFIT  |PTMEA|      |
|NUMBER  SCORE  COUNT  MEASURE  S.E. |MNSQ  ZSTD|MNSQ  ZSTD|CORR.| ITEM |
|-----------------------------------+---------+---------+-----+------|
|     7      2      7    1.71     .88| .89   -.1| .90    .0|  .44| ITEM7|
|     1      3      7    1.01     .81|1.39   1.3|1.39   1.2| -.14| ITEM1|
|     2      4      7     .38     .81| .92   -.2| .87   -.3|  .44| ITEM2|
|     5      5      7    -.31     .87|1.01    .1| .85   -.1|  .32| ITEM5|
|     6      5      7    -.31     .87| .97    .1| .99    .2|  .29| ITEM6|
|     3      6      7   -1.24    1.10| .73   -.2| .47   -.3|  .59| ITEM3|
|     4      6      7   -1.24    1.10|1.03    .3| .84    .2|  .21| ITEM4|
|-----------------------------------+---------+---------+-----+------|
| MEAN     4.4    7.0     .00     .92| .99    .2| .90    .1|     |      |
| S.D.     1.4     .0    1.03     .12| .18    .5| .25    .5|     |      |
+-------------------------------------------------------------------------+
```

В этом отчете результаты измерений (MEASURE) также приводятся по убыванию: от более сложных – к менее сложным заданиям (например, задание 7 на 0,7 логита сложнее задания 1 и т.д.). В столбце ENTRY NUMBER указаны номера заданий; RAW SCORE – суммарные баллы по каждому заданию; COUNT – количество испытуемых, по которым

проводился расчет сложности; MODEL S.E. – ошибка измерения сложности заданий по модели Раша; ITEM – имена (метки) заданий. В строках MEAN и S.D. указаны средние значения и стандартные отклонения по соответствующим значениям в столбцах.

Как видим, результаты измерений отличаются от тех, что были получены нами при расчете «вручную» по простейшему алгоритму, но это обстоятельство связано с тем, что в WINSTEPS применяются более точные алгоритмы. В алгоритмах, которые реализованы в WINSTEPS и других специальных программных продуктах по *Rasch Measurement*, учитывается много дополнительных факторов, например, делается поправка на размер выборки, распределение и разброс данных, учитываются случайные выбросы, рассчитываются ошибки калибровки данных и измерений.

Аналогично могут быть обработаны исследовательские данные, представленные в других шкалах. В главе 4 будет рассмотрен пример обработки данных по оценке организационного климата с помощью WINSTEPS и показано, как оценивается качество полученных измерений.

Модель Раша трактуется как модель «объективных измерений», которые не зависят (или зависят в значительно меньшей степени, чем оценочные результаты по данным в неметрических шкалах) от респондентов и измерительных инструментов. Эта модель также обеспечивает возможность корректного использования очень широкого круга статистических методов. Эти обстоятельства позволяют считать результаты, получаемые на основе применения модели Раша, более надежными, чем результаты, полученные по исходным данным, например в шкале Лайкерта.

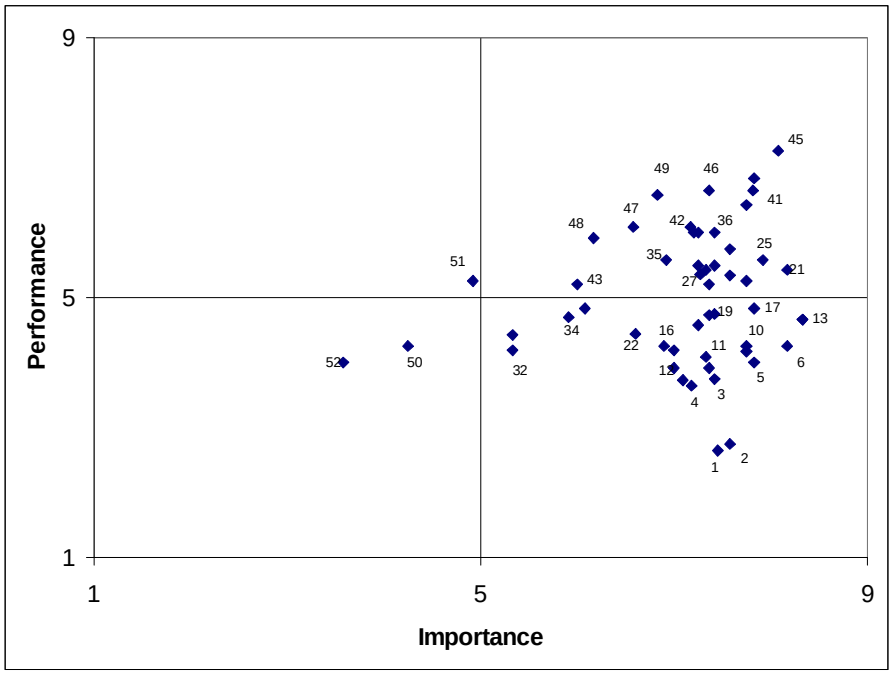
Поэтому когда результаты, полученные по данным в порядковой шкале, согласуются с результатами по данным, конвертированным в шкалу Раша, это можно считать дополнительным подтверждением надежности полученных результатов, используемого исследовательского инструмента, подхода, гипотезы и т.д. Как правило, при хорошо спланированном и проведенном эксперименте, при использовании обоснованных и надежных инструментов (см. главу 4), а также при правильном использовании процедуры конвертирования исходных данных в измерения по шкале Раша, результаты, полученные по исходным данным (оценки) и по данным, конвертированным в шкалу Раша (измерения), в целом хорошо согласуются[4]. Однако конвертация данных позволяет выявить менее заметные различия и связи, провести более глубокий и математически корректный анализ данных, сделать более обоснованные выводы.

[4] В то же время следует отметить, что тема согласованности результатов традиционных моделей измерения и модели Раша представлена в литературе довольно слабо.

Вместе с тем не следует забывать, что модель Раша – это прежде всего *математическая модель*, поэтому не следует руководствоваться только формальными соображениями при принятии тех или иных выводов. Всегда нужно помнить о содержательном наполнении полученных результатов и руководствоваться также содержательными критериями, в том числе здравым смыслом и практической важностью полученных результатов (об этом еще пойдет речь во второй части книги).

2.5. Метод QIPM: пример компромисса между точностью и простотой

Обсуждая виды шкал и, в частности, шкалу Лайкерта, мы уже указывали, что исследователи часто интерпретируют данные в порядковых шкалах как интервальные и используют различные методы обработки данных в ущерб строгости и корректности их использования. Очень наглядно эта ситуация проявляется в использовании подхода к выявлению приоритетных факторов, определяющих деятельность организации: QIPM (*Quality Improvement Priority Matrix* – матрица приоритетов улучшения качества). QIPM – это простой и наглядный метод, который заключается в оценке важности (*importance*) различных факторов, определяющих деятельность организации, и степени их осуществления-реализации (*performance*). Вариантами такого подхода является оценка факторов по таким параметрам, как, например, «важность» и «удовлетворенность» (*satisfaction*) клиентов компании или «влияние на бизнес» (*business value impact*) и «простота реализации» (*ease of implementation*). Оцененные факторы (после агрегирования индивидуальных оценок) отображаются на графике в соответствующих осях, например «важность-реализованность». Приоритетными, т.е. требующими мер по изменению и совершенствованию, считаются факторы с высокой важностью, но низкой реализованностью (факторы, которые попадают в юго-восточный квадрант). Пример использования этого подхода (при оценке «важности» и «реализованности» по шкале от 1 до 9 баллов) приведен на рисунке.



Если данные оцениваются в других категориях, то приоритетность может определяться попаданием в другой квадрант. Например, для оценки соотношения «влияние на бизнес – простота реализации» в качестве приоритетных будут рассматриваться факторы, попадающие в северо-восточный квадрант.

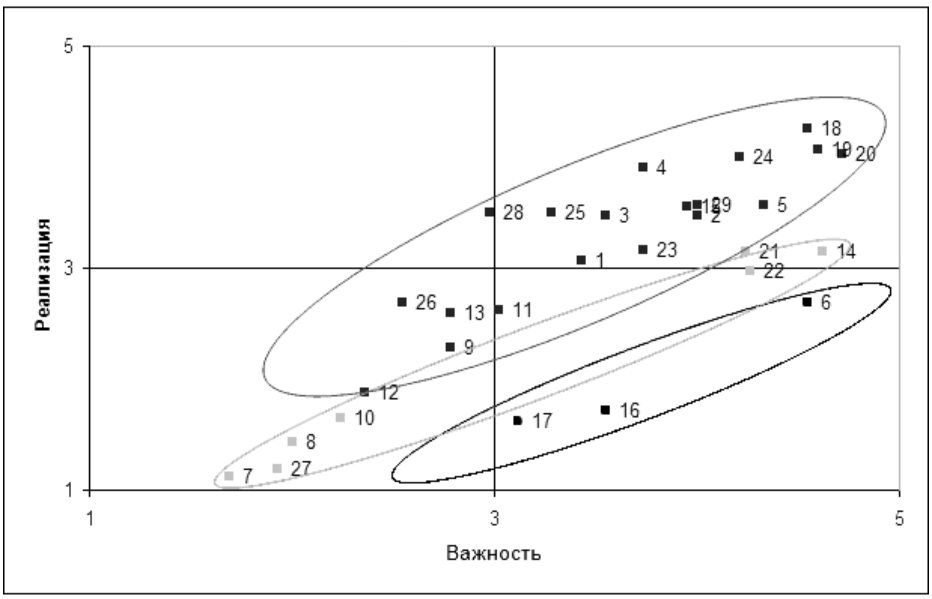
Такой подход не требует ни много времени и ресурсов для сбора данных, ни использования сложного аналитического инструментария для их обработки и анализа. В то же время метод нагляден и удобен для принятия решений, он широко используется в различных государственных и коммерческих организациях США и Европы. В частности, автору книги известны примеры использования такого подхода компанией IBM, Министерством финансов США, Университетом Дж. Вашингтона (США). Работы в области менеджмента качества с использованием QIPM были удостоены престижной премии Malcolm Baldrige National Quality Award в 1995 г.

Однако такой подход является довольно поверхностным и не обеспечивает возможности проведения серьезного анализа ситуации в организации и выделения приоритетов ее развития[5]. Более того, этот подход почти всегда используется некорректно с математической точки зрения, поскольку агрегированные оценки факторов, позиционируемые на графике, рассчитываются как среднее арифметическое индивидуальных оценок, отражаемых в порядковой шкале. Использование метода могло бы быть более корректным в случае преобразования данных в интервальную шкалу, например с помощью модели Раша. Однако практика показывает, что такое преобразование часто не изменяет значительным образом

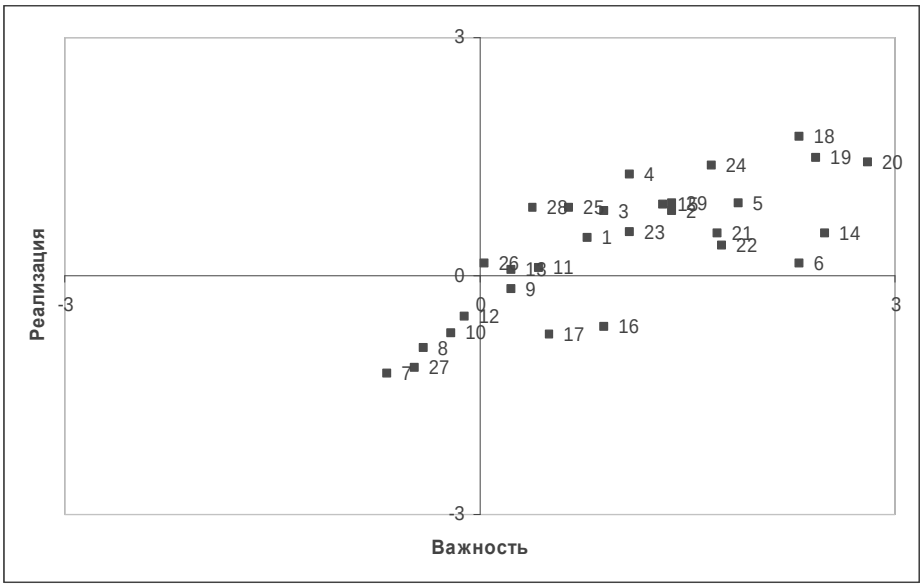
[5] Автор книги совместно с коллегами из Университета Дж. Вашингтона (Вашингтон, США) предложил несколько вариантов усовершенствования этого подхода (Dubina and Umpleby, 2006).

результаты (для такого подхода), поэтому исследователи часто жертвуют математической корректностью и точностью измерений в пользу простоты и удобства использования метода.

В качестве примера приведем результаты использования метода QIPM для мониторинга и совершенствования маркетинговой деятельности компании, занимающейся комплексным оснащением объектов торговли, общественного питания и пищевых производств. Для сбора исходных данных была разработана анкета, включающая в себя набор факторов, степень важности и качество реализации которых оценивались клиентами компании по шкале от 1 до 5. Анкета включала пункты по таким темам, как ассортиментная политика, качество обслуживания, конкурентные преимущества, ценовая политика, работа сервисной службы, эффективность рекламы и др. (всего 29 факторов). После обработки была построена матрица QIPM, а по определенному алгоритму выделены 3 кластера приоритетности факторов.



Для проверки сопоставимости результатов, исходные данные из шкалы Лайкерта конвертировались в шкалу Раша и тоже представлялись в матрице QIPM для визуального сравнение расположения факторов в матрице.



Результаты по оценке важности и реализации в двух шкалах дополнительно сравнивались после ранжирования факторов соответственно по важности и реализации. В таблице приведены идентификационные номера факторов, ранжированных по двум характеристиками (*I* и *P*).

№ в рейтинге	*I*		*P*	
	Шкала Лайкерта	Шкала Раша	Шкала Лайкерта	Шкала Раша
1	20	20	18	18
2	14	14	19	19
3	19	19	20	20
4	6	6	24	24
5	18	18	4	4
6	5	5	5	5
7	22	22	29	29
8	21	21	15	15
9	24	24	25	25
10	2	2	28	28
11	29	29	2	2
12	15	15	3	3
13	23	4	23	23
14	4	23	14	14
15	16	3	21	21
16	3	16	1	1
17	1	1	22	22
18	25	25	6	6
19	17	17	26	26
20	11	11	11	11
21	28	28	13	13
22	9	9	9	9
23	13	13	12	12
24	26	26	16	16
25	12	12	10	10
26	10	10	17	17

27	8	8	8	8
28	27	27	27	27
29	7	7	7	7

Сопоставление результатов показывает, что они почти полностью идентичны[6]. Поэтому в такого рода задачах, где высокая точность не требуется, нет необходимости использовать сложный математико-статистический инструментарий для измерений, а достаточно ограничиться оценками, что чаще всего и происходит на практике.

В качестве материала для выполнения задания, связанного с использованием QIPM, читателям предлагается выборка из данных, полученных автором книги совместно с профессором кафедры теории менеджмента Университета Дж. Вашингтона (США) С. Амплби. Метод QIPM использовался для мониторинга деятельности кафедры с целью определения приоритетных направлений ее развития. Преподаватели кафедры оценивали в категориях «важность–реализованность» такие факторы, как заработная плата, компьютерные лаборатории, финансовая поддержка научных исследований, расписание занятий, библиотечные ресурсы, помощь в подготовке заявок на получение грантов, зарубежные командировки, уровень знания английского языка студентами, функционирование веб-сайтов кафедры и университета и др. (всего оценивалось 52 фактора). Для оценки факторов по критериям «важность–реализованность» по методу QIPM использовалась порядковая шкала с позициями от 1 до 9. Предложенные подходы для анализа и интерпретации данных, полученных с помощью QIPM, и результаты использования этих подходов были представлены и обсуждены на годовом собрании преподавателей и сотрудников кафедры теории менеджмента Университета Дж. Вашингтона. По итогам обсуждения был разработан проект плана стратегического развития кафедры и сформулированы предложения по использованию разработанных методов на других кафедрах университета.

[6] Более строгие подходы и методы сопоставления распределений двух наборов данных с помощью специальных критериев приведены во второй части книги.

Глава 3. Формирование и анализ выборочной совокупности

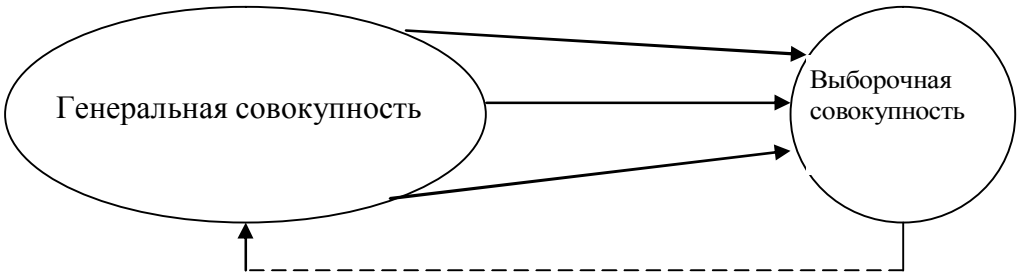
3.1. Особенности выборочного метода исследования

Основная проблема, часто возникающая при проведении исследований, заключается в том, что количество объектов, которые мы должны изучить, слишком велико и не поддается исследованию из-за ограниченности временных, финансовых, технических, человеческих и прочих ресурсов, имеющихся у исследователя. Все множество (полный набор) элементов, относительно которого осуществляется заключение по результатам измерения, называется *генеральной совокупностью* – *population* (иногда калька с англоязычного термина «популяция» используется и в русскоязычной литературе). Соответственно, объектами измерения являются элементы генеральной совокупности. Исследование *всех* элементов генеральной совокупности называется *сплошным исследованием*.

Например, необходимо оценить политические предпочтения граждан накануне выборов. Возможно, наиболее точные результаты были бы получены после опроса всех потенциальных избирателей (сплошное исследование). Но это невозможно в стране, электорат которой составляют несколько десятков миллионов человек. Даже для гораздо меньших групп (студенты университета, жители городского района и т.д.), сбор данных по всем членам группы чаще всего затруднен или невозможен. У исследователя практически всегда недостаточно времени и ресурсов для того, чтобы измерить характеристику каждого объекта генеральной совокупности.

Для решения этой проблемы исследователь с помощью определенных методов формирует выборку. *Выборка* (*sample*) – это ограниченная по численности группа объектов (например респондентов), отбираемая из генеральной совокупности для изучения ее свойств. Цель формирование выборки – получение достоверных выводов о свойствах исследуемой генеральной совокупности. Исследование объектов выборки называется *выборочным исследованием*. После анализа результатов выборочного исследования проверяется ряд гипотез для обоснования возможности переноса этих результатов на всю генеральную совокупность.

Методы формирования выборки



Проверка статистических гипотез

Выборочное исследование имеет целый ряд преимуществ по сравнению со сплошным. Во-первых, это существенное снижение затрат, во-вторых, сокращение времени на получение данных и их обработку. Часто результаты выборочного исследования рассматриваются как менее точные и надежные, чем результаты сплошного исследования. Однако это не совсем верное суждение. При применении корректных подходов выборочное исследование обеспечивает очень точное представление о генеральной совокупности. Более того, выборочное исследование подчас обеспечивает лучшее качество измерений, чем исследование всей генеральной совокупности, по причине более качественного сбора данных и их анализа. Например, выборочное исследование может обеспечить лучшее интервьюирование, наблюдение, обработку и другое, чем в случае измерения всех элементов генеральной совокупности. В некоторых ситуациях обследование всех элементов популяции вообще невозможно, например, если исследование предполагает разрушение или порчу объекта (контроль качества продукции и др.).

Выборочное исследование не имеет смысла, если *все* элементы генеральной совокупности различны (дисперсия измеряемых параметров значительна). Например, если мы выясняем мнение по определенной теме, а *все* представители генеральной совокупности имеют *разное* мнение, нам необходимо будет опросить всех представителей, т.е. провести сплошное исследование.

При формировании выборки необходим такой отбор элементов (наблюдений), чтобы сформированная выборка была *репрезентативной* (представительной) по отношению к генеральной совокупности, из которой она была извлечена. *Репрезентативность* выборки – способность выборки представлять изучаемые явления достаточно полно с точки зрения их изменчивости в генеральной совокупности. Эта характеристика показывает, можно ли распространить результаты, полученные на данной выборке, на всю генеральную

совокупность, из которой взята эта выборка. Репрезентативная выборка должна воспроизводить все важные характеристики изучаемой генеральной совокупности. Например, если мы определяем среднюю заработную плату работников отрасли, и генеральная совокупность на 60% состоит из женщин и на 40% – из мужчин, то и выборка из нее должна отражать те же пропорции. По выражению Дж. Хили, репрезентативная выборка – это генеральная совокупность в миниатюре. Для формулирования и обоснования статистических выводов репрезентативность выборок играет решающую роль: если выборки не являются репрезентативными, то обобщение выводов, полученных на основании выборки, на генеральную совокупность становится неприемлемым.

Репрезентативность характеризует соответствие показателей, полученных в результате выборочного наблюдения, показателям, свойственным для всей генеральной совокупности. Расхождение между указанными показателями представляет собой *ошибку репрезентативности*, которая может быть *случайной* (сбалансированной) или *систематической* (несбалансированной, направленной). Примером систематической ошибки репрезентативности выборки является прогнозирование президентских выборов в США в 1936 г., когда одним из главных кандидатов на этот пост был Ф. Рузвельт, претендовавший на второй срок. Для исследования была сформирована очень большая выборка, но с систематической ошибкой – проводилось изучение мнений верхнего и среднего классов, тогда как Рузвельта в основном поддерживал низший класс, который и обеспечил его сокрушительную победу на выборах. Причинами систематической ошибки могут быть также время, день или место проведения измерений, метод формирования выборки и др.

Репрезентативность – это не формализуемая характеристика. Однозначно и точно сказать, основываясь на четких показателях, что данная выборка является репрезентативной, невозможно. Но о репрезентативности выборки, т.е. о том, насколько результаты выборочного исследования соответствуют (или могут соответствовать) результатам сплошного исследования, можно *судить* по таким критериям, как способ формирования выборки, ее объем и вариативность (изменчивость) данных. Эти критерии будут подробно рассмотрены в следующих параграфах этой главы.

3.2. Методы формирования выборки

Известно несколько методов формирования выборок из генеральных совокупностей. В наиболее общем виде формируемые выборки можно разделить на две группы: *вероятностные* и *условно-вероятностные*. Вероятностная выборка часто называется также *случайной* выборкой. Однако термин «вероятностная выборка» все же предпочтительнее, так

как в повседневной речи под случайным часто подразумевается то, что происходит «по случайному стечению обстоятельств», «наугад» и т.п. Но на самом деле вероятностные выборки формируются с помощью таких подходов, которые предусматривают тщательное следование алгоритму, не оставляя место бессистемности и «случайности» в обыденном понимании этого слова.

Условно-вероятностные выборки формируются «по желанию», «по присутствию», «по доступности» и другим аналогичным принципам и критериям. Очевидно, что лишь к вероятностным выборкам можно в полной мере применить методы получения статистических выводов о генеральной совокупности. В то же время условно-вероятностные выборки также весьма часто используются в исследовательской практике, особенно применительно к небольшим группам. В частности, среди исследователей, работающих в вузах, распространена практика использования студентов в качестве респондентов для «отладки» исследовательских подходов и инструментов. При этом, как правило, опрашиваются студенты в тех студенческих группах, с которыми и работает преподаватель-исследователь. Такие «удобные» для исследователя выборки, безусловно, полезны для определенных целей (например, для предварительного формулирования исследовательских гипотез и тестирования различных форм опроса перед тем, как начать выполнение более сложного и дорогого проекта). Сформировать условно-вероятностные выборки обычно достаточно просто и недорого. Однако чаще всего такие выборки нерепрезентативны, и переносить результаты, полученные по таким выборкам, на другие группы нельзя. Следовательно, нельзя обосновывать статистические выводы и гипотезы только на выборках, не являющихся вероятностными, даже при наличии очень веских доказательств в пользу обосновываемых гипотез.

Методы формирования вероятностных выборок основаны на принципе равновероятностного отбора (*equal probability of selection method*): вероятность попадания в выборку должна быть одинакова для любого элемента (объекта) исследуемой генеральной совокупности. Вероятность того, что выборка, сформированная по такому принципу, будет репрезентативной, очень высока, однако гарантировать это все же невозможно.

Наиболее часто встречающиеся методы формирования выборок по такому принципу включают в себя:

а) список всех элементов (объектов) исследуемой генеральной совокупности;

б) систему отбора элементов из этого списка, которая бы обеспечивала равную вероятность включения в выборку любого элемента из списка.

Для отбора списочных элементов может использоваться процедура генерирования случайных чисел, включенная во многие статистические и математические программные

продукты. В Excel такая процедура может быть осуществлена выбором функции СЛЧИС() (в англоязычной версии – RAND()). Эта функция возвращает равномерно распределенное случайное число в диапазоне от 0 до 1. Чтобы получить случайное вещественное число в интервале между числами a и b, можно использовать следующую формулу: СЛЧИС()*$(b - a)$ + a. Если, например, элементами генеральной совокупности являются студенты факультета в количестве 2000 чел., и всем им присвоены соответствующие идентификационные номера (от 1 до 2000), то формирование вероятностной выборки может быть сделано с помощью функции СЛЧИС() * 1999 + 1. Полученный результат сопоставляется с номером из списка, и таким образом отбирается соответствующий студент в выборку. Когда формируемая выборка достигает желаемого объема, процедура завершается. Если идентификационный номер попадает в выборку только один раз, то получаем *вероятностную (случайную) бесповторную выборку*, иначе имеем *случайный повторный отбор*. Мы можем игнорировать (не включать в выборку) повторяющийся номер, чтобы получить *случайную бесповторную выборку*, но, строго говоря, такая практика вредит случайности процесса отбора. Однако если объем формируемой выборки намного меньше объема генеральной совокупности, то тогда вероятность извлечения одного и того же элемента дважды весьма мала, поэтому игнорирование повторений не приводит к искажению наших заключений о генеральной совокупности.

Для большого списка элементов и при отсутствии компьютерных средств, с помощью которых можно осуществить указанную процедуру, такой подход трудно реализуем. Чтобы сэкономить время и другие ресурсы, в этой ситуации исследователи часто формируют *систематическую вероятностную выборку*. При этом в выборку случайно отбирается только первый элемент, а затем каждый i-й элемент списка генеральной совокупности. Например, если есть список из 10000 элементов генеральной совокупности и необходима выборка объемом 200, нужно случайным образом отобрать один элемент списка, а затем взять элементы списка, отстоящие через каждые 10000 / 200, т.е. 50 номеров от него (вверх и вниз по списку).

При формировании систематических выборок нет строгого соответствия принципу равновероятностного отбора, так как после того, как отобран первый элемент, для остальных элементов вероятность отбора уже не является одинаковой, как это было ранее. Происходящий при этом рост вероятности ошибки очень незначителен, но лишь при условии, что список, из которого отбираются элементы, является сам по себе случайным (хорошо перемешанным) или по крайней мере нециклическим (непериодическим) по отношению к измеряемым признакам объекта.

Следующий метод формирования вероятностной выборки – это получение *стратифицированной* (*типической*, *районированной*) выборки. Этот метод гарантирует репрезентативность выборки в *отношении выбранных признаков объекта*. Список элементов генеральной совокупности разбивается на подсписки в соответствии с категориями нужного признака, и затем формируется выборка на основе этих подсписков. И если из каждого подсписка извлекается количество элементов, пропорциональное доле соответствующей категории в генеральной совокупности, то полученная выборка будет репрезентативной для данной генеральной совокупности в отношении выбранного признака.

Например, необходимо сформировать выборку объемом 50 студентов, в которой был бы пропорционально представлен каждый из 10 факультетов университета. Допустим, на одном факультете обучаются 30% студентов университета, а на другом – лишь 2% всех студентов, то ранее описанные процедуры формирования выборок дадут выборку с очень малым (или даже с нулевым) количеством студентов второго факультета. Если же генеральная совокупность будет предварительно разбита на подсписки (по факультетам), то тогда можно отобрать ровно 5 студентов соответствующего факультета.

Если же мы проводим стратифицированный отбор по *принципу пропорциональной численности единиц типических групп*, то объем выборки определяется по формуле

$$n_i = n\frac{N_i}{N},$$

где n_i – объем выборки из типической группы; n – общий объем выборки; N_i – объем типической группы; N – объем генеральной совокупности.

В этом случае объем типической выборки от первого факультета составит 15 студентов, от второго – 1 студент.

Главным ограничением на применение метода формирования стратифицированных выборок является то, что очень часто точный состав генеральной совокупности неизвестен. Если у нас нет никакой информации о генеральной совокупности, то мы не сможем построить схему стратифицирования и определить, сколько элементов следует отобрать из каждого подсписка.

Впрочем, эта же трудность осложняет применение *всех* описанных процедур для формирования выборок. Во многих случаях список элементов генеральной совокупности отсутствует или получить его крайне сложно (не существует полного списка граждан страны или даже области, эти списки устарели и т.д.). Еще более сложно изучить (измерить, оценить) все объекты (элементы списка), отобранные в выборку, например, часть отобранных респондентов могут отказаться участвовать в опросе.

В случае неполноты списков может формироваться *групповая (кластерная) выборка*, которая предполагает отбор *групп* элементов (кластеров), а не отдельных элементов. Такими кластерами могут являться, например, административные районы города. Из них можно сформировать выборку. Кластеры должны быть примерно одного объема, так как (это будет показано во второй части книги) большинство тестов для проверки гипотез работают при предположении примерно равного (по крайней мере однопорядкового) объема выборок.

Затем внутри каждого кластера (района, попавшего в выборку) формируется подвыборка по микрорайонам (избирательным округам, кварталам и т.п.). Подобная процедура может быть продолжена и дальше, вплоть до жителей отдельных домов. Если все подвыборки являются вероятностными, то в результате будет получена выборка, которая с большой вероятностью окажется репрезентативной для всего города (причем она будет получена без использования списка всех жителей города!).

Однако групповая выборка менее надежна, чем «действительно» случайные выборки сравнимого объема. Отчасти это является результатом наличия нескольких этапов отбора. Отбор элементов на каждом этапе вносит свою вероятность ошибки формирования нерепрезентативной, но если использование других методов формирования выборки невозможно, то формирование групповой выборки может оказаться единственной альтернативой.

3.3. Анализ точности результатов выборочного исследования: доверительный интервал

Довольно распространенным является мнение о том, что увеличение точности результатов выборочного исследования пропорционально увеличению числа элементов выборки. Это мнение не совсем справедливо. Более 90% ошибок измерений имеют «невыборочные» источники, и лишь 10% ошибок появляются в результате неправильного формирования выборки (Cooper & Shindler, 1998). Мнение об обязательном повышении точности с ростом численности выборки основывается на трех заблуждениях, к сожалению, разделяемых многими исследователями.

Заблуждение 1: Чем больше выборка, тем она репрезентативнее.

То, что это не всегда так, демонстрирует пример с прогнозированием исхода президентских выборов в США в 1936 г.: выборка объемом свыше 2 млн респондентов не обеспечила репрезентативности, прогноз оказался не верен. Выборка, состоящая из 100 элементов, может обеспечить лучшую репрезентативность, чем выборка, состоящая из 10000 элементов.

Заблуждение 2: Выборка должна состоять из как минимум 10% элементов генеральной совокупности.

На самом деле *абсолютный* объем выборки – фактор гораздо более важный, чем объем выборки *относительно* объема генеральной совокупности. Так, например, выборка объемом 100 элементов, извлеченная из 1000 элементов генеральной совокупности, может обеспечить примерно такую же точность оценки, как выборка объемом 1000 элементов, отобранных из 100 миллионов элементов, формирующих генеральную совокупность. Например, при опросах общественного мнения о возможном исходе выборов президента США обычно формируется выборка, насчитывающая «всего» 1500–2000 респондентов, тогда как весь электорат США насчитывает около 150 млн чел. Однако при формировании выборки из большой популяции возникает проблема отбора «соответствующих» элементов.

Заблуждение 3: Заранее невозможно ответить на вопрос о необходимой и достаточной численности выборки.

Это отчасти так. Действительно, требуемый объем выборки можно определить после того, как проведен анализ хотя бы нескольких случайным образом предварительно отобранных элементов генеральной совокупности. Но это может (и должно) быть сделано на стадии предварительного тестирования, что является обязательным этапом любого серьезного исследования. Кроме того, сделать заключение о необходимом объеме выборки можно по результатам предыдущих исследований.

Необходимый объем выборки представляет собой функцию вариации измеряемых параметров генеральной совокупности и точности оценки этих параметров, требуемой исследователем. Например, в зависимости от требуемой точности для одной и той же совокупности может быть достаточно 50, 500 или 5000 элементов выборки. И напротив, заданную точность могут обеспечить 100, 1000 или 10000 элементов в зависимости от того, насколько велика дисперсия полученных измерений.

Общим правилом при формировании выборки является следующее: чем больше дисперсия оцениваемых параметров генеральной совокупности, тем больший объем выборки требуется для того, чтобы обеспечить требуемую точность. Например, если мы выясняем мнение по определенной теме, когда все представители популяции (элементы генеральной совокупности) имеют предположительно одинаковое мнение, нам достаточно опросить одного представителя. Если возможны 2 мнения, нам необходима выборка из по меньшей мере 2-х элементов (но при этом мы должны позаботиться о том, чтобы в выборку попали носители этих различных мнений, и нам понадобится большая выборка, чтобы определить частотность встречаемости этих мнений). Если возможны 10 точек зрения, нужна еще большая выборка, и т.д.

С практической точки зрения главными факторами, определяющими объем выборки, являются желаемая точность оценки и бюджет исследования. По поводу бюджета особые комментарии не требуются. Если, скажем, в бюджете исследователя имеется 1000 руб., а телефонный опрос каждого респондента требует в среднем 10 руб., то максимальный объем выборки – 100 элементов. Если нужна выборка большего объема, то необходимо выбирать другие, более дешевые формы опроса. Более интересным в контексте тематики всей книги и данной главы в частности является вопрос об объеме выборки с позиций точности оценки измеряемых параметров.

Оценка параметра может быть *точечной* или *интервальной*. Примером точечной оценки является результат гипотетического выборочного исследования, установившего, что 61% опрошенных пользователей телефона марки «Brand Name» остались довольны сделанной покупкой, или 73% граждан, участвовавших в опросе общественного мнения, поддерживают новые инициативы правительства. Как мы уже говорили, результаты выборочного исследования нельзя напрямую переносить на генеральную совокупность, т.е. мы не можем сказать, что *именно* 61% *всех* пользователей данной марки телефонов ими довольны, или что *точно* 73% *всех* граждан страны поддерживают новый закон. Очевидно, что здесь должен быть некоторый допуск, или интервал, в пределах которого можно формулировать результат, например: от 45 до 75% пользователей телефона остались довольны сделанной покупкой и т.п. Последний подход, безусловно, более оправдан при выборочном исследовании.

Точность интервальной оценки параметра, измеряемого при выборочном исследовании, определяется двумя показателями:

а) интервалом, в котором ожидается обнаружить оцениваемый параметр;

б) вероятностью обнаружения этого параметра в данном интервале.

Эти два показателя объединяет понятие *доверительного интервала*. Процедура определения доверительного интервала основана на *центральной предельной теореме* – одной из основных теорем теории вероятностей и статистики. Согласно этой теореме, распределение средних значений выборок, извлекаемых из одной и той же совокупности, соответствует нормальному распределению. Более того, когда выборки становятся достаточно большими, то выборочные средние подчиняются нормальному закону, даже если исходная переменная не является нормальной. Среднее значение всех выборочных средних равно среднему значению генеральной совокупности (μ), а стандартное отклонение выборочных средних (s_x) определяется по следующей формуле:

$$s_x = \frac{s}{\sqrt{n}},$$

где s – стандартное отклонение по генеральной совокупности; n – объем выборки.

Величина σ_x называется также *стандартной ошибкой среднего*[7]. Вычисление стандартной ошибки среднего основывается на предположении нормальности измеряемой переменной. Если это предположение не выполнено, то оценка может оказаться неверной, особенно для малых выборок.

Естественным образом возникает вопрос о том, какой объем выборки может считаться «достаточно большим». Известно эмпирическое правило, согласно которому принимается, что если объем выборки (n) равен 100 или более, то применима центральная предельная теорема, и допущение о нормальности распределения всех возможных выборочных средних может быть принято. Показано, что при увеличении объема выборки до 100 и более качество оценки стандартной ошибки среднего улучшается и без предположения нормальности выборки. Если же n меньше 100, то нужно иметь веские доказательства нормальности распределения генеральной совокупности, и только в этом случае можно полагать, что распределение, которому подчиняются выборочные статистики, является нормальным.

Поскольку в большинстве случаев значение стандартного отклонения по генеральной совокупности (σ) неизвестно, его заменяют выборочным стандартным отклонением (s), и стандартная ошибка среднего рассчитывается как

$$s_x = \frac{s}{\sqrt{n}}$$

(предполагается, что выборка формируется в результате случайного повторного отбора).

Отсюда следует, что стандартное отклонение по выборке определяет интервал попадания среднего по всей генеральной совокупности. Стандартная ошибка среднего зависит от стандартного отклонения по выборке и ее объема. Например, если стандартное отклонение по выборке $s = 0{,}74$ и $n = 10$, то стандартное отклонение выборочных средних равно

$$s_x = \frac{0{,}74}{\sqrt{10}} = 0{,}234.$$

Если стандартное отклонение по выборке уменьшается в *два* раза, то оцениваемое изменение измеряемого параметра по генеральной совокупности также уменьшается в *два* раза:

$$s_x = \frac{0{,}37}{\sqrt{10}} = 0{,}117.$$

[7] Термин был впервые введен Юлом (Yule) в 1897 г. Также используется термин *средняя ошибка выборки.*

Предположим, что по результатам выборочного исследования мы получаем стандартную ошибку среднего

$$s_x = \frac{0{,}8}{\sqrt{25}} = 0{,}16.$$

При увеличении количества респондентов в *четыре* раза, при том же самом значении стандартного отклонения по выборке, мы можем обеспечить увеличение точности лишь в *два* раза:

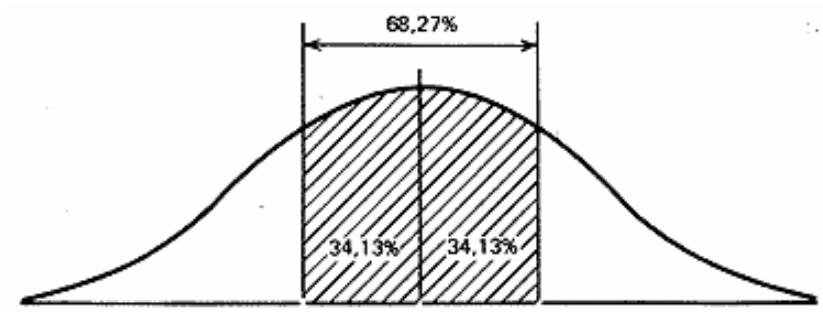
$$s_x = \frac{0{,}8}{\sqrt{100}} = 0{,}08.$$

При бесповторном случайном отборе стандартное отклонение выборочных средних рассчитывается как

$$s_x = \frac{s}{\sqrt{n}}\sqrt{1-\frac{n}{N}}.$$

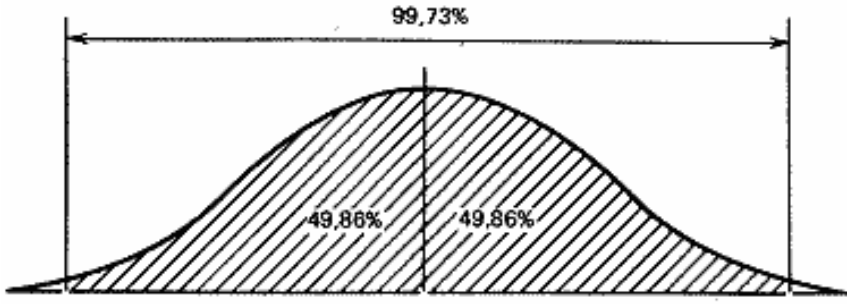
Очевидно, что для применения этой формулы (и этого механизма отбора соответственно) нам должна быть известна численность генеральной совокупности N.

При нормальном распределении данных большая их часть (68,27%) располагается в пределах одного стандартного отклонения (z = 1) по обе стороны от среднего значения ($\bar{x}$), вне зависимости от величины стандартного отклонения[8]. Это означает, что с вероятностью 0,68 среднее значение параметра по генеральной совокупности будет попадать в интервал $\bar{x} \pm \sigma_x$.



В пределах трех стандартных отклонений (z = 3) умещается почти вся генеральная совокупность – 99,73%. Это означает, что с вероятностью >0,99 среднее значение параметра по генеральной совокупности будет попадать в интервал $\bar{x} + 3\sigma_x$.

[8] Это следует из свойств нормального распределения случайной величины (см. также материал по теме 1).



Используя формулу, определяющую закон нормального распределения (функция Лапласа), можно определить значения *z*, которые соответствуют различным значениям вероятности попадания среднего значения параметра по генеральной совокупности в интервал $\Delta = \bar{x} + z\sigma_x$ (см. табл.).

z	Площадь покрытия, %	Вероятность попадания в интервал, %
1,00	68,27	68,27
1,65	90,10	90,10
1,96	95,00	95,00
3,00	99,73	>99

Значения *z* для других уровней вероятности можно определить из таблицы в приложении 1.1 или с помощью функции Excel НОРМСТОБР(*P*), возвращающей значение *z* при заданной вероятности *P* (более подробно особенности работы с этой функцией будут рассмотрены во второй части, посвященной процедурам обоснования статистических выводов).

Величина Δ называется *доверительным интервалом*, а величина $\delta = \pm z * \sigma_x$ называется *предельной ошибкой среднего* (или *предельной ошибкой выборки*). Таким образом, доверительный интервал фактически характеризует *точность оценки* измеряемой величины. Очевидно, что с увеличением значения *z* возрастает вероятность попадания среднего в доверительный интервал Δ, но при этом диапазон оценки становится неопределеннее и размытее, что уменьшает точность оценки[9]. Поэтому не следует стремиться задавать очень большое значение вероятности. Вполне достаточным является 90% или 95% уровень.

В случае, когда выборка состоит из менее 100 элементов или нет достаточных оснований считать выборочное распределение нормальным, для определения доверительного интервала рекомендуется использовать другое теоретическое распределение – *t*-распределение Стьюдента. Это распределение подробно будет рассмотрено во второй

[9] Чем менее определенным является прогноз, тем с большей вероятностью он осуществится.

части книги, а пока скажем, что в этом случае процедура определения доверительного интервала аналогична случаю больших выборок, но вместо значения z используется значение t-критерия Стьюдента. Это значение зависит от объема выборки и задаваемого уровня вероятности. Значение t-критерия можно определить из таблицы в приложении 1.2 или с помощью функции Excel СТЬЮДРАСПОБР(1 – P; n – 1).

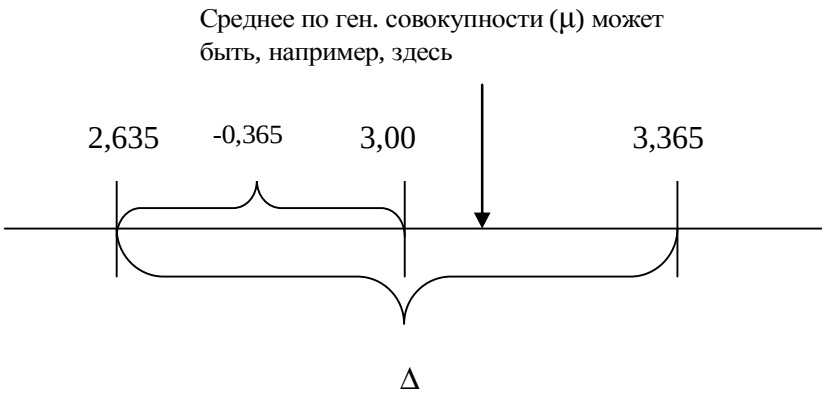
Для иллюстрации метода определения доверительного интервала рассмотрим следующий пример. Группа студентов университета разрабатывает проект по созданию университетского ланч-клуба (гибрида студенческой столовой и фешенебельного ресторана). Поскольку студенты помимо фонтана творческих идей и безудержного энтузиазма имеют еще и достаточно здравого смысла, подкрепленного знаниями основ математики, они решают провести мониторинг ряда параметров, на основании чего и принять решение об организации работы предполагаемого заведения. Пусть в качестве измеряемого параметра выступает «Предпочтительное время обеда». Проведено предварительное выборочное измерение (объем выборки – 10 элементов, выборки формировались в результате случайного повторного отбора). Респондентам предлагалось выбрать время обеда из 5 предложенных вариантов: 11:00, 11:30, 12:00, 12:30, 13:00. Отклики респондентов кодировались как 1, 2, 3, 4, 5, соответственно выбранному времени. Поскольку интервалы между позициями выбранной шкалы равны и соответствуют 30 мин, используемую шкалу можно рассматривать как интервальную. Результаты предварительного выборочного исследования представлены в таблице.

Респонденты	Отклики
1	3
2	2
3	3
4	2
5	5
6	1
7	4
8	3
9	4
10	3
$\bar{x}$ (среднее значение)	3,0
s (ст. отклонение)	1,155

Определяя стандартную ошибку среднего по характеристикам выборки, получим

$$s_{\bar{x}} = \frac{1{,}155}{\sqrt{10}} = 0{,}365.$$

Значение стандартного отклонения переводится во временной диапазон исходя из дизайна анкеты: интервал в 1 пункт соответствует 30 мин. Поэтому значение 0,5 соответствует 0,5 * 30 = 15 мин, а 0,365 соответствует 0,365 * 30 = 11 мин. Таким образом, по результатам выборочного исследования мы можем оценить интервал изменения среднего по генеральной совокупности: = $\pm$0,365 (12:00 $\pm$ 11 мин).



При этом вероятность попадания среднего по генеральной совокупности в этот интервал равна 68%, так как мы принимаем z = 1. Аналогично с вероятностью 95% по характеристикам первой выборки можно утверждать, что интервал изменения среднего по генеральной совокупности: Δ = 3,0 $\pm$ 0,365 * 1,96 = 3,0 $\pm$ 0,715 (12:00 $\pm$ 21,5 мин).

Однако, как отмечалось ранее, такой подход к определению доверительного интервала не совсем корректен, так как объем выборки мал (n < 100). Поэтому в данном случае необходимо использовать значение не z-критерия, а t-критерия. Для n = 10 и P = 0,95 (95%) t = 2,262 (см. приложение 1.2). Соответственно Δ = 3,0 $\pm$ 0,365 * 2,262 = 3,0 $\pm$ 0,826 (12:00 $\pm$ 25 мин).

3.4. Использование статистических пакетов для определения базовых характеристик выборки

Для большинства статистических методов и процедур, рассматриваемых в книге, показывается, как они реализованы в пакетах SPSS и STATISTICA. Параллельное рассмотрение возможностей применения обоих пакетов для решения исследовательских задач не связано с тем, что эти пакеты являются взаимоисключающими или, наоборот, взаимодополняющими. Область их функционального пересечения очень широка, и большинство процедур, реализованных в одном пакете, имеется и в другом. Однако каждый из этих пакетов имеет свои особенности, которые делают решение одних задач более

удобным с помощью одного пакета, других – с помощью другого. Например, выходные формы SPSS более удобны для анализа и интеграции в исследовательские отчеты, а пакет STATISTICA по некоторым видам задач предоставляет возможность сделать более развернутый анализ. В ряде случаев в этих пакетах используются различные варианты реализации некоторых статистических методов. Например, метод контрастов, реализованный в SPSS, использует статистику *t*-критерия Стьюдента, а аналогичный метод, реализованный в пакете STATISTICA, – статистику *F*-критерия Фишера.

При изложении материала, связанного с использованием статистических пакетов, несколько большее внимание будет все же уделено SPSS, поскольку, по мнению автора этой книги, этот пакет является более простым для освоения и – во многих случаях – более удобным для работы.

В данном параграфе мы опишем основные процедуры работы с этими пакетами для определения таких базовых (элементарных) статистик, как среднее значение, стандартное отклонение, доверительный интервал и др. Этот параграф является в определенной степени введением в работу со статистическими пакетами SPSS и STATISTICA, поэтому читатель, знакомый с основами работы с этими программами, может пропустить этот параграф без ущерба для понимания дальнейшего материала, изложенного в книге.

Описание процедуры расчета элементарных статистик мы предварим общими сведениями о структуре данных и переменных, с которыми работает SPSS. Мы делаем это для облегчения процесса освоения SPSS теми читателями, которые не имеют даже первоначального опыта работы с этим статистическим пакетом. В пакете STATISTICA правила и структура ввода и представления данных аналогичны, поэтому мы ограничимся рассмотрением вопросов о подготовке данных в среде SPSS.

Исходные данные для обработки могут быть введены в SPSS несколькими способами: путем набора данных в таблице SPSS, копированием данных из других приложений, например *Excel* (используются стандартные команды *Copy–Paste*), открытием файла с данными, конвертированием данных из файлов других форматов и др. После введения данных рекомендуется определить переменные в соответствии со структурой матрицы данных. Это можно осуществить, перейдя с вкладки *Data View* на вкладку *Variable View* (это также можно сделать, дважды щелкнув левой кнопкой мыши на заголовке выбранного столбца). Далее в соответствующих полях можно задать или изменить имя переменной, ее тип, шкалу измерения, формат представления, метку переменной.

По умолчанию каждая переменная имеет имя, состоящее из букв *var* и пяти цифр (00001, 00002 и т. д.). При изменении имени необходимо соблюдать следующие правила:

- имя должно начинаться с буквы;

- имя не может содержать больше восьми символов;
- в нерусифицированных версиях SPSS не рекомендуется использовать кириллицу (по крайней мере при определении переменных);
- допускается использование цифр и некоторых специальных символов (”_“, ”.“, ”$“ и ряда других);
- не допускается использование пробелов;
- в имени не различаются прописные и строчные буквы.

По умолчанию новая переменная является числовой. Тип и формат переменной можно изменить, щелкнув по правому краю поля *Type*. Здесь можно изменить количество цифр после десятичного знака, использовать в качестве десятичного знака точку или запятую, задать параметры переменной в денежном и других форматах и т.д.

Для улучшения идентификации переменных можно и полезно ввести метку (*label*) для используемых переменных. Метки применяются при формировании различных выходных форм и играют роль краткого содержательного описания. Например, в качестве метки можно использовать математическое или логическое выражение, которое применяется для вычисления переменной. При анализе данных метка оказывается очень полезной, так как она позволяет лучше ориентироваться в файле данных.

При сборе данных редко удается избежать отсутствия ряда показателей (измерений) по некоторым наблюдениям (объектам). При использовании многих статистических процедур, прежде всего связанных с анализом одной переменной, это не приводит к сколько-нибудь серьезным осложнениям: просто уменьшается число используемых наблюдений. Однако применение некоторых операций, например вычисление коэффициента корреляции, создает проблемы, так как отсутствие хотя бы одного значения в наблюдении исключает его из дальнейшего анализа. Для обозначения отсутствующих данных в SPSS используется специальный символ (« , »). Пользователь пакета может определить и другой код для пропущенных данных. Исследователь также может сам указать, что те или иные данные следует считать пропущенными. Например, все значения, лежащие в определенном интервале, указанном пользователем, будут считаться пропущенными (*Range of missing values*). Можно задать фиксированные значения переменной (до трех значений), которые будут считаться пропущенными (*Discrete missing values*). Также можно задать комбинацию интервала и фиксированного значения (*Range plus one discrete missing value*). Все эти варианты станут доступны в диалоговом окне *Missing Values*, если Вы щелкнете по правому краю поля *Missing*. По умолчанию установлена опция *No missing values*, т.е. все значения переменной считаются допустимыми для обработки.

Иногда исследователь предпочитает ограничить анализ имеющихся данных

определенным подмножеством, с которым и проводятся желаемые статистические процедуры. Для этого можно воспользоваться командой *Select Cases* из основного меню *Data*. В диалоговом окне *Select Cases* можно определить условия отбора наблюдений. Здесь можно задать несколько режимов отбора:

All cases – все элементы выборки используются для анализа (установлено по умолчанию);

Random sample of cases – из имеющихся данных случайным образом формируется подвыборка (задается определенный процент элементов выборки либо фиксированное число элементов);

Based on time or case range – подмножество данных определяется указанием диапазона изменений номеров наблюдений (элементов исходной выборки);

Use filter variable – подмножество данных формируется на основе задаваемой маски (фильтра). Маска может представлять собой отдельную переменную, включающую такое же количество элементов, что и исходный набор наблюдений. В подмножество данных отбираются те наблюдения, для которых элемент в фильтре имеет ненулевое значение.

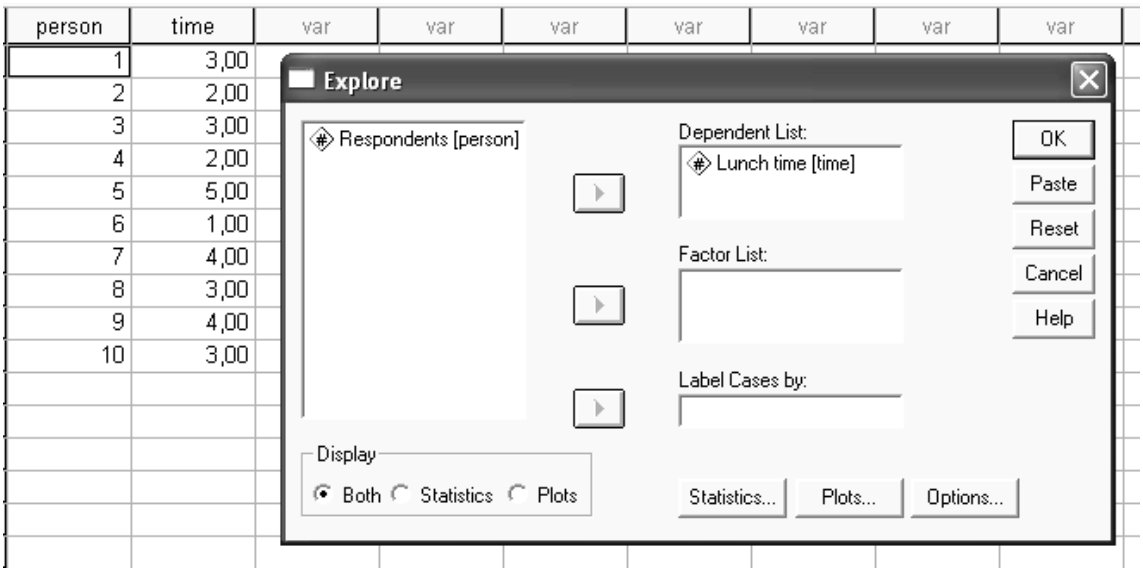
If condition is satisfied – отбираются элементы, удовлетворяющие некоторому условию, которое задается логическим выражением (с помощью элементов на панели «калькулятора» в диалоговом окне *Select Cases: If* и/или из списка функций *Functions* в этом же окне). Если условие истинно, то наблюдение выбирается, в противном случае – отбрасывается.

Предположим, что когда мы опрашиваем обучающихся о предпочитаемом времени обеда, мы отмечаем также их статус. Пусть это будет переменная status, измеряемая в номинативной шкале и имеющая три значения 1 – студент, 2 – магистрант, 3 – аспирант. Допустим, мы хотим ограничить анализ группой, содержащей только студентов и магистрантов. В этом случае мы можем задать в диалоговом окне *Select Cases: If* следующее логическое выражение: (status = 1) | (status = 2). Знак «|» означает логическое «или». Точно так же можно формулировать условия, используя две и более переменных. Например, мы отмечаем место проживания студентов (переменная living) как 1 – общежитие вблизи университета; 2 – в центре города; 3 – на окраине города; 4 – за городом. Допустим, нас по-прежнему интересуют студенты и магистранты, но только те, кто проживает в общежитии или в центре города. В этом случае мы можем задать следующее условие ((status = 1) | (status = 2)) & ((living = 1) | (living = 2)). Знак «&» означает логическое «и». Аналогичное условие можно для данного случая задать следующим образом: (~status = 3) & ((living = 1) | (living = 2)). Знак «~» означает логическое «не».

Элементы исходной выборки, не попавшие в подмножество, не используются в дальнейших вычислениях, но сохраняются в исходном файле данных. Необрабатываемые наблюдения легко отличить, так как, во-первых, соответствующий им номер строки

перечеркнут, а во-вторых, к массиву данных добавляется новая переменная-фильтр со значениями *Selected (1)* и *Not Selected (0)* соответственно для элементов, как вошедших, так и не вошедших в подмножество. Данная операция действует только в пределах текущего сеанса работы. При повторной загрузке файла вновь становится доступной вся исходная выборка. Переменная-фильтр при этом сохраняется, но она оказывается не активна (для ее активизации можно использовать режим *Use filter variable*. В случае необходимости можно удалить все элементы, не попавшие в анализируемую подгруппу. Для этого достаточно выделить в поле *Unselected Cases Are* режим *Deleted*. Однако это делать не рекомендуется, так как в процессе анализа приходится часто изменять или переконфигурировать исходный набор данных.

Теперь рассмотрим процедуру вычисления элементарных данных статистики в SPSS на ранее приведенном примере. Переменные обозначены как person (идентификатор респондента) и time (время обеда), переменным присвоены метки соответственно как Respondents и Lunch time. В основном меню *Analyze* выбирается опция *Descriptive Statistics*, затем *Explore*. В поле *Dependent List* переносится переменная *Lunch time*.



Далее в диалоговом окне *Statistics* (переход осуществляется нажатием на соответствующую кнопку в окне *Explore*) устанавливается уровень статистической значимости в процентах (по умолчанию 95%). В окне *Plots* устанавливаются режимы графического представления исходных данных (например полигон частот) и проверки распределения данных на нормальность. Для проверки нормальности в этом окне выбираем режим (устанавливаем флажок) *Normality plots with tests*.

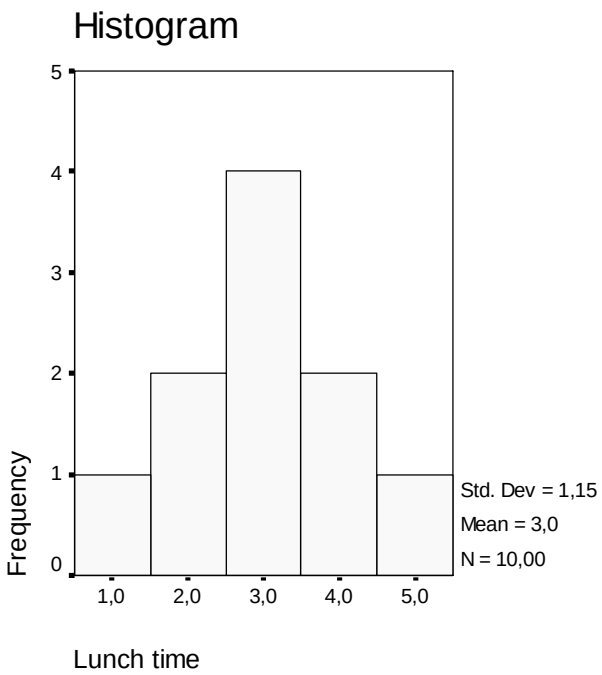
После нажатия на кнопку *Continue* в окне *Plots* и затем на кнопку *OK* в окне *Explore* формируется выходная форма, в которой представлены среднее значение по выборке (3,00),

стандартное отклонение (1,155), стандартная ошибка среднего (0,365), доверительный интервал (3,0 – 0,826 = 2,174; 3,0 + 0,826 = 3,826), минимальное (1,0) и максимальное (5,0) значения элементов выборки, их размах (4,0).

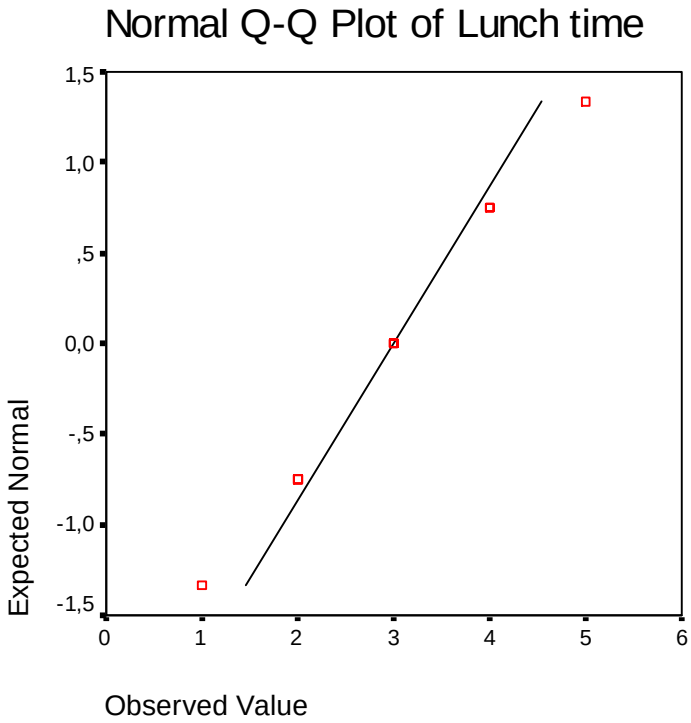
Descriptives

			Statistic	Std. Error
Lunch time	Mean		3,0000	,36515
	95% Confidence Interval for Mean	Lower Bound	2,1740	
		Upper Bound	3,8260	
	Median		3,0000	
	Variance		1,333	
	Std. Deviation		1,15470	
	Minimum		1,00	
	Maximum		5,00	
	Range		4,00	

На рисунке приведена гистограмма частот по рассматриваемой выборке (из отчета SPSS).



О нормальности распределения можно судить по графику *Normal Q-Q Plot*. В случае нормального распределения выборочных данных все точки должны лежать на одной прямой или рядом с ней. Чем больше отклонение точек от прямой линии, тем меньше выборочное распределение соответствует нормальному. В данном случае отклонение от нормального можно считать незначительным.

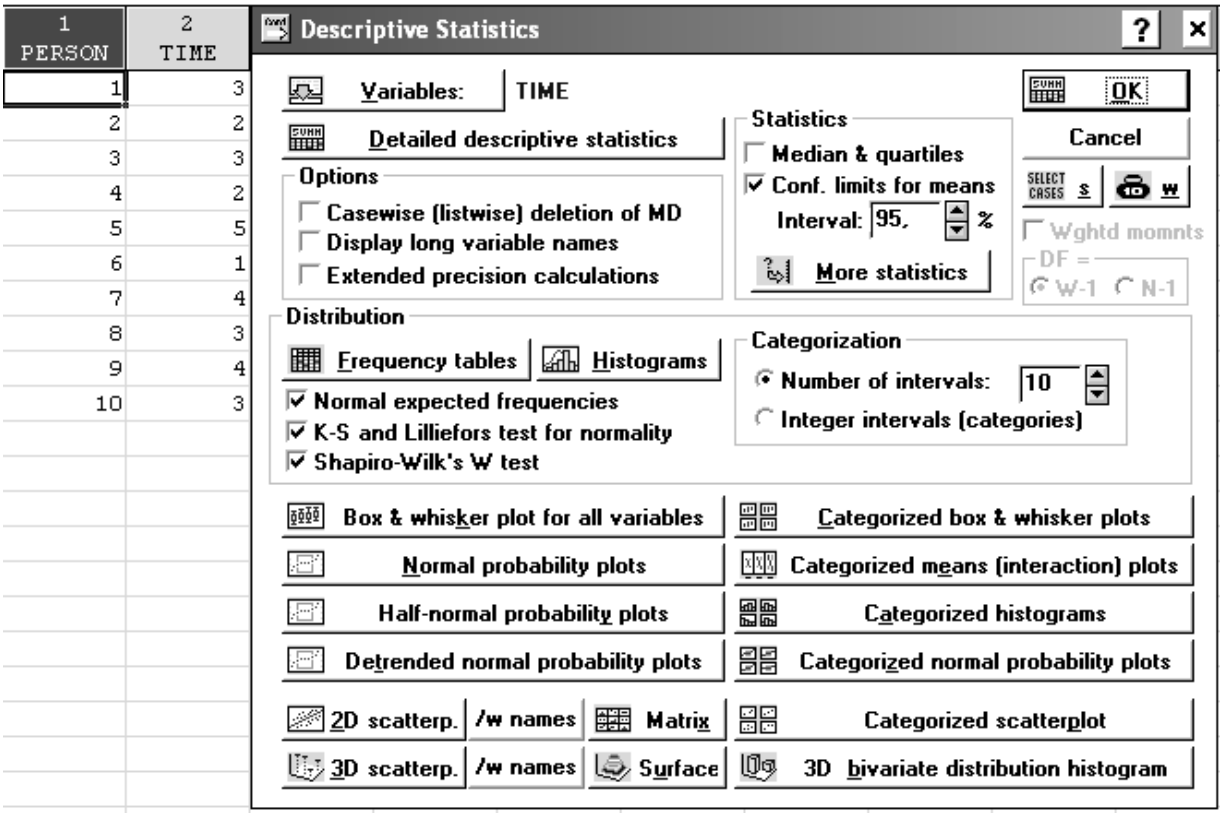


О нормальности распределения также можно судить с помощью специальных тестов. В выходной форме SPSS приводятся результаты проверки выборочного распределения на нормальность двумя методами: Колмогорова-Смирнова (Kolmogorov-Smirnov test) и Шапиро-Уилка (Shapiro-Wilk test). Последний тест считается более гибким и универсальным. В таблице приведена статистика по этим методам.

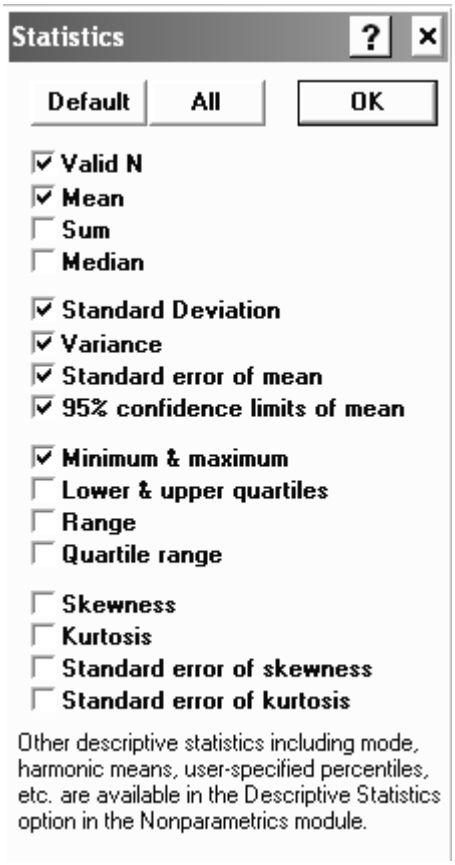
	Kolmogorov-Smirnov			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Lunch time	,200	10	,200	,953	10	,703

Если уровень значимости, отражаемый в столбце Sig. (Significance), больше 0,05, то можно считать, что вид анализируемого выборочного распределения существенно не отличается от нормального (подробно смысл этого показателя рассматривается во второй части книги).

Идентичные результаты могут быть получены с помощью пакета STATISTICA (хотя и в менее удобной форме представления). Для этого нужно включить модуль Basic Statistics (выбирается при запуске пакета), и в меню Analysis выбрать Descriptive Statistics. Для того, чтобы указать переменную для обработки, нужно нажать на кнопку *Variables* и в открывшемся окне выбрать соответствующую переменную (в данном случае – TIME).



После нажатия на кнопку *More Statistics* можно выбрать статистические показатели для расчета:



Valid N – число элементов выборки, пригодных для обработки;

Mean – среднее значение;

Sum – сумма значений переменной;

Median – медиана;

Standard Deviation – стандартное отклонение;

Variance – дисперсия;

Standard error of mean – стандартная ошибка среднего;

95% confidence limits of mean – 95% доверительный интервал (уровень статистической значимости устанавливается в поле *Interval*);

Minimum & maximum – минимальное и максимальное значения выборки;

Lover & upper quartiles – верхний и нижний квартили;

Range – размах;

Quartiles range – разность между верхним и нижним квартилями;

Skewness – коэффициент асимметрии;

Kurtosis – коэффициент эксцесса;

Standard error of skewness – стандартная ошибка коэффициента асимметрии;

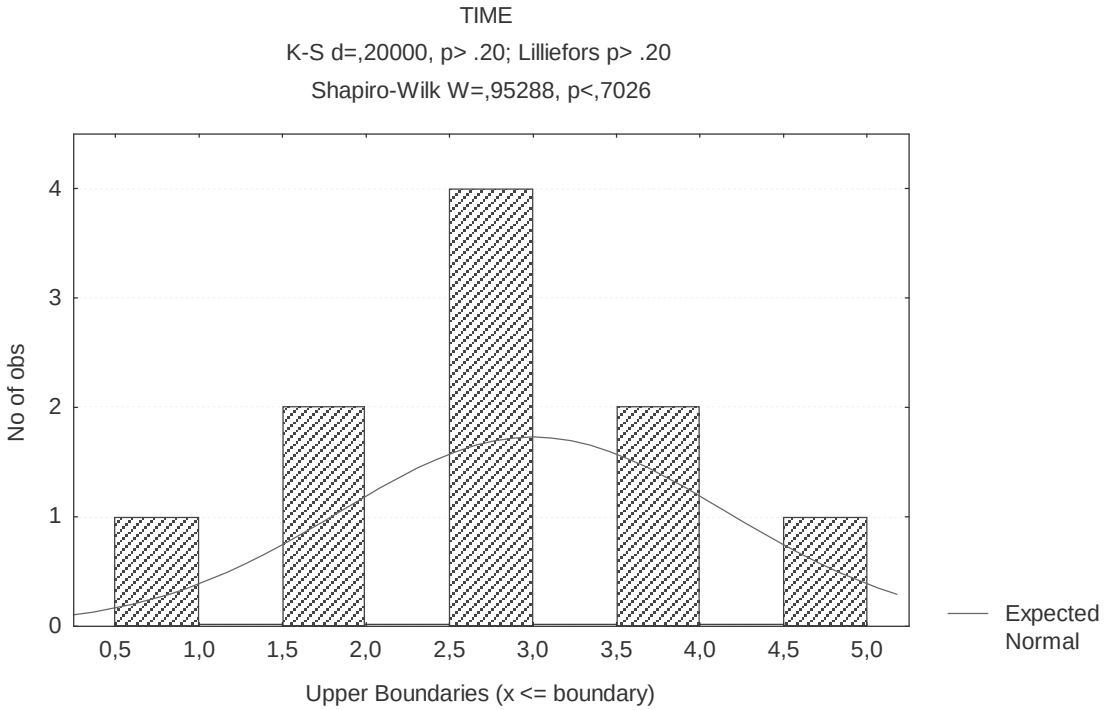
Standard error of kurtosis – стандартная ошибка коэффициента эксцесса.

Результаты вычислений представляются в выходной форме.

Descriptive Statistics (lunch_time.sta)

Continue...	Valid N	Mean	Confid. -95,000%	Confid. +95,000%	Minimum	Maximum	Variance	Std.Dev.	Standard Error
TIME	10	3,0000	2,1740	3,8260	1,0000	5,0000	1,3333	1,1547	,3651

Гистограмма частот может быть получена нажатием на кнопку *Histograms*. Кривая на гистограмме (*Expected Normal*) – график плотности нормального распределения со средним значением 3,0 и стандартным отклонением 1,15. На гистограмме также представлены статистики критериев Колмогорова-Смирнова и Шапиро-Уилка.



3.5. Определение необходимого объема выборки

Теперь мы можем сделать содержательный вывод о необходимом объеме выборки. Для определения необходимого количества элементов в выборке нужно задать желаемую точность, характеризуемую предельной ошибкой среднего (σ_x). Поскольку этот параметр

связан с объемом выборки, мы можем оценить необходимое количество элементов в выборке исходя из требуемой точности оценки исследуемого параметра генеральной совокупности.

Для простоты иллюстрации будем использовать *z*-критерий (ограничения на его использование были обозначены выше). Вернемся к примеру параграфа 3.3. Предположим, для нашего примера нам необходимо обеспечить точность оценки исследуемого нами параметра (предпочтительное время обеда) $\pm$15 мин с вероятностью 95%. Тогда мы задаем предельную ошибку среднего $\delta = \pm 0{,}5$ ($\pm$15 мин) и $z = 1{,}96$. Далее определяем минимально необходимый размер выборки.

$$\delta = \sigma_x * z\ ,\ \sigma_x = \delta / z, \qquad d / z = \frac{s}{\sqrt{n}}, \qquad n = (\frac{s}{d / z})^2.$$

По данным, полученным в результате выборочного исследования, находим $n = (1{,}15 / 0{,}255)^2 = 20{,}3$.

Таким образом, размер выборки должен быть не менее 21 для того, чтобы обеспечить точность оценки $\pm$15 мин с вероятностью 95%. Если необходимо увеличить точность, например до $\pm$10 мин с той же вероятностью, получим большее значение *n*. В качестве упражнения читателю предлагается определить *n* для обеспечения точности $\pm$10 мин с использованием *t*-критерия и сравнить результаты.

Данная формула для нахождения необходимой численности выборки и результаты по приведенному примеру получены при предположении, что выборка была сформирована при повторном случайном отборе. При формировании выборки методом случайного бесповторного отбора необходимый объем выборки определяется так:

$$n = \frac{s^2 N}{\left(d / z\right)^2 N + s^2}.$$

Читателю рекомендуется получить этот результат самостоятельно из формулы для определения предельной ошибки выборки при бесповторном случайном отборе, приведенной выше.

Рассмотрим еще один пример, вновь связанный с гипотетическим студенческим проектом «Университетский ланч-клуб». В данном случае в качестве измеряемого параметра будет выступать «частота посещений ланч-клуба»: «Как часто (сколько раз в месяц) Вы бы посещали ланч-клуб?»).

Допустим, по результатам предварительного тестирования ($n = 64$) получено: $\bar{x} = 10{,}0$, $s = 4{,}1$. Отсюда находим

$$s_x = \frac{s}{\sqrt{n}} = \frac{4{,}1}{\sqrt{64}} = 0{,}51.$$

Таким образом, с вероятностью 95% (z = 1,96) среднее по генеральной совокупности будет попадать в интервал 10 $\pm$ 1 визит в месяц (или от 9 до 11 визитов). С вероятностью 99,73% среднее по генеральной совокупности будет попадать в интервал 10 $\pm$ 3 * 0,51 визит в месяц (от 8,47 до 11,53 визитов).

Предположим, необходимо обеспечить точность оценки параметра $\pm$0,5 с вероятностью 95%. Тогда $n = (4{,}1 / 0{,}255)^2 = 258{,}5$, и необходимый объем выборки – 259 чел. Если исследователя устраивает меньшая точность, например $\pm$1 визит, тогда для выборочного исследования потребуется 65 чел.

Если необходимо определить требуемый объем выборки при одновременном измерении нескольких параметров (например, в опросный лист включены два пункта: «предпочтительное время обеда» и «предполагаемое количество посещений»), то определяется количество респондентов, требуемое для получения результатов заданной точности для каждого пункта. Максимальное из найденных значений и будет определять необходимую численность выборки. Допустим, для оценки «предпочтительного времени обеда» с точностью ±10 мин (вероятность – 95%) требуется опросить 42 респондента, а для оценки «предполагаемое количество посещений в месяц» с точностью ±1 посещение (вероятность – 95%) – 65 чел. Тогда необходимый объем выборки – 65 чел.

Средняя ошибка выборки при формировании стратифицированной (типической) выборки и бесповторном случайном отборе внутри типических групп рассчитывается по формуле

$$s_x = \frac{\bar{s}}{\sqrt{n}}\sqrt{1-\frac{n}{N}},$$

где $\bar{s}$ – среднее стандартное отклонение по типическим группам.

Аналогично, стандартная ошибка среднего при формировании стратифицированной (типической) выборки и повторном случайном отборе внутри типических групп рассчитывается по формуле

$$s_x = \frac{\bar{s}}{\sqrt{n}}.$$

Тогда необходимый объем стратифицированной выборки при повторном случайном отборе определяем как

$$n = \left(\frac{\bar{s}}{d / z}\right)^2.$$

Аналогично определяется необходимый объем стратифицированной выборки при бесповторном случайном отборе (читателю предлагается сделать это в качестве упражнения).

Определение численности каждой типической группы при условии пропорциональной численности единиц типических групп в стратифицированной выборке находим из условия

$$n_i = n\frac{N_i}{N},$$

где n_i – объем выборки из типической группы; n – общий объем выборки; N_i – объем типической группы; N – объем генеральной совокупности.

Достаточно часто выборочное исследование используется для определения параметров, представляющих собой *пропорции* от генеральной совокупности. Для иллюстрации вновь вернемся к примеру со студенческим проектом ланч-клуба. Предположим, инициаторы этого проекта задались вопросом: «Какой процент студентов университета будет посещать ланч-клуб?»

Стандартная ошибка определения пропорции при повторном отборе определяется как

$$s_p = \sqrt{\frac{pq}{n}},$$

где p – процент элементов генеральной совокупности, обладающих измеряемым атрибутом; q – процент элементов генеральной совокупности, не обладающих измеряемым атрибутом; $q = 1 - p$.

Сравнивая эту формулу с формулами для определения ошибки среднего, рассмотренными ранее, легко увидеть, что произведение pq имеет смысл дисперсии по выборке. Поэтому в литературе это произведение часто называется *групповой дисперсией*.

Отсюда получаем

$$n = \frac{pq}{s_p^2}.$$

Предположим, что предварительное тестирование выявило, что 30% студентов выказали желание посещать клуб. Установленная точность оценки: $\delta = \pm 0{,}1$ (т.е. $\Delta = \bar{p} \pm 10\%$), z = 1,96 (95%). Тогда $\sigma_p = 0{,}1 / 1{,}96 = 0{,}051$ и $n = 81$.

В случае, если предварительное тестирование не проводится, можно руководствоваться следующей логикой: произведение pq никогда не превысит 0,25, поэтому размер выборки можно оценить как

$$n = \frac{0{,}25}{s_p^2}.$$

При бесповторном отборе стандартная ошибка определения выборочной доли составит

$$s_p = \sqrt{\frac{pq}{n}\left(1 - \frac{n}{N}\right)}.$$

Отсюда легко может быть получено значение необходимого объема выборки для изучения выборочной доли (пропорции) в случае формирования выборки методом случайного бесповторного отбора:

$$n = \frac{pqN}{\left(d/z\right)^2 N + pq}.$$

Аналогично при стратифицированном способе отбора с пропорциональным отбором внутри типических групп средняя ошибка выборки для пропорции определяется по формуле (отбор внутри групп случайный повторный)

$$s_p = \sqrt{\frac{\overline{pq}}{n}}.$$

Здесь $\overline{pq}$ – среднее значение по группам (рассчитывается как среднее арифметическое взвешенное).

Отсюда легко получаем необходимую численность выборки для определения пропорции

$$n = \frac{\overline{pq}}{s_p^2} = \frac{\overline{pq}}{(d/z)^2}.$$

При случайном бесповторном отборе формула стандартной ошибки определения выборочной доли модифицируется:

$$s_p = \sqrt{\frac{\overline{pq}}{n}(1-\frac{n}{N})}.$$

Соответственно, значение необходимого объема стратифицированной выборки при определении пропорции в случае бесповторного отбора определяется как

$$n = \frac{\overline{pq}N}{\left(d/z\right)^2 N + \overline{pq}}.$$

Глава 4. Оценка качества измерений и измерительных инструментов в социально-экономических исследованиях

В предыдущей главе обсуждался вопрос проектирования и характеристик выборки для исследования, и мы уже отчасти рассматривали вопросы оценки качества измерений, а именно их точности. Теперь мы переходим к вопросу о проектировании и оценке характеристик *инструментов* измерения в социально-экономических сферах. Отметим сразу, что характеристики измерительных инструментов оцениваются через анализ и оценку самих измерений. Поэтому многие понятия, рассматриваемые в этой главе (например *надежность*), могут характеризовать как инструмент, так и измерения.

Как уже обсуждалось в первой главе, социально-экономические измерения часто являются косвенными. Непосредственно мы фиксируем некоторые признаки (время решения задачи, ответы на определенные вопросы, оценки респондентов и т.д.) и на этой основе делаем выводы о некотором скрытом (*латентном*) свойстве, недоступном прямому измерению (например, агрессивность, коммуникативность, креативность, благоприятные или неблагоприятные качества организационного климата для творчества и инноваций и т.п.). В частности, в социально-экономических эмпирических исследованиях широко используется измерение отношений респондентов, на основе чего и принимается или отвергается та или иная исследовательская гипотеза. Наиболее распространенными инструментами для таких измерений выступают разнообразные анкеты, опросные листы и т.п. Косвенный характер измерений осложняет оценку их качества.

В психологии, социологии и теории измерений разработано довольно много методов и подходов к оценке обоснованности, надежности и точности как результатов измерений, так и измерительных инструментов. Использование таких подходов для обоснования полученных результатов в зарубежной исследовательской практике считается обязательным. К сожалению, в нашей литературе этим методам и подходам уделяется явно недостаточное внимание. Это обстоятельство и явилось причиной того, что данный вопрос стал предметом рассмотрения в рамках отдельной главы. В основном эта глава подготовлена на основе англоязычной литературы и ресурсов.

Качество проектируемых и используемых инструментов исследования определяется их *обоснованностью* (*validity*)[10] и *надежностью* (*reliability*). Качество полученных результатов социально-экономических измерений зависит как от этих характеристик инструментов

[10] В русскоязычной литературе (преимущественно переводной), связанной с рассматриваемой темой, используется калька с англоязычного термина *«валидность»*. Поскольку этот термин представляется уже достаточно известным, закрепившимся и привычным, мы также будем использовать его в нашем пособии.

измерения, так и от *точности* (*accuracy*) проводимых измерений. Стандартная ошибка среднего по истинным значениям (*стандартная ошибка измерений*), которая характеризует точность измерений, может быть определена исходя из тех же принципов, которые использовались для определения доверительного интервала и рассматривались в главе 3. Для оценки точности измерений определяется среднее значение по оцениваемой группе показателей (пунктов анкеты) и стандартное отклонение средних по этой группе. Поскольку стандартное отклонение средних значительно меньше стандартного отклонения индивидуальных откликов, считается достаточным 68% доверительный интервал.

Кроме того, измерительный инструмент характеризуется *практичностью* (*practicality*), т.е. экономичностью применения (затратностью), удобством и простотой использования, а также возможностью наглядно интерпретировать полученные результаты. Практичность – это достаточно очевидная характеристика, поэтому в данном разделе мы остановимся на оценке таких основных характеристик измерительных инструментов, как обоснованность и надежность.

4.1. Обоснованность как характеристика измерительных инструментов

Обоснованность, или *валидность* (*validity*), – это эквивалентность измерений характеристикам измеряемого объекта. Другими словами, это мера соответствия оценок, получаемых в процессе измерения, представлениям о сущности свойств исследуемых объектов и их роли в исследуемых процессах. Оценивая обоснованность измерительного инструмента, мы отвечаем на вопрос: «Действительно ли мы измеряем то, что предполагаем измерять?» Хотя в общем случае обоснованность – это не формализуемая характеристика, и для ее оценки какие-либо специальные математические методы используются редко, эта характеристика представляется чрезвычайно важной при проектировании измерительного инструмента.

В литературе встречается достаточно большое количество (свыше 10) различных терминов, обозначающих типы валидности. При этом их четкая классификация отсутствует. Иногда одним и тем же термином в разных источниках обозначается разное содержание, а иногда, наоборот, различные термины наполняются одинаковым содержанием, характеризующим валидность. В данном параграфе выделяются и рассматриваются следующие типы валидности: внешняя, содержательная, эмпирическая и концептная (дискриминантная) валидность.

Внешняя валидность (*face validity*) характеризует восприятие вопросов анкеты непрофессионалами в той области, в которой планируется проводить исследования. В то же

время это люди – объекты той генеральной совокупности, которую предполагается исследовать. Это понятие характеризует, как пункты анкеты (вопросы, утверждения и т.д.) воспринимаются и понимаются респондентами.

Содержательная валидность (*content validity*) показывает, насколько пункты или вопросы анкеты, проектируемой для проведения измерения, соответствуют сути (содержанию) измеряемых характеристик. Это, как и внешняя валидность, не формализуемая характеристика, но в отличие от предыдущей, она оценивается экспертами, т.е. специалистами в той области, в которой проводится исследование. Если, например, мы разрабатываем инструмент для оценки инновационности нового продукта, выводимого на рынок, то для оценки содержательной валидности мы можем обратиться к специалистам в области творчества, инновационного менеджмента, маркетинга. Как правило, на самом первом этапе проектирования измерительного инструмента для обеспечения содержательной валидности проводится интервьюирование (структурированное или полуструктурированное) нескольких специалистов в области предполагаемого исследования для того, чтобы определить основные факторы и характеристики, которые являются наиболее важными с точки зрения предмета предполагаемого исследования.

Проверка внешней и содержательной валидности – это первые и обязательные элементы при разработке любого измерительного инструмента. Возможно, что обоснование содержательной валидности будет предшествовать проверке внешней валидности. Предположим, что отдел внеучебной работы намерен провести оценку того, насколько студенты довольны проживанием в общежитии университета. Для обеспечения содержательной валидности можно сформировать фокус-группу из нескольких (5–10) студентов, проживающих в общежитии, чтобы выяснить, какие факторы и аспекты наиболее важны для них. После этого может быть составлен черновой вариант анкеты и предложен этим же студентам для ознакомления с тем, чтобы они высказали свое мнение.

Эмпирическая валидность (*criterion-related validity*) рассматривает качество измерительного инструмента с позиций двух эмпирических критериев, а именно:

а) возможность предсказывать те или иные результаты (*predictive validity*) на основе измерений, полученных этим инструментом;

б) соответствие полученных результатов неким «золотым стандартам», т.е. уже проверенным и испытанным инструментам, используемым параллельно с разрабатываемым и тестируемым инструментом (*concurrent validity*).

Приведем пример из исследовательской практики автора книги. Последние 10–15 лет в практике управления персоналом достаточно широко используется такой инструмент оценки стилей решения проблем, как KAI (Kirton Adaption Innovation Inventory). Этот инструмент,

разработанный английским исследователем М. Кертоном в 1970-80-х гг., предварительно тестировался на нескольких тысячах человек для оценки его надежности и обоснованности. В частности, данные, полученные с помощью этого инструмента, статистически значимо показывают, что люди разных профессий имеют разные стили решения проблем. Например, бухгалтеры имеют более адаптивный стиль, а маркетологи – более инновативный (в терминах концепции М. Кертона). Для разработки русскоязычной версии KAI автором данного учебного пособия проводилось тестирование адаптируемой версии в различных профессиональных группах. Соответствие полученных результатов тем данным, которые были получены ранее М. Кертоном и его коллегами, рассматривалось как одно из свидетельств валидности адаптируемого инструмента.

Термином *концептная валидность* мы обозначили соответствие измерительного инструмента тому концепту (свойству), для измерения или оценки которого этот инструмент разрабатывался. Другими словами, этот инструмент характеризует соответствие инструмента его целям. В англоязычной специальной и учебной литературе подобный смысл передается термином *construct validity*. Нам представляется не совсем адекватной калька с этого термина («конструктная валидность»), которую используют некоторые авторы в русскоязычных изданиях, поэтому в этой книге используется термин «концептная валидность».

Эта характеристика измерительного инструмента связана с его способностью выделять и отделять различные концепты, поэтому для обозначения этого типа валидности также используется термин *дискриминантная валидность* (*discriminant validity*). На практике эта характеристика проверяется и оценивается по способности каждого пункта измерительного инструмента (вопроса или утверждения используемой анкеты или теста) отражать изменчивость измеряемых характеристик. Такая проверка проводится для выделения и исключения пунктов анкеты (например вопросов), не обеспечивающих достаточную степень «уверенного» разделения откликов. Например, если на некоторый вопрос все респонденты отвечают «Да», а на другой вопрос все отвечают «Нет», то такие ответы никакой информации фактически не несут, поэтому они не вносят никакой вклад в изучение того концепта, который интересует исследователя. Следовательно, такие вопросы не нужны в разрабатываемой анкете. Или другой пример. Предположим, мы проводим экзамен в студенческой группе и предлагаем такой тест, по которому все студенты выполняют все задания и получают «отлично», затем предлагаем другой тест, и в результате никто не выполняет ни одного задания, и все получают «неудовлетворительно». Способны ли такие тесты дать представление о знаниях студентов? Вряд ли. Поэтому такие тесты нельзя считать обоснованными (валидными) измерительными инструментами.

Для оценки концептной (дискриминантной) валидности могут использоваться более или менее сложные математические процедуры и индикаторы. Рассмотрим одну из таких процедур.

После тестирования инструмента (например проведения анкетирования) отклики всех респондентов по каждому из вопросов суммируются. Например, 1-й респондент по откликам на все пункты анкеты имеет 35 баллов, 2-й – 43, 3-й – 12 и т.д. Затем суммарные отклики ранжируются по величине. В итоге мы можем отобрать 20–25% респондентов с наименьшим суммарным откликом и столько же с наибольшим суммарным откликом. (В контексте измерения отношений (восприятий) первая группа соответствует респондентам с наименее благоприятным отношением к предмету исследования, вторая группа соответствует респондентам с наиболее благоприятным отношением к предмету исследования. В контексте тестирования первая группа соответствует респондентам с наихудшими результатами, вторая группа соответствует респондентам с наилучшими результатами.

Таким образом, мы формируем две группы по n человек: группу с низким суммарным откликом (группа L) и группу с высоким суммарным откликом (группа H). Оставшиеся респонденты (50%) со «средним» откликом не рассматриваются. Далее *для каждого* пункта анкеты определяются следующие величины:

f – число респондентов, выбравших соответствующую позицию шкалы (например, в 5-позиционной шкале Лайкерта это 1, 2, 3, 4 или 5, что соответствует различной степени согласия с предлагаемым утверждением);

$fX = f * X;$

$fX^2 = f * X * X;$

$$\overline{X} = \frac{\sum fX}{n},$$

где X – кодировка позиции шкалы (например 1, 2, 3, 4 или 5); $n = \Sigma f$ – число респондентов в группах L и H (в каждой группе это число должно быть одним и тем же).

Для каждого пункта анкеты определяется модифицированный t-критерий[11].

$$t = \frac{\overline{X}_H - \overline{X}_L}{\sqrt{\frac{(\sum fX_L^2 - \frac{(\sum fX_L)^2}{n}) + (\sum fX_H^2 - \frac{(\sum fX_H)^2}{n})}{n(n-1)}}}.$$

В этой формуле индексы L (*low*) и H (*high*) соответствуют первой и второй группам соответственно.

[11] Этот подход к оценке дискриминантной валидности связан с методами проверки статистических гипотез, которые рассматриваются во второй части книги. Там будет более подробно разбираться и t-критерий.

После определения *t*-критерия пункты ранжируются по величине *t*-критерия. Большее значение *t*-критерия соответствует лучшей дифференцирующей (разделяющей) способности пункта. В качестве критерия пригодности вопросов шкалы по степени различения принимается $t_{критическое}$ = 1,75 для n >= 25 (Cooper & Schindler, 1998). Пункты с t < 1,75 должны быть исключены из анкеты или переформулированы.

Использование этого подхода продемонстрируем на примере оценки дифференцирующей способности одного из пунктов анкеты: «Я считаю свою работу интересной». В анкете используется 5-позиционная шкала Лайкерта, отражающая степень согласия с предлагаемыми утверждениями. Пусть анкета распространяется среди 292 человек, входящих в группу, сформированную для предварительного тестирования разрабатываемой анкеты[12].

Пусть после распространения нашей анкеты и выделения в *L*- и *H*-группы по 25% респондентов (по 73 чел.) имеем следующие результаты:

	Группа с наименьшим суммарным откликом (L)				**Группа с наибольшим суммарным откликом (H)**			
Категории шкалы	X	f_L	fX_L	fX_L^2	X	f_H	fX_H	fX_H^2
Полностью согласен	5	3	15	75	5	22	110	550
Согласен	4	4	16	64	4	30	120	480
Нейтральное отношение	3	29	87	261	3	15	45	135
Не согласен	2	22	44	88	2	4	8	16
Полностью не согласен	1	15	15	15	1	2	2	2
		73	*177*	*503*		*73*	*285*	*1183*
		n	ΣfX_L	ΣfX_L^2		n	ΣfX_H	ΣfX_H^2

Для этих данных определяем *t*-критерий:

$$t = \frac{3{,}90 - 2{,}42}{\sqrt{\frac{(503 - \frac{177^2}{73}) + (1183 - \frac{285^2}{73})}{73(73-1)}}} = \frac{3{,}90 - 2{,}42}{\sqrt{\frac{73{,}84 + 70{,}33}{73(73-1)}}} = 8{,}92.$$

На основании полученного результата можно сделать вывод о том, что анализируемый пункт шкалы обладает достаточно высокой разделительной способностью. Аналогично можно провести анализ всех пунктов анкеты и на этой основе сделать вывод о дискриминантной валидности всего инструмента.

[12] Перед распространением анкеты в группе необходимо обеспечить по крайней мере ее внешнюю и содержательную валидность.

Если выборка состоит из нечетного числа респондентов (например 291), то при формировании *L*- и *H*-групп не обязательно добавлять или удалять респондентов, чтобы получить четное число, также не нужно включать в группы одних и тех же респондентов. Соотношение 25–25–50% – условное и может варьироваться. После ранжирования суммарных откликов всех респондентов отбирается равное количество респондентов «сверху» и «снизу» (приблизительно по 25% от численности всей выборки), какое именно количество остается в средней группе (четное или нечетное) – не принципиально. Отбираются *относительно* «высокие» и «низкие» суммарные отклики, т.е. определенное количество откликов из нижней и верхней частей списка, безотносительно к их абсолютным значениям.

Оценка и аргументация валидности измерительных инструментов носит преимущественно описательный характер, и в научных публикациях (к сожалению, в основном зарубежных), в которых оценивается валидность используемых инструментов, соответствующий раздел может занимать несколько страниц.

4.2. Подходы к оценке надежности измерений и измерительных инструментов

Надежность (*reliability*) – это характеристика, отражающая устойчивость и согласованность получаемых результатов измерения. В повседневном общении мы очень часто используем слово «надежность» (надежный человек, надежный компьютер, надежный автомобиль и т.д.). Основные смыслы, которые при этом вкладываются в эту характеристику, – это стабильность, безотказность, повторяемость, предсказуемость, регулярность. Примерно такие же смыслы вкладываются и в понятие «надежность» как характеристики инструмента измерения. Это, например, возможность использования того же инструмента для одной и той же выборки и получения при этом аналогичных результатов при условии, что измеряемая характеристика не изменилась.

Надежность является составляющей обоснованности (валидности) инструмента, это необходимый, но недостаточный элемент валидности. Связь этих двух характеристик может иллюстрировать простой пример с измерением веса. Если мы правильно измеряем наш вес (и проверяем это измерение разными весами), это измерение обоснованно и надежно. Если наши весы всегда показывают на 10 кг больше (что опять же проверяется с помощью других весов), то эти измерения могут рассматриваться как надежные, но не обоснованные (не валидные). Если наши весы каждый день показывают заметно разные значения, то эти измерения не являются ни надежными, ни обоснованными.

Надежность инструмента характеризует, насколько наши измерения свободны от случайных ошибок. В отличие от оценки обоснованности, оценка надежности измерительного инструмента всегда осуществляется с помощью математических операций. Общий подход к оценке надежности заключается в оценке степени связанности результатов измерения с помощью либо *параллельных испытаний*, либо разнесения измерений во времени, либо соотнесения данных из разных фрагментов одного инструмента.

Для определения надежности используются три основных подхода, основанных на трех разных вариантах понимания надежности:

1. *Надежность-устойчивость* (*stability*) характеризует стабильность результатов во времени.

2. *Надежность-эквивалентность* (*equivalence*) характеризует идентичность результатов, полученных несколькими аналогичными инструментами.

3. *Надежность-согласованность* (*internal consistency*) характеризует согласованность (гомогенность) пунктов инструмента (например анкеты).

Рассмотрим эти подходы подробнее. Для оценки надежности в смысле устойчивости результатов во времени один и тот же инструмент повторно тестируется по одной и той же выборке через определенный промежуток времени (*метод «тест-ретест»*). Результаты двух тестов, как правило, сравниваются путем определения коэффициента корреляции или другой меры связи (см. главу 9), а также средних значений по двум испытаниям. В случае, если получается высокий коэффициент корреляции (близкий к единице) и средние значения по первому и второму тестированию близки, то это свидетельствует о надежности измерений в смысле их воспроизводимости и стабильности. Если по результатам первого и второго испытаний средние значения различаются достаточно сильно, но в целом те респонденты, которые имели высокие баллы при первом тестировании, получили примерно также высокие баллы во втором, то в этом случае коэффициент корреляции принимает достаточно высокие значения, что указывает на определенную надежность инструмента. На практике статистически значимый коэффициент корреляции выше 0,7 считается свидетельством достаточной надежности тестируемого инструмента. При этом также следует указывать уровень статистической значимости полученного результата (методы и критерии проверки статистической значимости мер связи рассматриваются в главе 9 учебного пособия). Плохая воспроизводимость результатов предыдущего тестирования приводит к низкой корреляции результатов, что свидетельствует о низкой надежности.

Достоинство этого метода заключается в сравнительной простоте его использования, ясности основных посылок, лежащих в определении надежности, и простоте расчетов. Сложности возникают при определении временного интервала между двумя испытаниями.

Если ретест проводится слишком рано, респонденты могут запомнить ответы, которые они давали при первом испытании. При слишком позднем проведении повторного испытания измеряемые характеристики могут измениться (например знания, способности, опыт, отношения респондентов и т.д.). Приемлемым считается интервал между тестированиями от 2-х недель до 2-х месяцев. Кроме того, сама по себе высокая корреляция не может однозначно свидетельствовать о воспроизводимости результатов, поэтому результаты повторного тестирования рекомендуется контролировать другими методами. Например, можно сравнивать ранги респондентов, и если они в основном не изменились, то появляются дополнительные основания в пользу надежности измерений, но только в смысле их стабильности, так как возможен тренд, т.е. систематическое увеличение или уменьшение результатов от одного тестирования к другому. Возможно использование процедур проверки статистической гипотезы о равенстве средних значений и достоверности различий дисперсий по первому и повторному тестированиям (см. вторую часть книги).

Для оценки надежности-эквивалентности используется метод параллельного тестирования, или альтернативных тестов (*parallel forms*), проводимых либо одновременно, либо с небольшим интервалом. Данный метод оценки надежности применим только тогда, когда имеются параллельные (сходные, но не одинаковые) формы одного инструмента. Одной и той же группе респондентов предлагается вначале одна форма, затем после некоторого перерыва (до одной-двух недель) – другая. Коэффициент корреляции, полученный по результатам двух тестов, называется *коэффициентом эквивалентности измерения*. Если между предъявлением обеих форм имеется значительный временной интервал (свыше двух недель), то полученный коэффициент корреляции называется *коэффициентом эквивалентности и стабильности*.

Статистически значимый коэффициент корреляции > 0,7 считается свидетельством достаточной надежности тестируемого инструмента. Однако, как уже обсуждалось и дополнительно будет показано в следующем параграфе, вычисление коэффициента корреляции может оказаться недостаточным в случае больших различий в средних значениях и дисперсиях по параллельным тестам. Еще одна сложность применения данного метода заключается в невозможности обеспечить *полную* эквивалентность двух разных тестов. Рекомендуется в качестве альтернативного теста использовать тот же самый инструмент с переформулированными пунктами (вопросами или утверждениями) и, соответственно, обратной кодировкой откликов (см. главу 2). Одним из вариантов такого подхода является повторное тестирование тем же самым тестом, но администрируемым другим человеком.

Наиболее часто надежность измерительных инструментов оценивается по *согласованности* составляющих элементов инструмента. Анкеты и опросные листы как основные измерительные инструменты эмпирических социологических и экономических исследований состоят из большого числа отдельных составляющих – пунктов, вопросов, утверждений и т.п. Каждый из пунктов направлен на косвенное выяснение какой-то одной стороны, отдельного фрагмента общего целого, вследствие чего он является частичным индикатором измеряемого шкалой фактора (свойства). Предполагается, что когда мы принимаем во внимание всю совокупность индикаторов и определенным образом интегрируем косвенную информацию, которую несет каждый из индикаторов, наши выводы становятся более надежными и обоснованными. Считается, что согласованный инструмент должен состоять из внутренне непротиворечивых пунктов.

Как было показано в материалах двух первых глав, существуют и используются различные способы интегрирования информации из частных индикаторов (суммирование баллов, полученных по каждому пункту шкалы; использование модели Раша и др.). Однако прежде чем интегрировать данные по индикаторам, необходимо соблюдение условия, что все пункты шкалы измеряют одно и то же. Если это не так, т.е. у пунктов нет ничего общего, тогда операция получения комплексной оценки просто не имеет смысла. Надежность-согласованность как раз и показывает, в какой степени различные пункты могут определять нечто единое.

Для оценки надежности-согласованности пунктов инструмента разработано несколько методов. Рассмотрим базовые предпосылки, лежащие в основе этих методов. Убедиться в том, что пункты шкалы измеряют нечто общее, можно путем определения коэффициента корреляции между откликами на пункты инструмента. Высокое значение коэффициента корреляции (r) между двумя переменными может свидетельствовать о том, что имеется какой-то скрытый (латентный) фактор, общая причина, которая стоит за каждой из них. Именно на этом соображении может строиться проверка такого качества шкалы, как согласованность: *если у пунктов есть нечто общее, они должны коррелировать друг с другом*. Но такой подход позволяет сравнивать пункты попарно. Как можно на этой основе получить универсальный показатель для всего инструмента?

Одним из первых методов, разработанных для решения этой задачи, был *метод* раздельного коррелирования (*split-half*). Он заключается в разбиении всех пунктов инструмента на две половины и расчете коэффициента корреляции по соответствующим двум наборам данных – суммарным результатам по каждому пункту. Суммирование баллов в двух сформированных группах дает два набора данных, корреляция между которыми и характеризует надежность-согласованность инструмента. Если пункты инструмента

совершенно надежны, то следует ожидать, что обе части абсолютно скоррелированны (т.е. $r = 1.0$). Впрочем, такая «абсолютная надежность» является гипотетической и на практике встречается исключительно редко. Если пункты инструмента не являются абсолютно согласованными, то коэффициент корреляции будет меньше 1.

Преимущество метода раздельного коррелирования перед методом параллельного тестирования заключается в том, что он позволяет оценить надежность при однократном тестировании. Однако использование этого метода предполагает допущение об эквивалентности не только отдельных форм, но и пунктов теста. Еще одна сложность использования метода раздельного коррелирования заключается в том, что два набора пунктов анкеты можно получить разными способами. Если у нас, например, 20 пунктов в анкете, можно первые 10 включить в одну группу, а остальные 10 – в другую, а также можно в первую группу включать пункты с нечетными номерами, а во вторую – с четными (это наиболее распространенная на практике процедура) и т.д. Тест, состоящий, например, из 20 пунктов, может быть поделен на две половины 92378 разными способами. Понятно, что коэффициент корреляции будет зависеть от способа разбиения. Наиболее правильным считается разбиение, производимое случайным образом, что позволяет избежать искусственных эффектов. Тем не менее показатель надежности-согласованности, полученный таким методом, будет варьироваться всякий раз при формировании групп. Проблемами этого подхода является также неэквивалентность пунктов (например, одни вопросы могут быть сложнее, чем другие, и наоборот), а также то обстоятельство, что респондент может вообще не ответить на какие-то пункты или, дойдя до половины теста, ответить небрежно на оставшиеся вопросы.

Еще одной проблемой применения этого метода является то, что коэффициент корреляции рассчитывается не по всем пунктам анкеты, а по половине. Для корректировки полученного значения используется *формула Спирмена-Брауна*, предложенная независимо друг от друга К. Спирменом (C. Spearmen) и У. Брауном (W. Brown) в 1910 г.:

$$r_{SB_2} = \frac{2r}{1+r},$$

где r_{SB2} – скорректированный показатель надежности-согласованности по методу ***split-half***;
r – коэффициент корреляции между двумя наборами пунктов инструмента.

Более полезной для оценки согласованности измерений является *обобщенная формула Спирмена-Брауна*, в которой используется усредненный коэффициент корреляции:

$$r_{SB} = \frac{kr}{1+(k-1)r},$$

где k – количество пунктов в шкале (вопросов, утверждений и т.п.); r – усредненный коэффициент корреляции между всеми пунктами инструмента.

Проиллюстрируем использование этой формулы на простом примере. Пусть мы имеем анкету из 3 пунктов, на которые получены отклики 4 респондентов.

Респонденты	**Вопросы**		
	1	**2**	**3**
1	0	1	1
2	1	2	2
3	2	1	2
4	0	1	1
	0	1	1

	1	*2*	*3*
Корреляционная матрица	1		
	0,25	1	
	0,918559	0,612372	1

Усредн. коэф. корреляции	0,593644

Вычисляется корреляционная матрица (матрица коэффициентов корреляции). Это можно сделать, например, с помощью функции Excel для вычисления коэффициента парной корреляции КОРРЕЛ(), а еще проще – с помощью специальной опции Excel «Анализ данных» («Сервис», «Анализ данных», «Корреляция»)[13]. Входным интервалом при использовании этой опции будет выступать матрица откликов респондентов (способ группировки – по столбцам). В результате будет получена корреляционная матрица, как представлено в таблице. Для определения среднего коэффициента корреляции необходимо посчитать среднее значение коэффициентов корреляции в одной половине матрицы (кроме диагональных, поскольку это корреляционные коэффициенты пунктов с собой).

Для нашего случая среднее арифметическое трех коэффициентов корреляции можно определить, например, с помощью функции СРЗНАЧ(). (Для простоты использования этой функции можно предварительно удалить единицы, расположенные по диагонали корреляционной матрицы, и в качестве параметра функции указать массив коэффициентов 2x2 с одним пустым элементом). Для нашего примера усредненный коэффициент корреляции равен 0,594. Далее по обобщенной формуле Спирмена-Брауна определяется индекс согласованности пунктов анкеты. Для нашего случая он составит 0,814 (читателям рекомендуется проверить это самостоятельно).

[13] Если в Excel, установленном на вашем компьютере, такая опция отсутствует, необходимо зайти в меню Excel «Сервис», «Надстройки» и отметить «Пакет анализа».

Такой подход предполагает равенство дисперсий в двух коррелируемых группах. Известен аналог коэффициента раздельного коррелирования Спирмена-Брауна, который не предполагает такого равенства дисперсий. Это *коэффициент раздельного коррелирования Гутмана (Guttman split-half coefficient)*, который вычисляется по формуле

$$r_G = 2\frac{s_t^2 - s_1^2 - s_2^2}{s_t^2},$$

где s_t^2 – общая дисперсия по всем данным (первой и второй группам); s_1^2 – дисперсия по первой группе; s_2^2 – дисперсия по второй группе.

Другие подходы к определению внутренней согласованности основаны на вычислении коэффициентов ***KR*$_{20}$** *Кадера-Ричардсон*, *альфа Кронбаха* и *лямбда Гутмана*. Рассмотрим базовые аксиомы, на основе которых разработаны формулы для вычисления этих коэффициентов.

Каждое измерение (ответ на вопрос, отклик на утверждение и т.п.) включает в себя как истинное значение, так и частично неконтролируемую, случайную погрешность

$O = T + E$,

где O – наблюдаемое (измеряемое) значение (*observed score*); T – истинное значение (*true score*); E – случайная ошибка (*random error*).

Более полно, с учетом систематической ошибки (B), имеем: $O = T + E + B$.

Например, мы измеряем рост 100 случайно выбранных человек. Вероятно, измеренные значения будут соответствовать нормальному распределению. А как будут связаны изменения (вариации) этих слагаемых?

Изменчивость измеряемого признака может быть связана с «естественной» изменчивостью самого признака (различие в росте), но определенный вклад может внести то, как мы измеряем, т.е. изменчивость ошибки измерения. Запишем это как

$$s_O^2 = s_T^2 + s_E^2$$

(систематическая ошибка не учитывается, так как считается, что ее изменчивость равна 0).

Тогда надежность измерения (ρ) может характеризоваться отношением изменчивости истинных значений к изменчивости наблюдаемых значений, т.е. $r = \frac{s_T^2}{s_O^2}$.

Разумеется, мы не можем знать истинные значения и их изменчивость (иначе нам бы не пришлось проводить никакие измерения), но мы можем исключить их из рассмотрения, представив как $s_T^2 = s_O^2 - s_E^2$. Тогда получаем

$$r = \frac{s_O^2 - s_E^2}{s_O^2} = 1 - \frac{s_E^2}{s_O^2}.$$

В знаменателе мы имеем не что иное, как дисперсию измеряемых значений. Определить значение в числителе сложнее, но понятно, что оно должно иметь смысл дисперсии ошибок наших измерений. Однако заметим сразу, что использование подобного подхода не предполагает разбиения пунктов инструмента на группы, поэтому снимается проблема зависимости результата от способа разбиения. Это весьма абстрактная идея о надежности измерений воплотилась в нескольких конкретных вариантах расчетных моделей.

Впервые конкретная реализация подобных рассуждений была предложена Кадером и Ричардсон (1937 г.) и получила название *формулы Кадера-Ричардсон-20* (*Kuder-Richardson 20*), или KR-20. Несколько необычное название формулы связано с тем, что авторы предложили несколько различных формул, обозначаемых разными индексами; двадцатая оказалась наиболее удачной. Эта формула была предложена для вычисления индекса согласованности для дихотомической шкалы (т.е. для переменных, принимающих только два значения, например для ответов истинно/ложно).

$$KR_{20} = \frac{k}{k-1}\left(1 - \frac{\sum_{i=1}^{k} p_i q_i}{s_t^2}\right),$$

где p_i – доля первого варианта ответа на i-й вопрос; $q_i = (1 - p_i)$ – доля второго варианта ответа на i-й вопрос; s_t^2 – дисперсия сумм измеряемых значений (суммирование осуществляется по всем вопросам анкеты для каждого респондента); k – количество вопросов.

Например, получены ответы 5 испытуемых по трем заданиям теста с дихотомической шкалой. В таблице приведены ответы, а также все промежуточные результаты для вычисления коэффициента KR_{20} . Полученное значение (0,86) свидетельствует о хорошей согласованности заданий данного теста.

Респонденты	Вопросы			Сумма	
	1	2	3		
1	0	1	1	2	
2	1	1	1	3	
3	0	0	0	0	
4	1	1	1	3	
5	1	1	0	2	
s_t (ст. отклонение суммарных откликов)					1,22
s_t^2 (дисперсия суммарных откликов)					1,5
p_i	0,6	0,8	0,6		
q_i	0,4	0,2	0,4		

$p_i q_i$	0,24	0,16	0,24	0,64	
KR_{20}					0,86

Для порядковых шкал с большим количеством позиций (например шкалы Лайкерта), а также для более мощных шкал (например интервальных) Л. Кронбах предложил другую формулу для определения согласованности измерений (1951 г.). Показатель согласованности, рассчитанный по этой формуле, получил название *коэффициент альфа Кронбаха* (*Cronbach's Coefficient Alpha*). Большинство современных статистических пакетов (SPSS, SAS, STATISTICA и др.) включают процедуры вычисления коэффициента альфа Кронбаха. Несложно посчитать этот коэффициент и с помощью стандартных функций Excel. Формула выглядит следующим образом:

$$a = \frac{k}{k-1}\left(1 - \frac{\sum_{i=1}^{k} s_i^2}{s_t^2}\right).$$

где σ_i^2 – дисперсия откликов по каждому пункту анкеты; σ_t^2 – дисперсия суммарной шкалы (дисперсия суммы откликов каждого респондента на пункты анкеты); k – количество пунктов.

Эта формула отражает следующую идею. Если есть несколько субъектов, отвечающих на вопросы анкеты, то можно вычислить дисперсию для каждого вопроса и суммарной шкалы. Дисперсия для суммарной шкалы будет меньше, чем сумма дисперсий каждого отдельного вопроса в том случае, когда вопросы измеряют (оценивают) *одну и ту же* изменчивость между субъектами, т.е. если они измеряют некоторое истинное значение. Если не измеряется реальное значение, а только случайная погрешность в ответах на вопросы (следовательно, ответы полностью некоррелированны между субъектами), то дисперсия суммы будет такой же, как сумма дисперсий отдельных пунктов. Поэтому коэффициент альфа будет равен нулю. Если все вопросы измеряют один и тот же объект (истинную метку), то коэффициент альфа равен 1. Формула Кронбаха является расширенной аналогией формулы Кадера-Ричардсон, но не прямым ее обобщением, поэтому для дихотомических шкал следует использовать коэффициент KR_{20}, а не коэффициент альфа Кронбаха.

Рассмотрим использование формулы Кронбаха на примере. Возьмем те же данные, которые мы использовали для иллюстрации применения обобщенной формулы Спирмена-Брауна. Определим дисперсии по откликам на каждый вопрос и их сумму. Получим значение 1,3. Дисперсия агрегированных оценок по каждому вопросу составит 2,7. Отсюда значение коэффициента альфа Кронбаха 0,778.

Если мы сравним полученное значение с индексом согласованности, определенным по обобщенной формуле Спирмена-Брауна, то увидим, что коэффициент альфа меньше: 0,778 <

0,814. Это связано с тем, что обобщенная формула Спирмена-Брауна вычисляет коэффициент согласованности как если бы измерения были стандартизованы, т.е. приведены к одной шкале с нулевым средним значением и единичной дисперсией (см. главу 2). Часто (но не всегда) стандартизация исходных данных приводит к возрастанию надежности измерений.

Респонденты	**Вопросы**			**Сумма дисперсий**	**Сумма**	**Альфа**
	1	**2**	**3**			
1	0	1	1		2	
2	1	2	2		5	
3	2	1	2		5	
4	0	1	1		2	
5	0	1	1		2	
Среднее значение					3,2	
Ст. отклонения	0,89	0,45	0,55		1,643168	
Дисперсии	0,8	0,2	0,3	1,3	2,7	**0,778**

Дисперсия сумм

Коэффициент альфа Кронбаха принимает значения в диапазоне от 0 до 1. Приемлемым считается значение > 0,7. (Некоторые специалисты считают приемлемыми значения альфа Кронбаха > 0,8.) Высокое значение коэффициента указывает на наличие общего основания у набора вопросов, но не говорит о том, что за ними стоит один единственный фактор, поэтому одномерность шкалы следует подтверждать дополнительными методами, т.е. предварительно проверять валидность инструмента. Понятно, что надежность-согласованность всего инструмента, определяемая по формуле Кронбаха, будет зависеть от количества и качества пунктов, входящих в этот инструмент. При исключении любого пункта из анкеты коэффициент альфа будет изменяться (уменьшаться или увеличиваться). При исключении пунктов, которые не противоречат другим пунктам анкеты, коэффициент альфа Кронбаха уменьшается. И напротив, при исключении пунктов, которые не согласуются с другими пунктами анкеты, значение коэффициента альфа будет увеличиваться. Теоретически при оценке надежности используемого (тестируемого) инструмента (и каждого пункта, входящего в этот инструмент) мы должны определить коэффициент альфа Кронбаха при условии, что один из

пунктов исключен (и так для всех пунктов). Это весьма трудоемкая задача, особенно если в измерительном инструменте много пунктов. Поэтому для решения этой задачи используют специальные статистические пакеты. Подробно технологию их применения мы рассмотрим в следующем параграфе, а пока лишь приведем пример, иллюстрирующий дальнейшее изложение последовательности улучшения надежности инструмента.

Пусть мы получили результаты анализа надежности-согласованности анкеты, состоящей из 10 пунктов[14]. Наибольший интерес представляют три крайних правых столбца таблицы. Они показывают корреляцию между откликами по соответствующему пункту и суммарным показателем по всем пунктам, квадрат множественного коэффициента корреляции между соответствующим пунктом и другими пунктами, внутреннюю согласованность пунктов (коэффициент альфа Кронбаха), если соответствующий пункт будет удален.

STATISTICA АНАЛИЗ НАДЕЖНОСТИ	**Итоги для шкалы: Среднее=46.1100 Ст.откл.=8.26444 N набл:100 Альфа Кронбаха: .794313 Стандартизованная альфа: .800491 Средняя межпозиционная корреляция: .297818**					
Переменная	**Среднее при удал.**	**Дисперсия при удал.**	**Ст.откл. при удал.**	**Общ-поз. коррел.**	**Квадрат мн. регр.**	**Альфа при удал.**
ITEM1	41.61000	51.93790	7.206795	.656298	.507160	.752243
ITEM2	41.37000	53.79310	7.334378	.666111	.533015	.754692
ITEM3	41.41000	54.86190	7.406882	.549226	.363895	.766778
ITEM4	41.63000	56.57310	7.521509	.470852	.305573	.776015
ITEM5	41.52000	64.16961	8.010593	.054609	.057399	.824907
ITEM6	41.56000	62.68640	7.917474	.118561	.045653	.817907
ITEM7	41.46000	54.02840	7.350401	.587637	.443563	.762033
ITEM8	41.33000	53.32110	7.302130	.609204	.446298	.758992
ITEM9	41.44000	55.06640	7.420674	.502529	.328149	.772013
ITEM10	41.66000	53.78440	7.333785	.572875	.410561	.763314

По этим показателям пункты *5* и *6* заметно выделяются: они не согласуются с остальными пунктами, поэтому эти два пункта можно исключить. Но после удаления этих пунктов количество оставшихся пунктов может быть недостаточным для того, чтобы создать надежный инструмент (чем меньше пунктов, тем менее надежен измерительный инструмент). Поэтому возможно, что эти пункты потребуется переформулировать или заменить другими.

Пункты, которые плохо согласуются с остальными пунктами инструмента (т.е. увеличение коэффициента альфа при их удалении достаточно сильное), следует убрать из анкеты. К сожалению, однозначных критериев того, что значит «достаточно сильное увеличение», не существует. Однозначно нельзя сказать, что если при удалении пункта

[14] Использован пример из мультимедийного учебника по пакету STATISTICA, доступного на сервере StatSoft Russia (www.statsoft.ru).

коэффициент альфа увеличился на столько-то, то этот пункт должен быть исключен. Единственное общепринятое правило заключается в том, что если альфа (или другой показатель, характеризующий надежность-согласованность измерений) меньше 0,7, то инструмент не может считаться надежным. Те его пункты, при исключении которых надежность инструмента в целом увеличивается, необходимо удалить. Если, например, мы определили, что коэффициент альфа при исключении некоторого вопроса анкеты возрастает с 0,65 до 0,75, этот вопрос лучше исключить. Но если коэффициент альфа для исходного набора вопросов 0,8, а при исключении какого-то вопроса увеличивается до 0,95, нужно обратить особое внимание на этот вопрос (например, на то, как он сформулирован), но исключать его не обязательно, так как и с данным вопросом надежность-согласованность измерений приемлема.

Важным показателем качества пункта инструмента (в смысле его согласованности с другими пунктами для обеспечения однородности инструмента в целом) является также коэффициент корреляции между откликами респондентов по этому пункту и общей суммарной шкалой (суммарным показателем по всем пунктам). Считается, что этот показатель не должен быть меньше 0,3.

Если исследователь считает плохо согласованный пункт особо важным для проектируемого и тестируемого измерительного инструмента, то этот пункт необходимо переформулировать. Но после замены или переформулировки хотя бы одного из пунктов необходимо вновь проводить тестирование инструмента и затем снова проверять надежность-согласованность измерений. Следует помнить, что модификацию инструмента можно проводить только по одному пункту на каждом шаге, так как после изменения любого из пунктов все значения коэффициентов надежности изменятся. На практике исследователь часто несколько раз проходит через этапы создания, удаления и замены пунктов анкеты до тех пор, пока не придет к окончательному набору пунктов, формирующих надежный измерительный инструмент.

Коэффициент альфа Кронбаха можно рассматривать как оценку корреляции измерений данным инструментом с измерениями всеми другими инструментами, составленными из такого же числа пунктов, которые случайным образом извлекли из множества всех возможных индикаторов измеряемого свойства. Его можно также интерпретировать как корреляцию между измерениями данным инструментом и «истинными» измерениями, полученными, если бы респондент ответил на *все* возможные вопросы или пункты, направленные на измерение изучаемого свойства.

Коэффициент альфа может также применяться и для решения гораздо более широкого круга задач. Например, с его помощью можно измерять степень согласованности экспертов,

оценивающих тот или иной объект, стабильность данных при многократных измерениях, качество различных шкал и т.д.

Еще один подход к оценке согласованности данных был предложен в 1945 г. Л. Гутманом, который составил формулы для вычисления шести коэффициентов, наиболее важными из них являются первые три (L_1, L_2, L_3). Первый коэффициент определяет нижнюю границу надежности, второй коэффициент – «лучшую» из возможных оценок нижней границы надежности, а третий формально эквивалентен коэффициенту альфа Кронбаха. Доказано, что коэффициент L_2 всегда больше либо равен коэффициенту альфа Кронбаха. Мы не будем приводить формулы для расчета коэффициента Гутмана в силу их достаточной громоздкости, что делает весьма трудоемким расчет этих коэффициентов без специальных статистических пакетов. По-видимому, громоздкость формул и является основной причиной того, что этот подход получил значительно меньшее распространение на практике, чем формула Кронбаха, хотя подход Гутмана был описан в литературе на 6 лет раньше, чем подход Кронбаха. С помощью современных статистических программ коэффициенты Гутмана вычисляются так же просто, как и коэффициент альфа Кронбаха, но в силу «привычки» и того обстоятельства, что в литературе они описаны гораздо реже, коэффициенты Гутмана в исследовательской практике используются не так часто, как коэффициент альфа Кронбаха.

Отметим, что рассмотренные показатели (α Кронбаха, KR_{20}, L_1, L_2, L_3 и др.) не обязательно всегда неотрицательны. Возможны ситуации, когда каждый из этих коэффициентов будет иметь отрицательные значения (это произойдет в случае, если сумма ковариаций между компонентами инструмента отрицательна). В отличие, например, от коэффициента корреляции, отрицательные значения коэффициентов надежности (внутренней согласованности пунктов анкеты) не несут никакой дополнительной информации, кроме той, что компоненты тестируемого измерительного инструмента не согласованы друг с другом и этот инструмент, соответственно, не может считаться надежным.

При использовании процедур проверки обоснованности и надежности измерений может возникнуть определенная сложность, связанная с тем, что инструменты, используемые для измерения тех или иных интересующих исследователя признаков, часто включают в себя несколько различных блоков (например пунктов анкеты), которые не только сформулированы по-разному, но и используют различные измерительные шкалы. Как в таком случае оценить обоснованность и надежность?

На этот вопрос можно дать несколько ответов. Во-первых, рекомендуется измерительный инструмент делать гомогенным (однородным). Прежде всего необходимо,

чтобы инструмент измерял некий единый концепт. Должны использоваться одинаковые шкалы для каждого пункта. Желательно, чтобы пункты анкеты были примерно одинаковы по форме (например, только вопросы или только утверждения). Также желательно, чтобы направление пунктов было одинаковым (например, только утвердительные или только отрицательно сформулированные утверждения), если только исследователь не намерен объединить в одном опросе два параллельных теста для оценки надежности-эквивалентности измерений.

Если исследователь все же использует разнородный инструмент, то можно оценивать обоснованность и надежность по блокам. Для некоторых показателей обоснованности это некритично. Например, дискриминантная валидность всего инструмента оценивается по различительной способности каждого пункта. Но для корректной оценки надежности измерений рекомендуется все же использовать однородный инструмент.

4.3. Использование статистических пакетов для оценки надежности измерений и измерительных инструментов

Статистические пакеты SPSS и STATICTICA включают в себя процедуры для оценки надежности несколькими разными методами. Данные должны вводиться таким образом, чтобы столбцы соответствовали пунктам анкеты, а строки – респондентам. В SPSS для доступа к пакету оценки надежности необходимо в основном меню *Analyze* выбрать *Scale*, затем – *Reliability Analysis*. В этом диалоговом окне выбираются необходимые пункты для анализа и переносятся в поле *Items*. В списке *Model* выбирается нужный метод оценки: *Alpha* (коэффициент альфа Кронбаха), *Split-Half* (метод раздельного коррелирования), *Parallel* (метод параллельного тестирования), *Strict Parallel* (метод параллельного тестирования при предположении равенства средних значений по каждому тесту), *Guttman* (оценка границ надежности методом Гутмана).

В окне *Statistics* можно установить режимы вывода базовых статистических показателей (среднее, стандартное отклонение и др.) для каждого пункта (*Items*) или для всего инструмента в целом (*Scale*). Полезно также выбрать режим, показывающий изменение базовых статистик и индикаторов, характеризующих надежность в случае удаления одного из пунктов (*Scale if item deleted*).

Для иллюстрации работы и сравнения с результатами вычисления «вручную» используем данные из приведенного ранее простого примера (3 пункта анкеты, 5 респондентов, порядковая шкала).

Из отчетов, сформированных SPSS после выполнения процедуры *Reliability Analysis* с использованием разных методов оценки, имеем следующие результаты (выходные формы по разным методам оценки объединены с сокращениями; пункты анкеты обозначены как ITEM1, ITEM2, ITEM3; знаком «*» отмечены комментарии автора учебного пособия, для сравнения результатов полужирным шрифтом отмечены значения, вычисленные ранее с помощью Excel).

* Базовая статистика для каждого пункта анкеты (среднее, стандартное отклонение, количество наблюдений):

		Mean	Std Dev	Cases
1.	ITEM1	,6000	,8944	5,0
2.	ITEM2	1,2000	,4472	5,0
3.	ITEM3	1,4000	,5477	5,0

* Корреляционная матрица пунктов анкеты:

Correlation Matrix

	ITEM1	ITEM2	ITEM3
ITEM1	1,0000		
ITEM2	**,2500**	1,0000	
ITEM3	**,9186**	**,6124**	1,0000

* Базовая статистика по всему инструменту в целом (определяется на основе сумм откликов респондентов на пункты анкеты): среднее значение, дисперсия, стандартное отклонение, количество пунктов:

Statistics for Scale	Mean	Variance	Std Dev	N of Variables
	3,2000	**2,7000**	**1,6432**	3

* Статистика для средних значений пунктов анкеты:

Item Means	Mean	Minimum	Maximum	Range	Max/Min	Variance
	1,0667	,6000	1,4000	,8000	2,3333	,1733

* Статистика для дисперсий по пунктам анкеты:

Item Variances	Mean	Minimum	Maximum	Range	Max/Min	Variance
	,4333	,2000	,8000	,6000	4,0000	,1033

* Статистика для корреляций пунктов анкеты (усредненный коэффициент корреляции, минимальный и максимальный коэффициенты корреляции, диапазон изменения, отношение максимальный/минимальный коэффициенты, дисперсия):

Inter-item Correlations	Mean	Minimum	Maximum	Range	Max/Min	Variance
	,5936	,2500	,9186	,6686	3,6742	,0896

* Статистика по инструменту при условии удаления одного из пунктов (среднее, дисперсия, коэффициент альфа):

```
Item-total Statistics

               Scale          Scale
               Mean         Variance         Alpha
             if Item        if Item        if Item
             Deleted        Deleted        Deleted

ITEM1         2,6000          ,8000         ,7500
ITEM2         2,0000         2,0000         ,9000
ITEM3         1,8000         1,2000         ,3333
```

* Коэффициент альфа Кронбаха:

```
Alpha =   ,7778
```

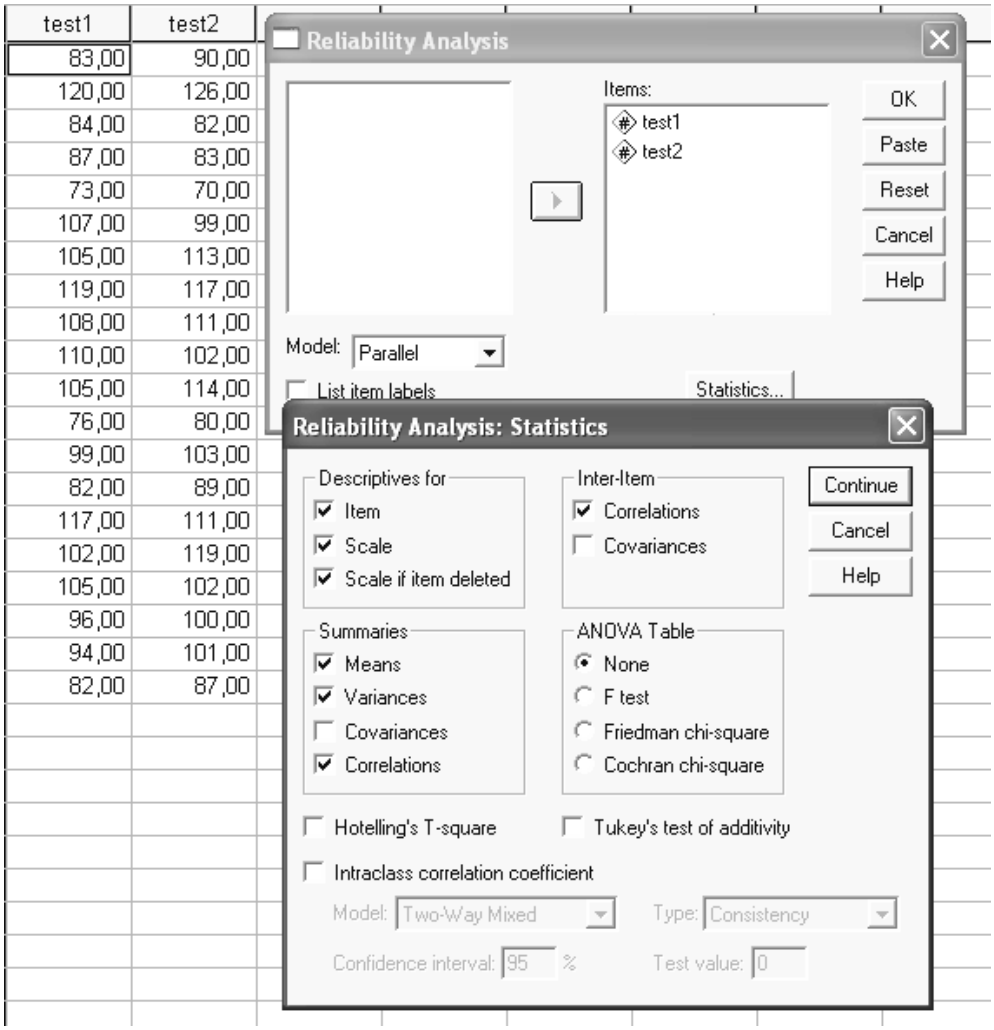
* Коэффициент альфа Кронбаха с учетом стандартизации измерений:

```
Standardized item alpha =   ,8142
```

* Коэффициенты Гутмана:

```
   Lambda 1 =  ,5185    Lambda 2 =  ,8295    Lambda 3 =  ,7778
```

Приведем пример использования метода параллельного тестирования. Предположим, что для оценки надежности-эквивалентности инструмента для оценки творческих стилей (стилей решения нестандартных проблем) использовались 2 инструмента, в которых вопросы имели приблизительно одинаковый смысл, но различались по формулировкам. Тестирование проводилось с интервалом в один день на одной и той же группе респондентов (n = 20). Результаты измерений обрабатывались с помощью метода параллельного тестирования в SPSS.



Ниже представлен фрагмент отчета по выполнению процедуры оценки (некоторые показатели опущены).

```
                          Mean        Std Dev       Cases

  1.     TEST1           97,7000        14,4299        20,0
  2.     TEST2           99,9500        14,9648        20,0

                  Correlation Matrix

               TEST1       TEST2
TEST1          1,0000
TEST2           ,9027      1,0000

Item Variances       Mean    Minimum    Maximum      Range    Max/Min   Variance
                  216,0829   208,2211   223,9447    15,7237     1,0755   123,6171

          Parameter Estimates

      Estimated common variance =     216,0829
                  Error variance =      21,1513
                   True variance =     194,9316
      Estimated common inter-item correlation =     ,9021
```

Обратите внимание на то, что при выполнении этой процедуры помимо коэффициента парной корреляции вычисляются и приводятся в отчете дисперсия истинных значений s_T^2

(*true variance*), дисперсия ошибок измерения s_E^2 (*error variance*) и дисперсия наблюдаемых значений s_O^2 или общая дисперсия (*common variance*). Эти величины, как было показано в предыдущем параграфе, связаны друг с другом:

$$s_O^2 = s_T^2 + s_E^2.$$

Как отмечалось ранее, надежность измерения может определяться как отношение

$$r = \frac{s_T^2}{s_O^2} = \frac{s_O^2 - s_E^2}{s_O^2}.$$

Как видно из отчета, это значение равно 0,9021, что очень близко значению коэффициента корреляции (0,9027). Доказано, что эти два значения будут равны (или почти равны) при равенстве (или приблизительном равенстве) средних двух наборов данных. Соответственно, чем больше будет различие средних значений, тем больше будут различаться коэффициент корреляции и оценка надежности. Это различие учитывается при использовании метода оценки надежности *Strict Parallel*. Для тех же самых данных мы получим следующие значения при выборе метода *Strict Parallel* в окне *Statistics*:

```
  1.     TEST1           97,7000        14,4299        20,0
  2.     TEST2           99,9500        14,9648        20,0

                   Correlation Matrix

               TEST1       TEST2
TEST1          1,0000
TEST2           ,9027      1,0000

Item Variances       Mean    Minimum    Maximum      Range    Max/Min   Variance
                  216,0829   208,2211   223,9447    15,7237     1,0755   123,6171

          Parameter Estimates

      Estimated common mean     =      98,8250
      Estimated common variance =     217,3485
                 Error variance =      22,6250
                  True variance =     194,7235
      Estimated common inter-item correlation =     ,8910
```

Мы видим, что при использовании данной процедуры оценка надежности уменьшается (коэффициент корреляции, естественно, остается тем же). Если же средние значения по результатам двух тестирований будут различаться в большей степени, то и различие между коэффициентом корреляции и оценкой надежности увеличится. Для демонстрации этого увеличим среднее значение по первому тесту на 20 (для этого достаточно к каждому элементу добавить 20). В этом случае получаем следующие результаты:

```
                           Mean        Std Dev       Cases

  1.     TEST1          117,7000        14,4299        20,0
  2.     TEST2           99,9500        14,9648        20,0
```

```
                    Correlation Matrix

                TEST1       TEST2
TEST1           1,0000
TEST2            ,9027      1,0000

Statistics for       Mean   Variance    Std Dev  Variables
      Scale      217,6500   822,0289    28,6710          2

Item Means           Mean    Minimum    Maximum      Range    Max/Min   Variance
                  108,8250    99,9500   117,7000    17,7500     1,1776   157,5313

Item Variances       Mean    Minimum    Maximum      Range    Max/Min   Variance
                  216,0829   208,2211   223,9447    15,7237     1,0755   123,6171

          Parameter Estimates

      Estimated common mean     =     108,8250
      Estimated common variance =     294,8485
                  Error variance =     177,6250
                   True variance =     117,2235
      Estimated common inter-item correlation =     ,3940
```

В этом случае оценка надежности значительно снизилась (0,3940), тогда как коэффициент корреляции не изменился. Таким образом, вычисление лишь коэффициента корреляции может оказаться недостаточным при оценке надежности-эквивалентности методом параллельного тестирования. Методы *Parallel* и *Strict Parallel* также могут использоваться для проверки надежности-стабильности при повторном тестировании.

Для иллюстрации использования процедуры *Split-half* возьмем фрагмент массива данных по оценке организационного климата, полученных Дж. Кабра (John Cabra) из Международного центра исследования творчества (Баффоло, США). Анкета включала в себя 87 пунктов, опрос проводился среди 317 сотрудников 11 компаний Колумбии. Предлагаемый фрагмент данных содержит отклики лишь 30 первых респондентов (из одной компании) на 8 пунктов анкеты. Данные представлены в 5-позиционной шкале Лайкерта (1–5).

№	Пункты анкеты							
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
1	5	4	4	4	4	4	4	5
2	4	4	2	2	4	3	4	4
3	5	2	1	4	5	2	4	5
4	5	5	4	5	4	2	4	5
5	2	2	2	4	2	3	1	4
6	1	1	3	1	2	2	4	3
7	3	2	4	5	3	2	3	4
8	5	4	3	4	4	2	1	4
9	3	4	3	4	4	3	3	4
10	5	3	2	3	3	3	2	4
11	4	4	4	4	4	3	4	4
12	2	1	2	2	2	2	1	2
13	4	3	1	1	1	1	1	5

№	Пункты анкеты							
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
16	5	2	4	5	5	2	2	4
17	4	2	2	3	4	2	4	4
18	4	1	2	4	5	2	2	4
19	5	2	3	2	2	2	2	4
20	4	3	3	4	5	2	3	5
21	4	2	4	4	5	4	4	4
22	5	3	5	5	5	2	4	4
23	3	1	2	5	2	4	3	2
24	3	2	3	2	3	4	2	4
25	3	5	5	5	5	2	3	5
26	5	3	2	2	4	2	2	4
27	5	2	4	4	5	1	3	4
28	3	3	1	5	4	2	4	4

14	3	2	3	2	3	1	3	5	29	4	3	4	1	5	2	2	4
15	3	2	1	2	3	2	2	5	30	4	4	4	2	3	3	3	3

Безусловно, на основе обработки только этих данных нельзя сделать никаких содержательных выводов ни об инструменте в целом, ни о характеристиках организационного климата компании. Однако этот набор данных вполне может быть использован для иллюстрации применения процедур SPSS для оценки надежности измерений. Мы будем рассматривать каждый пункт анкеты в качестве индикатора, характеризующего организационный климат компании, и оценим согласованность этих индикаторов.

Ниже приведены результаты оценки согласованности измерений методами ***Reliability Analysis*** пакета SPSS (пункты анкеты обозначены как ITEM1 – ITEM8; к отчету по использованию ***Split-half*** добавлены результаты использования процедуры ***Alpha*** и ***Guttman***).

		Mean	Std Dev	Cases
1.	ITEM1	3,8333	1,0854	30,0
2.	ITEM2	2,7000	1,1492	30,0
3.	ITEM3	2,9000	1,1847	30,0
4.	ITEM4	3,3333	1,3730	30,0
5.	ITEM5	3,6667	1,1842	30,0
6.	ITEM6	2,3667	,8503	30,0
7.	ITEM7	2,8000	1,0635	30,0
8.	ITEM8	4,0667	,7849	30,0

Statistics for	Mean	Variance	Std Dev	N of Variables
Part 1	12,7667	10,5989	3,2556	4
Part 2	12,9000	5,6793	2,3831	4
Scale	25,6667	25,8851	5,0877	8

Item Means	Mean	Minimum	Maximum	Range	Max/Min	Variance
Part 1	3,1917	2,7000	3,8333	1,1333	1,4198	,2529
Part 2	3,2250	2,3667	4,0667	1,7000	1,7183	,6069
Scale	3,2083	2,3667	4,0667	1,7000	1,7183	,3688

Item Variances	Mean	Minimum	Maximum	Range	Max/Min	Variance
Part 1	1,4468	1,1782	1,8851	,7069	1,6000	,0940
Part 2	,9681	,6161	1,4023	,7862	2,2761	,1330
Scale	1,2075	,6161	1,8851	1,2690	3,0597	,1628

Inter-item Correlations	Mean	Minimum	Maximum	Range	Max/Min	Variance
Part 1	,2825	,1967	,3825	,1858	1,9445	,0067
Part 2	,1277	-,2962	,4107	,7069	-1,3865	,0633
Scale	,2258	-,2962	,4918	,7881	-1,6604	,0335

Item-total Statistics

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Squared Multiple Correlation	Alpha if Item Deleted
ITEM1	21,8333	20,6264	,4139	,3382	,6869

ITEM2	22,9667	19,6195	,4857	,3909	,6708
ITEM3	22,7667	19,7023	,4544	,3140	,6778
ITEM4	22,3333	18,5747	,4584	,2785	,6783
ITEM5	22,0000	18,0690	,6371	,5194	,6335
ITEM6	23,3000	24,7000	,0547	,2038	,7446
ITEM7	22,8667	20,6713	,4222	,2630	,6853
ITEM8	21,6000	23,0069	,3004	,4025	,7081

Reliability Coefficients 8 items

Correlation between forms = ,6191 Equal-length Spearman-Brown = ,7648

Guttman Split-half = ,7423

Alpha for part 1 = ,6053 Alpha for part 2 = ,4242

4 items in part 1 4 items in part 2

Alpha = ,7164 Standardized item alpha = ,6999

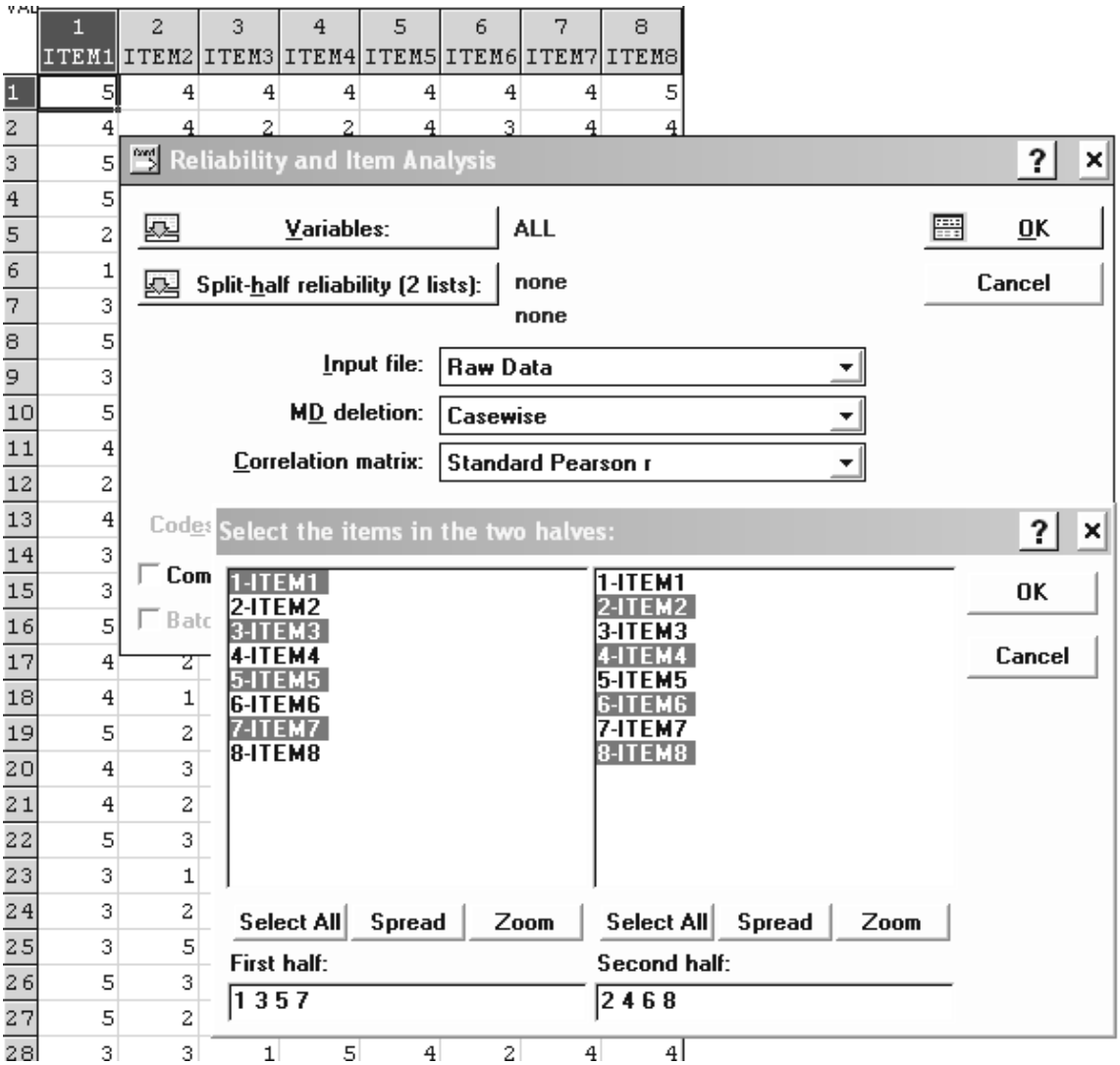
Lambda 1 = ,6268 Lambda 2 = ,7405 Lambda 3 = ,7164

Как видно из этой выходной формы, для использовании метода раздельного коррелирования исходный набор измерений разделен на две группы по 4 пункта в каждой. Статистические показатели как по анкете в целом, так и по ее отдельным пунктам сгруппированы по двум сформированным группам. Коэффициент корреляции между двумя группами (*correlation between forms*): 0,6191. Коэффициент надежности-согласованности (коэффициент межгрупповой корреляции, скорректированный по формуле Спирмена-Брауна (*equal-length Spearman-Brown*)): 0,7648. Коэффициент раздельного коррелирования Гутмана (*Guttman Split-half*) несколько ниже: 0,7423 (различия в дисперсиях по группам довольно велики). Коэффициент альфа Кронбаха (*Alpha*): 0,7164; коэффициент альфа Кронбаха с учетом стандартизации измерений (*Standardized item alpha*): 0,8142. Коэффициент альфа отдельно рассчитывается также для каждой из двух групп (*Alpha for part 1* и *Alpha for part 2*): 0,6053 и 0,4242 соответственно. Коэффициенты лямбда Гутмана: 0,6268; 0,7405; 0,7164.

Мы видим, что если бы мы рассматривали данный фрагмент анкеты как самостоятельный измерительный инструмент, мы должны были бы сделать вывод о том, что он требует корректировки, так как полученные показатели надежности хотя и имеют значения на уровне приемлемых, но все же недостаточно высоки, чтобы можно было бы говорить о хорошей надежности-согласованности измерений. Очевидно, что пункт 6 (ITEM6) в наименьшей степени согласован с остальными пунктами. При удалении этого пункта из анкеты коэффициент альфа увеличивается с 0,7164 до 0,7446. Именно на этот пункт необходимо обратить внимание в первую очередь. Его необходимо удалить, заменить или переформулировать (это обсуждалось в предыдущем параграфе).

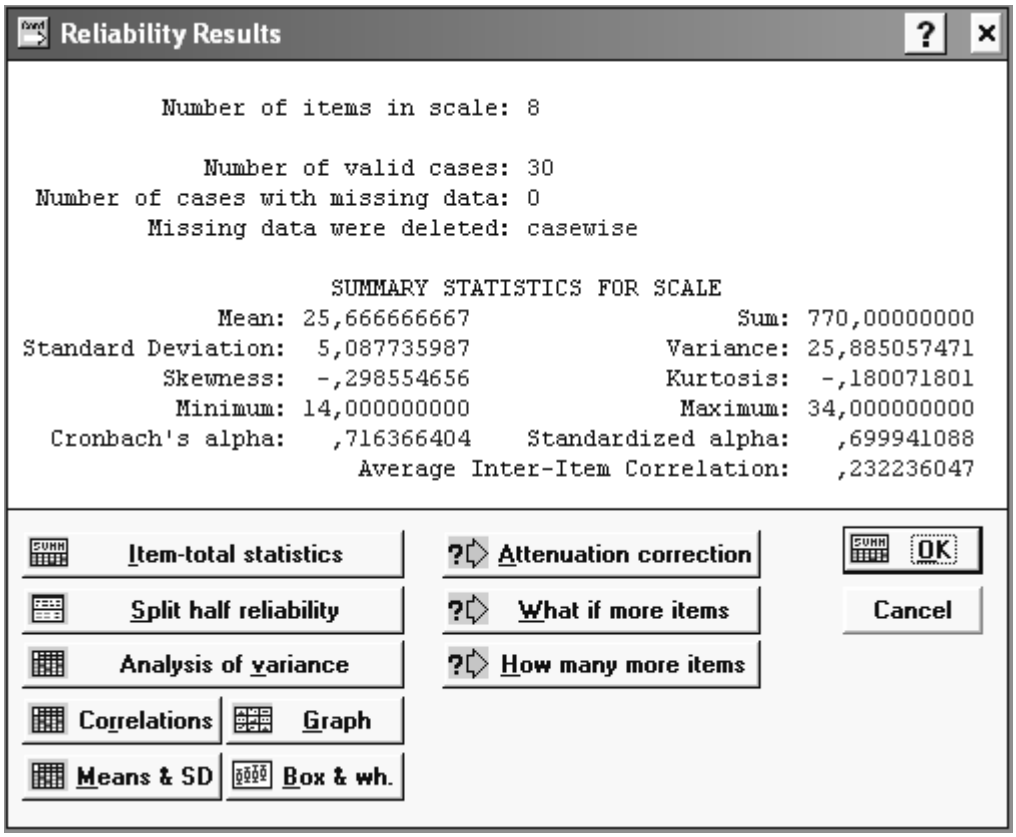
Далее мы кратко опишем возможности пакета STATISTICA для оценки надежности измерений. После запуска пакета необходимо выбрать модуль *Reliability/Item Analysis* (из

основного меню *Analysis, Other Statistics*), затем ввести (набрать, скопировать или загрузить из файла) исходные данные. При необходимости можно добавить или удалить число строк и (или) столбцов для ввода данных (*Edit, Cases* и (или) *Variables, Add* или *Delete* соответственно). В окне *Variables* указать переменные для обработки (пункты анкеты): ITEM1-ITEM8. В окне *Split-half reliability* выбираются переменные двух групп для использования метода раздельного коррелирования (если предполагается использование этого метода). Обычно в одну группу отбираются переменные с нечетными номерами, в другую – с четными.



После нажатия на ОК открывается окно *Review Descriptive Statistics*, где нажатием на соответствующие кнопки можно вывести формы, содержащие базовые статистики для переменных (средние значения, стандартные отклонения, коэффициенты корреляции и др.) в табличной и графической формах. Если такие промежуточные результаты не требуются (они

будут доступны и на следующем этапе), можно, нажав кнопку *OK*, сразу перейти к итоговым результатам расчетов.

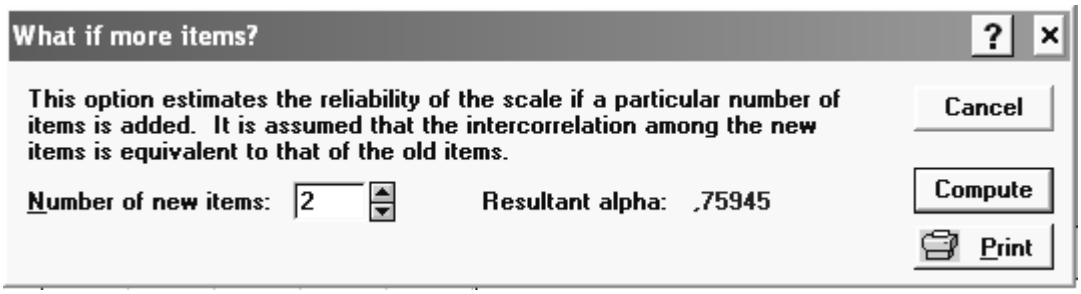


Как видим, получены те же оценки надежности, что и в SPSS. Так же, как в SPSS, можно проследить изменение надежности инструмента при удалении одного из пунктов (кнопка *Item-total statistics*).

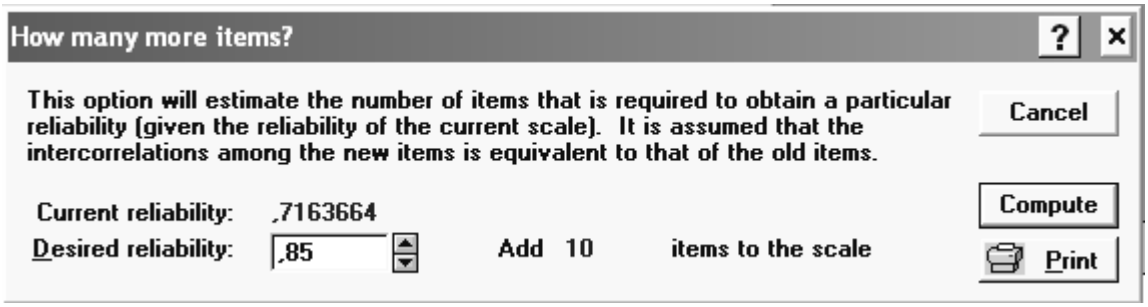
RELIABL. ANALYSIS	Cronbach alpha: ,716366 Standardized alpha: ,699941 Average inter-item corr.: ,232236				
variable	Mean if deleted	Var. if deleted	StDv. if deleted	Itm-Totl Correl.	Alpha if deleted
ITEM1	21,83333	19,93889	4,465298	,413871	,686932
ITEM2	22,96667	18,96556	4,354946	,485709	,670787
ITEM3	22,76667	19,04556	4,364121	,454442	,677771
ITEM4	22,33333	17,95555	4,237400	,458426	,678342
ITEM5	22,00000	17,46667	4,179314	,637086	,633503
ITEM6	23,30000	23,87667	4,886375	,054672	,744551
ITEM7	22,86667	19,98222	4,470148	,422184	,685313
ITEM8	21,60000	22,24000	4,715930	,300417	,708067

Ценным дополнением к анализу надежности в пакете STATISTICA является ряд интерактивных процедур типа «что-если», помогающих при разработке инструментов измерения. Например, при добавлении (удалении) некоторого количества вопросов пользователь может вычислить ожидаемую надежность или же оценить количество вопросов, которые нужно внести в анкету или исключить из нее, чтобы добиться нужной надежности (опции *What if more items* и *How many more items* в окне *Reliability Results*). Если,

например, планируется добавить в анкету два новых пункта (с согласованностью, соответствующей средней по анкете), то можно ожидать повышения надежности до 0,76.



Если, допустим, исследователь стремится обеспечить надежность инструмента на уровне 0,85, легко определить, что необходимо добавить 10 дополнительных пунктов с согласованностью, соответствующей средней по анкете.

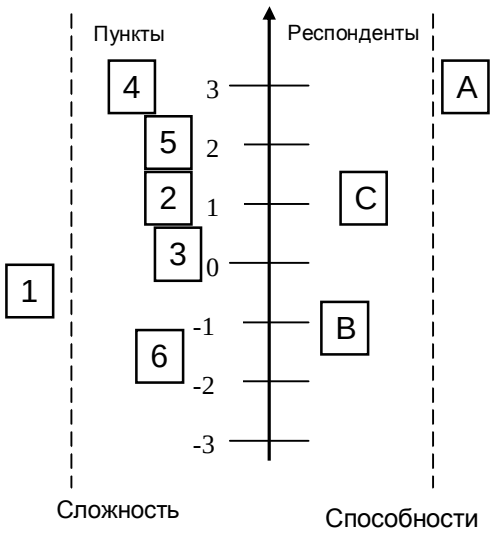


4.4. Оценка согласованности измерений на основе модели Раша

Рассмотренные выше процедуры оценки надежности-согласованности измерений были разработаны в рамках классической теории измерений. Серьезным ее недостатком является то, что во многих случаях при использовании процедур оценки не принимается во внимание вид измерительной шкалы. В частности, для данных в порядковых шкалах используются те же процедуры, что и для интервальных шкал.

Поэтому другой (не альтернативный, но дополняющий) подход к оценке качества измерений и измерительных инструментов может быть построен на основе модели Раша. С помощью этой модели можно ответить на вопросы: «Насколько пункты анкеты согласованы в плане измерения единого конструкта (концепта)?», «Измеряют ли они некий единый фактор или различные факторы?», «Насколько исходные данные подходят для измерения на основе используемой модели?». Модель Раша показывает, насколько каждый пункт анкеты подходит (*fits*) для измерения той или иной характеристики предмета исследования. Показатель надежности пункта (*item reliability index*) характеризует повторяемость результатов: если этот же пункт будет предложен другой группе респондентов, будут ли получены аналогичные результаты? Чем ближе к центральной оси диаграммы Раша расположены пункты анкеты, тем лучше они согласованы (см. рис.). Несогласованные

пункты, расположенные за пунктирной линией, должны быть исключены или переформулированы.



Модель Раша также позволяет оценить надежность измерений для респондентов (*person reliability index*). Этот показатель характеризует повторяемость результатов: если этой же группе респондентов будут предложены вопросы, измеряющие тот же конструкт, будут ли получены аналогичные результаты?

На основе модели могут быть получены ошибки измерений по респондентам, а также степень соответствия откликов респондентов модели. Чем ближе к центральной оси диаграммы Раша расположены измерения, тем они точнее. Измерения, расположенные за пунктирной линией, не могут рассматриваться как надежные и должны быть исключены из анализа.

Рассмотрим процедуру оценки результатов по данным, использованным в предыдущем параграфе. Исходный файл для обработки выглядит следующим образом.

```
ITEM1=1 ; Starting column of item responses
NI=8          ; Number of items
; SPSS Cases processed = 30
; datum: 1          count: 20
; datum: 2          count: 61
; datum: 3          count: 46
; datum: 4          count: 75
; datum: 5          count: 38
XWIDE = 1          ; this matches the biggest data value observed
CODES = "12345" ; matches the data
NAME1 = 10         ; Starting column for person label in data record
; Person Label variables: in label: in line
NAMLEN = 0          ; Length of person label
&END ; Item labels follow:
ITEM1     ; Item 1
ITEM2     ; Item 2
ITEM3     ; Item 3
ITEM4     ; Item 4
ITEM5     ; Item 5
ITEM6     ; Item 6
```

```
ITEM7     ; Item 7
ITEM8     ; Item 8
END NAMES
54444445 1
44224344 2
52145245 3
```

и т.д. (всего 30 респондентов).

Поскольку используется порядковая шкала с позициями 1–5, в поле CODES указаны соответствующие коды (12345). Результаты обработки данных (измерения по пунктам анкеты) представлены в таблице.

ITEM STATISTICS: MEASURE ORDER

ENTRY NUMBER	RAW SCORE	COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PTMEA CORR.	ITEM
6	71	30	1.04	.21	1.13	.6	1.23	.9	.23	ITEM6
2	81	30	.62	.20	.88	-.5	.89	-.4	.65	ITEM2
7	84	30	.50	.20	.85	-.6	.96	-.1	.59	ITEM7
3	87	30	.39	.20	.93	-.2	.92	-.2	.63	ITEM3
4	100	30	-.12	.20	1.30	1.2	1.20	.9	.65	ITEM4
5	110	30	-.54	.21	.79	-.8	.78	-.8	.75	ITEM5
1	115	30	-.77	.22	1.03	.2	1.07	.4	.58	ITEM1
8	122	30	-1.13	.24	.86	-.5	.74	-.9	.46	ITEM8
MEAN	96.3	30.0	.00	.21	.97	-.1	.98	.0		
S.D.	17.0	.0	.71	.01	.16	.7	.17	.7		

Как указывалось ранее, согласно классической модели Раша пункты анкеты с большим количеством исходных баллов получают меньшие значения в шкале Раша. В специальной литературе, посвященной оценке организационного климата, широко используется «метеорологическая» аналогия: организационный климат, менее благоприятный для творчества сотрудников, рассматривается как более «холодный». Поэтому для данной задачи с целью более «естественной» интерпретации перед измерениями пунктов следует поменять знак с тем, чтобы отрицательные значения соответствовали менее благоприятным («более холодным») факторам организационного климата, которые требуют изменений в первую очередь.

В столбцах `INFIT` и `OUTFIT` указаны параметры, характеризующие соответствие данных модели Раша. В поле `MNSQ` (*mean-square statistic*) приводится статистика соответствия исходных данных измерительной модели, полученная на основе средних сумм квадратов отклонений теоретических значений от эмпирических. Значения `MNSQ` характеризуют степень «случайности» результатов или несоответствие данных используемой модели измерения. «Ожидаемые» значения `MNSQ` находятся вблизи 1. Высокие значения `MNSQ` `OUTFIT` могут быть связаны со «случайными» откликами респондентов. Высокие значения `MNSQ` `INTFIT` обычно интерпретируются как индикатор низкой валидности инструмента, т.е. низкой пригодности

инструмента для тех задач, для решения которых он разрабатывался. Например, если в результате тестирования обнаруживаются высокие значения MNSQ INTFIT, то это свидетельствует о том, что данный тест непригоден для группы испытуемых, в которой он предъявлялся.

Значения MNSQ больше 2 рассматриваются как несоответствующие модели измерения и не могут быть использованы при анализе результатов. Такие пункты рекомендуется исключать из инструмента. Наиболее качественными и значимыми (*productive*) считаются измерения, для которых значения MNSQ лежат в диапазоне от 0,5 до 1,5. Более высокие значения (>1,5) указывают на неопределенность и «шум» в исходных данных. Слишком низкие значения (<0,5) также не очень желательны, поскольку они свидетельствуют об избыточности, «информационной перегруженности» инструмента.

Более критичными для измерений являются высокие значения MNSQ, поэтому анализ результатов следует начинать с пунктов, характеризуемых высоким MNSQ. Более важными с точки зрения характеристики качества результатов, как указывалось, являются значения INTFIT. Рассчитываемая статистика соответствия зависит от объема данных. Если количество наблюдений (респондентов) меньше 30, модель может оказаться не очень чувствительной к «неподходящим» пунктам («подходит все»). В случае, если количество наблюдений больше 300, модель, напротив, может оказаться слишком чувствительной («ничто не подходит»). В поле ZSTD приводятся стандартизованные значения MNSQ (со средним 0 и стандартным отклонением 1). Приемлемым считается значение $-2{,}0 \leq$ ZSTD $\leq +2{,}0$.

Для данного примера статистика соответствия по измерениям всех пунктов находится в данном диапазоне, поэтому все они могут использоваться для дальнейшего анализа.

В поле PTMEA CORR. (или PTBIS CORR.) приведено значение *коэффициента бисериальной корреляции*, который показывает, насколько согласуются отклики всех респондентов по данному пункту с их откликами по всем пунктам. Данный коэффициент может принимать значения от –1 до 1. Этот коэффициент рассматривается как показатель надежности-согласованности и может быть использован для определения и возможного исключения слабосогласованных пунктов. В данном примере значение коэффициента для пункта 6 является довольно низким (0,23), поэтому этот пункт должен быть переформулирован или исключен.

Аналогично интерпретируются результаты измерений по респондентам.

```
   PERSON STATISTICS:  MEASURE ORDER

+-------------------------------------------------------------------------+
|ENTRY    RAW                    MODEL|   INFIT   |  OUTFIT  |PTMEA|       |
|NUMBER  SCORE  COUNT  MEASURE  S.E. |MNSQ  ZSTD|MNSQ  ZSTD|CORR.| PERSON|
```

|-----------------------------------+---------+---------+-----+-------|
1 34 8 1.70 .49	.20 -2.0	.27 -1.7	.77	1
4 34 8 1.70 .49	1.09 .4	.87 .0	.63	4
22 33 8 1.48 .46	1.07 .3	1.03 .2	.63	22
25 33 8 1.48 .46	1.80 1.4	1.93 1.5	.39	25
11 31 8 1.10 .42	.20 -2.4	.25 -2.0	.55	11
21 31 8 1.10 .42	.75 -.4	.71 -.5	.40	21
16 29 8 .77 .40	1.04 .2	1.02 .2	.77	16
20 29 8 .77 .40	.30 -2.0	.37 -1.6	.93	20
3 28 8 .62 .39	1.39 .9	1.32 .8	.81	3
9 28 8 .62 .39	.44 -1.5	.49 -1.2	.41	9
27 28 8 .62 .39	.89 -.1	.90 -.1	.82	27
2 27 8 .47 .39	.96 .1	.85 -.2	.32	2
8 27 8 .47 .39	.96 .1	.88 -.1	.69	8
7 26 8 .32 .38	.86 -.2	.82 -.3	.48	7
28 26 8 .32 .38	1.29 .7	1.16 .5	.47	28
30 26 8 .32 .38	1.11 .4	1.13 .4	-.10	30
10 25 8 .18 .38	.54 -1.1	.56 -1.0	.67	10
17 25 8 .18 .38	.43 -1.5	.39 -1.6	.76	17
29 25 8 .18 .38	1.34 .8	1.19 .5	.54	29
18 24 8 .03 .38	.82 -.3	.77 -.4	.83	18
26 24 8 .03 .38	.59 -.9	.56 -1.0	.79	26
24 23 8 -.11 .38	.89 -.1	1.07 .3	.25	24
14 22 8 -.26 .39	.63 -.8	.68 -.6	.78	14
19 22 8 -.26 .39	.77 -.4	.67 -.6	.67	19
23 22 8 -.26 .39	2.42 2.4	2.61 2.5	-.12	23
5 20 8 -.57 .40	1.16 .5	1.22 .6	.34	5
15 20 8 -.57 .40	.60 -.8	.58 -.9	.80	15
6 17 8 -1.07 .43	2.10 1.9	2.43 2.2	-.09	6
13 17 8 -1.07 .43	2.26 2.0	2.08 1.8	.64	13
12 14 8 -1.71 .50	.34 -1.4	.50 -.9	.46	12
-----------------------------------+---------+---------+-----+-------
MEAN 25.7 8.0 .29 .41	.98 -.1	.98 -.1		
S.D. 5.0 .0 .80 .03	.56 1.2	.59 1.1		
+---+

Более высокие значения в столбце MEASURE соответствуют более «положительному» восприятию организационного климата в компании. По результатам статистики соответствия (INFIT, OUTFIT) отклики респондентов 23, 6, 13 (а также, возможно, 11) должны быть исключены из анализа.

Необходимо отметить, что качество измерений, определяемое «классическими» методами (в том числе альфа Кронбаха), и качество измерений в модели Раша имеют разные смыслы (разную внутреннюю логику и разные вычислительные процедуры). Например, коэффициент альфа Кронбаха основан на идее о внутренней согласованности элементов измерительного инструмента, а модель Раша предлагает инструментарий для оценки соответствия данных модели (*fit statistics*). Поэтому прямо сопоставлять эти подходы нельзя, как нельзя напрямую сопоставить результаты их применения. В частности, нельзя сравнить согласованность измерений (например, вычислением коэффициента Кронбаха) по исходным данным и после их преобразования в шкалу Раша. Дело в том, что применение модели Раша дает интегрированные результаты и по пунктам, и по респондентам. В итоге осуществляется переход к вероятностным оценкам (например, может быть оценена вероятность правильного

ответа на определенный вопрос определенным респондентом). Коэффициент альфа Кронбаха рассчитывается по фиксированным эмпирическим оценкам каждого респондента по каждому пункту[15]. Поэтому при оценке надежности измерений и измерительных инструментов следует для полноты анализа не заменять, а дополнять одни подходы другими.

[15] Такая несопоставимость в моделях в определенном смысле отражает нестыковку, несопоставимость научных парадигм, проанализированную Т. Куном в его знаменитой книге «Структура научных революций».

Вопросы, задания и упражнения (часть 1)

В данном разделе вопросы, задания и упражнения разделены по тематике соответствующих глав.

Задания к главе 1

1. В чем заключаются различия номинативной, порядковой, интервальной и относительной шкал? Как эти различия влияют на выбор математических методов, которые могут быть использованы для обработки данных?

2. Какая шкала может быть использована для измерения?

- Регионы, в которых расположены офисы компании.
- Количество сотрудников.
- Доход компании.
- Затраты на осуществление рекламной кампании.
- Профессиональный статус (должность) в анкете.
- Профессиональный статус (должность) как мера продвижения по службе.
- Телефонный номер.
- Время решения задачи.
- Профессия.
- Номера домов.
- Сортность продукции.
- Образование (начальное, среднее, высшее, ученая степень).
- Образование (количество классов школы).
- Возраст.
- Год рождения.
- Место рождения.
- Успеваемость (средний балл).
- Профилирующая дисциплина (специальность).
- Знак Зодиака.
- Отношение к принципу «Лучший способ борьбы с наркотиками – это запрещение наркотиков».
- Количество братьев и сестер.
- Национальность.

- Количество учителей (врачей, детей, государственных служащих и др.) на душу населения.
- Физическая привлекательность (оценивается экспертами в баллах).
- Количество аварий по городу в целом за неделю.
- Количество аварий (на каждой из 10 главных магистральных трасс региона).
- Количество аварий в городе при классификации по 3 категориям: «минимальные повреждения», «средние повреждения», «тяжелые повреждения».
- Удовлетворенность клиентов магазина покупкой (покупками).

3. Какие характеристики следующих объектов для исследования могут быть измерены и в какой шкале: а) стиральный порошок; б) сотрудники фирмы; в) деятельность компании? Укажите не менее 4 характеристик по каждому объекту, измеряемых в 4 различных шкалах.

4. Вы получили данные о заработной плате каждого из сотрудников компании. Как эти данные могут быть представлены с помощью номинативной, порядковой, интервальной и относительной шкал? Объясните потерю информации при переходе от более мощных к менее мощным шкалам (относительная – интервальная – порядковая – номинативная).

5. Вы собираетесь подготовить и издать книгу «100 лучших предприятий города N». Какие характеристики вы будете измерять и по какой шкале? Будете ли вы использовать какой-либо комплексный показатель? Обоснуйте свой ответ.

6. Вам необходимо разработать анкету для опроса студентов, на основе которого можно было бы оценить качество преподаваемых курсов. Что именно вы бы оценивали? Какие вопросы могли бы быть включены в анкету? Какие шкалы вы бы использовали?

7. Маркетолог авиакомпании разрабатывает исследование для оценки мнения пассажиров. Какие вопросы могут быть включены в анкету? Предложите вопросы, отклики на которые могут быть измерены по каждой из 4 типов шкал.

8. Разрабатывается исследование для оценки мнения студентов о высшем учебном заведении. Какие вопросы могут быть включены в анкету? Предложите вопросы, ответы на которые могут быть измерены по каждой из 4 типов шкал.

Задания к главе 2

1. Выберите из второго задания предыдущего раздела свойство, которое может быть измерено в шкале Лайкерта. На примере выбранного свойства покажите преимущества и недостатки использования этой измерительной шкалы.

2. Для выполнения упражнения по преобразованию данных читателю предлагается выборка из данных по оценке организационного климата двух компаний (приложение 2.1). Данные представлены в шкале Лайкерта. Рассматривая каждый пункт анкеты в качестве индикатора, характеризующего организационный климат, постройте и сравните профили организационного климата двух компаний. Представьте данные на двух графиках. На одном графике представьте профили по данным в шкале Лайкерта, на другом – после конвертации исходных данных в шкалу Раша. Проследите изменение результатов. Предложите возможные объяснения, сделайте выводы, объясните сложность использования *z*-шкалы и *T*-шкалы для данного случая.

Цель этого задания – освоение методики конвертирования данных, а не получение содержательного вывода по конкретному исследованию. Поэтому содержательный смысл вопросов (индикаторов) не приводится, а используются формальные обозначения индикаторов: «пункт1», «пункт2» и т.д.

Рекомендуется выполнить задание в среде Excel с использованием функций СРЗНАЧ (вычисление среднего значения), СТАНДОТКЛОН (вычисление стандартного отклонения) и стандартных графических инструментов. Конвертирование данных может быть осуществлено с помощью WINSTEPS или в Excel (по алгоритму, представленному в параграфе 2.3). При преобразовании данных в среде Excel полезным может оказаться использование функции СЧЕТЕСЛИ (диапазон поиска; значение) для расчета частот (количества респондентов, поставивших тот или иной балл).

3. Для выполнения упражнения по работе с методом QIPM и преобразованию данных читателю предлагается выборка (приложение 2.2), содержащая оценки 33 факторов 11 респондентами в категориях «важность» (*I*) и «реализованность» (*P*). Необходимо агрегировать оценки по каждой из категорий в двух шкалах: порядковой шкале и шкале Раша. Учитывайте, что от некоторых респондентов не получены отклики по ряду пунктов. Результаты необходимо представить на двух графиках в осях «важность» (*I*) и «реализованность» (*P*). На одном графике должны быть представлены результаты, полученные по исходным данным, на другом – после преобразования данных в шкалу Раша.

Рекомендуется выполнить задание в среде MS Excel с использованием функции вычисления среднего значения СРЗНАЧ и стандартных графических инструментов.

Преобразование данных в шкалу Раша может быть осуществлено с помощью WINSTEPS или в Excel (см. предыдущее задание). Помимо графиков результаты необходимо представить также в табличном виде:

Фактор	I	P	Квадрант

Квадранты матрицы QIPM можно обозначить как «юго-восточный» (правый нижний) – 1, «северо-восточный» (правый верхний) – 2, «юго-восточный» (левый верхний) – 3, «юго-западный» (левый нижний) – 4. «Попадание» в квадрант можно определить либо «вручную» (определяя положение фактора относительно средних значений «важности» и «реализованности»), либо используя логическую функцию Excel ЕСЛИ.

Сравните полученные результаты с точки зрения приоритетности факторов. Изменилась ли приоритетность факторов при переходе к другой шкале? (Для оценки изменения приоритетности факторы можно ранжировать по характеристикам (I и P) и проследить изменение рангов при изменении шкалы). Сделайте выводы.

Цель этого задания – освоение методики QIPM и конвертирования данных, а не получение содержательного вывода по конкретному исследованию. Поэтому в задании используются формальные обозначения факторов: «фактор1», «фактор2» и т.д.

Задания к главе 3

1. Это задание направлено на закрепление понимания центральной предельной теоремы. В таблице приведен список элементов генеральной совокупности, представляющий собой возраст участников студенческих строительных отрядов, сформированных на базе вузов города в 200х году (N = 100).

№	возраст	№	возраст	№	возраст	№	возраст	№	возраст
1	17	21	19	41	22	61	20	81	21
2	20	22	22	42	19	62	18	82	20
3	20	23	19	43	19	63	20	83	25
4	19	24	23	44	20	64	19	84	20
5	20	25	19	45	20	65	20	85	22
6	18	26	20	46	23	66	22	86	18
7	21	27	25	47	17	67	17	87	21
8	19	28	18	48	18	68	21	88	20
9	20	29	20	49	21	69	21	89	22
10	19	30	20	50	20	70	21	90	21
11	17	31	19	51	22	71	20	91	21
12	20	32	22	52	19	72	18	92	20
13	20	33	19	53	19	73	20	93	24
14	19	34	23	54	20	74	19	94	20

15	20	35	19	55	20	75	20	95	22
16	18	36	20	56	23	76	22	96	18
17	21	37	23	57	17	77	17	97	21
18	19	38	18	58	18	78	21	98	20
19	20	39	26	59	21	79	24	99	22
20	19	40	20	60	20	80	21	100	21

Методом, обеспечивающим случайный отбор (например, с помощью функции-генератора случайных чисел Excel) сформируйте по меньшей мере 10 выборок объемом 10 элементов, рассчитайте среднее для каждой выборки. С помощью стандартных графических инструментов Excel постройте по этим данным полигон частот (ломаную линию, получаемую при соединении точек, лежащих по оси абсцисс на верхних границах каждого класса, а по оси ординат – частот классов)[16]. Построенная кривая (ломаная) будет представлять собой распределение всех возможных выборочных средних. Соответствует ли форма полученной кривой центральной предельной теореме?

Рассчитайте среднее значение всех выборочных средних, среднее генеральной совокупности и сравните результаты. Соответствуют ли эти результаты центральной предельной теореме?

Рассчитайте стандартное отклонение для распределения выборочных средних, стандартное отклонение для генеральной совокупности. Сравните полученные значения. Соответствует ли этот результат центральной предельной теореме?

2. В одном из районов города N проживает 2000 семей. С целью определения среднего размера семьи района было проведено 3% выборочное обследование семей методом случайного повторного отбора. В результате обследования получены следующие данные:

Размер семьи, чел.	0	1	2	3	4	5	6	7	8
Число семей	4	8	14	16	8	4	3	2	1

С вероятностью 0,95 определите пределы, в которых находится средний размер семьи в районе. (При расчете среднего значения и стандартного отклонения по выборке используйте метод расчета взвешенных показателей).

Сделайте предположения о возможных сферах практического использования результатов, полученных в ходе решения задачи. С какой целью могли проводиться

[16] Это упражнение будет более показательным при извлечении большего количества выборок (20, 50, 100) или при формировании выборки с большей численностью (например 15 или 20). Однако в этом случае упражнение будет более трудоемким.

подобные исследования? Насколько полученные результаты могут быть полезны для достижения поставленных целей?

3. Определите необходимое количество респондентов в примерах, приведенных в параграфе 3.5 (измеряемые параметры – «предпочтительное время обеда» и «количество посещений») для обеспечения точности $\pm$10 мин и $\pm$1 визит соответственно с вероятностью 95%. Для решения использовать t-критерий. Выполнить это же задание при предположении, что выборка формируется путем случайного бесповторного отбора (принять, что N = 15000). Сравните результаты.

4. Принимается решение о целесообразности строительства парка семейного отдыха в одном из районов города N. Для установления среднего числа детей в семье проводится соответствующее исследование. Известно, что в районе проживает 2500 семей. Сформирована 2% случайная бесповторная выборка семей. В результате обследования были получены следующие данные:

Число детей в семье	Количество семей
0	10
1	20
2	12
3	4
4	2
5	2

С вероятностью 0,95 определите границы, в которых будет находиться среднее число детей в семье во всем районе. (При расчете среднего значения и стандартного отклонения по выборке используйте метод расчета взвешенных показателей.)

5. В городе N с целью определения средней продолжительности поездки населения на работу предполагается провести выборочное обследование методом случайного повторного отбора. Какова должна быть численность выборки, чтобы с вероятностью 0,95 предельная ошибка среднего не превышала 5 мин при стандартном отклонении 20 мин (известно из предыдущего обследования). С какой целью могли проводиться подобные исследования?

6. В городе N ориентировочно проживают 10 тыс. семей (точные данные неизвестны). С помощью повторной случайной выборки предполагается определить долю семей в городе с

числом детей три и более. Какова должна быть численность выборки, чтобы с вероятностью 0,95 предельная ошибка среднего не превышала 0,02 чел., если на основе предыдущих обследований известно, что дисперсия равна 0,2. Сделайте предположения о возможных сферах практического использования результатов, полученных в ходе решения задачи. С какой целью могли проводиться подобные исследования? Насколько полученные результаты могут быть полезны для достижения поставленных целей?

7. На предприятии с числом рабочих 15 тыс. чел. предполагается определить долю рабочих со стажем работы 20 лет и более. Данные по рабочим еще не внесены в компьютер и в основном находятся в бумажных папках, поэтому определить за короткое время точное значение требуемого параметра невозможно. Предполагается сформировать выборку из имеющихся данных методом бесповторного случайного отбора. Какова должна быть численность выборки, чтобы с вероятностью 0,95 предельная ошибка среднего не превышала 0,03? Из предыдущих обследований известно, что дисперсия равна 0,2. Выполните это же задание при предположении, что выборка формируется методом повторного случайного отбора. Сравните результаты и сделайте выводы.

8. Для определения средней заработной платы работников трех различных категориях была произведена 20% типическая выборка с отбором единиц пропорционально численности типических групп (внутри типов применялся метод случайного бесповторного отбора). Результаты выборки представлены в таблице.

Тип работы	Средняя заработная плата, руб.	Стандартное отклонение, руб.	Число работников, чел.
1	5100	650	150
2	6500	1900	500
3	7800	1150	350

С вероятностью 0,95 определите пределы, в которых находится средняя заработная плата всех работников трех типов исследуемых компаний.

9. Для выявления затрат времени на обработку деталей рабочими разных квалификаций на предприятии была произведена 10% типическая выборка пропорционально численности

выделенных групп (внутри типичных по специальности групп произведен случайный повторный отбор). Результаты обследования представлены в таблице.

Квалификация рабочих	Число рабочих в выборке	Средние затраты времени на обработку одной детали, мин	Стандартное отклонение, мин
I	60	10	1
II	120	14	4
III	80	20	2
IV	40	25	6

С вероятностью 0,95 определите пределы, в которых находятся средние затраты времени на обработку деталей рабочими завода.

10. В городе N проживают 12 тыс. жителей. Из них 7 тыс. женщин и 5 тыс. мужчин. С целью определения доли жителей в возрасте старше 60 лет предполагается провести типическую выборку жителей с пропорциональным отбором. Отбор внутри типов – случайный бесповторный. Какое количество жителей необходимо отобрать, чтобы с вероятностью 0,95 предельная ошибка не превышала 5%? На основе предыдущих исследований принимается, что среднегрупповая дисперсия составляет 1600.

11. В городе N проживают 5 тыс. семей. Из них 1 тыс. составляют одинокие и 4 тыс. – семейные. С целью определения доли семей, имеющих отдельные квартиры, предполагается провести стратифицированную выборку семей с пропорциональным отбором внутри типических групп. Отбор внутри групп случайный повторный. Какое количество семей необходимо отобрать, чтобы с вероятностью 0,95 предельная ошибка не превышала 5%? На основе предыдущих исследований принимается, что среднегрупповая дисперсия по типическим группам составляет 1000.

Задания к главе 4

1. Для проверки надежности-эквивалентности использовались два параллельных теста с интервалом в один день. Коэффициент корреляции по результатам тестирований составил 0,77. Если бы временной интервал был увеличен до 1 недели, как бы изменился результат (коэффициент корреляции увеличился бы, уменьшился или остался прежним)? Объясните ответ.

2. Если в научном журнале вы сталкиваетесь с упоминанием о вычислении коэффициента альфа Кронбаха, вы предполагаете, что оценивалась стабильность, эквивалентность или согласованность?

3. Коэффициент альфа Кронбаха для некоторого набора измерений равен 0,6. Как изменится коэффициент при удвоении количества вопросов за счет добавления параллельного теста?

4. Согласитесь ли вы со следующими утверждениями? Объясните ответ.

Обоснованность (валидность) – более критичная характеристика для качества измерений, чем надежность.

Оценить содержательную валидность сложнее, чем другие виды валидности.

Стабильность и эквивалентность характеризуют одни и те же аспекты надежности измерений.

5. Представьте, что вам необходимо разработать анкету для опроса студентов, на основе которого можно было бы оценить качество преподаваемых курсов (см. аналогичный вопрос к первой главе). Каким образом вы бы проверили (или обосновали) следующие характеристики разработанной вами анкеты:

- содержательная валидность;
- эмпирическая валидность;
- концептная валидность;
- стабильность;
- эквивалентность;
- согласованность?

Постарайтесь дать максимально обоснованный и детальный ответ.

6. Представители одного из крупных автозаводов разработали инструмент для оценки уровня безопасности нового автомобиля. Анкета содержала 12 вопросов относительно нескольких функциональных систем автомобиля, которые обеспечивают безопасность пассажиров. Эти вопросы касались, например, подушек безопасности, системы компьютерного тестирования и др. Проверка анкеты методом тест-ретест и методом параллельного тестирования показало высокую стабильность и эквивалентность измерений,

однако коэффициент альфа Кронбаха оказался ниже 0,35. Какой вывод можно из этого сделать?

7. Дисперсия «истинных» значений проведенных измерений равняется 190, а дисперсия наблюдаемых измерений равна 225. Можно ли, основываясь только на этих данных, судить о надежности измерений? Обоснуйте ответ.

8. Если стандартная ошибка измерений определена как 4,5, а стандартное отклонение истинных значений равно 10, можно ли, основываясь только на этих данных, судить о надежности измерений? Обоснуйте ответ.

9. По данным, представленным в приложении 2.1, оцените дискриминантную валидность вопросов фрагмента анкеты для оценки организационного климата (7 пунктов, 60 респондентов). Проведите анализ на основе расчета модифицированного *t*-критерия ($t_{критическое}$ = 1,75).

10. По данным, представленным в приложении 2.1, оцените надежность-согласованность вопросов анкеты по формуле Спирмена-Брауна. Исключите последовательно каждый вопрос анкеты и вновь оцените согласованность по формуле Спирмена-Брауна. Сделайте выводы.

11. По данным, представленным в приложении 2.1, оцените надежность-согласованность вопросов анкеты на основе коэффициента альфа Кронбаха. Исключите последовательно каждый вопрос анкеты и вновь определите коэффициент альфа Кронбаха. Сделайте выводы. Сравните результаты оценки согласованности двумя методами (формула Спирмена-Брауна и формула Кронбаха).

12. По собственным данным проанализируйте согласованность ваших измерений. Опишите измеряемый фактор (концепт). Определите 2 наиболее согласованных и 2 наименее согласованных пункта. Соответствует ли описание фактора смыслу этих пунктов? Какие пункты можно удалить, чтобы повысить согласованность? Какие новые пункты и сколько можно добавить в анкету, чтобы повысить согласованность? Измените оценки нескольких респондентов (не менее 10) по нескольким пунктам (3–4), проследите, как эти изменения отразились на согласованности пунктов анкеты. Сделайте выводы.

Часть 2. Анализ результатов исследования и статистические выводы

Статистика — это, прежде всего, способ мышления, и для ее применения нужно лишь иметь немного здравого смысла и знать основы математики.

К.Р. МакКоннелл

Глава 5. Основные принципы проверки статистических гипотез

Тема проверки статистических гипотез является очень обширной. Вместе с тем эта тема довольно хорошо разобрана в учебной и специальной литературе, доступной русскоязычному читателю. Существует свыше 100 различных методов и критериев для проверки статистических гипотез. Мы не ставим себе цель рассмотреть в этом издании все существующие методы и критерии. Следуя основному принципу данной книги, основное внимание уделяется логике и методологии проверки гипотез. Также довольно подробно рассматриваются процедуры использования основных статистических критериев, набор которых охватывает большинство ситуаций, возникающих в практике формулирования и проверки статистических гипотез.

5.1. Статистическая достоверность связи и статистическая гипотеза

Любое исследование направлено на определение некоторой характеристики изучаемой генеральной совокупности или выявление связи между признаками (факторами). Такая связь часто исследуется в причинно-следственном аспекте, когда некоторые факторы рассматриваются как причины (*независимые переменные*), а другие – как следствия или результаты (*зависимые переменные*). Связь может характеризоваться не только величиной (степенью связи) и направлением (что показывает, например, коэффициент корреляции), но также и *статистической достоверностью* (*statistical confidence*). Последняя характеристика связи показывает, можно ли распространить результаты, полученные по данной выборке, на всю генеральную совокупность, из которой взята эта выборка.

Любое заключение, полученное из статистического наблюдения / исследования / анализа, индуктивно и строится на конечном числе наблюдений, поэтому оно в принципе не полно и может быть недостоверно. Необходимо обоснование получаемого заключения, т.е. проверка результатов, на которых строится гипотеза, на статистическую достоверность.

Достоверность непосредственно связана с репрезентативностью выборки, т.е. с тем, насколько уверенно данные, полученные по выборке, позволяют судить о соответствующих параметрах генеральной совокупности. Целью исследования почти никогда не является изучение данной конкретной выборки; выборка представляет интерес лишь постольку, поскольку она дает информацию обо всей генеральной совокупности. Статистическая достоверность связи определяется тем, насколько вероятно, что обнаруженная в выборке связь подтвердится (будет вновь обнаружена) на другой выборке той же генеральной совокупности. Оценивая статистическую достоверность связи, мы должны ответить на вопрос: какова вероятность случайного получения результата, подтверждающего наличие связи, которой нет в генеральной совокупности?

Фактически, оценивая статистическую достоверность связи, мы проверяем ту или иную статистическую гипотезу. *Статистическая гипотеза* – утверждение относительно неизвестного параметра (параметров) генеральной совокупности на основе выборочного исследования. Часто гипотеза формулируется как утверждение наличия или отсутствия связи между признаками (зависимыми и независимыми переменными). Предположим, мы сформулировали некую гипотезу. Понятно, что измеряемые значения по выборке всегда отличаются от измеряемых значений по генеральной совокупности, поэтому нужно определить, насколько велико (значимо) это различие. Разница в измерениях может считаться значимой, если есть основания полагать, что она не представляет случайную флуктуацию. Если мы имеем такие основания, мы можем отклонить или признать невозможность отклонения сформулированной гипотезы.

Например, мы проверяем монетку по правилу «орел-решка»: фальшивая она или настоящая. Мы предполагаем, что нефальшивая монетка должна упасть «орлом» примерно столько же раз, сколько и «решкой». Предположим, монетка была подброшена 100 раз, из которых 57 раз выпал орел, а 43 – «решка». Что можно сказать о ее подлинности? Достаточно ли такого количества подбрасываний, чтобы судить о подлинности монетки? С какой вероятностью мы можем судить о том, что монетка настоящая (или фальшивая)? Какова вероятность ошибки? Вот те вопросы, на которые мы должны уметь дать ответ при проверке статистической гипотезы и обосновании статистической достоверности.

Другой пример. Предположим, директора завода интересует, различается ли производительность труда у работников разного возраста на заводе при выполнении одних и тех же видов работ. Допустим, было осуществлено измерение производительности труда в двух разных возрастных группах (с объемом выборок 37 и 59 чел. соответственно). Получен следующий результат: средняя производительность труда в первой группе – 47 деталей/час, во второй – 55 деталей/час со стандартным отклонением 8 и 10 соответственно. Можно ли на

основании полученных результатов сделать вывод о том, что производительность труда рабочих во второй группе выше, чем в первой? Какова вероятность того, что такой вывод может быть ошибочным?

Еще один пример. Проверяется гипотеза о том, что женщины тратят больше времени на разговоры по телефону, чем мужчины. Предположим, что в исследовании принимали участие 52 мужчины и 43 женщины. Среднее время разговора составило 25 мин в день у мужчин и 35 мин в день у женщин (стандартное отклонение 7 и 15 мин соответственно). На первый взгляд, различия обнаружены, и эти результаты подтверждают гипотезу. Однако такой результат может быть получен случайно, даже если в генеральной совокупности различий нет, как и наоборот, когда различия на самом деле существуют. Поэтому закономерен вопрос: достаточно ли полученного различия в средних значениях для того, чтобы утверждать, что вообще *все* женщины в среднем говорят по телефону дольше, чем *все* мужчины? Какова вероятность, что это не так? Является ли это различие *статистически значимым*?

Как видно из приведенных примеров, *абсолютно точный* ответ о различиях или связях в отношении генеральной совокупности по результатам выборочного исследования получить невозможно. Поэтому необходимо определить, достаточно ли велика разность между средними значениями, чтобы эту разность можно было бы объяснить действием независимой переменной, а не случайностью, связанной с малым объемом выборки. Многократное проведение исследования на разных выборках трудоемко (иногда невозможно) и не может обеспечить абсолютно точного ответа, пока не проведено сплошное исследование. Методы статистики позволяют оценить вероятность *случайного* получения такого различия при условии, что на самом деле различий в генеральной совокупности нет. Сферы применения этих методов (статистических тестов) очень широки и разнообразны. В качестве примеров можно назвать несколько приложений:

- сравнение отказов компьютеров разных производителей после 20-дневного тестирования;
- оценка уровня доходов разных групп населения;
- анализ предпочтения товаров в разных демографических группах;
- сравнение числа подписчиков на различные журналы;
- сравнение социально-экономических характеристик в разных регионах;
- сопоставление производительности труда в разных группах работников предприятия;
- анализ общественного мнения (выборы и др.).

5.2. Нулевая и альтернативная гипотезы

Первым шагом проверки статистической гипотезы является формулировка допущений. При решении любой задачи путем применения статистических методов требуется принять определенные предположения относительно данных, например о способе формирования выборки, используемой измерительной шкале, виде распределения и т.д. Такие предположения называются *статистической моделью*. Очень важно обоснованно сформулировать статистическую модель, поскольку она обусловливает выбор конкретных статистических процедур, которые могут использоваться для проверки статистической гипотезы.

Далее формулируется нулевая гипотеза. *Нулевая гипотеза* (*null hypothesis*) – это утверждение об *отсутствии* различий или связи. Согласно нулевой гипотезе (H_0), различие между значениями недостаточно велико, чтобы принять, что это различие вызвано неслучайными факторами. Точная формулировка нулевой гипотезы зависит от проводимой проверки. В случае с одной выборкой нулевая гипотеза может формулироваться следующим образом: изучаемая характеристика по выборке не отличается от определенной характеристики по генеральной совокупности.

Например, производительность труда в среднем по заводу составляет 50 деталей/час. Допустим, выясняется, отличается ли производительность труда рабочих определенной возрастной группы от средних показателей в целом по заводу. Тогда нулевая гипотеза может быть сформулирована как H_0: $\mu = 50$. Такая формулировка означает, что проверяется равенство среднего значения по исследуемой группе рабочих (генеральной совокупности, из которой будет формироваться выборка) определенному значению (в данном случае – средней производительности труда).

Альтернативная гипотеза (*alternative hypothesis*) – гипотеза о значимости различий (утверждает наличие различий или существование связи). В соответствии с альтернативной гипотезой, различия достаточно значимы и обусловлены влиянием неслучайных факторов. *Статистически значимое различие* – это различие, которое настолько велико, что вероятность его возникновения вследствие простой случайности крайне мала. Альтернативная гипотеза (H_A) является, как правило, «рабочей» гипотезой исследования, поскольку часто цель исследователя заключается в том, чтобы собрать доказательства, которые говорили бы в пользу альтернативной гипотезы и на основании которых можно было бы отклонить нулевую гипотезу.

Альтернативная гипотеза может быть *ненаправленной* и *направленной*. Если целью исследователя является показать, что значение изучаемого параметра по выборке «просто»

отличается от значения параметра по генеральной совокупности, то альтернативная гипотеза символически формулируется с помощью знака «≠». Например, H_A: $\mu \neq 50$. В таком случае мы имеем дело с ненаправленной альтернативной гипотезой.

Если исследователя интересуют различия в определенном направлении, может использоваться направленная альтернативная гипотеза, учитывающая направление изменений (различий). То есть если исследователь считает, что истинное значение генеральной совокупности выше или ниже, чем значение, указанное в нулевой гипотезе, то это мнение будет отражено в альтернативной гипотезе. Например, мы можем предположить, что производительность труда в изучаемой возрастной группе больше, чем в среднем по заводу. Тогда альтернативная гипотеза будет иметь следующий вид: H_A: $\mu > 50$. Если мы предполагаем, что производительность труда в изучаемой возрастной группе меньше, чем в среднем по заводу, то альтернативная гипотеза формулируется как H_A: $\mu < 50$.

Если альтернативная гипотеза формулируется как направленная (например, H_A: $\mu > 0$), то нулевая гипотеза может формулироваться в следующем виде: H_0: $\mu \leq 50$. Однако символ строгого или нестрогого равенства (например, «=» или «≤», «≥») должен всегда присутствовать в формулировке нулевой гипотезы и никогда – в формулировке альтернативной гипотезы. В литературе существуют разные мнения по поводу возможности включения в формулировку нулевой гипотезы символов «≤» и «≥», но с точки зрения процедуры проверки и содержательной интерпретации результатов такие, например, формулировки, как «H_0: $\mu = 50$; H_A: $\mu > 0$» и «H_0: $\mu \leq 50$; H_A: $\mu > 0$», полностью эквивалентны. Очень важно, чтобы выбор направления альтернативной гипотезы был основан на априорных теоретических предпосылках, а не только на эмпирических результатах. Поэтому если таких предпосылок нет, то следует формулировать альтернативную гипотезу как ненаправленную.

Нулевая и альтернативная гипотезы представляют полную группу несовместных событий: отклонение одной влечет принятие другой. Поэтому основной принцип проверки статистических гипотез состоит в том, что выдвигается нулевая гипотеза H_0 с тем, чтобы попытаться опровергнуть ее и тем самым подтвердить альтернативную гипотезу H_A. Если результаты статистического теста окажутся таковы, что позволят отклонить H_0, это будет означать, что верна H_A, т.е. выдвинутая рабочая гипотеза подтверждается. Однако всегда следует помнить о том, что если мы не можем отклонить нулевую гипотезу, это не значит, что мы получили ее подтверждение, так как нулевая гипотеза никогда не может быть абсолютно подтверждена. В то же время из отклонения нулевой гипотезы всегда следует принятие альтернативной гипотезы.

При невозможности отклонения нулевой гипотезы нельзя декларировать отсутствие связи или различий, но выводом исследования будет являться именно то, что статистически значимая связь или различия не обнаружены. С другой стороны, при отклонении нулевой гипотезы всегда остается вероятность того, что она верна. Поэтому нельзя *категорично* утверждать, что связь есть или что результаты *доказывают* наличие связи. Более корректным будет утверждение *статистически значимой* связи, что *свидетельствует* или *указывает* на наличие возможной связи.

Невозможность полного подтверждения нулевой гипотезы можно объяснить, например, следующими рассуждениями. Чем слабее зависимость между переменными, тем большего объема требуется выборка, чтобы ее обнаружить. Например, нужно сделать очень много бросков монеты, чтобы доказать, что отклонение от равной вероятности выпадения «орла» и «решки» составляет тысячные доли процента. Если различия или связь близки к нулю, то необходимый объем выборки для возможности их статистически значимого обнаружения приближается к объему всей генеральной совокупности.

Таким образом, при принятии нулевой гипотезы (точнее, невозможности ее отвергнуть) всегда остается вероятность того, что различия или связь все же существуют. Это означает лишь то, что не обнаружены *статистически значимые* различия. Некоторые авторы на основе этого утверждают, что невозможность отклонения нулевой гипотезы является отсутствием какого-либо результата. Но автор данной книги не согласен с тем, что мы не можем сделать содержательный вывод на основании невозможности отклонения нулевой гипотезы. Действительно, «рабочей» гипотезой исследователя чаще всего становится альтернативная гипотеза. Однако бывают случаи, когда исследователь заинтересован в подтверждении (свидетельстве) отсутствия связи. Например, проверка «тест-ретест», оценка различий в заработной плате, эффективности работы, ожидаемых и фактических результатов и т.д. На основе необнаружения статистически значимых различий также можно принимать решения как в исследовательской работе, так и бизнес-практике.

Еще раз обращаем внимание читателя на то, что нулевые гипотезы всегда формулируются в варианте «отсутствие различий» или «отсутствие связей», т.е. фактически в варианте нулевого различия (например, средних значений). При иной формулировке нулевые гипотезы было бы невозможно опровергнуть. Ведь если нулевая гипотеза формулировалась бы в терминах *наличия* связи или различий, то степень связи или различий была бы неопределенной, с бесконечным количеством вариантов. Предположим, что мы бы сформулировали H_0: $\mu \neq 0$. Что бы мы ожидали получить, если бы нулевая гипотеза была истинной? Все, что угодно! Нулевую гипотезу можно опровергнуть (с определенным

уровнем значимости) только в том случае, если она сформулирована в виде строгого или нестрогого равенства.

5.3. Виды статистических ошибок и уровни статистической значимости

Тот или иной вывод, полученный на основе выборочного исследования, с некоторой вероятностью может оказаться ошибочным, и обычно вероятность ошибки тем меньше, чем больше выборка. В общем случае, чем больше получено результатов, тем в большей степени по различиям, например между двумя выборками, можно судить о том, что действительно имеет место в тех генеральных совокупностях, из которых взяты эти выборки. Однако обычно используемые выборки относительно невелики, и тогда вероятность ошибки может быть значительной.

Выделяют два главных типа статистических ошибок, возможных при проверке статистических гипотез. *Статистическая ошибка первого рода* (*Type I Error*), или α-ошибка, – ошибка *обнаружить* различия или связи, которые на самом деле *не существуют*. То есть ошибка первого рода – это ошибка отклонения истинной нулевой гипотезы. *Статистическая ошибка второго рода* (*Type II Error*), или β-ошибка, – это ошибка *не обнаружить* различия или связи, которые на самом деле *существуют*. То есть это ошибка принятия решения, в результате которого ложная нулевая гипотеза не может быть отклонена[17].

Более «критичной» считается статистическая ошибка первого рода. В книге Купера и Шиндлера (Cooper & Shindler, 1998) приводится «судебная» аналогия статистических ошибок двух видов. Предположим, присяжные заседатели выносят вердикт «не виновен» или «виновен». Согласно этой аналогии, ошибкой первого рода будет ситуация, когда невинный обвинен, а ошибкой второго рода – когда виновный освобожден.

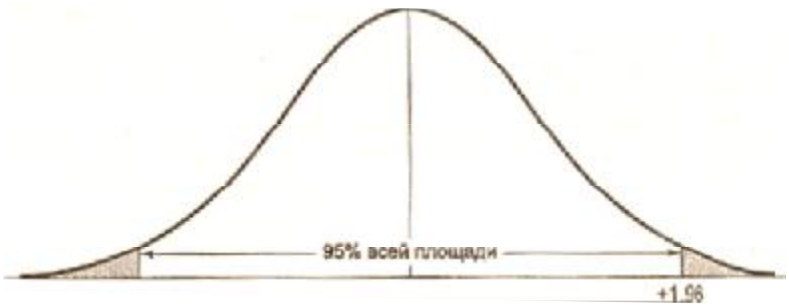
С понятиями ошибок первого и второго рода связано решение важного вопроса о том, что можно считать критерием отклонения или принятия статистической гипотезы, т.е. по каким критериям мы можем судить о том, что статистическая ошибка *маловероятна*, или о том, что связь *статистически значима*.

Уровень значимости (*level of significance*) (также используются термины *уровень достоверности*, *уровень надежности*, *доверительный уровень*, *вероятностный порог*) – это

[17] М. Фокс, преподаватель Международного центра изучения творчества в Баффоло (США), говорит еще об «ошибке третьего рода», которую он шутливо называет «Mike Fox Type Error» – «когда неправильная проблема решается правильными методами». Эта шутка прямого отношения к нашей теме не имеет, но она указывает на ситуацию, когда исследователи действительно получают ошибочный результат только потому, что недостаточно внимания уделили анализу разрабатываемой проблемы.

пороговая (критическая) вероятность ошибки, заключающейся в отклонении нулевой гипотезы, когда она верна. Другими словами, это допустимая (с точки зрения исследователя) вероятность совершения статистической ошибки первого рода – ошибки того, что различия сочтены существенными, а они на самом деле случайны. Обычно используют уровни значимости (обозначаемые α), равные 0,05, 0,01 и 0,001. Например, уровень значимости, равный 0,05, означает, что допускается не более чем 5-процентная вероятность ошибки.

Понятие уровня статистической значимости связано также со свойством нормального распределения выборочных средних, которое мы рассматривали в первой части книги. Из центральной предельной теоремы нам фактически заранее известно распределение средних для случая, когда верна нулевая гипотеза. Это распределение позволяет установить, насколько вероятно то или иное случайное отклонение среднего по выборке от среднего по генеральной совокупности. Например, мы знаем, что примерно 95% площади под кривой нормального распределения будет находиться в диапазоне ±σ от среднего значения. Вероятность того, что выборочное значение попадет в этот диапазон, составляет 95%. Соответственно вероятность случайного отклонения выборочного среднего определяется площадью в «хвостах» под кривой нормального распределения – за пределами найденного диапазона. В контексте проверки статистических гипотез эта область в «хвостах» называется *критической* (на рисунке эта область заштрихована). Следовательно, вероятность того, что мы получим случайное отклонение выборочного среднего от истинного значения (при условии истинности нулевой гипотезы), составляет 1 – 0,95 = 0,05. Это вероятность того, что результат, на основе которого может быть отклонена нулевая гипотеза, мог быть получен случайно, когда на самом деле для генеральной совокупности нулевая гипотеза является истинной. Для данного примера вероятность ошибки первого рода – 0,05.



Таким образом, нулевую гипотезу можно отвергнуть в пользу альтернативной гипотезы, если по результатам статистического теста вероятность ошибки, т.е. вероятность случайного возникновения обнаруженного различия (*p-уровень*), не превышает 5 из 100 (имеется лишь 5 шансов из 100 ошибиться) для заданного уровня значимости в 5%. Если же этот уровень значимости не достигается (вероятность ошибки выше 5%), то считают, что разница вполне может быть случайной и поэтому нельзя отклонить нулевую гипотезу. Таким

образом, *p-уровень значимости* (*p-value*) соответствует риску совершения ошибки первого рода (отклонения истинной нулевой гипотезы). Если $p < \alpha$, нулевая гипотеза отклоняется.

Чем меньше значение *p*-уровня, тем выше статистическая значимость результата исследования. Уровень значимости выше (значение *p*-уровня меньше), если:

- величина связи или различия выборочных параметров выше;
- изменчивость признака меньше;
- объем выборки больше.

Соответствие некоторых значений α и значений *z* приведено в таблице, более подробно – в приложении 1.1.

α	***z***
0,10	±1,65
0,05	±1,96
0,01	±2,58
0,001	±3,29

Критическая область включает все маловероятные и редкие значения выборочных результатов. Получение при обработке данных значения из этой области позволяет отклонить нулевую гипотезу. Остальная область состоит из всех значений выборочных результатов, являющихся «нередкими». Чем меньше значение альфа, тем меньше критическая область и тем больше расстояние между средним значением распределения выборочных статистик и граничными точками критической области.

Вопрос о приемлемом значении альфа, т.е. вопрос о том, при каком уровне можно отклонить H_0, не имеет однозначного ответа. Чтобы свести к минимуму ошибку первого рода, следует использовать очень малые значения альфа. По мере уменьшения значения альфа критическая область становится меньшей и все больше удаляется от среднего значения распределения выборочных статистик, которому подчиняется выборка. Чем меньше значение альфа, тем меньше вероятность ошибки первого рода.

Однако по мере того, как уменьшается размер критической области (уменьшаются значения альфа), оставшаяся область (область между двумя критическими значениями) увеличивается. Следовательно, при увеличении значения альфа (например с 0,01 до 0,05) вероятность ошибки второго рода уменьшается. Для установленного значения альфа вероятность ошибки второго рода уменьшается с ростом объема выборки. Если все остальное остается неизменным, тогда чем меньше значение альфа, тем меньше вероятность того, что значение выборочного результата попадет в критическую область. Это повышает

вероятность принятия ошибки второго рода: принятие нулевой гипотезы, которая в действительности является ложной.

Итак, с уменьшением значения альфа снижается вероятность ошибки первого рода, но возрастает вероятность ошибки второго рода. Таким образом, обе ошибки обратно зависят друг от друга, и при одной проверке невозможно свести к минимуму обе ошибки. По мере уменьшения вероятности одной ошибки увеличивается вероятность другой, и наоборот. Таким образом, проверка гипотез всегда содержит элемент риска, поскольку невозможно одновременно свести к минимуму вероятность ошибки и первого, и второго рода.

В конечном итоге это означает, что необходимо подбирать такое значение альфа, которое было бы балансом между двумя ошибками. Более высокие значения альфа сведут к минимуму вероятность ошибки второго рода (признание ложного утверждения истинным), а более низкие значения альфа сведут к минимуму вероятность ошибки первого рода (признание истинного утверждения ложным). Как мы отмечали, более критичной считается ошибка первого рода, поэтому обычно применяются низкие значения альфа (0,05; 0,01; 0,001 и ниже).

На практике значение α устанавливается исходя из «научных конвенций» – соглашений, принятых в научном сообществе на основе практического опыта в различных областях исследования. В частности, значение 0,05 считается общепринятым показателем значимого результата. Традиционная интерпретация различных уровней значимости, исходящая из α = 0,05, приведена в таблице.

Уровень значимости	*Решение*	*Возможный статистический вывод*
$p > 0{,}1$	H_0 не может быть отклонена	«Статистически достоверные различия не обнаружены»
$p \leq 0{,}1$	Неопределенность в истинности H_0	«Различия обнаружены на уровне статистической тенденции»
$p \leq 0{,}05$	Отклонение H_0	«Обнаружены статистически достоверные (значимые) различия»
$p \leq 0{,}01$	Отклонение H_0	«Различия обнаружены на высоком уровне статистической значимости»

Значение α = 0,05 рекомендовано для небольших выборок (когда высока вероятность ошибки второго рода). Если объемы выборок $n \geq 100$, то порог отклонения H_0 лучше снизить до α = 0,01, т.е. принимать решение о наличии связи (различий) при $p \leq 0{,}01$ (причины этого рассматриваются ниже).

Однако распространенное использование значения 0,05 является лишь общепринятой условностью, и нет причин, по которым нельзя было бы дать альфа любое другое разумное значение (например 0,02, 0,03 или 0,07). Исследователь может выбирать и обосновывать такое значение альфа, которое выглядит наиболее разумным с точки зрения конкретного исследовательского проекта. С другой стороны, статистические пакеты (SPSS, STATISTICA и другие) вычисляют точное значение *p*-уровня (например $p = 0{,}023$), и это значение, в принципе, можно указывать в исследовательских работах. Но с точки зрения принятых правил представления результатов статистических исследований корректнее указывать, что *p*-уровень меньше одного из принятых значений (0,01, 0,05 или 0,1), например $p < 0{,}05$.

Если при проверке гипотезы формулируется ненаправленная альтернативная гипотеза, то критические области, соответствующие отклонению нулевой гипотезы и принятию альтернативной, поровну распределяются по обоим «хвостам». Поэтому такие кривые называют *двусторонними* (*2-tailed*). В этом случае если принят уровень статистической значимости α, то существуют два критических порога: α / 2 справа на кривой распределения и –α / 2 слева. Если проверяется направленная альтернатива, то говорят об *одностороннем* (*1-tailed*) критерии. В этом случае существует один критический порог α справа или –α слева в зависимости от направления альтернативы. При двусторонней альтернативе нулевая гипотеза отвергается при больших значениях силы связи (корреляции, различиях средних и т.д.), чем при односторонней, т.е. проверка ненаправленной гипотезы является более «строгой» при прочих равных условиях.

Существует простое соотношение между *p*-уровнями для направленных и ненаправленных альтернатив: для одного и того же эмпирического значения критерия *p*-уровень для направленной альтернативы в два раза меньше *p*-уровня для ненаправленной альтернативы. Критические значения *z* для односторонних и двусторонних критериев для некоторых распространенных значений альфа приведены в таблице.

Альфа	Двусторонний критерий	Односторонний критерий
0,10	±1,65	±1,29
0,05	±1,96	±1,65
0,01	±2,58	±2,33
0,001	±3,29	±3,10

Значения *z* для других уровней значимости можно определить из таблицы в приложении 1.1 или с помощью функции Excel НОРМСТОБР(*P*), возвращающей значение

для *одностороннего* критерия z при заданной вероятности $P = 1 - \alpha$. Другая функция Excel – НОРМСТРАСП(z) возвращает значение P, из которого легко определяется α.

Например, в двусторонней проверке с альфа, равной 0,05, критическая область начинается с критического стандартизованного значения $z = \pm 1{,}96$. В случае одностороннего критерия при том же значении альфа критическое значение z равно +1,65, если задан правый «хвост» распределения, и –1,65, если задан левый «хвост» распределения. Критические значения для односторонних критериев расположены ближе к среднему значению распределения выборочных статистик. Это еще раз показывает, что при одностороннем критерии повышается вероятность отклонения нулевой гипотезы без изменения значения альфа. Как мы уже говорили, односторонние критерии следует использовать, когда направление различий можно с уверенностью предсказать, исходя из неких априорных теоретических предпосылок, позволяющих *заранее* (до эксперимента) предположить направленное различие. Если существуют сомнения при выборе между направленной и ненаправленной альтернативой, то всегда следует выбирать ненаправленную альтернативную гипотезу.

Из подтверждения альтернативной гипотезы на уровне $p < \alpha$ нельзя сделать вывод, что вероятность аналогичного результата при повторном исследовании будет больше 1 – α. Из-за соотношения ошибок первого и второго рода о вероятности получения того или иного результата в повторных исследованиях сказать невозможно. Также на основании статистически значимой связи нельзя делать вывод о наличии причинно-следственной связи между явлениями (признаками). Для утверждения причинно-следственной связи необходимо обоснование двух дополнительных признаков: а) одно явление предшествует другому во времени, б) отсутствует общая причина изменения признаков (явлений).

5.4. Общая процедура проверки статистических гипотез

Для принятия решений о том, можно ли отклонить нулевую гипотезу и принять альтернативную, используют *статистические критерии*, которые включают в себя методы расчета определенного показателя, на основании которого принимается решение об отклонении нулевой гипотезы, а также правила (условия) принятия решения. Этот рассчитываемый показатель называется *эмпирическим (или экспериментальным) значением критерия.* Найденное эмпирическое значение сравнивается с известным (например, заданным таблично или определенным с помощью той или иной статистической программы) эталонным числом, именуемым *критическим значением критерия*. В статистических

таблицах критические значения приводятся, как правило, для нескольких уровней значимости: 5% (0,05), 1% (0,01) и др.

Статистический критерий зависит также от числа степеней свободы. *Число степеней свободы* (*degrees of freedom*) – это количество значений в распределении, которые свободны для изменения. Как правило, число степеней свободы линейно зависит от объема выборки (например $df = n - 1$), а также от числа признаков или их градаций – чем больше эти показатели, тем больше число степеней свободы. Не существует единой формулы для определения числа степеней свободы для всех возможных случаев, поэтому статистический критерий также устанавливает формулу для расчета числа степеней свободы. Все показатели, определяемые для нахождения эмпирического и критического значений критерия, называют *статистиками критерия.*

Для большинства статистических критериев действует следующее правило: если эмпирическое значение критерия для данного числа степеней свободы оказывается ниже критического уровня, соответствующего выбранному значению α (порогу вероятности), то нулевая гипотеза не может считаться опровергнутой. Это означает, что выявленная разница (или связь) недостоверна. Другими словами, если эмпирическое значение критерия оказывается меньше или равно критическому, то можно сделать вывод, что характеристики распределений совпадают на уровне значимости α. Если эмпирическое значение критерия, напротив, оказывается строго больше критического, то нулевая гипотеза отклоняется и принимается альтернативная гипотеза – характеристики распределений считаются различными с достоверностью различий $1 - \alpha$. Например, если $\alpha = 0{,}05$ и принята альтернативная гипотеза, то достоверность различий равна 0,95, или 95%. Чем больше эмпирическое значение критерия критического значения, тем сильнее различаются характеристики сравниваемых объектов. Для некоторых статистических критериев нулевая гипотеза отклоняется при ином соотношении эмпирического и критического значений (см. главу 8).

Таким образом, общая процедура проверки статистической гипотезы включает в себя следующие шаги.

1. Сформулировать статистическую модель (допущения относительно способа формирования выборки, используемой шкалы, вида распределения и т.д.).
2. Сформулировать нулевую и альтернативную гипотезы.
3. Выбрать соответствующий статистический критерий.
4. Выбрать требуемый уровень значимости (α = 0,1; 0,05; 0,01; 0,001 и т.д.).
5. Вычислить эмпирическое значение критерия.
6. Сравнить эмпирическое значение критерия с критическим значением.

7. Принять решение об отклонении или невозможности отклонения нулевой гипотезы.

5.5. Статистические методы и критерии проверки гипотез

Для того чтобы судить о том, какова вероятность ошибки при отклонении или не отклонении нулевой гипотезы, применяют статистические критерии, соответствующие задачам исследования и особенностям выборки (выборок). Для данных, полученных в метрических шкалах (интервальных или относительных) при распределениях, близких к нормальным, используют *параметрические критерии*, основанные на таких показателях, как среднее и стандартное отклонение.

В частности, для определения достоверности разницы средних для двух выборок применяют критерий Стьюдента, а для того чтобы судить о различиях между тремя или большим числом выборок, – дисперсионный анализ (ANOVA) или ANOVA с повторными измерениями. Корректное использование многих параметрических методов ограничивается определенными требованиями, накладываемыми на объем выборок, вид распределения, однородность дисперсий и др.

Если исследователь имеет дело с данными, полученными в неметрических (номинативных или порядковых) шкалах, или если выборки слишком малы для уверенного заключения о том, что генеральные совокупности, из которых они взяты, подчиняются нормальному распределению, используют *непараметрические критерии* («свободные от распределения») – χ^2 (хи-квадрат), Манна-Уитни, Уилкоксона и другие. Эти критерии довольно просты с точки зрения как расчетов, так и применения.

Непараметрические критерии менее чувствительны, чем параметрические, поэтому, если важно обнаружить даже слабые различия (например, оцениваются возможные негативные эффекты влияния тех или иных лекарственных средств на здоровье людей), следует особенно внимательно выбирать метод и уровень значимости. Преимуществом непараметрических критериев является то обстоятельство, что их можно использовать и при сравнительно небольших выборках, в отличие от многих параметрических критериев.

Выбор статистического критерия также зависит от того, являются ли выборки, средние которых сравниваются, независимыми (например взятыми из двух разных групп испытуемых) или зависимыми (например отражающими результаты одной и той же группы испытуемых до и после воздействия или после двух различных воздействий).

В зависимости от условий проверки возможно использование свыше 100 различных вариантов тестирования. Поэтому перед исследователем может возникнуть проблема выбора и корректного использования соответствующего статистического метода. Для демонстрации

многообразия статистических критериев и различий их использования в зависимости от количества выборок и типов используемых шкал в таблице приводится несколько основных критериев и методов.

	Одна выборка	**Две выборки**		**Больше 2-х выборок**	
Тип шкалы		*Зависимые*	*Независимые*	*Зависимые*	*Независимые*
Номинативная	Одновыборочный хи-квадрат	Тест Мак-Немара	Двухвыборочный хи-квадрат	*Q*-тест Кокрена	Многовыборочный хи-квадрат
Порядковая	Одновыборочный тест Колмогорова-Смирнова	Тест Уилкоксона; Критерий знаков	*U*-критерий Манна-Уитни; Двухвыборочный тест Колмогорова-Смирнова; Критерий серий Уальда-Волфовица	Критерий хи-квадрат Фридмана	Тест Краскела-Уоллиса
Метрическая	*t*-критерий Стьюдента; *z*-тест	*t*-тест для зависимых выборок	*t*-критерий для независимых выборок; *z*-тест	ANOVA с повторными измерениями	ANOVA

Приведенная таблица может помочь в выборе необходимого метода проверки в зависимости от имеющихся у исследователя данных, поставленной цели и задач исследования, но она не определяет этот выбор полностью. Дело в том, что даже в одной и той же формальной ситуации (тип данных, количество выборок, характер связи, вид распределения и т.д.) может быть применено несколько методов и критериев, каждый из которых имеет как достоинства и недостатки, так и особенности применения. Например, критерий Уилкоксона предполагает, что можно ранжировать различия между сравниваемыми наблюдениями. Если это не так, то следует использовать критерий знаков. Такие нюансы будут оговариваться при рассмотрении параметрических и непараметрических методов в следующих главах, а пока можно предложить общее правило: если исследование является практически важным и экономически затратным, то всегда целесообразно использовать различные методы, применение которых корректно в данной ситуации. Если результаты проверки различными методами будут существенно различаться, следует попытаться понять причины этих различий с точки зрения как содержания исследовательской ситуации, так и анализа примененных процедур (выполнены ли необходимые и достаточные предпосылки их использования, является ли данный критерий

достаточно чувствительным, соответствует ли выбранный уровень значимости данной задаче и т.д.).

5.6. Статистическая значимость и практическая важность

Во введении мы отмечали, что многие исследователи иногда не обращают должного внимания на проверку и обоснование статистической значимости исследовательских результатов. Но есть и другая крайность – исключительно формальная проверка полученных выводов. Корректно применив тот или иной статистический критерий проверки гипотез (с точки зрения выполнения всех предпосылок использования критерия) и подтвердив статистическую значимость искомой связи, различий и так далее, исследователь относится к такому формальному подтверждению как к некоему магическому знаку, который автоматически гарантирует качество результатов и выводов. Однако случается, что результаты, имеющие статистическую значимость, лишены содержательного смысла в контексте предмета исследования, т.е. они не имеют содержательной значимости. Это может быть вызвано многими причинами: плохо поставленными целями и задачами исследования; ошибками в сборе и обработке данных; неправильной интерпретацией результатов; революционностью получения результатов (выход за пределы той или иной научной парадигмы) и т.д. Но анализ таких вариантов не входит в задачи данной книги, поэтому мы рассмотрим лишь математические причины того, что результаты исследования могут быть статистически значимы, однако при этом не имеют содержательной значимости и практической важности.

Подтверждение статистически значимых различий осуществляется путем отклонения нулевой гипотезы. Мы знаем, что вероятность отклонения нулевой гипотезы связана с уровнем значимости, определяемым выбранным значением альфа. Чем выше значение альфа, тем больше критическая область и тем выше вероятность отклонения нулевой гипотезы. Например, при значении альфа 0,05 отклонить H_0 проще, чем при значении альфа 0,01, а при значении альфа 0,10 это сделать еще проще. С другой стороны, увеличение значения альфа ведет к повышению вероятности совершения ошибки первого рода, т.е. может оказаться, что мы сочтем несущественные различия статистически значимыми.

Также нам известно, что проверка ненаправленной гипотезы является более «строгой», чем направленной, поскольку в последнем случае уровень значимости фактически уменьшается ровно в два раза (значение альфа не распределяется пополам по «хвостам», а фиксирует критическую область с одной стороны). Поэтому неоправданное применение одностороннего критерия повышает вероятность отклонения нулевой гипотезы.

Еще одним фактором риска совершения ошибки первого рода является размер выборки: при увеличении объема выборки повышается вероятность отклонения нулевой гипотезы. Как будет видно из следующих глав, эмпирические значения критерия возрастают с увеличением количества элементов выборки, а чем сильнее отличается эмпирическое значение критерия от его критического значения, тем более значимыми полагаются различия. То есть на больших выборках совершить ошибку первого рода более вероятно. Сказанное, конечно, не означает, что исследователь не должен работать с большими выборками. Напротив, большая выборка может быть лучшим приближением генеральной совокупности, которую она представляет, поэтому к решениям, основанным на выборках большего размера, можно относиться с большим доверием, чем к решениям, основанным на выборках малого размера. Но по большим выборкам даже сравнительно небольшие различия, не представляющие никакого интереса для исследователя, могут быть статистически значимыми. Именно поэтому при больших выборках ($n \geq 100$) значение альфа целесообразно уменьшить до 0,01.

Таким образом, и в случае больших, и в случае малых выборок обнаружение статистически значимых различий не гарантирует их теоретическую или практическую важность. С одной стороны, различие, которое не является статистически значимым, не может считаться важным, поскольку, с точки зрения статистики, этого различия вообще не существует, точнее, оно имеет случайный характер. Но, с другой стороны, статистически значимое различие может не иметь никакой важности для исследования. Поэтому помимо статистической значимости необходимо всегда обосновывать содержательное наполнение обнаруженных различий и связей.

Глава 6. Параметрические методы проверки статистических гипотез

6.1. Критерий z

Самым простым и известным параметрическим методом проверки статистических гипотез является *z-тест* (*z-критерий*). Этот метод может использоваться при выполнении базовых предположений о нормальности распределения и представленности данных в метрической шкале.

Рассмотрим вначале *случай одной выборки*. Как мы уже знаем, эмпирическое значение *z*-критерия показывает, насколько выборочное среднее отличается от среднего по генеральной совокупности в единицах стандартного отклонения, и определяется по формуле

$$z = \frac{\bar{x} - m}{s_x}.$$

В случае формирования выборки методом случайного повторного отбора

$$z = \frac{\bar{x} - m}{s/\sqrt{n}}.$$

Вычислив эмпирическое значение z и сравнив его с критическим значением при заданном уровне значимости (приложение 1.1), можно сделать вывод о том, является ли случайным различие средних значений по выборке и генеральной совокупности. Пусть, например, получено эмпирическое значение $z = 2$. Тогда вероятность того, что данная выборка принадлежит генеральной совокупности со средним μ, составляет $p = 1 - 0{,}954 = 0{,}046$. Это значение соответствует p-уровню значимости, т.е. вероятности того, что данный выборочный результат мог быть получен случайно, когда на самом деле в генеральной совокупности верна H_0. Таким образом, при $\alpha = 0{,}05$ нулевая гипотеза отклоняется, поскольку $p < \alpha$.

Для иллюстрации этого подхода рассмотрим пример. Для случайной выборки, состоящей из 152 уголовных дел, рассматривавшихся в районном суде, средняя продолжительность срока лишения свободы составила 27,3 мес. ($s = 3{,}7$). Отличается ли это в значительной мере от среднего срока лишения свободы для лиц, совершивших уголовные преступления, по стране (28,7 мес.)?

Нулевая гипотеза (H_0: $\mu = 28{,}7$) утверждает, что выборка извлечена из генеральной совокупности, среднее значение которой равно 28,7. По имеющимся данным получаем

эмпирическое значение z = –4,67 (читателям предлагается проверить это самостоятельно). Выбираем значение альфа, равное 0,05. Тогда критическая область будет начинаться с критического значения z = ±1,96. Поскольку эмпирическое значение критерия равно –4,67, мы можем отклонить нулевую гипотезу. Различие между сроками лишения свободы для преступников, признанных виновными в данном суде, и преступников по стране в целом, статистически значимо. Различие настолько велико, что мы можем заключить, что оно не явилось следствием простой случайности. Вероятность того, что решение отклонить нулевую гипотезу окажется неверным, меньше 0,05 ($p < 0,05$).

Во многих случаях измерение представляющей интерес характеристики выборки проводится таким образом, что нельзя при этом подтвердить предположение о принадлежности полученных данных к интервальной или относительной шкале измерений. Один из выходов в данной ситуации заключается в том, чтобы использовать в качестве критерия не среднее значение выборки, а выборочную пропорцию. Общая процедура проверки гипотезы в этом случае идентична процедуре, используемой при проверке средних значений. В этом случае основным для исследователя по-прежнему является вопрос «имеет ли генеральная совокупность, из которой была извлечена выборка, данную характеристику?» Следовательно, мы по-прежнему выполняем проверку, основываясь на предположении, что нулевая гипотеза истинна. Если полученное эмпирическое значение z попадает в критическую область (что маловероятно, если предположить, что H_0 истинна), тогда нулевая гипотеза отклоняется.

Формула для расчета экспериментального значения z при сравнении пропорций:

$$z = \frac{P_s - P_p}{\sqrt{P_p(1-P_p)/n}},$$

где P_p – доля элементов генеральной совокупности, обладающих определенным свойством; P_s – доля элементов выборочной совокупности, не обладающих этим свойством.

Рассмотрим использование этого подхода на примере. В случайной повторной выборке из жителей самого неблагополучного района города 76% респондентов во время последних выборов голосовали за представителей партии «Партия». В целом по городу за эту партию проголосовало 66% электората. Можно ли сделать вывод, что среди жителей неблагополучного района значительно больше людей голосовало за «Партию»? Для проверки гипотезы имеем следующие данные: P_s = 0,76; P_p = 0,66; n = 103.

Нулевая гипотеза (H_0: P_p = 0,66) утверждает, что выборка была получена из генеральной совокупности, 66% которой проголосовало за «Партию». Эмпирическое значение z = 2,13 (читателям предлагается проверить это самостоятельно). В данном случае

рассматривается гипотеза, действительно ли среди жителей неблагополучного района значительно *больше* людей голосовало за «Партию», поэтому должна использоваться односторонняя проверка. При значении альфа, равном 0,05, критическая область начнется с критического значения, равного +1,65. Критическая область располагается в правом «хвосте», так как мы предполагаем, что выборка извлечена из генеральной совокупности, в которой за «Партию» проголосовало *больше* людей, чем в целом по городу.

Так как эмпирическое значение равно 2,13, следует отклонить нулевую гипотезу. Различие между жителями района и жителями города в целом является статистически значимым и имеет предсказанное направление.

Теперь рассмотрим возможности применения *z*-теста для *случая двух выборок*. В случае с одной выборкой предполагается, что она была получена на основе метода случайного отбора. Это подразумевает, что у всех наблюдений в генеральной совокупности должны быть равные шансы на то, чтобы попасть в выборку. В ситуации с двумя выборками они должны быть сформированы не только методом случайного отбора, но еще и независимо друг от друга. Это требование удовлетворяется в случаях, когда выбор конкретного объекта для одной выборки не оказывает влияния на вероятность того, что некоторый другой объект будет включен в другую выборку. Для того чтобы удовлетворить этому условию, можно составить одну случайную выборку из общего списка генеральной совокупности, а затем разделить ее элементы на отдельные группы. Если исходная выборка была получена методом случайного отбора, то все подвыборки будут удовлетворять предположению о независимых случайных выборках.

Нулевая гипотеза в случае с двумя выборками по-прежнему представляет собой утверждение об отсутствии различий. Однако теперь вместо того, чтобы проверять значимость различия между средними по выборке и генеральной совокупности, из которой эта выборка была извлечена, проверяется значимость различия между средними по двум выборкам. Нулевая гипотеза утверждает, что между двумя генеральными совокупностями, из которых извлечены эти две выборки, нет различий. Если статистика критерия попадает в критическую область, то нулевую гипотезу, утверждающую отсутствие различий между генеральными совокупностями, можно отклонить.

Если нулевая гипотеза истинна, то разность между средними значениями генеральных совокупностей должна быть близкой к нулю и различия между средними значениями выборок будут небольшими. Чем больше различия между средними значениями выборок, тем больше вероятность того, что действительно существует различие между генеральными совокупностями, представленными этими выборками.

Если выборки имеют сравнительно большой размер (суммарное количество наблюдений в двух выборках превышает 100), то распределение разностей средних значений выборок будет очень близко к нормальному и для установления критических областей можно использовать кривую нормального распределения. В случае двух выборок эмпирическое значение *z* определяется по формуле

$$z = \frac{\overline{x_1} - \overline{x_2}}{s_{x1-x2}} = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{s_1^2/(n_1-1) + s_2^2/(n_2-1)}},$$

где $\overline{x_1} - \overline{x_2}$ – разность средних значений выборок; s_{x1-x2} – стандартное отклонение распределения значений разностей средних значений выборок; s_1, s_2 – стандартные отклонения по первой и второй выборкам; n_1, n_2 – объемы первой и второй выборок.

В остальном, как мы уже отметили, процедура тестирования не отличается от случая одной выборки.

Проверка значимости различий между двумя пропорциями выборок аналогична проверке значимости различий между выборочными средними. Нулевая гипотеза утверждает, что между генеральными совокупностями, из которых были извлечены выборки, нет различий по проверяемой переменной. При больших размерах выборок (сумма двух выборок превышает 100) для нахождения границы критической области можно по-прежнему использовать z-распределение. В данном случае эмпирическое значение z определяется по формуле

$$z = \frac{P_{s1} - P_{s2}}{s_{s1-s2}} = \frac{P_{s1} - P_{s2}}{\sqrt{P_p(1-P_p)}\sqrt{(n_1+n_2)/(n_1 n_2)}},$$

где $P_p = \frac{n_1 P_{s1} + n_2 P_{s2}}{n_1 + n_2}$; $P_{s1} - P_{s2}$ – разность между пропорциями выборок; s_{s1-s2} – стандартное отклонение распределения разностей выборочных пропорций; n_1, n_2 – объемы выборок.

Дальнейшая процедура тестирования аналогична случаю с одной выборкой.

6.2. Критерий t Стьюдента для независимых выборок

В больших выборках (содержащих 100 и более наблюдений) выборочное стандартное отклонение дает адекватную оценку s_x. Для выборок небольшого объема (сумма элементов выборок меньше 100) и при неизвестном s_x *z*-критерий применять нельзя, так как распределение средних соответствует другому теоретическому распределению – *t*-

распределению Стьюдента. Форма *t*-распределения изменяется в зависимости от размера выборки. При маленьких размерах выборки кривая *t*-распределения является намного более пологой, чем кривая *z*-распределения, но при увеличении размера выборки *t*-распределение все больше напоминает *z*-распределение. При объеме выборки больше 120 эти распределения практически идентичны, поэтому для больших выборок *z*- и *t*-тесты обеспечивают почти одинаковые результаты. По этой причине в большинство статистических пакетов не включается процедура проверки гипотез на основе z-критерия, но там обязательно присутствует процедура *t*-тестирования.

Наиболее часто *t*-критерий Стьюдента применяется для обнаружения различий между средними *двух* выборок. Важным преимуществом *t*-критерия является возможность его использования на малых выборках (с 10 и даже меньшим количеством элементов). Для применения *t*-критерия необходимо определить число степеней свободы (*df*), что требуется для правильного определения критических значений *t*-критерия. В случае одной выборки $df = n - 1$, для двух выборок $df = n_1 + n_2 - 2$, где n_1, n_2 – объемы выборок. Критические значения для *t*-распределения при заданных уровне значимости и числе степеней свободы можно взять из соответствующих статистических таблиц (см. приложение 1.2) или рассчитать, например, с помощью функции Excel СТЬЮДРАСПОБР(α, *df*).

Общая логика проверки и даже формула расчета эмпирического значения *t*-критерия остаются аналогичными тем, что и при использовании *z*-теста. Для небольших выборок критическое значение *t* будет больше, чем критическое значение *z* при том же уровне значимости. Это связано с тем, что *t*-распределение является более пологим, чем *z*-распределение. При использовании *t*-распределения начальные точки критических областей будут находиться дальше от среднего значения выборочного распределения, поэтому нулевую гипотезу будет сложнее отклонить. Более того, чем меньше размер выборки (чем меньше степеней свободы), тем более высоким будет эмпирическое значение *t*, необходимое для отклонения нулевой гипотезы.

Рассмотрим использование *t*-теста в *случае одной выборки*. Эмпирическое значение *t* определяется по формуле (при повторном случайном отборе)

$$t = \frac{\bar{x} - m}{s / \sqrt{n}},$$

где $\bar{x}$ – среднее значение изучаемой характеристики, определенное по выборочной совокупности; μ – среднее значение по генеральной совокупности (предполагаемое и проверяемое в ходе тестирования); *s* – стандартное отклонение выборки; *n* – объем выборки.

Очевидно, что такой подход напрямую связан с процедурой определения доверительного интервала, которая была подробно рассмотрена выше. При решении вопроса

о возможности отклонения нулевой гипотезы на основе *t*-критерия его критические и экспериментальные значения сравниваются по абсолютному значению.

Обратимся к примеру о различиях в производительности труда. Известно, что производительность труда в среднем по заводу составляет 50 деталей/час. Необходимо установить, отличается ли производительность труда рабочих определенной возрастной группы от этого значения. Формулируем нулевую и альтернативную гипотезы:

H_0: $\mu = 50$ – нулевая гипотеза;

H_A: $\mu \neq 50$ – альтернативная гипотеза.

В ходе исследования получены следующие данные:

№	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*15*
x	61	80	40	71	39	20	40	80	80	60	80	29	51	43	47
№	*16*	*17*	*18*	*19*	*20*	*21*	*22*	*23*	*24*	*25*	*26*	*27*	*28*	*29*	*30*
x	40	41	37	41	60	40	60	31	40	65	61	40	59	60	80

$\bar{x} = 52{,}5$ – среднее по выборке,

$s = 17{,}1$ – стандартное отклонение по выборке,

$n = 30$ – объем выборки.

Подставляя эти значения в формулу *t*-критерия, получаем $t = 0{,}81$ – эмпирическое значение критерия. Для уровня значимости $\alpha = 0{,}05$ и числа степеней свободы $df = n - 1 = 29$ критическое значение критерия составляет 2,05. Поскольку эмпирическое значение меньше критического, то H_0 не может быть отклонена (проверить эти расчеты читателю предлагается самостоятельно).

Аналогичные результаты можно получить, воспользовавшись соответствующей процедурой пакета SPSS. Для этого необходимо в меню *Analyze* выбрать *Compare Means*, затем – *One-Sample T Test*. В поле *Test Value* устанавливается значение, с которым сравнивается выборочное среднее (в данном случае устанавливается значение 50), затем *OK*. В результате получаем базовые статистики по выборке (*One-Sample Statistics*) (среднее, стандартное отклонение и стандартную ошибку среднего) и статистику одновыборочного *t*-критерия (*One-Sample Test*). Мы видим, что эмпирическое значение критерия составляет 0,81, а *p*-уровень значимости при двунаправленной проверке – 0,425 (*df* = 29), что значительно выше $\alpha = 0{,}05$. Следовательно, H_0 не может быть отклонена.

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
SAMPLE1	30	52,5333	17,12980	3,12746

One-Sample Test

	Test Value = 50			
	t	df	Sig. (2-tailed)	Mean Difference
SAMPLE1	,810	29	,425	2,5333

При использовании *t*-критерия для случая *с двумя независимыми выборками* помимо предположения о нормальности распределения изучаемого признака в выборках вводится дополнительное предположение о равенстве (не слишком большом различии) дисперсий генеральных совокупностей, из которых формируются выборки. Для проверки нормальности распределения можно применить подходы, описанные в главе 3. Предположение о равенстве дисперсий генеральных совокупностей можно проверить с помощью критериев однородности дисперсий, которые доступны в SPSS и STATISTICA, например критерий Левена (*Levene test*) или критерий Брауна-Форсайта (*Brown-Forsythe test*). Однако если проверка этого предположения специально не проводится, можно считать, что имеет место равенство дисперсий генеральных совокупностей, если размеры выборок приблизительно одинаковы. В случае, если распределение признака хотя бы в одной из выборок заметно отличается от нормального и/или дисперсии статистически значимо различаются, альтернативой данному методу может быть непараметрический критерий Манна-Уитни (описан в главе 8).

Формула для расчета экспериментального значения *t*-критерия:

$$t = \frac{\left|\overline{x_1} - \overline{x_2}\right|}{s_{x1-x2}},$$

где $s_{x1-x2} = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}} \sqrt{\frac{n_1+n_2}{n_1 n_2}}$; s_1, s_2 – стандартные отклонения по первой и второй выборкам; n_1, n_2 – объемы выборок.

Эта формула применяется для точных расчетов, когда выборки заметно различаются по численности. Для приближенных расчетов в случае, если численность выборок примерно одинакова, можно воспользоваться более простой формулой, в которой

$$s_{x1-x2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}.$$

Чтобы проиллюстрировать эту процедуру, предположим, что проверяется гипотеза о том, что семьи, проживающие в центре города, имеют большее количество детей, чем семьи, живущие в пригородах. Были сформированы случайные выборки жителей обоих районов и вычислены статистические показатели выборок.

Выборка 1	**Выборка 2**
Пригороды	*Центр города*
$\overline{x_1}$ = 2,37	$\overline{x_2}$ = 2,78
s_1 = 0,63	s_2 = 0,95
n_1 = 42	n_2 = 37

Размер выборок невелик, но их объемы различаются незначительно, поэтому при формулировке статистической модели делаем предположение о равенстве дисперсий генеральных совокупностей. Также предполагаются независимость случайных выборок, измерение в метрической шкале, нормальное распределение выборочных статистик.

Нулевая гипотеза формулируется как H_0: $\mu_1 = \mu_2$. Поскольку заранее предположено направление различий (семьи, проживающие в центре города, больше по размеру), альтернативная гипотеза будет направленной – H_A: $\mu_1 < \mu_2$.

Выбираем $\alpha = 0,05$ и используем односторонний критерий. Число степеней свободы $df = 42 + 37 - 2 = 77$; t критическое равно 1,671. Критическая область отнесена в левый «хвост» распределения в соответствии с направлением, определенным альтернативной гипотезой. Эмпирическое значение t-критерия составляет 2,29 (читателям рекомендуется это проверить). Сравнивая эмпирическое и критическое значения, видим, что эмпирическое значение попадает в критическую область. Если бы нулевая гипотеза была истинной, такое значение выборочной статистики было бы очень маловероятным, поэтому мы можем отклонить нулевую гипотезу. Имеет место статистически значимое различие в размерах семей, проживающих в центре города, и семей, проживающих в пригородах. Более того, семьи, проживающие в центре города, значительно больше по количеству детей.

Рассмотрим еще один пример. Предположим, что менеджер крупной торговой компании намерен сравнить эффективность обучения персонала двумя разными методами (метод А и метод Б). Были отобраны 22 вновь поступивших сотрудника, которые случайным образом поделены на две группы. Одна группа прошла обучение по методу А, другая – по методу Б. В конце года менеджер компании получил данные об объемах продаж в каждой группе на одного сотрудника (в тыс. руб.):

	Группа А	*Группа Б*
Среднемесячные продажи	1500	1300
Стандартное отклонение	225	251

Следуя стандартной процедуре проверки статистических гипотез, проверяем, действительно ли метод А дает лучший результат по сравнению с методом Б.

H_0: объемы продаж за рассматриваемый период не различаются в двух группах (эффективность методов одинакова).

H_A: объемы продаж в первой группе больше, чем во второй (метод А обеспечивает лучшие продажи по сравнению с методом Б).

Используем *t*-критерий Стьюдента для проверки гипотезы, поскольку данные представлены в метрической шкале и выборки рассматриваются как независимые. Выбранный уровень значимости – 0,05, проверяется направленная альтернатива. Подстановка имеющихся данных в формулу для расчета экспериментального значения *t*-критерия дает значение t = 200 / 101,63 = 1,97 (читателям рекомендуется получить это значение самостоятельно). Критическое значение *t*-критерия – 1,725, при df = (11 – 1) + (11 – 1) = 20. Это значение можно определить с помощью функции Excel СТЬЮДРАСПОБР(0,1; 20). Обратите внимание на то, что выбран уровень значимости α = 0,05 для односторонней альтернативы, поэтому в функцию Excel в качестве первого параметра подставляем значение 0,1. Таким образом, мы можем отклонить нулевую гипотезу и принять, что метод А обеспечивает лучшую эффективность тренингов по сравнению с методом Б.

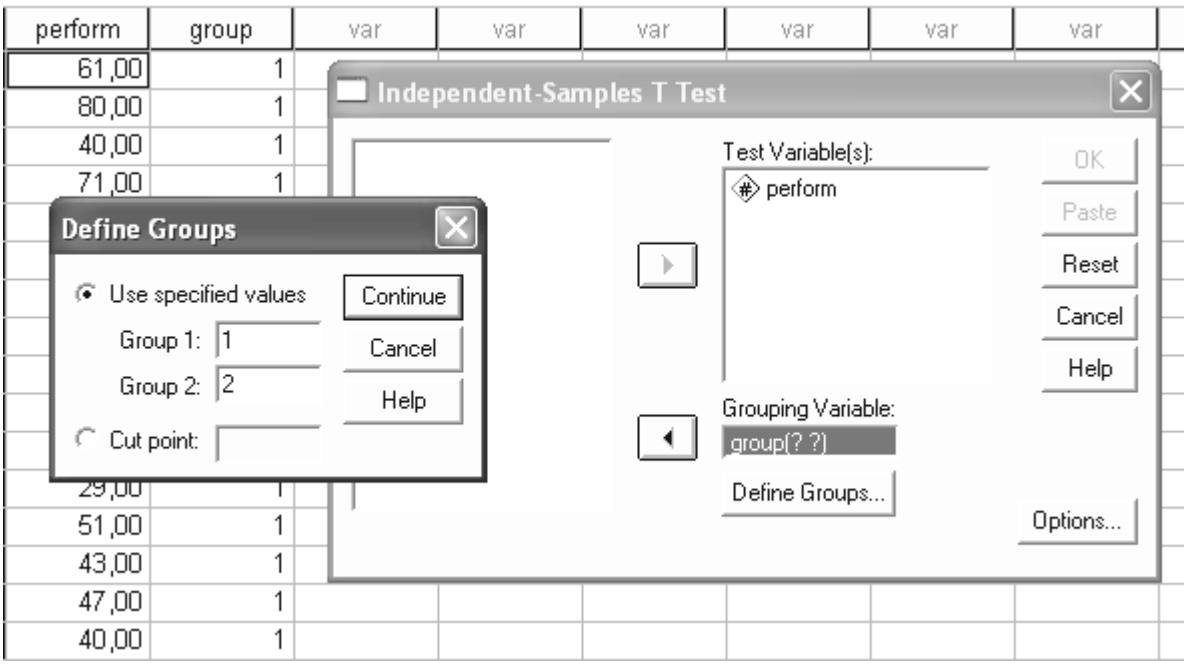
Рассмотрим, как может быть осуществлена процедура проверки статистической гипотезы на основе *t*-критерия (для двух независимых выборок) с помощью статистических пакетов. Допустим, нам известны результаты измерения производительности труда рабочих двух разных возрастных групп (Sample1 и Sample2).

№	*Sample1*	*Sample2*	*№*	*Sample1*	*Sample2*	*№*	*Sample1*	*Sample2*
1	59	61	11	56	80	21	58	40
2	56	80	12	37	29	22	71	60
3	85	40	13	56	51	23	31	31
4	92	71	14	42	43	24	42	40
5	48	39	15	49	47	25	34	65
6	59	20	16	70	40	26	51	61
7	42	40	17	30	41	27	70	40
8	87	80	18	58	37	28	42	59
9	42	80	19	79	41	29	67	60
10	73	60	20	84	60	30	56	80

Для обработки этих данных в SPSS необходимо сгруппировать их, а именно представить в виде двух переменных. Первая (зависимая) переменная соответствует

изучаемому признаку (в данном случае – производительность труда), а вторая (группирующая) – принадлежности к той или иной группе. Объединим результаты измерения производительности труда в одну переменную, содержащую 60 значений (назовем эту переменную, например, perform), в качестве второй переменной будем использовать столбец, первые 30 элементов которого будут иметь значение (код) 1, а оставшиеся 30 – значение 2 (дадим этой переменной имя group).

После ввода данных в меню *Analyze* нужно выбрать *Compare Means*, затем – *Independent Samples T Test*. В поле *Test Variable(s)* переносится переменная perform, а в поле *Grouping Variable* – переменная group. В окне *Define Groups* указывается кодировка группирующей переменной (в данном примере одна группа обозначена кодом 1, другая – 2).



Статистика критерия показывает, что, несмотря на, казалось бы, довольно большое различие в средних (52,5 и 57,5), это различие не является статистически значимым: p = 0,256 > α = 0,05. Результатом теста Левена является то, что гипотеза однородности дисперсий двух выборок не может быть отклонена (Sig. > 0,05), поэтому применение *t*-критерия вполне корректно. Незначительные различия выборочных стандартных отклонений приводят к практически неразличимым результатам при предположении равных дисперсий (*Equal variances assumed*) и отсутствии такого предположения (*Equal variances not assumed*).

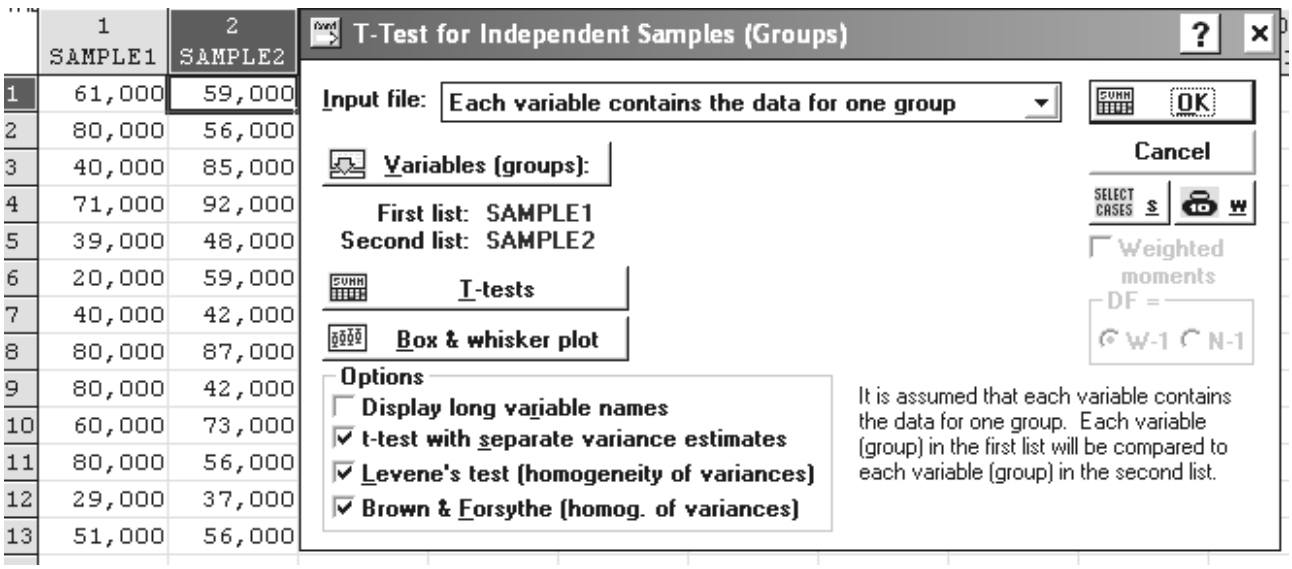
Group Statistics

	GROUP	N	Mean	Std. Deviation	Std. Error Mean
PERFORM	1	30	52,5333	17,12980	3,12746
	2	30	57,5333	17,27812	3,15454

Independent Samples Test

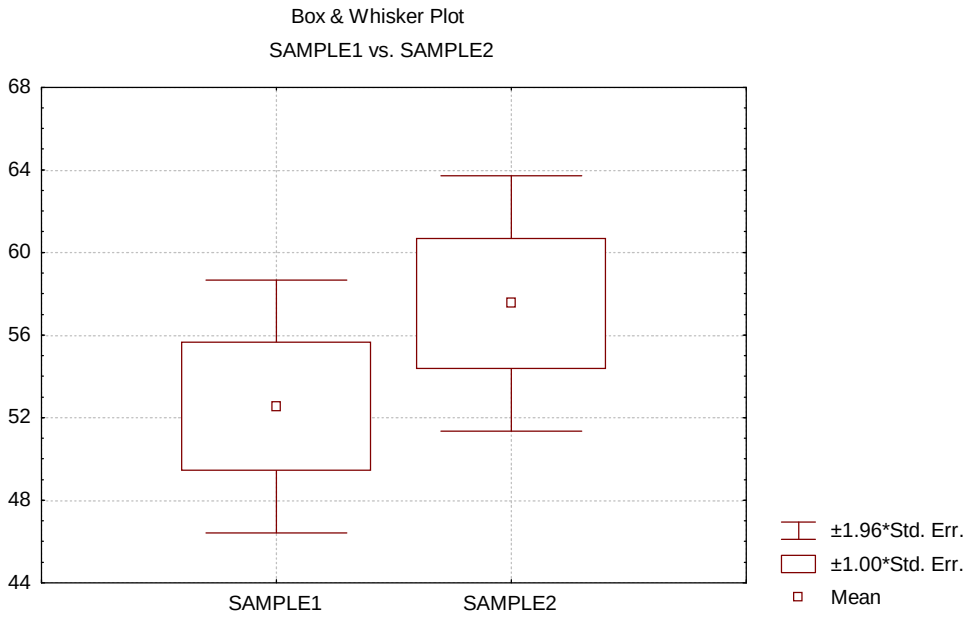
		Levene's Test for Equality of Variances		t-test for Equality of Means		
		F	Sig.	t	df	Sig. (2-tailed)
PERFORM	Equal variances assumed	,176	,676	-1,126	58	,265
	Equal variances not assumed			-1,126	57,996	,265

Идентичные результаты (но представленные в другой форме) могут быть получены в пакете STATISTICA. Для этого необходимо запустить модуль *Basic Statistics and Tables*, где в меню *Analysis* выбрать *t-test for independent samples*. В этом пакете данные могут быть представлены как в сгруппированном виде (*One record per case (use a grouping variable)*), подобно тому, как это было сделано для работы в SPSS, так и в исходном, когда каждая переменная содержит измерения по соответствующей выборке (*Each variable contains the data for one group*). В режиме *Variables (groups)* в левом и правом окнах выбирается по одной переменной для анализа. Дополнительно можно включить в выходную форму результаты тестов на однородность дисперсий выборок: тест Левена (*Levene's test*) и тест Брауна-Форсайта (*Brown & Forsythe*). Интерпретация результатов теста Брауна-Форсайта полностью аналогична интерпретации результатов теста Левена. Выходная форма с результатами этих тестов и статистикой *t*-критерия представляется после нажатия на кнопку *T-test*. Чтобы не загромождать текст, мы не приводим эту форму.



Полезной опцией для анализа результатов являются диаграммы размаха (*Box & Whisker plot*). С их помощью можно визуально оценить степень различия между двумя переменными путем сравнения средних и различных мер отклонения от среднего (стандартная ошибка

среднего, стандартное отклонение, размах и др.) в выборках. Можно, например, сравнить 95% доверительные интервалы по двум выборкам.



6.3. Критерий t Стьюдента для связанных выборок

Для связанных (зависимых) выборок подход к определению эмпирического значения *t*-критерия незначительно отличается от случая независимых выборок. Допущение зависимости выборок чаще всего означает, что признак измерен на одной и той же выборке дважды, например, до воздействия и после него (состояние пациентов до и после лечения; производительность труда рабочих до и после отпуска и т.п.). Применение *t*-критерия для зависимых выборок может также оказаться полезным при сравнении результатов предварительного и повторного тестирования для оценки надежности измерительного инструмента (метод «тест-ретест», см. главу 4).

Исходные данные должны быть таковы, чтобы имелось по два значения изучаемого признака для каждого объекта (или для каждой пары объектов). Очевидно, что измерения по обеим выборкам должны быть содержательно сопоставимы, т.е. должны использоваться аналогичные методы получения и обработки данных и одинаковые единицы измерения.

Для расчета *t*-критерия для зависимых выборок рассматриваются разности между двумя измерениями («до» и «после») для каждого элемента выборки (респондента). Таким образом, в формуле *t*-критерия для зависимых выборок основным элементом является разность значений для каждой из n пар объектов: $d_i = x_{1i} - x_{2i}$. Соответственно определяются среднее значение (M_d) и стандартное отклонение (s_d) этих разностей, после чего используется формула

$$t = \frac{|M_d|}{s_d / \sqrt{n}}, \qquad df = n - 1,$$

где $M_d = \frac{\sum d_i}{n}$,

$$s_d = \sqrt{\frac{\sum d_i^2 - \frac{(\sum d_i)^2}{n}}{n-1}},$$

Для иллюстрации этого подхода используем данные по объемам продаж (млн долл.) 10 американских компаний в 1988 и 1989 гг. из списка Forbes Super 500 (Cooper & Shindler, 1998).

Компания	*Продажи в 1989 г.*	*Продажи в 1988 г.*
General Motors	126932	123505
General Electric	54574	49662
Exxon	86656	78944
IBM	62710	59512
Ford	96146	92300
AT&T	36112	35173
Mobil	50220	48111
DuPont	35099	32427
Sears	53794	49975
Amoco	23966	20779

Есть ли статистически значимые различия в объемах продаж ведущих компаний за год? Для ответа на этот вопрос можно применить *t*-тест для двух зависимых выборок, поскольку имеются результаты повторных измерений для каждой компании, а данные представлены в относительной шкале.

Для анализа используем соответствующую процедуру SPSS (*Analyze*, *Compare Means*, *Paired Samples T Test*). В данном случае группировать переменные не нужно, достаточно указать 2 переменные, по которым проводится сравнение, в поле *Paired Variables*.

В итоге мы получаем, что средние объемы продаж (62620,9 млн долл. и 59038,8 млн долл.) статистически значимо различаются ($p < 0{,}001$). Поэтому нулевая гипотеза отклоняется, что позволяет сделать вывод о росте торгового оборота ведущих американских компаний за анализируемый период.

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	S1989	62620,9000	10	31777,64941	10048,97508
	S1988	59038,8000	10	31072,87052	9826,10443

Paired Samples Test

		Paired Differences			t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean			
Pair 1	S1989 S1988	3582,1000	1803,1592	570,2090	6,282	9	,000

Процедура вычисления статистики *t*-критерия для зависимых выборок в STATISTICA (*t-test for dependent samples*) аналогична процедуре вычисления статистик этого критерия для независимых выборок. В результате выполнения этой процедуры мы получаем такие же эмпирические значения *t*-критерия.

T-test for Dependent Samples (new-2.sta)

BASIC STATS — Marked differences are significant at p < ,05000

Variable	Mean	Std.Dv.	N	Diff.	Std.Dv. Diff.	t	df	p
SAMPLE1	62620,90	31777,65						
SAMPLE2	59038,80	31072,87	10	3582,100	1803,159	6,282082	9	,000144

Предположения и ограничения на использование *t*-критерия для зависимых выборок аналогичны статистической модели *t*-критерия для независимых выборок. Поскольку предполагается, что распределение исследуемого признака существенно не отличается от нормального, то в данном случае это означает, что попарные разности тоже должны быть нормально распределены. Также дополнительно предполагается, что данные двух измерений положительно коррелируют (это свидетельствует о зависимости выборок).

Если эти условия не выполняются, то можно воспользоваться одним из альтернативных непараметрических критериев, например *T*-критерием Уилкоксона. Если данные для двух выборок не коррелируют положительно, можно использовать критерий Стьюдента для независимых выборок.

Закономерен вопрос о том, как будут различаться результаты проверки при замене *t*-критерия для зависимых выборок соответствующим критерием для независимых выборок. В случае зависимых выборочных совокупностей допустимо применение критерия Стьюдента для независимых выборок, но не наоборот. Причем это рекомендуется осуществлять, если

корреляция между двумя измерениями отрицательна. Если корреляция положительна, то такая замена приведет к существенной недооценке достоверности различий. Это связано с тем, что в критерии для зависимых выборок парное вычитание значений по двум выборкам и последующий анализ этих разностей позволяют исключить ту часть вариации, которая является результатом различия в исходных уровнях респондентов. Поэтому по сравнению с *t*-критерием для независимых выборок критерий Стьюдента для зависимых выборок является гораздо более чувствительным. Действительно, если в нашем примере к двум выборкам, характеризующим объемы продаж, применить *t*-критерий для независимых выборок, нулевая гипотеза не может быть отклонена (p = 0,802), т.е. статистически значимые различия не обнаруживаются. Читателям рекомендуется самостоятельно проверить этот результат. Этот пример еще раз иллюстрирует необходимость корректного использования статистических методов.

Глава 7. Дисперсионный анализ

Основные идеи дисперсионного анализа (***Analysis of variance – ANOVA***) были разработаны и предложены в 1920 г. Р. Фишером (R. Fisher), английским ученым, основоположником математической генетики, внесшим значительный вклад в разработку теории математической статистики и ее приложений для анализа экспериментальных исследований. В определенном смысле ANOVA может рассматриваться как развитие *t*-критерия Стьюдента. Но в отличие от критерия Стьюдента, который применим, когда независимая переменная имеет только две категории (вспомните процедуру группирования данных для обработки в SPSS), дисперсионный анализ применим для независимых переменных, имеющих более чем две категории. Фактически в случае сравнения двух групп дисперсионный анализ дает результаты, идентичные *t*-критерию. Но при сложных исследовательских планах следует использовать дисперсионный анализ, а не последовательность *t*-тестов.

Дисперсионный анализ применяется, когда зависимая переменная измеряется в метрической шкале. Независимые переменные (факторы) могут измеряться в неметрических шкалах и представляться в нескольких градациях (уровнях).

В зависимости от количества факторов, влияние которых на зависимую переменную анализируется, выделяют однофакторный дисперсионный анализ (*one-way ANOVA*), двухфакторный дисперсионный анализ (*two-way ANOVA*), трехфакторный дисперсионный анализ (*three-way ANOVA*) и т.д. Многофакторный дисперсионный анализ позволяет проверить гипотезы не только о влиянии каждого фактора по отдельности, но и об их взаимодействии, когда влияние одного фактора на зависимую переменную проявляется по-разному при разных уровнях других факторов. То есть многофакторный дисперсионный анализ позволяет проследить, насколько влияние одного фактора зависит от градаций других факторов. Это свойство делает многофакторный метод уникальным и часто незаменимым.

Процедура использования дисперсионного анализа значительно сложнее, чем применение рассмотренных выше методов, в силу того, что ANOVA предполагает проведение большого количества довольно громоздких вычислений. Большинство известных статистических пакетов (SPSS, SAS, STATISTICA, MICROCASE, NCSS, STATA и др.) включают в себя процедуры ANOVA. Но работа с этими процедурами и, главное, правильная интерпретация результатов требуют ясного понимания того, как этот метод работает и что означают основные показатели, получаемые в результате его использования. Поэтому мы будем приводить много довольно простых примеров и решений для иллюстрации

«внутренних» механизмов работы метода, а затем сравнивать эти результаты с теми, которые получены на основе применения статистических пакетов.

7.1. Однофакторный дисперсионный анализ

Метод ANOVA основан на соотнесении межгрупповой и внутригрупповой изменчивости изучаемой зависимой переменной, т.е. на сравнении величины изменчивости между категориями с величиной изменчивости внутри категорий. Чем больше различия между категориями по отношению к различиям внутри категорий, тем больше вероятность того, что нулевая гипотеза об *отсутствии различий* является ложной и может быть отклонена.

Рассмотрим простой пример, иллюстрирующий этот принцип. Предположим, необходимо выяснить, зависит ли время выполнения определенного задания (операции) работниками предприятия от возможных методов выполнения этой операции. Случайным образом были сформированы 4 группы работников, в каждой из которых демонстрировался один из четырех методов.

Допустим, что получены следующие результаты (в минутах).

	Метод 1	*Метод 2*	*Метод 3*	*Метод 4*
Среднее значение	11,2	10,1	10,5	9,5
Ст. отклонение	2,7	1,5	2,3	1,7

Средние значения и стандартные отклонения по разным методам отличаются незначительно. Вероятность нулевой гипотезы об отсутствии различий между методами достаточно велика. А теперь предположим, что мы получили другие результаты.

	Метод 1	*Метод 2*	*Метод 3*	*Метод 4*
Среднее значение	16,2	12,1	9,5	6,5
Ст. отклонение	1,1	0,9	0,8	0,6

В этом случае средние значения заметно различаются, а стандартные отклонения невелики и сравнительно одинаковы, что указывает на небольшую изменчивость внутри каждой категории. Такие результаты с большой вероятностью позволяют отклонить нулевую

гипотезу и указывают на действительные различия эффективности методов выполнения задания.

Нулевая гипотеза, проверяемая методом ANOVA, утверждает равенство *всех* средних значений (для *всех* изучаемых генеральных совокупностей). Для приведенного выше примера нулевая гипотеза формулируется в виде H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$. При отклонении нулевой гипотезы принимается альтернативная гипотеза о том, что по крайней мере два средних значения различаются (среднее по крайней мере для одной генеральной совокупности отличается от остальных). Часто альтернативная гипотеза формулируется как H_A: $\mu_1 \neq \mu_2 \neq \mu_3 \neq \mu_4$. Пусть читателя не смущает такая формулировка, так как она не означает, что *все* средние значимо отличаются друг от друга. Дело в том, что если нулевая гипотеза отклоняется, то без дополнительных методов все равно невозможно определить, какое именно среднее (или средние) значимо отличается от остальных. Для этого используются критерии множественных сравнений, рассматриваемые в следующем параграфе.

Изучаемый признак (зависимая переменная) измеряется у объектов, каждый из которых принадлежит к *одной* из нескольких сравниваемых выборок. Для принятия решения на основе дисперсионного анализа необходимо вычислить следующие величины.

Общая сумма квадратов, или SS_T (*total sum of squares*):

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

Общая сумма квадратов складывается из двух компонентов: $SS_T = SS_B + SS_W$. Первый компонент характеризует случайную изменчивость внутри категорий и называется внутригрупповой суммой квадратов SS_W (*sum of squares within groups*). Другой компонент характеризует изменчивость между категориями и называется межгрупповой суммой квадратов SS_B (*sum of squares between groups*).

Внутригрупповая сумма квадратов определяется как

$$SS_W = \sum_{j=1}^{k} \sum_{i=1}^{n_k} (x_i - \overline{x_k})^2,$$

где $\overline{x_k}$ – среднее значение k-й группы (категории); n_k – количество наблюдений в k-й категории.

Межгрупповая сумма квадратов определяется как

$$SS_B = \sum_{j=1}^{k} n_k (\overline{x_k} - \overline{x})^2.$$

Отношение межгрупповой и общей сумм квадратов показывает долю общей дисперсии зависимой переменной, обусловленную влиянием исследуемого фактора. Этот показатель имеет тот же смысл, что и *коэффициент детерминации*, т.е. квадрат коэффициента корреляции $r^2 = \frac{SS_B}{SS_T}$.

Если нулевая гипотеза верна (различий между категориями нет), то изменчивость от категории к категории не должна быть большой по сравнению с изменчивостью внутри категорий, т.е. SS_W и SS_B должны быть примерно равны. Чем больше различие между этими двумя оценками, тем с большей вероятностью нулевая гипотеза должна быть отклонена. Если же средние значения в группах приблизительно одинаковы, то тогда различия не будут значимыми. Чем больше различие между средними значениями категорий и чем более однородны категории, тем больше вероятность, что различия являются статистически значимыми.

Критерием принятия решения в методе дисперсионного анализа является *F-отношение* (*критерий Фишера*):

$$F = \frac{MS_B}{MS_W}, \qquad df_B = k - 1, \qquad df_W = n - k,$$

где MS_B – межгрупповой средний квадрат (*mean square between*):

$$MS_B = \frac{SS_B}{df_B};$$

величина MS_W – внутригрупповой средний квадрат (*mean square within*):

$$MS_W = \frac{SS_W}{df_W}.$$

Величины df_B и df_W являются числом степеней свободы для соответствующих сумм квадратов (SS_B и SS_W) и рассчитываются следующим образом:

число степеней свободы для межгрупповой суммы квадратов $df_B = k - 1$;

число степеней свободы для внутригрупповой суммы квадратов $df_W = n - k$.

F-отношение характеризует отношение величины изменчивости между категориями к величине изменчивости внутри категорий. Чем больше изменчивость между категориями по сравнению с изменчивостью внутри категорий, тем больше будет значение критерия Фишера и тем больше вероятность того, что нулевая гипотеза будет отклонена.

Для определения критических значений *F*-распределения (при соответствующих уровне значимости и числах степеней свободы df_B и df_W) можно воспользоваться

специальными статистическими таблицами (приложение 1.3) или рассчитать их с помощью функции FРАСПОБР(α, df_B, df_W). Если проверка гипотезы на основе ANOVA осуществляется без программных продуктов, реализующих соответствующую процедуру, то для упрощения расчетов можно использовать другую, более удобную для вычислений формулу для общей суммы квадратов:

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

После определения SS_B значение SS_W находится из формулы $SS_W = SS_T - SS_B$.

Проиллюстрируем всю процедуру на примере о том, зависит ли время выполнения производственной операции от применяемого метода. Для простоты возьмем выборку из 16 чел. (по 4 чел. в каждой группе). Использование метода дисперсионного анализа для такой группы, как будет показано ниже, не совсем корректно, но в данном случае это сделано исключительно для иллюстрации работы метода, чтобы не утруждать читателя повторением излишне громоздких расчетов. Предположим, что получены следующие результаты по выполнению задания.

№	*Метод 1*	*Метод 2*	*Метод 3*	*Метод 4*
1	10	25	14	16
2	13	19	16	17
3	15	20	16	23
4	16	17	24	15

Проверим гипотезу о том, что время на выполнение задания не зависит от метода (примем уровень значимости α = 0,05).

H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$ (средние значения по методам статистически значимо не различаются).

H_A: $\mu_1 \neq \mu_2 \neq \mu_3 \neq \mu_4$ (среднее значение по крайней мере для одной выборки статистически значимо отличается от остальных).

Сделаем необходимые вычисления для определения *F*-отношения.

	x_i				x_i^2			
№	Метод 1	Метод 2	Метод 3	Метод 4	Метод 1	Метод 2	Метод 3	Метод 4
1	10	25	14	16	100	625	196	256
2	13	19	16	17	169	361	256	289
3	15	20	16	23	225	400	256	529
4	16	17	24	15	256	289	576	225
Сумма	54	81	70	71	750	1675	1284	1299
$\bar{x}$	17,25							
$\bar{x}_k$	13,5	20,25	17,5	17,75				

$SS_T = (750 + 1675 + 1284 + 1299) – 16(17,25)^2 = 247$

$SS_B = 4(13,5 – 17,25)^2 + 4(20,25 – 17,25)^2 + 4(17,5 – 17,25)^2 + 4(17,75 – 17,25)^2 = 93,5$

$SS_W = 247 – 93,5 = 153,5$

$df_W = 16 – 4 = 12$

$df_B = 4 – 1 = 3$

$MS_B = 93,5 / 3 = 31,17$

$MS_W = 153,5 / 12 = 12,79$

$F\text{эксп} = 31,17 / 12,79 = 2,44$

$F\text{крит} = $ FРАСПОБР(0,05;3;12) = 3,49

Сравнение *Fэксп* и *Fкрит* (*Fэксп* < *Fкрит*) не позволяет отклонить нулевую гипотезу, поэтому необходимо сделать вывод о том, что наблюдаемые различия выборочных средних не являются статистически значимыми и, следовательно, при применении разных методов производительность труда в данном случае статистически значимо не различается.

Статистической моделью при проверке гипотез методом ANOVA являются предположения о том, что выборки случайные и независимые, зависимая переменная измеряется в интервальной шкале или шкале отношений, распределение зависимой переменной в сравниваемых генеральных совокупностях характеризуется нормальным законом и равными дисперсиями. (Последнее предположение называется предположением об *однородности дисперсий*.) Формально численность выборок не должна быть меньше 2 объектов, но рекомендуется иметь не менее 5 объектов в каждой выборке.

Это достаточно жесткие предположения. В частности, требование метрического характера шкалы, как мы обсуждали в первой и второй главах, не всегда можно обеспечить в социально-экономических исследованиях, поэтому данный метод следует использовать только для таких зависимых переменных, которые были тщательно и точно измерены.

В теории статистики показано, что *F*-критерий довольно устойчив к небольшим отклонениям от статистической модели, в частности к отклонению от нормальности. Если число наблюдений достаточно большое, то отклонение от нормальности не имеет особого значения в силу *центральной предельной теоремы*, в соответствии с которой распределение среднего значения при большом объеме выборки близко к нормальному независимо от начального распределения.

Предположение об однородности дисперсий связано с тем, что вычисление суммы квадратов ошибок производится внутри каждой группы. Если дисперсии в двух группах значительно отличаются друг от друга, то сложение их некорректно и не дает верной оценки общей внутригрупповой дисперсии (так как в этом случае общей дисперсии вообще не существует). Доказано, что *F*-критерий также достаточно устойчив относительно нарушения предположений однородности дисперсий. Поэтому считается, что пока объемы выборок равны или приблизительно равны между собой, метод дисперсионного анализа можно использовать без дополнительной проверки однородности дисперсий выборок, иначе необходимо проверять однородность дисперсий в сравниваемых выборках. Для этого применяется тест Левена (*Levene test of homogeneity of variances*), который реализован в SPSS, STATISTICA и других статистических пакетах.

Несмотря на то, что все названные ограничения на применение метода дисперсионного анализа не являются критическими и на практике этот метод используется при небольших отклонениях от указанной статистической модели, следует помнить, что такие ограничения существуют. И это обстоятельство нужно всегда принимать во внимание как при планировании и обработке результатов собственных исследований, так и при вынесении суждений о корректности результатов, полученных другими исследователями. Если статистическая модель нарушается в значительной степени или исследователь не уверен в том, что отклонения от модели незначительны, то рекомендуется применить непараметрический аналог этого метода (критерий хи-квадрат или *H*-критерий Краскела-Уоллиса).

7.2. Критерии множественных сравнений

Альтернативная гипотеза, формулируемая при применении метода ANOVA и принимаемая в случае отклонения нулевой гипотезы, неконкретна. Она лишь утверждает, что среднее по крайней мере для одной генеральной совокупности отличается от остальных. Очевидно, что исследователю важно знать, по каким именно категориям различия значимы, какие из групп значимо отличаются от других. Принятие решения только

по величине средних может привести к неверным выводам. Поэтому для корректного заключения о том, какие средние значимо отличаются от остальных, необходимо использовать дополнительные статистические критерии.

Казалось бы, что для решения этой задачи можно провести серию *t*-тестов Стьюдента и путем попарного сравнения выборок определить значимые различия для всех возможных пар средних. Однако использование *t*-критерия Стьюдента в данном случае не совсем корректно и обычно не рекомендуется. Дело в том, что *t*-критерий проверяет, являются ли два средних значимо отличающимися друг от друга, в предположении, что рассматриваются всего две выборки. Поэтому многократное применение критерия Стьюдента существенно повышает вероятность ошибки первого рода, так как в связи с большим числом парных сравнений такая процедура случайно увеличивает шансы получения значимого результата.

Разработаны процедуры для более корректного сравнения пар средних с целью выявления статистически значимых различий. Эти процедуры называются *критериями множественных сравнений*. Некоторые из этих критериев могут применяться *после* использования ANOVA и отклонения нулевой гипотезы, поэтому они называются *апостериорными критериями* (*post hoc tests* – от лат. «после того, как»). Другие критерии могут применяться и *до* использования ANOVA, поэтому они именуются *априорными критериями* (*a priori tests*).

Пример апостериорных процедур – *тест Шеффе* (*Scheffe's test*). Этот метод должен использоваться в случае, если нулевая гипотеза уже отклонена на основе применения ANOVA. Оценка значимости различий основана на вычислении *t-критерия Шеффе*, являющегося модификацией *t*-критерия Стьюдента:

$$t_s = \frac{\left|\overline{x_1} - \overline{x_2}\right|}{\sqrt{\frac{MS_W}{n_1} + \frac{MS_W}{n_2}}},$$

где $\overline{x_1}, \overline{x_2}$ – сравниваемые средние значения для двух анализируемых выборок; n_1, n_2 – объемы соответствующих выборок; MS_W – внутригрупповой средний квадрат.

Критическое значение *t*-критерия Шеффе определяется как $t_s^{'} = \sqrt{F'(k-1)}$, где *F'* – критическое значение *F*-критерия, определяемое для заданного уровня значимости α и соответствующих чисел степеней свободы df_B и df_W. Тест Шеффе имеет те же допущения по использованию, что и ANOVA.

Апостериорные критерии особенно полезны в случаях, когда дисперсионный анализ обнаружил различия, которые являются неожиданными для исследователя, поэтому

необходимо определить, по каким именно факторам эти различия наблюдаются. Апостериорные критерии позволяют надежно идентифицировать значимые различия между парами выборочных средних.

Если исследователь заранее (до осуществления процедуры проверки методом ANOVA) решает, что необходимо проводить множественную проверку выборок, то можно воспользоваться методом *априорных контрастов* (*a priori contrasts*).

Контрастом называется сумма

$$C = \sum_{j=1}^{k} c_j \overline{x_j},$$

где $\overline{x_j}$ – средние значения исследуемой зависимой переменной по группам (категориям изменения фактора); k – количество групп (категорий изменения фактора); c_j – коэффициенты контраста, такие, что $\sum_{j=1}^{k} c_j = 0$.

Если, например, фактор имеет три градации, то коэффициентами контраста могут быть: $c_1 = 2$, $c_2 = -1$, $c_3 = -1$ или $c_1 = 1$, $c_2 = -0{,}5$, $c_3 = -0{,}5$. В случае игнорирования какой-либо категории соответствующему коэффициенту присваивается нулевое значение. Несмотря на возможность выбора различных сочетаний значений коэффициентов контраста, эмпирическое значение критерия, на основе которого проверяется гипотеза, не изменяется.

Существуют две модификации этого метода. Одна из них (реализованная в SPSS) основана на использовании статистики *t*-критерия Стьюдента. Эмпирическое значение *t*-критерия определяется по формуле

$$t_c = \frac{C}{\sqrt{MS_W \sum_{j=1}^{k} \frac{c_j^2}{n_j}}}.$$

При изменении (ином выборе) коэффициентов контраста величина t_c изменяться не будет. Найденное значение сравнивается с критическим значением двустороннего *t*-распределения Стьюдента для $df = df_W = n - k$.

Другая модификация метода контрастов (реализованная в STATISTICA) основана на использовании статистики *F*-критерия Фишера. Эмпирическое значение *F*-критерия вычисляется по формуле

$$F_c = \frac{C^2}{MS_W \sum_{j=1}^{k} \frac{c_j^2}{n_j}}$$

и сравнивается с критическим значением F-критерия Фишера для $df_B = k - 1$ *и* $df_W = n - k$.

Проиллюстрируем использование метода контраста на рассмотренном выше примере (проверка различий времени выполнения задания 4 различными методами). Проверим статистическую значимость различия времени выполнения задания методом 2 от результатов, полученных другими методами (это имеет смысл сделать, поскольку среднее значение по второй выборке в наибольшей степени отличается от средних значений по другим выборкам).

Выберем, например, следующие коэффициенты контраста: 0, 1, –1, 0 (сравниваются второй и четвертый методы). Значение контраста в этом случае составит 2,5. Величина MS_W была определена в предыдущем параграфе. Подстановка всех величин в формулу дает нам эмпирическое значение t-критерия: 0,99. Критическое значение t = 2,18 (α = 0,05 и df = 16 – 4 = 12). Сравнение эмпирического и критического значений (0,99 < 2,18) не позволяет отклонить нулевую гипотезу, поэтому приходим к выводу о том, что среднее по второй выборке не отличается статистически значимо от среднего по четвертой выборке (различие в эффективности методов не обнаружено). Вычисление эмпирического значения F-критерия Фишера дает величину 0,98. Критическое значение F-критерия: 3,49 (α = 0,05, df_B = 3, df_W = 12). Таким образом, нулевая гипотеза также не может быть отклонена.

А теперь сравним средние значения первой и второй выборок (между средними этих выборок максимальное различие). Определим коэффициенты контраста, например, как: –1, 1, 0, 0. Значение контраста в этом случае составит 6,75. Эмпирическое значение t-критерия: 2,67; критическое значение t = 2,18; эмпирическое значение F-критерия: 7,12; критическое значение F-критерия: 3,49.

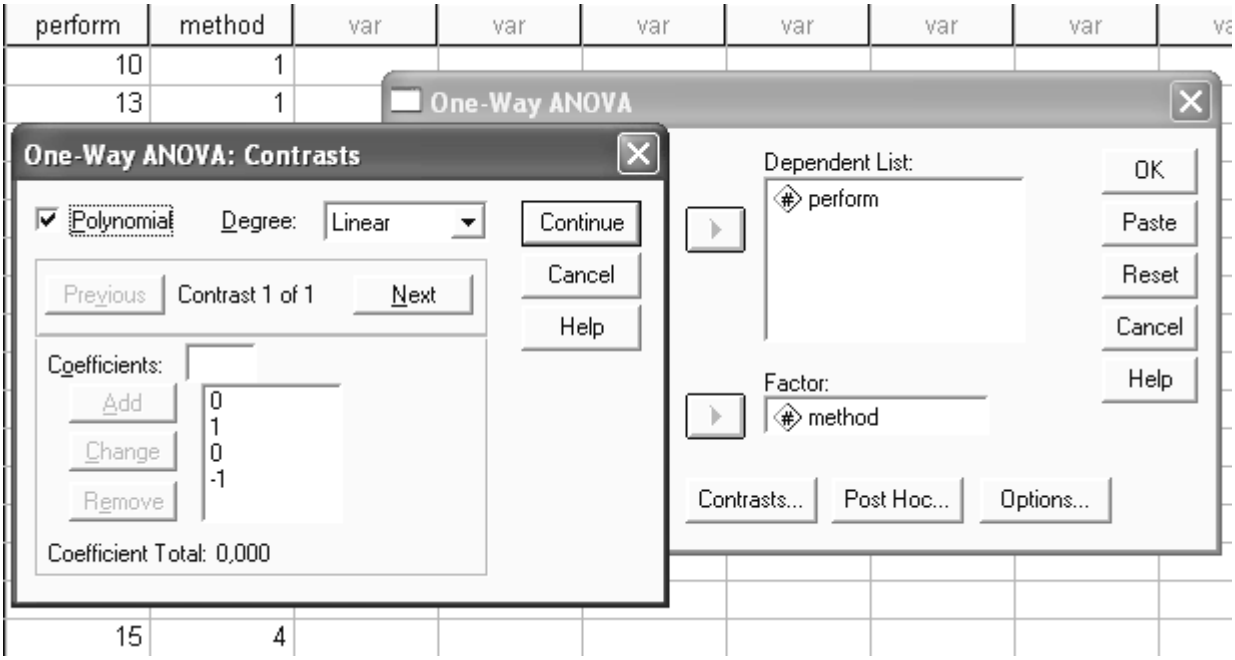
Этот пример демонстрирует одну важную особенность применения априорных методов множественного сравнения. Мы видим, что в данном случае эмпирические значения критериев по двум вариантам метода контрастов превышают критические значения. Казалось бы, можно отклонить нулевую гипотезу и сделать вывод о том, что среднее по второй выборке статистически значимо отличается от среднего по первой выборке (эффективность соответствующих методов различается). Но такой вывод будет не совсем корректным. Очень важно помнить, что мы используем априорный метод, который дает нам информацию о возможных различиях еще до применения метода дисперсионного анализа. И если в результате проведенного дисперсионного анализа нулевая гипотеза о равенстве всех средних отклоняется, т.е. подтверждается статистическая значимость различий хотя бы по одной паре, то тогда мы можем рассматривать информацию, полученную на основе метода контрастов, как достоверную. Но если в результате дисперсионного анализа нулевая гипотеза не может быть отклонена, различия,

обнаруженные методом контрастов, не могут рассматриваться как статистически значимые. Поэтому целесообразнее использовать апостериорные критерии множественных сравнений, которые являются более консервативными и надежными. Несмотря на такой нюанс, метод априорных контрастов полезен тем, что его применение увеличивает вероятность обнаружения различий, если они действительно существуют.

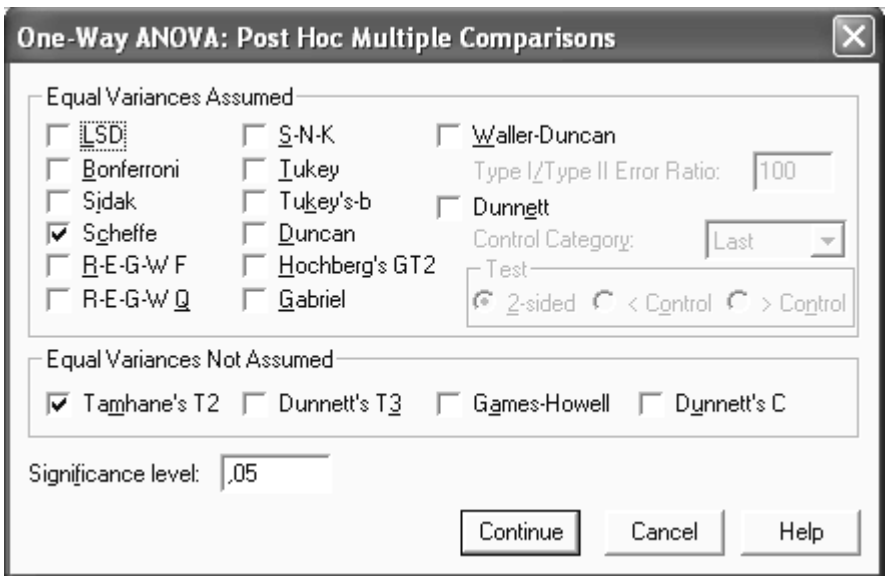
7.3. Реализация метода однофакторного дисперсионного анализа в статистических пакетах

Рассмотрим процедуру использования методов однофакторного дисперсионного анализа и множественных сравнений в SPSS и STATISTICA. Для сравнения результатов проведем обработку данных уже рассмотренного примера. Предварительно необходимо сгруппировать данные подобно тому, как мы это делали при использовании *t*-критерия для независимых выборок. Формируется столбец из 16 элементов, соответствующий зависимой переменной (время выполнения операции), и столбец группирующей переменной (фактор), принимающей четыре значения (например, 1, 2, 3 и 4), которые соответствуют применяемому методу. Обозначим зависимую переменную как perform (от performance – выполнение), а группирующую (факторную) переменную – как method (метод). После подготовки массива данных в меню *Analyze* выбирается *Compare Means*, затем *One Way Anova*. В окно *Dependent Lists* переносим зависимую переменную (perform), в поле *Factor* – группирующую переменную (method).

В окне *Contrasts* указывается, что для вычисления контраста будет использоваться линейный полином (выбрать опции *Polynomial Coefficients* и *Linear*). Затем, используя кнопку *Add*, нужно ввести коэффициенты контраста. Если предполагается вычисление нескольких контрастов, то, завершив ввод всех коэффициентов одного контраста, нужно нажать кнопку *Next* и ввести коэффициенты следующего контраста и т.д. (в данном примере в учебных целях рассчитываются два контраста для сравнения результатов).

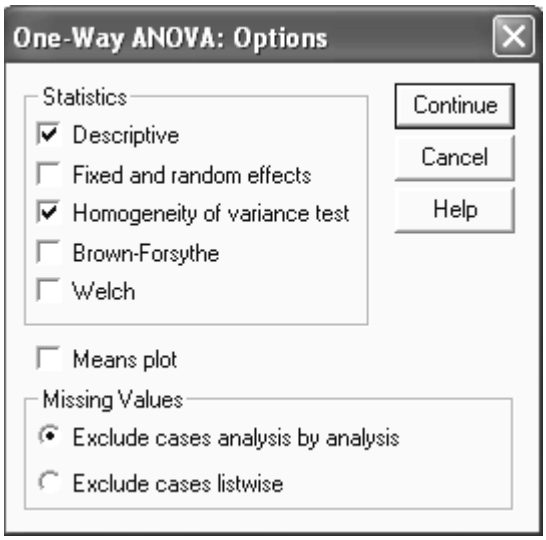


Для выбора апостериорных критериев множественного сравнения нужно нажать на кнопку *Post Hoc* в окне *One-Way Anova*. И SPSS, и STATISTICA предлагают широкий выбор апостериорных критериев, позволяющих определить различия между изучаемыми группами. Наиболее часто используемыми являются тест Шеффе (*Scheffe*) при предположении равенства дисперсий и T2-тест Тэмхейна (*Tamhane's T2*) при предположении о неравенстве дисперсий выборок. В поле *Significance level* указывается уровень значимости. Переход в окно *One-Way Anova* осуществляется нажатием кнопки *Continue*.



Для выбора параметров, которые будут выведены в выходную форму, необходимо нажать *Options* в окне *One-Way Anova*. В появившемся окне выбрать *Descriptive Statistics* для вывода базовых статистических показателей и *Homogeneity of variance tests* для вывода

результатов теста на однородность дисперсий. При необходимости можно выбрать и другие тесты (Брауна-Форсайта и Уэлша).



Выходная форма генерируется после нажатия на кнопку *OK*. Результаты обработки данных приведены ниже. В отчете ***Descriptives*** для каждой из 4 выборок указаны их объемы (*N*), средние значения (*Mean*), стандартные отклонения (*Std. Deviation*), стандартная ошибка среднего (*Std. Error*), нижняя (*Lower Bound*) и верхняя (*Upper Bound*) границы 95% доверительного интервала, минимальное (*Minimum*) и максимальное (*Maximum*) значения элементов выборки.

В отчете *Test of Homogeneity of Variances* представлены результаты проверки равенства дисперсий критерием Левена. На основе статистики критерия получен *p*-уровень (*Sig.*). Это значение много выше уровня значимости $\alpha = 0{,}05$, поэтому нулевая гипотеза не может быть отклонена. Следовательно, предположение о равенстве дисперсий выполняется и использование метода дисперсионного анализа корректно (при условии, что все другие предпосылки для его применения также выполнены).

Descriptives

PERFORM

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
1	4	13,50	2,646	1,323	9,29	17,71	10	16
2	4	20,25	3,403	1,702	14,83	25,67	17	25
3	4	17,50	4,435	2,217	10,44	24,56	14	24
4	4	17,75	3,594	1,797	12,03	23,47	15	23
Total	16	17,25	4,058	1,014	15,09	19,41	10	25

Test of Homogeneity of Variances

PERFORM

Levene Statistic	df1	df2	Sig.
,294	3	12	,829

В отчете ANOVA приведены результаты вычисления *F*-критерия (сравните значения, представленные в таблице, с результатами, полученными ранее при расчете «вручную»). Значение *p*-уровня (0,115) превышает α = 0,05, поэтому нулевая гипотеза не может быть отклонена. Следовательно, наблюдаемые различия выборочных средних не являются статистически значимыми.

В отчете *Contrast Coefficients* указаны введенные коэффициенты двух контрастов. В отчете *Contrast Tests* приведены уже знакомые нам значения статистик *t*-критерия: 0,342 для второй и четвертой выборок и 0,020 для первой и второй (мы проводили расчеты при предположении о равенстве дисперсий, а в отчете даны также значения статистик при предположении неравенства дисперсий выборок).

ANOVA

PERFORM

			Sum of Squares	df	Mean Square	F	Sig.
Between Groups	(Combined)		93,500	3	31,167	2,436	,115
	Linear Term	Contrast	20,000	1	20,000	1,564	,235
		Deviation	73,500	2	36,750	2,873	,096
Within Groups			153,500	12	12,792		
Total			247,000	15			

Contrast Coefficients

	METHOD			
Contrast	1	2	3	4
1	0	1	0	-1
2	-1	1	0	0

Contrast Tests

		Contrast	Value of Contrast	Std. Error	t	df	Sig. (2-tailed)
PERFORM	Assume equal variances	1	2,50	2,529	,989	12	,342
		2	6,75	2,529	2,669	12	,020
	Does not assume equal variances	1	2,50	2,475	1,010	5,982	,352
		2	6,75	2,155	3,132	5,656	,022

В отчете *Multiple Comparisons* приведены результаты парных сравнений средних значений всех выборок на основе двух апостериорных критериев множественных сравнений

– Шеффе (*Scheffe*) и Тэмхейна (*Tamhane*). Здесь представлены статистики этих тестов при сравнении *i*-й выборки с *j*-й выборкой (различие средних, стандартная ошибка различия, *p*-уровень значимости, нижняя и верхняя границы доверительного интервала). Наиболее информативен столбец, представляющий *p*-уровень значимости (*Sig.*). Как следует из отчета, по всем парным сравнениям это значение превышает $\alpha = 0{,}05$, поэтому нулевая гипотеза о различии средних в парах выборок отклоняется, что согласуется с полученным ранее выводом о том, что наблюдаемые различия выборочных средних не являются статистически значимыми.

Multiple Comparisons

Dependent Variable: PERFORM

Scheffe

(I) METHOD	(J) METHOD	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
1	2	-6,75	2,529	,121	-14,93	1,43
	3	-4,00	2,529	,501	-12,18	4,18
	4	-4,25	2,529	,451	-12,43	3,93
2	1	6,75	2,529	,121	-1,43	14,93
	3	2,75	2,529	,760	-5,43	10,93
	4	2,50	2,529	,807	-5,68	10,68
3	1	4,00	2,529	,501	-4,18	12,18
	2	-2,75	2,529	,760	-10,93	5,43
	4	-,25	2,529	1,000	-8,43	7,93
4	1	4,25	2,529	,451	-3,93	12,43
	2	-2,50	2,529	,807	-10,68	5,68
	3	,25	2,529	1,000	-7,93	8,43

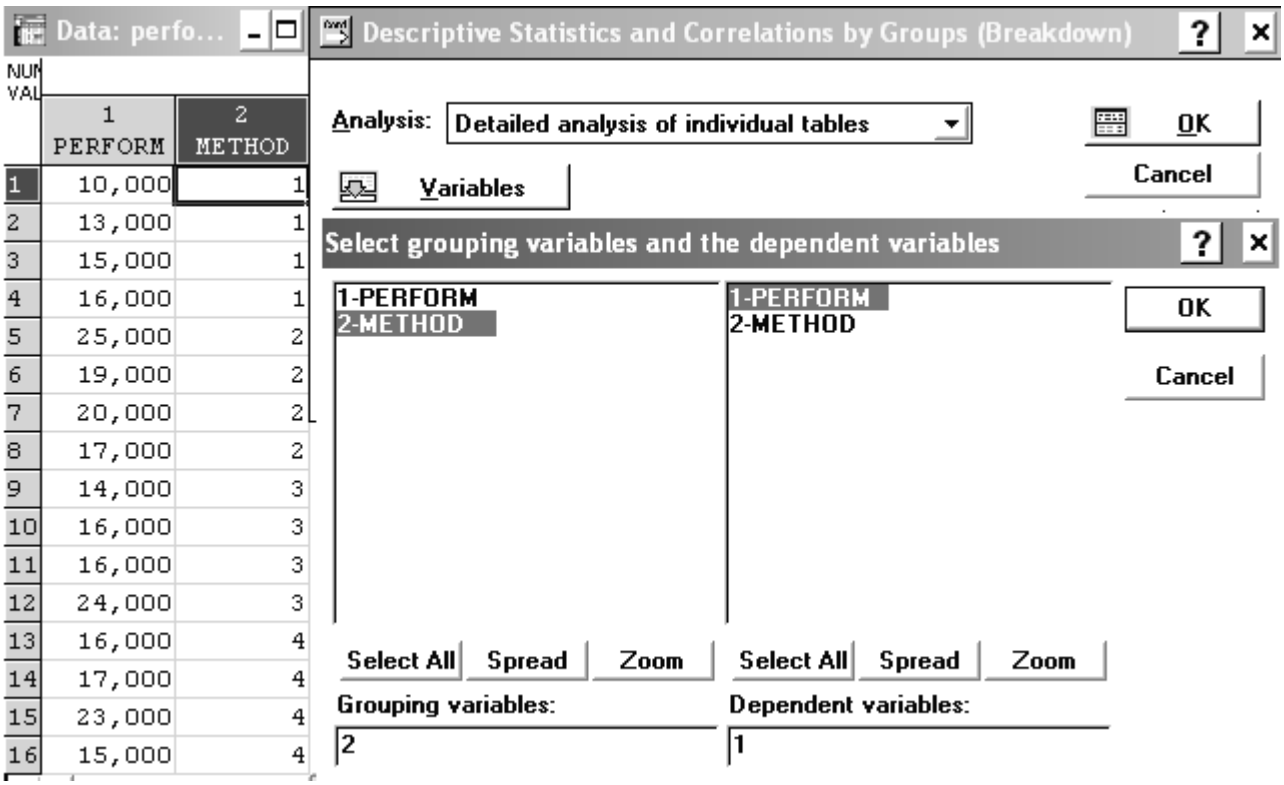
Multiple Comparisons

Dependent Variable: PERFORM

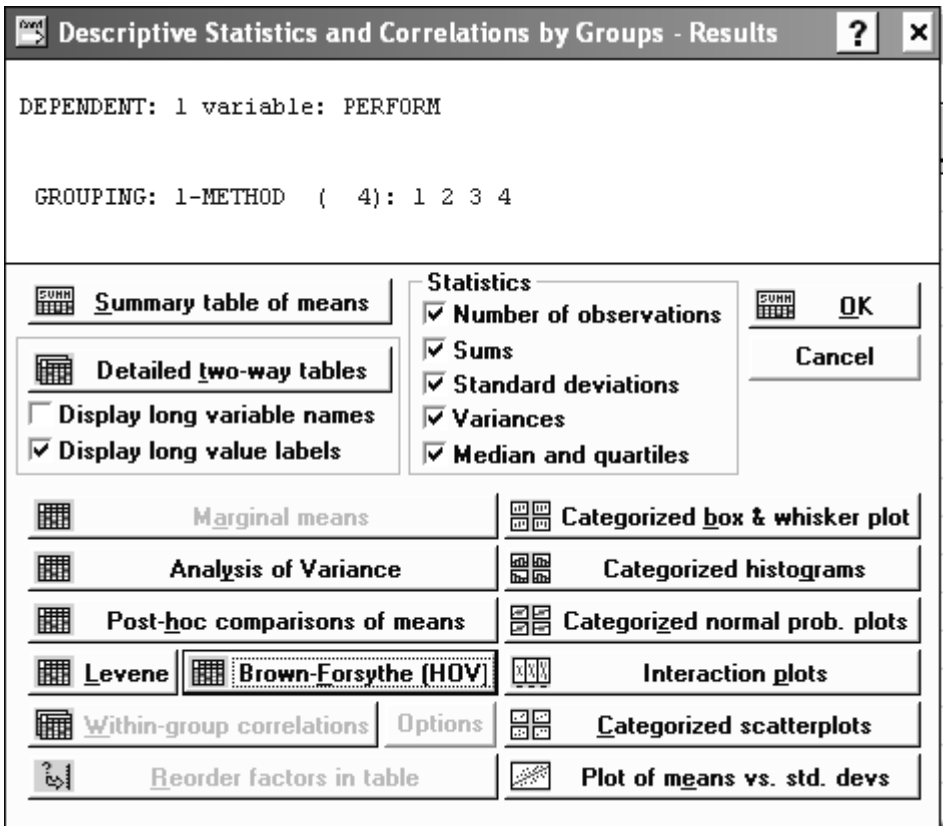
Tamhane

(I) METHOD	(J) METHOD	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
1	2	-6,75	2,155	,125	-15,26	1,76
	3	-4,00	2,582	,703	-14,96	6,96
	4	-4,25	2,231	,502	-13,16	4,66
2	1	6,75	2,155	,125	-1,76	15,26
	3	2,75	2,795	,935	-8,31	13,81
	4	2,50	2,475	,926	-7,03	12,03
3	1	4,00	2,582	,703	-6,96	14,96
	2	-2,75	2,795	,935	-13,81	8,31
	4	-,25	2,854	1,000	-11,43	10,93
4	1	4,25	2,231	,502	-4,66	13,16
	2	-2,50	2,475	,926	-12,03	7,03
	3	,25	2,854	1,000	-10,93	11,43

В среде STATISTICA задача решается аналогично. В модуле *Basic Statistics* выбирается режим *Breakdown & One-Way ANOVA*, затем *Variables*. Здесь выбираются группирующая и зависимая переменные.



Процедура определения переменных завершается нажатием на кнопку *OK* в окнах *Select grouping variables and the dependent variables* и *Descriptive Statistics and Correlation by Groups (Breakdown).*



В появившемся диалоговом окне *Descriptive Statistics and Correlation by Groups – Results* выбираются результаты, которые будут представлены в соответствующих выходных формах. Могут быть рассчитаны все базовые статистики (*Summary table of means* и *Statistics*), статистики для расчета *F*-критерия (*Analysis of Variance*), апостериорные критерии множественных сравнений (*Post-hoc comparisons of means*), критерии Левена и Брауна-Форсайта (*Levene* и *Brown-Forsythe*). В правой нижней части окна можно выбрать различные инструменты для визуального анализа результатов (например диаграмму размаха и др.). Мы не приводим результаты расчетов, поскольку они полностью эквивалентны результатам, полученным ранее в среде SPSS.

Более широкий набор аналитических и графических инструментов для проведения дисперсионного анализа, в том числе метод контрастов (*Planned Comparisons*), доступен в модуле ANOVA/MANOVA пакета STATISTICA. Последовательность этапов осуществления анализа в этом модуле в целом аналогична процедуре *Breakdown & One-Way ANOVA.*

Рассмотрим еще одну (более сложную и интересную) ситуацию. Крупное маркетинговое агентство получило заказ на оценку качества сервиса трех ведущих авиакомпаний страны, совершающих международные перелеты. Случайным образом было отобрано по 20 пассажиров каждой авиакомпании (10 пассажиров экономического класса и 10 пассажиров бизнес-класса), которые по определенной методике оценивали качество сервиса соответствующей компании (Рейтинг 1). Через месяц после исследования всем

опрошенным ранее пассажирам был доставлен аналогичный опросный лист для повторной оценки, чтобы проследить изменение результатов во времени (Рейтинг 2). Результаты опросов приведены в таблице. (Категория «класс» означает класс полета: 1 – экономический, 2 – бизнес.)

Нужно определить, различаются ли статистически значимо рейтинги авиакомпаний (по результатам первой оценки). Принимается, что оценки представлены в интервальной шкале; уровень значимости – 0,05. Оценка различия сервисов авиакомпаний осуществляется без учета класса полета.

Респон-дент	Авиаком пания	Класс	Рейтинг 1	Рейтинг 2		Респон-дент	Авиаком пания	Класс	Рейтинг 1	Рейтинг 2
1	1	1	40	36		31	2	2	52	65
2	1	1	28	28		32	2	2	70	80
3	1	1	36	30		33	2	2	73	79
4	1	1	32	28		34	2	2	72	88
5	1	1	60	40		35	2	2	73	89
6	1	1	12	14		36	2	2	71	72
7	1	1	32	26		37	2	2	55	58
8	1	1	36	30		38	2	2	68	67
9	1	1	44	38		39	2	2	81	85
10	1	1	36	35		40	2	2	78	80
11	1	2	40	42		41	3	1	92	95
12	1	2	68	49		42	3	1	56	60
13	1	2	20	24		43	3	1	64	70
14	1	2	33	35		44	3	1	72	78
15	1	2	65	40		45	3	1	48	65
16	1	2	40	36		46	3	1	52	70
17	1	2	51	29		47	3	1	64	79
18	1	2	25	24		48	3	1	68	81
19	1	2	37	23		49	3	1	76	69
20	1	2	44	41		50	3	1	56	78
21	2	1	56	67		51	3	2	88	92
22	2	1	48	58		52	3	2	79	85
23	2	1	64	78		53	3	2	92	94
24	2	1	56	68		54	3	2	88	93
25	2	1	28	69		55	3	2	73	90
26	2	1	32	74		56	3	2	68	67
27	2	1	42	55		57	3	2	81	85
28	2	1	40	55		58	3	2	95	95
29	2	1	61	80		59	3	2	68	67
30	2	1	58	78		60	3	2	78	83

Проведем решение этой задачи в среде SPSS. Результаты анализа показывают, что средние значения оценок для трех авиакомпаний (38,95; 58,90; 72,90) различаются статистически значимо. Полученное значение *F*-критерия (28,3) значительно превышает

критическое ($p < 0{,}001$). Тест Левена не позволяет отклонить гипотезу о равенстве дисперсий ($p = 0{,}747 > \alpha = 0{,}05$).

Descriptives

RATING1

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
1	20	38,95	14,006	3,132	32,40	45,50	12	68
2	20	58,90	15,089	3,374	51,84	65,96	28	81
3	20	72,90	13,902	3,108	66,39	79,41	48	95
Total	60	56,92	19,902	2,569	51,78	62,06	12	95

ANOVA

RATING1

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	11644,033	2	5822,017	28,304	,000
Within Groups	11724,550	57	205,694		
Total	23368,583	59			

Test of Homogeneity of Variances

RATING1

Levene Statistic	df1	df2	Sig.
,293	2	57	,747

Таким образом, нулевая гипотеза об отсутствии различий может быть отклонена. Поэтому можно сделать вывод: гипотеза о том, что различия между тремя агентствами вызваны только случайными отклонениями, очень маловероятна. Рейтинг авиакомпаний статистически значимо различается.

В данном случае у нас не было *теоретических* причин ***a priori*** (заранее) предполагать различия в уровне сервиса, предоставляемого авиакомпаниями. Поэтому используем апостериорный критерий множественных сравнений Шеффе для определения, по каким именно авиакомпаниям существуют различия. По результатам теста Шеффе мы видим, что все попарные комбинации (по всем компаниям) значимо различаются между собой.

Multiple Comparisons

Dependent Variable: RATING1

Scheffe

(I) AIRLINE	(J) AIRLINE	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
1	2	-19,95*	4,535	,000	-31,35	-8,55
	3	-33,95*	4,535	,000	-45,35	-22,55
2	1	19,95*	4,535	,000	8,55	31,35
	3	-14,00*	4,535	,012	-25,40	-2,60
3	1	33,95*	4,535	,000	22,55	45,35
	2	14,00*	4,535	,012	2,60	25,40

*. The mean difference is significant at the .05 level.

7.4. Многофакторный дисперсионный анализ

Многофакторный дисперсионный анализ, как уже отмечалось, позволяет оценить не только влияние на исследуемую (зависимую) переменную каждого фактора по отдельности (так называемые *главные эффекты*), но и зависимость влияния одних факторов от уровня других (*взаимодействие* факторов). Это преимущество ANOVA иногда дает возможность получать интересные исследовательские результаты, которых невозможно достичь другими методами.

Рассмотрим случай двух факторов, для чего вернемся к задаче об авиакомпаниях. Если в модель задачи добавить еще один фактор, например класс полета, мы получим двухфакторную модель для дисперсионного анализа. Данная модель позволяет ответить на три вопроса:

1. Есть ли статистически значимое различие в уровнях сервиса, предоставляемого тремя авиакомпаниями?
2. Есть ли статистически значимое различие в уровнях сервиса, предоставляемого при полете экономическим и бизнес-классом?
3. Влияют ли *совместно* вид авиакомпании и класс полета на оценку сервиса?

Общая логика проведения двухфакторного дисперсионного анализа принципиально не отличается от логики однофакторного анализа, за исключением того, что межгрупповая изменчивость включает в себя также изменчивость взаимодействия факторов:

$$SS_B = SS_a + SS_b + SS_{ab},$$

где SS_a, SS_b – суммы квадратов для факторов a и b; SS_{ab} – сумма квадратов для взаимодействия факторов.

Для каждого источника изменчивости вычисляются соответствующие квадраты и степени свободы, на основе чего определяется эмпирическое значение *F*-отношения, которое сравнивается с критическим значением.

На аналогичном принципе построен дисперсионный анализ, в котором количество факторов превышает два. Отличие заключается лишь в том, что помимо двухфакторных взаимодействий также анализируются взаимодействия больших порядков. Например, в случае трех факторов (*a*, *b*, *c*) вычисляются SS_a, SS_b, SS_c, SS_{ab}, SS_{bc}, SS_{ac}, а также сумма квадратов для трехфакторного взаимодействия SS_{abc}. Мы не будем приводить формулы для расчета этих сумм из-за их громоздкости и нецелесообразности использования без применения статистических пакетов.

Программные продукты, реализующие многофакторный метод ANOVA, сняли техническую проблему громоздкости вычислений, но не проблему содержательной интерпретации результатов в случае многофакторного анализа. Дело в том, что количество формулируемых и проверяемых гипотез (*h*) растет с увеличением числа факторов (*k*) в геометрической прогрессии: $h = 2^k - 1$. То есть в двухфакторной модели проверяются три гипотезы, в трехфакторной – семь и т.д. Именно по причине сложности формирования исследовательской модели и содержательной интерпретации результатов дисперсионный анализ с количеством факторов больше двух используется редко.

Для иллюстрации процедуры использования двухфакторного дисперсионного анализа и интерпретации результатов рассмотрим двухфакторную модель задачи об авиакомпаниях. Для этого воспользуемся возможностями, предоставляемыми SPSS. В главном меню *Analyze* выбирается *General Linear Model*, затем *Univariate*. Как и в случае однофакторного анализа, определяется зависимая переменная (*Dependent Variable*). В данном случае это переменная «Рейтинг 1» (rating1). В качестве анализируемых факторов (*Fixed Factor(s)*) выбираются переменные «авиакомпания» (airline) и «класс полета» (class). В окне *Options* рекомендуется выбрать *Descriptive Statistics* для вывода базовых статистик и *Homogeneity tests* для вывода результатов проверки однородности выборочных дисперсий, а также установить требуемый уровень значимости (по умолчанию $\alpha = 0{,}05$).

Для визуализации результатов в окне *Plots* в качестве горизонтальной оси графика (*Horizontal Axis*) рекомендуется указать фактор class, а для отображения рейтингов авиакомпаний отдельными линиями (*Separate Lines*) – фактор airline. После выбора факторов нужно нажать на кнопку *Add* и затем *Continue*.

Для использования апостериорных критериев множественных сравнений в окне *Post Hoc* из поля *Factor(s)* выбирается фактор airline и переносится в поле *Post Hoc Tests for*. В качестве критерия проверки выбирается тест Шеффе. (Фактор class не выбирается, так как он

содержит лишь две категории, а для факторов с количеством категорий меньше трех критерий множественных сравнений применять нет смысла.)

Базовая статистика по выборкам приведена в отчете *Descriptive Statistics.*

Descriptive Statistics

Dependent Variable: RATING1

AIRLINE	CLASS	Mean	Std. Deviation	N
1	1	35,60	12,140	10
	2	42,30	15,550	10
	Total	38,95	14,006	20
2	1	48,50	12,501	10
	2	69,30	9,166	10
	Total	58,90	15,089	20
3	1	64,80	13,037	10
	2	81,00	9,603	10
	Total	72,90	13,902	20
Total	1	49,63	17,165	30
	2	64,20	20,035	30
	Total	56,92	19,902	60

В отчете по тесту Левена значение *Sig.* больше 0,05, поэтому нулевая гипотеза о равенстве дисперсий не может быть отклонена и ANOVA может применяться.

Levene's Test of Equality of Error Variances[a]

Dependent Variable: RATING1

F	df1	df2	Sig.
,712	5	54	,617

Tests the null hypothesis that the error variance of the dependent variable is equal across groups.

a. Design: Intercept+AIRLINE+CLASS+AIRLINE * CLASS

Из отчета *Tests of Between-Subjects Effects* по значениям *F*-критерия и *p*-уровня значимости мы можем получить ответы на все три поставленных вопроса о влиянии на оценку сервиса компаний двух анализируемых факторов и их совместном влиянии.

Tests of Between-Subjects Effects

Dependent Variable: RATING1

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	15343,883[a]	5	3068,777	20,650	,000
Intercept	194370,417	1	194370,417	1307,962	,000
AIRLINE	11644,033	2	5822,017	39,178	,000
CLASS	3182,817	1	3182,817	21,418	,000
AIRLINE * CLASS	517,033	2	258,517	1,740	,185
Error	8024,700	54	148,606		
Total	217739,000	60			
Corrected Total	23368,583	59			

a. R Squared = ,657 (Adjusted R Squared = ,625)

В пакете STATISTICA получаются полностью аналогичные результаты (*ANOVA/MANOVA – Variables – OK – All effects*).

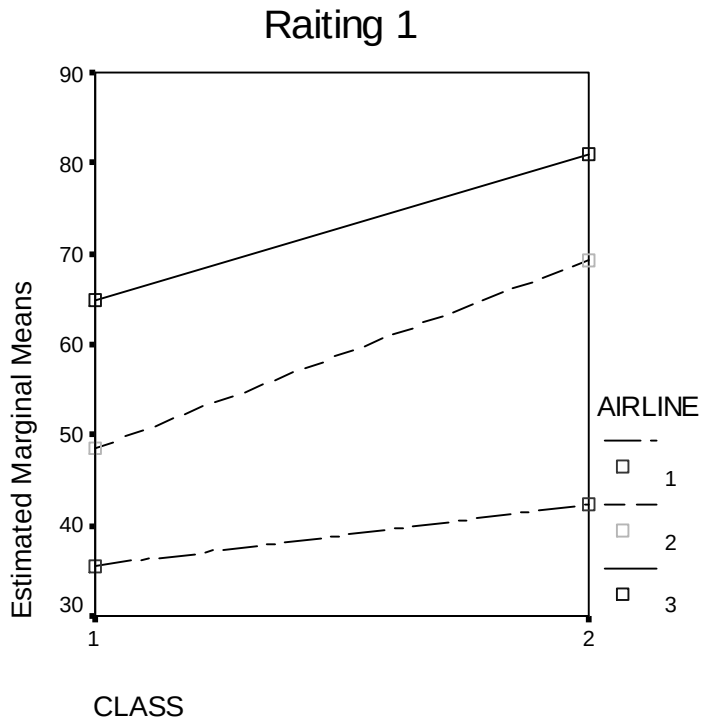
Summary of all Effects; design: (2-way-anova.sta)

Continue... 1-AIRLINE, 2-CLASS

Effect	df Effect	MS Effect	df Error	MS Error	F	p-level
1	2	5822,017	54	148,6056	39,17765	,000000
2	1	3182,817	54	148,6056	21,41788	,000024
12	2	258,517	54	148,6056	1,73962	,185283

Оценим вначале эффект взаимодействия факторов, поскольку анализ воздействия главных эффектов по отдельности будет неполным, если обнаружится эффект взаимного влияния. Как следует из отчета, эффект взаимодействия не существенен на уровне значимости 0,05 (F = 1,74; p = 0,185). Нулевая гипотеза относительно взаимодействия факторов (точнее, их отсутствия) не может быть отклонена. Поэтому теперь можно оценить влияние главных эффектов. Так же, как в однофакторной модели, нулевая гипотеза о неразличимости по компаниям отвергается на уровне 0,001, различия по классам также значимы на уровне 0,001. В отчете приведено значение R^2 = 0,657 (*R Squared*). Это значение показывает, что изменчивость зависимой переменной на 66% зависит от изменчивости анализируемых факторов.

График подтверждает, что между категориями «класс полета» и «авиакомпания» нет взаимодействия, иначе линии были бы не параллельными, а пересекались. Это могло бы означать, например, что по экономическому классу рейтинг одной из компаний выше рейтинга другой (других), а по бизнес-классу – ниже.



Отчет *Multiple Comparisons* представляет результаты применения теста Шеффе. Эти результаты показывают статистически значимое различие среднего рейтинга всех компаний на уровне $p < 0,05$.

Multiple Comparisons

Dependent Variable: RATING1

Scheffe

(I) AIRLINE	(J) AIRLINE	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
1	2	-19,95*	3,855	,000	-29,65	-10,25
	3	-33,95*	3,855	,000	-43,65	-24,25
2	1	19,95*	3,855	,000	10,25	29,65
	3	-14,00*	3,855	,003	-23,70	-4,30
3	1	33,95*	3,855	,000	24,25	43,65
	2	14,00*	3,855	,003	4,30	23,70

Based on observed means.

*. The mean difference is significant at the ,05 level.

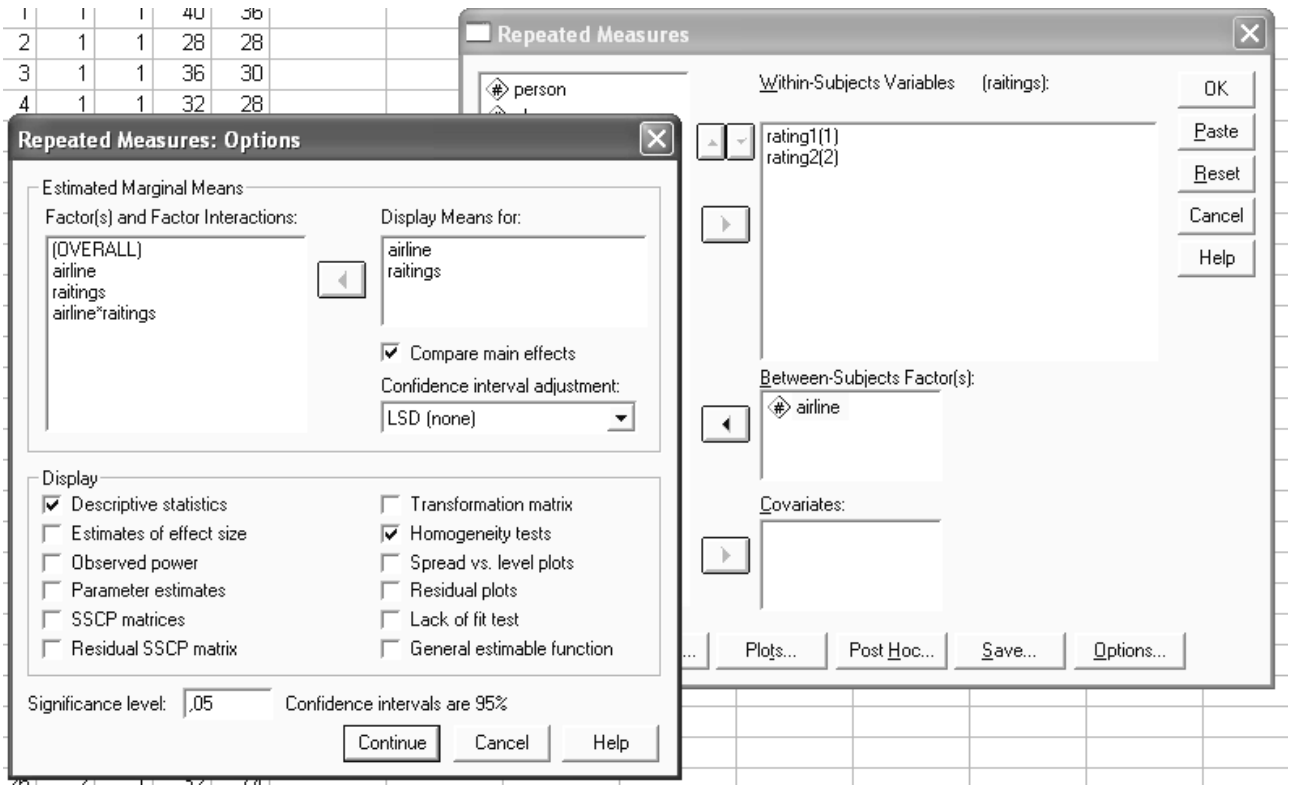
7.5. Дисперсионный анализ с повторными измерениями

В случае, когда осуществляются повторные измерения (хотя бы по одному фактору) по одной и той же группе объектов (т.е. выборки связанные), целесообразно применять специальную модификацию ANOVA – *дисперсионный анализ с повторными измерениями* (*Repeated Measures ANOVA*). В принципе, в данной ситуации можно использовать и

рассмотренный выше вариант многофакторного ANOVA для независимых выборок, но метод с повторными измерениями более чувствителен к влиянию изучаемых факторов на зависимую переменную, поэтому с большей надежностью обнаруживает эффекты такого воздействия. В остальном процедура проверки и анализа результатов аналогична рассмотренным выше.

Дисперсионный анализ с повторными измерениями можно рассматривать как обобщение *t*-критерия Стьюдента для зависимых выборок, позволяющее сделать анализ не только корректным, но и более чувствительным и гибким. Например, можно одновременно контролировать влияние нескольких факторов.

Для примера воспользуемся данными повторной оценки сервиса трех авиакомпаний (в данном случае фактор «класс полета» не будем принимать во внимание). В SPSS в меню *Analyze* выбирается *General Linear Model*, затем *Repeated measures.* В окне *Repeated measures Define Factor(s)* определяем переменные с повторными измерениями. Для этого задается имя фактора, включающего переменные с повторными измерениями. Имя фактора выбирается произвольно. Для нашего примера возьмем имя «raitings». В поле *Number of levels* указывается количество уровней (повторных измерений). Для данного случая – 2. После нажатия на кнопок *Add* и *Define* осуществляется переход в окно *Repeated measures.* В левом поле из списка выбираются переменные с повторными измерениями (в данном случае – «rating1» и «rating2»), затем они переносятся в поле *Within-Subject Variables.*



В поле *Between Subject Factor* переносится анализируемый фактор (в данном случае – переменная «airline»). В окне *Options* отмечаются опции *Descriptive Statistics* и *Homogeneity tests*, а также указывается требуемый уровень значимости. В окне *Plots* для визуализации результатов в качестве оси абсцисс (*Horizontal Axis*) в данном примере указан фактор «airline», а в поле *Separate Lines* – новый фактор («ratings»), объединяющий рейтинги авиакомпаний. Далее нажимаем кнопки *Add, Continue, OK* и получаем выходную форму с несколькими отчетами.

Descriptive Statistics

	AIRLINE	Mean	Std. Deviation	N
RATING1	1	38,95	14,006	20
	2	58,90	15,089	20
	3	72,90	13,902	20
	Total	56,92	19,902	60
RATING2	1	32,40	8,268	20
	2	72,25	10,572	20
	3	79,80	11,265	20
	Total	61,48	23,208	60

В отчете *Descriptive Statistics* приведены средние значения и стандартные отклонения для трех компаний для первого и повторного измерений.

Тест Левена демонстрирует отсутствие статистически достоверных различий для двух измерений зависимой переменной. *Тест Бокса* (*Test of Equality of Covariance Matrices*) также важен с точки зрения корректности применения используемого метода. Как и в тесте Левена, если значение *Sig.* превышает установленный уровень статистической значимости (например 0,05), то применение данного подхода является корректным.

Levene's Test of Equality of Error Variances[a]

	F	df1	df2	Sig.
RATING1	,293	2	57	,747
RATING2	1,472	2	57	,238

Tests the null hypothesis that the error variance of the dependent variable is equal across groups.

a. Design: Intercept+AIRLINE
Within Subjects Design: RAITINGS

Box's Test of Equality of Covariance Matrices[a]

Box's M	7,500
F	1,187
df1	6
df2	80975,077
Sig.	,310

Tests the null hypothesis that the observed covariance matrices of the dependent variables are equal across groups.

a.
Design: Intercept+AIRLINE
Within Subjects Design: RAITINGS

Для содержательной интерпретации результатов наиболее важным является отчет *Multivariate Tests*. В этом отчете приводятся результаты многомерного тестирования данных по нескольким критериям. Из отчета следует, что статистически значимо (*Sig.* < 0,001) различаются оценки, сдвинутые во времени (RATINGS). Тем самым доказывается, что фактор времени в значительной степени обусловливает различия в оценках сервиса авиакомпаний. Вторая половина этого отчета (RATINGS * AIRLINE) демонстрирует, что взаимодействие факторов «время измерения» и «авиакомпания» также достоверно на высоком уровне статистической значимости. Другими словами, сочетание эффекта времени и авиакомпании, сервис которой оценивается, значительно определяет различия в рейтингах.

Multivariate Tests[b]

Effect		Value	F	Hypothesis df	Error df	Sig.
RATINGS	Pillai's Trace	,201	14,318[a]	1,000	57,000	,000
	Wilks' Lambda	,799	14,318[a]	1,000	57,000	,000
	Hotelling's Trace	,251	14,318[a]	1,000	57,000	,000
	Roy's Largest Root	,251	14,318[a]	1,000	57,000	,000
RATINGS * AIRLINE	Pillai's Trace	,453	23,592[a]	2,000	57,000	,000
	Wilks' Lambda	,547	23,592[a]	2,000	57,000	,000
	Hotelling's Trace	,828	23,592[a]	2,000	57,000	,000
	Roy's Largest Root	,828	23,592[a]	2,000	57,000	,000

a. Exact statistic

b.
Design: Intercept+AIRLINE
Within Subjects Design: RATINGS

Главный эффект влияния фактора «airline» на оценки сервиса компаний можно оценить из отчета *Univariate Tests*. Из отчета следует, что и этот фактор является значимым источником различий в оценках ($p < 0,001$).

Univariate Tests

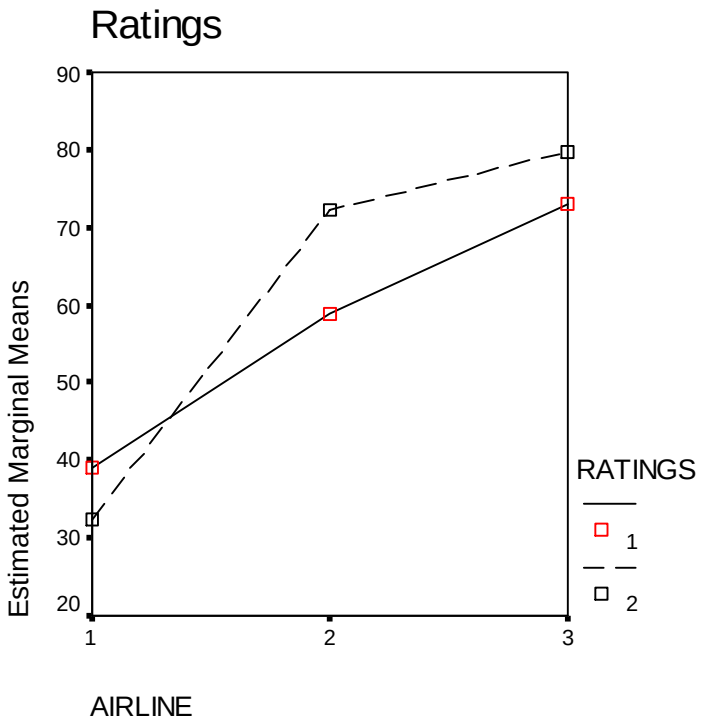
Measure: MEASURE_1

	Sum of Squares	df	Mean Square	F	Sig.
Contrast	17763,775	2	8881,887	67,199	,000
Error	7533,825	57	132,172		

The F tests the effect of AIRLINE. This test is based on the linearly independent pairwise comparisons among the estimated marginal means.

Таким образом, получено, что оценки услуг значимо зависят от компаний; временной интервал оказывает значимое влияние на оценки; сочетание временного интервала и авиакомпании также оказывает значимое влияние на оценки.

Полученные результаты иллюстрирует график средних значений. Прямая 1 соответствует первоначальным измерениям, прямая 2 – повторным. Видно, что оценки услуг компаний отличаются во времени. Оценки второй и третьей компаний по прошествии времени улучшились, а оценки первой – ухудшились, что и отражается пересечением прямых на графике.



В пакете STATISTICA аналогичные результаты могут быть получены с помощью модуля *ANOVA/MANOVA*. В окне *Variables* в качестве независимого фактора (*Independent factor*) указывается переменная «airline», в качестве зависимых (*Dependent*) переменных – переменные «rating1» и «rating2». (Выделить сразу две и более переменных можно, удерживая клавишу *Ctrl* на клавиатуре компьютера.) Нажатием на кнопку *Repeated measures (within SS) design* можно определить факторы, объединяющие зависимые переменные,

подобно тому, как это было сделано при решении задачи в SPSS. С помощью кнопки *All effects* можно получить выходную форму, содержащую результаты, аналогичные полученным в SPSS.

Summary of all Effects; design: (2-way-anova.sta)

GENERAL MANOVA | 1-AIRLINE, 2-RFACTOR1

Effect	df Effect	MS Effect	df Error	MS Error	F	p-level
1	2	17763,78	57	264,3447	67,19928	,000000
2	1	625,63	57	43,6956	14,31799	,000373
12	2	1030,86	57	43,6956	23,59180	,000000

Из данной формы следует, что все эффекты (влияние фактора авиакомпаний (F = 67,199; p < 0,000001), влияние фактора времени (F = 14,318; p = 0,0004) и совместное влияние этих факторов (F = 23,592; p < 0,000001)) достоверны на высоком уровне значимости. Графики средних значений для всех эффектов могут быть построены после нажатия *Mean / Graphs* и выбора анализируемого эффекта в окне *Table of All Effects*.

Глава 8. Непараметрические методы проверки статистических гипотез

8.1. Критерий C^2 Пирсона

Критерий c^2 (хи-квадрат) – один из самых широко используемых критериев для проверки статистических гипотез в социально-экономических и гуманитарных исследованиях. Это наиболее простой критерий проверки значимости связи между категоризованными переменными. Популярность этого критерия связана как с его простотой, так и с возможностями очень гибкого применения. В отличие от параметрических методов, для применения критерия хи-квадрат не требуется выполнения многих допущений и ограничений, с которыми связано использование, например, *z*- и *t*-тестов. Данные могут представляться в любых шкалах, в том числе номинативной шкале; форма распределения также несущественна для этого теста (поэтому такие критерии еще называются *критериями, свободными от распределения* (*distribution free tests*)).

В качестве примеров использования этого критерия можно назвать следующие исследовательские задачи:

- Кто чаще обращается в службу знакомств: мужчины или женщины?
- Влияет ли на частоту посещений магазина уровень доходов покупателей?
- Кто чаще совершает аварии: мужчины или женщины?
- Зависит ли количество аварий от дня недели?

Различаются два варианта применения этого критерия. Первый вариант, который называется *критерием согласия*, предназначен для проверки различий эмпирического и теоретического распределений исследуемого свойства по одной выборке. Сравнивается численность элементов выборки, обладающих определенным свойством, с предполагаемой (теоретической) численностью, согласно определенному виду распределения случайной величины. На практике чаще всего сопоставляются численности долей откликов по выборке с равномерным распределением, хотя вид теоретического распределения, с которым сравнивается эмпирическое распределение, может быть и любым другим. Эмпирические частоты откликов по определенной градации изменения изучаемого свойства называются *наблюдаемыми*, а теоретические частоты – *ожидаемыми*.

Второй вариант использования критерия хи-квадрат, называемый *критерием независимости*, применяется для сравнения распределений *нескольких независимых выборок*. В этом случае нулевая гипотеза формулируется как отсутствие зависимости (связи) между изучаемыми переменными. Применительно к критерию хи-квадрат понятие независимости имеет следующее значение: две переменные являются независимыми, если отнесение наблюдения к определенной категории одной переменной не оказывает влияния на вероятность того, что это наблюдение попадет в определенную категорию другой переменной. Другими словами, при формулировке нулевой гипотезы в терминах независимости истинная нулевая гипотеза означает, что влияние одной выборки на другую (другие) случайно.

В случае использования критерия согласия эмпирическое значение хи-квадрат вычисляется путем сравнения эмпирического, или наблюдаемого (*observed*), распределения частот (O) и теоретического, или ожидаемого (*expected*), распределения (E) по следующей формуле:

$$c^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i},$$

где k – количество градаций изменения изучаемого признака, т.е. это количество сравниваемых категорий (или количество сравнений по определенным категориям).

Следуя общей логике и процедуре проверки статистических гипотез, эмпирическое значение хи-квадрат сравнивается с критическим при выбранном уровне значимости α и числом степеней свободы для исследуемой ситуации. Число степеней свободы для критерия согласия хи-квадрат определяется как $df = k - 1$. Критическое значение хи-квадрат можно получить из специальных статистических таблиц (приложение 1.4), а также с помощью функции Excel ХИ2ОБР(α; *df*).

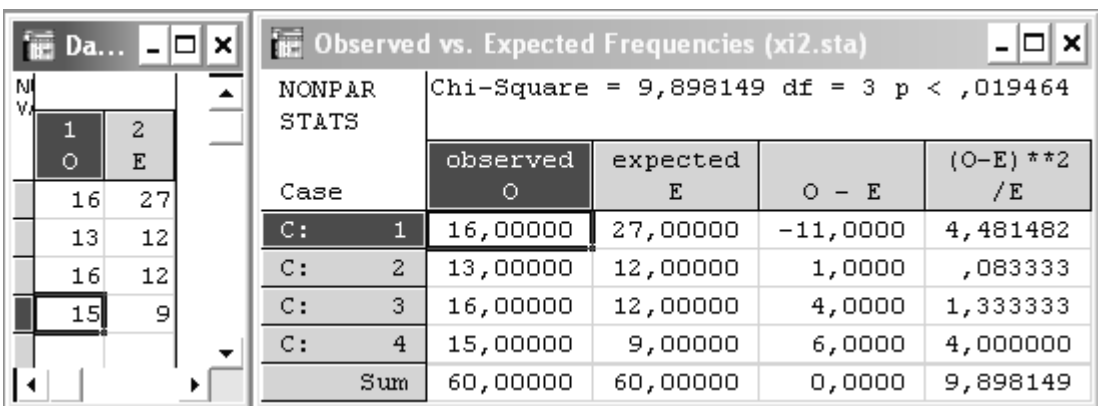
Проиллюстрируем процедуру вычисления статистики критерия согласия хи-квадрат на примере, связанном с мониторингом проекта создания студенческого ланч-клуба (этот пример рассматривался в предыдущих главах). Предположим, что поставлена задача оценить возможное посещение ланч-клуба студентами в зависимости от условий их проживания. Допустим, что было опрошено 200 студентов. Нулевая гипотеза H_0 формулируется следующим образом: желание посещать ланч-клуб не зависит от условий проживания, т.е. наблюдаемое распределение частот соответствует ожидаемому (в данном случае – равномерному) распределению ($O = E$).

В результате опроса получены данные, приведенные в таблице, где указаны частоты откликов респондентов по определенным группам (категориям). В данном случае имеются 4 категории изменения признака, т.е. $k = 4$.

	Намерены посещать (*O*)	Опрошено всего	Ожидаемое значение (*E*)
Общежитие	16	90	27
Квартира в центре	13	40	12
Квартира на окраине	16	40	12
За городом	15	30	9
Всего	60	200	60

В данном случае принцип вычисления ожидаемых частот (*E*) очевиден: количество опрошенных по каждой из 4 категорий умножается на отношение суммы наблюдаемых частот *O* (сколько всего человек намерены посещать клуб) к общему количеству опрошенных (предполагается равномерное распределение). Рассчитывая статистику хи-квадрат, имеем эмпирическое значение критерия $\chi^2 = 9{,}89$ (рекомендуется проверить самостоятельно). Количество степеней свободы $df = 4 - 1 = 3$. При $\alpha = 0{,}05$ критическое значение $\chi^2 = 7{,}82$. Сравнивая критическое и эмпирическое значения критерия, принимаем решение об отклонении нулевой гипотезы и делаем вывод о статистически значимом влиянии условий проживания студентов на их вероятное посещение ланч-клуба.

Такого рода задачи можно легко решить в среде STATISTICA. В модуле *Nonparametrics/Distrib.* выбирается процедура *Observed versus expected XI*, затем в окне *Variables* – переменные, содержащие наблюдаемые и ожидаемые частоты, после чего нажимается кнопка *OK*. Исходные данные из рассмотренного выше примера и результаты вычисления критерия согласия хи-квадрат представлены на рисунке.



При вычислении критерия согласия хи-квадрат в SPSS исходные данные должны быть представлены в виде столбца, содержащего значения номинативной переменной по нескольким категориям (кодировка значений каждой категории может быть произвольной). Процедура расчета критерия согласия запускается из меню *Analyze*, *Nonparametric tests*, *Chi-square*. В открывшемся окне выбирается необходимая переменная и переносится в поле *Test Variable List*. После нажатия кнопки *ОК* вычисляются наблюдаемые и ожидаемые частоты, статистика хи-квадрат, уровень значимости, которые представляются в соответствующих отчетах формируемой выходной формы.

Применение критерия независимости хи-квадрат основано на использовании *таблицы сопряженности*, которая отражает перекрестное распределение частот представленности изучаемого признака по градациям его изменения в одной или нескольких сопоставляемых выборках. Для иллюстрации понятия таблицы сопряженности предположим, что поставлена задача оценить различие в предполагаемом посещении ланч-клуба студентами и студентками. Всего было опрошено 50 студентов и 50 студенток. Предпочтения мужчин и женщин по посещению ланч-клуба отражены в таблице.

	Вероятно, буду посещать	Вероятно, не буду посещать	Всего
Мужчины	26	24	50
Женщины	35	15	50
Всего	61	39	100

Это таблица сопряженности 2 х 2, в которой количество сопоставляемых выборок равно 2 (мужчины и женщины) и количество градаций изучаемого признака тоже равно 2 («буду посещать» и «не буду посещать»). Если бы все студентки выбрали, предположим, вариант «буду посещать», а все студенты – вариант «не буду», то было бы очевидно, что между предпочтениями студентов и студенток есть явно выраженные различия. В данном случае различия в предпочтениях выражены слабее, поэтому наблюдаемое различие необходимо проверять на статистическую значимость.

Для проверки статистической значимости различий такого рода может быть использован критерий независимости хи-квадрат, который, как и критерий согласия, тестирует значимость различия между *наблюдаемыми* и *ожидаемыми* частотами. В нашем примере, если между предпочтениями студентов разного пола нет различий (или, что то же самое, между предпочтением и полом нет связи), то естественно ожидать равного количества частот выбора вариантов «посещать» и «не посещать» для каждой выборки. Относительно

малые отклонения наблюдаемых частот от ожидаемых будут доказывать значимость, если число наблюдений велико.

Формула для определения экспериментального значения критерия независимости хи-квадрат незначительно отличается от формулы для критерия согласия. Отличие заключается в том, что сравнение наблюдаемых частот (O) и ожидаемых частот (E) осуществляется по всем ячейкам таблицы сопряженности.

$$c^2 = \sum_i^k \sum_j^l \frac{(O_{ij} - E_{ij})^2}{E_{ij}},$$

где k – число градаций изменения изучаемого признака; l – количество сопоставляемых распределений (выборок); индексы i,j соответствуют ячейкам таблицы сопряженности.

Ожидаемые частоты в таблице сопряженности при гипотезе «нет различий / нет связи» можно вычислить непосредственно. Для этого для каждой ячейки таблицы предварительно определяется так называемая краевая, или маргинальная (*marginal*), частота по каждой строке и каждому столбцу таблицы (в таблице рассматриваемого примера она обозначена как «Всего»). Ожидаемая частота для каждой ячейки таблицы равна произведению краевой частоты по столбцу на краевую частоту по строке, деленному на общее количество наблюдений в таблице (n). В нашем случае n = 100. Тогда для левой верхней ячейки имеем 61 * 50 / 100 = 30,5. В общем случае ожидаемая частота в ячейке ij рассчитывается как

$$E_{ij} = \frac{f_i \cdot f_j}{n},$$

где f_i – сумма частот i-й строки; f_j – сумма частот j-го столбца таблицы сопряженности.

Таким образом, значение статистики хи-квадрат зависит от общего числа наблюдений и количества ячеек в таблице сопряженности. Число степеней свободы для критерия независимости хи-квадрат определяется как $df = (k - 1)(l - 1)$, где k – число градаций изменения изучаемого признака, l – количество сопоставляемых распределений.

Рассмотрим еще один пример. Директор фабрики, который пытается ввести на фабрике запрет на курение, намерен установить, связано ли курение с браком на производстве. По результатам учета бракованных изделий случайным образом сформированы списки рабочих, соответственно допускавших и не допускавших брак в работе. После интервьюирования рабочих об их пристрастии к курению (случайная выборка) и сопоставления со сведениями о браке построена таблица сопряженности, в которой указаны также определенные ожидаемые частоты.

	Брак	*Нет*	*Всего*
Заядлый курильщик	12 8,24	4 7,75	16
Умеренный курильщик	9 7,73	6 7,27	15
Не курит	13 18,03	22 16,97	35
Всего	34	32	66

Читателю рекомендуется проверить, что эмпирическое значение хи-квадрат составляет 6,86, а критическое значение – 5,59 (α = 0,05, df = 2). Нулевая гипотеза отклоняется, и принимается, что случаи брака связаны с курением.

В случае, когда количество сравниваемых выборок превышает 2, использование критерия независимости хи-квадрат не отличается от случая двух независимых выборок. Все расчеты производятся аналогично.

Альтернативная гипотеза, принимаемая при отклонении нулевой гипотезы об отсутствии связи, неконкретна, в том смысле, что можно утверждать, что связь проявляется по крайней мере для одной ячейки таблицы сопряженности, но неизвестно, для какой именно. Поэтому, как и в случае применения метода ANOVA, возникает необходимость конкретизации результатов. Для этого исходную таблицу сопряженности обычно разбивают на подтаблицы 2 x 2, после чего проводят множественные сравнения с помощью критерия хи-квадрат.

Имеется существенное ограничение на использование критерия хи-квадрат (кроме предположения о случайном выборе наблюдений и независимости выборок). Это ограничение состоит в том, что ожидаемые частоты не должны быть очень малы. Существует эмпирическое правило, помогающее определить минимальное значение ожидаемых частот, при котором применение критерия хи-квадрат можно считать корректным. Если количество степеней свободы df = 1, то ожидаемые частоты в каждой ячейке не должны быть меньше 5. Если количество степеней свободы df > 1, то в таблице сопряженности не должно быть свыше 20% ячеек с ожидаемыми частотами ниже 5 и ни одной ячейки с ожидаемой частотой ниже 1.

Для таблиц сопряженности 2 x 2 с малыми частотами (меньше 10) рекомендуется рассчитывать эмпирическое значение критерия хи-квадрат, используя *поправку на непрерывность Йетса* (*Yates' correction for continuity*):

$$c^2 = \sum_i^2 \sum_j^2 \frac{(|O_{ij} - E_{ij}| - 0{,}5)^2}{E_{ij}}.$$

Поскольку поправка Йетса применяется лишь для таблиц 2 x 2, то $df = 1$. Поправка Йетса делает оценку связей более умеренной. Такая поправка может использоваться, если объем выборки $n > 40$ **или** n находится между 20 и 40 **и** все значения ожидаемых частот выше 5. Если эти требования не выполняются, можно применить так называемый *точный критерий Фишера* (*Fisher's exact test*), расчет которого целесообразно выполнять с помощью статистических пакетов.

Точный критерий Фишера, как и поправка Йетса, применим только для таблиц сопряженности 2 x 2. Этот критерий основан на вычислении вероятности получения наблюдаемых частот исходя из маргинальных частот, заданных в таблице сопряженности. Такая вероятность *точно* вычисляется подсчетом всех таблиц, которые можно построить исходя из известных маргинальных частот. Таким образом, критерий Фишера вычисляет вероятность появления наблюдаемых частот при нулевой гипотезе (отсутствие связи между табулированными переменными).

Для таблиц размерности большей, чем 2 x 2, не существует формулы для корректировки эмпирического значения критерия хи-квадрат в случае малых выборок. Выход может заключаться в том, чтобы объединить некоторые категории переменной, тем самым увеличив ожидаемые частоты. Но такой подход можно использовать только тогда, когда исследователь считает возможным (с точки зрения здравого смысла или теоретического обоснования) игнорировать различия в тех или иных категориях. Если категории не удается объединить, а количество ячеек с частотами 5 и менее мало, то критерий можно применять, но результаты следует интерпретировать с осторожностью.

Как и все статистические тесты, хи-квадрат чувствителен к объему выборки, даже в большей степени, чем другие тесты (так как эмпирическое значение критерия изменяется пропорционально объему выборки). Поэтому использование больших выборок может привести к решению об отклонении нулевой гипотезы, хотя связи между переменными могут быть несущественны. Это обстоятельство еще раз подчеркивает различие между статистической значимостью и практической важностью. Критерий хи-квадрат, как и все другие статистические критерии, может служить индикатором того, являются ли результаты статистически значимыми. Но он не может свидетельствовать о том, *важны* ли результаты в любом другом смысле. Отчасти это может разрешить оценка мер связи, рассматриваемая в следующей главе.

В SPSS для вычисления критерия независимости хи-квадрат, а также других критериев, рассчитываемых по таблицам сопряженности, используется процедура *Crosstabs*,

запускаемая из меню *Analyze, Descriptive Statistics*. Исходные данные для работы должны быть в виде столбцов, соответствующих изменяемым номинативным переменным. Например, для нашего примера первый столбец (defect) может включать выборки работников, допускавших и не допускавших брак. В этот столбец могут входить значения 1 и 2, которые соответствуют работе с браком и без брака. В столбце будут 34 элемента со значением 1 и 32 элемента со значением 2. Второй столбец (smoke) – степень пристрастия работника к курению (например, 1 – «заядлый курильщик», 2 – «умеренный курильщик», 3 – «не курит»). Всего в этом столбце будет 16 элементов со значением 1, 15 элементов со значением 2 и 35 элементов со значением 3 (всего 66 элементов). В поля *Row(s)* и *Column(s)* необходимо перенести переменные, соответствующие строкам и столбцам таблицы сопряженности. В данном случае переменная defect переносится в поле *Column(s)*, а переменная smoke – в поле *Row(s)*. После нажатия на кнопку *Statistics* выбирается режим *Chi-square*. После нажатия на кнопки *Continue* и *OK* осуществляется подсчет наблюдаемых и ожидаемых частот и статистик критерия хи-квадрат. Результаты расчетов представляются в отчетах выходной формы. Результаты полностью совпадают с результатами, полученными выше.

SMOKE * DEFECT Crosstabulation

Count

		DEFECT		
		1	2	Total
SMOKE	1	12	4	16
	2	9	6	15
	3	13	22	35
Total		34	32	66

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	6,860[a]	2	,032
Likelihood Ratio	7,070	2	,029
Linear-by-Linear Association	6,688	1	,010
N of Valid Cases	66		

a. 0 cells (,0%) have expected count less than 5. The minimum expected count is 7,27.

В пакете STATISTICA процедура работы с таблицами сопряженности осуществляется в модуле *Basic statistics* (меню *Tables and banners*). Исходные данные также задаются в виде столбцов, содержащих значения номинативной переменной по категориям изменения. Для таблиц сопряженности 2 x 2 статистики всех критериев выводятся в одной выходной форме после ввода наблюдаемых частот (модуль *Nonparametrics/Distrib.*, процедура *2 x 2 Tables…*).

8.2. Тест Мак-Немара

Если гипотеза проверяется для зависимых (связанных) выборок с данными, представленными в неметрической шкале, то используется *тест Мак-Немара* (*McNemar test*). Особенно полезен этот тест при повторных измерениях. Этот метод позволяет сравнить долю респондентов, которые не имели некоторой характеристики (0) во время первого обследования (например до воздействия), но стали обладать ею во время повторного обследования (1), с долей тех, кто имел это свойство (1), но перестал иметь его после (0).

До	*После*	
	0	1
1	A	B
0	C	D

Например, метод может применяться для оценки предпочтений некоторого товара одними и теми же респондентами *до* и *после* рекламы. В таком случае «0» может означать, что товар не нравится, а «1» – что товар нравится.

Вычисляются два значения хи-квадрат: для ячеек A и D (1,0 и 0,1) и для ячеек B и C (1,1 и 0,0). A/D хи-квадрат проверяет гипотезу о том, что частоты в ячейках A и D (верхняя левая, нижняя правая) одинаковы. B/C хи-квадрат проверяет гипотезу о равенстве частот в ячейках B и C (верхняя правая, нижняя левая). Метод не может использоваться для выборок, содержащих меньше 10 элементов.

Для ячеек A и D эмпирическое значение хи-квадрат для теста Мак-Немара вычисляется по формуле

$$c^2 = \frac{(|A-D|)^2}{A+D}, \qquad df = 1.$$

Нулевая гипотеза формулируется как H_0: P(A) = P(D).

Соответственно для ячеек B и C эмпирическое значение хи-квадрат определяется как

$$c^2 = \frac{(|B-C|)^2}{B+C},$$

а нулевая гипотеза формулируется как H_0: P(B) = P(C).

Подобно тому, как для независимых выборок вводится поправка Йетса, при применении теста Мак-Немара также рекомендуется использовать модифицированную

формулу для эмпирического значения хи-квадрат, дающую более умеренную оценку. Например, для ячеек A и D эмпирическое значение хи-квадрат должно определяться следующим образом:

$$c^2 = \frac{(|A-D|-1)^2}{A+D}, \qquad df = 1.$$

В пакетах SPSS и STATISTICA метод реализован с использованием критерия хи-квадрат, но реализация метода Мак-Немара может быть основана и на применении z-критерия. В этом случае проверка гипотез (а) (1,0 и 0,1) и б) (1,1 и 0,0)) осуществляется на основе одной из формул для соответствующих гипотез:

а) $$z = \frac{|A-D|}{\sqrt{A+D}};$$ б) $$z = \frac{|B-C|}{\sqrt{B+C}}.$$

Найденные по этим формулам эмпирические значения *z*-критерия сравниваются с критическими, на основе чего делается вывод об отклонении или невозможности отклонения нулевой гипотезы.

Из данных формул для эмпирических значений следует, что значение *z*, возведенное в квадрат, будет равняться значению хи-квадрат, определенному по формуле без корректировки. Для иллюстрации взаимосвязи результатов, полученных при использовании хи-квадрат и *z*-критерия в методе Мак-Немара, приведем два примера.

Пример 1. В компании проводился опрос, насколько сотрудники ценят коллективную работу. Была случайным образом сформирована выборка из 200 сотрудников. По результатам опроса сотрудники были разделены на 2 группы с условным названием «Нравится коллективная работа» и «Не нравится коллективная работа». В компанию был приглашен специалист, который провел 2-дневный тренинг для сотрудников, включающий элементы совместного принятия решений. После тренингов по той же выборке был осуществлен повторный опрос. Результаты представлены в таблице.

До	*После*	
	Не нравится	Нравится
Нравится	10	90
Не нравится	60	40

Нулевая гипотеза формулируется как

H_0: P(A) = P(D);

H_A: P(A) ≠ P(D).

Используется тест Мак-Немара, поскольку данные представлены в номинативной шкале, две выборки являются зависимыми (проводятся повторные измерения). При подсчете «вручную» получаем эмпирическое значение хи-квадрат, равное $29^2 / 50 = 16{,}82$. Критическое значение (при $\alpha = 0{,}05$ и $df = 1$) равно 3,84. Нулевая гипотеза отклоняется. Фактически эмпирическое значение таково, что $p < 0{,}001$. Отметим, что при применении формулы без корректировки получим $30^2 / 50 = 18$.

Определим теперь эмпирическое значение z-критерия:

$$z = \frac{|10 - 40|}{\sqrt{10 + 40}} = 4{,}243.$$

При $\alpha = 0{,}05$ критическое значение z составляет 1,96. Следовательно, и на основании этого критерия мы отклоняем нулевую гипотезу. Заметим теперь, что если возвести полученное эмпирическое значение z в квадрат, то мы получим эмпирическое значение хи-квадрат, определенное по формуле без корректировки: $4{,}243^2 = 18$.

Пример 2. С помощью этого примера продемонстрируем еще один случай, когда выборки должны рассматриваться как связанные. Компьютерная компания тестирует две различные антивирусные программы (А и Б) на определенном количестве зараженных файлов (156). Результаты тестирования (количество файлов, «излеченных» от вирусов) представлены в таблице.

«Излечено» программой А	*«Излечено» программой Б*	
	Нет	Да
Да	33	45
Нет	20	58

Можно ли, основываясь на этих испытаниях, сказать, что программы работают приблизительно с одинаковой эффективностью?

Данные представлены в номинативной шкале («излечен» – «не излечен»), но применять в этом случае критерий хи-квадрат для независимых выборок нельзя, поскольку программы тестируются на одной и той же выборке зараженных файлов. Фактически мы работаем с одной выборкой, подвергая ее разным воздействиям, поэтому мы должны принять предположение связанности выборок и, следовательно, использовать тест Мак-Немара.

В качестве индикатора эффективности работы антивирусной программы выберем следующий показатель: «количество файлов, излеченных одной программой, но не излеченных другой». В таком случае мы должны сравнивать ситуации (1,0) и (0,1), т.е. нас

интересуют ячейки A и D. Нулевая гипотеза формулируется как H_o: P(A) = P(D) – показатели эффективности программ равны. Соответственно альтернативная гипотеза формулируется как H_A: P(A) ≠ P(D) – показатели эффективности программ не равны.

Определяем эмпирическое значение *z*:

$$z = \frac{|33-58|}{\sqrt{33+58}} = 2{,}62.$$

При α = 0,05 критическое значение *z* составляет 1,96. Следовательно, мы можем отклонить нулевую гипотезу и признать, что эффективность программ статистически значимо различается.

Определим теперь эмпирическое значение хи-квадрат по формуле без корректировки:

$$c^2 = \frac{(|33-58|)^2}{33+58} = 6{,}87.$$

Отметим, что квадратный корень из этого значения равен эмпирическому значению *z*: $\sqrt{6{,}87} = 2{,}62$.

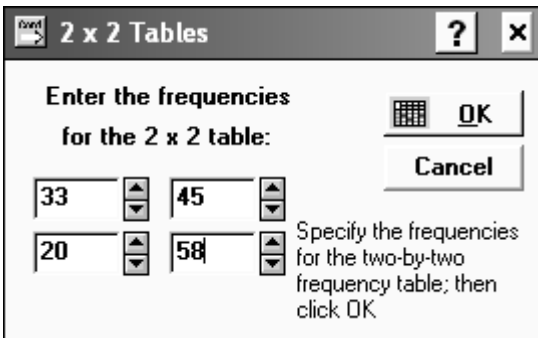
Критическое значение хи-квадрат при *df* = 1 и α = 0,05 равно 3,84. Таким образом, нулевая гипотеза отклоняется и на основе этого критерия. Применив модифицированную формулу, получим

$$c^2 = \frac{(|33-58|-1)^2}{33+58} = 6{,}33.$$

Нулевая гипотеза также отвергается, но уже при несколько меньшем *p*-уровне значимости.

В случае, если результаты измерений сразу представлены в виде таблицы сопряженности, то удобнее и быстрее определить эмпирические значения хи-квадрат или *z* по формулам метода Мак-Немара «вручную» или с помощью программы STATISTICA, чем с помощью SPSS. Дело в том, что, как отмечалось в предыдущем параграфе, исходные данные для работы с таблицами сопряженности в SPSS должны представляться в виде столбцов, соответствующих изменяемым номинативным переменным (например состояниям «до» и «после»). Эти переменные содержат значения, например 0 или 1, в зависимости от наличия или отсутствия интересующей исследователя характеристики. На основе данных, введенных таким образом, программа SPSS формирует таблицы сопряженности и вычисляет необходимые статистические показатели, которые можно использовать для принятия решения об отклонении или невозможности отклонения нулевой гипотезы.

Программа STATISTICA позволяет работать с таблицами сопряженности 2 x 2 непосредственно по заданию наблюдаемых частот в окне *2 x 2 Tables*, выводимом выбором процедуры *2 x 2 Tables…* модуля *Nonparametrics/Distrib.*



В отчете выводятся все статистические критерии, рассчитываемые для таблиц сопряженности 2 x 2, в том числе статистика теста Мак-Немара.

2 x 2 Table (xi2.sta)

Continue...	Column 1	Column 2	Row Totals
Frequencies, row 1	33	45	78
Percent of total	21,154%	28,846%	50,000%
Frequencies, row 2	20	58	78
Percent of total	12,821%	37,179%	50,000%
Column totals	53	103	156
Percent of total	33,974%	66,026%	
Chi-square (df=1)	4,83	p= ,0280	
V-square (df=1)	4,80	p= ,0285	
Yates corrected Chi-square	4,12	p= ,0425	
Phi-square	,03096		
Fisher exact p, one-tailed		p= ,0210	
two-tailed		p= ,0420	
McNemar Chi-square (A/D)	6,33	p= ,0119	
Chi-square (B/C)	8,86	p= ,0029	

Если же исходные данные представлены в виде двух наборов данных в номинативной шкале, то целесообразно использовать программу SPSS. Для этого в меню *Analyze* необходимо выбрать *Descriptive Statistics*, затем *Crosstabs*. Одна из переменных переносится в поле *Row(s)*, другая – в поле *Column(s)*. В *Statistics* выбирается нужный тест *(McNemar)*. Тест Мак-Немара можно запустить также через *Analyze, Nonparametric Tests, 2 related samples, McNemar*.

8.3. Критерий Манна-Уитни

Для сравнения двух независимых выборок может быть применен *U-критерий Манна-Уитни* (*Mann-Whitney U-test*). Этот статистический критерий является самым мощным (чувствительным) среди непараметрических аналогов *t*-критерия Стьюдента для независимых выборок. Для использования этого критерия данные должны быть представлены, по меньшей мере, в порядковой шкале (но тест может быть применен и для шкал большей мощности).

Эмпирическое значение критерия показывает, насколько совпадают два ряда значений измеряемого признака. Для вычисления статистики критерия значения двух выборок ранжируются: представляются в виде одной упорядоченной последовательности значений – от наименьших к наибольшим. Наибольшие отрицательные значения имеют наименьший ранг. Если несколько элементов имеют одинаковое значение, им приписываются средние арифметические значения рангов. Например, два элемента имеют значение 1300 и находятся на 8 и 9 месте. Каждому из этих элементов приписывается ранг 8,5.

Эмпирическое значение *U*-критерия вычисляется по формуле

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1$$

или

$$U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2,$$

где n_1 – объем первой выборки; n_2 – объем второй выборки; R_1 – сумма рангов по первой выборке; R_2 – сумма рангов по второй выборке.

Правильность вычислений проверяется по формуле

$$U_1 + U_2 = n_1 n_2.$$

В качестве эмпирического значения принимается наименьшее значение из U_1, U_2. Критические значения этого критерия можно взять из соответствующих статистических таблиц (см. приложение 1.5). Обратите внимание, что для этого теста нулевая гипотеза отклоняется, если эмпирическое значение *меньше* критического.

Приведем пример. Предположим, оценивается эффективность работы персонала филиалов компании в двух регионах. Для такой оценки среди клиентов распространяется специально разработанная анкета. Результаты анкетирования представляются в порядковой шкале по нескольким показателям. (Это может быть и метрическая шкала, например объемы продаж, но будем считать, что по каким-то причинам *t*-критерий Стьюдента неприменим). Результаты оценки даны в таблице. В таблице также приведены ранги оцененных показателей.

Филиал 1	Ранг	Филиал 2	Ранг
1500	15	1340	10
1540	16	1300	8,5
1860	22	1620	18
1230	6	1070	3
1370	12	1210	5
1550	17	1170	4
1840	21	1770	20
1250	7	950	1
1300	8,5	1380	13
1350	11	1460	14
1710	19	1030	2

Нулевая гипотеза: рейтинги двух филиалов не различаются. Альтернативная гипотеза: рейтинг Филиала 1 выше рейтинга Филиала 2. Определим суммы рангов (R_1 = 154,5; R_2 = 98,5), эмпирические значения критерия:

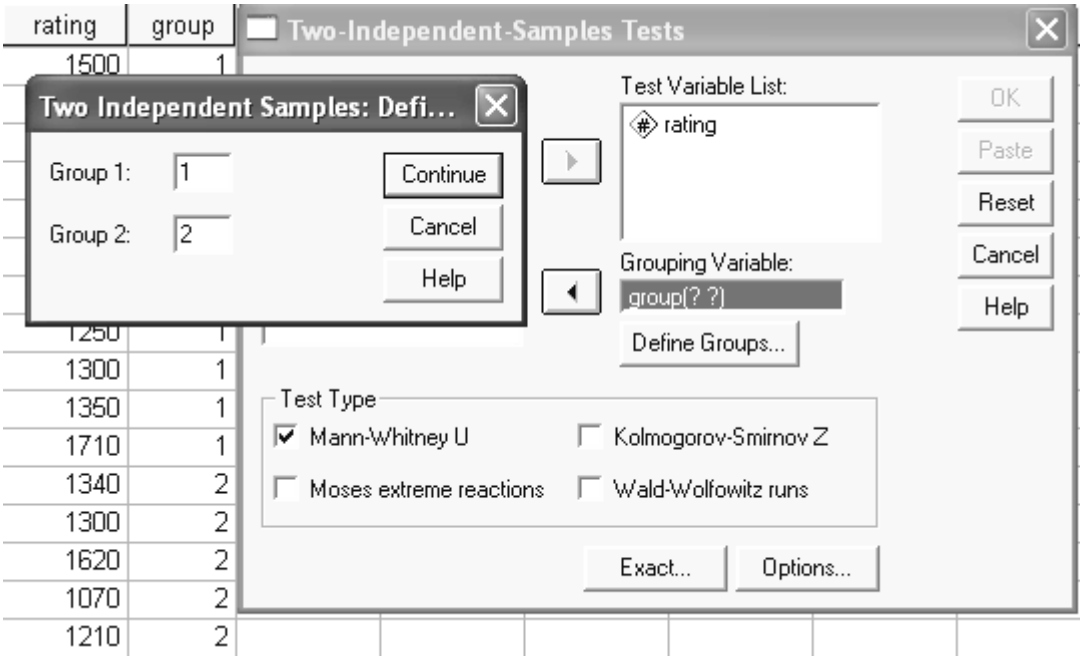
$U1$ = 11 * 11 + (11(11 + 1)) / 2 – 154,5 = 32,5;

$U2$ = 11 * 11 + (11(11 + 1)) / 2 – 98,5 = 88,5.

Критическое значение критерия составляет 34 (α = 0,05, $n_1 = n_2 = 11$, однонаправленный тест). Наименьшее значение эмпирического значения критерия меньше критического значения, следовательно, нулевая гипотеза отклоняется.

Для использования *U*-критерия Манна-Уитни в статистических пакетах данные из двух выборок объединяются в одну переменную, в качестве второй (группирующей) переменной берутся идентифицирующие значения (например 1 и 2). В пакете SPSS из меню *Analyze* выбирается опция *Nonparametric Tests*, затем *2-Independent Samples.*

Независимая переменная (в данном примере – rating) переносится в поле *Test Variable(s)*, группирующая переменная (в данном примере – group) – в поле *Grouping Variable(s)*. Для того чтобы задать коды группирующей переменной, соответствующие двум выборкам, необходимо нажать на кнопку *Define Groups* и установить идентифицирующие значения (в данном случае – 1 и 2).



После нажатия кнопок *Continue* и *OK* генерируются отчеты. В отчете *Ranks* приводятся объемы выборок (*N*), средние ранги (*Mean Rank*) и суммы рангов (*Sum of Ranks*) по выборкам (сравните эти значения с определенными нами выше).

Ranks

	GROUP	N	Mean Rank	Sum of Ranks
RATING	1	11	14,05	154,50
	2	11	8,95	98,50
	Total	22		

Test Statistics[b]

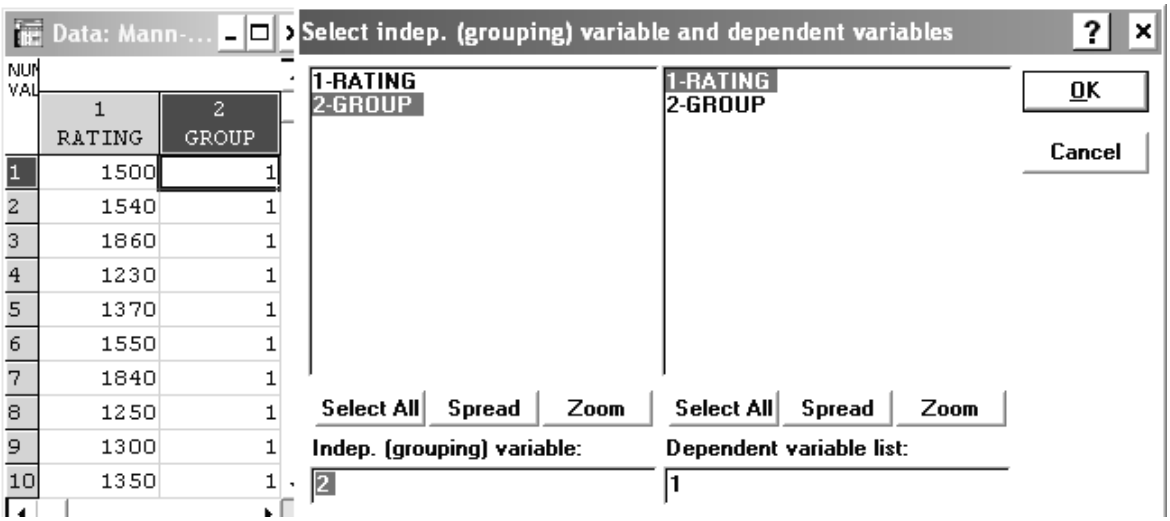
	RATING
Mann-Whitney U	32,500
Wilcoxon W	98,500
Z	-1,839
Asymp. Sig. (2-tailed)	,066
Exact Sig. [2*(1-tailed Sig.)]	,065[a]

a. Not corrected for ties.

b. Grouping Variable: GROUP

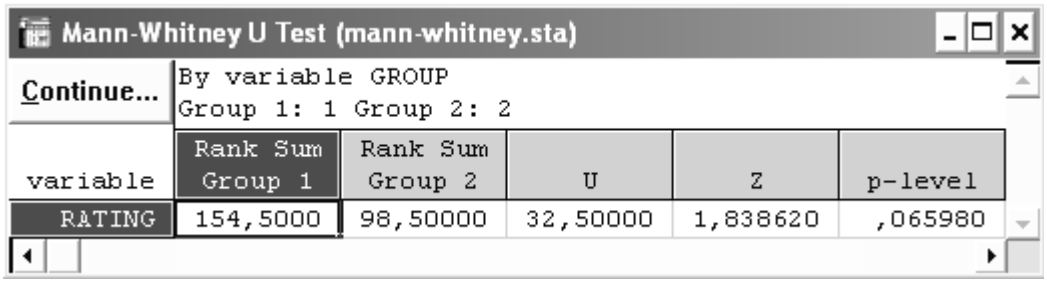
В отчете *Test Statistics* приводятся минимальное значение из эмпирических значений *U*-критерия Манна-Уитни (*Mann-Whitney U*) и *p*-уровень для *двусторонней* проверки (*Asymp. Sig. (2-tailed)*). В нашем примере альтернативная гипотеза формулировалась как однонаправленная, поэтому полученный *p*-уровень должен быть поделен пополам. Если бы осуществлялась двусторонняя проверка, то нулевая гипотеза не могла бы быть отклонена на уровне значимости 0,05.

Вычисления в пакете STATISTICA осуществляются аналогично. После запуска модуля *Nonparametrics/Distrib.* в меню *Analysis* запускается *Startup Panel* и выбирается процедура *Mann-Whitney U test.* В окне *Variables* в качестве группирующей переменной (*Indep. (grouping) variable*) указывается переменная group, а в качестве зависимой (*Dependent variable list*) – переменная rating.



	1 RATING	2 GROUP
1	1500	1
2	1540	1
3	1860	1
4	1230	1
5	1370	1
6	1550	1
7	1840	1
8	1250	1
9	1300	1
10	1350	1

Решение, которое приводится в выходной форме, идентично полученному выше.



variable	Rank Sum Group 1	Rank Sum Group 2	U	Z	p-level
RATING	154,5000	98,50000	32,50000	1,838620	,065980

8.4. Тест Колмогорова-Смирнова

Тест Колмогорова-Смирнова (*Kolmogorov-Smirnov test*) используется для данных, представленных по крайней мере в порядковой шкале, для сравнения наблюдаемого распределения и теоретического распределения, например нормального, равномерного, пуассоновского или биноминального. Одновыборочный тест Колмогорова-Смирнова (*Kolmogorov-Smirnov one-sample test*) является более мощным, чем хи-квадрат, и, более того, он может быть применен для выборок, объем которых даже меньше, чем минимально допустимый объем выборки для корректного использования хи-квадрат. Тест Колмогорова-Смирнова основан на принципе подсчета *накопленных* теоретических (F_T) и наблюдаемых (F_0) частот и определения их максимального расхождения (D).

$$D = max\ |F_0(x) - F_T(x)|.$$

Найденное значение *D* сравнивается с критическим значением (приложение 1.6).

Проиллюстрируем использование этой процедуры на примере знакомой ситуации с проектом ланч-клуба. Предположим, необходимо оценить, влияет ли курс обучения на желание студентов посещать ланч-клуб. Нулевая гипотеза формулируется как H_0: Нет связи между курсом обучения студентов и их желанием посещать ланч-клуб. Допустим, что было опрошено 60 чел. Результаты опроса и обработки данных по методу Колмогорова-Смирнова представлены в таблице. Принцип определения накопленных наблюдаемых и теоретических частот (предполагается равномерное распределение) достаточно наглядно показан в таблице.

Курс обучения	1	2	3	4	5
Количество студентов, выразивших желание посещать клуб	5	9	11	16	19
Накопленные наблюдаемые частоты	5/60	14/60	25/60	41/60	60/60
Накопленные теоретические частоты	12/60	24/60	36/60	48/60	60/60
$\|F_0(x) - F_T(x)\|$	7/60	10/60	11/60	7/60	0

Из данных таблицы определяем максимальное расхождение накопленных наблюдаемых и теоретических частот *D* = 11 / 60 = 0,183. Из таблицы критических значений (приложение 1.6) находим критическое значение критерия, которое для α = 0,05 и n = 60 составляет $1{,}36 / \sqrt{60} = 0{,}175$. Эмпирическое значение критерия превышает его критическое значение, поэтому нулевая гипотеза должна быть отклонена.

Работа с одновыборочным тестом Колмогорова-Смирнова в среде SPSS осуществляется с помощью процедуры *1-Sample K-S.* (меню *Analyze, Nonparametric Tests*).

Двухвыборочный тест Колмогорова-Смирнова (*Kolmogorov-Smirnov two-sample test*) применяется для проверки гипотезы о том, что две независимые выборки $x = (x_1, x_2, \ldots, x_{n1})$ и $y = (y_1, y_2, \ldots, y_{n2})$ получены из одной генеральной совокупности, т.е. функции распределения $F_1(x)$ и $F_2(y)$ двух генеральных совокупностей одинаковы (генеральные совокупности *однородны*). Для использования теста должны выполняться предположения о независимости выборок и представленности данных в порядковой шкале.

Эмпирическим значением критерия является максимальное расхождение частот

$$D = max\ |F_1(x) - F_2(x)|.$$

Значение *D* вычисляется так, как и в случае одновыборочного теста, но критические значения критерия будут другими (приложение 1.7).

Процедура работы с двухвыборочным тестом Колмогорова-Смирнова в пакетах SPSS и STATISTICA не отличается от процедуры использования *U*-критерия Манна-Уитни, рассмотренной в предыдущем параграфе, за исключением того, что выбирается не *Mann-Whitney U test*, а *Kolmogorov-Smirnov test*.

Обратим внимание, что если применить тест Колмогорова-Смирнова к данным примера, рассмотренного в предыдущем параграфе, то статистически значимое различие между рейтингами двух филиалов компании не обнаруживается, в отличие от использования *U*-критерия Манна-Уитни.

Test Statistics[a]

		RATING
Most Extreme Differences	Absolute	,455
	Positive	,000
	Negative	-,455
Kolmogorov-Smirnov Z		1,066
Asymp. Sig. (2-tailed)		,206

a. Grouping Variable: GROUP

8.5. Критерий знаков и Т-критерий Уилкоксона

Для сравнения *двух зависимых* выборок, элементы которых представлены в неметрических шкалах, или вследствие других причин, определяющих невозможность использования соответствующих параметрических критериев, применяются *критерий знаков* и *Т-критерий Уилкоксона*.

Критерий знаков (*sign test*) применяется для проверки гипотезы об однородности генеральных совокупностей по *попарно связанным выборкам*. Такая задача возникает, когда по одной и той же выборке сравниваются эффективность двух методов, точность измерения двумя приборами и т.д. Критерий знаков основан на подсчете количества раз, когда элементы одной выборки превышают парные элементы другой выборки ($x_a > x_b$) и наоборот ($x_b > x_a$). Нулевая гипотеза предполагает равенство этих случаев.

Более чувствительным, чем критерий знаков, при сравнении двух связанных выборок является *Т*-критерий Уилкоксона (*Wilcoxon matched pairs test*[18]). Его целесообразно использовать в случаях, когда неприменим *t*-критерий Стьюдента.

Критерий Уилкоксона основан на вычислении разности (d_i) измерений в каждой *i*-й паре элементов связанных выборок. Затем эти разности ранжируются от меньших к большим по модулю безотносительно знака. Если несколько разностей имеют одинаковое значение, им приписываются средние арифметические значения рангов. Нулевые значения не

[18]Другое название, встречающееся в литературе, – *Wilcoxon signed-rank test.*

учитываются. Затем подсчитываются суммы рангов положительных (T^+) и отрицательных разностей (T^-). Меньшая из сумм принимается в качестве эмпирического значения и сравнивается с критическим значением. Нулевая гипотеза отклоняется, если эмпирическое значение *меньше* критического.

Приведем пример. Предположим, исследуется эффект влияния известных торговых марок (*brand name*) на восприятие качества продукта. Десяти респондентам предлагается попробовать, сравнить и оценить в порядковой шкале (баллы) два вида алкогольных напитков, один из которых представляется как продукт известного производителя, другой – как продукт компании, недавно вышедшей на рынок. В действительности оба образца идентичны. Полученные данные и результаты их обработки представлены в таблице.

№	Баллы		d_i	Ранг d_i
	Брэнд	Новый продукт		
1	52	48	-4	-4
2	37	32	-5	-5,5
3	50	52	+2	+2
4	45	32	-13	-9
5	56	59	+3	+3
6	51	50	-1	-1
7	40	29	-11	-8
8	59	54	-5	-5,5
9	38	38	0	-
10	40	32	-8	-7

Сумма отрицательных разностей T^- = 40, сумма положительных разностей T^+ = 5. Следовательно, за эмпирическое значение критерия Уилкоксона принимается T = 5. Критическое значение находится из таблицы приложения 1.8 при α = 0,05 и n, равном количеству ненулевых разностей (для данного примера n = 10 – 1 = 9). Критическое значение критерия равно 6. В этом тесте эмпирическое значение должно быть *меньше* критического для отклонения нулевой гипотезы. Следовательно, нулевая гипотеза отклоняется, и делается содержательный вывод о статистически значимом влиянии брэндов на восприятие качества продукта.

В SPSS расчет критерия знаков и критерия Уилкоксона осуществляется с помощью процедуры *2 Related Samples* (из меню *Analyze, Nonparametric Tests*). В окне *Two-Related-Samples Tests* необходимо в поле *Test Pair(s) List* перенести переменные, соответствующие двум связанным выборкам, из списка переменных справа (для переноса необходимо выделить *две* переменные). В разделе *Test Type* нужно выбрать требуемый тест: критерий знаков (*Sign*) или критерий Уилкоксона (*Wilcoxon*). Нажатие на кнопку *OK* приводит к генерации выходной формы с отчетами.

Ranks

		N	Mean Rank	Sum of Ranks
NEW - BRAND	Negative Ranks	7[a]	5,71	40,00
	Positive Ranks	2[b]	2,50	5,00
	Ties	1[c]		
	Total	10		

a. NEW < BRAND

b. NEW > BRAND

c. BRAND = NEW

В отчете *Ranks* указаны количество отрицательных (7), положительных (2) и нулевых (1) разностей, средние ранги (*Mean Rank*) и суммы положительных и отрицательных разностей (*Sum of Ranks*) (сравните эти значения с определенными нами выше).

Test Statistics[b]

	NEW - BRAND
Z	-2,075[a]
Asymp. Sig. (2-tailed)	,038

a. Based on positive ranks.

b. Wilcoxon Signed Ranks Test

В отчете *Test Statistics* указан *p*-уровень для двусторонней проверки критерием Уилкоксона (*Asymp. Sig. (2-tailed)*). Он меньше выбранного уровня значимости ($\alpha = 0{,}05$), что позволяет отклонить нулевую гипотезу.

В отчетах *Frequencies* и *Test Statistics* приведены результаты использования критерия знаков для рассматриваемого примера. Как видим, этот критерий не обнаружил статистически значимых различий: *p*-уровень (*Exact Sig. (2-tailed)*) превышает $\alpha = 0{,}05$, что не позволяет отклонить нулевую гипотезу.

Frequencies

		N
NEW - BRAND	Negative Differences[a]	7
	Positive Differences[b]	2
	Ties[c]	1
	Total	10

a. NEW < BRAND

b. NEW > BRAND

c. BRAND = NEW

Test Statistics[b]

	NEW - BRAND
Exact Sig. (2-tailed)	,180[a]

a. Binomial distribution used.

b. Sign Test

8.6. Критерии Краскела-Уоллиса и χ^2 Фридмана

Критерий Краскела-Уоллиса (*Kruskal-Wallis test*) может использоваться как альтернатива однофакторному дисперсионному анализу, поэтому еще одним названием этого критерия является следующее: однофакторный дисперсионный анализ Краскела-Уоллиса (*Kruskal-Wallis one-way analysis of variance*). Он может применяться для трех и более независимых выборок с данными в порядковой или интервальной шкале, если не удовлетворяются предположения об использовании дисперсионного анализа.

Тест Краскела-Уоллиса является обобщением критерия Манна-Уитни для случая, когда исследуется более 2 выборок по уровню выраженности признака. Так же, как и критерий Манна-Уитни, тест Краскела-Уоллиса основан на положении о том, что чем меньше совпадений, тем больше различаются сравниваемые выборки. Как и в случае критерия Манна-Уитни, все значения сравниваемых выборок представляются в виде одной последовательности упорядоченных значений с последующем вычислением среднего ранга для каждой из выборок.

Эмпирическое значение критерия определяется как

$$H = \frac{12}{N(N+1)} \sum_{j=1}^{k} \frac{R_j^2}{n_j} - 3(N+1),$$

где R_j – сумма рангов для j-й выборки; n_j – объем j-й выборки; N – суммарная численность всех анализируемых выборок; k – количество анализируемых выборок.

Если объем каждой изучаемой выборки больше 5 или количество выборок больше 3, то эмпирическое значение критерия сравнивается с критическим (табличным) значением c^2 для d$f = k - 1$, иначе эмпирическое значение сравнивается с критическим значением критерия Краскела-Уоллиса, приведенным в специальных таблицах. Такие таблицы представлены во многих доступных учебных изданиях по статистике, в том числе в книгах, рекомендованных в разделе «Библиографический список» данного учебного пособия.

Проиллюстрируем использование критерия Краскела-Уоллиса на примере. Определяется, есть ли различие в количестве покупок, совершаемых ежедневно в трех магазинах, принадлежащих одной компании.

Магазин 1	Ранг	Магазин 2	Ранг	Магазин 3	Ранг
60	1	80	5	90	8,5
70	2,5	90	8,5	90	8,5
80	5	80	5	110	14
70	2,5	100	11,5	100	11,5
90	8,5	110	14	140	18
110	14	130	16,5	130	16,5
R_j	33,5		60,5		77

Подставив данные в формулу, получаем эмпирическое значение критерия H = 5,66. Критическое значение хи-квадрат для df = 2 и α = 0,05 – 5,99. Следовательно, нулевая гипотеза об отсутствии различий не может быть отклонена при заданном уровне значимости.

Для расчета критерия Краскела-Уоллиса в SPSS и STATISTICA данные должны быть сгруппированы в одну переменную, содержащую значения признака (для рассмотренного примера – переменная sale), и группирующую переменную, содержащую коды выборок (для данного примера – переменная group, включающая значения 1, 2, 3). В SPSS из меню *Analyze*, *Nonparametric Tests* выбирается процедура *K Independent Samples*. В открывшемся диалоговом окне *Tests for Several Independent Samples* необходимо в поле *Test Variable List* перенести зависимую переменную (sale), а в поле *Grouping Variable* – группирующую переменную (group). В разделе *Test Type* нужно отметить *Kruskal-Wallis H*, а в окне *Define Range* – указать минимальное и максимальное значения группирующей переменной (в данном случае – 1 и 3).

Результаты расчетов содержатся в отчетах *Ranks* и *Test Statistics*. Показатели, приведенные в отчетах, аналогичны тем, которые даны в отчетах других непараметрических тестов, которые были рассмотрены в предыдущих параграфах, поэтому на них мы останавливаться не будем. Отметим лишь, что эмпирическое значение критерия, рассчитанное в SPSS (5,765), несколько выше, чем рассчитанное вручную (5,66). Это связано с тем, что в пакете используется дополнительная корректировка критерия в случае, если несколько элементов выборки имеют один ранг. Тем не менее полученное значение *p*-уровня (0,056) также не позволяет отклонить нулевую гипотезу.

Ranks

	GROUP	N	Mean Rank
SALE	1,00	6	5,58
	2,00	6	10,08
	3,00	6	12,83
	Total	18	

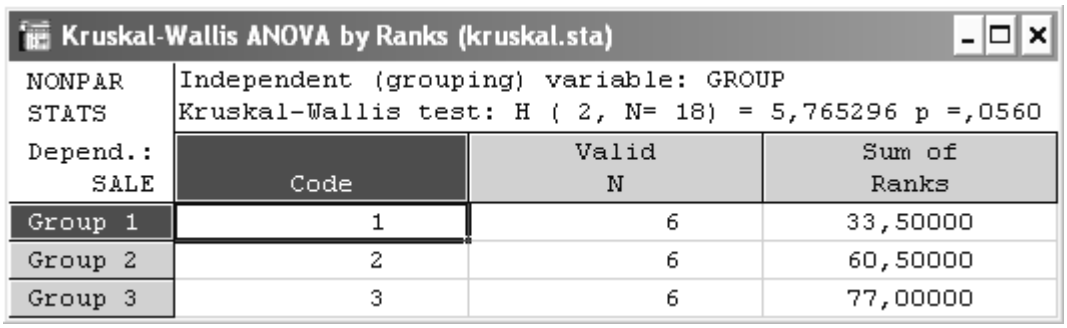
Test Statistics[a,b]

	SALE
Chi-Square	5,765
df	2
Asymp. Sig.	,056

a. Kruskal Wallis Test

b. Grouping Variable: GROUP

Аналогичные вычисления в пакете STATISTICA осуществляются в модуле *Nonparametrics/Distrib.* с помощью процедуры *Kruskal-Wallis ANOVA* (выбирается из меню *Analysis, Startup Panel*). Работа с этой процедурой аналогична работе с процедурой *Mann-Whitney U test.* Результаты идентичны результатам, полученным в SPSS.



Kruskal-Wallis ANOVA by Ranks (kruskal.sta)

NONPAR STATS

Independent (grouping) variable: GROUP
Kruskal-Wallis test: H (2, N= 18) = 5,765296 p =,0560

Depend.: SALE	Code	Valid N	Sum of Ranks
Group 1	1	6	33,50000
Group 2	2	6	60,50000
Group 3	3	6	77,00000

Как и в случае использования ANOVA, при отклонении нулевой гипотезы по критерию Краскела-Уоллиса нельзя сделать конкретный вывод о направлении различий. В такой ситуации можно использовать критерий Манна-Уитни для парного сравнения выборок.

Для сравнения *более двух зависимых* выборок используется критерий *хи-квадрат Фридмана* (*Friedman test*), который является непараметрическим аналогом однофакторного дисперсионного анализа с повторными измерениями. Этот критерий может быть даже более эффективен, чем ANOVA, при небольших выборках. Работа с критерием хи-квадрат Фридмана в статистических пакетах аналогична работе с рассмотренными выше методами. В SPSS эта процедура носит название *K Related Samples*, в STATISTICA – *Friedman ANOVA*. В случае отклонения нулевой гипотезы по критерию хи-квадрат Фридмана для парного сравнения выборок используется *T*-критерий Уилкоксона.

Глава 9. Меры связи и их проверка на значимость

Статистические критерии проверки гипотез позволяют выявить наличие статистически значимых связей, но не оценить их степень (величину). Поэтому в исследовательской практике очень часто определяется мера связи между двумя или несколькими признаками. Для этого, например, вычисляются коэффициент корреляции, коэффициент детерминации, строятся регрессионные уравнения разного вида и т.д. Достаточно высокое значение коэффициента корреляции рассматривается многими исследователями как индикатор сильной связи, что как будто предполагает ее безусловное присутствие. Однако, с другой стороны, для корректного анализа необходимо не только определить вид, степень и направление связи, но и проверить статистическую значимость полученного результата, т.е. проверить на значимость собственно найденные коэффициенты связи. Иными словами, требуется узнать, существует ли связь, обнаруженная в выборке и характеризуемая коэффициентом корреляции, для всей генеральной совокупности, из которой эта выборка была извлечена.

Проверяя любую статистическую гипотезу, исследователь ищет ответ на вопрос, существуют ли в генеральной совокупности те характерные особенности, которые наблюдаются в выборках. Проверке на значимость могут (и должны) быть подвергнуты и меры связи. Это абсолютно необходимая процедура с точки зрения научной строгости анализа и корректности представления исследовательских результатов. При выборочном исследовании всегда существует ошибка получения того или иного результата, в том числе коэффициента корреляции. Поэтому нужно оценивать и показывать эту возможную ошибку, т.е. уровень статистической значимости результата. В общем случае, чем сильнее связь, тем менее вероятна ошибка, но, тем не менее, определив меру связи, необходимо всегда указывать и уровень значимости, например $p < 0{,}01$.

Согласно общей логике проверки статистических гипотез, в данном случае нулевая гипотеза утверждает, что для генеральной совокупности, из которой была извлечена выборка, связь между переменными полностью отсутствует, т.е. значение проверяемого показателя меры связи для генеральной совокупности равно нулю. Дальнейшая логика проверки на статистическую значимость аналогична общей последовательности этапов проверки любой статистической гипотезы.

В зависимости от вида шкал, в которых измерены эмпирические данные, используются разные методы определения степени и направления связи. Например, для номинативных шкал в качестве меры связи вычисляют коэффициент φ (фи) и коэффициент *V* Крамера, для

порядковых – коэффициент γ (гамма), коэффициент ρ-Спирмена, коэффициент τ-Кендалла. Для метрических шкал вычисляется коэффициент корреляции (*r*) Пирсона. В качестве критерия для проверки значимости используется *t*-критерий (для ρ и *r*) или *z*-критерий (для γ и τ).

8.1. Меры связи для переменных в метрических шкалах

Рассмотрим подробнее процедуру проверки на значимость коэффициента корреляции Пирсона между двумя признаками *x* и *y*, представленными в метрической шкале. Это один из наиболее широко используемых статистических показателей. *Коэффициент корреляции Пирсона* вычисляется по формуле

$$r_{xy} = \frac{\sum \left(x - \bar{x}\right) \cdot \left(y - \bar{y}\right)}{(n-1) s_x s_y},$$

где $\bar{x}, \bar{y}$ – средние значения признаков; s_x, s_y – стандартные отклонения.

Коэффициент корреляции показывает, ассоциированы ли наборы данных по величине, т.е. большие значения из одного набора данных связаны с большими значениями другого набора (положительная корреляция) или, наоборот, малые значения одного набора связаны с большими значениями другого (отрицательная корреляция), или данные двух диапазонов никак не связаны (корреляция близка к нулю). По абсолютной величине коэффициента корреляции можно судить о степени связи между признаками:

Степень тесноты связи	Коэффициент корреляции
отсутствует	0 – 0,2
слабая	0,2 – 0,4
средняя (умеренная)	0,4 – 0,7
сильная	0,7 – 1

Коэффициент корреляции, возведенный в квадрат, называется *коэффициентом детерминации* и показывает, в какой степени (на сколько процентов) изменчивость одного признака определяется (детерминируется) изменчивостью другого. Или, другими словами, насколько результативный признак зависит от анализируемого признака. Коэффициент детерминации, в отличие от коэффициента корреляции, отражает связь между признаками линейно, он является аддитивным, т.е. допускается его усреднение для нескольких выборок.

Для определения коэффициента корреляции и выполнения проверки на значимость требуется принять ряд допущений. Предполагается следующее: во-первых, интервальный (или относительный) характер измеряемых данных; во-вторых, обе переменные подчиняются

нормальному закону распределения; в-третьих, зависимость между переменными приблизительно линейна; в-четвертых, *гомоскедастичность*, т.е. однородность дисперсий. Это понятие означает, что дисперсия значений *y* равномерна для всех значений *x*. На графике в осях *x*–*y* это должно проявляться следующим образом: значения *y* должны быть приблизительно равномерно распределены выше и ниже линии регрессии *y(x)* по всей ее длине, т.е. точки на графике должны образовывать достаточно симметричную овалообразную форму без значительных выбросов. Существенные признаки нелинейности или отклонения от гомоскедастичности свидетельствуют о необходимости использовать другую меру связи и другой критерий значимости.

Предположим, что все допущения выполняются и коэффициент корреляции определен. Тогда нулевая гипотеза H_0: $r_{\text{популяции}} = 0$ утверждает отсутствие связи для генеральной совокупности; H_A: $r_{\text{популяции}} \neq 0$ утверждает наличие этой связи.

Для оценки статистической значимости коэффициента корреляции (r) по выборке из n элементов используется следующая формула для определения эмпирического значения t-критерия:

$$t = r\sqrt{\frac{n-2}{1-r^2}}.$$

Критическое значение t-критерия вычисляется с помощью функции Excel СТЬЮДРАСПОБР(α, df) или определяется из таблицы значений t-распределения для выбранного уровня значимости α и числа степеней свободы $df = n - 2$ (приложение 1.2). Затем на основе сравнения критического и экспериментального значений принимается решение об отклонении или невозможности отклонения нулевой гипотезы, следуя тем же правилам, что и для других критериев, рассмотренных выше.

Аналогично можно оценить значимость коэффициента частной корреляции. Это необходимо тогда, когда изучается связь между тремя переменными *x, y, z* и нужно оценить связь между, например, *x* и *y* без учета влияния *z*. Коэффициент частной корреляции вычисляется по формуле

$$r_{xy-z} = \frac{r_{xy} - r_{xz}r_{yz}}{\sqrt{(1-r_{xz}^2)(1-r_{yz}^2)}}.$$

Нулевая гипотеза H_0: $r_{xy\text{-}z} = 0$. Эмпирическое значение t-критерия для проверки этой гипотезы определяется как

$$t = \frac{r_{xy-z}\sqrt{n-3}}{\sqrt{1-r_{xy-z}^2}}.$$

Критическое значение t-критерия определяется для выбранного уровня значимости α и числа степеней свободы $df = n - 3$.

9.2. Меры связи для переменных в неметрических шкалах

В случае, если хотя бы одна из переменных представлена в порядковой шкале, могут использоваться коэффициенты ρ-Спирмена, гамма и τ-Кендалла. *Коэффициент ρ-Спирмена* интерпретируется так же, как и коэффициент r Пирсона (в терминах объяснения изменчивости одной переменной от изменчивости другой), и может принимать значения в таком же диапазоне (от –1 до +1). Фактически это и есть коэффициент корреляции Пирсона, вычисленный по рангам переменных (после предварительного ранжирования). При отсутствии повторяющихся рангов коэффициент ρ-Спирмена определяется по формуле

$$r = \frac{6\sum_i d_i^2}{n(n^2-1)},$$

где d_i – разность рангов двух признаков i-го элемента выборки.

Проверка статистической значимости коэффициента ρ-Спирмена ничем не отличается от аналогичной процедуры для коэффициента r Пирсона (при числе наблюдений в выборке 10 и больше распределение выборочных значений ρ-Спирмена приближается к t-распределению, поэтому для проверки используется t-критерий).

Коэффициент гамма (γ) основан на проверке различия между вероятностью того, что наблюдаемые данные расположены в том же самом порядке для двух величин (*совпадение*), и вероятностью того, что они расположены в другом порядке (*инверсия*). Для вычисления этого коэффициента между элементами двух выборок попарно определяется число совпадений (P) и инверсий (Q).

Коэффициент гамма вычисляется по формуле

$$g = \frac{P-Q}{P+Q}.$$

Коэффициент τ-Кендалла является обобщением коэффициента гамма на случай возникновения одинаковых рангов (*tied ranks*) у элементов. Формула для вычисления этого коэффициента включает в себя поправки к коэффициенту гамма, но принцип расчета остается тем же. Обычно значения коэффициентов Спирмена и Кендалла различаются, так как у них разные принцип и процедура вычисления. Соотношение между этими коэффициентами выражается неравенством:

$$-1 <= 3\tau - 2\rho <= 1.$$

Значения коэффициентов Пирсона, Спирмена и Кендалла могут быть определены с помощью статистических пакетов. Для этого, например, в SPSS из меню ***Analyze*** выбирается процедура ***Correlate***. Затем для расчета парной корреляции выбирается опция ***Bivariate***, где указывается пара переменных для расчета и выбирается нужный коэффициент, а также вид проверки на значимость (одно- или двунаправленная). Для расчета частной корреляции выбирается опция ***Partial***.

Для оценки силы связи между двумя переменными в номинативных шкалах наиболее часто используется *коэффициент φ (фи)*. Его вычисление основано на статистике хи-квадрат для таблиц 2 x 2:

$$j = \sqrt{\frac{c^2}{n}},$$

где n – общее количество наблюдений.

Значения коэффициента фи изменяются от 0 (нет зависимости между переменными; хи-квадрат = 0) до 1 (абсолютная зависимость между двумя факторами в таблице). В случае таблиц большего размера верхний предел коэффициента фи может превысить 1, тогда интерпретация результатов затрудняется. Поэтому используется более общий показатель, который получил название *коэффициент V Крамера* (*Cramer's V*):

$$V = \sqrt{\frac{c^2}{n(k-1)}}.$$

В этой формуле k означает наименьшее значение из числа строк и столбцов в таблице сопряженности.

Коэффициенты фи и *V* Крамера вычисляются в SPSS с помощью процедуры ***Crosstabs*** (из меню ***Analyze, Descriptive Statistics***). После выбора необходимых переменных для формирования строк и столбцов таблицы сопряженности (как это осуществляется при расчете критерия хи-квадрат) необходимо в окне ***Statistics*** выбрать необходимые коэффициенты для расчета. В выходной форме представляются значения коэффициентов и *p*-уровень значимости.

В частности, для примера о связи курения и брака на производстве, рассмотренного в главе 8, имеем следующие результаты.

Symmetric Measures

		Value	Approx. Sig.
Nominal by Nominal	Phi	,322	,032
	Cramer's V	,322	,032
N of Valid Cases		66	

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

Эти результаты демонстрируют слабую значимую связь между двумя изучаемыми признаками при $p < 0{,}05$.

Вопросы, задания и упражнения (часть 2)

Задания по темам второй части книги предлагаются в одном разделе, без разбивки по главам. Это сделано для того, чтобы читатели могли самостоятельно решить, какие именно статистические критерии следует использовать при выполнении тех или иных заданий, поскольку выбор метода, применение которого было бы корректным в той или иной исследовательской ситуации, и составляет наибольшую сложность (вычислительные процедуры благодаря статистическим пакетам обычно больших трудностей не вызывают). Задачи 19–22 являются комплексными, они требуют использования материалов как второй, так и первой частей учебного пособия.

1. Для каждой ситуации определите z критическое.

Альфа	*Вид альтернативы*	*z критическое*
0,05	Односторонняя	
0,10	Двусторонняя	
0,06	Двусторонняя	
0,01	Односторонняя	
0,02	Двусторонняя	

2. Для каждой ситуации определите t критическое.

Альфа	*Вид альтернативы*	*df*	*t критическое*
0,10	Односторонняя	20	
0,02	Двусторонняя	30	
0,01	Двусторонняя	30	
0,01	Односторонняя	25	
0,05	Двусторонняя	15	

3. а) Студенты университета посещают в среднем 3,3 клубных вечеринок в месяц. Случайная выборка, состоящая из 97 студентов-социологов, посещает в среднем 3,8 вечеринок в месяц со стандартным отклонением 0,53. Можно ли сделать вывод, что студенты-социологи отличаются от студентов в целом в пристрастиях к посещению клубов?

б) Если вопрос исследования поменять на вопрос «Посещают ли студенты-социологи значительно большее количество вечеринок?», что изменится в процедуре проверки? Изменятся ли результаты?

4. Средний балл по результатам государственного экзамена по стране составляет 453. При поступлении на один из факультетов университета города N (142 абитуриента), средний балл составил 520 со стандартным отклонением 85. Является ли статистически значимым это различие? Можно ли сказать, что на этот факультет поступают более подготовленные выпускники школ? Принять допущение об интервальности шкалы.

5. Средняя зарплата 95 работников филиала зарубежной компании 24375 руб./месяц. Средняя зарплата всех работников этой компании составляет 22230 руб./мес. со стандартным отклонением 523 руб. Является ли статистически значимым это различие? Можно ли сказать, что работники филиала получают больше?

6. По данным УВД области, раскрывается 55% случаев грабежа и 62% случаев разбоя. В городе зафиксировано 207 случаев грабежа и 178 случаев разбоя в течение года. Было раскрыто 103 случая грабежа и 90 случаев разбоя. Можно ли сказать, что имеют место значимые различия между раскрываемостью преступлений по городу и области?

7. Можно ли сказать, что иногородние студенты, живущие в общежитиях, больше вовлечены в жизнь университета, чем городские студенты? Ниже приведены данные по двум выборкам о среднем количестве часов в неделю, которое студенты посвящают внеурочной деятельности. Являются ли значимыми эти различия?

Иногородние студенты	Городские студенты
$\bar{x}$ = 12,4	$\bar{x}$ = 10,2
s = 2,0	s = 1,9
n = 48	n = 33

8. Компания по производству лакокрасочных изделий разработала новый вид краски для внешних (уличных) покрытий. Прежде чем выпускать новую продукцию на рынок, компания намерена проверить устойчивость новой краски к внешним воздействиям. Компания обладает соответствующим тестирующим оборудованием, которое воздействует на окрашенную поверхность специальными щетками и фиксирует количество воздействий, необходимых для полного стирания краски. С помощью этого оборудования было проведено пять испытаний, в ходе которых сравнивалось качество краски, разработанной компанией, с аналогичными продуктами, предлагаемыми двумя конкурентами. Результаты (в тыс. воздействий) представлены в таблице.

№ испытания	Продукт компании	Продукт конкурента А	Продукт конкурента Б
1	37	34	24
2	30	19	25
3	34	22	23
4	28	31	20
5	29	27	20

Различается ли оцениваемое качество продуктов? Проверить гипотезу о различии средних на уровне α = 0,05.

9. На основании рейтингов двух авиакомпаний (см. таблицу) проверить гипотезу о различиях средних на уровне α = 0,05. Данные представлены в интервальной шкале.

Авиакомпания 1					Авиакомпания 2			
№	Балл	№	Балл		№	Балл	№	Балл
1	67	11	65		1	95	11	92
2	58	12	80		2	60	12	85
3	78	13	79		3	70	13	94
4	68	14	88		4	78	14	93
5	69	15	89		5	65	15	90
6	74	16	72		6	70	16	67
7	55	17	58		7	79	17	85
8	55	18	67		8	81	18	95
9	80	19	85		9	69	19	67
10	78	20	80		10	78	20	83

10. Решить предыдущую задачу при условии, что шкала не интервальная, а порядковая.

11. Производитель компьютерной и оргтехники разработал новый продукт, позиционируемый преимущественно на использование в «домашних» условиях, а не в офисе. Организованы тестовые продажи продукта в магазинах, торгующих бытовой электроникой, универмагах и специализированных компьютерных магазинах (случайно отобрано по 5 торговых точек каждой категории). Количество покупок продукта в магазинах каждого типа приведено в таблице.

Тип магазина	Точки продаж				
	1	*2*	*3*	*4*	*5*
Магазины электроники	5	4	3	3	3

Универмаги	9	7	8	6	5
Компьютерные магазины	7	4	8	4	3

Необходимо определить, в магазинах какого типа следует организовать основные продажи. Проверить гипотезу о различиях средних на уровне $\alpha = 0{,}05$.

12. Специалист в области финансового консалтинга анализирует изменение ситуации на национальном рынке услуг связи за определенный период времени. Ему известны данные о цене акций 11 компаний, работающих на этом рынке, на соответствующие даты. Выборка компаний осуществлена случайным образом.

№ компании	*Дата1*	*Дата2*
1	218,9	361,0
2	79,4	82,3
3	99,1	125,2
4	49,0	84,7
5	171,5	157,8
6	224,8	203,2
7	105,0	124,9
8	446,8	465,6
9	120,7	122,6
10	128,2	117,7
11	124,9	148,8

Необходимо проверить, увеличилась ли средняя цена акций компаний за анализируемый период. Принять $\alpha = 0{,}05$.

13. Автодилер исследует предпочтения покупателей, в частности связана ли повторная покупка автомобиля со страной производителя (Страна 1 и Страна 2). Результаты исследований (количество покупок) приведены в таблице. При проверке гипотезы принять $\alpha = 0{,}05$.

	Производитель	
	Страна 1	*Страна 2*
Повторная покупка	40	20
Первая покупка	8	32

14. Ответьте на вопросы:

а) что общего и различного между параметрическими и непараметрическими тестами?

б) какие общие предпосылки лежат в основе параметрических тестов?

в) какие тесты можно назвать «независимыми от распределения» и почему?

г) почему невозможность отклонения нулевой гипотезы не означает ее принятие?

д) когда риск совершения ошибки первого рода является более критичным, чем риск совершения ошибки второго рода (приведите примеры)?

е) когда риск совершения ошибки второго рода является более критичным, чем риск совершения ошибки первого рода (приведите примеры)?

ж) каковы пути снижения вероятности ошибки первого рода?

з) каковы пути снижения вероятности ошибки второго рода?

и) в чем различие статистической значимости и практической значимости? Приведите примеры.

к) каким образом могут быть проверены все предпосылки для применения дисперсионного анализа?

15. Какие методы проверки статистических гипотез могут быть корректно применены в следующих ситуациях?

А. Проверка зависимости карьерного роста сотрудников от пола (n = 200).

Б. Партия уплотнителей проверяется на соответствие техническим требованиям по толщине, которая должна составлять 2 мм (n = 25).

В. Проверяется различие общего расходования горючего в двух службах такси, принадлежащих одному владельцу (отобрано по 50 автомобилей от каждой компании).

Г. Оценивается результативность тренинговой программы «эффективные продажи». Компания имеет 20 торговых представителей, принимавших участие в тренинге. Изменились ли объемы продаж после проведения тренинга?

Д. Компания предполагает определить, связан ли объем продаж с повышением качества предлагаемой продукции, затратами на рекламу или совместным воздействием качества продукции и затрат на рекламу. Зависимая переменная – объем продаж, независимыми переменными (влияющими факторами) являются качество продукции и объем рекламы.

Е. Для оценки успешности карьерного старта выпускников двух ведущих университетов города проводится выборочный опрос о средней заработной плате

выпускников этих университетов, получивших диплом год назад. Количество опрошенных выпускников одного университета – 50 чел., другого – 70 чел.

16. Определить, существует ли статистически значимое различие между студентами и студентками в их предпочтениях по посещению университетского клуба (по данным, приведенным в таблице).

	Посещают	*Не посещают*	**Всего**
Мужчины	26	24	50
Женщины	35	15	50
Всего	61	39	100

17. Необходимо оценить эффективность, с которой каждая из трех социальных служб (агентств) реализует некоторую программу. В частности, заказчика этого исследования интересует скорость, с которой эти агентства работают с документами, определяя, подходит ли потенциальный клиент под условия данной программы. Для выполнения оценки работы агентств была собрана информация о количестве дней, требуемых для обработки документов. Информация была получена на основе формирования случайной выборки из 10 наблюдений для каждого агентства. Имеющиеся данные приведены в таблице (в днях).

№	*Агентство 1*	*Агентство 2*	*Агентство 3*
1	5	12	9
2	7	10	8
3	8	19	12
4	10	20	15
5	4	12	20
6	9	11	21
7	6	13	20
8	9	14	19
9	6	10	15
10	6	9	11

Существует ли значимое различие между тремя агентствами по скорости обработки документов? Принять уровень значимости $\alpha = 0{,}05$.

18. Центр социально-политических исследований проводит оценку различий активности избирателей при проведении выборов разного уровня (выборы депутатов в органы законодательной власти региона, выборы депутатов Государственной Думы и выборы Президента страны). Случайная выборка из избирательных участков показывает процент явки избирателей в зависимости от типа выборов.

Региональные выборы	*Выборы в Думу*	*Выборы Президента*
33	35	42
78	56	40
32	35	52
28	40	66
10	45	78
12	42	62
61	65	57
28	62	75
29	25	72
45	47	51
44	52	69
41	55	59

Изменяется ли значимо процент явки избирателей в зависимости от типа выборов?

19. По данным, приведенным в приложении 2.1, проверить, различается ли организационный климат в двух компаниях по показателям. Например, различается ли значение показателя 1 в двух компаниях, и т.д. (проверить различие каждого из 7 показателей). В каком направлении имеются различия?

А. Решить задачу при предположении, что данные представлены в шкале Лайкерта.

Б. Решить задачу при допущении о том, что данные представлены в интервальной шкале.

В. Решить задачу после преобразования данных в шкалу Раша.

Сравнить результаты А, Б, В.

20. Для оценки надежности-стабильности нового инструмента для оценки когнитивных стилей (И1) использовался метод «тест-ретест» с интервалом в 1 мес. (объем выборки –28 респондентов). Данные тестирования представлены в таблице.

Респондент	*Тест*	*Ретест*	*Респондент*	*Тест*	*Ретест*
1	84	98	15	97	95
2	88	93	16	102	98
3	88	92	17	102	95
4	89	94	18	103	97
5	89	77	19	105	99
6	90	89	20	107	112
7	93	91	21	107	104
8	93	91	22	107	97
9	95	111	23	109	109
10	95	99	24	111	114
11	95	82	25	111	115
12	96	106	26	115	112
13	96	95	27	116	115
14	97	86	28	130	129

Оценить надежность инструмента при предположении, что а) шкала порядковая; б) шкала интервальная. Сравнить результаты.

21. Для оценки надежности-стабильности другого инструмента (И2) также использовался метод «тест-ретест» (объем выборки – 34 респондента). Этот инструмент выделяет 4 разных доминирующих психологических типа (возможно сочетание нескольких типов). Данные тестирования приведены в таблице. Значение 1 соответствует наличию определенного типа, значение 0 означает несоответствие этому типу.

	Тест				**Ретест**			
Респондент	*Tun1*	*Tun2*	*Tun3*	*Tun4*	*Tun1*	*Tun2*	*Tun3*	*Tun4*
1	0	0	0	1	1	0	0	1
2	0	1	1	1	0	0	0	1
3	1	0	0	0	1	0	0	0
4	1	1	1	0	1	1	1	0
5	1	0	1	0	1	0	0	0
6	0	1	0	1	0	1	0	1
7	0	0	1	1	1	0	1	0
8	1	0	0	0	1	0	1	0
9	1	0	1	1	1	0	0	1
10	1	0	1	1	0	0	1	1
11	0	1	1	1	0	1	1	0
12	0	1	0	0	0	1	0	0
13	1	0	1	1	1	0	1	1
14	0	1	0	1	0	0	0	0
15	0	0	1	0	0	0	1	0
16	1	1	1	1	1	1	1	1
17	1	0	1	0	1	0	0	0
18	0	0	0	1	0	0	0	1
19	1	0	1	1	1	0	1	0
20	1	0	0	0	1	0	1	0
21	1	0	1	0	1	0	1	0
22	1	1	0	0	0	1	0	0
23	1	0	1	0	0	0	1	0
24	0	0	0	1	0	0	0	1
25	1	0	1	1	1	0	0	0
26	1	0	1	1	1	1	1	1
27	1	0	1	1	1	0	0	0
28	0	1	0	1	0	0	0	1
29	0	0	0	1	0	0	0	1
30	0	0	1	0	1	1	1	1
31	0	0	0	1	0	0	0	1
32	1	1	0	1	1	1	1	1
33	0	1	0	1	0	1	0	1
34	1	0	1	1	0	0	0	1

Оценить надежность инструмента.

22. Сопоставлялись результаты двух инструментов И1 и И2 (см. две предыдущие задачи). Данные группировки результатов И1 по 4 типам, выделенным с помощью И2, приведены в таблице.

Значения по И1			
Tun1	*Tun2*	*Tun3*	*Tun4*
96	105	105	84
93	130	93	96
90	95	107	105
107	111	94	95
96	109	109	95
107	116	103	107
109	120		88
103			95
97			116
95			89
98			111
			98
			120

В частности, проверялась гипотеза о том, что люди разных психологических типов (оцениваемых И2) характеризуются (в среднем) различными значениями по шкале И1. Проверить эту гипотезу при предположении, что а) шкала И1 порядковая; б) шкала И1 интервальная. Сравнить результаты.

Комментарии к заданиям

В данном разделе приводятся комментарии к некоторым заданиям, упражнениям и вопросам для того, чтобы облегчить читателям книги самостоятельное освоение материалов учебного пособия. Также для ряда задач приводятся ответы, чтобы читатель мог проверить собственное решение. Рассматриваются преимущественно те задания, по которым у студентов и слушателей учебных курсов по темам этой книги возникает наибольшее количество вопросов.

Индикативное задание

Одним из главных вопросов индикативного задания является вопрос о репрезентативности выборки. Этот вопрос, несмотря на кажущуюся простоту, является важным, сложным и отчасти – провокационным. Дело в том, что далеко не все авторы эмпирических социально-экономических исследований уделяют внимание обоснованию репрезентативности используемых в исследовании выборочных совокупностей. Более того, из сведений, приводимых авторами изысканий в публикациях разного уровня, подчас невозможно судить о репрезентативности. Следовательно, результаты таких исследований могут быть поставлены под сомнение. Вопросу о репрезентативности и ее обоснованию посвящена в значительной степени вторая глава книги.

Часть 1

Глава 1

3. Для выполнения задания можно рассмотреть, например, такие характеристики, как:

а) вес, объем, цвет, запах, цена, способность удалять загрязнения, воздействие на цвет стираемых вещей, скорость отстирывания, воздействие на элементы стиральной машины и т.д.;

б) возраст, пол, цвет волос, семейный статус, мотивация, ответственность, квалификация, производительность труда и т.д.;

в) торговый оборот, суммарные затраты, сезонность продаж, наиболее успешно продаваемые товары и многое другое.

4. Это задание вызывает сложности в части представления заработной платы в интервальной шкале. Действительно, для денежных величин невозможно указать «чисто»

интервальную шкалу. Но можно предложить, например, некое подобие интервальной шкалы («эрзац»-интервальную шкалу и для такой измеряемой величины, как заработная плата). Предположим, мы указываем зарплату в определенных диапазонах и отмечаем количество сотрудников в каждом диапазоне, например:

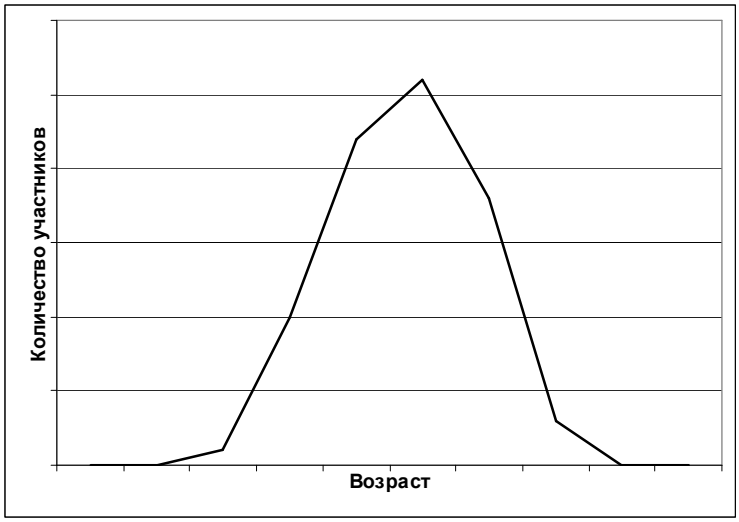
Заработная плата, руб.	Кол-во сотрудников
0–2999	17
3000–5999	39
6000–8999	27

Мы можем взять средние точки в каждом диапазоне, которые примем за позиции шкалы (1500, 4500, 7500), и интервалы между этими позициями будут, очевидно, равны. Хотя это «не настоящая» интервальная шкала, так как нулевая точка не условная, однако мы видим, что здесь присутствует потеря информации за счет осреднения. То есть в этом случае нельзя точно сказать, что конкретный сотрудник получает столько-то, можно сказать лишь приблизительно («в среднем»).

Для представления заработной платы в порядковой шкале можно, например, ранжировать список сотрудников в зависимости от их заработной платы. Для представления заработной платы в номинативной шкале можно, допустим, дифференцировать сотрудников с повременной и фиксированной оплатой.

Глава 3

1. Согласно центральной предельной теореме, распределение средних значений выборок, извлекаемых из одной и той же совокупности, при достаточно большом числе таких выборок соответствует нормальному распределению. При выделении 10 или более выборок по 10 или более элементам может быть построен полигон частот, например следующего вида (см. рисунок), на основании которого можно судить о приблизительном соответствии нормальному распределению.



Для того чтобы проверить утверждения центральной предельной теоремы относительно среднего значения и стандартного отклонения выборочных средних, можно построить таблицу следующего вида и сравнить результаты.

Показатель	По выборочным средним	Для генеральной совокупности
Среднее		
Стандартное отклонение		

Данные из списка рекомендуется ввести в два столбца: идентификационный номер (номер по порядку) и возраст. Осуществлять случайный отбор из списка можно с помощью функции Excel СЛЧИС(). Необходимо учитывать диапазон генерируемых чисел (1–100). Для этого используем выражение СЛЧИС()*99+1. В результате может получиться дробное число. Для определения номера студента из списка это число нужно округлять. Для этого можно воспользоваться функцией Excel ОКРУГЛ(). Для данной задачи округление производится до целого, поэтому можно использовать выражение =ОКРУГЛ(СЛЧИС()*99+1;0), которое помещается, например, в ячейку С1.

Далее для автоматизации процесса выбора элементов из списка, чтобы не искать вручную, можно воспользоваться функцией определения элемента в таблице по значению (функция вертикального переноса – ВПР(значение;массив;номер столбца)). В качестве первого аргумента функции можно указать ОКРУГЛ(СЛЧИС()*99+1;0), массив – это набор идентификационных номеров студентов по списку и их возраст (например, $A$1:$B$100; номер столбца для данного случая – 2). В итоге получаем: =ВПР(С1;$A$1:$B$100;2).

Эта формула копируется в 10 ячеек (если формируется выборка из 10 элементов) и определяется среднее значение =СРЗНАЧ (массив из 10 ячеек). Нажимая F9, можно изменять выборку и получать всякий раз различное среднее значение по выборке из 10 элементов.

Для построения графика можно разбить диапазон возраста на равные промежутки, например 10 интервалов ((максимальный возраст – минимальный возраст) / 10), или выбрать фиксированную длину интервала, например 0,5. Затем подсчитывается частота попадания определяемого выборочного среднего в интервал. Чем больше случайных выборок Вы сделаете и чем больше интервалов возьмете, тем ближе полигон частот будет напоминать нормальное распределение.

2. В данном случае среднее значение измеряемого параметра должно рассчитываться не как *среднее арифметическое* (сумма значений признаков объектов, деленная на количество объектов), а как *среднее арифметическое взвешенное*, поскольку данные сгруппированы[19]. Число одинаковых значений признака в ряду распределения называется *частотой* и обычно обозначается *f*.

Среднее арифметическое взвешенное определяется как

$$m = \frac{\sum x_i f_i}{\sum f_i},$$

а *стандартное отклонение взвешенное* – как

$$s = \sqrt{\frac{\sum (x_i - m)^2 f_i}{\sum f_i}}.$$

10, 11. Для решения этих задач напрямую используется формула для вычисления необходимого объема выборки при определении доли. Однако среднегрупповая дисперсия (среднее значение $\overline{pq}$ по группам) приведена с учетом того, что доли представляются в процентах. Поэтому при решении в формулу в качестве ошибки нужно подставлять число 5, а не 0,05.

Глава 4

4а) Валидный (обоснованный) измерительный инструмент всегда является надежным, но не всякий надежный инструмент является обоснованным (в смысле содержательной, эмпирической и концептной валидности).

4б) В большинстве случаев это не так. Содержательная обоснованность оценивается на основе квалификации, опыта, знания предмета (исследователя, экспертов или участников

[19] Читатели могут найти подробное изложение разных подходов к определению среднего (в том числе среднего арифметического, среднего арифметического взвешенного, среднего гармонического, среднего геометрического и др.) в учебной литературе по математической статистике.

фокус-группы). И если исследователь, эксперт или участники фокус-группы делают обоснованное заключение о содержательной валидности, то это, как правило, считается достаточным. Эмпирическая или концептная (дискриминантная) валидность может потребовать статистических сравнений, эмпирических оценок и дополнительных обоснований.

3в) Хотя оба термина характеризуют воспроизводимость измерений, стабильность связана с устойчивостью измерений во времени и в разных ситуациях (проверяется методом повторного тестирования тем же инструментом в той же группе), а эквивалентность – с устойчивостью измерений к изменению формы представления инструмента и с сопоставимостью результатов измерения с «каноническими» результатами.

6. Вероятно, вопросы анкеты не относятся к одной характеристике (безопасности), т.е. данным инструментом измеряется не один концепт, а сразу несколько. Необходимо удалить плохо согласованные вопросы и добавить другие.

Часть 2

1. Критическое значение z можно посчитать в Excel с помощью функции НОРМСТОБР(P), возвращающей значение для *одностороннего* критерия z при заданном $P = 1 - \alpha$. Для двустороннего критерия следует в качестве аргумента функции подставлять $P = 1 - \alpha / 2$. Например, для $\alpha = 0{,}06$ для одностороннего критерия имеем $P = 0{,}94$ и $z = 1{,}55$, для двустороннего – $P = 0{,}97$ и $z = 1{,}88$.

8. Для решения задачи можно использовать однофакторный дисперсионный анализ для k независимых выборок. Предварительно необходимо проверить выполнение предпосылок для применения этого метода (однородность дисперсий в выборках и нормальность распределения). Первая предпосылка подтверждается тестом Левена, вторая – Q-Q графиками для каждой выборки и тестом Шапиро-Уилка. Критическое значение F-критерия: 3,89 при $df = (2; 12)$ и $\alpha = 0{,}05$; эмпирическое значение F-критерия: 5,5. Нулевая гипотеза отклоняется, $p = 0{,}02$. Далее необходимо использовать апостериорные сравнения (метод Шеффе). Статистически значимое различие обнаружено между компанией и конкурентом Б.

9. В данном случае имеем две независимые выборки, при этом предполагается интервальный характер шкалы, в которой измерен рейтинг компаний. Поэтому для проверки гипотезы используется t-критерий Стьюдента. Критическое значение – 2,03 ($df = 38$; $\alpha = 0{,}05$). Эмпирическое значение – 2,19. Нулевая гипотеза об отсутствии различий отклоняется.

11. Поскольку данные представлены в шкале отношений и имеется k независимых выборок, может быть использован однофакторный дисперсионный анализ (при условии выполнения остальных необходимых предпосылок). F критическое: 3,89 при df = (2; 12) и α = 0,05; F эмпирическое: 5,43. Нулевая гипотеза отклоняется, p=0,021. Дальнейшее применение метода Шеффе обнаруживает статистически значимое различие между магазинами электроники и универмагами.

12. В данном случае может быть применен однонаправленный t-критерий для зависимых выборок, поскольку данные представлены в шкале отношений, а выборки на две даты являются связанными. Критическое значение t-критерия: 1,81 (df = 10; α = 0,05). Эмпирическое значение t-критерия: 1,53. Нулевая гипотеза не может быть отклонена. Следовательно, статистически значимых различий в ценах акций компаний, работающих на рынке услуг связи, за анализируемый период не обнаружено.

13. Нулевая гипотеза в этой задаче может быть сформулирована следующим образом: предпочтения покупателем страны-изготовителя не связано с тем, имел ли он машину ранее, т.е. наблюдаемые частоты равны ожидаемым: H_0: $O_i = E_i$. Для проверки этой гипотезы может быть использован критерий хи-квадрат. Критическое значение критерия: 3,84 (df = 1, α = 0,05). Эмпирическое значение критерия: 20,94. В случае использования поправки на непрерывность получим 19,11. В любом случае нулевая гипотеза отклоняется.

14. Для ответа на вопросы д) и е) рассмотрите, например, ситуации венчурного инвестирования (инвестирования в технологически новые и рискованные проекты), приобретения крупной партии оборудования и подобные; представьте, что исследуются такие вопросы, как: оказывает ли людям помощь некоторая весьма дорогостоящая и болезненная терапия, является ли данная пищевая добавка опасной для людей и т.п.; сформулируйте для таких ситуаций соответствующие гипотезы и определите возможные ошибки первого и второго родов.

15. А. Целесообразно применение критерия хи-квадрат.

Б. Проверяется отличие средних значений по выборке и генеральной совокупности. Могут использоваться z- и t-критерий, но в данном случае из-за небольшого объема выборки целесообразно использовать двусторонний t-критерий.

В. В данном случае выборки могут рассматриваться как независимые. Объем выборок позволяет использовать двунаправленный z-критерий.

Г. Может быть применен *t*-критерий для двух связанных выборок (данные представлены в шкале отношений). Поскольку предполагается увеличение продаж после тренинга, целесообразно использование однонаправленной альтернативной гипотезы.

Д. Если зависимая переменная (объем продаж) измеряется в шкале отношений, а влияющие факторы представлены в порядковой шкале (например, качество продукции ранжируется как «высокое», «среднее» и «низкое», а рекламный бюджет, например, в диапазонах «0-10000 руб.», «10000-20000 руб.», «выше 20000 руб.»), то корректным будет использование дисперсионного анализа. В этом случае могут быть оценены влияние главных эффектов (качество и реклама) и эффект их взаимодействия.

Е. Может использоваться *z*- или *t*-критерий для независимых выборок.

17. Для решения задачи целесообразно применить процедуру ANOVA и критерии множественных сравнений. Средние значения по выборкам: 7, 13, 15 дней соответственно. Эмпирическое значение *F*-критерия: 12,66. Критическое значение *F*-критерия: 3,35 ($\alpha = 0,05$; $df_B = 2$, $df_W = 27$). Для $\alpha = 0,01$ критическое значение *F*-критерия: 5,49. Нулевая гипотеза об отсутствии различий может быть отклонена при $p < 0,01$. Используя функцию FРАСПОБР(α;2;57), можно показать, что нулевая гипотеза может быть отклонена и при $p < 0,001$. Поэтому делаем вывод: гипотеза о том, что различия между тремя агентствами вызваны только случайными отклонениями, маловероятна. Скорость обработки документов статистически значимо различается по агентствам.

18. Для решения задачи целесообразно использовать процедуру ANOVA и критерии множественных сравнений.

Библиографический список

Литература, рекомендуемая для дополнительного изучения тем учебного пособия

1. Вуколов Э.А. Основы статистического анализа. Практикум по статистическим методам и исследованию операций с использованием пакетов STATISTICA и EXCEL. М.: ФОРУМ: ИНФРА-М, 2004.
2. Наследов А.Д. SPSS. Компьютерный анализ данных в психологии и социальных науках. СПб.: Питер, 2005.
3. Орлов А.И. Прикладная статистика. М.: Экзамен, 2004.
4. Хили Дж. Статистика: Социологические и маркетинговые исследования. СПб.: Питер, 2005.
5. Bond, T.G. and Fox, C.M. (2001) Applying the Rasch Model: Fundamental Measurement in the Human Science, Lawrence Erlbaum.
6. Cooper, D.R. and Shindler, P.S. (1995) Business Research Methods. Irwin/McGraw-Hill.
7. Wright, B.D. and Masters, G.N. (1982) Rating Scale Analysis, MESA Press.

Литература и ресурсы, использованные при подготовке учебного пособия

1. Аванесов В.С. Методологические и теоретические основы тестового педагогического контроля: дис. ... д-ра пед. наук. М.: Исследовательский центр проблем качества подготовки специалистов при Московском институте стали и сплавов, 1994.
2. Аванесов В.С. Применение тестовых форм в Rasch Measurement // Педагогические измерения. 2004. №4. http://testolog.narod.ru
3. Айвазян С.А., Енюков И.С., Мешалкин Л.Д. Прикладная статистика. Основы моделирования и первичная обработка данных: Справочное издание. М.: Финансы и статистика, 1983.
4. Бююль А., Цефель П. SPSS: искусство обработки информации. М.: Diasoft, 2005.
5. Виноградов А.Г. Надежность / HR-Лаборатория Human Technologies. www.ht.ru/on-line/forum/?forum=7&topic=190
6. Вуколов Э.А. Основы статистического анализа. Практикум по статистическим методам и исследованию операций с использованием пакетов STATISTICA и EXCEL. М.: ФОРУМ: ИНФРА-М, 2004.

7. Гласс Дж., Стэнли Дж. Статистические методы в педагогике и психологии. М.: Прогресс, 1976.
8. Дубнов П.Ю. Обработка статистической информации с помощью SPSS. М.: АСТ, 2005.
9. Дюк В.А. Конструирование психодиагностических тестов: традиционные математические модели и алгоритмы. http://psyfactor.org/lib/
10. Крыштановский А.О. Анализ социологических данных с помощью пакета SPSS. М.: ГУ ВШЭ, 2006.
11. Наследов А.Д. SPSS. Компьютерный анализ данных в психологии и социальных науках. СПб.: Питер, 2005.
12. Наследов А.Д. Математические методы психологического исследования. Анализ и интерпретация данных. СПб.: Речь, 2006.
13. Нейман Ю.М. Как измерять учебные достижения? // Вопросы тестирования в образовании. 2001. №1.
14. Нейман Ю.М., Хлебников В.А. Введение в теорию моделирования и параметризации педагогических тестов. М.: Прометей, 2000.
15. Орлов А.И. Прикладная статистика. М.: Экзамен, 2004.
16. Орлов А.И. Репрезентативная теория измерений и ее применения // Заводская лаборатория. 1999. Т. 65. No. 3.
17. Орлов А.И. Часто ли распределение результатов наблюдений является нормальным? // Заводская лаборатория. 1991. Т. 57. No. 7.
18. Пациорковский В.В., Пациорковская В.В. SPSS для социологов. М.: ИСЭПН РАН, 2005.
19. Плис А.И., Сливина Н.А. Практикум по прикладной статистике в среде SPSS: В 2 ч. М.: Финансы и статистика, 2004.
20. Практикум по общей теории статистики / Под ред. Н.Н. Ряузова. М.: Финансы и статистика, 1981.
21. Пфанцагль И. Теория измерений. М.: Мир, 1976.
22. Ростовцев П. Анализ социологических данных с применением статистического пакета SPSS
23. Таганов Д.Н. SPSS: Статистический анализ в маркетинговых исследованиях. СПб.: Питер, 2005.
24. Тюрин Ю.Н., Макаров А.А. Статистический анализ данных на компьютере. М.: ИНФРА-М, 1998.

25. Хили Дж. Статистика: Социологические и маркетинговые исследования. СПб.: Питер, 2005.
26. Amabile, T.M., Burnside, R.M. and Gryskiewicz, S.S. (1999). User's manual for assessing the climate for creativity: A survey from the Center for Creative Leadership. Greensboro, NC.
27. Anastasi, A. (1988). Psychological testing (6th ed.). New York: MacMillan Publishing Company.
28. Baker, F. (2001). The Basics of Item Response Theory, University of Maryland, College Park, MD.
29. Bond, T.G. and Fox, C.M. (2001) Applying the Rasch Model: Fundamental Measurement in the Human Sciences, Lawrence Erlbaum.
30. Campbell, D.T. (2001) Social measurement. Thousand Oaks: SAGE Publications.
31. Cannel, C., Fowler, F.J., Kalton, G., Oksenberg, L. and Bishoping, K. New quantitative techniques for pretesting survey questions, in Bulmer, M. (Ed) (2004) Questionnaries. Thousand Oaks: SAGE Publications. Vol. 4, pp. 187–201.
32. Carlson, M. (1995) 'GTE Directories: Customer Focus and Satisfaction', The Quest for Excellence VII, The Official Conference of the Malcolm Baldrige National Quality Award, February 6–8, 1995, Washington, DC.
33. Cooper, D.R. and Shindler, P.S. (1995) Business Research Methods. Irwin/McGraw-Hill.
34. Dubina, I.N. and Umpleby, S.A. (2006) 'Agenda Setting and Improvement Monitoring in a University Department', The Proceedings of the 12th International Deming Research Seminar, New York: The W. Edwards Deming Institut.
35. Ekvall, G. (1996) 'Organizational climate for creativity and innovation', European Journal of Work and Organizational Psychology, Vol. 5, pp. 105–123.
36. Ekvall, G. (1999) 'Creative climate', in Runco, M.A. and Pritzker, S.R. (Eds.) Encyclopedia of Creativity, San Diego: Academic Press, pp. 403–412.
37. El-Murad, J. and West, D.C. (2004) 'The definition and measurement of creativity: what do we know?', Journal of Advertising Research, Vol. 44, No. 2, pp. 188–201.
38. Feick, L.F. (2004) 'Latent class analysis of survey questions that include don't know responses', in Bulmer, M. Questionnaries. Vol. 4, Thousand Oaks: SAGE Publications, pp. 145–166.
39. Forman, E.H. and Selly, M.A. (2001) Decision By Objectives. World Scientific. http://www.expertchoice.com/dbo/

40. Hess, J., Singer, E. and Bushrey, J. (2004) ‘Predicting Test-Retest reliability from behavior coding’, in Bulmer, M. Questionnaries. Vol. 4, Thousand Oaks: SAGE Publications, pp. 129–143.
41. Karabatsos G. (1999) ‘Axiomatic measurement theory as a basis for model selection in item response theory’. Paper presented at 32nd annual conference of the Society for Mathematical Psychology, Santa Cruz, CA.
42. Kirton, M.J. (1987) Kirton Adaption-Innovation Inventory (KAI) manual (2nd edition), Hatfield: Occupational Research Centre.
43. Kirton, M.J. (1999) ‘Adaptors and innovators: a description and measure’, in Puccio, G.J. and Murdock, M.C. (Eds.) Creativity Assessment: Readings and Resources, Buffalo: Creative Education Foundation Press, pp. 121–132.
44. Litwin, M.S. (1995) How to measure survey reliability and validity. Thousand Oaks: SAGE Publications.
45. Mathisen, G.R. and Einarsen, S. (2004) ‘A review of instruments assessing creative and innovative environments within organizations’, Creativity Research Journal, Vol. 16, No. 1, pp. 119–140.
46. Maxwell, J.A. (1996) Qualitative research design: An interactive approach. Thousand Oaks: SAGE Publications.
47. Parry, H.J. and Crossley, H.M. (2004) ‘Validity of responses to survey questions’, in Bulmer, M. Questionnaries. Vol. 3, Thousand Oaks: SAGE Publications, pp. 351–372.
48. Shuman, H. (2004) ‘The random probe: A technique for evaluating the validity of closed questions’, in Bulmer, M. Questionnaries. Vol. 3, Thousand Oaks: SAGE Publications, pp. 389–396.
49. Tewari, J., Cooper, D.R. and Shindler, P.S. (1995) Instructor’s Resource Guide to accompany Business Research Methods. Irwin/McGraw-Hill.
50. Traub, R.E. (1994) Reliability for the social sciences: theory and applications. Thousand Oaks: SAGE Publications.
51. Winston, W.L., Albright, S.C. (2001) Practical Management Science. Duxbury.
52. Wright, B.D. and Masters, G.N. (1982) Rating Scale Analysis, MESA Press.

53. Образовательный портал «Экономика, Социология, Менеджмент: Методы сбора и анализа социологических данных». http://ecsocman.edu.ru/db/sect/124/36.html
54. Сайт В.С. Аванесова. http://testolog.narod.ru
55. Социологические исследования. http://ecsocman.edu.ru/socis/
56. Социология и маркетинг в сети. http://socionet.narod.ru/stat.html

57. Социология: методология, методы, математические модели. http://www.nir.ru/socio/scipubl/4M.htm
58. Статистика и обработка данных в психологии. http://psyfactor.org/lib/
59. Статистические методы. Сайт А.И. Орлова. http://orlovs.pp.ru/stat.php
60. Центр маркетинга и информационного менеджмента РГГУ. Словарь терминов. http://marketing.rsuh.ru
61. Экономика и математические методы. http://www.cemi.rssi.ru/emm/
62. Cumulative Item Response Theory Models. http://www.education.umd.edu/Depts/EDMS/tutorials/CIRT.html
63. Expert Choice. www.expertchoice.com
64. Exponenta.ru. Образовательный математический сайт. http://www.exponenta.ru/SOFT/STATIST/STATIST.asp
65. Institute for Objective Measurement. http://www.rasch.org
66. Palisade Decision Tools. www.palisade.com
67. StatSoft Russia. www.statsoft.ru
68. Wikipedia, the free encyclopedia. http://en.wikipedia.org

ПРИЛОЖЕНИЯ

Приложение 1. Статистические таблицы

1.1. Стандартное нормальное распределение

Для значений z указана площадь (S_R) под кривой стандартного нормального распределения между средним значением (0) и значением z (правая половина). В силу симметричности кривой нормального распределения площади слева (0 и z) будут иметь такое же значение. Общая площадь под кривой определяется как $2S_R$. В таблице также приведены значения α (для двусторонней проверки). При односторонней проверке значения α следует делить пополам.

z	S_R	α	z	S_R	α	z	S_R	α	z	S_R	α
0,60	0,226	0,549	0,91	0,319	0,363	1,22	0,389	0,222	1,53	0,437	0,126
0,61	0,229	0,542	0,92	0,321	0,358	1,23	0,391	0,219	1,54	0,438	0,124
0,62	0,232	0,535	0,93	0,324	0,352	1,24	0,393	0,215	1,55	0,439	0,121
0,63	0,236	0,529	0,94	0,326	0,347	1,25	0,394	0,211	1,56	0,441	0,119
0,64	0,239	0,522	0,95	0,329	0,342	1,26	0,396	0,208	1,57	0,442	0,116
0,65	0,242	0,516	0,96	0,331	0,337	1,27	0,398	0,204	1,58	0,443	0,114
0,66	0,245	0,509	0,97	0,334	0,332	1,28	0,400	0,201	1,59	0,444	0,112
0,67	0,249	0,503	0,98	0,336	0,327	1,29	0,401	0,197	1,60	0,445	0,110
0,68	0,252	0,497	0,99	0,339	0,322	1,30	0,403	0,194	1,61	0,446	0,107
0,69	0,255	0,490	1,00	0,341	0,317	1,31	0,405	0,190	1,62	0,447	0,105
0,70	0,258	0,484	1,01	0,344	0,312	1,32	0,407	0,187	1,63	0,448	0,103
0,71	0,261	0,478	1,02	0,346	0,308	1,33	0,408	0,184	1,64	0,449	0,101
0,72	0,264	0,472	1,03	0,348	0,303	1,34	0,410	0,180	1,65	0,451	0,099
0,73	0,267	0,465	1,04	0,351	0,298	1,35	0,411	0,177	1,66	0,452	0,097
0,74	0,270	0,459	1,05	0,353	0,294	1,36	0,413	0,174	1,67	0,453	0,095
0,75	0,273	0,453	1,06	0,355	0,289	1,37	0,415	0,171	1,68	0,454	0,093
0,76	0,276	0,447	1,07	0,358	0,285	1,38	0,416	0,168	1,69	0,454	0,091
0,77	0,279	0,441	1,08	0,360	0,280	1,39	0,418	0,165	1,70	0,455	0,089
0,78	0,282	0,435	1,09	0,362	0,276	1,40	0,419	0,162	1,71	0,456	0,087
0,79	0,285	0,430	1,10	0,364	0,271	1,41	0,421	0,159	1,72	0,457	0,085
0,80	0,288	0,424	1,11	0,367	0,267	1,42	0,422	0,156	1,73	0,458	0,084
0,81	0,291	0,418	1,12	0,369	0,263	1,43	0,424	0,153	1,74	0,459	0,082
0,82	0,294	0,412	1,13	0,371	0,258	1,44	0,425	0,150	1,75	0,460	0,080
0,83	0,297	0,407	1,14	0,373	0,254	1,45	0,426	0,147	1,76	0,461	0,078
0,84	0,300	0,401	1,15	0,375	0,250	1,46	0,428	0,144	1,77	0,462	0,077
0,85	0,302	0,395	1,16	0,377	0,246	1,47	0,429	0,142	1,78	0,462	0,075
0,86	0,305	0,390	1,17	0,379	0,242	1,48	0,431	0,139	1,79	0,463	0,073
0,87	0,308	0,384	1,18	0,381	0,238	1,49	0,432	0,136	1,80	0,464	0,072
0,88	0,311	0,379	1,19	0,383	0,234	1,50	0,433	0,134	1,81	0,465	0,070
0,89	0,313	0,373	1,20	0,385	0,230	1,51	0,434	0,131	1,82	0,466	0,069
0,90	0,316	0,368	1,21	0,387	0,226	1,52	0,436	0,129	1,83	0,466	0,067

z	S_R	α	z	S_R	α	z	S_R	α
1,84	0,467	0,066	2,35	0,491	0,019	2,86	0,498	0,004
1,85	0,468	0,064	2,36	0,491	0,018	2,87	0,498	0,004
1,86	0,469	0,063	2,37	0,491	0,018	2,88	0,498	0,004
1,87	0,469	0,061	2,38	0,491	0,017	2,89	0,498	0,004
1,88	0,470	0,060	2,39	0,492	0,017	2,90	0,498	0,004
1,89	0,471	0,059	2,40	0,492	0,016	2,91	0,498	0,004
1,90	0,471	0,057	2,41	0,492	0,016	2,92	0,498	0,004
1,91	0,472	0,056	2,42	0,492	0,016	2,93	0,498	0,003
1,92	0,473	0,055	2,43	0,492	0,015	2,94	0,498	0,003
1,93	0,473	0,054	2,44	0,493	0,015	2,95	0,498	0,003
1,94	0,474	0,052	2,45	0,493	0,014	2,96	0,498	0,003
1,95	0,474	0,051	2,46	0,493	0,014	2,97	0,499	0,003
1,96	0,475	0,050	2,47	0,493	0,014	2,98	0,499	0,003
1,97	0,476	0,049	2,48	0,493	0,013	2,99	0,499	0,003
1,98	0,476	0,048	2,49	0,494	0,013	3,00	0,499	0,003
1,99	0,477	0,047	2,50	0,494	0,012	3,01	0,499	0,003
2,00	0,477	0,046	2,51	0,494	0,012	3,02	0,499	0,003
2,01	0,478	0,044	2,52	0,494	0,012	3,03	0,499	0,002
2,02	0,478	0,043	2,53	0,494	0,011	3,04	0,499	0,002
2,03	0,479	0,042	2,54	0,494	0,011	3,05	0,499	0,002
2,04	0,479	0,041	2,55	0,495	0,011	3,06	0,499	0,002
2,05	0,480	0,040	2,56	0,495	0,010	3,07	0,499	0,002
2,06	0,480	0,039	2,57	0,495	0,010	3,08	0,499	0,002
2,07	0,481	0,038	2,58	0,495	0,010	3,09	0,499	0,002
2,08	0,481	0,038	2,59	0,495	0,010	3,10	0,499	0,002
2,09	0,482	0,037	2,60	0,495	0,009	3,11	0,499	0,002
2,10	0,482	0,036	2,61	0,495	0,009	3,12	0,499	0,002
2,11	0,483	0,035	2,62	0,496	0,009	3,13	0,499	0,002
2,12	0,483	0,034	2,63	0,496	0,009	3,14	0,499	0,002
2,13	0,483	0,033	2,64	0,496	0,008	3,15	0,499	0,002
2,14	0,484	0,032	2,65	0,496	0,008	3,16	0,499	0,002
2,15	0,484	0,032	2,66	0,496	0,008	3,17	0,499	0,002
2,16	0,485	0,031	2,67	0,496	0,008	3,18	0,499	0,001
2,17	0,485	0,030	2,68	0,496	0,007	3,19	0,499	0,001
2,18	0,485	0,029	2,69	0,496	0,007	3,20	0,499	0,001
2,19	0,486	0,029	2,70	0,497	0,007			
2,20	0,486	0,028	2,71	0,497	0,007			
2,21	0,486	0,027	2,72	0,497	0,007			
2,22	0,487	0,026	2,73	0,497	0,006			
2,23	0,487	0,026	2,74	0,497	0,006			
2,24	0,487	0,025	2,75	0,497	0,006			
2,25	0,488	0,024	2,76	0,497	0,006			
2,26	0,488	0,024	2,77	0,497	0,006			
2,27	0,488	0,023	2,78	0,497	0,005			
2,28	0,489	0,023	2,79	0,497	0,005			
2,29	0,489	0,022	2,80	0,497	0,005			
2,30	0,489	0,021	2,81	0,498	0,005			
2,31	0,490	0,021	2,82	0,498	0,005			
2,32	0,490	0,020	2,83	0,498	0,005			
2,33	0,490	0,020	2,84	0,498	0,005			
2,34	0,490	0,019	2,85	0,498	0,004			

1.2. *t*-распределение Стьюдента

	Уровни значимости для одностороннего критерия					
	0,1	*0,05*	*0,02*	*0,01*	*0,001*	*0,0005*
	Уровни значимости для двустороннего критерия					
df	*0,2*	*0,1*	*0,05*	*0,02*	*0,01*	*0,001*
1	3,078	6,314	12,706	31,821	63,656	636,578
2	1,886	2,920	4,303	6,965	9,925	31,600
3	1,638	2,353	3,182	4,541	5,841	12,924
4	1,533	2,132	2,776	3,747	4,604	8,610
5	1,476	2,015	2,571	3,365	4,032	6,869
6	1,440	1,943	2,447	3,143	3,707	5,959
7	1,415	1,895	2,365	2,998	3,499	5,408
8	1,397	1,860	2,306	2,896	3,355	5,041
9	1,383	1,833	2,262	2,821	3,250	4,781
10	1,372	1,812	2,228	2,764	3,169	4,587
11	1,363	1,796	2,201	2,718	3,106	4,437
12	1,356	1,782	2,179	2,681	3,055	4,318
13	1,350	1,771	2,160	2,650	3,012	4,221
14	1,345	1,761	2,145	2,624	2,977	4,140
15	1,341	1,753	2,131	2,602	2,947	4,073
16	1,337	1,746	2,120	2,583	2,921	4,015
17	1,333	1,740	2,110	2,567	2,898	3,965
18	1,330	1,734	2,101	2,552	2,878	3,922
19	1,328	1,729	2,093	2,539	2,861	3,883
20	1,325	1,725	2,086	2,528	2,845	3,850
21	1,323	1,721	2,080	2,518	2,831	3,819
22	1,321	1,717	2,074	2,508	2,819	3,792
23	1,319	1,714	2,069	2,500	2,807	3,768
24	1,318	1,711	2,064	2,492	2,797	3,745
25	1,316	1,708	2,060	2,485	2,787	3,725
26	1,315	1,706	2,056	2,479	2,779	3,707
27	1,314	1,703	2,052	2,473	2,771	3,689
28	1,313	1,701	2,048	2,467	2,763	3,674
29	1,311	1,699	2,045	2,462	2,756	3,660
30	1,310	1,697	2,042	2,457	2,750	3,646
35	1,306	1,690	2,030	2,438	2,724	3,591
40	1,303	1,684	2,021	2,423	2,704	3,551
45	1,301	1,679	2,014	2,412	2,690	3,520
50	1,299	1,676	2,009	2,403	2,678	3,496
55	1,297	1,673	2,004	2,396	2,668	3,476
60	1,296	1,671	2,000	2,390	2,660	3,460
65	1,295	1,669	1,997	2,385	2,654	3,447
70	1,294	1,667	1,994	2,381	2,648	3,435
75	1,293	1,665	1,992	2,377	2,643	3,425
80	1,292	1,664	1,990	2,374	2,639	3,416
85	1,292	1,663	1,988	2,371	2,635	3,409
90	1,291	1,662	1,987	2,368	2,632	3,402
95	1,291	1,661	1,985	2,366	2,629	3,396
100	1,290	1,660	1,984	2,364	2,626	3,390
150	1,287	1,655	1,976	2,351	2,609	3,357
200	1,286	1,653	1,972	2,345	2,601	3,340
250	1,285	1,651	1,969	2,341	2,596	3,330
300	1,284	1,650	1,968	2,339	2,592	3,323

1.3. *F*-распределение
α = 0,05

df_W	df_B 1	2	3	4	5	6	8	10	15	20	∞
2	18,51	19,00	19,16	19,25	19,30	19,33	19,37	19,40	19,43	19,45	19,50
3	10,128	9,552	9,277	9,117	9,013	8,941	8,845	8,785	8,703	8,660	8,527
4	7,709	6,944	6,591	6,388	6,256	6,163	6,041	5,964	5,858	5,803	5,628
5	6,608	5,786	5,409	5,192	5,050	4,950	4,818	4,735	4,619	4,558	4,365
6	5,987	5,143	4,757	4,534	4,387	4,284	4,147	4,060	3,938	3,874	3,669
7	5,591	4,737	4,347	4,120	3,972	3,866	3,726	3,637	3,511	3,445	3,230
8	5,318	4,459	4,066	3,838	3,688	3,581	3,438	3,347	3,218	3,150	2,928
9	5,117	4,256	3,863	3,633	3,482	3,374	3,230	3,137	3,006	2,936	2,707
10	4,965	4,103	3,708	3,478	3,326	3,217	3,072	2,978	2,845	2,774	2,538
11	4,844	3,982	3,587	3,357	3,204	3,095	2,948	2,854	2,719	2,646	2,405
12	4,747	3,885	3,490	3,259	3,106	2,996	2,849	2,753	2,617	2,544	2,297
13	4,667	3,806	3,411	3,179	3,025	2,915	2,767	2,671	2,533	2,459	2,207
14	4,600	3,739	3,344	3,112	2,958	2,848	2,699	2,602	2,463	2,388	2,131
15	4,543	3,682	3,287	3,056	2,901	2,790	2,641	2,544	2,403	2,328	2,066
16	4,494	3,634	3,239	3,007	2,852	2,741	2,591	2,494	2,352	2,276	2,010
17	4,451	3,592	3,197	2,965	2,810	2,699	2,548	2,450	2,308	2,230	1,961
18	4,414	3,555	3,160	2,928	2,773	2,661	2,510	2,412	2,269	2,191	1,917
19	4,381	3,522	3,127	2,895	2,740	2,628	2,477	2,378	2,234	2,155	1,879
20	4,351	3,493	3,098	2,866	2,711	2,599	2,447	2,348	2,203	2,124	1,844
21	4,325	3,467	3,072	2,840	2,685	2,573	2,420	2,321	2,176	2,096	1,812
22	4,301	3,443	3,049	2,817	2,661	2,549	2,397	2,297	2,151	2,071	1,784
23	4,279	3,422	3,028	2,796	2,640	2,528	2,375	2,275	2,128	2,048	1,758
24	4,260	3,403	3,009	2,776	2,621	2,508	2,355	2,255	2,108	2,027	1,734
25	4,242	3,385	2,991	2,759	2,603	2,490	2,337	2,236	2,089	2,007	1,712
30	4,171	3,316	2,922	2,690	2,534	2,421	2,266	2,165	2,015	1,932	1,623
35	4,121	3,267	2,874	2,641	2,485	2,372	2,217	2,114	1,963	1,878	1,559
40	4,085	3,232	2,839	2,606	2,449	2,336	2,180	2,077	1,924	1,839	1,510
45	4,057	3,204	2,812	2,579	2,422	2,308	2,152	2,049	1,895	1,808	1,471
50	4,034	3,183	2,790	2,557	2,400	2,286	2,130	2,026	1,871	1,784	1,439
60	4,001	3,150	2,758	2,525	2,368	2,254	2,097	1,993	1,836	1,748	1,390
70	3,978	3,128	2,736	2,503	2,346	2,231	2,074	1,969	1,812	1,722	1,354
80	3,960	3,111	2,719	2,486	2,329	2,214	2,056	1,951	1,793	1,703	1,326
90	3,947	3,098	2,706	2,473	2,316	2,201	2,043	1,938	1,779	1,688	1,303
100	3,936	3,087	2,696	2,463	2,305	2,191	2,032	1,927	1,768	1,676	1,284
∞	3,842	2,996	2,605	2,372	2,214	2,099	1,939	1,831	1,666	1,571	1,000

α = 0,01

df_W	df_B 1	2	3	4	5	6	8	10	15	20	∞
2	98,50	99,00	99,16	99,25	99,30	99,33	99,38	99,40	99,43	99,45	99,50
3	34,12	30,82	29,46	28,71	28,24	27,91	27,49	27,23	26,87	26,69	26,13
4	21,20	18,00	16,69	15,98	15,52	15,21	14,80	14,55	14,20	14,02	13,46
5	16,26	13,27	12,06	11,39	10,97	10,67	10,29	10,05	9,722	9,553	9,022
6	13,75	10,93	9,780	9,148	8,746	8,466	8,102	7,874	7,559	7,396	6,881
7	12,25	9,547	8,451	7,847	7,460	7,191	6,840	6,620	6,314	6,155	5,651
8	11,26	8,649	7,591	7,006	6,632	6,371	6,029	5,814	5,515	5,359	4,860
9	10,56	8,022	6,992	6,422	6,057	5,802	5,467	5,257	4,962	4,808	4,312
10	10,04	7,559	6,552	5,994	5,636	5,386	5,057	4,849	4,558	4,405	3,910
11	9,646	7,206	6,217	5,668	5,316	5,069	4,744	4,539	4,251	4,099	3,604
12	9,330	6,927	5,953	5,412	5,064	4,821	4,499	4,296	4,010	3,858	3,362
13	9,074	6,701	5,739	5,205	4,862	4,620	4,302	4,100	3,815	3,665	3,166
14	8,862	6,515	5,564	5,035	4,695	4,456	4,140	3,939	3,656	3,505	3,005
15	8,683	6,359	5,417	4,893	4,556	4,318	4,004	3,805	3,522	3,372	2,870
16	8,531	6,226	5,292	4,773	4,437	4,202	3,890	3,691	3,409	3,259	2,754
17	8,400	6,112	5,185	4,669	4,336	4,101	3,791	3,593	3,312	3,162	2,654
18	8,285	6,013	5,092	4,579	4,248	4,015	3,705	3,508	3,227	3,077	2,567
19	8,185	5,926	5,010	4,500	4,171	3,939	3,631	3,434	3,153	3,003	2,490
20	8,096	5,849	4,938	4,431	4,103	3,871	3,564	3,368	3,088	2,938	2,422
21	8,017	5,780	4,874	4,369	4,042	3,812	3,506	3,310	3,030	2,880	2,361
22	7,945	5,719	4,817	4,313	3,988	3,758	3,453	3,258	2,978	2,827	2,307
23	7,881	5,664	4,765	4,264	3,939	3,710	3,406	3,211	2,931	2,780	2,257
24	7,823	5,614	4,718	4,218	3,895	3,667	3,363	3,168	2,889	2,738	2,212
25	7,770	5,568	4,675	4,177	3,855	3,627	3,324	3,129	2,850	2,699	2,171
30	7,562	5,390	4,510	4,018	3,699	3,473	3,173	2,979	2,700	2,549	2,008
35	7,419	5,268	4,396	3,908	3,592	3,368	3,069	2,876	2,597	2,445	1,892
40	7,314	5,178	4,313	3,828	3,514	3,291	2,993	2,801	2,522	2,369	1,806
45	7,234	5,110	4,249	3,767	3,454	3,232	2,935	2,743	2,464	2,311	1,739
50	7,171	5,057	4,199	3,720	3,408	3,186	2,890	2,698	2,419	2,265	1,685
60	7,077	4,977	4,126	3,649	3,339	3,119	2,823	2,632	2,352	2,198	1,602
70	7,011	4,922	4,074	3,600	3,291	3,071	2,777	2,585	2,306	2,150	1,542
80	6,963	4,881	4,036	3,563	3,255	3,036	2,742	2,551	2,271	2,115	1,496
90	6,925	4,849	4,007	3,535	3,228	3,009	2,715	2,524	2,244	2,088	1,459
100	6,895	4,824	3,984	3,513	3,206	2,988	2,694	2,503	2,223	2,067	1,429
∞	6,635	4,605	3,782	3,319	3,017	2,802	2,511	2,321	2,039	1,878	1,000

1.4. χ^2-распределение

	Уровни значимости			
df	*0,1*	*0,05*	*0,01*	*0,001*
1	2,706	3,841	6,635	10,827
2	4,605	5,991	9,210	13,815
3	6,251	7,815	11,345	16,266
4	7,779	9,488	13,277	18,466
5	9,236	11,070	15,086	20,515
6	10,645	12,592	16,812	22,457
7	12,017	14,067	18,475	24,321
8	13,362	15,507	20,090	26,124
9	14,684	16,919	21,666	27,877
10	15,987	18,307	23,209	29,588
11	17,275	19,675	24,725	31,264
12	18,549	21,026	26,217	32,909
13	19,812	22,362	27,688	34,527
14	21,064	23,685	29,141	36,124
15	22,307	24,996	30,578	37,698
16	23,542	26,296	32,000	39,252
17	24,769	27,587	33,409	40,791
18	25,989	28,869	34,805	42,312
19	27,204	30,144	36,191	43,819
20	28,412	31,410	37,566	45,314
21	29,615	32,671	38,932	46,796
22	30,813	33,924	40,289	48,268
23	32,007	35,172	41,638	49,728
24	33,196	36,415	42,980	51,179
25	34,382	37,652	44,314	52,619
26	35,563	38,885	45,642	54,051
27	36,741	40,113	46,963	55,475
28	37,916	41,337	48,278	56,892
29	39,087	42,557	49,588	58,301
30	40,256	43,773	50,892	59,702
35	46,059	49,802	57,342	66,619
40	51,805	55,758	63,691	73,403
45	57,505	61,656	69,957	80,078
50	63,167	67,505	76,154	86,660
60	74,397	79,082	88,379	99,608
70	85,527	90,531	100,425	112,317
80	96,578	101,879	112,329	124,839
90	107,565	113,145	124,116	137,208
100	118,498	124,342	135,807	149,449
200	226,021	233,994	249,445	267,539
300	331,788	341,395	359,906	381,424
400	436,649	447,632	468,724	493,131
500	540,930	553,127	576,493	603,446

1.5. Критические значения критерия Манна-Уитни

(для ненаправленных проверок)

α = 0,05

n_2	n_1													
	7	8	9	10	11	12	13	14	15	16	17	18	19	20
3	1	2	2	3	3	4	4	5	5	6	6	7	7	8
4	3	4	4	5	6	7	8	9	10	11	11	12	13	13
5	5	6	7	8	9	11	12	13	14	15	17	18	19	20
6	6	8	10	11	13	14	16	17	19	21	22	24	25	27
7	8	10	12	14	16	18	20	22	24	26	28	30	32	34
8	10	13	15	17	19	22	24	26	29	31	34	36	38	41
9	12	15	17	20	23	26	28	31	34	37	39	42	45	48
10	14	17	20	23	26	29	33	36	39	42	45	48	52	55
11	16	19	23	26	30	33	37	40	44	47	51	55	58	62
12	18	22	26	29	33	37	41	45	49	53	57	61	65	69
13	20	24	28	33	37	41	45	50	54	59	63	67	72	76
14	22	26	31	36	40	45	50	55	59	64	67	74	78	83
15	24	29	34	39	44	49	54	59	64	70	75	80	85	90
16	26	31	37	42	47	53	59	64	70	75	81	86	92	98
17	28	34	39	45	51	57	63	67	75	81	87	93	99	105
18	30	36	42	48	55	61	67	74	80	86	93	99	106	112
19	32	38	45	52	58	65	72	78	85	92	99	106	113	119
20	34	41	48	55	62	69	76	83	90	98	105	112	119	127

Источник: Наследов А.Д. Математические методы психологического исследования. Анализ и интерпретация данных. СПб.: Речь, 2006. 368 с.

1.6. Критические значения *D* одновыборочного теста Колмогорова-Смирнова

n	**Уровни значимости**	
	0,05	*0,01*
1	0,98	0,99
2	0,78	0,93
3	0,64	0,83
4	0,56	0,73
5	0,51	0,67
6	0,47	0,62
7	0,44	0,58
8	0,41	0,54
9	0,39	0,51
10	0,37	0,49
11	0,35	0,47
12	0,34	0,45
13	0,33	0,43
14	0,31	0,42
15	0,30	0,40
17	0,29	0,38
19	0,27	0,36
25	0,24	0,32
30	0,22	0,29
35	0,21	0,27
свыше 35	$1{,}36/\sqrt{n}$	$1{,}63/\sqrt{n}$

Источник: Cooper, D.R. and Shindler, P.S. (1995) Business Research Methods. Irwin/McGraw-Hill. P. 687.

1.7. Критические значения *D* двухвыборочного теста Колмогорова-Смирнова

N	**Уровни значимости**			
	направленная проверка		ненаправленная проверка	
	0,05	*0,01*	*0,05*	*0,01*
3	3	-	-	-
4	4	-	4	-
5	4	5	5	5
6	5	6	5	6
7	5	6	6	6
8	5	6	6	7
9	6	7	6	7
10	6	7	7	8
11	6	8	7	8
12	6	8	7	8
13	7	8	7	9
14	7	8	8	9
15	7	9	8	9
20	8	10	9	11
25	9	11	10	12
30	10	12	11	13
35	11	13	12	-
40	11	14	13	-
свыше 40			$1{,}36\sqrt{\frac{n_1+n_2}{n_1 n_2}}$	$1{,}63\sqrt{\frac{n_1+n_2}{n_1 n_2}}$

Источник: Cooper, D.R. and Shindler, P.S. (1995) Business Research Methods. Irwin/McGraw-Hill. Pp. 688–689.

1.8. Критические значения *T*-критерия Уилкоксона

n	**направленная проверка** *0,025*	*0,01*	*0,005*
	ненаправленная проверка *0,05*	*0,02*	*0,01*
7	2	0	-
8	4	2	0
9	6	3	2
10	8	5	3
11	11	7	5
12	14	10	7
13	17	13	10
14	21	16	13
15	25	20	16
16	30	24	20
17	35	28	23
18	40	33	28
19	46	38	32
20	52	43	38
21	59	49	43
22	66	56	49
23	73	62	55
24	81	69	61
25	89	77	68

Источник: Cooper, D.R. and Shindler, P.S. (1995) Business Research Methods. Irwin/McGraw-Hill. P. 686.

Приложение 2. Данные для выполнения практических задач

2.1. Фрагмент массива данных по оценке организационного климата в компаниях

№	Компания 1								Компания 2						
	Пункты анкеты								*Пункты анкеты*						
	п1	*п2*	*п3*	*п4*	*п5*	*п6*	*п7*		*п1*	*п2*	*п3*	*п4*	*п5*	*п6*	*п7*
1	4	3	5	4	3	3	4		4	4	5	2	2	2	2
2	2	3	4	3	4	1	5		4	3	4	4	4	4	4
3	4	2	5	1	1	1	4		3	4	5	4	3	2	3
4	4	4	4	4	4	4	4		3	5	5	5	5	5	5
5	3	5	3	1	3	1	2		3	4	4	4	3	2	2
6	1	3	1	4	2	2	2		4	5	4	5	3	2	3
7	3	4	4	5	3	2	3		4	3	3	3	3	3	4
8	2	5	3	2	2	3	2		2	1	3	5	1	5	2
9	3	5	4	1	4	4	4		3	5	4	4	4	4	3
10	4	4	4	2	3	3	3		3	5	4	1	2	2	3
11	4	5	4	3	3	4	4		3	4	2	3	2	1	2
12	1	2	1	3	2	3	1		3	4	4	5	3	5	4
13	5	3	1	1	1	1	1		2	5	5	5	1	2	4
14	3	4	1	3	3	2	3		4	5	4	4	3	3	3
15	3	3	3	1	2	1	1		3	5	4	5	5	4	4
16	4	4	4	4	4	2	4		3	2	2	1	1	2	1
17	3	3	4	4	3	3	2		4	4	4	4	4	4	4
18	4	4	2	1	2	2	2		1	2	1	1	1	1	2
19	3	5	3	4	3	3	3		3	4	3	4	3	2	3
20	4	4	4	4	4	4	4		4	4	4	4	3	4	3
21	4	4	4	4	4	4	4		3	4	3	3	2	3	3
22	5	4	5	3	3	4	4		3	4	5	5	3	5	2
23	2	4	2	4	2	2	1		5	5	5	5	3	5	4
24	2	2	2	3	3	3	2		3	3	5	5	3	5	4
25	3	5	3	4	4	1	1		3	3	4	2	3	1	3
26	3	3	2	3	2	1	1		4	1	5	3	2	4	4
27	5	4	5	4	4	4	4		3	3	3	4	3	3	3
28	3	4	2	5	3	4	4		1	4	4	4	1	4	1
29	1	5	4	4	3	3	5		4	4	5	4	3	5	4
30	3	4	3	2	2	3	3		3	4	5	1	4	5	3

2.2. Матрица QIPM

Факторы	Респонденты																					
	1		*2*		*3*		*4*		*5*		*6*		*7*		*8*		*9*		*10*		*11*	
	I	*P*	*I*	*P*	*I*	*P*	*I*	*P*	*I*	*P*	*I*	*P*	*I*	*P*	*I*	*P*	*I*	*P*	*I*	*P*	*I*	*P*
1	9	7			9	6	9	7	5	5	6	7	7	6	6	7	6	6	8	8	7	7
2	8	6			9	6	9	7	8	5	7	7	7	5	8	8	6	7	8	8	7	7
3	8	8	5	2	9	1	9	5	9	1	9	9	4	5	7	4	6	6	7	8	7	5
4	8	5			8	2	9	2	9	1	6	5	5	5	5	5	6	7	7	2	7	4
5	7	7	7	3	9	3	9	5	7	1	8	5	7	3	9	3	7	7	8	5	6	6
6	8	7	9	6	8	5	9	9	9	4	8	7	5	5	7	7	7	8	8	7	6	8
7	8	7	2	9	1	1	7	7	3	5	8	4	6	6	5	5	6	6	2	2	6	6
8	3	6	1	5	5	1	5	5	6	3	5	5	5	5	6	6	4	5	2	2	4	5
9	1	7	1	5	2	1	5	5	5	3	5	5	5	5	4	4	5	5	2	2	3	3
10	7	5	5	1	7	6	9	4	9	1	8	7	6	4	7	2	6	7	8	6	7	3
11	8	7	9	8	7	7	9	5	5	5	5	5					6	7	8	8	6	6
12	9	8	7	9	8	6	9	2	9	1	7	4	6	4	9	5	6	8	8	1	7	7
13	8	7	9	1	9	1	9	2	9	1	7	8	6	4	9	4	7	8	9	6	7	5
14	8	8	9	1	9	1	9	8	9	1	8	8	6	6	9	7	6	7	9	9	7	5
15	8	8	1	5	7	4	8	8	8	3	7	7	5	5	9	9	8	6	8	8	8	7
16	5	9	1	8	6	4	8	5	7	5	9	3	7	7	6	6	5	5	6	9	5	5
17	8	7	3	4	8	5	6	6	9	5	9	7	7	5	9	7	8	9	8	5	7	7
18	7	7			6	3	8	2	9	3	6	4	7	3	6	5	7	7	7	2	9	5
19	8	8	6	1	9	7	9	8	8	1	9	8	5	5	7	7	7	7	9	8	9	5
20	8	8	1	7	2	3	9	9	7	3	9		4	5	7	7	7	7	8	8	8	5
21	8	6	9	1	9	1	7	7	7	5	9	3	5	5	6	4	8	8	7	7	9	3
22	7	7	7	2	9	3	9	5	8	2	9	3	5	5	4	1	7	6	8	2	7	7
23	8	7	5	2	7	3	9	2	6	3	9	8	3	3			8	6	5	1	8	6
24	8	2	4	6	8	8	9	5	8	6	7	7	5	5	9	6	6	7	8	6	8	4
25	6	6	2	2	6	5	9	8	8	5	9	5	7	5	9	9	5	6	2	2	1	1
26	3	6	5	6	8	5	7	5	6	1	9	4	6	6	2	2	8	8	1	1	5	5
27	9	7	7	2	9	1	9	5	7	3	9	1	6	4	7	5	8	8	7	3	7	7
28	9	7	6	7	9	5	9	8	8	3	9	3	6	4	5	5	7	6	8	7	6	5
29	8	2	9	1	6	1	7	5	9	1	9	1	7	2	5	2	8	8	8	1	6	6
30	8	3	9	1	6	1	5	5	9	1	9	1	7	2			7	7	8	1	5	4
31	9	1	6	7	5	1	5	5	6	5	9	1	6	4			7	5	9	1	7	8
32	8	6	1		6	5	8	7	6	5	8	3	3	3			8	6	4	7	5	7
33	7	7	1		5	5	8	6	7	5	8	3	3	3			7	5	4	4	6	6

Сведения об авторе

Игорь Николаевич Дубина – кандидат философских наук (Омский государственный университет, 1999 г.), доцент кафедры информационных систем в экономике Алтайского государственного университета (с 2000 г.). Имеет физико-математическое и экономическое образование и 14-летний опыт исследовательской работы, связанной с проектированием, проведением, обработкой и анализом результатов эмпирических исследований. В период с 1992 по 1995 г. работал в Центре космического мониторинга и на кафедре теоретической кибернетики и прикладной математики АлтГУ, где применял методы статистического анализа для дешифрирования спутниковых снимков.

В 2004–2005 гг. автор проходил стажировку в Университете Дж. Вашингтона (Вашингтон, США), где знакомился с преподаванием курсов по методам статистического анализа (Statistical Methods and Models, Advanced Statistical Analysis), методам психометрических исследований и разработке измерительных шкал для гуманитарных и социально-экономических исследований (Survey Research, Rating Scale Construction).

Автором пособия разработан учебный курс «Методология научного творчества», который читался студентам исторического и социологического факультетов Алтайского государственного университета в 1996–1999 гг. С 2005 г. автор ведет специальный учебный курс «Математические основы маркетинговых исследований» для студентов экономического факультета АлтГУ. Также читает авторские курсы «Математические методы и модели в экономике», «Информационные технологии прогнозирования и оценки рисков», «Теория экономических игр», «Системы поддержки творческих решений».

Во время научно-исследовательской стажировки в Международном центре исследования творчества (Баффоло, США) в 2005 г. автор изучал и разрабатывал методы количественной оценки и анализа организационного климата для творчества и инноваций. Результаты этих исследований отражены в нескольких статьях, опубликованных в международных реферируемых журналах. Общее количество публикаций автора книги, непосредственно касающихся применения методов статистического анализа, социально-экономических измерений и методологии научных исследований, превышает 35, включая монографию, изданную при поддержке Российского гуманитарного научного фонда в Издательстве СО РАН.

Учебное издание

Игорь Николаевич Дубина

Математические основы эмпирических социально-экономических исследований

Учебное пособие

Редакторы: *Л.И. Базина, И.С. Васенко*

Изд. лиц. ЛР 020261 от 14.01.1997 г.
Подписано к печати 7.08.2006. Формат 60x84 1/16.
Бумага офсетная. Печать офсетная.
Усл.-печ. л. . Тираж 100. Заказ

Издательство Алтайского государственного университета:
656049, Барнаул, ул. Димитрова, 66