Высшее образование

Учебник

А. В. Пашкевич

ТЕОРИЯ ВЕРОЯТНОСТЕЙ И МАТЕМАТИЧЕСКАЯ СТАТИСТИКА

ДЛЯ СОЦИОЛОГОВ И МЕНЕДЖЕРОВ

Под редакцией профессора А. А. Макарова

ЕСТЕСТВЕННЫЕ НАУКИ

ACADEMIA

Высшее образование

БАКАЛАВРИАТ

А. В. ПАШКЕВИЧ

ТЕОРИЯ ВЕРОЯТНОСТЕЙ И МАТЕМАТИЧЕСКАЯ СТАТИСТИКА

ДЛЯ СОЦИОЛОГОВ И МЕНЕДЖЕРОВ

УЧЕБНИК

Под редакцией профессора А. А. Макарова

Рекомендовано
Учебно-методическим объединением вузов
Российской Федерации в области экономики,
менеджмента, логистики и бизнес-информатики
в качестве учебника для студентов высших учебных заведений,
обучающихся по направлениям подготовки «Социология»
и «Менеджмент» (квалификация «бакалавр»)

ACADEMIA

Москва
Издательский центр «Академия»
2014

УДК 519.21:519.22/25(075.8)
ББК 22.171:22.172я73
П222

Автор —
кандидат социологических наук, доцент общеуниверситетской кафедры высшей математики, доцент кафедры методов сбора и анализа социологической информации НИУ «Высшая школа экономики»
А. В. Пашкевич

Научный редактор —
кандидат физико-математических наук, заведующий общеуниверситетской кафедрой высшей математики НИУ «Высшая школа экономики», профессор
А. А. Макаров

Рецензенты:
канд. физ.-мат. наук, д-р философ. наук *С. Н. Бычков* (профессор кафедры философии образования Московского ин-та открытого образования);
канд. физ.-мат. наук *О. В. Иванов* (доцент факультета глобальных процессов Московского государственного ун-та им. М. В. Ломоносова, управляющий партнёр исследовательской компании Movie Research);
канд. соц. наук, доцент *Т. Ю. Черкашина* (зав. кафедрой общей социологии экономического факультета Новосибирского государственного ун-та);
канд. техн. наук *Л. Б. Косова* (доцент кафедры методов сбора и анализа социологической информации НИУ «Высшая школа экономики», директор Единого архива экономических и социологических данных)

Пашкевич А. В.

П222 Теория вероятностей и математическая статистика для социологов и менеджеров : учебник для студ. учреждений высш. образования / А. В. Пашкевич ; под ред. А. А. Макарова. — М. : Издательский центр «Академия», 2014. — 336 с. — (Сер. Бакалавриат).

ISBN 978-5-4468-0561-7

Учебник создан в соответствии с Федеральным государственным образовательным стандартом по направлениям подготовки «Социология» и «Менеджмент» (квалификация «бакалавр»).

В учебнике представлены материалы по методам описательной статистики, коэффициентам корреляции, построению доверительных интервалов и проверке статистических гипотез, о нормальном распределении и распределении Стьюдента, точечной оценке параметров. Учебник отличает направленность на профессиональные компетенции социологов, а также наличие примеров из практики социальных исследований.

Для студентов учреждений высшего образования.

УДК 519.21:519.22/25(075.8)
ББК 22.171:22.172я73

Оригинал-макет данного издания является собственностью Издательского центра «Академия», и его воспроизведение любым способом без согласия правообладателя запрещается

© Пашкевич А. В., 2014
© Образовательно-издательский центр «Академия», 2014
© Оформление. Издательский центр «Академия», 2014

ISBN 978-5-4468-0561-7

ПРЕДИСЛОВИЕ АВТОРА

Предмет «Теория вероятностей и математическая статистика» является неотъемлемым компонентом образовательной программы для многих специальностей в высших учебных заведениях и обязательным курсом для студентов, обучающихся по направлениям «Социология», «Менеджмент». Успешное овладение методологией и инструментарием этой дисциплины предопределяет дальнейшие шансы студентов на приобретение высокого уровня квалификации по большинству смежных курсов, касающихся стратегии обработки и анализа информации, и определяет статус выпускников на рынке труда по окончании университета.

Цель учебника — стать ***настольной книгой-гидом*** для студентов, обучающихся по направлениям социально-гуманитарного профиля (например, «Социология», «Менеджмент»). На страницах учебника представлено большое количество примеров из практики социальных исследований: они содержат иллюстрации разнообразных проектов, реализуемых в России и мире. В учебнике нашёл отражение мой преподавательский опыт работы в Национальном исследовательском университете «Высшая школа экономики» и компетенции автора в области проведения эмпирических социальных исследований.

На рынке книжной продукции этот учебник отличает одна важная деталь: он написан социологом, поэтому включает в себя задачи и упражнения, тематика которых соотносится с профессиональными компетенциями социологов, исследователей. В учебник включены задачи, которые прошли апробацию на лекционных и практических занятиях в ходе моей преподавательской деятельности. Подбор задач осуществлялся с учётом двух аспектов: *во-первых*, какие темы при изучении теории вероятностей и математической статистики наиболее трудны для понимания, а *во-вторых*, какие задачи делают процесс обучения интересным, «живым», ориентированным «на специальность» и помогают студентам успешно освоить этот предмет.

Структура учебника включает шесть глав, задачи для самостоятельной работы, образцы контрольной и экзаменационной работ, а также приложения, включающие статистические таблицы распределений и синтаксис статистических функций в Microsoft Excel, который можно использовать для решения задач из учебника. Содержание представлено темами, соответствующими принятому Федеральному государственному образовательному стандарту. Материал учебника

обеспечивает студентов базовыми знаниями и формирует практические навыки. Учебник содержит около 200 задач и практических упражнений, разработанных мною (1/3 сопровождаются подробным решением, 2/3 предлагаются читателям для самостоятельного решения). В Приложении 3 приводится синтаксис статистических функций в Microsoft Excel. Все расчёты в задачах могут быть выполнены как вручную, так и с помощью компьютера.

Логика изложения выстроена исходя из соображений естественности и лёгкости освоения информации.

Особую благодарность и уважение выражаю ординарному профессору НИУ «Высшая школа экономики», доктору социологических наук ***Юлиане Николаевне Толстовой***, книги и лекции которой позволили сформироваться мне как социологу. Примером этого человека во многом объясняется моя любовь к профессии и моё отношение к «социологической» математике в частности.

Я очень признательна консультанту и научному редактору данной книги — профессору ***Алексею Алексеевичу Макарову***, заведующему общеуниверситетской кафедрой высшей математики НИУ «Высшая школа экономики», и рецензентам: ***Сергею Николаевичу Бычкову*** — кандидату физико-математических наук, доктору философских наук, профессору кафедры философии образования Московского института открытого образования; ***Олегу Валентиновичу Иванову*** — кандидату физико-математических наук, доценту факультета глобальных процессов в МГУ им. М. В. Ломоносова, управляющему партнёру исследовательской компании Movie Research; ***Татьяне Юрьевне Черкашиной*** — кандидату социологических наук, доценту, заведующей кафедрой общей социологии экономического факультета Новосибирского государственного университета; ***Ларисе Борисовне Косовой*** — кандидату технических наук, доценту кафедры методов сбора и анализа социологической информации НИУ «Высшая школа экономики», директору Единого архива экономических и социологических данных.

Благодарю Единый архив экономических и социологических данных (ЕАЭСД) и его директора ***Ларису Борисовну Косову*** за предоставление массивов эмпирических исследований различной тематики, которые легли в основу многих задач учебника, предложенных читателям для решения.

Сердечную симпатию хочу выразить ***студентам факультетов социологии и прикладной политологии*** НИУ «Высшая школа экономики», у которых я веду занятия, и поблагодарить за интерес к предмету «Теория вероятностей и математическая статистика», за плодотворное, интересное общение и эффективную работу в семестрах.

ПРЕДИСЛОВИЕ НАУЧНОГО РЕДАКТОРА

Не каждое учебное издание предваряют слова научного редактора. Как человек, много лет участвующий в подготовке специалистов социально-экономического профиля в Национальном исследовательском университете «Высшая школа экономики» и Московском государственном университете им. М. В. Ломоносова, хочу обратить внимание читателя на особенности замысла и стиля изложения данной книги. К этому меня подталкивает ряд существенных обстоятельств.

При всём многообразии вузовских учебников по теории вероятностей и математической статистике, появившихся в последние годы, далеко не все отвечают требованиям времени и учитывают особенности специальности. Речь идёт не о мелких замечаниях и вкусовых предпочтениях. Изложение материала во многих вузовских учебниках по данной тематике прежде всего делает упор на разделы теории вероятностей, добавляя материал по математической статистике фактически по остаточному принципу. При этом не учитывается, что начиная с 2003 года, элементы теории вероятностей и статистики вошли в программу школьного математического образования. Учащиеся 7—9 классов сейчас знакомятся с понятиями комбинаторики, случайных событий и их вероятностей, получают первое представление о простых случайных величинах и их характеристиках. Простые задачи на эту тему включает и Единый государственный экзамен по математике. (Эта тенденция в школьном математическом образовании будет развиваться и усиливаться). Несомненно, материал по теории вероятностей нуждается в повторении, углублении и закреплении и в вузовской аудитории. Однако, на мой взгляд, он не должен уже доминировать в подобных курсах и соответствующих учебниках.

Исходя из этого и выстроено изложение материала в данной книге. Её прежде всего следует считать учебником по математической статистике для социологов, менеджеров и экономистов. В социально-экономической практике специалисты прежде всего сталкиваются именно со статистическими задачами, а не с задачами теории вероятностей. Именно поэтому в данной книге основные понятия теории вероятностей даются *сквозь призму их использования в практических прикладных статистических исследованиях*, повторяются *ретроспективно*. Читателя не должно удивлять отсутствие в книге традиционных глав из классической теории вероятностей. Простое

изложение этих тем следует искать в школьных учебниках[1] или вузовских учебниках для экономических и гуманитарных специальностей[2]. Другим ограничением учебной литературы, существующей по данному направлению, является несколько сухое, формализованное, оторванное от практического применения изложение материала. Это не удивительно. Обычно подобные учебники пишутся профессиональными математиками, не всегда чувствующими специфику и особенности восприятия нового знания, новой информации студентами социально-экономического профиля. Автор данного учебника — Анна Валерьевна Пашкевич — профессиональный социолог, хорошо знающий особенности и специфику своей специальности, с одной стороны, а с другой — много преподающий «методные» курсы студентам-социологам. Надо сказать, что книг по теории вероятностей и математической статистике и анализу социологических данных, написанных профессиональными социологами для социологов, в России не так уж и много. Это, прежде всего, учебники А. О. Крыштановского[3] и Ю. Н. Толстовой[4] — непосредственных учителей А. В. Пашкевич.

Отличительной чертой данного учебника является подробный, неторопливый разговор автора с читателем. Этот разговор богато проиллюстрирован социологическими примерами и задачами. Автор, не ограничиваясь формальным математическим изложением, а порой и намеренно избегая его, не ленится уточнять и разъяснять те моменты, которые обычно вызывают трудности у студентов. Очевидно, что на таком пути неизбежны компромиссы между математической строгостью формулировок и их содержательной понятностью. Не исключаю, что кому-то такой стиль изложения не всегда придётся по вкусу, но по существу дела — крайне трудно освоить профессию без обсуждения её специфики и деталей на простых доходчивых примерах. Именно такое

---

[1] См., например: *Тюрин Ю. Н.* Теория вероятностей и статистика / Ю.Н. Тюрин, А.А. Макаров, И.Р. Высоцкий, И. В. Ященко. — 2-е изд., перераб. — М.: МЦНМО, Московские учебники, 2008; *Макарычев Ю. Н.* Алгебра. Элементы статистики и теории вероятностей. 7—9 классы: учеб. пособие / Ю.Н. Макарычев, Н. Г. Миндюк. — 3-е изд. — М.: Просвещение, 2005; *Бунимович Е.А.* Математика: вероятность и статистика. 5—9 классы / Е.А. Бунимович, В. А. Булычев. — М.: Дрофа, 2002.

[2] См., например: *Тюрин Ю. Н.* Анализ данных на компьютере: учеб. пособие. — 4-е изд., перераб. / Ю. Н. Тюрин, А. А. Макаров. — М.: Издат. дом «ФОРУМ», 2008; *Тюрин Ю. Н.* Теория вероятностей: учебник для экономических и гуманитарных специальностей / Ю. Н. Тюрин, А. А. Макаров, Г. И. Симонова. — М.: МЦНМО, 2009.

[3] *Крыштановский А. О.* Анализ социологических данных с помощью пакета SPSS: учеб. пособие для вузов. — М.: Издат. дом ГУ-ВШЭ, 2007.

[4] *Толстова Ю. Н.* Математико-статистические модели в социологии: учеб. пособие. — М.: Издат. дом ГУ-ВШЭ, 2007; *Толстова Ю. Н.* Анализ социологических данных: методология, дескриптивная статистика, изучение связей между номинальными признаками: учеб. пособие. — М.: Научный мир, 2000; *Математическая статистика для социологов. Задачник*. — М.: Издат. дом ГУ-ВШЭ, 2010.

обсуждение и предлагает автор читателю, и это, на мой взгляд, с лихвой оправдывает присутствие в тексте «живых» разговорных оборотов и обращений (с точки зрения лингвистики речи), напоминающих беседу лектора со студенческой аудиторией тет-а-тет.

Значительный упор на математическую статистику и её использование при анализе социологических данных, сделанный в данной книге, оправдан и тем, что в последнее десятилетие стремительно расширяется объём эмпирической информации, которая становится доступной социологам и менеджерам. Это связано не только с тем, что первичные данные традиционных социологических и маркетинговых исследований стали оформляться в виде файлов и баз данных, но и с тем, что современные электронные средства в автоматическом режиме позволили собирать огромные массивы статистической информации о различных сторонах человеческой деятельности. Например, сканирование на кассах шрихкодов покупаемых товаров многое может рассказать нам о вкусовых предпочтениях и притязаниях, а привязка покупок к банковской или дисконтной карте потребителя даёт нам представление и о самом потребителе — *что* и *почему* он ценит. Так, автоматически собираемая статистика при использовании мобильного телефона или Интернета, дополненная традиционной опросной информацией, может многое поведать о жизни и настроениях социума. В этих условиях заметно возрастают требования к профессиональной подготовке социологов, их умению работать с различными эмпирическими данными и делать корректные, обоснованные выводы. Именно этому посвящена данная книга.

Следует сказать несколько слов о не вполне традиционном порядке расположения материала в книге: сразу после гл. 2 «Методы описательной статистики» происходит переход к разбору коэффициентов корреляции, отодвигая тему оценивания параметров выборки в гл. 5. С прикладной точки зрения такая логика может быть вполне оправданна.

Хочу отметить, что, как показывает многолетний опыт работы преподавателей по этой дисциплине, понятия теории вероятностей и математической статистики даются студентам нелегко, можно сказать — не с первого раза. Это связано с тем, что данные понятия и концепты поздно появились в общечеловеческой культуре, и с тем, что эти дисциплины требуют определённой математической подготовки слушателей. Последняя у студентов социологических факультетов в нашей стране далеко не всегда находится на должном уровне. Это ещё раз оправдывает стиль изложения, выбранный в книге. Пошаговый рецептурный стиль изложения вопросов порой критикуется в зарубежной литературе и даже получил специальное название «cooking book» (поваренная книга). Однако мне думается, что авторы подобных книг выбирают этот стиль не случайно, а заботясь о студентах при изложении концептуально сложного материала. Я надеюсь, что эта книга будет полезна многим студентам и поможет им не только успешно сдать экзамены, но и стать хорошими специалистами.

Глава 1

ВВЕДЕНИЕ В МАТЕМАТИЧЕСКУЮ СТАТИСТИКУ

Знакомство с математической статистикой начнём с небольшого экскурса в историю социологии и проследим, что происходило в прикладных исследованиях в течение четырёх последних веков. Отметим, что одной из наиболее распространённых (доминирующих) в науке точек зрения о том, каково «происхождение» математической статистики, является «*англосаксонская*» точка зрения[1].

§ 1.1. История становления математической социологии

В XVII веке вышла в свет работа Джона Граунта (1620 — 1674) «Естественные и политические наблюдения, перечисленные в прилагаемом оглавлении и сделанные на основе бюллетеней о смертности по их отношению к управлению, религии, торговле, росту, воздуху, болезням и другим изменениям гор. Лондона» (Natural and Political Observations Made upon the Bills of Mortality), которая наряду с работой Уильяма Петти (1623 — 1687) «Политическая арифметика» (Political Arithmetic) предопределила судьбу математической статистики в целом и математической социологии, в частности.

Сферу исследовательских интересов Джона Граунта составляли социальные закономерности, события повседневной жизни и, в частности, вопрос смертности среди разных социальных классов, возрастных групп, профессий (при этом особое внимание он уделял

[1] Крайне интересными, глубочайшими по своему содержанию являются книги и лекции профессора *Юлианы Николаевны Толстовой*, профессора *Дмитрия Семёновича Шмерлинга*, посвящённые рассмотрению такого вопроса, как *история математики* в социальных науках. В данном параграфе учебника автором приводится краткий обзор событий и фамилий, сыгравших наиболее значимую роль в становлении математической социологии.

социальной обусловленности заболеваний)[1]. Джон Граунт первым составил таблицы смертности на основе данных о рождении и смерти, регистрируемых в то время церковью, и стал основоположником так называемой *статистики социальной гигиены*. Им впервые были рассчитаны *вероятности дожития* людей до определённого возраста и реализованы попытки инструментального применения в социальном исследовании достижений теории вероятностей.

Весьма важную страницу в социальной статистике открывает работа другого английского учёного — Уильяма Петти[2] — «Политическая арифметика». Полное название работы исчерпывающе указывает на широту выполненных автором исследований: «Политическая арифметика, рассуждение о величине и ценности земель, населения, строений, земледелия, мануфактур, торговли, рыболовства, ремесленников, солдат, государственных доходов, процентов, налогов, ростовщичества, кораблей, банков, об оценке людей, увеличении числа матросов, о милиции, гаванях, позициях, мореплавании, морском могуществе и т. д., насколько всё это относится ко всем странам вообще и в особенности к территориям его величества короля Великобритании и его соседей — Голландии, Зеландии и Франции». На основе анализа и подсчёта разного рода показателей, в цифрах описывающих «изнанку» жизни Лондона (и других городов), Великобритании (и других государств), учёный *эмпирически* (т. е. на опыте) *установил* так называемую *постоянную величину частот* (*процентов*) и пришёл к выводу о том, что те или иные социальные показатели из года в год колеблются вокруг одной и той же цифры (определённого процента). Это заявление прозвучало в то время весьма неожиданно. На примере анализа эмпирических данных У. Петти наглядно показал, что социум характеризуется весьма устойчивыми закономерностями. Проведя статистико-экономические и социальные обследования, Петти зафиксировал, что многие показатели, лежащие в основе системы национальных счетов и характеризующие

---

[1] См. *Граунт Дж.* Начала статистики населения, медицинской статистики, математики страхового дела / Дж. Граунт, Э. Галлей; пер. с англ. О. Б. Шейнина. — Берлин: NG Verl., 2005 — 133 с.

[2] См. *Petty W.* Political arithmetic or A discourse concerning the extent and value of lands, people, buildings... As the same relates to every country in general, but more particularly to the territories... of Great Britain and... of Holland, Zealand, and France / W. Petty. — London: Printed for Robert Clavel at the Peacock, and Hen. Mortlock at the Phoenix..., 1691; *Petty W.* Essays on manking and political arithmetic / W. Petty. — Cassel & Company, 1888. — 192 с.; *Петти В.* Экономические и статистические работы / Пер. под ред. д-ра экон. наук М. Смит; Предисл. д-ра экон. наук Д. Розенберга. — М.: Соцэкгиз, 1940. — 324 с.; *Слуцкий Е. Е.* Сэр Вилльям Петти: Краткий очерк его экономических воззрений с приложением нескольких важнейших отрывков из его произведений / Евгений Слуцкий. — Киев: Тип. А. И. Гросман, 1914. — 48 с. (Евгений Евгеньевич Слуцкий (1880 — 1948)).

социальную сторону жизни определённых категорий людей, сохраняют своё постоянство.

Для анализа эмпирических фактов (сведений) о социуме были в некотором смысле впервые привлечены основные инструменты анализа информации, которыми на тот момент располагала *теория вероятностей*. Будучи разделом математики, теория вероятностей изучает закономерности (т. е. устойчивые взаимосвязи между явлениями) *случайных событий* и *случайных величин*. Поначалу *теория вероятностей* была слабо структурированной наукой (вплоть до середины XVII в.), многие понятия не имели строгих научных определений, а эмпирическим материалом для математических задач служили головоломки азартных игр — каковы шансы на выигрыш при тех или иных условиях.

Однако после публикации работ Дж. Граунта и У. Петти, а также исследований других авторов, начиная с середины XVII в. и до середины XIX в., теория вероятностей постепенно «подпитывалась» социальными данными, материалом прикладных, «жизненно важных» социальных исследований. Это привело к возникновению социальной статистики и демографии. Статистические исследования XVII—XIX вв., пожалуй, можно рассматривать как *предтечу* эмпирической социологии (в истории социологии возникновение эмпирической социологии связывают с Чикагской школой, второй четвертью XX в.). Постепенно начала́ формироваться ***культура счёта*** — люди начали считать и *придавать новый смысл имеющимся данным об обществе*, в котором живут. Так, к изучению социума постепенно были приложены понятия, математические методы и обобщения теории вероятностей.

Наряду с этим в конце XVIII в. Мари Жан Антуан Николя Кондорсе (1743—1794) — французский учёный и политический деятель — предложил и развил идею о том, что *мнение людей*, в дополнение к данным официальной государственной статистики, *можно и до́лжно изучать*, *измерять*, *подвергать анализу*[1]. Следует отталкиваться от суждения о том, что мнение человека (или группы людей), *если придать этому мнению числовое выражение* (*т. е. формализовать*, *зафиксировать с помощью математических символов*, *чисел*), *то его можно ассоциировать с распределением вероятностей и смотреть на мнение как на случайную величину*. Так была сформулирована (пусть и не конкретизирована) идея опросов общественного мнения, которые в XXI в. стали привычной чертой современности. Пользуясь массовыми опросами населения,

---

[1] См., например, *Старосельская-Никитина О. А.* Кондорсе как социолог / (Соч.) О. Старосельской; М. В. Ж. К. (Моск. высш. жен. курсы). — М.: Печатня А. И. Снегиревой, 1915. — 123 с.; *Бадентэр Э.* Кондорсе (1743—1794): Учёный в политике / Э. Бадентэр, Р. Бадентэр (Вступ. ст., сокращ. пер. с франц. О. В. Захаровой). — М.: Ладомир, 2001. — 398 с.

социологи не делают выводов относительно каждого человека в отдельности, а анализируют картину ***в среднем***, ***в общем***.

«Математика счёта» (статистика) в социальных науках оказывается важным инструментом научного поиска, позволяющим выявить ***типичные*** модели поведения людей, найти ***взаимосвязи*** между событиями, признаками, явлениями или подтвердить их отсутствие. Цели проведения такого поиска — выявить ***тенденции***, обозначить ***социальные взаимосвязи***, зафиксировать некие законы, по которым живет социум, а затем использовать это знание в прогнозировании и управлении.

§ 1.2. Объект и предмет математической статистики

Как только разговор заходит о понятии «статистика», нередко в предисловиях разного рода учебной литературы, да и на форумах в Интернете можно встретить упоминание и ссылку на вторую часть книги «Двенадцать стульев». Многие из нас ещё со школьной скамьи помнят, что одна из глав этого литературного произведения начинается с крылатой фразы о том, что «статистика знает всё»[1]. Далее в книге подробно изложены многие преимущества статистики (поскольку «ей под силу посчитать количество пахотной земли, сколько пищи съедает в год средний гражданин, <...>, количество охотников, балерин, <...>, собак всех пород, велосипедов, девушек, маяков и швейных машинок»). Имеются, конечно, и другие (противоположные и менее лестные) шутки, высказывания, которые приходится слышать в повседневной жизни (например, известная многим фраза о том, что существует три знания: «истина», «ложь» и «статистика»...).

Дадим строгое определение главного понятия для данного учебника.

Математической статистикой называется раздел математики, который посвящен ***методологии и методам обработки и анализа данных***.

Одна из ***главных задач***, которая стоит перед этой наукой, — снабдить исследователей и аналитиков надлежащими ***правилами*** (алгоритмами) обработки информации и научить их ***культуре обоснованного высказывания***. Эта наука опирается на ключевые понятия теории вероятностей, такие как: случайное событие, случайная величина, вероятность события, условная вероятность события, зависимые и независимые события, закон распределения вероят-

---

[1] *Ильф И.* Двенадцать стульев / И. Ильф, Е. Петров. — М.: АСТ: Астрель, 2011. (Ч. вторая, Гл. 9. Среди океана стульев).

ностей случайной величины, числовые характеристики случайных величин и др.

Одним из основных ***объектов*** математической статистики являются случайные величины, а ***предметом*** — параметры их распределений. Случайной величиной в теории вероятностей называют величину, которая в результате случайного опыта (эксперимента, испытания) может принимать те или иные *числовые* значения, а *каждому* из её возможных значений может быть противопоставлена (т. е. указана, заявлена, определена) своя вероятность встречаемости. Вероятность — это количественная оценка возможности того, что определённое событие может иметь место в реальности, может появиться. Это так называемый *шанс* случиться, произойти на практике некоторому событию.

Приведём несколько примеров случайных величин из практики исследований:

- *среднемесячный доход* наудачу (т. е. случайно) выбранного «в толпе» человека можно рассматривать как случайную величину и измерять её, например, в рублях;
- *возраст* наудачу выбранного «в толпе» человека можно рассматривать как случайную величину и измерять её, например, как число полных лет;
- *дневная выручка магазина*, случайно выбранного среди магазинов определённого сегмента в городе, может быть измерена, например, в тыс. евро;
- *количество тюбиков зубной пасты*, купленных за последний месяц случайно выбранным «в толпе» человеком, — в шт.;
- *количество медалей*, которые получила выбранная наудачу спортивная команда в ходе соревнований на XXVII Всемирной летней Универсиаде — 2013 в Казани;
- *количество зарубежных стран*, в которых успел побывать человек (наудачу выбранный «в толпе») за последние три года;
- *количество верно решённых заданий* по математике абитуриентом (выбранным наудачу) в ходе Единого государственного экзамена (ЕГЭ);
- *количество банков*, чьими услугами пользуются члены домохозяйства, выбранного наудачу (случайно) среди других домохозяйств;
- *количество человек* (в случайной выборке исследования), которые *планируют голосовать за определённую политическую партию* на предстоящих выборах в парламент;
- *количество человек* (в случайной выборке исследования), которые скажут, что *умеют кататься на горных лыжах*;
- *количество студентов* (в случайной выборке исследования), которые ответят, что *приезжают на занятия в университет на личном автомобиле*;
- *количество человек* (в случайной выборке исследования), которые *имеют карий цвет глаз*.

Можно привести множество других примеров.

Из приведённого перечня случайных величин видно, что если речь идёт об опросах общественного мнения или о прикладных исследованиях вообще, то случайными величинами могут выступать самые разнообразные характеристики (свойства) изучаемых объектов, измеренные с помощью чисел, т. е. имеющие числовое (количественное) выражение. В эмпирических исследованиях случайными величинами выступают ***признаки***, они могут присутствовать в анкете в виде вопросов, или в карточке наблюдения (протоколе, базе данных) в виде каких-либо атрибутов. Интересующие исследователя свойства объектов могут регистрироваться с помощью чисел или каких-то других символов, позволяющих формализовать (конкретизировать, строгим образом обозначить, зарегистрировать) неструктурированную информацию, собранную в ходе исследования (см. гл. 3).

Вместе с тем в фокусе внимания математической статистики находятся *не просто случайные величины* сами по себе, ***а законы их распределения***, которые задаются конкретным ***распределением вероятностей***, т. е. распределением *вероятностей появления* их значений. Закон распределения вероятностей позволяет указать область значений случайной величины, которые она принимает очень часто, довольно часто или, напротив, редко или крайне редко. Вероятность в социальных исследованиях обычно оценивается как ***частота появления*** того или иного события, как ***частота встречаемости*** (упоминания, наблюдения) того или иного свойства (черты) в рассматриваемой совокупности данных. Точнее, ***в эмпирических исследованиях вероятность может быть приблизительно определена (найдена) как доля опытов (случаев), завершившихся интересующим исследователя событием.***

Пример 1.1. Предположим, что в ходе некоторого исследования в университете было опрошено 780 студентов, отобранных случайно. Вопрос, который задавался респондентам, звучал так: «Скажите, пожалуйста, сколько примерно раз в неделю Вы ходите в спортзал/фитнес-клуб/спортивные секции и другие спортивные заведения (включая самостоятельные занятия спортом)?» Случайной величиной в данном примере выступает *число посещений спортзала* (*в течение недели*). Пусть исследователей интересовало, какова доля студентов, которые не посещают никакие спортивные заведения и не занимаются спортом, и на вопрос исследователей ответили «0 раз». Тогда *событие*, *наступление которого фиксируется в ходе данного исследования*, *состоит в том*, *что студент* (*участник опроса*) *не посещает спортзал*. Доля таких студентов, которые сказали, что не посещают спортзал, может рассматриваться как *примерная вероятность* того, что *случайно выбранный студент* университета не посещает спортивные заведения. Предположим, что среди опрошенных 780 человек оказалось, что вообще не посещают спортзал 162 человека. Следовательно, искомая вероятность может быть при-

близительно найдена как $\frac{162}{780} \approx 0,208 \approx 0,21$, т. е. *примерно каждый пятый* студент университета не посещает спортивные заведения (не занимается спортом). Итак, можно считать, что вероятность того, что случайно выбранный «в толпе» студент не уделяет время этим занятиям, равна 0,21.

Аналогичным образом можно рассчитать приблизительные вероятности для таких событий, как «Студент посещает спортзал один раз в неделю», «Студент посещает спортзал два раза в неделю», «Студент посещает спортзал три раза в неделю»... Так, если бы исследователей интересовала доля студентов, которые посещают спортзал *не менее трех раз* в неделю, то вероятность, соответствующая событию «Студент посещает спортзал не менее трёх раз в неделю», могла бы быть найдена как отношение числа человек, которые в ответе указали 3, 4, 5 или более раз, к общему числу опрошенных. Предположим, что среди опрошенных 780 человек оказалось, что три раза посещают спортзал 201 человек, четыре раза — 56 человек, пять раз — 12 человек, шесть раз — 4 человека, семь раз — 3 человека. Тогда приблизительная вероятность интересующего события может быть найдена как $\frac{201+56+12+4+3}{780} = \frac{276}{780} \approx 0,35$. Итак, можно считать, что вероятность того, что случайно выбранный «в толпе» студент посещает спортивные занятия не менее трех раз в неделю, равна 0,35.

Следовательно, из эмпирических исследований несложно получить информацию о том, каковы *примерные вероятности* тех или иных событий, а, значит, каково *примерное* (*приблизительное*) *распределение вероятностей* для интересующей случайной величины. Приведём ещё несколько примеров:

- вероятность того, что городской житель вообще не посещает театр, посещает его один, два, три, четыре и более четырёх раз в год может быть получена в ходе исследования;
- вероятность того, что выпускник престижного университета в первый год своей трудовой деятельности получает доход менее 10 тыс. руб., от 10 до 20 тыс. руб., от 20 до 30 тыс. руб., от 30 до 50 тыс. руб., более 50 тыс. руб. в месяц, может быть получена из данных социологического исследования;
- вероятность того, что курящий человек употребляет одну — две, три — пять, шесть — десять, одиннадцать — пятнадцать, более пятнадцати сигарет в день, может быть получена из данных эмпирического исследования;
- вероятность того, что автомобилист по дороге домой попадает в пробку и задерживается не более чем на полчаса, на час, на полтора часа, на два часа, более чем на два часа, может быть получена из данных эмпирического исследования.

Выявляемые в результате исследований законы распределения случайных величин имеют свои ***параметры*** (основные числовые характеристики, показатели), наиболее распространёнными из них являются ***математическое ожидание*** (***среднее***), ***дисперсия*** (***вариация***), ***стандартное отклонение***, ***медиана***, ***квартильный размах*** и ***многие другие***, которые будут рассмотрены в данном учебнике.

§ 1.3. Выборка и генеральная совокупность

Все, но не каждый[1].

Э. Ноэль-Нойманн, Т. Петерсен

Выборка может рассматриваться как ключ к поиску статистических закономерностей (закономерностей *в среднем*), как основа прикладных исследований: изучение *проявления* статистических закономерностей *на выборке и переносе* (*обобщении*) *результатов с выборки на генеральную совокупность*.

Интуитивные (в ряде случаев повседневные) представления о понятии «выборка» связаны с тем, что выводы и заключения об окружающем мире могут быть сделаны на основе *сравнительно небольшого* числа наблюдений, опираясь лишь на доступный фрагмент данных, обладая ограниченным объёмом информации и знаний об интересующей проблеме. Если вспомнить слова Джорджа Гэллапа, организовавшего в 30-е годы XX в. в США первый в мире Институт общественного мнения, то ***выборка*** — это та ***«ложка супа»***, которая призвана «рассказать» о вкусе всей кастрюли (это известная метафора, предложенная Гэллапом). Ещё одной наглядной иллюстрацией, которую можно часто услышать в качестве примера, поясняющего понятие «выборка», будет область клинических медицинских исследований: *всего несколько миллилитров крови* несут в себе информацию о состоянии здоровья. Для того чтобы получить первое, базовое, представление о том, здоров ли человек, достаточно провести медицинские исследования на небольшом фрагменте данных (т. е. выборке), осуществив забор крови из вены или из пальца. Кровь в пробирке — это выборка, своего рода *уменьшенная копия всей интересующей нас совокупности* — материи (в данном случае — крови всего организма человека).

Для начала введём три основных понятия, которыми нам предстоит активно пользоваться в социальных исследованиях, — генеральная совокупность, выборка и единица отбора.

---

[1] Эпиграф взят из заглавия книги *Ноэль-Нойманн Э.* Все, но не каждый. Введение в методы демоскопии / Э.Ноэль-Нойманн, Т. Петерсен; Пер. с нем. Л. Н. Рыбаковой и Е. В. Невмержицкой. Науч. ред. и автор вступ. ст. Н. С. Мансуров. — М.: МГУКИ, 2007. — 616 с.

Генеральной совокупностью будем называть совокупность (или множество) *всех* социальных объектов, которые подлежат изучению в рамках программы исследовательского проекта. Иными словами, это люди, которые потенциально могли бы принять участие в данном исследовании, т. е. все те, кого в исследовании мы считаем информантами — сведения *от кого* мы планируем получить; чьё мнение мы стремимся узнать в ходе исследования; та категория людей, которая нас интересует.

В прикладных исследованиях выделяют ***идеальную*** (теоретическую, воображаемую) и ***реальную*** (физически досягаемую, доступную в соответствии с ресурсами, которыми располагаем) генеральную совокупность. Например, всероссийские опросы населения предполагают изучение мнения только той категории граждан России, которые отвечают следующим критериям: граждане, которые достигли возраста 18 лет, имеют определённое место жительства, не находятся на непрерывном лечении в медицинских учреждениях, не несут службу в военных гарнизонах.

Следовательно, потенциальным участником всероссийского опроса общественного мнения может стать любой, кто отвечает этим характеристикам (требованиям). Вместе они составляют *идеальную* (*теоретическую*) *генеральную совокупность*. Однако *по факту* в опросах общественного мнения не принимают участия (хотя и удовлетворяют вышеперечисленным критериям) жители труднодоступных районов России, например, некоторых территорий Якутии, полуострова Ямал и др. Также опрос не проводится в местах, где ведутся военные действия (до недавнего времени жители Чечни не были участниками общероссийских опросов). В таком случае речь идет *о реальной* (*фактической*) *генеральной совокупности* — множестве людей, в отношении которых социологи *могут* обеспечить доступ к их мнению, не подвергая опасности жизнь интервьюеров, не расходуя колоссальные средства на транспортные расходы с целью добраться до труднодоступных территорий, экономя время.

Важно отметить, что есть еще и вторая традиция в понимании понятия «генеральная совокупность». Так, в качестве генеральной совокупности может рассматриваться и ***мнение отдельно взятого человека*** **на** ***временно́м континууме***. В основе этой идеи лежит представление о том, что мнение каждого человека — это не статичная, а, напротив, изменчивая величина, «колеблющаяся» в разные отрезки времени. Мнение отдельно взятого человека может быть разным (подвижным) или, напротив, весьма постоянным, устойчивым. А это значит, что ***мнение человека***, ***зафиксированное в некий конкретный момент времени***, ***— выборка***.

Выборкой (или ***выборочной совокупностью***) будем называть часть объектов генеральной совокупности (её *под*множество), отобранных *по определённым правилам* и подлежащих изучению

в пределах программы исследовательского проекта[1]. Выделяют два типа выборок — ***случайные (вероятностные)*** и ***неслучайные (невероятностные)***[2]. Названия говорят сами за себя. К вероятностным (случайным) выборкам относят выборки, процедура отбора респондентов в которых является случайной: когда *всем* потенциальным участникам исследования гарантировано *равенство шансов* быть отобранными. Невероятностные выборки — те, в которых нарушены эти гарантии. В вероятностной выборке процедура отбора устроена по принципу лотереи или «крутящегося барабана», в котором объекты *тщательно* перемешаны и каждому из них присвоен уникальный (т. е. неповторяющийся) номер и из которого *совершенно произвольно* (наудачу, случайно) извлекается нужное количество объектов (номеров выигрышных билетов, участников и т. п.). Аналогично мы поступаем в социальном исследовании. Необходимо непредвзято, независимо от субъективных предпочтений и прочих «удобств» для исследователя отобрать нужное количество участников опроса, руководствуясь незамысловатым правилом: «как карта ляжет» или «будь, что будет». Для того чтобы обеспечить выполнение основного условия — ***равенство шансов*** быть отобранным (что означает отсутствие дискриминации) — исследователю требуется иметь предварительный список (так называемый *реестр*) всех потенциальных участников исследования. В него должны войти *все* те, кто составляет фактическую генеральную совокупность. Этот список будет ***основой*** выборочного исследования.

Пример 1.2. Если планируется выборочный опрос постоянных клиентов некоторой корпорации, являющихся держателями клубной карты, то для начала можно воспользоваться общим списком постоянных клиентов компании, которые, например, имеют карту постоянного покупателя и потому зарегистрированы в клиентской базе. Далее, имея на руках этот перечень (в виде электронной или бумажной таблицы), нужно обеспечить совершенно случайный (не

---

[1] Рекомендуем студентам ознакомиться с классическими работами по проектированию выборки. См., например: *Kish L.: selected papers*. Edited by Steven Heeringa and Graham Kalton. John Wiley and Sons, Inc., Hoboken, New Jersey, 2003; *Кокрен У.* Методы выборочного исследования. — М.: Статистика, 1976; *Gallup G. A.* Guide to Public Opinion Polls. Princeton: Princeton University Press, 1948; *Чуриков А. В.* Основы формирования выборки. — М.: Издат. дом ГУ-ВШЭ, 2005; *Ноэль Э.* Массовые опросы. Введение в методику демоскопии. — М.: Ава-Эстра, 1993. — С. 85 — 134.

[2] См., например: *Чуриков А. В.* Случайные и неслучайные выборки в социологических исследованиях // Социальная реальность. — 2007. — №4. — С.89 — 109; *Петренко Е. С.* Социологический опрос и выборка // Социальная реальность. 2007. — № 4. — С. 87 — 88. Рекомендуем также ознакомиться с публикациями: *Quota versus Probability Sampling* // Survey Methods Center Newsletter. — London, 1994. — Vol. 15. — No. 1. — P. 3 — 31; *Шляпентох В. Э.* Проблемы качества социологической информации: достоверность, репрезентативность, прогностический потенциал. — М.: ЦСП, 2006. — С. 294 — 303.

зависящий от нашего решения и предпочтений) отбор участников опроса. Для этого следует пронумеровать всех людей в списке, присвоив каждому уникальный (неповторяющийся) порядковый номер. Далее требуется воспользоваться датчиком случайных чисел, по которому будут выбраны порядковые номера клиентов, и эти люди будут приглашены к участию в исследовании.

В качестве датчика случайный чисел выступает ***Таблица случайных чисел или Датчик (генератор) случайных чисел*** (например, в программе Excel это функции «***СЛЧИС***» или «***СЛУЧМЕЖДУ***»). *Датчиком случайных чисел выступает математическая функция*, которая генерирует (создаёт) случайное число в некотором диапазоне (интервале) между двумя заданными числами. Классическим примером датчика случайных чисел может стать игральный шестигранный кубик. Подробно о технике случайного отбора можно прочитать в фундаментальных работах Лесли Киша (Leslie Kish, Мичиганский университет, США), Алексея Владимировича Чурикова (Фонд «Общественное мнение», Россия) и др.[1]

В результате из всей совокупности мы отбираем только её часть (фрагмент) и черпаем информацию лишь *о некотором её подмножестве*. А поскольку этот отбор является вероятностным, то выборка становится мини-копией генеральной совокупности, повторяя её структуру и обладая таким свойством, которое принято называть репрезентативностью. ***Репрезентативность*** — способность выборки *воспроизводить* характеристики всей популяции в целом, *адекватно представлять* генеральную совокупность по определенным параметрам (признакам)[2].

Это ключевое свойство вероятностной выборки — давать представление об изучаемых характеристиках, *близких к их истинным значениям*, как если бы мы имели возможность проанализировать мнение *всех* тех, кто составляет генеральную совокупность.

Пример 1.3. Пусть нас интересует процент (%) людей, которые готовы купить автомобиль в кредит, и если таких клиентов в генеральной совокупности, предположим, 60 %, то ожидается, что и в опросе, выборка для которого проектируется вероятностным (случайным) образом, оценка этой доли людей будет примерно равна

---

[1] *Kish L.* Survey Sampling. — N. Y.: John Wiley and Sons, 1965. — P. 35—47, 49—53; *Чуриков А. В.* Основы формирования выборки. — М.: Издат. дом ГУ ВШЭ, 2005. — С. 25—29.

[2] Кстати, часто нарушение репрезентативности (т. е. неадекватность воспроизведения структуры генеральной совокупности и свойств объектов, её составляющих) возникает в случае выборок малого объёма. В следующих главах мы, конечно, коснёмся вопроса, какие выборки следует считать большими или малыми. См., например: *Давыдов А. А.* Репрезентативность выборки // Социологические исследования, 1990. — № 1. — С. 115—121; *Козырева П. М.* Международные стандарты оценки качества социологических обследований / П. М. Козырева, М. С. Косолапов, М. С. Сваффорд // Мир России. — 1999. — №1—2.

этой цифре, а погрешность будет незначительной. Таким образом, выборка *повторит* (отразит) состав генеральной совокупности, а ошибка (погрешность) будет невелика. Это вполне поддаётся здравому смыслу и совершенно естественным рассуждениям: если таких людей, которые готовы купить автомобиль в кредит, 60 %, то это значит, что и их вероятность попадания в выборку будет выше, чем у других, и составит 0,6 против 0,4.

Пример 1.4. Если в пруду плавают 200 рыб двух видов: 150 карпов и 50 окуней, то вероятность выловить карпа будет в три раза выше вероятности выловить окуня и составит 0,75 против 0,25. А это означает, что если мы выловим 40 рыб, то следует ожидать, что в нашем улове (выборке) пропорция, отражающая численное превосходство карпов, будет примерно повторять ту, которая имеет место в реальности (т. е. в генеральной совокупности).

Единицей отбора будем называть элемент (группу элементов), подлежащий отбору на определённой ступени выборки.

В социальных исследованиях выделяют четыре основных вида вероятностной выборки — простая случайная, гнездовая (кластерная), механическая, стратифицированная (районированная).

Выборка, механизм случайного (вероятностного) отбора в которой реализован в один шаг (в одну ступень), называется ***простой случайной выборкой*** (***simple random sample***, ***SRS***). *Сначала берётся общий массив* (*реестр*, *список*, *основа*) — в виде электронной или бумажной таблицы, в которой перечислены все объекты, составляющие фактическую генеральную совокупность. *Далее осуществляется случайный отбор* (с помощью датчика или таблицы случайных чисел, кубика) и в результате *формируется выборка* (*подмножество*).

Пример 1.5. Пусть для реализации исследования требуется провести опрос 10 человек среди студентов I — IV курсов некоторого факультета. Для этого берём общий список студентов (его можно взять/составить в деканате). Создаём реестр/электронную таблицу (например, в программе Excel). Предположим, что на I — IV курсах факультета обучается 347 человек. Студенты будут пронумерованы от 1 до 347. Каждый студент, который обучается на факультете, будет зарегистрирован в таблице, и каждому будет присвоен уникальный (неповторяющийся) порядковый номер, соответствующий его месту в общем списке. С помощью вызова функции «СЛУЧМЕЖДУ» можно сгенерировать случайные числа. Для этого программе следует указать диапазон, внутри которого должны создаваться случайные числа. В данном примере нижней границей будет число 1, а верхней границей — число 347. Так, функция «СЛУЧМЕЖДУ» генерирует случайные числа, принадлежащие именно этому интервалу. Предположим, что были сгенерированы десять чисел, соответствующие требуемому объёму выборки: 31, 172, 288, 233, 65, 10, 275, 19, 323, 197. Следовательно, студенты, чьи фамилии присутствуют в списке под этими порядковыми номерами, и составят выборочную сово-

купность исследования. Далее с ними можно связаться по телефону, электронной почте или при личной встрече (в учебных аудиториях, буфете, спортзале и пр.). Если кто-то из студентов отказывается от участия в исследовании, то такому респонденту ищется замена (эта процедура осуществляется по аналогичным правилам: в таблице генерируется случайной число, которое укажет на нового участника исследования).

Итак, процедура простого случайного отбора реализована в одну ступень (один шаг) — из списка вероятностным образом выбраны участники (студенты). Никаких дополнительных процедур отбора не потребовалось. Однако в социальных исследованиях далеко не всегда отбор нужных элементов может быть осуществлен столь просто. Иногда это весьма трудоёмкий процесс, который требует реализации нескольких ступеней.

Гнездовая (кластерная) выборка (*clustering sampling*) строится по принципу случайного отбора, но осуществляется *в несколько этапов* (*как минимум в два*), поскольку в качестве единиц отбора на первой ступени (первом шаге) выступают *относительно более крупные* (бóльшие) общности объектов. Гнездом в практике исследований принято называть *единицы отбора более высокой ступени*, вбирающей в себя более мелкие (меньшие).

Пример 1.6. Пусть необходимо осуществить опрос среди школьников крупного города А*, при этом не обязательно затрачивать колоссальные силы на составление фактического списка всех школьников, обучающихся на данный момент в этом городе, и затем случайным образом отбирать из этого списка участников для опроса. Такая задача в большинстве случаев оказалась бы нерезультативной, так как потребовала бы больших затрат — привлечения административных ресурсов, временны́х расходов и пр. В этой ситуации оптимальным решением станет реализация «гнездового» отбора. В качестве гнёзд могут служить учебные заведения (школы): как и у каждого корабля есть свой порт приписки, так и каждый школьник «прикреплён» к конкретной школе. Это позволяет осуществить случайный отбор из списка школ (т. е. более крупных общностей), а не самих учащихся. Далее в тех школах, которые попадут в выборку, можно реализовать ещё одну попытку отбора — случайно выбрать классы, которые выступают гнёздами (*под*уровнями) следующей ступени. И уже на последнем этапе можно осуществить случайный отбор учащихся в отобранных классах (или включить всех школьников отобранных классов в выборку). В результате такая выборка строится как минимум в два (три) шага: на первом гнёздами (самыми крупными общностями объектов) служат школы; на втором — более «мелкие» общности — классы учащихся; и, наконец, на третьем шаге — сами люди (учащиеся).

Гнездовая выборка крайне ***эффективна по экономическим (прагматическим) соображениям*** и потому часто применяется

на практике, когда реализация простой случайной выборки (SRS) не представляется возможной. Отбор осуществляется по принципу «*сужающейся воронки*»: начиная с крупных кластеров, мы постепенно, шаг за шагом, доходим непосредственно до отбора самих участников исследования. В итоге отпадает необходимость составления списка фамилий всех потенциальных участников исследования — достаточно составить список гнёзд, к которым они принадлежат. Кстати, именно так поступают социологи и в случае всероссийских опросов общественного мнения. Представьте, что вам требуется спроектировать (продумать и реализовать на практике) общенациональную выборку населения[1] (учтите, что численность взрослого населения России крайне велика).

Может ли здесь идти речь о том, что сначала следует составить список всех граждан в возрасте 18 лет и более, постоянно проживающих на территории РФ? Едва ли воплощение такой идеи покажется многообещающим. Даже если на минуту представить себе, что это можно сделать быстро, тут же возникнет другая проблема, куда менее разрешимая. Отбирая людей из этого огромного списка фамилий *случайно*, мы будем иметь примерно такую картину: в выборку попадут люди из самых разных уголков страны — от Калининграда до Магадана, от Мурманска до Астрахани, находящихся на разном удалении друг от друга. Ведь работает принцип лотереи! И для проведения интервью нам пришлось бы затратить огромные усилия и понести большие издержки (транспортные, временны́е). *Простая случайная выборка теряет свою привлекательность на фоне больших по размеру генеральных совокупностей, географически распределённых на большой территории*. Сто́ит ли такой вопрос большой цены? Тем более, в ряде случаев при длительной процедуре сбора данных информация может попросту утратить свою актуальность, а качество исполнения каждой из операций, которые придётся реализовать исследовательской команде, заметно снизится. Поэтому исследователи поступают следующим образом: на первом шаге из общего списка

---

[1] Подробно об истории, теории и методологии, а также о современных практиках проектирования общенациональной выборки можно прочитать в кн.: *Kish L.* Survey Sampling. — N. Y.: John Wiley and Sons, 1965.— P. 301—304, 313—322, 396—404; *Heeringa S. G., Connor J. H., Darrah D. C.* 1980 SRC National Sample: Design and Development. Ann Arbor (MI): Institute for Social Research, 1986; *Hess I.* Sampling for Social Research Surveys 1947—1980. Ann Arbor (MI): Institute for Social Research, 1985; *Свафорд М. С., Косолапов М. С., Козырева П. М.* Российский мониторинг экономического положения и здоровья населения (РМЭЗ): измерение благосостояния россиян в 90-е годы // Мир России. — 1999. — Т. VIII. — № 3. — С. 153—163; *Чуриков А. В.* Основы формирования выборки. — М.: Издат. дом ГУ-ВШЭ, 2005. — С. 86—109; *Козеренко Е. В.* Проектирование территориальной выборки для регулярных обследований населения субъекта Российской Федерации: Автореф. дисс. на соиск. звания канд. социолог. наук. — М., 1999 и др.

случайно отбираются административные районы (после федеральных округов и субъектов РФ они представляют собой наиболее крупные структурные элементы территориального деления РФ). Их число в России составляет около двух тысяч. Это гнёзда первой ступени. На втором шаге в тех административных районах, которые по итогам вероятностного отбора на первом шаге попали в выборку, случайно отбираются более мелкие единицы — гнёзда второй ступени — избирательные участки или почтовые отделения. На третьем шаге случайным образом из числа домашних хозяйств, входящих в состав избирательных участков и/или почтовых отделений, отобранных на втором шаге, выбираются номера квартир (с соответствующими улицами и номерами домов). Наконец, на четвёртом шаге происходит отбор респондентов — участников исследования, достигших 18 лет, выбираемых из членов домашних хозяйств.

Механическая выборка (другое название — систематический отбор, *systematic sampling*) предусматривает выбор единиц из списка генеральной совокупности с некоторым *заданным шагом* (через равные промежутки).

В реализации систематического отбора имеется два принципиальных момента: во-первых, это выбор так называемой *стартовой точки* (т. е. номера единицы, с которой следует начать отбор), а во-вторых, *шага отбора* (т. е. через какой интервал следует проводить выбор элементов). Стартовую точку и шаг отбора следует определять, исходя из численности (размера) генеральной совокупности (пусть это будет число N), а также объёма выборки (n). Шаг отбора (k) рассчитывается как отношение размера генеральной совокупности к размеру выборки и равен $k = \frac{N}{n}$. Таким образом, каждый k-й объект (элемент) в генеральной совокупности подлежит включению в выборку и, следовательно, изучению. Случайность этого вида выборки проявляется ***в вероятностном выборе стартовой точки*** — совершенно непредвзято, полагаясь на волю случая, следует случайно выбрать номер объекта, который первым войдет в выборку. Здесь следует руководствоваться теми же процедурами и техниками, что и в простой случайной выборке. Со стартовой точки следует начинать отбор 2-го, 3-го, ..., k-го элемента, последовательно прибавляя шаг отбора к номеру ранее выбранного элемента. Принцип реализации систематического отбора сопоставим с разметкой железнодорожного полотна километровыми столбами или с улицами, на которых с заданным шагом установлены фонари. Поскольку выбор первого элемента случаен (!), то и вероятностный отбор всех последующих элементов исследования может быть обоснован этой случайностью.

Пример 1.7. Пусть в одном из маркетинговых исследований предстоит провести около 40 телефонных интервью среди клиентов компании, заинтересованных в покупке товаров определённой категории.

Имея на руках списки фамилий клиентов и их телефонные номера, на практике довольно легко воспользоваться систематическим отбором респондентов. Предположим, что этот список состоит из 570 человек. Требуемый объём выборки для исследования составляет, например, 38 человек. Следовательно, шаг отбора равен $k = \frac{570}{38} = 15$,

т. е. каждый 15-й человек в списке подлежит включению в выборку исследования. Основной вопрос: с кого же начать? Поскольку весь реестр состоит из 570 человек, то с помощью датчика случайных чисел можно случайно выбрать (сгенерировать) целое число в диапазоне от 1 до 570. Допустим, этим числом оказалось 84. Это означает, что клиент, чей порядковый номер в списке равен 84, и будет первым элементом отбора. Остальные участники исследования определяются так: 84 + 15 (шаг отбора) = 99 (это номер второго участника исследования); 99 + 15 = 84 + 2 · 15 = 114 (номер третьего участника исследования), и как только мы исчерпаем список, следует перейти в его начало. В данном примере 84 + 15 · 32 = 564, т. е. клиент под № 564 завершает круг, и нам следует вернуться в начало списка: 84 + 15 · 33 = 579, т. е. это будет клиент под номером 9 в нашем списке. В завершении отбора в выборку будут включены клиенты под номерами 24, 39, 54, 69.

Проектирование ***стратифицированной (районированной) выборки*** (*stratified sampling*) состоит из двух шагов: первый (теоретический) заключается в осмысленном разделении генеральной совокупности на подгруппы, называемые стратами (или слоями). При этом *каждый элемент генеральной совокупности должен быть отнесён к одной и только одной страте*, *а сами выделенные подгруппы* (*страты*) *не должны пересекаться*, и в сумме составляют генеральную совокупность. На втором шаге *из каждого* (*без исключения!*) слоя случайным образом выбираются элементы (с помощью SRS, кластерного или механического принципов отбора). Такая процедура проводится с единственной целью — ***увеличить точность выборки*** и, следовательно, ***повысить качество данных***, собираемых с её помощью, так как в исследовании принимают участие представители всех без исключения страт (слоёв). Критерии, которые кладутся в основу стратификации в социальных исследованиях, могут быть разными. Но, как правило, они охватывают базовые социально-демографические характеристики людей:

- *пол* (мужчины и женщины);
- *тип населённого пункта*, место жительства (например, жители города и жители села);
- *уровень образования* (например, люди с высшим образованием и без высшего образования);
- *возраст* (например, люди старшего, среднего и молодого возраста);

• *материальное благополучие* (например, люди с высоким, средним и низким достатком).

Критерий, выбираемый в качестве основания для стратификации, должен быть тесно связан с характеристиками, подлежащими изучению в исследовательском проекте. В противном случае процедура стратификации (расслоения) теряет свою актуальность для исследования.

Пример 1.8. Если в ходе некоторого маркетингового исследования изучается ценовая эластичность спроса, то разумно выделить несколько групп клиентов в соответствии с их покупательской способностью (покупательской активностью): этими категориями могут стать, например, такие сегменты, как «эконом», «стандарт», «премиум». Каждый клиент компании принадлежит только одной из выделенных групп. В данном примере стратификационным критерием отнесения каждого из клиентов к «своей» страте может послужить «средняя сумма чека за последние полгода» или «общая сумма покупок за последние полгода/год». Это своего рода маркеры, по которым можно довольно точно «распознать» и классифицировать покупательский выбор клиентов и их активность. Идеи и предложения здесь могут быть разными. Главное, что стратификационный критерий и факт принадлежности клиента к определённой страте должны быть тесно связаны с исследовательскими характеристиками покупательского поведения — с изучением ценовой эластичности спроса. В результате в выборку войдут представители *каждой* из выделенных страт (как, правило, *пропорционально* численности этих страт в генеральной совокупности). Это позволяет ***повысить качество собираемых данных***, особенно в случае ***небольшого объёма выборки***.

Очевидно, что в страту объединяются похожие между собой элементы. В этом смысле каждая страта ***однородна*** по своему составу.

Пример 1.9. Крайне часто в практике социологических исследований выделяют две страты: городское и сельское население, и обеспечивают представительство каждой из групп в нужной пропорции. Если, например, планируется опрос жителей и известно, что доли городского и сельского населения соответственно равны 70 % и 30 %, а объём выборки при этом равен 1 200 человек, то на горожан придётся 840 интервью ($0{,}7 \cdot 1\,200$), а число интервью в сельской местности составит 360 ($1\,200 - 840$ или $0{,}3 \cdot 1\,200$). Имеются и другие стратегии размещения общего объёма выборки по стратам, такие как *оптимальное*, *размещение по Нейману*, *равное*. О них подробно можно прочитать в специальной литературе[1].

---

[1] См., например, классические работы: *Чуриков А. В.* Основы формирования выборки. — М.: Издат. дом ГУ-ВШЭ, 2005; *Kish L.* Survey Sampling. — N. Y.: John Wiley and Sons, 1965. — P. 75 — 106, 123 — 144; *Кокрен У.* Методы выборочного исследования. — М.: Статистика, 1976. — С. 103 — 126, 131 — 167.

§ 1.4. Выборка: немного истории

В повседневной жизни выборка играет колоссальную и в то же время привычную роль. Однако не более чем сто лет назад выборочный метод исследования, как способ получения информации об окружающем мире, не признавался практиками, которые считали, что достоверностью обладают исключительно сплошные обследования и переписи населения, предполагающие участие всех (или почти всех) граждан определённой страны. На начальном этапе своего становления выборка (как, впрочем, и всё новое, непривычное) снискала себе славу ненадежного источника эмпирических данных, а исследователи, которые прибегали к её использованию, считались людьми ленивыми и недобросовестными в профессии. Но об этом по порядку.

В России история выборочного метода берёт свое начало в царское время: реформой 1864 г. были введены земства — органы местного самоуправления. Круг задач, решаемых земствами, охватывал такие сферы, как просвещение, здравоохранение и др. Наряду с этим важной социальной функцией деятельности земств были сбор и систематизация данных о жизни крестьянских дворов (хозяйств), структуре их быта.

Нужно отметить, что земствами была создана весьма развитая инфраструктура: статистические исследования проводились во многих губерниях, а крестьянские хозяйства изучались всесторонне. До того как в России были учреждены земства, информация о жизни населения собиралась Центральным статистическим комитетом при Министерстве внутренних дел.

Основным методом сбора данных на протяжении длительного времени были переписи. Заметим, что в царской России не так уж часто и регулярно проводились переписи[1]. Переписи, хоть и позволяли охватить довольно широкий спектр вопросов о крестьянских хозяйствах, ввиду серьёзных ресурсных ограничений (финансовых, временны́х, кадровых), не давали возможности проводить детализированные исследования. Поэтому данные переписей позволяли получить лишь ориентировочное и отчасти поверхностное представление об укладе жизни населения. В результате те данные, которые собирались на основе сплошных обследований, не отличались глубиной и детализацией, а потому не позволяли сделать исчерпывающие выводы применительно к отдельным губерниям и областям.

---

[1] Вообще реализация сплошных исследований возможна далеко не всегда и встречается на практике редко. Это связано с колоссальными издержками различного характера (материальные, временны́е затраты и пр.). Так, национальные переписи населения во многих странах мира и в настоящее время проводятся не чаще одного раза в 10 — 15 лет.

Земская реформа открыла новые возможности сбора эмпирических сведений о крестьянских хозяйствах.

Деятельность земств осуществлялась на территории разных губерний. Это позволило нивелировать ограничения переписей: теперь на достаточно регулярной основе проводились эмпирические исследования крестьян.

Пик активности этих исследований пришёлся на конец XIX в. Переписи позволили получать более или менее регулярные сведения о повседневной жизни людей.

Необходимо отметить, что к тому времени сложились многие предпосылки к тому, чтобы уйти от сплошных подворных переписей как основного источника сбора информации. Этому послужил *ряд причин*[1]. Коротко перечислим их.

Плохие дороги России (или вовсе их отсутствие) не позволяли добраться до некоторых населённых пунктов.

Ещё одним важным фактором стало *резкое сокращение финансирования* со стороны государства; причиной этому послужил страх правительства перед нарастающим революционным движением, а также активной деятельностью земств, подстёгнутый трагическими событиями 1881 г., когда в результате террористического акта погиб царь Александр II.

«Чрезмерно» активная исследовательская работа земств настораживала и вызывала у власти тревожные опасения, что ситуация, вероятно, выходит из-под контроля, и возможностей влиять на то, что делают земства на местах, становится всё меньше. События того времени вынуждали земских статистиков искать другие инициативы, пути и тактики, реализуя социальные обследования населения.

Важным стало и то, что в те времена произошли *серьёзные прорывы в науке*: академические достижения математиков, развитие теории вероятностей послужили тому, что была открыта новая страница в методологии прикладных исследований.

С конца XIX в. до 1916 — 1917 гг. земскими исследователями было реализовано немалое число выборочных опросов населения. И этот первый опыт проектирования выборки оказался чрезвычайно успешным.

Статистиками закладывались основы работы целой *научной школы*. Именно в то время были сформулированы многие идеи механического отбора и стратифицированной (районированной) вы-

---

[1] Подробно о предпосылках возникновения выборочного метода и о методологии земских исследований см.: *Ермолаев А. В.* К вопросу о методике земских статистических исследований / *А. В. Ермолаев*, *И. В. Забаев* // Социологические исследования. — 2001. — № 11. — С.129 — 135; *Абрамов В. Ф.* Земская корреспондентская сеть в Московской губернии // Социологические исследования, 1996. — № 5. — С. 129 — 132.

борки[1]. Ряд исследований проводился параллельно как на сплошном массиве данных (с помощью подворных переписей), так и на основе выборочных опросов населения. Сравнивая результаты, исследователи на своём опыте постепенно убеждались в высоком качестве и достоверности информации, которую обеспечивает выборка. И по этим критериям она не уступает сплошным обследованиям генеральной совокупности. ***Скептицизм и опасения за качество данных постепенно сменялись доверием и интересом***.

В России к 1917 г. планировалась серьёзная работа по модернизации исследовательских технологий, преобразованию процедур сбора данных на территории страны, но Временное правительство в 1918 г. подписало декрет о ликвидации земств, в результате их деятельность была упразднена и они довольно быстро *утратили свои функции*. Так и *не получили развития* и широкой огласки достижения российских исследователей того времени. Об опыте земских статистиков было мало что известно в мире. Активный интерес на Западе к их методологии, идеям в реализации достоверных выборочных обследований населения на территории столь большой страны появился сравнительно недавно — с середины 1990-х годов. И, возможно, если бы об истории можно было говорить в сослагательном наклонении, то становление выборочного метода в социальных науках могло бы пойти совершенно по другому пути...

На Западе первая волна исследований, в основе которых лежал выборочный метод, пришлась на 1950—1960-е годы. Статистическое бюро при Организации Объединенных Наций в 1960 г. в издании «Sample Survey of Current Interest» впервые опубликовало перечень актуальных выборочных обследований, проведённых на территории нескольких десятков стран. Данные этих исследовательских проектов, которые носили национальный масштаб, охватывали самые разнообразные социально-экономические темы: от классических исследований безработицы и проблем промышленности до изучения стиля

---

[1] Крайне любопытны работы Александра Аркадьевича Кауфмана (1864—1919) — русского экономиста и исследователя, который посвятил многие годы своей жизни изучению крестьянских дворов в Сибири. Можно ознакомиться, например, с такими его работами, как: *Кауфман А. А.* К вопросу о выборочном исследовании // Теория и методы статистики. — 3-е изд., 1916; *Кауфман А. А.* Статистическая наука в России. Теория и методология. 1806—1917 гг. Историко-критический очерк. — М., 1922; *Кауфман А. А.* Статистика, её приёмы и её значение для общественных наук. — М.: Научное изд., 1912; *Кауфман А. А.* Крестьянская община в Сибири по местным исследованиям 1886—1892 гг.. — СПб., 1897; *Кауфман А. А.* Крестьянское землепользование и хозяйство Тобольской и Томской губерний. — СПб., 1894. См. также работы Владимира Густавовича Громана (1874—1940) — статистика и экономиста — о стратифицированной выборке (выборке типичных единиц), например: *Громан В. Г.* О применении типического метода в текущей статистике // Тр. Императорского вольного экономического общ-ва. — СПб.: Стат. комиссия вольного экон. общ-ва., февраль 1900.

жизни разных категорий людей, исследований в области рекламы и медиа.[1] Помимо этого выборка стала постоянной спутницей переписей населения в разных странах. Часть дополнительных опросных листов была адресована не всем гражданам, а лишь части из них — выборке. После этого выборочные данные экстраполировались (т. е. переносились, обобщались) с выборки на всю генеральную совокупность. В итоге выборочные исследования существенным образом дополняли переписи и помогали государствам собирать детальную, развёрнутую информацию о повседневной жизни людей. Начиная со второй половины XX в. стали постепенно появляться исследовательские организации, в чьи компетенции входило профессиональное проведение выборочных обследований. Поначалу воспринятая в штыки, выборка (как метод сбора данных) рано или поздно сумела заслужить авторитетное к себе отношение со стороны специалистов различных профессий.

---

[1] Подробно об истории выборочного метода на Западе можно прочитать в классической работе Уильяма Кокрена, см.: *Кокрен У.* Методы выборочного исследования. — М.: Статистика, 1976. — С. 15—24.

Глава 2

МЕТОДЫ ОПИСАТЕЛЬНОЙ СТАТИСТИКИ

В этой главе речь пойдет о том, *что* необходимо делать в первую очередь, приступая к анализу данных, собранных в ходе исследования; какие *главные* показатели, характеризующие эти данные, следует *рассчитать*. Будем учиться рассчитывать для выборки такие показатели, как среднее арифметическое, медиана и мода, дисперсия, стандартное отклонение и коэффициент вариации, квартили и квартильный размах, деци́ли и коэффициент деци́льной вариации, а также визуализировать информацию с помощью графиков.

§ 2.1. Вариационный ряд. Графическое представление данных

Итак, основным понятием, с которого начинается любое математико-статистическое исследование, является выборочная совокупность, состоящая из отдельных наблюдений. Для работы с данными нам потребуется ещё несколько базовых понятий.

Первоначальным рядом будем называть такую совокупность данных (в частности, ответов респондентов на поставленный вопрос анкеты), которые записаны в том порядке, как они были собраны (получены) в ходе исследования (интервьюирования, наблюдения и пр.).

Эта запись не предполагает *никакой сортировки или упорядочения* ответов (по убыванию или возрастанию, совпадению или различию). *Отдельные значения* исследуемой характеристики, как правило, являются числовыми и их принято называть вариантами.

Варианта — это отдельное значение переменной.

Обозначают буквой x с нижними индексами 1, 2, 3, ..., соответствующими конкретным значениям этой характеристики: x_1, x_2, x_3, ..., x_n.

Количество наблюдений — это размер выборки (n).

Упорядоченные* (*отсортированные*)** в порядке возрастания или убывания значения первоначального ряда превращают его в так называемый ***ранжированный ряд.

Число, в абсолютном выражении показывающее, *сколько раз* (т. е. *как часто*) в выборочной совокупности встречается то или иное значение признака (т. е. вариáнта), будем называть ***частотой*** (или ***абсолютной частотой***) f_i: $f_1, f_2, f_3, \ldots, f_k$. Сумма всех частот равна общей численности наблюдений выборки.

Число, в *относительном* выражении показывающее, *как часто* в выборочной совокупности встречается то или иное значение признака, называется ***относительной частотой*** (или ***весом***) и обозначается как w_i: $w_1, w_2, w_3, \ldots, w_k$.

Относительная частота каждой варианты ряда указывается *в долях или процентах* от общего числа наблюдений, содержащихся в выборке. Сумма всех относительных частот (весов) равна единице (1) или 100 %.

Вариационный ряд — это упорядоченный (ранжированный) ряд значений, в котором каждой варианте противопоставлена частота (относительная частота), с которой она встречается (присутствует) в выборке.

Пример 2.1. «ТВиМС». Пусть в результате итоговой аттестации знаний студентов (на одном из факультетов) по курсу «Теория вероятностей и математическая статистика» (ТВиМС), успешно сдавших экзамен (оценки ≥ 4 баллов по 10-балльной системе), преподаватель получил такие данные (ниже представлен фрагмент):

№ п/п	Студент	Оценка	Интерпретация
1	А	6	Хорошо
2	Б	4	Вполне удовлетворительно
3	В	5	Удовлетворительно
4	Г	4	Вполне удовлетворительно
5	Д	10	Блестяще
6	Е	8	Почти отлично
7	Ж	8	То же
8	З	7	Очень хорошо
9	И	9	Отлично
10	К	5	Удовлетворительно
11	Л	6	Хорошо

Перед нами выборка (X, оценки), которая представлена значениями: **6, 4, 5, 4, 10, 8, 8, 7, 9, 5, 6.** Первоначальный ряд состоит из 11 наблюдений. Чтобы перейти к упорядоченному ряду значений, выстроим имеющиеся наблюдения, например, в порядке возрастания. Получим: **4, 4, 5, 5, 6, 6, 7, 8, 8, 9, 10**. Для того чтобы получить вариационный ряд, составим таблицу (каждому значению X поставим в соответствие свою абсолютную или относительную частоту):

X_i	**4**	**5**	**6**	**7**	**8**	**9**	**10**
f_i	2	2	2	1	2	1	1
Всего	11 человек						

Вариационный ряд может быть записан в таком виде:

X_i	**4**	**5**	**6**	**7**	**8**	**9**	**10**
w_i	2/11 (0,18)	2/11 (0,18)	2/11 (0,18)	1/11 (0,09)	2/11 (0,18)	1/11 (0,09)	1/11 (0,09)
Всего	1						

Ряд может отражать распределение дискретных или непрерывных по своей вариации признаков.

Дискретные величины отличаются от непрерывных тем, что принимают счётное (т. е. поддающееся калькуляции) число отдельных значений, которые мы наблюдаем в результате опыта (например, число детей в семье, число парт в лекционной аудитории, число автомобилей на университетской парковке, число яблок в корзине и др.).

Так, изменчивость отдельных значений ряда может быть рассчитана и охарактеризована *конкретной величиной*, которая *конечна*.

Непрерывными величинами называются такие величины, изменчивость отдельных значений которых может отличаться на крайне малую величину и не поддаваться конечной оценке со 100%-й точностью.

Например, ежедневная выручка магазина (может фиксироваться с точностью до рубля или копейки), возраст человека (с точностью до числа дней, часов, минут), масса человека (с точностью до килограмма, грамма, миллиграмма) и многое другое.

Рассмотрим основные методы ***визуализации*** вариационных рядов. К ним можно отнести построение ***полигона распределения*** *и* ***гистограммы***, а также ***кумулятивной кривой*** **(*кривой сумм*)**.

Визуализация вариационных рядов с помощью полигона распределения или гистограммы реализуется просто. Рисуются две оси — горизонтальная и вертикальная. По горизонтальной оси откладываются

значения признака в виде его отдельных, дискретных значений или в виде интервалов значений; по вертикальной оси — частоты (или относительные частоты, веса), соответствующие каждому из дискретных значений (или интервалу значений) рассмотренной величины. Как правило, полигон распределения строится для вариационного ряда, содержащего дискретную величину, и в котором число вариант относительно невелико, а гистограмма — для вариационного ряда, содержащего непрерывную величину. При построении гистограммы по горизонтальной оси откладываются не дискретные значения величины, а диапазоны её значений (интервалы).[1] Иногда гистограмма строится и для дискретной величины: это происходит в том случае, если число вариант исследуемого признака довольно велико и исходные значения признака разбивают на диапазоны (интервалы) для лучшей визуализации, иллюстрации данных. На практике часто эти два метода визуализации — полигон распределения и гистограмму — совмещают на одной картинке. Соединяя вершины столбцов гистограммы ломаной линией, получают полигон распределения.

Пример 2.2. «ТВиМС-2». Дополним *пример 2.1* эмпирическими данными. Приведём вариационный ряд, характеризующий распределение оценок 78 студентов, успешно аттестованных по курсу «Теория вероятностей и математическая статистика» на факультете социологии НИУ «Высшая школа экономики» в 2011 — 2012 учебном году:

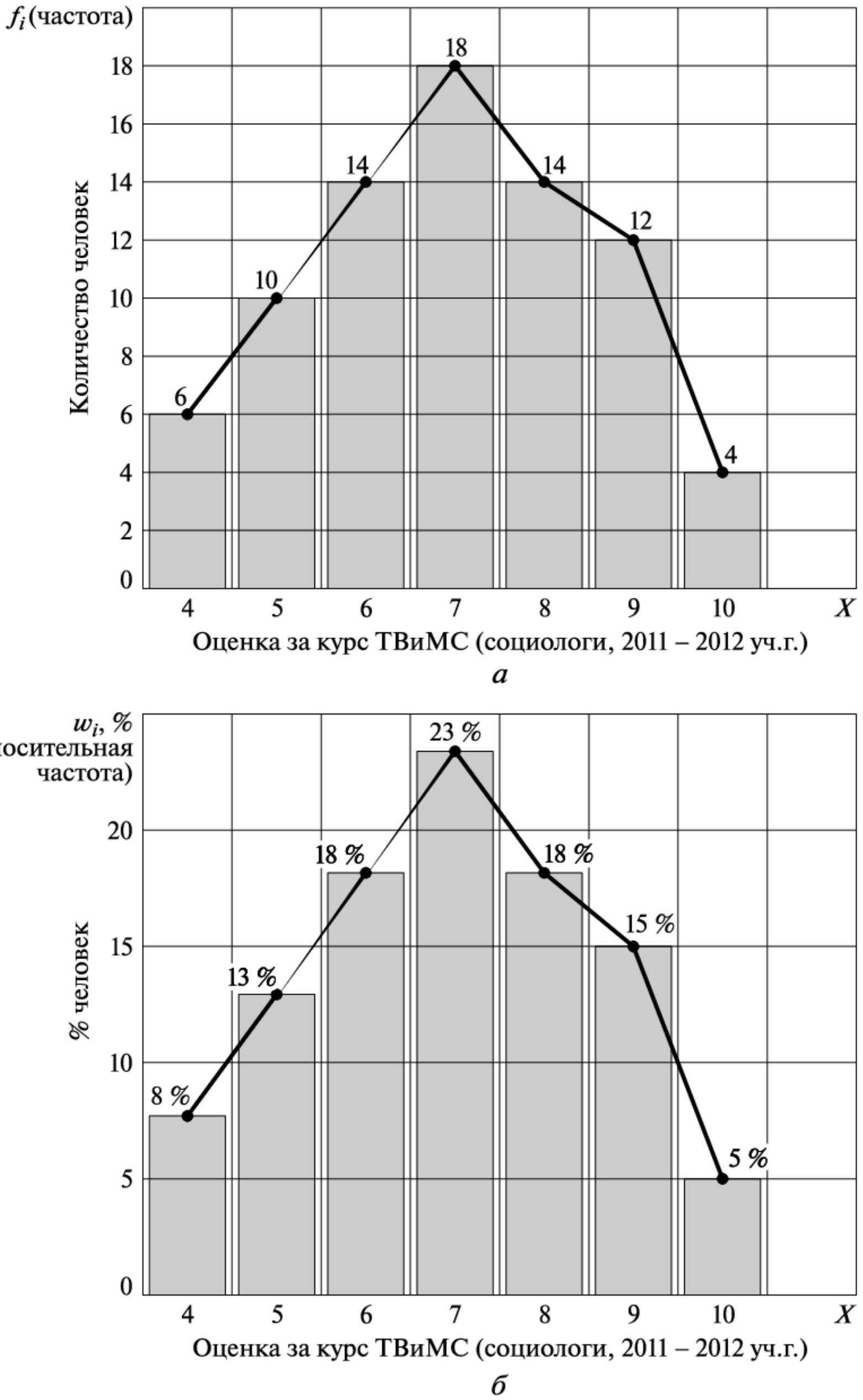
X_i	**4**	**5**	**6**	**7**	**8**	**9**	**10**
f_i	6	10	14	18	14	12	4
Всего	78 человек						

Вариационный ряд может быть записан в таком виде:

X_i	**4**	**5**	**6**	**7**	**8**	**9**	**10**
w_i	6/78 (0,08)	10/78 (0,13)	14/78 (0,18)	18/78 (0,23)	14/78 (0,18)	12/78 (0,15)	4/78 (0,05)
w_i ,%	8	13	18	23	18	15	5
Всего	1 (или 100 %)						

Гистограмма распределения оценок (столбцы-прямоугольники) и полигон (в виде ломаной линии), позволяющие визуализировать имеющиеся в нашем распоряжении данные, изображены на рис. 2.1, *а*, *б*.

[1] Хотя, конечно, это требование не исключает возможности построения гистограммы для дискретного вариационного ряда. В этом случае гистограмма просто превратится в обычную столбчатую диаграмму.



а

б

Рис. 2.1. Гистограмма и полигон распределения оценок студентов

Наконец, ещё одним способом изображения вариационных рядов, служит так называемая ***кумулятивная кривая*** (*кривая сумм*, или *кривая скопления*). Она позволяет изобразить вариационный

ряд в виде ряда ***накопленных значений исследуемого признака***. По горизонтальной оси (оси абсцисс) откладываются значения исследуемой величины, а по вертикальной (оси ординат) — накопленные частоты (накопленные относительные частоты). Построение кумулятивной кривой опирается на расчётную таблицу.

Вернёмся к *примеру 2.2* и подготовим данные для визуализации с помощью кривой сумм:

X_i	**4**	**5**	**6**	**7**	**8**	**9**	**10**
f_i	6	10	14	18	14	12	4
$f_{i\,(\text{накопл})}$	**6**	10 + 6 = = **16**	6 + 10 + + 14 =16 + + 14 = **30**	6 + 10 + + 14 + 18 = = 30 + 18 = = **48**	6 + 10 + + 14 + 18 + + 14 = 48 + + 14 = **62**	6 + 10 + + 14 + 18 + + 14 + 12 = = 62 + 12 = = **74**	6 + 10 + + 14 + 18 + + 14 + 12 + + 4 = 74 + + 4 = **78**
Всего	78 человек						

Кумулятивное распределение может быть записано так:

X_i	**4**	**5**	**6**	**7**	**8**	**9**	**10**
w_i	0,08	0,13	0,18	0,23	0,18	0,15	0,05
$w_{i\,(\text{накопл})}$	**0,08**	**0,21**	**0,39**	**0,62**	**0,80**	**0,95**	**1,00**
$w_{i\,(\text{накопл})}, \%$	**8**	**21**	**39**	**62**	**80**	**95**	**100**
Всего	1 (или 100 %)						

Расчёт накопленных (или накопленные относительных) частот позволяет зафиксировать, сколько человек (студентов) в рассматриваемой совокупности имеет оценку, ***меньшую определённого значения или равную ему***. Так, процент студентов, которые получили оценку 7 баллов и ниже (т. е. 4, 5, 6 или 7), составляет 62 % всей совокупности. Доля студентов, получивших оценку не более 5 баллов (т. е. 4 или 5), равна 21 %. Кумулята (столбцы-прямоугольники и ломаная линия, соединяющая их вершины) изображена на рис. 2.2, *а*, *б*.

Пример 2.3. «Опрос курильщиков». Пусть в некотором исследовании, которое проводилось среди курильщиков, в ходе опроса респондентам мужского пола был задан вопрос «Сколько примерно сигарет Вы выкуриваете в течение дня? Укажите, пожалуйста, числом». Было зафиксировано распределение количества сигарет, выкуриваемых мужчинами (объём выборки составил 520 человек). Поскольку в исследовании рассматривается такая дискретная величина, как количество сигарет, выкуриваемых мужчиной в течение

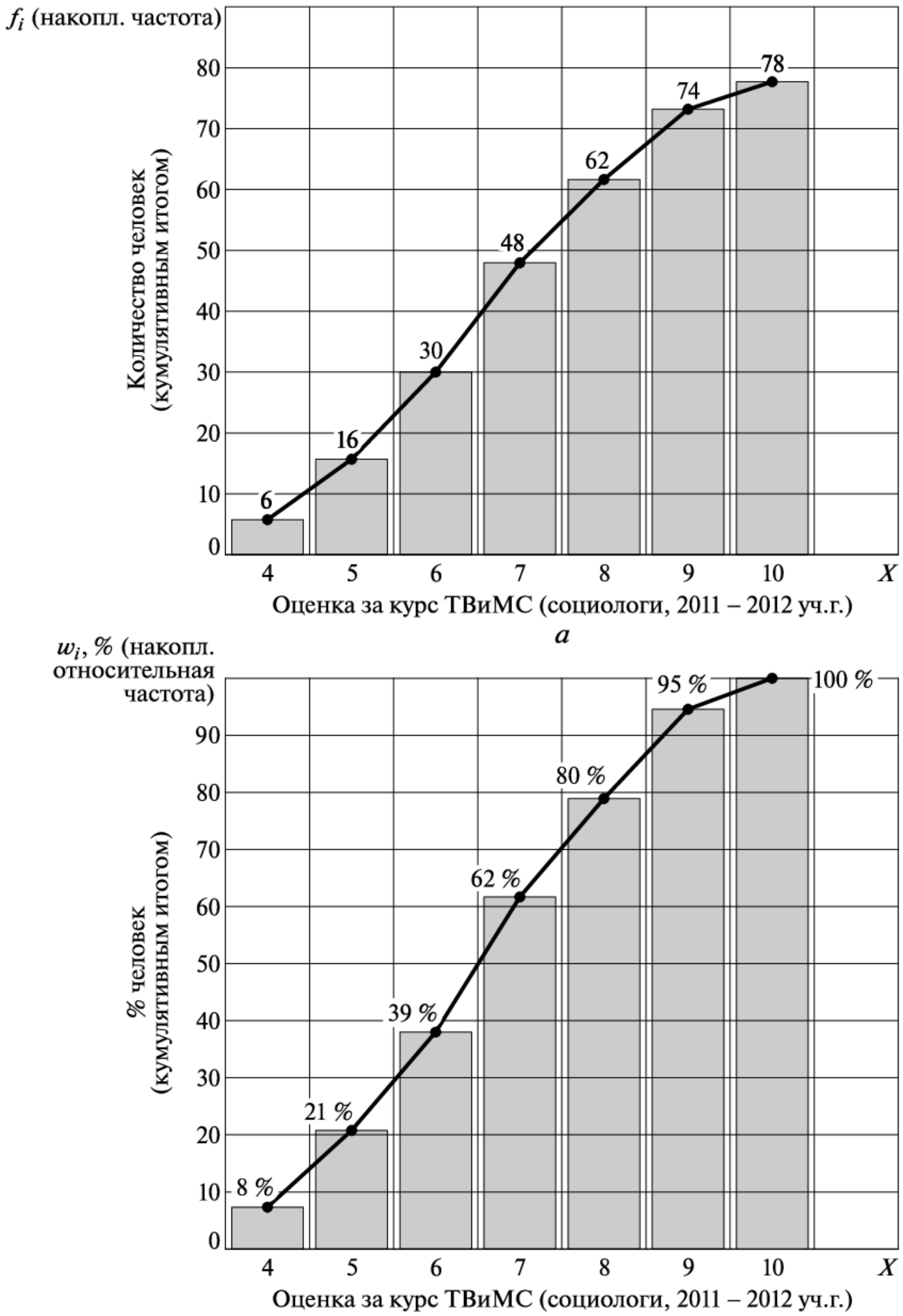


Рис. 2.2. Кумулята распределения оценок студентов

дня, то очевидно, что она может принимать довольно много значений: 1, 2, 3, 4, ..., 15, 16, 17, ... (верхняя граница, конечно, разумна). Поэтому, чтобы иллюстрировать данные с помощью гистограммы,

можно исходные значения разбить на диапазоны (интервалы) — для лучшей визуализации:

X_i	**1 — 3**	**4 — 6**	**7 — 9**	**10 — 12**	**13 — 15**	**16 — 18**	**19 — 21**	**22 — 24**	**25 и более**
f_i	15	23	28	46	49	185	161	11	2
Всего	520 человек								

Вариационный ряд может быть записан и в таком виде:

X_i	**1 — 3**	**4 — 6**	**7 — 9**	**10 — 12**	**13 — 15**	**16 — 18**	**19 — 21**	**22 — 24**	**25 и более**
w_i	15/520 (0,029)	23/520 (0,044)	28/520 (0,054)	46/520 (0,088)	49/520 (0,094)	185/520 (0,356)	161/520 (0,310)	11/520 (0,021)	2/520 (0,004)
w_i, %	2,9	4,4	5,4	8,8	9,4	35,6	31,0	2,1	0,4
Всего	1 (или 100 %)								

Гистограмма распределения количества сигарет (столбцы-прямоугольники) и полигон (в виде ломаной линии), позволяющие визуализировать имеющиеся в нашем распоряжении данные, изображены на рис. 2.3.

Рис. 2.3. Гистограмма и полигон распределения количества сигарет

Подготовим данные для визуализации с помощью кривой сумм (кумуляты). Получим расчётную таблицу:

X_i	**1—3**	**4—6**	**7—9**	**10—12**	**13—15**	**16—18**	**19—21**	**22—24**	**25 и более**
f_i	15	23	28	46	49	185	161	11	2
f_i (накопл)	15	38	66	112	161	346	507	518	520
Всего	520 человек								

Кумулятивное распределение может быть записано так:

X_i	**1—3**	**4—6**	**7—9**	**10—12**	**13—15**	**16—18**	**19—21**	**22—24**	**25 и более**
w_i, %	2,9	4,4	5,4	8,8	9,4	35,6	31,0	2,1	0,4
w_i (накопл), %	2,9	7,3	12,7	21,5	30,9	66,5	97,5	99,6	100
Всего	100 %								

Расчёт накопленных относительных частот позволяет зафиксировать, сколько человек в рассматриваемой совокупности выкуривает число сигарет, меньшее или равное определённого диапазона

Рис. 2.4. Кумулята распределения количества сигарет

значений. Так, например, процент мужчин, которые в течение дня выкуривают 13 — 15 сигарет и менее (от 1 до 15 шт.) составляет 30,9 % всей совокупности. Доля мужчин, которые в течение дня выкуривают не более 4 — 6 сигарет (от 1 до 6 шт.), равна 7,3 %. Кумулятивная кривая (столбцы-прямоугольники и ломаная линия, соединяющая их вершины) изображена на рис. 2.4.

Итак, кумулятивное распределение позволяет увидеть, сколько наблюдений (единиц) в исследуемой совокупности имеют значение признака, которое не превышает рассматриваемое. По своей форме кумулята напоминает ***S-образную кривую***, показывающую ***скорость приращения частот (относительных частот) при изменении значений вариационного признака на определённую величину***.

Рассмотренные графические методы изображения вариационных рядов позволяют наглядно представить сведения, содержащиеся в статистических таблицах, и визуально охарактеризовать в агрегированном (обобщённом) виде распределение исследуемых величин.

§ 2.2. Меры центральной тенденции

В математике выделяют три основных показателя, с помощью которых можно охарактеризовать положение центра исследуемой совокупности. Ими выступают медиана, среднее арифметическое и мода. Рассмотрим каждую из мер центральной тенденции в отдельности.

Медианой **(Me)** называют такое ***число***, ***которое делит упорядоченное множество данных пополам***, ***на две равные по своей численности части***.

Иными словами, это такое число, которое ***находится в середине распределения***, ***упорядоченного от минимума к максимуму***. *Ровно половина* (50 %) всех наблюдений выборки имеют значения, *бóльшие этого срединного значения либо равные ему*, и *ровно половине* (50 %) оставшихся наблюдений выборки свойственны значения, *меньшие либо равные этому срединному значению*.

Выборка может состоять как из чётного, так и нечётного числа наблюдений, поэтому выделяют понятия чётного упорядоченного ряда и нечётного упорядоченного ряда. ***Чётным*** называется ряд, содержащий чётное количество наблюдений (т. е. ряд, в котором количество наблюдений в выборке равно чётному числу); ***нечётным*** — ряд, содержащий нечётное число наблюдений (т. е. ряд, в котором количество наблюдений в выборке равно нечётному числу).

- В случае ***нечётного*** числа наблюдений в упорядоченном ряду медиане будет соответствовать наблюдение под порядковым номером $\frac{(n+1)}{2}$.

• В случае ***чётного*** числа наблюдений в упорядоченном ряду медиана будет равна полусумме соседних вариант, делящих между собой «серединное» место: $\text{Me} = 0{,}5\left(X_{\frac{n}{2}} + X_{\frac{n}{2}+1}\right)$.

Пример 2.4. «Покупка воды». Компания, продающая бутилированную воду, оценивает, каковы еженедельные расходы её постоянных клиентов на покупку воды. Выборка, состоящая из 11 элементов, представлена следующими значениями:

X (руб.): **500, 700, 700, 500, 600, 4 000, 3 000, 1 000, 800, 700, 600**.

Найдите медиану выборки.

Решение. Сначала выстроим по возрастанию все данные по клиентам компании. Получим упорядоченный ряд: **500, 500, 600, 600, 700, 700, 700, 800, 1 000, 3 000, 4 000.** Этот ряд нечётный, так как состоит из 11 элементов (т. е. нечётного числа наблюдений). Медиане соответствует наблюдение с порядковым номером 6 — это значение 700 руб., так как оно делит упорядоченное множество имеющихся данных ровно на две половины, содержащие по 5 наблюдений каждая:

500, 500, 600, 600, 700 , **700**, **700, 800, 1 000, 3 000, 4 000**

Значение 700 показывает, чему равен центр совокупности исследуемых данных о еженедельных расходах на покупку воды. При этом обратите внимание на ***особенность интерпретации медианы***: половина клиентов в выборке еженедельно расходуют на покупку бутилированной воды ***не более*** *700 руб.* (*т. е. 700 руб. и меньше*). Можно сказать иначе: половина клиентов в выборке еженедельно тратят на покупку воды ***не менее*** *700 руб.* (*т. е. 700 руб. и больше*).

Пример 2.5. «Товары повседневного спроса». Жителям (членам домашних хозяйств) из крупного города был адресован вопрос: «Какую часть семейного бюджета Вы тратите на покупку продуктов питания и предметов первой необходимости? Укажите, пожалуйста, примерную долю этих расходов в процентах». Ответы первых 12 респондентов представлены так:

X (%): **40, 60, 25, 15, 10, 10, 20, 40, 30, 35, 40, 85**.

Найдите медиану выборки.

Решение. Итак, выборка содержит 12 наблюдений. Напомним, что первое, с чего следует начинать решение задач на нахождение медианы, — это упорядочение имеющихся значений выборки по возрастанию — от минимума к максимуму. Тогда получим упорядоченный ряд: **10, 10, 15, 20, 25, 30, 35, 40, 40, 40, 60, 85**. В данном примере имеем чётный ряд, состоящий из 12 наблюдений. Следовательно, право называться медианой делят между собой два соседних значения: 30 % и 35 % — с порядковыми номерами 6 и 7. Медианой будет значение,

равное их полусумме: $(30 + 35) \cdot 0{,}5 = 32{,}5\,\%$, так как именно значение 32,5 % делит упорядоченное множество значений пополам:

10, 10, 15, 20, 25, 30, Me = ?, 35, 40, 40, 40, 60, 85

Интерпретировать медиану следует таким образом: половина всех домохозяйств в выборке тратит на покупку товаров повседневного спроса ***не более*** *32,5 %* своего семейного бюджета, или иначе: половина всех домохозяйств тратит на покупку товаров повседневного спроса ***не менее*** *32,5 %* своего семейного бюджета. Можно сформулировать и такой вывод: 50 % домашних хозяйств расходуют на покупку товаров первой необходимости ***не менее*** 32,5 % семейного бюджета.

Пример 2.6. «Летние каникулы». Учитель начальной школы выборочно спросил учеников своего класса, сколько произведений художественной литературы они прочитали во время летних каникул. Вот их ответы: 5, 10, 0, 2, 3, 2, 0, 6, 3, 5, 6, 4, 7 (книг). Оцените по данной выборке медиану распределения прочитанных школьниками за лето книг. Интерпретируйте полученный ответ.

Решение. Случайная величина (переменная X) в этом примере — число книг, прочитанных учащимся за время летних каникул. Ответы школьников представлены вперемешку. Их следует упорядочить. Получим ранжированный ряд: **0, 0, 2, 2, 3, 3, 4, 5, 5, 6, 6, 7, 10**. Он содержит нечётное количество наблюдений (13). Следовательно, медианой будет значение под номером 7 — это значение «4»:

0, 0, 2, 2, 3, 3, 4, 5, 5, 6, 6, 7, 10

В результате можно сказать, что половина всех учащихся, составляющих выборку, прочитали за летние каникулы ***не более*** 4 книг, т. е. 4 книги и менее. Либо можно сказать, что другая половина школьников прочитали на каникулах ***не менее*** 4 книг, т. е. 4 книги и более.

Пример 2.7. «Безработные». Пусть нам известны сведения о численности безработных определенной профессии в населённых пунктах некоторого региона за прошлый год:

Населённый пункт	А	Б	В	Г	Д	Е	Ж	З	И	К	Л	М
Количество безработных	1 100	910	1 030	1 080	990	1 100	1 000	970	980	920	1 180	1 240

Найдите по данной выборке медиану распределения численности безработных в рассматриваемом регионе.

Решение. Исследуемая переменная X — численность безработных в одном населённом пункте. Выборка состоит из 12 населённых пун-

ктов. Чтобы определить медиану, необходимо упорядочить имеющиеся 12 значений. Перейдём от выборки к ранжированному ряду:

910, 920, 970, 980, 990, 1 000, **Me=?**,

1 030 , 1 080, 1 100, 1 100, 1 180, 1 240

Теперь можем найти медиану как полусумму чисел 1 000 и 1 030. Получим, что медиана равна 1 015 (человек). Следовательно, вывод будет таким: в половине населённых пунктов рассматриваемого региона численность безработных ***не превысила*** 1 015 человек (т. е. 1 015 и меньше), либо таким: в половине населённых пунктов рассматриваемого региона численность безработных определенной профессии была ***не меньше*** 1 015 человек (т. е. 1 015 и больше). Число 1 015 — это медиана распределения.

В социальных науках иногда встречается и другое название медианы —***позиционное среднее***, поскольку это значение явно указывает на среднюю позицию в имеющихся эмпирических данных. В практике исследований, посвящённых изучению социально-экономического неравенства, социального расслоения по доходам, часто встречается и такое понятие, как «*медианный доход*». Им исследователи обозначают *срединную* величину среднедушевых доходов населения, которая служит индикатором материального положения населения. Наряду с этим в практике маркетинговых исследований специалисты могут использовать понятие «*медианная цена*» — при изучении ценовой эластичности спроса, например, или анализе истории продаж товаров определённой категории. ***Медиана очень информативна***. Этим во многом и объясняется популярность её использования как меры центральной тенденции.

Среднее арифметическое **($\bar{X}$)** — основной и наиболее часто применяемый на практике показатель центральной тенденции, наблюдаемой в имеющейся выборке данных. Среднее арифметическое (или просто ***среднее***) является аналогом *центра тяжести* или *равновесия* (наверняка, многие из вас хорошо помнят эти понятия из курса школьной физики)[1]. Это образное сравнение происходит вот по какой причине. Дело в том, что если рассчитать разность между средним и каждым из значений в выборке, то сумма этих разностей (или отклонений) будет равна нулю. И, более того, если взять и рассчитать квадраты отклонений между значениями выборки и средним арифметическим, то их сумма будет *строго меньше* суммы квадратов отклонений этих значений *от какой-либо другой точки* (взятой взамен среднего). Это и делает среднее арифметиче-

---

[1] Вспомните, что центром тяжести в физике называется геометрическая точка, которая связана с твёрдым телом и через которую проходит *равнодействующая всех сил тяжести*, действующих на частицы этого тела. Данная точка вовсе не обязательно совпадает с точками этого тела.

ское столь уникальным показателем, позволяющим говорить о своего рода центре тяжести, наблюдаемом в данных.

Среднее арифметическое вычисляется по формулам (две записи абсолютно идентичны):

$$1)\ \bar{X} = \frac{\sum_{i=1}^{n} x_i}{n};$$

$$2)\ \bar{X} = \frac{\sum_{i=1}^{k} x_i f_i}{n},$$

где i — номер наблюдения в выборке; n — размер выборки; x_i — определенное значение (вариáнта) переменной (признака); f_i — частота встречаемости значения x_i в выборке наблюдений (указывающая на удельный вес отдельных значений во всей совокупности); k — количество различных значений, которые принимает X.

Обратите внимание, что размер выборки равен сумме всех частот $n = \sum_{i-1}^{k} f_i$, т. е. сумме частот ряда распределения, с которыми отдельные вариáнты встречаются в совокупности. Вторая формула является более компактной записью первой, в случае если в таблице каждому из значений переменной X уже противопоставлена своя частота встречаемости значений в выборке наблюдений.

Пример 2.8. «Покупка воды (продолжение)». Вернёмся к рассмотренному ранее примеру (см. пример 2.4). Производственно-торговая компания оценивает, каковы еженедельные расходы её клиентов на покупку бутилированной воды. Выборка, состоящая из 11 элементов, представлена следующими значениями:

X (руб.): **500, 700, 700, 500, 600, 4 000, 3 000, 1 000, 800, 700, 600**.

Найдите среднее арифметическое выборки.

Решение.

$$\bar{X} = \frac{500 + 700 + 700 + 500 + 600 + 4\,000 + 3\,000 + 1\,000 + 800 + 700 + 600}{11} \approx$$

$$\approx 1\,190{,}9 \text{ руб.}$$

В среднем каждый клиент затрачивает 1 190,9 руб. в неделю на покупку бутилированной воды. При этом обратите внимание, что медиана равна 700 руб. Среднее и медиана далеко не всегда совпадают. И у каждой из этих мер центральной тенденции — *собственный содержательный смысл*.

Пример 2.9. «Туризм». Пусть проводилось исследование, в рамках которого респондентов просили указать количество зарубежных

стран, в которых каждый из них побывал за последние пять лет. Полученные данные следующие:

Число стран	1	2	3	4	5	6	7	8	9	10	11	12	17	18	24	25	30	36	42
Количество респондентов	2	10	9	3	4	5	3	1	1	1	2	6	1	8	16	2	6	3	1

Найдите среднее арифметическое.

Решение. Из таблицы видно, что в опросе приняло участие 84 человека (что соответствует сумме частот $2+10+9+3+\ldots+6+3+1$). Число стран — это исследуемая характеристика (признак). Каждому из значений соответствует своя частота — сколько человек дали одинаковый ответ. Следовательно, среднее арифметическое находится как сумма значений, *взвешенных на свою частоту* (*вес*) и приведённых к общему объёму выборки:

$$\bar{X}=\frac{1\cdot 2+2\cdot 10+3\cdot 9+4\cdot 3+\ldots+30\cdot 6+36\cdot 3+42\cdot 1}{84}=\frac{1178}{84}\approx 14 \text{ (стран)},$$

т. е. *в среднем* один человек (из числа опрошенных) посетил за последние пять лет 14 зарубежных стран.

В ряде случаев исходные значения исследуемого признака могут быть разбиты на ***диапазоны (интервалы) — от ...и до ...***, а каждому из диапазонов соответствует своя частота. В такой ситуации, для того чтобы вычислить среднее арифметическое, в каждом из интервалов сначала потребуется взять свои ***центры (середины)***. Эти центры интервалов находятся как ***полусумма левой и правой границ интервала***. Так, если дан интервал $2-6$, то серединой (центром) этого интервала будет число 4 $\left(\frac{2+6}{2}=4\right)$. Итак, для случаев, когда исходная информация о распределении значений признака представлена в виде интервалов, среднее арифметическое определяется по формуле

$$\bar{X}=\frac{\sum_{i=1}^{k} x_{c_i} f_i}{n},$$

где c_i — центральное значение переменной (признака) в i-м интервале (т. е. середина интервала); n — размер выборки, равный $\sum_{i-1}^{k} f_i$; f_i — количество наблюдений, содержащихся в i-м интервале; k — количество интервалов, на которые разбит вариационный ряд.

Пример 2.10. «Канцелярские товары». Допустим, нас интересует средний объём продаж в течение выходного дня в магазинах

фирменной сети, продающей канцелярские товары (тетради, папки, офисные принадлежности и др.). Выборка состоит из 73 магазинов. Получены следующие данные:

Объём продаж выходного дня, X_i, тыс. руб.	10 — 20	20 — 30	30 — 40	40 — 50	50 — 60
Количество магазинов, f_i	5	12	22	24	10

По имеющимся данным требуется определить средний объём продаж магазина фирменной сети.

Решение. Исследуемый признак — объём продаж магазина фирменной сети в выходной день. Для расчёта среднего арифметического в данной задаче нам потребуется сначала определить центральное значение признака в каждом интервале (диапазоне). Напомним, что среднее значение интервала находится как полусумма левой и правой границ данного интервала. Тогда получим таблицу:

Объём продаж выходного дня, X_i, тыс. руб.	10 — 20	20 — 30	30 — 40	40 — 50	50 — 60
Середины интервалов, Xc_i	$\frac{10+20}{2}=15$	$\frac{20+30}{2}=25$	$\frac{30+40}{2}=35$	$\frac{40+50}{2}=45$	$\frac{50+60}{2}=55$

Исходные данные можно оформить в несколько дополненном виде:

Объём продаж выходного дня, X_i, тыс. руб.	10 — 20	20 — 30	30 — 40	40 — 50	50 — 60
Середины интервалов, Xc_i	$\frac{10+20}{2}=15$	$\frac{20+30}{2}=25$	$\frac{30+40}{2}=35$	$\frac{40+50}{2}=45$	$\frac{50+60}{2}=55$
Количество магазинов, f_i	5	12	22	24	10

Средний объём продаж магазина фирменной сети в выходной день равен:

$$\bar{X}=\frac{15\cdot 5+25\cdot 12+35\cdot 22+45\cdot 24+55\cdot 10}{5+12+22+24+10}=\frac{2\,775}{73}\approx 38 \text{ (тыс. руб.)}.$$

Обратите внимание на следующее обстоятельство: бывают случаи, когда имеются ***«открытые» интервалы***. Это такие интервалы, одна из границ которых — левая или правая — не закрыта.

«Открытыми» интервалами бывают первый и/или последний в вариационном ряду. Для вычисления центральных значений «открытых» интервалов сначала требуется задать их крайние границы, т. е. *условно* закрыть.

Для этого на каждый из открытых интервалов распространяется величина соседнего «закрытого» интервала:

- чтобы определить нижнюю границу первого интервала, следует посмотреть на шаг интервала, который характерен для последующего интервала;
- последний интервал в вариационном ряду закрывается в соответствии с тем, чему равен шаг интервала в предпоследнем интервале.

Пример 2.11. «Остров Сахалин». В произведении А. П. Чехова «Остров Сахалин (из путевых записок)» приводится следующий фрагмент: «К 1 января 1890 г. во всех трёх сахалинских округах состояло каторжных обоего пола 5730. Из них...»[1]:

Срок осуждения, лет	До 8	8 — 12	12 — 15	15 — 20	20 — 50	50 и более
Количество арестантов	2 124	1 567	747	731	175	386

Рассчитайте среднее арифметическое (найдите среднее количество лет ссылки).

Решение. В этой задаче требуется найти средний срок осуждения. Данные представлены в сгруппированном виде. И нам доступна только информация о диапазонах и их частотах.

В таких случаях требуется, во-первых, закрыть крайние границы первого и последнего интервалов, указанных в таблице.

А во-вторых, задать середины (центры) каждого из интервалов. После этого приступить к расчёту среднего арифметического.

Шаг интервала, следующего после первого интервала, равен 4 (12 – 8 = 4).

Следовательно, нижняя граница первого открытого интервала будет равна 4 (8 – 4 = 4). Шаг интервала, предшествующего последнему интервалу в вариационном ряду, равен 30 (50 – 20 = 30). Тогда и верхняя граница последнего интервала будет условно закрыта значением 80 (50 + 30 = 80).

Теперь расширим исходную таблицу, добавив еще одну строку — середины интервалов:

---

[1] *Чехов А. П.* Остров Сахалин. — Новосибирск: Сиб. унив. изд-во, 2009. — С. 146.

Срок осуждения, лет	4 — 8	8 — 12	12 — 15	15 — 20	20 — 50	50 — 80
Середины интервалов	$\frac{4+8}{2}=6$	$\frac{8+12}{2}=10$	$\frac{12+15}{2}=13{,}5$	$\frac{15+20}{2}=17{,}5$	$\frac{20+50}{2}=35$	$\frac{50+80}{2}=65$
Количество арестантов	2 124	1 567	747	731	175	386

Получим, что средний срок осуждения для каторжных составил:

$$\bar{X}=\frac{6\cdot 2\,124+10\cdot 1\,567+13{,}5\cdot 747+17{,}5\cdot 731+35\cdot 175+65\cdot 386}{2\,124+1\,567+747+731+175+386}=$$

$$=\frac{82\,506}{5\,730}\approx 14{,}4 \text{ лет.}$$

Вполне очевидны те ограничения, которые могут повлиять на корректность вычислений среднего арифметического. Наверняка, вы успели обратить внимание, что в случае вариационных рядов, где значения признака представлены в виде интервалов, центр (середина) диапазона могут в некоторой степени отличаться от фактического (истинного) среднего по интервалу. Это происходит в тех случаях, когда ***распределение отдельных вариант в пределах интервала окажется неравномерным***. Степень такого расхождения между расчётным значением среднего и его истинным значением обусловлена двумя основными причинами:

1) *способ* (*правило*) *построения интервалов*. Если в социальных обследованиях мы по каким-либо соображениям группируем исходные данные, то эти диапазоны должны быть выделены, *во-первых*, теоретически (опираясь на глубоко осмысленные, содержательные доводы), а, *во-вторых*, границы интервалов (левые и правые) следует выбирать так, чтобы максимум (пик) распределения на этом участке приходился, по возможности, на середину отрезка;

2) *шаг* (*величина*) *интервала*. Чем шире интервал (т. е. чем больше его шаг), тем более ощутимой и заметной может стать ошибка в вычислениях.

Заметим, что в некоторых задачах вариационные ряды могут содержать *относительные, а не абсолютные частоты*. В таких случаях относительные частоты — это удельный вес численности отдельных групп в выборке (доли или проценты к итогу, указывающие на соотношение между группами по числу наблюдений). Это естественным образом никак не влияет на результат расчёта среднего арифметического и не меняет значение среднего.

Пример 2.12. «Доходы». Пусть компания провела исследование с целью получить классификацию потребителей. Респондентам был

задан один из вопросов: «Каким был доход Вашей семьи в прошлом месяце в расчёте на одного члена семьи (включая детей)?». Фрагмент данных представлен в таблице:

Доход, тыс. усл. ед.	0 — 2	2 — 4	4 — 6	6 — 8	8 — 10	Итого
Середины интервалов	1	3	5	7	9	—
Количество респондентов, % к итогу	7	47	26	13	7	100

Средний доход семьи в расчёте на одного человека составляет:

$$\bar{X} = \frac{1 \cdot 7\,\% + 3 \cdot 47\,\% + 5 \cdot 26\,\% + 7 \cdot 13\,\% + 9 \cdot 7\,\%}{100\,\%} = \frac{432}{100} \approx 4{,}3 \text{ тыс. усл. ед.}$$

или запись можно оформить таким образом:

$$\bar{X} = \frac{1 \cdot 0{,}07 + 3 \cdot 0{,}47 + 5 \cdot 0{,}26 + 7 \cdot 0{,}13 + 9 \cdot 0{,}07}{1} = \frac{4{,}32}{1} \approx 4{,}3 \text{ тыс. усл. ед.}$$

Мода* (Mo).** Понятие «мода» красноречиво говорит само за себя: как и в повседневной жизни, мы ассоциируем слово «мода» с чем-то, что свойственно большинству, что является наиболее типичным (распространённым), так и в математике — модой называется такое значение изучаемого признака, которое ***встречается чаще всех остальных в рассматриваемой совокупности данных. Если, к примеру, взять некоторую совокупность значений: 6, 7, 5, 5, 2, 5, 8, 1, то модой будет значение 5, так как оно встречается чаще других. Теперь решим более содержательную задачу.

Пример 2.13. «Театр». Пусть в некотором городе в преддверии открытия театра, находившегося на реставрации, супружеским парам задавали вопрос: «Сколько примерно раз в год Вы ходите в театр?». В опросе участвовало 20 супружеских пар. Выборка представлена следующими ответами: 0, 2, 2, 1, 0, 8, 3, 3, 4, 3, 0, 3, 5, 3, 2, 3, 1, 5, 3, 0. Чему равна мода в этой выборке ответов?

Решение. Случайная величина, рассматриваемая в задаче, — число посещений театра. По итогам опроса видно, что она принимает следующие значения: 0, 1, 2, 3, 4, 5, 8 — и каждое из значений описывается своей частотой:

X, число посещений	0	1	2	**3**	4	5	8
Частота (количество ответивших супружеских пар)	4	2	3	7	1	2	1

Значит, для данного вариационного ряда мода равна 3, т. е. ***большинство*** супружеских пар посещают театр три раза в год. Это значение и можно рассматривать в качестве одной из характеристик центра исследуемой совокупности.

Ещё на одном примере покажем, что мода — это не всегда удачная характеристика центра исследуемой совокупности, так как у многих совокупностей может просто не быть моды.

Пример 2.14. «Театр-2». Супружеским парам задавали вопрос: «Сколько примерно раз в год Вы ходите в театр?». В опросе участвовало 20 супружеских пар. Выборка представлена следующими ответами: 1, 3, 2, 2, 0, 5, 4, 8, 3, 0, 1, 2, 0, 4, 4, 3, 0, 2, 4, 3. Чему равна мода в этой выборке ответов?

Решение. По итогам опроса видно, что случайная величина (число посещений театра) принимает следующие значения: 0, 1, 2, 3, 4, 5, 8 — и каждое из значений описывается своей частотой:

X, число посещений	0	1	2	3	4	5	8
Частота (количество ответивших супружеских пар)	4	2	4	4	4	1	1

Видим, что в данном распределении нет одной моды — ни одно из значений рассматриваемой величины не может быть названо модой, ни одно не встречается чаще остальных. Модой не может быть названо ни значение 0, ни значение 2, ни значение 3, ни значение 4. Моды в этом примере нет. А, значит, и центр совокупности не может быть охарактеризован с помощью моды.

Пример 2.14 обращает наше внимание на особенности (и в ряде случаев на ограничения) моды. *Во-первых*, бывают ситуации, когда в вариационном ряду присутствуют две и более моды. Такие распределения принято называть *бимодальными* (т. е. содержащими *две* моды) и *полимодальными* (т. е. содержащими *три* моды *и более*).

Пример 2.15. «Автомобили». Пусть в группе, состоящей из 10 человек, 4 не имеют личного автомобиля, 4 имеют один автомобиль, 1 человек имеет 2 автомобиля и ещё 1 человек имеет 3 автомобиля. В этих данных содержится два значения, которые встречаются чаще остальных — 0 (нет автомобиля) и 1 (есть один автомобиль). Это пример бимодального распределения.

Во-вторых, мода не может быть использована как адекватная мера центральной тенденции в случае, если частоты для всех наблюдаемых значений переменной *X* почти равны. Если, например, 49 % людей любят мороженое, а 51 % — нет, то использование моды здесь неуместно. Доли практически равны, и нельзя сделать однозначный вывод о том, кого же всё-таки больше — тех, кто любят мороженое, или же тех, кому оно не нравится.

§ 2.3. Выборочные квартúли и децúли

При изучении закономерностей, анализируя имеющиеся распределения, исследователи прибегают и к другим важным характеристикам вариационных рядов, например квартúлям (Q) и децúлям (D).

Выборочные квартили (Q, quartile) — это такие числа, которые разбивают (делят) упорядоченное от минимума к максимуму множество данных на четыре равные по численности части, т. е. *равновеликие между собой по объёму наблюдений*.

Всё упорядоченное множество объектов выборки можно разбить на четыре равновеликие части, прибегая к использованию ***трёх выборочных квартилей***. Поэтому выделяют нижнюю (первую), среднюю (вторую) и верхнюю (третью) квартили (рис. 2.5).

Интерпретировать значения квартилей следует аналогично тому, как мы интерпретировали медиану. Применительно к опросам общественного мнения, нижней квартилью будет считаться такая точка на шкале исследуемого признака, значения ***меньше (либо равные)*** которой отметили 25 % опрошенных. Тогда верхней квартилью будет называться точка на шкале, значения ***больше (либо равные)*** которой тоже отметили 25%.

Метод Дж. Тьюки. Джоном Тьюки (1915—2000), американским математиком, был предложен метод (или принцип), согласно которому при нахождении квартилей для упорядоченных рядов, содержащих чётное или нечётное число наблюдений, исследователям нужно придерживаться следующего алгоритма, состоящего из двух последовательных этапов (или шагов).

Если ряд чётный. В упорядоченном ряду значений признака определяется медиана. Это первый шаг. Напомним, что медиане чётного ряда будет соответствовать значение, равное полусумме соседних вариант (т. е. делящих между собой «срединное» место):

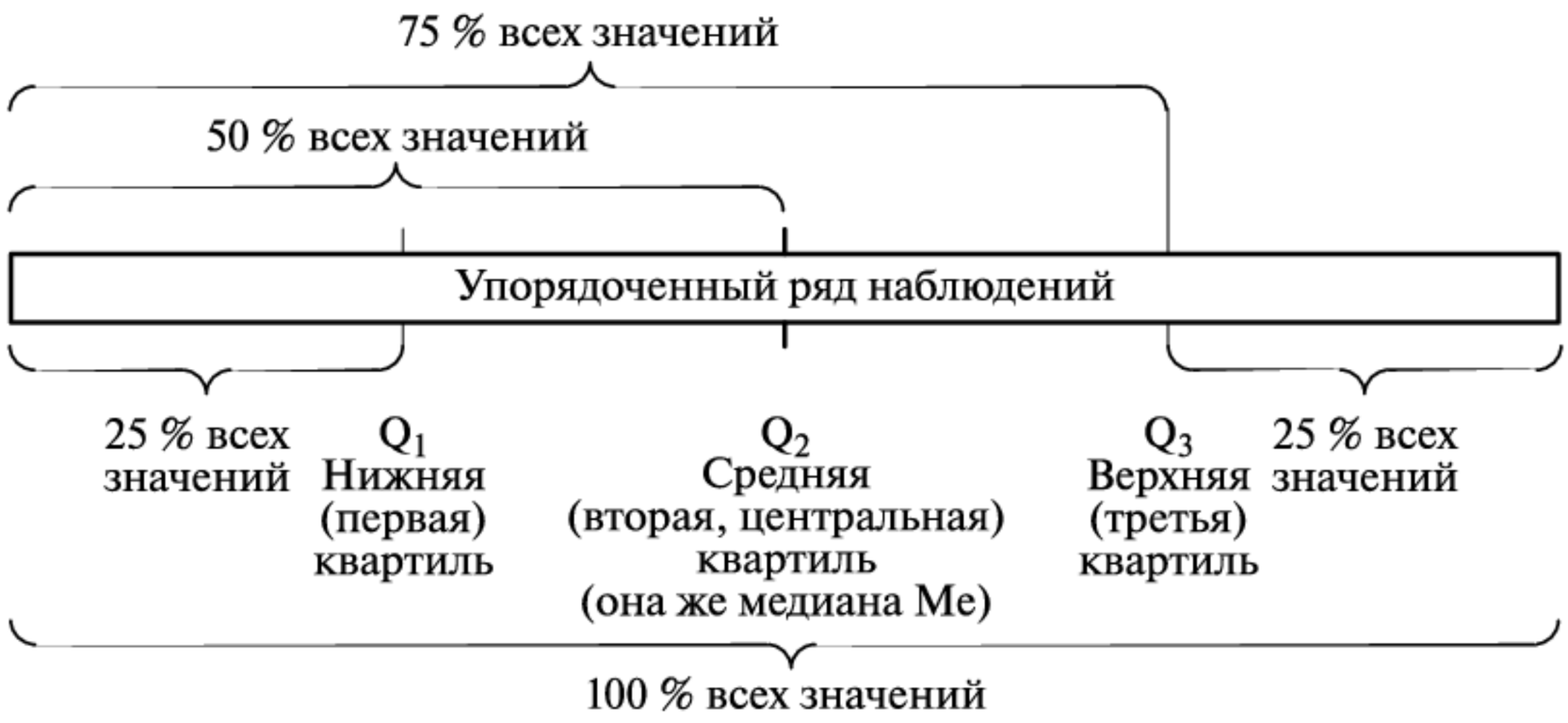


Рис. 2.5. Выборочные квартили

$\text{Me} = 0{,}5\left(X_{\frac{n}{2}} + X_{\frac{n}{2}+1} \right)$. Это расчётное значение *как бы условно* (искусственно) создаётся нами и помещается между соседними вариантами с порядковыми индексами $\frac{n}{2}$ и $\left(\frac{n}{2}+1\right)$. Тем самым выборка делится медианой на две равновеликие части — ***две подвыборки*** наблюдений. Для нахождения нижней и верхней квартилей потребуется *в отдельности* рассмотреть две эти *под*совокупности, внутри каждой из них найти свою медиану. *Медиана первой подвыборки* будет соответствовать значению *нижней выборочной квартили*, а *медиана второй подвыборки* — значению *верхней выборочной квартили*. Это второй шаг. Если говорить образно, прибегая к использованию метафоры, то предложенный Дж.Тьюки метод «работает» по принципу *матрёшка-в-матрёшке*. Обратите внимание на важное правило: если исходный ряд чётный, то для дальнейшего вычисления квартилей медиана общей выборки ***не включается ни в первую, ни во вторую подвыборки***, *т.е. не принадлежит ни одной* из двух выделенных совокупностей.

Поясним реализацию метода на конкретном примере.

Пример 2.16. «Товары повседневного спроса (продолжение)». Вернёмся к задаче про расходы населения на товары повседневного спроса, рассмотренного в теме «Медиана» (см. пример 2.5). Напомним, что выборка состояла из 12 наблюдений.

X (%): **40, 60, 25, 15, 10, 10, 20, 40, 30, 35, 40, 85.**

Найдём нижнюю и верхнюю квартили выборки.

Решение. Итак, выборка содержит 12 наблюдений (ряд чётный). Имеем упорядоченный ряд: **10, 10, 15, 20, 25, 30, 35, 40, 40, 40, 60, 85**. Медиана равна значению 32,5%, которое делит упорядоченное множество ответов респондентов пополам:

10, 10, 15, 20, 25, 30 | **Me = 32,5** | **35, 40, 40, 40, 60, 85**

Далее нам следует рассмотреть две *под*выборки наблюдений: $(10,10,15,20,25,30)$ и $(35,40,40,40,60,85)$. Для каждой из них найдём свою медиану. Каждая из подвыборок состоит из 6 наблюдений. Следовательно, полученные новые упорядоченные ряды также *чётные*. Значит, медианой в первой группе будет значение $17{,}5 = \frac{15+20}{2}$; а медианой во второй совокупности станет значение $40 = \frac{40+40}{2}$. Итак, $Q_1 = 17{,}5\,\%$, а $Q_3 = 40\,\%$, в то время как Q_2 (медиана всей совокупности) = 32,5 %. Итак, получаем, что ***четверть*** жителей (членов домашних хозяйств) тратит на покупку товаров повседневного спроса ***не более*** 17,5 % от общего семейного бюджета, ***ещё одна***

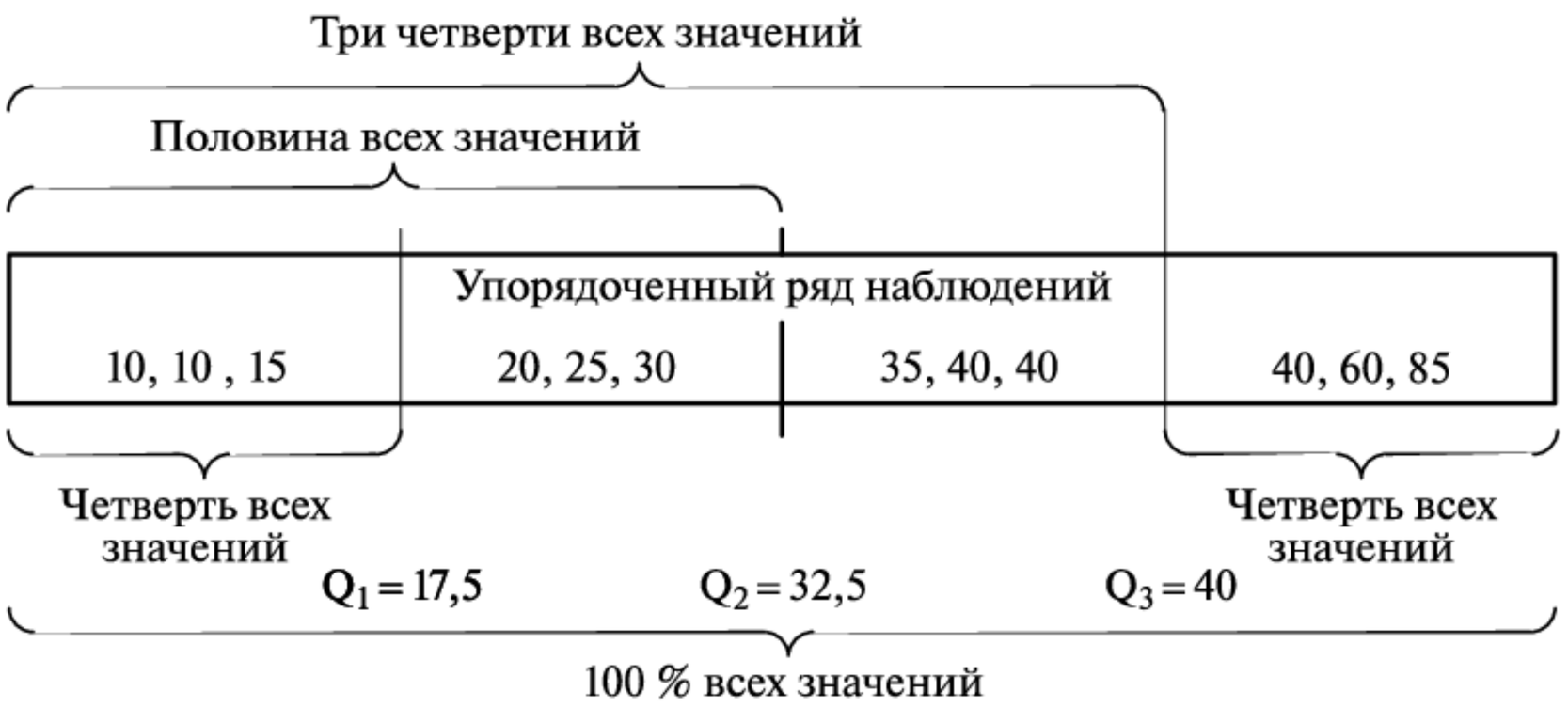


Рис. 2.6. Выборочные квартили (см. пример 2.16 «Товары повседневного спроса»)

четверть — не менее* 40 %**. Либо можно сказать так: ***75 % (т. е. три четверти) жителей (членов домашних хозяйств) тратят на покупку товаров первой необходимости ***не более 40 %*** бюджета или расходы на покупку товаров повседневного спроса у ***75 %*** граждан ***не ниже 17,5 %***. Все эти интерпретации взаимозаменяемы и корректны (рис. 2.6).

Если ряд нечётный. Медиане нечётного ряда будет соответствовать наблюдение выборки под порядковым номером $\left(\frac{n}{2}+1\right)$. Это расчётное значение *является одним из фактических наблюдений выборки и принадлежит ей*. Обратите внимание на важное правило: если исходный вариационный ряд *не*чётный, то для дальнейшего вычисления квартилей медиана общей выборки ***включается в обе подвыборки*** *— как в первую, так и во вторую, т. е. принадлежит обеим* из двух выделенных совокупностей.

Пример 2.17. «Покупка воды (продолжение)». Вернёмся к примеру 2.4. Компания оценивает, каковы еженедельные расходы её клиентов на покупку бутилированной воды. Выборка, состоящая из 11 элементов, представлена следующими значениями:

X (руб.): **500, 700, 700, 500, 600, 4 000, 3 000, 1 000, 800, 700, 600**.

Найдём нижнюю и верхнюю квартили выборки.

Решение. Имеем упорядоченный ряд: **500, 500, 600, 600, 700, 700, 700, 800, 1 000, 3 000, 4 000.** Он состоит из 11 наблюдений, т. е. является нечётным. Медиане соответствует наблюдение с порядковым номером 6 — это значение 700 руб. Рассматриваются две совокупности (500, 500, 600, 600, 700, 700) и (700, 700, 800, 1 000, 3 000, 4 000), в обоих случаях *захватывая* медиану выборки (700). Обе подвыборки содержат по 6 наблюдений. Следовательно, полученные упорядоченные ряды *чётные*:

500, 500, 600, 600, 700, 700, 700, 800, 1000, 3000, 4000

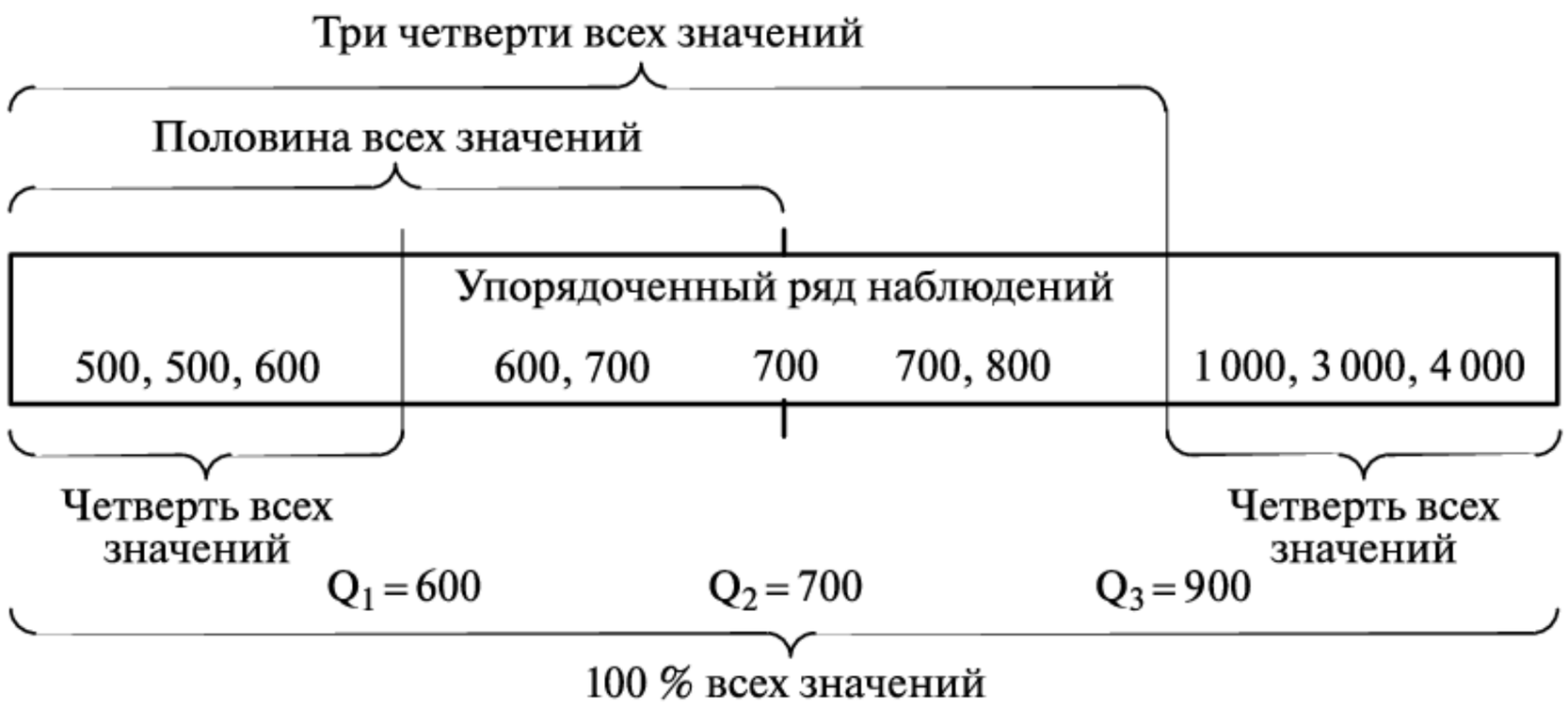


Рис. 2.7. Выборочные квартили (см. пример 2.17)

Тогда медианой в первой группе наблюдений будет значение $600 = \frac{600+600}{2}$, а медианой во второй совокупности станет значение $900 = \frac{800+1\,000}{2}$. Итак, $Q_1 = 600$ руб., а $Q_3 = 900$ руб., в то время как Q_2 (медиана всей совокупности) = 700 руб. (рис. 2.7).

Итак, получаем, что ***четверть*** клиентов в выборке тратит на покупку воды ***не более 600*** руб. в неделю. ***Ещё одна четверть*** клиентов расходует ***не менее 900*** руб. Либо можно сказать так: ***75 %*** (т. е. три четверти) клиентов компании еженедельно расходуют на покупку воды ***не более 900*** руб., или: еженедельные расходы на покупку воды у ***75 %*** клиентов компании ***не ниже 600*** руб.

Подчеркнем, что все эти интерпретации взаимозаменяемы и корректны.

Децили (D, decile), ***десятые доли*** — это такие числа, которые разбивают (делят) упорядоченное от минимума к максимуму множество данных на десять равных по численности частей, т. е. *равновеликих между собой по объёму наблюдений*.

Всё упорядоченное множество значений выборки можно разбить на десять равновеликих частей, прибегая к использованию ***девяти выборочных децилей***. Каждая из 10 частей составляет $\frac{1}{10}$ всей совокупности значений. Расчёт децилей основан на тех же идеях, что и нахождение медианы, квартилей. Только теперь выделяются не две, не четыре, а десять частей (десять *под*совокупностей). *Расчёт децилей опирается на кумулятивные частоты, т. е. на частоты, рассчитанные накопительным итогом.*

Первый дециль (D_1) отсекает 10 % объектов выборки с наименьшими значениями переменной, а девятый дециль (D_9) выделяет

10 % объектов выборки с наибольшими значениями переменной. Применительно к опросам общественного мнения, первым децилем будет считаться такая точка на шкале исследуемого признака, значения меньше (либо равные) которой отметили 10 % опрошенных. Тогда верхним децилем будет называться точка на шкале, значения больше (либо равные) которой тоже отметили 10 %. Пятым децилем (D_5) будет являться медиана ряда.

Пример 2.18. «Абитуриенты». Пусть распределение оценок абитуриентов (по 10-балльной шкале), полученных ими на вступительных экзаменах по профильному предмету при поступлении в бизнес-школу, выглядит следующим образом:

Оценка, балл	0	1	2	3	4	5	6	7	8	9	10	Всего
Число абитуриентов	80	140	160	180	180	200	540	300	140	60	20	2 000

Найдите первый и девятый децили выборки.

Решение. Задача решается крайне просто путём вычисления частот накопительным итогом. Расширим исходную таблицу, добавив в неё две дополнительные строки:

Оценка, балл	0	1	2	3	4	5	6	7	8	9	10	Всего
Число абитуриентов	80	140	160	180	180	200	540	300	140	60	20	2 000
Число абитуриентов, % к итогу	4	7	8	9	9	10	27	15	7	3	1	100
Число абитуриентов, кумулятивный % к итогу	4	**11**	19	28	37	47	74	89	**96**	99	100	—

Теперь следует обратить внимание на накопленные относительные частоты. Нужно увидеть, в каком именно значении вариационного ряда содержится (накопительным итогом) ***10 % носителей наименьших значений*** переменной. Очевидно, что этим значением является оценка 1 балл: 10 % всех абитуриентов сдали экзамен по профильному предмету на оценку, ***не превышающую*** 1 балл (т. е. получили оценки 0 или 1). Далее следует определить, какое из значений вариационного ряда охватывает («вбирает») ***10 % носителей наибольших значений*** переменной. Из таблицы видно, что этим значением выступает оценка 8 баллов. Она и будет тем рубежным значением, через которое «пройдут» накопленные относительные частоты 90 % и оставшиеся 10 %. Обратите внимание, что девятый дециль — это не оценка 7. Если мы посмотрим на накопленную относительную частоту, то она равна 89 % (и мы всё еще «не дотягиваем» до 90 %, которые отсека-

ли бы 10 % значений справа). Итак, $D_1 = 1$ (балл), $D_9 = 8$ (баллов). В выборке о результатах вступительных испытаний значения децилей указывают на то, что ***среди 10 % абитуриентов***, *хуже всех* сдавших экзамен, ***максимальный балл составляет (т. е. не превысил) 1***, а ***среди 10 % абитуриентов с наилучшим результатом минимальный балл составил 8*** (*8 и выше*), т. е. *в 8 раз больше*.

Заметим, что медиана в данном примере будет равна 6 баллам (что соответствует накопленной относительной частоте 74 %); $Me = Q_2$ (и это не 5 баллов, так как кумулятивный процент 47 % < 50 %; а это значит, что «срединное» значение лежит в следующей ячейке и равно 6). Аналогично определяются квартили: нижняя квартиль $Q_1 = 3$ балла (охватывает первые 25 % наблюдений), а верхняя квартиль $Q_3 = 7$ баллов (охватывает 75 % наблюдений слева и 25 % наблюдений справа).

В заключение важно подчеркнуть, что все рассмотренные характеристики выборки (медиана, среднее, квартили или децили) будут в дальнейшем активно применяться для определения степени и характера вариации данных.

§ 2.4. Выборочные меры разброса (вариации)

Рассмотрим основные показатели, характеризующие вариабельность (т. е. рассеивание, колеблемость) данных в выборке. Напомним, что вариацией называется ***изменчивость значений*** переменной (признака). Вариация напрямую связана со способностью среднего арифметического характеризовать центр совокупности. Чем выше изменчивость значений в выборке, тем более *непохожими между собой* являются значения (они *широко разбросаны вокруг среднего*). И наоборот, чем меньше вариация признака, тем компактнее и более «скученно» лежат друг к другу значения в выборке (они *слабо разбросаны вокруг среднего*). Говорят, что один ряд вариабельнее другого, если вариация признака в первом ряду значений значительно более ощутимая, чем во втором.

Изменчивость является ***следствием комплекса причин***, действующих на изучаемую совокупность и её единицы. Вариация обусловлена различными факторами, требующими от аналитиков ***объяснения***, детального изучения, рассмотрения. Для социальных наук понятие вариации имеет одно из приоритетных значений и представляет бóльший интерес, чем меры центральной тенденции. Это объясняется тем, что большая вариация — прямое указание на отсутствие сходств и присутствие социальных различий между людьми, наличие «разрывов». А это означает, что важно понять, чем (какими причинами) обусловлены такие дифференциация, расслоение и непохожесть. В поиске факторов (предпосылок), объясняющих наблюдаемую дис-

персию (т. е. социальную дифференциацию: почему она имеет место?) и лежит ключевая задача социальных наук.

В математике принято выделять две группы показателей вариации: ***абсолютные меры вариации*** и ***относительные меры вариации***. В качестве абсолютных мер вариации выборки рассмотрим *размах вариации*, *дисперсию*, *стандартное отклонение*, *квартильный размах*. В качестве относительных мер вариации выборки изучим *коэффициент вариации*, *коэффициент дециль́ной вариации*.

***Выборочный размах вариации* (*R*, *диапазон вариации*)** — наиболее простая (и в некотором смысле даже грубая) мера изменчивости.

Он равен разности между наибольшим и наименьшим значениями в выборке наблюдений: $R = X_{max} - X_{min}$.

Размах вариации — весьма неустойчивая величина, так как обусловлена всего двумя значениями выборки — минимумом и максимумом. Вместе с тем размах даёт информативное представление об изучаемом признаке, поскольку указывает на ***широту разнообразия*** значений признака, диапазон его возможных значений. Перед тем как вычислить диапазон вариации, нам потребуется исключить из выборочной совокупности нехарактерные (нетипичные, странные, «аномальные») наблюдения (подробнее о методе их выявления будет рассказано в § 2.5).

Пример 2.19. «Карманные деньги». Допустим, в ходе исследования у респондентов-школьников были зафиксированы данные о том, какую примерно сумму денежных средств им еженедельно дают родители на карманные расходы. В выборке минимальным значением было 1 000 руб., максимальным — 6 500 руб. Чему равен размах вариации?

Решение. $R = X_{max} - X_{min} = 6\,500 - 1\,000 = 5\,500$ руб. Это и есть диапазон разнообразия значений исследуемой переменной.

Выборочная дисперсия* (s^2) —** основной показатель вариации, используемый в статистике. Дисперсия отражает меру разброса значений выборки вокруг среднего арифметического. Дисперсия — очень информативная величина и несёт в себе колоссальный содержательный смысл. *Чем она меньше, тем о большей однородности данных она свидетельствует. И, напротив, чем она больше, тем менее однородна совокупность имеющихся данных.* В основе её расчета лежит оценка ***суммы квадратов отклонений каждого из значений выборки от общего среднего, а именно:

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

где i — номер наблюдения в выборке; n — размер выборки; x_i — определенное значение переменной (признака); $\bar{x}$ — среднее значение признака в выборке данных.

Разность между отдельным значением x_i и средним $\bar{x}$ отражает абсолютное отклонение (расстояние, дистанцию) между конкретным наблюдением и «статистическим центром» выборки. Очевидно, что складывая все эти *частные* квадраты разности, мы получим *суммарную* оценку отклонения значений выборки от среднего арифметического. А чтобы определить, чему тогда *в среднем равен квадрат отклонения* между произвольно выбранным значением и средним арифметическим, достаточно *разделить сумму квадратов отклонений на объём выборки*, т. е. нормировать сумму на число наблюдений. Это к тому же позволит сравнивать между собой вариацию выборок разного объёма. Тогда формула дисперсии примет вид:

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n} \text{ или } s^2 = \frac{\sum_{i=1}^{k}(x_i - \bar{x})^2 f_i}{n},$$

где f_i — частота встречаемости значения x_i в выборке наблюдений (в случае если данные представлены в виде вариационного ряда, в котором каждому значению признака соответствует своя частота).

Следует обратить внимание на то обстоятельство, что каждое отклонение возводится в квадрат для того, чтобы все частные отклонения стали положительными числами. Это в свою очередь позволит не допустить «взаимоуничтожения» положительных и отрицательных отклонений при их суммировании — ведь какие-то значения в выборке больше среднего (лежат правее на числовой оси), а какие-то меньше среднего (лежат левее на числовой оси). Здесь можно вспомнить основное свойство среднего арифметического: сумма всех разностей между $\bar{x}$ и x_1, x_2, x_3, x_4, ..., x_n равна нулю. Именно по этой причине, чтобы исключить «обнуление» числителя, прибегают к использованию ***квадрата разности***[1]. В результате деления суммы квадратов отклонений на объём выборки (число наблюдений в совокупности) получают ***средний квадрат отклонений***. Единицей измерения дисперсии служит квадрат единицы измерения исходного признака (например, квадратные рубли, квадратные баллы, квадратные штуки, квадратные проценты, квадратные килограммы и пр.).

Дисперсия является промежуточным показателем для вычисления более интерпретируемой меры вариации — стандартного отклонения.

***Выборочное стандартное отклонение* (*s*).** Понятия дисперсии и стандартного отклонения неразрывно связаны друг с другом.

[1] Некоторые из вас, наверное, могут задаться вопросом, а почему квадрат, а не модуль? Модуль как функция в математике не очень предпочтительна, поскольку не имеет производной в точке ноль. А поскольку дисперсия участвует во многих видах статистического анализа, модуль доставлял бы много неудобств практического характера. Поэтому конвенциональным решением в среде математиков стал *квадрат* отклонений. Так исторически сложилось.

Стандартное отклонение равно ***корню квадратному из дисперсии***. Это позволяет перейти *от квадратов единицы измерения* к тем *первоначальным единицам*, в которых и выражался исследуемый признак (т. е. рубли, баллы, штуки, проценты, килограммы и пр.). Чтобы избежать взаимоуничтожения слагаемых числителя, мы прибегаем к использованию квадрата разности в вычислении дисперсии. А теперь, чтобы вернуть возможность интерпретации, мы словно «возвращаемся обратно» — вычисляем корень квадратный из дисперсии:

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}}, \text{ или } s = \sqrt{\frac{\sum_{i=1}^{k}(x_i - \bar{x})^2 f_i}{n}},$$

Стандартное отклонение на практике легко интерпретируется. Оно обладает ***способностью оценивать «типичность» («характе́рность») среднего значения***, позволяет зафиксировать, насколько плотно значения признака находятся относительно центра совокупности — среднего арифметического (на рис. 2.8 представлены выборочные данные об уровне безработицы в федеральных округах (ФО) РФ за 2010 г.[1]). И наоборот. Стандартное отклонение — глубоко содержательный показатель положения каждого конкретного наблюдения относительно среднего.

Математиками хорошо изучены свойства этого показателя. Самым известным свойством стандартного отклонения является так называемое ***правило «68-95-99»*** (или говорят: ***«правило трёх сигм»***). В теории вероятностей это правило справедливо для нормального (гауссова) распределения. Согласно этому правилу, вероятность значений случайной величины концентрируется в пределах интервала $\mu \pm \sigma$ (где μ — математическое ожидание случайной величины; σ — её стандартное отклонение (корень квадратный из дисперсии)), и на этот интервал приходится вероятность, равная примерно 0,68. На интервал значений случайной величины в пределах $\mu \pm 2\sigma$ приходится вероятность, равная примерно 0,95, а на интервал $\mu \pm 3\sigma$ приходится вероятность, равная приблизительно 0,99.

Так, согласно этому правилу, если распределение случайной величины описывается нормальным законом («колокольчиком Гаусса»), то можно сказать, что из 100 значений этой величины 99 будут принадлежать отрезку ($\mu \pm 3\sigma$), т. е. 99 % значений случайной величины будут находиться *не далее*, *чем на три сигмы* по обе стороны от математического ожидания (рис. 2.9).

---

[1] Уровень безработицы зафиксирован по данным выборочных обследований населения по проблемам занятости. 2010 г. (См. мат-лы Федеральной службы государственной статистики [http://www.gks.ru/bgd/regl/b11_14p/IssWWW.exe/Stg/d01/04-15.htm] (проверено 26.07.2012)).

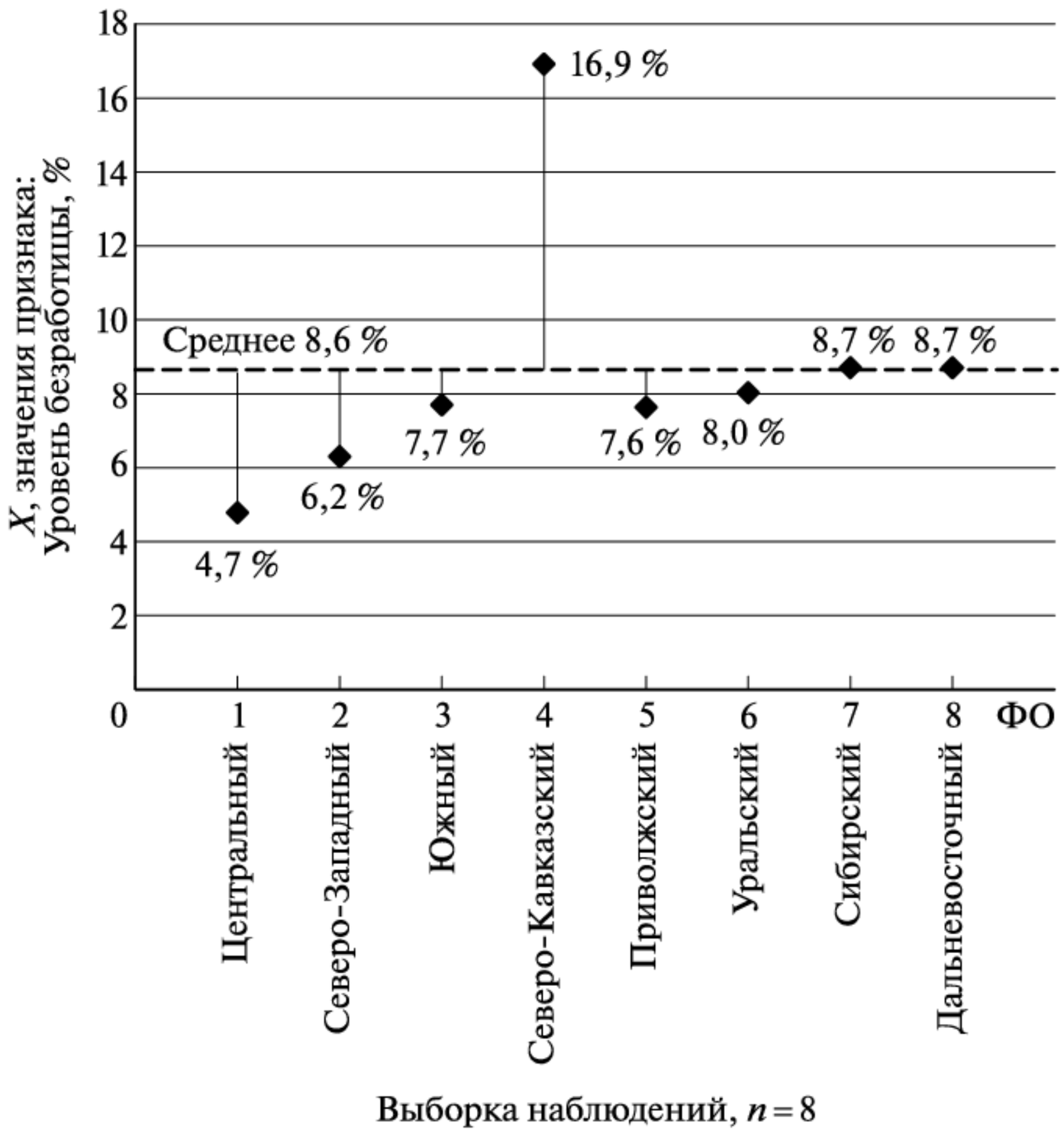


Рис. 2.8. Данные об уровне безработицы в РФ за 2010 г.

Рис. 2.9. Нормальное распределение и правило «68-95-99» («трёх сигм»)

Стандартное отклонение участвует во многих статистических расчётах. С его помощью мы будем определять степень точности различных оценок, полученных на выборке, и осуществлять прогнозы.

Пример 2.20. «Преступность среди несовершеннолетних». Динамика числа преступлений, совершённых в Москве несовершеннолетними и при их участии, за последнее десятилетие выглядит так[1]:

Годы	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010
Число преступлений	3 181	3 359	2 480	2 905	3 193	3 398	2 637	2 148	2 047	1 488	1 145

Вычислите выборочную дисперсию и выборочное стандартное отклонение.

Решение. Прежде чем вычислить выборочную дисперсию, найдём среднее арифметическое:

$$\bar{X} = (3\,181 + 3\,359 + 2\,480 + 2\,905 + 3\,193 + 3\,398 + 2\,637 + \\ + 2\,148 + 2\,047 + 1\,488 + 1\,145) / 11 \approx 2\,543{,}7,$$

т. е. в среднем каждый год подростками и при их участии в Москве совершается около 2 544 преступлений.

Вычисление дисперсии осуществляется так:

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n} = \\ = [(3\,181 - 2\,543{,}7)^2 + (3\,359 - 2\,543{,}7)^2 + (2\,480 - 2\,543{,}7)^2 + \ldots + \\ + (1\,488 - 2\,543{,}7)^2 + (1\,145 - 2\,543{,}7)^2] / 11 = \frac{5\,839\,738{,}19}{11} = \\ = 530\,885{,}29 \approx 530\,885.$$

Выборочная дисперсия равна 530 885 (преступлений)^2.
Тогда стандартное отклонение составит:

$$s = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}} = \sqrt{530885{,}29} = 728{,}6 \approx 729 \text{ (преступлений)}.$$

Следовательно, в течение каждого года в Москве регистрируется $2\,544 \pm 729 \cdot 3$ преступлений, совершённых детьми и подростками (в том числе и при их участии).

---

[1] Согласно данным Федеральной службы государственной статистики: см. Регионы России. Социально-экономические показатели-2011. (http://www.gks.ru/bgd/regl/b11_14p/Main.htm) (проверено 26.07.2012).

Пример 2.21. «Африканская саванна». Предположим, что зоологами в ходе исследования фиксировалась численность (размер) прайдов в африканской саванне. Прайдом в зоологии называется семейная стая львов, состоящая из нескольких особей (численность прайда может достигать трёх-четырёх десятков животных). Зоологами указывалось количество животных в каждом прайде. Пусть нас интересует среднее число особей в прайде и стандартное отклонение этого числа. Выборка состоит из 27 наблюдений (прайдов, охваченных исследованием):

Численность прайда (число особей в группе)	1	3	5	6	7	8	9	11	13	20	Всего
Число прайдов с такой численностью особей	2	1	3	2	4	4	5	3	2	1	$n = 27$

Решение. Переменной (признаком) X в задаче является численность прайда (количество особей в стае). Обратите внимание, что вторая строка таблицы указывает на абсолютные частоты значений X. Найдём среднее арифметическое выборки, взвешенное на наблюдаемые частоты:

$$\bar{X} = \frac{\sum_{i=1}^{k} x_i f_i}{n} = \frac{1\cdot 2 + 3\cdot 1 + 5\cdot 3 + 6\cdot 2 + \ldots + 13\cdot 2 + 20\cdot 1}{2 + 1 + 3 + 2 + \ldots + 2 + 1 = 27} = \frac{216}{27} = 8 \text{ (особей)}$$

— средний размер одного прайда.

Вычислим выборочную дисперсию и выборочное стандартное отклонение:

$$s^2 = \frac{\sum_{i=1}^{k} (x_i - \bar{x})^2 f_i}{n} =$$

$$= \frac{(1-8)^2\cdot 2 + (3-8)^2\cdot 1 + (5-8)^2\cdot 3 + (6-8)^2\cdot 2 + \ldots + (13-8)^2\cdot 2 + (20-8)^2\cdot 1}{27} =$$

$$= \frac{388}{27} \approx 14{,}4 \text{ (особей)}^2.$$

Тогда стандартное отклонение составит: $s = \sqrt{14{,}4} \approx 3{,}8$(особей). Округлять дробные значения стандартного отклонения вовсе не обязательно.

Значит, численность львов в одном прайде примерно колеблется в пределах $8 \pm 3{,}8\cdot 3$ особей.

Пример 2.22. «Кофейня». Компанию, владеющую сетью кофеен в городе, интересует средняя сумма чека посетителей и показатель

вариации — стандартное отклонение. Предположим, что аналитики компании изучают показатели за прошлый месяц. Имеются следующие данные:

Сумма чека, руб.	До 300	300 — 500	500 — 1 000	1 000 — 1 500	1 500 и выше	Итого
Середины интервалов	200	400	750	1250	1750	—
Доля посетителей с такой суммой чека, % к итогу	15	45	25	10	5	100

Средняя сумма чека в кофейне составит:

$$\bar{X} = \frac{200 \cdot 15\,\% + 400 \cdot 45\,\% + 750 \cdot 25\,\% + 1\,250 \cdot 10\,\% + 1\,750 \cdot 5\,\%}{100\,\%} = 610 \text{ руб.}$$

Запись можно оформить и таким образом:

$$\bar{X} = \frac{200 \cdot 0{,}15 + 400 \cdot 0{,}45 + 750 \cdot 0{,}25 + 1\,250 \cdot 0{,}1 + 1\,750 \cdot 0{,}05}{1} = 610 \text{ руб.}$$

Выборочная дисперсия будет равна:

$$s^2 = [(200 - 610)^2 \cdot 15\,\% + (400 - 610)^2 \cdot 45\,\% + (750 - 610)^2 \cdot 25\,\% + (1\,250 - 610)^2 \cdot 10\,\% + (1\,750 - 610)^2 \cdot 5\,\%] / 100\,\% = 155\,900.$$

Запись можно оформить и таким образом:

$$s^2 = [(200 - 610)^2 \cdot 0{,}15 + (400 - 610)^2 \cdot 0{,}45 + (750 - 610)^2 \cdot 0{,}25 + (1\,250 - 610)^2 \cdot 0{,}10 + (1\,750 - 610)^2 \cdot 0{,}05] / 1 = 155\,900.$$

Тогда выборочное стандартное отклонение составит:

$s = \sqrt{155\,900} \cong 395$ руб.

Значит, сумма одного чека большей части посетителей (но не всех!) колеблется в пределах 610 ± 395 руб.

Выборочный квартильный размах (RQ), точно так же, как и стандартное отклонение, служит мерой разброса (изменчивости) данных. Квартильный размах является характеристикой разброса *50 % центральных значений упорядоченного ряда* и вычисляется как ***разность между верхней (третьей) и нижней (первой) квартилями***: $RQ = Q_3 - Q_1$. И выполняет ту же функцию, что и остальные меры вариации — характеризует изменчивость (рассеивание) данных (рис. 2.10).

Рис. 2.10. Выборочный квартильный размах

Теперь перейдём к рассмотрению относительных мер вариации. Два коэффициента, которые мы изучим, свидетельствуют о том, ***в какой мере (степени)*** изучаемая совокупность является ***однородной***.

Выборочный коэффициент вариации **(V_X)** — главный показатель, характеризующий относительную меру рассеивания значений выборки. Он указывает на ***степень (интенсивность, выраженность)*** разброса значений переменной относительно среднего. Коэффициент вариации вычисляется как отношение выборочного стандартного отклонения к среднему значению, и показывает, *во сколько раз* стандартное отклонение и среднее арифметическое различаются между собой: $V_x = \frac{s}{\overline{x}} 100\,\%$.

Коэффициент вариации измеряется в долях или процентах и не зависит от исходных единиц измерения признаков. И, в результате, позволяет судить о том, в какой мере стандартное отклонение можно считать большим или малым, а исследуемую совокупность — ***разнородной*** или, напротив, ***однородной***. Благодаря этому обстоятельству мы обладаем возможностью сравнивать различные вариационные ряды между собой и определять, какой более вариабелен, а какой — менее.

Среди людей, занимающихся обработкой и анализом данных, имеются некоторые договорённости в интерпретации значений выборочного коэффициента вариации. Так, если коэффициент вариации не превышает ***30—35 %***, то говорят, что данные обладают ***высокой однородностью***; если находится в пределах ***35—65 % — умеренной однородностью***; если равен или превышает ***65—70 % — высокой разнородностью***. Заметим, что чем выше стандартное отклонение, тем сильнее выражена непохожесть значений выборки друг на друга, и тем большее значение принимает числитель дроби в формуле для вычисления коэффициента вариации. Но это вполне объяснимо. В таком случае исследуемая совокупность разнородна, а среднее не отражает *типичную* картину, так как значения имеют широкий разброс и весьма далеко отстоят от среднего. И наоборот. Чем меньше стандартное отклонение, тем у́же диапазон рассеивания, значения

находятся довольно «кучно» друг к другу. В этом случае числитель дроби стремится к нулю. Однородность выборки (*похожесть* значений между собой) высока.

Вернёмся к фрагменту данных одного из примеров (см. пример 2.20), рассчитаем и интерпретируем коэффициент вариации.

***Пример 2.23.* «*Преступность среди несовершеннолетних (продолжение)*».** Напомним, ранее было вычислено, что в среднем каждый год (рассматривалась выборка последних 11 лет) подростками (и при их участии) в Москве совершается в среднем около 2 544 преступлений. Стандартное отклонение составляет около 729 преступлений. Чему равен коэффициент вариации?

Решение. Коэффициент вариации равен: $V_x = \frac{729}{2\,544} 100\,\% \approx 28{,}6\,\%$.

Полученное значение (28,6 %) свидетельствует о том, что данные о подростковой преступности за последние 11 лет однородны, годовые показатели преступности весьма похожи между собой. Из года в год (с 2000 по 2010 г.) ситуация кардинально *не менялась*, *оставаясь весьма стабильной*. Ощутимого разброса значений не наблюдается.

Выборочный коэффициент дециль́ной вариации **(V_D)** также указывает на степень однородности совокупности по изучаемому признаку. Коэффициент децильной вариации равен отношению *девятого дециля к первому децилю*. Он показывает, в какой мере (т.е. во сколько раз) выражена дифференциация ***между крайними децилями***, и вычисляется по формуле

$$V_D = \frac{D_9}{D_1}.$$

Чаще всего коэффициент децильной вариации применяется при изучении социально-экономического неравенства (расслоения) между людьми (дифференциации их доходов) и рассчитывается как соотношение средних (или совокупных) доходов 10% граждан с наибольшими доходами и средних (или совокупных) доходов 10% наименее обеспеченных граждан. Иными словами, выборочный коэффициент децильной вариации показывает, ***во сколько раз минимальные доходы 10 % наиболее обеспеченных (самых богатых) граждан превышают максимальные доходы 10 % наименее обеспеченных (самых бедных)***. Чем выше этот коэффициент, тем выше степень экономического расслоения (расслоения по доходу). Чем он меньше, тем более сглаженным (одинаковым, похожим) является финансовое положение людей. Уровень социально-экономической стратификации — крайне важная характеристика социума, поскольку слишком ощутимое различие в доходах между социальными группами («богатыми» и «бедными») может приводить к росту протестных настроений и всевозможных акций недовольства в обществе.

Если говорить более развёрнуто, то можно отметить, что среди коэффициентов социально-экономической дифференциации (как правило, деци́льных) различают *два* типа: ***коэффициент уровней*** и ***коэффициент фондов***. ***Децильный коэффициент уровней*** — это как раз отношение девятого дециля к первому, т. е. уровня дохода, который отсекает 10 % самых богатых индивидов (или домохозяйств) к величине дохода, меньше которого имеют 10% самых бедных индивидов (или домохозяйств). А ***коэффициент фондов*** — это соотношение *средней* величины дохода у 10 % самых богатых и *среднего* дохода у 10 % самых бедных индивидов (домохозяйств). При строгом равенстве числа объектов в крайних деци́льных группах можно, например, оперировать не средним доходом, а суммой всех доходов, попавших в эти децили.

Пример 2.24. Если коэффициент деци́льной вариации (коэффициент уровней) по доходам равен 15, это означает, что ***минимальные доходы*** 10 % самых обеспеченных граждан выше ***максимальных доходов*** 10 % самых малоимущих граждан в 15 раз.

Показатели деци́льной вариации часто используются в сравнительных исследованиях городов, регионов, стран.

§ 2.5. Коробчатая диаграмма («Ящик с усами»)

В этом параграфе мы научимся визуализировать данные с помощью графика, изображающего одномерное распределение во многих его деталях. «Ящик с усами» (от англ. *box-and-whiskers plot* или просто *box plot*) — график, позволяющий быстро проводить визуальный анализ имеющихся данных. Это очень удобный способ сделать большие объёмы (массивы) данных наглядными и компактными и представить их на одном рисунке (рис. 2.11).

Построение коробчатой диаграммы («Ящика с усами») для одной выборки включает в себя реализацию следующих подготовительных этапов:

- вычисление ***медианы*** (Me) выборки;
- вычисление ***квартилей*** выборки — нижней и верхней (Q_1 и Q_3), ***квартильного размаха*** (RQ);
- определение так называемого ***диапазона характе́рных значений*** выборки. Его ещё называют ***областью характе́рных (типичных) значений*** выборки. Это такая область, которая охватывает значения признака, принадлежащие интервалу $\mathbf{[Q_1 - 1{,}5RQ;\ Q_3 + 1{,}5RQ]}$. У этого интервала есть две границы — нижняя и верхняя. Нижняя граница называется ***статистическим минимумом***: $X_{\text{стат.min}} = Q_1 - 1{,}5RQ$. Верхняя граница называется ***статистическим максимумом***: $X_{\text{стат.max}} = Q_3 + 1{,}5RQ$. Итак, видим, что для определения области характерных значений нужно из соответствующих квартилей

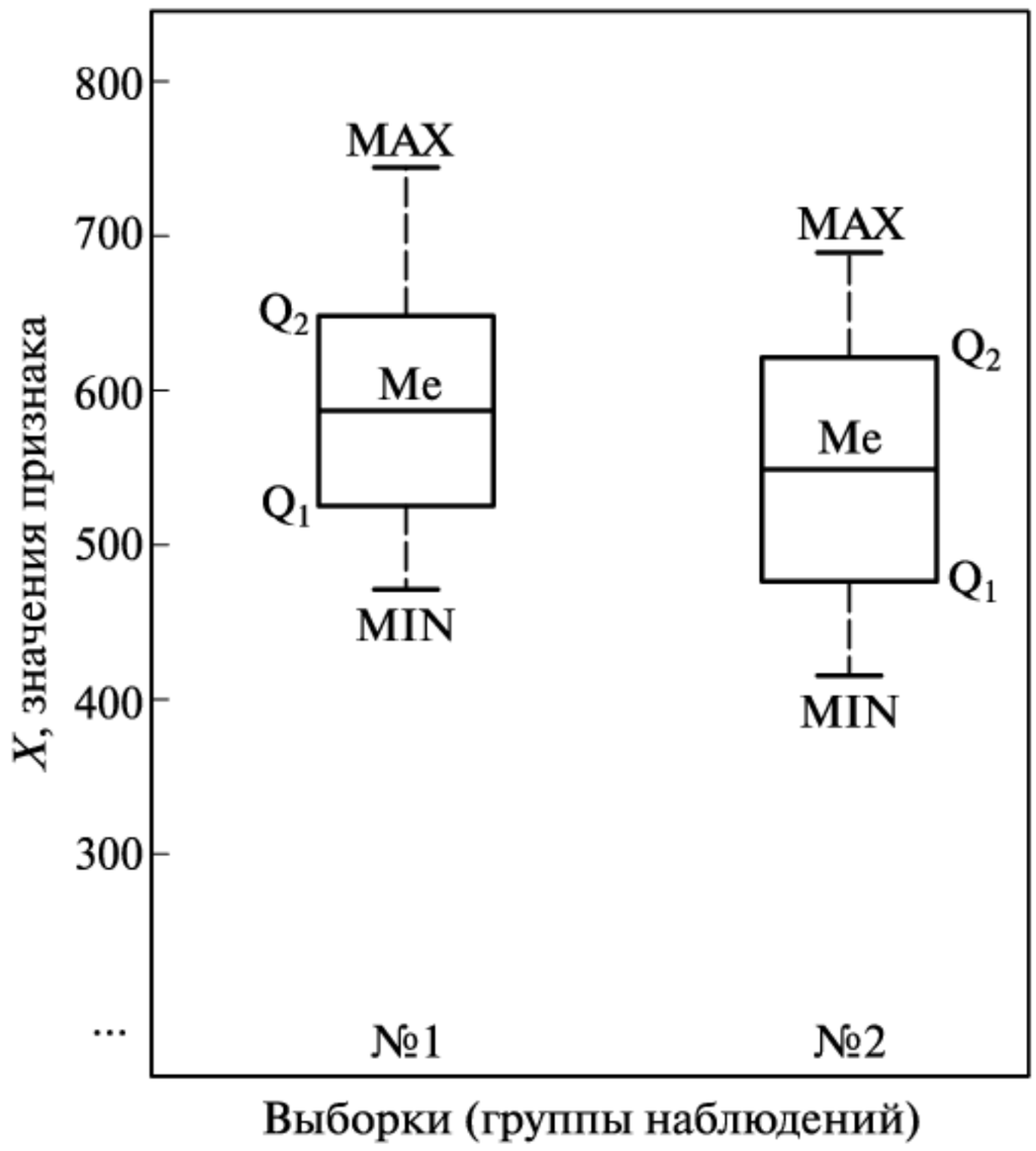


Рис. 2.11. Схематичный внешний вид диаграмм «Ящик с усами»

вычесть или прибавить 1,5 квартильных размаха. В результате формируется интервал $[Q_1 - 1{,}5RQ;\ Q_3 + 1{,}5RQ]$, в который попадает подавляющее большинство наблюдений в выборке. Этот интервал ***не*** охватывает ***крайне большие*** или ***крайне малые*** значения. Все те значения, которые не попадают в данный диапазон, называются ***нетипичными*** (*редкими*, *странными*, *крайне большими либо крайне малыми в сравнении с остальными значениями*);

• *выделение* из группы нетипичных значений «***выбросов***», т. е. ***крайне редких***, ***аномальных***, ***экстремально больших*** *либо* ***экстремально малых*** *в сравнении с остальными*. Нетипичные наблюдения в выборке классифицируются на две группы (две категории) — «***подозрительные***» (*suspected*) *и* «***выбросы***» (*outliers*). Чтобы дифференцировать их между собой, необходимо провести дополнительные расчёты, определив *ещё более широкий диапазон* как $[Q_1 - 3RQ;\ Q_3 + 3RQ]$. Нетипичные значения, которые «не укладываются» даже в столь широкий интервал, т. е. лежат за его пределами, классифицируются как *выбросы*, а значения, которые отстоят от квартилей (нижней или верхней) ***более чем на 1,5RQ***, ***но менее чем 3RQ***, считаются «*подозрительными*»;

• *после выполненных расчётов* можно приступать к графическому построению «Ящика с усами».

Коробчатая диаграмма строится в следующей последовательности операций (шагов):

1) рисуется прямоугольник (так называемый «ящик», коробка). Его нижней планкой будет первая квартиль выборки (Q_1), а верх-

ней — третья квартиль (Q_3). Высота прямоугольника соответствует квартильному размаху (RQ);

2) внутри прямоугольника в соответствующем месте проводится полоска (линия) — медиана (Me). Она может совпадать с центром «ящика», располагаться ближе к одной из квартилей или касаться одной из его границ;

3) рисуются «отростки» (так называемые усы, *whiskers*), направления которых зависят от того, есть ли в выборке нетипичные значения:

а*) *если в выборке нет нетипичных значений (т. е. все наблюдения в выборке охватываются, «накрываются» интервалом $[Q_1 - 1{,}5RQ; Q_3 + 1{,}5RQ]$), то «усы» отходят от Q_3 вверх и от Q_1 вниз до точек X_{min} (*фактический* минимум) и X_{max} (*фактический* максимум) соответственно;

б*) *если в выборке присутствуют нетипичные значения («подозрительные» и/или «выбросы»), то «усы» отходят от Q_3 вверх и от Q_1 вниз до точек X'_{max} (максимальное *из тех значений выборки, которые принадлежат диапазону* $[Q_3 + 1{,}5RQ]$) и X'_{min} (минимальное *из тех значений выборки, которые принадлежат диапазону* $[Q_1 - 1{,}5RQ]$);

4) на диаграмме особым образом помечаются нетипичные наблюдения выборки: точками (или кружочками) — подозрительные наблюдения, звёздочками — выбросы.

В результате получается коробчатая диаграмма, которая в компактном виде ***отражает структуру*** выборки и ***её основные ха-***

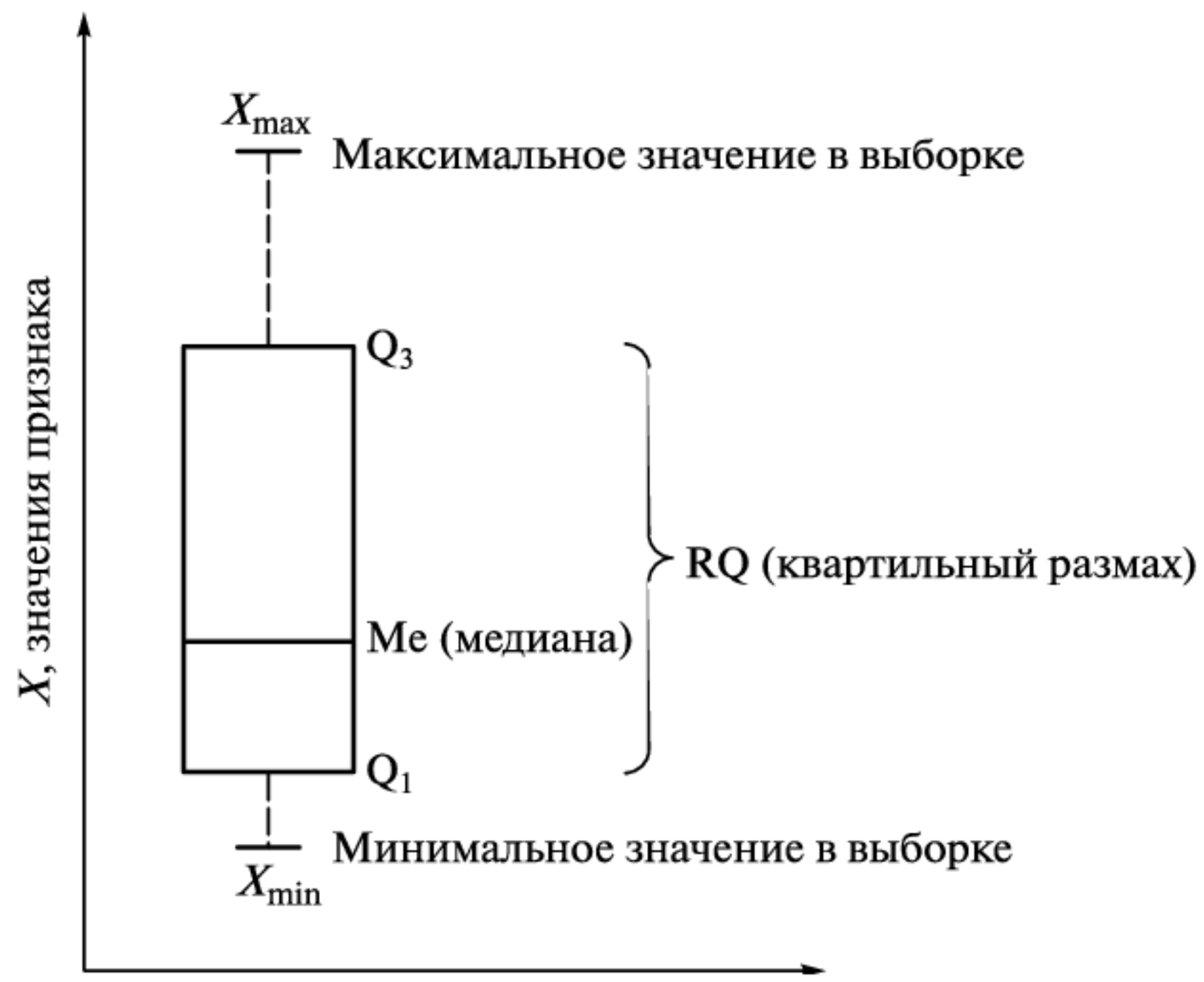


Рис. 2.12. Схематичный внешний вид диаграммы «Ящик с усами» в отсутствие нетипичных значений в выборке

рактеристики: медиану, квартильный размах, минимум, максимум, нетипичные значения. (Кстати, на занятиях студенты часто приводят много других образных аналогий и метафор, пытаясь подобрать аль-

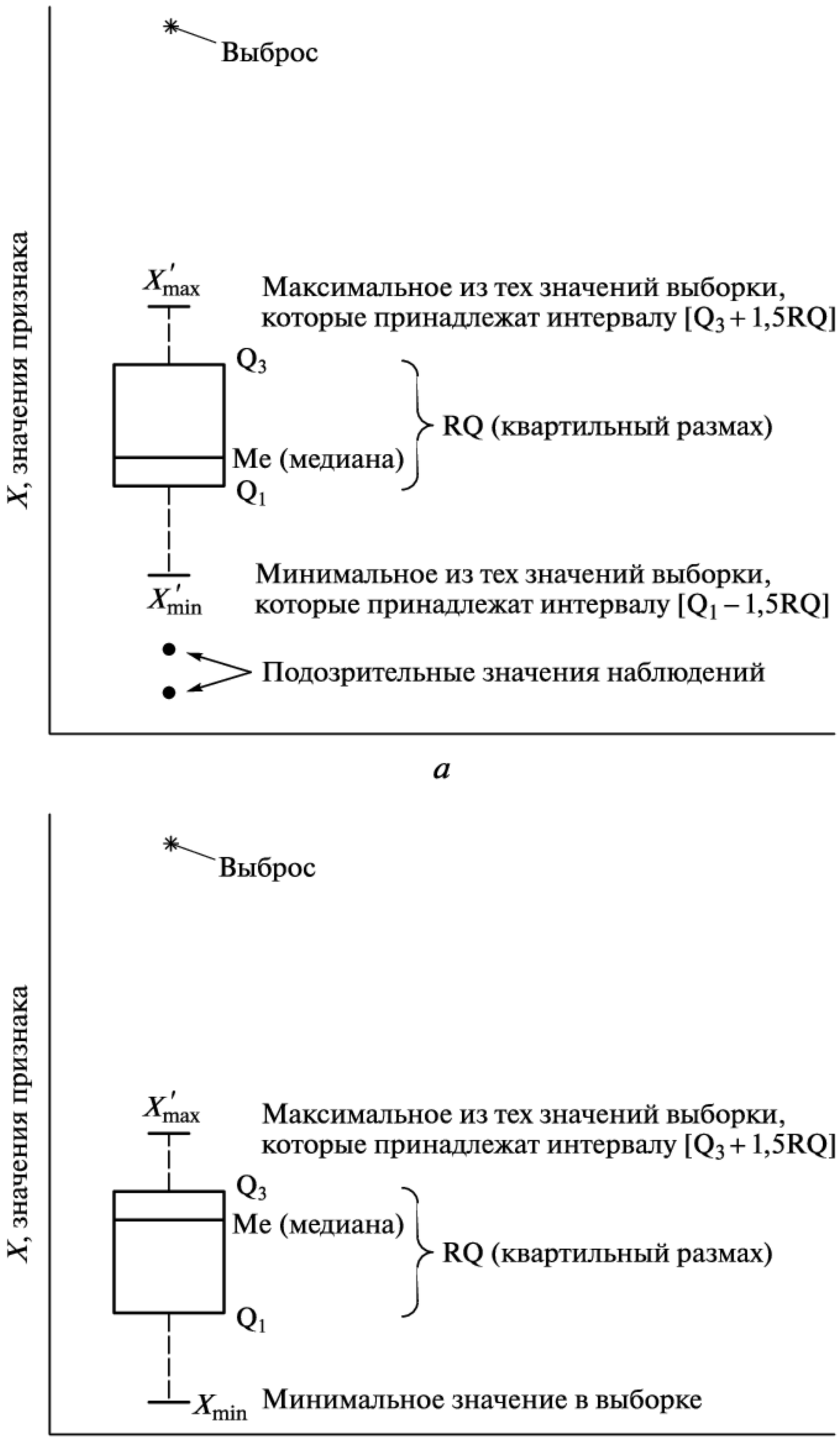


Рис. 2.13. Схематичный внешний вид диаграммы «Ящик с усами» при наличии нетипичных значений в выборке:

а — «подозрительных» и «выбросов»; *б* — «выбросов»

тернативное название «Ящику с усами»: кто-то сравнивает диаграмму с шашлыком, кто-то — с медицинским шприцем, а кто-то — с музыкальной гармонью). Схематично разные виды коробчатых диаграмм представлены на рис. 2.12 и 2.13.

Пример 2.25. «Рестораны». Рассмотрим финансовые результаты деятельности двух ресторанов быстрого питания — А и Б. Пусть нас интересует такая величина, как сумма чека в одном заказе посетителя. Имеются две выборки данных:

- ресторан А, X (руб.): 50, 600, 500, 400, 300, 400, 500, 450, 500 — суммы чеков;
- ресторан Б, X (руб.): 700, 200, 400, 500, 200, 500, 700, 500 — суммы чеков;

Проведём подготовительные расчёты, требующиеся для визуализации данных.

- Рассмотрим первую выборку, упорядочим её значения от минимума к максимуму:

$X_А$ (руб.): 50, 300, 400, 400, 450, 500, 500, 500, 600;

- Вычислим медиану, квартили и квартильный размах.

$X_А$ (руб.): 50, 300, 400, 400, 450, 500, 500, 500, 600;

$Me = 450$, $Q_1 = 400$, $Q_3 = 500$, $RQ = 500 - 400 = 100$.

- Найдём область характе́рных (типичных) для этой выборки значений: $X_{стат.min}$ (нижняя граница диапазона) $= 400 - 1{,}5 \cdot 100 = 400 - 150 = 250$; $X_{стат.max}$ (верхняя граница диапазона) $= 500 + 1{,}5 \cdot 100 = 500 + 150 = 650$. Получим интервал [250; 650].
- Теперь посмотрим, есть ли в выборке наблюдения, не вошедшие в этот интервал, т. е. являющиеся нетипичными. Да, значение 50 не принадлежит области характерных значений.
- Определим, к какой категории нетипичных значений — «подозрительным» или «выбросам» — относится значение «50 руб.». Для этого рассчитаем более широкий диапазон, сформированный точками, отстоящими от квартилей на 3RQ. Получим: $[400 - 3 \cdot 100; 500 + 3 \cdot 100] = [100; 800]$. Полученный интервал не захватывает значение 50. Следовательно, значение 50 — «выброс».

Теперь все проделанные операции повторим для второй выборки (данные по ресторану Б).

- Сформируем упорядоченный ряд:

$X_Б$ (руб.): 200, 200, 400, 500, 500, 500, 700, 700;

$Me = 500$ (полусумма соседних значений 500 и 500), $Q_1 = 300$, $Q_3 = 600$, $RQ = 600 - 300 = 300$.

- Область характерных значений: $X_{стат.min} = 300 - 1{,}5 \cdot 300 = 300 - 450 = -150$ (значение «–150» можно округлить до нуля, так как сумма заказа в чеке не может быть отрицательной) $X_{стат.max} = 600 + 1{,}5 \cdot 300 = 600 + 450 = 1\,050$. Получен интервал [0; 1 050], охватывающий значения от 0 до 1 050.
- В этой выборке отсутствуют нетипичные значения наблюдений.

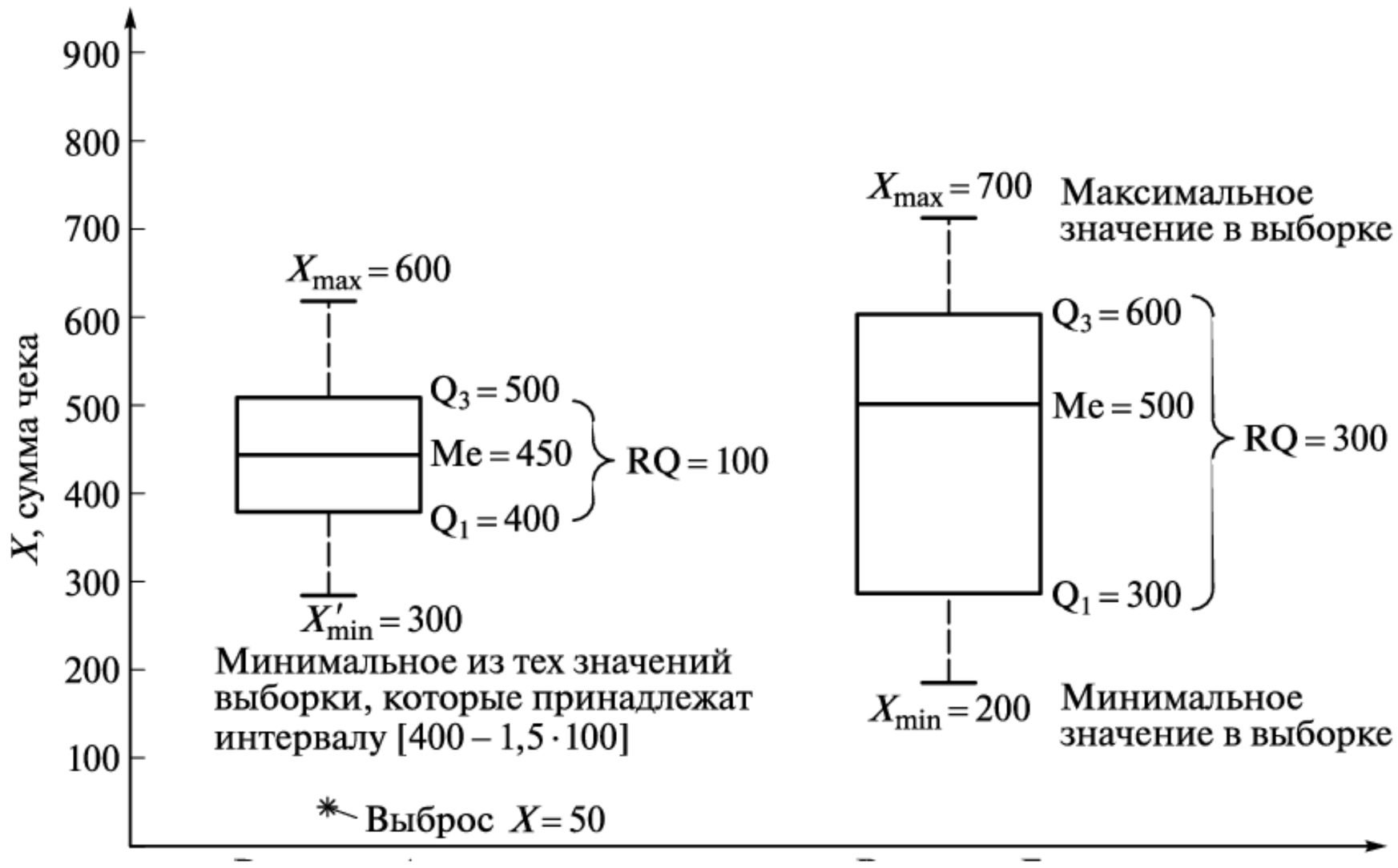


Рис. 2.14. Диаграмма «Ящик с усами» (см. пример 2.25 «Рестораны»)

Построим на одном рисунке (рис. 2.14) коробчатые диаграммы («Ящик с усами») для ресторанов А и Б. Эти изображения дают довольно исчерпывающую информацию о структуре данных в рассматриваемых выборках (ресторан А и ресторан Б). Мы нарисовали «Ящики с усами» бок о бок, что позволяет визуально сравнить финансовые результаты деятельности обоих ресторанов между собой и сделать *предварительные* предположения о том, в каком из них сумма чека является более вариабельной, в каком — менее. Данные первой выборки более «скучены» вокруг медианы, а сумма чека — менее «разбросана», так как 50% всех заказов лежат в интервале [400, 500] руб., в то время как в ресторане Б сумма чека в половине заказов, рассмотренных в выборке, находится в интервале [300, 600] руб., и этот интервал существенно шире.

Диаграмма «Ящик с усами» — наглядная форма презентации данных, с помощью которой можно на одном рисунке отобразить до нескольких десятков диаграмм и сопоставить разные выборки друг с другом. Это создаёт колоссальные удобства в визуализации, «прочтении» и сравнительном анализе информации.

ЗАДАЧИ ДЛЯ САМОСТОЯТЕЛЬНОГО РЕШЕНИЯ

Задача 1. «Мечты». Студентами факультета социологии на одном из семинаров по математической статистике был инициирован блиц-опрос на семейную тему. Один из вопросов, на которые требовалось дать ответ, звучал так: «О каком количестве детей в своей будущей семье Вы мечтаете? Укажите, пожалуйста, числом». Ответы молодёжи были сведены в таблицу:

X, число детей	0	1	2	3	4	5	Всего
Число ответивших	6	8	45	21	5	3	88

Вычислите выборочное среднее. Охарактеризуйте разброс значений с помощью разных мер вариации, таких как выборочное стандартное отклонение, коэффициент вариации. Визуализируйте имеющийся вариационный ряд, построив полигон распределения и кумулятивную кривую.

Задача 2. «Студенты и автомобиль». Пусть имеются следующие данные о том, какова доля студентов университета, приезжающих на занятия на личном автомобиле:

Факультет	А	Б	В	Г	Д	Е	Ж
Доля студентов в общей численности факультета	0,12	0,06	0,17	0,09	0,10	0,08	0,08

Определите, чему равна, в среднем, на каждом факультете доля студентов, приезжающих на своем автомобиле. Рассчитайте выборочное стандартное отклонение и выборочный коэффициент вариации. О чём свидетельствует полученное значение коэффициента вариации? Что можно сказать о доле автомобилистов среди студентов разных факультетов?

Задача 3. «Выборы в США». В таблице представлены показатели явки избирателей на выборы в США в 2012 г. по некоторым штатам (объём выборки $n = 14$)[1].

Штат США	Явка избирателей, %
Арканзас	47,5
Делавэр	58,0
Флорида	55,2
Джорджия	52,2
Гавайи	40,0
Айдахо	55,5
Айова	67,1
Луизиана	57,1
Нью-Гемпшир	67,8

[1] По данным United States Election Project [http://elections.gmu.edu/] (проверено 19.11.2012).

Окончание табл.

Штат США	Явка избирателей, %
Северная Дакота	59,4
Южная Каролина	53,8
Южная Дакота	57,7
Вермонт	59,3
Вайоминг	56,8

Найдите медиану выборки, выборочные квартили, квартильный размах. Проинтерпретируйте полученные значения. По методу Дж. Тьюки определите, есть ли в этой выборке нетипичные значения наблюдений — «подозрительные» и/или «выбросы». Постройте коробчатую диаграмму «Ящик с усами».

Задача 4. «Стоимость жилья». Имеются следующие данные о стоимости вторичного жилья, предлагаемого к продаже на территории Южного федерального округа РФ[1]. В таблице приведены средние цены 1 м^2 общей площади жилых помещений типовых квартир на конец I квартала 2012 г.

Регион России	Средняя цена 1 м^2, тыс. руб.
Республика Адыгея	24 216
Республика Калмыкия	33 432
Краснодарский край	47 466
Астраханская обл.	29 852
Волгоградская обл.	31 508
Ростовская обл.	47 758

Найдите среднюю цену 1 м^2 вторичного жилья в регионах Южного федерального округа. Рассчитайте выборочное стандартное отклонение и выборочный коэффициент вариации. О чём свидетельствует полученное значение коэффициента вариации? Можно ли считать, что стоимость вторичного жилья в регионах ЮФО однородная (похожая)? Или нет?

Задача 5. «Конференции». Имеются следующие выборочные данные о том, в скольких конференциях принял участие тот или иной преподаватель университета за последние три года. Выборка исследования состоит из тридцати участников. Получены данные: 5, 4, 3, 3, 5, 4, 4, 4, 3, 4, 4, 5, 4, 4, 3, 2, 5, 3, 4, 4, 4, 3, 2, 5, 2, 5, 5, 2, 3, 3. Вычислите моду, медиану, среднее ариф-

[1] Данные Федеральной службы государственной статистики по Волгоградской области [http://www.volgastat.ru/federal/ufo14/default.aspx] (проверено 05.05.2012).

метическое, нижнюю и верхнюю квартили, размах вариации, дисперсию, стандартное отклонение, квартильный размах. Визуализируйте имеющуюся информацию, построив полигон распределения и кумулятивную кривую.

Задача 6. «Кино». В исследовании участникам опроса был задан вопрос: «Сколько примерно раз в месяц Вы ходите в кинотеатр?» Ответы распределились следующим образом:

Частота посещений	0	1 — 3	4 — 6	7 — 9	10 — 12	13 и более	Всего
Количество респондентов	6	11	8	8	4	3	40

Для выборки найдите среднее арифметическое и стандартное отклонение. Вычислите коэффициент вариации. О чём свидетельствует полученное значение коэффициента вариации?

Задача 7. «Офисы продаж». На территории города у торговой компании А* имеется пять офисов продаж. В таблице для каждого офиса продаж указаны число постоянных клиентов и величина их средней суммы покупок, совершённых за последний год (тыс. руб.):

Офисы продаж	№1	№2	№3	№4	№5
Количество постоянных клиентов	1 320	1 250	1 570	800	2 010
Средняя сумма покупок за последний год, тыс. руб.	28	20	32	25	25

Вычислите среднюю сумму покупок клиентов торговой компании. Найдите и интерпретируйте выборочный коэффициент вариации.

Задача 8. «Курение». Имеются две выборки наблюдений:

X: 5, 6, 6, 7, 6, 7, 7, 7, 8, 9, 8, 8 — число сигарет, выкуриваемых юношами в течение дня;

Y: 2, 4, 3, 6, 5, 7, 7, 9, 8, 10, 12, 9 — число сигарет, выкуриваемых девушками в течение дня.

Рассчитайте для двух выборок меры центральной тенденции — среднее арифметическое, медиану, моду.

Охарактеризуйте вариабельность *X* и *Y* с помощью таких показателей, как выборочное стандартное отклонение и коэффициент вариации.

Задача 9. «Стаж работы». Бригада экспериментального швейного цеха состоит из 12 швей, стаж работы которых на фабрике описывается следующей таблицей:

Швея	А	Б	В	Г	Д	Е	Ж	З	И	К	Л	М
Стаж работы на фабрике, лет	13	7	5	12	10	4	8	8	5	7	6	2

С помощью медианы охарактеризуйте стаж работы швей этого цеха. Интерпретируйте полученный ответ.

Задача 10. «Возрастной состав». В книге А. П. Чехова «Остров Сахалин (из путевых записок)» приводится следующий фрагмент о возрастном составе каторжного населения женского пола[1]:

Возраст, лет	10 и менее	10 — 20	20 — 35	35 — 45	45 — 55	55 — 65	65 — 75	75 и выше
Количество женщин	787	330	816	578	236	56	12	1

Рассчитайте средний возраст женщин с помощью среднего арифметического и дайте характеристику разброса значений с помощью выборочного стандартного отклонения. Рассчитайте выборочный коэффициент вариации. Интерпретируйте полученные ответы.

Задача 11. «Малиновое варенье». Винни-Пух признался, сколько банок малинового варенья он съел в течение 30 дней. Вот протокол:

День месяца	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Число банок, шт.	5	2	0	1	1	1	4	0	1	3	3	0	5	1	2

День месяца	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
Число банок, шт.	2	1	2	6	0	2	1	4	4	8	8	6	5	0	5

1. Что в данном случае является случайной величиной X?

2. Найдите среднее число банок варенья, которое он съедал в течение одного дня, с помощью различных мер центральной тенденции (медиана, среднее арифметическое, мода).

3. Охарактеризуйте разброс с помощью различных мер вариации (рассеивания): выборочный размах вариации, выборочные дисперсия и стандартное отклонение, выборочный квартильный размах.

4. В какой мере число ежедневно съедаемых им банок варенья можно считать одинаковым? (найдите коэффициент вариации).

5. Интерпретируйте ответы, полученные в результате расчётов.

Задача 12. «Обувной магазин». Работа обувного магазина торговой марки «Каблук» (*название условное*) строится по принципу индивидуальных продаж консультантов. Управляющего сетью этих обувных магазинов интересует, сколько пар обуви продаёт, в среднем, один консультант в течение рабочего дня. Ему известно, что число проданных пар обуви в течение дня одним консультантом является случайной величиной с законом распределения (дана выборка):

[1] *Чехов А. П.* Остров Сахалин. — Новосибирск: Сиб. унив. изд-во, 2009.

X, шт. (продано пар обуви за день)	0	1	2	3	4	5	6	7	8
Количество консультантов, демонстрирующих такой результат	5	20	65	40	10	8	8	3	1

1. Оцените результаты продаж персонала: с помощью различных мер центральной тенденции (медиана, среднее арифметическое, мода) найдите средний объём продаж, приходящийся на одного консультанта. Охарактеризуйте разброс с помощью различных мер вариации (рассеивания): выборочный размах вариации, выборочные дисперсия и стандартное отклонение, выборочный квартильный размах.

2. В какой мере можно считать похожими (однородными) результаты продаж персонала? (найдите коэффициент вариации).

3. Интерпретируйте ответы, полученные в результате расчётов.

Задача 13. «Репродуктивные стратегии». Во всероссийском исследовании «Молодёжь новой России: образ жизни и ценностные приоритеты», проведённом в 2007 г. Центром социального прогнозирования под руководством Ф. Э. Шереги, молодому населению в возрасте 17 — 26 лет был адресован следующий вопрос: «Сколько детей Вы планируете иметь с учётом Ваших реальных обстоятельств (учитывая тех детей, которые уже есть)?». Вот данные опроса[1]:

Количество детей	0	1	2	3	4	5	6
Количество респондентов	130	655	833	141	28	5	4

1. Найдите, чему равно *в среднем* планируемое число детей (с помощью среднего арифметического).

2. Дайте характеристику разброса с помощью выборочного стандартного отклонения.

3. Рассчитайте выборочные коэффициент вариации и коэффициент децильной вариации.

4. Интерпретируйте полученные ответы.

Задача 14. «Турагентство». Турагентство спрашивает своих клиентов: «В каком количестве зарубежных стран Вы успели побывать к настоящему моменту своей жизни? Назовите, пожалуйста, число». Ответы первых 11 клиентов представлены так:

X (стран): 11, 6, 1, 4, 2, 6, 0, 7, 6, 3, 17.

1. Найдите медиану выборки, нижнюю и верхнюю квартили выборки (по методу Дж. Тьюки).

---

[1] Данные Единого архива экономических и социологических данных [http://sophist.hse.ru/db/oprosy.shtml?ts=121&en=0] (проверено 04.06.2013).

2. Определите по методу Дж. Тьюки, есть ли в этой выборке нетипичные значения наблюдений — «подозрительные» и/или «выбросы».

3. Визуализируйте результаты своих расчётов, построив коробчатую диаграмму («Ящик с усами»).

Задача 15. «Товары первой необходимости». Жителям двух стран — N и M — был адресован вопрос: «Какую часть ежемесячного семейного бюджета вы обычно тратите на покупку продуктов питания и предметов первой необходимости? Укажите, пожалуйста, примерную долю этих расходов в процентах». Получены следующие результаты двух выборок:

Ответы жителей страны N, %: 10, 70, 65, 50, 55, 50, 80, 100, 60, 60, 50, 90, 100.

Ответы жителей страны M, %: 10, 20, 25, 40, 30, 30, 90, 20, 25, 95, 10, 5.

1. Найдите в обеих выборках медиану, нижнюю и верхнюю квартили (по методу Дж. Тьюки).

2. Определите по методу Дж. Тьюки, есть ли в этих выборках нетипичные значения наблюдений — «подозрительные» и/или «выбросы».

3. Визуализируйте результаты своих расчётов, построив коробчатые диаграммы («Ящик с усами»).

Задача 16. «Алиса и Борис». Алиса и Борис сравнивают количество чашек чая, которые они потребляли в течение дня на протяжении последних двух недель (14 дней).

Количество чашек чая (Алиса): 8, 7, 6, 10, 10, 9, 5, 3, 11, 7, 9, 4, 6, 4.

Количество чашек чая (Борис): 8, 8, 8, 7, 10, 8, 8, 2, 10, 7, 8, 8, 9, 7.

1. Охарактеризуйте потребление чая (для Алисы и Бориса), рассчитав медианы обеих выборок.

2. Найдите квартильный размах для каждой выборки.

3. Укажите область характерных (типичных) значений для Алисы и Бориса.

4. Проверьте, есть ли в выборках нетипичные значения наблюдений — «подозрительные» и/или «выбросы».

5. Визуализируйте результаты расчётов, построив диаграммы «Ящик с усами».

Задача 17. «Без названия». Выборка представлена значениями *X*: 10, 0, 12, 4, 5, 2, 2, 7, 13, 6, 7, 5, 6, 8, 15, 3. По методу Дж. Тьюки найдите медиану, нижнюю и верхнюю квартили, диапазон типичных значений. Укажите нетипичные значения в выборке, если они есть. Постройте диаграмму «Ящик с усами».

Задача 18. «Рост юношей и девушек». В ходе небольшого опроса, проведённого на одном из семинаров по математической статистике среди студентов факультета прикладной политологии НИУ «Высшая школа экономики», участникам семинара было предложено ответить на следующий вопрос: «Укажите, пожалуйста, Ваш рост (напишите с точностью до сантиметра)».

Объём выборки юношей составил 8 человек, объём выборки девушек — 11 человек. Получены следующие результаты:

Юноши: 169, 173, 181, 178, 172, 178, 178, 192;

Девушки: 172, 166, 174, 164, 176, 159, 161, 176, 175, 176, 167.

По методу Дж. Тьюки для обеих выборок найдите медиану, нижнюю и верхнюю квартили, диапазон типичных значений. Укажите нетипичные значения в выборках, если они есть. Постройте диаграммы «Ящик с усами».

Задача 19. «Время на спорт». В ходе того же опроса (см. задачу 18), проведённого среди студентов факультета прикладной политологии, участникам было предложено ответить на следующий вопрос: «Укажите, пожалуйста, сколько примерно часов в неделю Вы тратите на спорт и различные физические упражнения (включая бег, бассейн, утреннюю зарядку, ролики, танцы и прочие виды активности). Напишите с точностью до 0,5 ч. Объём выборки юношей составил 8 человек, объём выборки девушек — 11 человек:

Юноши: 2; 5,5; 6,5; 10; 2,5; 5; 7; 7;

Девушки: 7, 4, 2, 5, 4, 0, 0, 2, 3, 2, 5.

По методу Дж. Тьюки для обеих выборок найдите медиану, нижнюю и верхнюю квартили, диапазон типичных значений. Укажите нетипичные значения в выборках, если они есть. Постройте диаграммы «Ящик с усами».

Глава 3

КОЭФФИЦИЕНТЫ СВЯЗИ

Переходим к одному из интереснейших сюжетов в математической статистике — поиску *статистических* взаимосвязей между двумя признаками, явлениями, событиями и оценке тесноты этой связи с помощью коэффициентов корреляции.

В этой главе нам предстоит изучить коэффициент линейной корреляции К. Пирсона и коэффициент ранговой корреляции Ч. Э. Спирмена.

Однако прежде чем приступить к их описанию, необходимо рассмотреть шкалы измерения случайных величин. Напомним основные сведения, принципиально важные для понимания дальнейшего материала.

Надеемся, что базовые знания о шкалах уже приобретены читателями в курсах по методологии и методам социологических исследований.

§ 3.1. Шкалы

При проведении социологического исследования необходимо измерять те или иные свойства изучаемых объектов, их характеристики.

Какие-то из характеристик измерить (обнаружить и зафиксировать) достаточно просто, не составляет особого труда (например, пол человека, его масса, возраст, место проживания, число чашек кофе, выпиваемых им в течение дня, наличие личного автомобиля и многое другое).

Гораздо сложнее измерить, например, его ценностные ориентации, толерантность к окружающим людям, стиль жизни, социальный статус, религиозность, образованность, удовлетворенность жизнью, структуру бюджета времени, предпочтения в выборе книг для чтения.

Чтобы осуществить измерение исследуемого признака требуется спроектировать (создать) так называемую ***линейку*** (шкалу)[1]. При проектировании вопросов анкеты (или гайда интервью), анализе полученных данных в исследовании приходится непрерывно сталкиваться с необходимостью *присвоения чисел* (*числовых значений*) или каких-либо других *символов* ответам респондентов, руководствуясь чёткими *определёнными правилами* или *критериями*. Такая процедура измерения, позволяющая с помощью чисел или иных символов *моделировать* реальность, окружающий мир, называется *шкалированием*. Это позволяет измеряемым в ходе исследования характеристикам людей (чаще всего *качественным* характеристикам — например, предпочтениям, ценностям, установкам и др.) поставить *в соответствие* некоторые числа[2]. Так, процедура шкалирования служит средством «перевода» качественной характеристики в числовой признак (или какой-то символ), и лежит в основе математического анализа полученных в ходе исследования эмпирических сведений, данных. Выделяют ***три основных типа шкал*** в практике социологических исследований — номинальную, порядковую, количественную. Все они различаются по тем отношениям, которые *можно установить между объектами с их помощью*. Тип шкалы напрямую определяет *правомерность* и *целесообразность* применения тех или иных математических методов обработки и анализа данных. Рассмотрим основные положения, раскрывающие особенности этих типов шкал.

Номинальная шкала (шкала наименований или ***классификаций)*** — самый распространённый тип шкалы в практике исследований. Она считается низшим уровнем измерения, поскольку те числа, которые присваиваются в ходе исследования характеристикам объектов, *условны* — их нельзя воспринимать как числа в подлинном смысле слова. Они не несут в себе той смысловой нагрузки,

---

[1] Рекомендуем читателям ознакомиться с иллюстрированными лекциями Н. И. Даудрих по курсу «Методология и методы социологического исследования» (НИУ ВШЭ, факультет социологии). Тема лекции «Измерение. Виды шкал. Критерии качества» (визуальное сопровождение лекции размещено по ссылке: [http://www.hse.ru/data/2012/02/29/1265871084/03 Измерение.pdf] (проверено 07.05.2013). Н. И. Даудрих приводит следующее определение шкалы: «*Шкала в социологии — специально сконструированная линейка для измерения интересующих исследователя свойств объекта, т. е. присваивания им числовых значений — меток, шкальных оценок*».

[2] Рекомендуем читателям ознакомиться, например, с такими фундаментальными работами об измерении в социологии, как *Толстова Ю. Н.* Измерение в социологии. — М.: Инфра-М, 1998. — С. 9—68, 111—123; *Девятко И. Ф.* Методы социологического исследования. — М: КД «Университет», 2002. — С. 166—198; *Батыгин Г. С.* Лекции по методологии социологических исследований. — М.: Аспект Пресс, 1995. — С. 62—93; *Татарова Г. Г.* Методология анализа данных в социологии (введение): Учеб. пособие для вузов. — М.: Издательский дом «Стратегия», 1998. — С. 43—100.

которая обычно стоит за понятием числа. Числа выступают здесь всего лишь как *коды*, *метки*, *лейблы*, *маркеры*, *имена*, позволяющие классифицировать объекты, обозначить их подобие или, напротив, различие по некоторому признаку. Чаще всего здесь мы имеем дело с качественными признаками. Числа в данном случае (будучи просто именами) ***не выражают никакого превосходства*** одной категории ответов над другой. Так, когда в ходе исследования мы присваиваем число «1» жителям Москвы, число «2» — жителям Санкт-Петербурга, а число «3» — жителям Владивостока, то с их помощью мы вносим лишь возможность дифференциации людей по месту проживания — мы не судим о числах как о математических абстракциях. Здесь они служат не для количественной оценки объектов (людей), а выражают только их сходство или, напротив, различие по ***качественному*** признаку. Поэтому к этим числам (являющимся всего лишь кодами) не применимы какие-либо математические преобразования и вычисления — мы не можем складывать их, умножать или делить, находить разность, ставить знаки «<» или «>». Это не несет в себе ни малейшего содержательного смысла. Вопросы анкеты, в которых мы уточняем у респондентов *место жительства*, *пол*, *семейное положение*, *вероисповедание*, *политические предпочтения* (*например*, *за какую партию они голосовали на минувших выборах*), *профессию*, *национальность*, *марку личного автомобиля*, *любимый бренд одежды*, *номер водительского удостоверения и т. д.*, — это всё ***примеры*** признаков, имеющих не количественную, а качественную природу. Поэтому все числа, с помощью которых мы классифицируем категории ответов, *равноценны* и *равнозначны* между собой.

Что касается адекватности используемых по отношению к номинальным шкалам математических методов анализа, то здесь уместны и применимы только те виды математических методов анализа, которые основаны на *подсчёте частот* (*процентное* содержание каждой из категорий ответа, *мода*, а также критерий *хи*-квадрат, с которым мы познакомимся на дальнейших страницах учебника).

Порядковая шкала (шкала рангов) является следующим уровнем измерения (более высоким в сравнении со шкалой наименований). Мы располагаем возможностью сопоставлять объекты между собой, а также судить о том, как они расположены ***относительно друг друга по порядку (рангу)***. С помощью чисел, присваиваемых объектам, эта шкала позволяет исследователю оценивать *степень выраженности признака*, а значит, и сравнивать объекты между собой в терминах «больше» или «меньше», «лучше» или «хуже», «сильнее» или «слабее». Так, для порядковой шкалы справедливы отношения равенства и порядка. В порядковых шкалах у исследователей появляется возможность воспринимать числа не только как метки, но и *различать их по порядку*, смотреть на них как на *индикаторы рейтинга* объектов. В анкетных вопросах часто фигурируют категории ответов типа:

1) «*Полностью не* согласен»; 2) «*Скорее не* согласен»; 3) «*Скорее согласен*»; 4) «*Полностью согласен*». Или вводятся шкалы «от –2 до +2», «от 1 до 7» и другие, а респондентов просят оценить (выразить) ***степень*** своего согласия. Порядковые шкалы применяются для выявления позиции людей: ***выраженности*** их мнения, предпочтений, оценки, отношения к кому-либо или чему-либо. В отношении порядковых шкал правомерными в использовании являются те же математические методы, что и в отношении номинальных шкал, но ещё добавляются возможности вычисления таких параметров, как *медиана*, *квартили*, *децили*, *квартильный размах*, *коэффициент децильной вариации*, *коэффициент ранговой корреляции* (вычислять который мы научимся уже в данной главе).

Так, ***например***, студентов можно ранжировать по уровню академической успеваемости. Заметим, что шкала оценок, используемых преподавателями университетов или школьными учителями, — это классический пример использования именно порядковой шкалы. Почему? Потому что всё, что мы можем сказать о студенте, получившем по некоторому предмету оценку «10» (из 10 баллов) и о студенте, получившем по тому же предмету оценку «8» (из 10 баллов), так это то, что первый студент знает предмет ***лучше***, чем второй. Но мы не можем определить, *насколько* лучше или *во сколько раз* знания первого студента превосходят знания второго. Точно так же нельзя утверждать, что различие в знаниях этих студентов такое же, как и в знаниях тех, кто получили, например, «8» и «6» баллов. Расстояния между баллами не осмыслены, они не одинаковы: «10 минус 8» — не то же самое, что «8 минус 6». *Восприятие расстояний между баллами шкалы носит* ***субъективный характер*** *для тех, кто с помощью этой шкалы проводит оценивание*. Шкала оценок является порядковой (шкалой рангов), поэтому не вполне корректно выставлять итоговую оценку студента как, скажем, среднюю арифметическую текущих оценок[1].

Некоторыми наглядными ***примерами*** порядковой шкалы являются следующие признаки:

- «*Ступень* высшего образования» (например: «1» — «неполное высшее», «2» — «высшее», «3» — «учёная степень»);
- «Профессионально-должностной *статус*»[2], занимаемый человеком в управлении организацией, в которой он работает: «1» — «директор», «2» — «заместитель директора», «3» — «начальник отдела (руководитель направления)», «4» — «руководитель рабочей группы»;
- «*Степень* поддержки программы реформ в стране» (например: «+1» — «полностью поддерживаю», «0» — «поддерживаю частично», «–1» — «полностью не поддерживаю»);

---

[1] Хотя на практике часто пренебрегают этим обстоятельством.

[2] Если речь идёт о менеджменте компании.

• «*Степень* доверия к действующему президенту» (например: «+2» — «очень высокое доверие», «+1» — «высокое доверие», «–1» — «низкое доверие», «–2» — «крайне низкое доверие»);

• *Оценка качества знаний* студентов, оценка качества освоения учебной дисциплины (например: «1» — «крайне неудовлетворительно», «2» — «очень плохо», «3» — «плохо», «4» — «вполне удовлетворительно», «5» — «удовлетворительно», «6» — «хорошо», «7» — «очень хорошо», «8» — «почти отлично», «9» — «отлично», «10» — «блестяще» [1]);

• *Оценка качества услуг* туроператора (например: «–4» — «ужасное», «–3» — «неприемлемое», «–2» — «очень плохое», «–1» — «плохое», «0» — «среднее», «+1» — «хорошее», «+2» — «очень хорошее», «+3» — «отличное», «+4» — «великолепное»).

И других примеров можно привести огромное количество.

***Количественная шкала*: *интервальная* и *шкала отношений*.** Это высокие уровни измерения, обеспечивающие бóльшую информативность, в сравнении с двумя предыдущими типами шкал.

Интервальная шкала. Интервалы (расстояния) между числами (или какими-либо другими символами), используемыми в шкале, отражают *одинаковую меру величины* признака. Таким образом, для интервальных шкал осмысленны отношения равенства, порядка и самое главное (!) — осмысленна *разница между значениями шкалы, с помощью которой измеряется интересующая характеристика*. Здесь уже становится уместным и справедливым вопрос типа: «*На сколько?* измеряемый показатель больше (или меньше) у разных объектов». Промежутки (градации) этой шкалы количественно равны между собой и свидетельствуют о равных отрезках между значениями измеряемого признака. Так, расстояния между соседними числами (делениями) шкалы равны: расстояние между «5» и «2» такое же, как и между «7» и «4», т. е. в эти расстояния (дистанции) вкладывается один и тот же смысл. Однако на такой шкале *произвольно* выбирается (фиксируется, назначается) *точка отсчёта* (*начальная* точка, «условный нуль») и *произвольно* конструируется *единица измерения*. Самыми типичными и распространёнными примерами интервальной шкалы, поясняющими её смысл, служат: шкала температуры (шкала Цельсия[2], шкала Фаренгейта[3], шкала Рёмера[4] и др.), шкала

---

[1] Такова, например, шкала, которая используется преподавателями при аттестации и оценивании знаний студентов, обучающихся в НИУ «Высшая школа экономики».

[2] В качестве «условного нуля» (точки отсчёта) принимается температура замерзания пресной воды.

[3] В качестве «условного нуля» (точки отсчёта) берётся температура отвердевания (замерзания) смеси, состоящей из равных пропорций воды, соли и хлорида аммония (нашатырного спирта), взятых в соотношении 1:1:1.

[4] В качестве «условного нуля» (точки отсчёта) принимается температура замерзания солёной воды.

времени[1]. В социальных науках это могут быть специальные шкалы сумматорного типа (шкала Лайкерта, шкала Гуттмана, шкала Терстоуна и др.)[2] шкалы интеллекта IQ и пр.

Шкала отношений (*пропорций*). Самый высокий уровень измерения признаков, так как значению «0» на шкале соответствует реальный факт, что измеряемый признак отсутствует *вовсе*. Это, в свою очередь, означает, что точка отсчёта на шкале не является «условно назначенной» — напротив, она является ***фиксированной***, однозначной, единственно возможной. Шкала отношений (пропорций) — ***подлинно числовая*** шкала, в которой числа — ***результат измерений*** (а не результат процедуры кодирования, конструирования, моделирования). Числа воспринимаются именно как числа (со всеми их свойствами и возможностями интерпретации различий между ними). Ещё раз обратите внимание на то обстоятельство, что начало отсчёта исследователь выбирает на шкале далеко *не произвольно*.

Примеры шкалы отношений:

- *уровень безработицы в стране/регионе/городе* (*измеряемый в процентах*);
- *величина семейных расходов на покупку продуктов питания*;
- *величина объёма продаж магазина*;
- *доля рынка*;
- *количество решённых на экзамене задач*;
- *уровень младенческой смертности* (количество умерших в возрасте до одного года на 1 000 родившихся);
- *величина прожиточного минимума*;
- *удельный вес городского и сельского населения*;

---

[1] Так, говоря о шкале времени, можно дать следующие пояснения: в качестве точки отсчета («условного» нуля) могут быть выбраны разные даты/события, могут быть предложены разные подходы, традиции, календари для измерения времени. В них по-разному осуществляется выбор точки, фиксирующей начало временно́й оси. И в каждом из этих случаев, независимо от выбора шкалы, структура временны́х отрезков и соотнесение исторических событий на временно́й шкале (временно́м континууме) будут одинаковыми. Разным будет «условный нуль» шкалы, ее «условное» начало.

[2] В основе этих шкал лежит измерение латентного (скрытого, неявного, неизвестного, «загадочного», трудно поддающегося выявлению) признака с помощью ***серии суждений*** (с их помощью измеряются разные ***аспекты*** исследуемого признака), предложенных респонденту для оценки степени своего согласия. Суждения чаще всего отражают положительное или отрицательное отношение к чему-либо/кому-либо. Результатом такого измерения становится ***индекс*** отношения, выраженный в баллах. Чаще всего придерживаются следующего правила в интерпретации: чем выше значение индекса, тем более выраженным является положительное отношение; чем ниже индекс, тем более выраженным является отрицательное отношение к чему-либо/кому-либо. Линейка (шкала) имеет свое «условное» начало (точку отсчета), конструируемое исследователем.

- *урожайность пшеницы* (*ц/га*);
- *артериальное давление* и многое другое[1].

В отношении интервальных шкал и шкал пропорций правомерными в использовании являются практически *все* без исключения методы математической статистики. К числам справедливо применять все возможные математические операции, процедуры, вычисления.

§ 3.2. Коэффициент линейной корреляции Пирсона

В основе вычисления коэффициента корреляции Пирсона лежит идея нахождения (обнаружения) линейной связи между двумя признаками (переменными), уровень измерения которых ***не ниже*** интервального, ***например***:

- между *затратами на рекламу* и *объёмом валовой прибыли* предприятия;
- между *уровнем преступности* и *уровнем безработицы* в регионах страны;
- между *стажем работы* и *размером денежного вознаграждения* (заработной платы);
- между *процентом верно решённых в ходе ЕГЭ задач по физике* и *процентом верно выполненных заданий по химии*;
- между *частотой посещений театра* и *частотой посещений кинотеатра*;
- между *числом дорожно-транспортных происшествий* и *стажем вождения*.

Примеров и содержательных задач подобного типа (которые иллюстрируют возможность нахождения взаимосвязи между двумя признаками) можно привести множество. Прежде чем разобраться в математическом содержании понятия «коэффициент линейной корреляции», уточним одно из базовых понятий — понятие ковариации двух случайных величин (коснёмся определения этого понятия), а также свойства ковариации.

Ковариация служит мерой оценивания ***линейной*** взаимосвязи между двумя случайными величинами — X и Y — мерой их ***совместной*** (сопряжённой) ***изменчивости***. В основе нахождения ковариации лежит двумерное (совместное) распределение вероят-

---

[1] Говоря про количественные шкалы, конечно, стоит задуматься, например, над такими примерами, как возраст человека, количество детей в семье и др.. С одной стороны, один год равен 365 дням и в возрасте 0—10 лет, и в возрасте 30—40 лет. Но в содержательном отношении (в плане физического и интеллектуального развития) это очень разные (!) годы. И, анализируя разные участки числовых отрезков подобных признаков (переменных), важно отдавать себе отчёт в том, а одинаковы ли расстояния между значениями в социологическом смысле?

ностей этих величин, позволяющее получить представление о том, с какой вероятностью встречается *каждое из возможных сочетаний* (комбинаций) X и Y, а именно: какова вероятность (шанс) каждой потенциально возможной пары $(X_i;\ Y_i)$ в реальности. Ковариация показывает *линейную согласованность* двух распределений, отражает, *что*, в среднем, происходит со значениями одной величины (увеличиваются они или уменьшаются?), в зависимости от того, как значения другой величины меняются (увеличиваются или же уменьшаются). Ковариация — это конкретное ***число***, которое может быть найдено следующим образом[1]:

$$\mathrm{Cov}(X;Y) = E[(X - E(X))\cdot(Y - E(Y))] =$$
$$= E(XY - X\cdot E(Y) - Y\cdot E(X) + E(X)\cdot E(Y)) =$$
$$= E(XY) - E(X\cdot E(Y)) - E(Y\cdot E(X)) + E(E(X)\cdot E(Y)) = ?$$

Теперь, чтобы упростить полученное выражение, воспользуемся знанием одного из свойств математического ожидания: о том, что математическое ожидание константы (конкретного числа) равно самой константе. В нашем случае этими константами (конкретными числами) выступают математические ожидания X и Y (ведь они фиксированы и принимают определённые значения). Получим:

$$E(XY) - E(X\cdot E(Y)) - E(Y\cdot E(X)) + E(E(X)\cdot E(Y)) =$$
$$= E(XY) - \underline{E}(X\cdot E(Y)) - \underline{E}(Y\cdot E(X)) + \underline{E}(E(X)\cdot E(Y)) =$$
$$= E(XY) - \underline{E(Y)}\cdot E(X) - \underline{E(X)}\cdot E(Y) + \underline{E(X)}\cdot\underline{E(Y)} =$$
$$= \langle \text{упрощаем выражение} \rangle = E(XY) - E(X)\cdot E(Y).$$

Следовательно, ***ковариация*** может быть вычислена как ***разность между математическим ожиданием произведения двух случайных величин ($E(XY)$) и произведением математических ожиданий двух величин в отдельности ($E(X)\cdot E(Y)$)***.

Теперь уточним (вспомним) ***свойства ковариации***.

• ***Во-первых***, ковариация обладает свойством симметричности: $\mathrm{Cov}(X; Y) = \mathrm{Cov}(Y; X)$.

• ***Во-вторых***, ***если*** величины X и Y <u>не</u>зависимы, ***то*** их ковариация однозначно равна нулю. Обратное утверждение ложно: нельзя сказать, что ***если*** ковариация двух величин равна нулю, ***то*** две величины могут считаться независимыми. Из равенства ковариации нулю еще не следует независимость двух величин. Можно лишь утверждать, что между величинами *нет линейной* взаимосвязи. Однако вопрос о том, есть ли между ними какая-либо другая функциональная зависимость, остается открытым: зависимость другого рода вполне может присутствовать (например, квадратичная), однако ковариация не измеряет её. Важно помнить, что ковариация обна-

[1] Здесь буквой E обозначено математическое ожидание случайной величины.

руживает только наличие (или отсутствие) линейной связи, и никакой больше.

• ***В-третьих***, ковариация случайной величины с самой собой совпадает со значением её дисперсии:

$$\mathrm{Cov}(X;\, X) = E(X^2) - E(X)\cdot E(X) = E(X^2) - (E(X))^2 = D(X).$$
$$\mathrm{Cov}(Y;\, Y) = E(Y^2) - E(Y)\cdot E(Y) = E(Y^2) - (E(Y))^2 = D(Y).$$

• ***В-четвертых***, значение ковариации меняется в зависимости от того, применены ли по отношению к X и/или Y какие-либо *произвольные линейные преобразования*. Это означает, что ковариация ***чувствительна*** к смене «масштаба» единицы измерения X и/или Y. Например, пусть к рассматриваемым величинам X и Y, мы применим следующие линейные операции: все значения величины X умножим на константу a и ко всем полученным значениям прибавим число c; все значения величины Y умножим на константу b и ко всем полученным значениям прибавим число d. Далее найдем ковариацию новых величин: $(aX + c)$ и $(bY + d)$. Укажем, что $\mathbf{Cov}(\boldsymbol{aX} + \boldsymbol{c}; \boldsymbol{bY} + \boldsymbol{d}) = \boldsymbol{ab}\,\mathbf{Cov}(\boldsymbol{X}; \boldsymbol{Y})$. Итак, до того, как мы воспользовались линейными преобразованиями, ковариация была равна $\mathrm{Cov}(X; Y)$, а после — *увеличилась в* (ab) *раз*. Видим, что такие преобразования исходных значений X и Y приводят к тому, что само значение ковариации (в абсолютном выражении) *меняется*, становится другим — это уже *другое число* — меньшее или большее исходного. Но ведь при этом такие линейные преобразования ***никак не меняют сути взаимозависимости между двумя величинами*** X и Y (!), они «не затрагивают» её. Разве можно сказать, что мы как-то «исказили» структуру отношений между значениями двух величин X и Y, применив к имеющимся величинам X и Y линейные комбинации вида $\boldsymbol{aX} + \boldsymbol{c}$ и $\boldsymbol{bY} + \boldsymbol{d}$? Конечно, нет. Поэтому ***ковариацию нельзя рассматривать как величину безотносительную***, «***безразмерную***». Её значение, напротив, целиком определяется единицами измерения самих случайных величин. А потому полученное в ходе вычислений абсолютное значение ковариации очень трудно интерпретировать «в лоб»: само по себе оно не позволяет говорить о том, в какой мере (т. е. насколько тесно — сильно или слабо) две величины взаимосвязаны. Так, большое значение ковариации ещё не указывает на высокую степень зависимости между величинами, а малое значение не означает, что между величинами почти нет никакой линейной зависимости. *Значение ковариации* напрямую *зависит от единиц измерения* X *и* Y, *а также от их дисперсий*. Так, если X измеряется в процентах, а Y — в рублях, то ковариация будет измеряться в таких единицах, как (%·руб.). Поэтому, чтобы устранить это «неприятное обстоятельство», связанное с ограничениями в интерпретации значения ковариации, полученное число нормируют (т. е. учитывают «масштаб») с помощью стандартных отклонений величин X и Y. Тем самым «избавляются» от первоначальных единиц измерения.

Этот «переход» делает ковариацию величиной безразмерной. Теперь пользуются не значением ковариации в «чистом виде», а её нормированным значением — ***корреляцией*** **(*коэффициентом корреляции*** *или* ***теоретической корреляцией*****)**:

$$\mathrm{Cor}(X;Y) = \frac{\mathrm{Cov}(X;Y)}{\sqrt{D(X)}\sqrt{D(Y)}}.$$

Коэффициент корреляции показывает ***тесноту*** *связи между двумя признаками* (*переменными*), *интенсивность её присутствия*. *Теснота связи* определяется тем, *насколько плотно* наблюдения (точки) лежат по отношению к некоторой прямой (вектору). Так, коэффициент линейной корреляции — это мера «плотности расположения» точек по отношению к некоторой прямой (т. е. линии). Коэффициент корреляции *тем выше*, чем *с большей плотностью* точки сосредоточены относительно *наклонной* прямой — как бы «вырисовывают» её, обнаруживая при этом линейную зависимость между двумя величинами.

Сравним две диаграммы рассеивания, представленные на рис. 3.1. На диаграмме № 1 (рис. 3.1, *а*) точки расположены крайне хаотично. Нельзя выделить никакой закономерности между величинами X и Y, в то время как на диаграмме № 2 (рис. 3.1, *б*) расположение наблюдений довольно плотное — они словно «вычерчивают» линию, относительно которой скапливаются, обнаруживая при этом довольно высокую тесноту связи между величинами X и Y.

Итак, коэффициент линейной корреляции указывает на *степень выраженности тесноты связи* между двумя переменными.

Укажем ***свойства коэффициента корреляции***.

• ***Во-первых***, в том случае если две величины независимы (т. е. между ними вообще нет никакой связи), то для них коэффициент

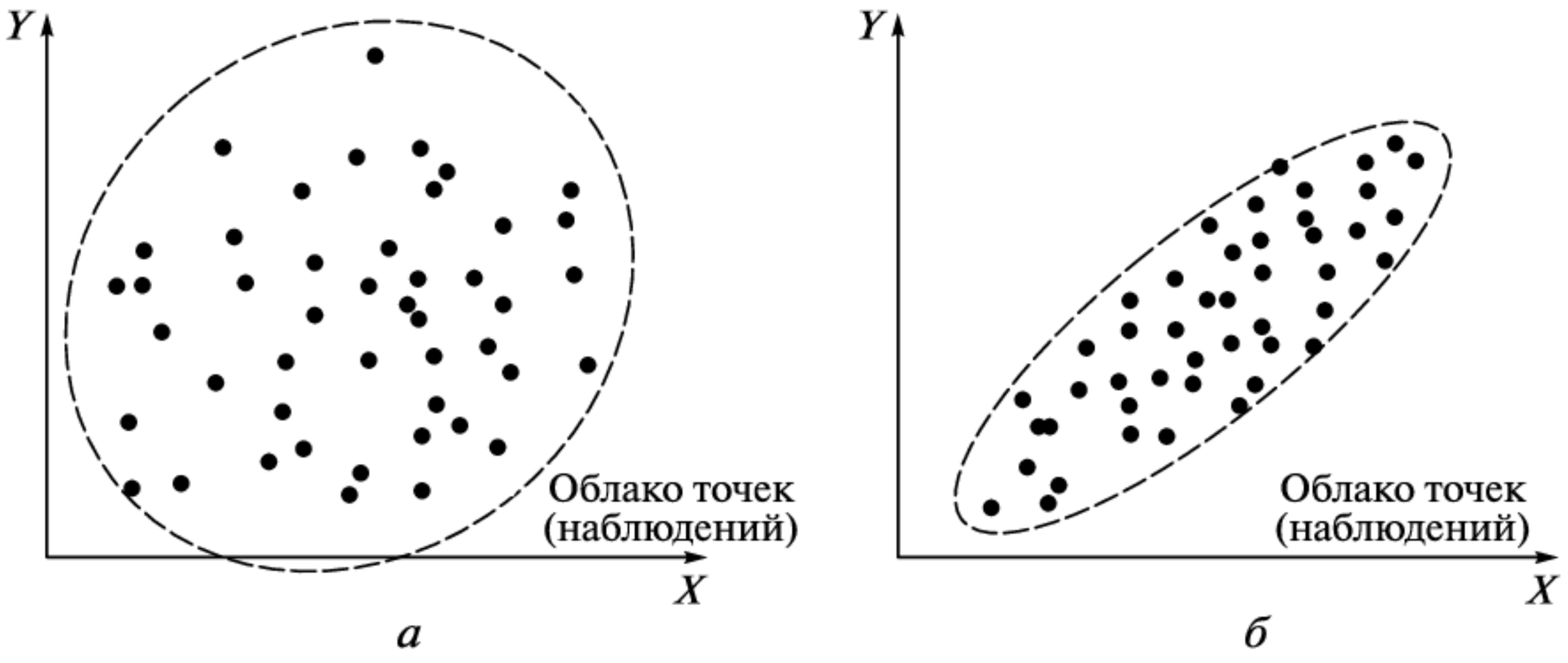


Рис. 3.1. Примеры диаграмм рассеивания:
а — диаграмма №1; *б* — диаграмма №2

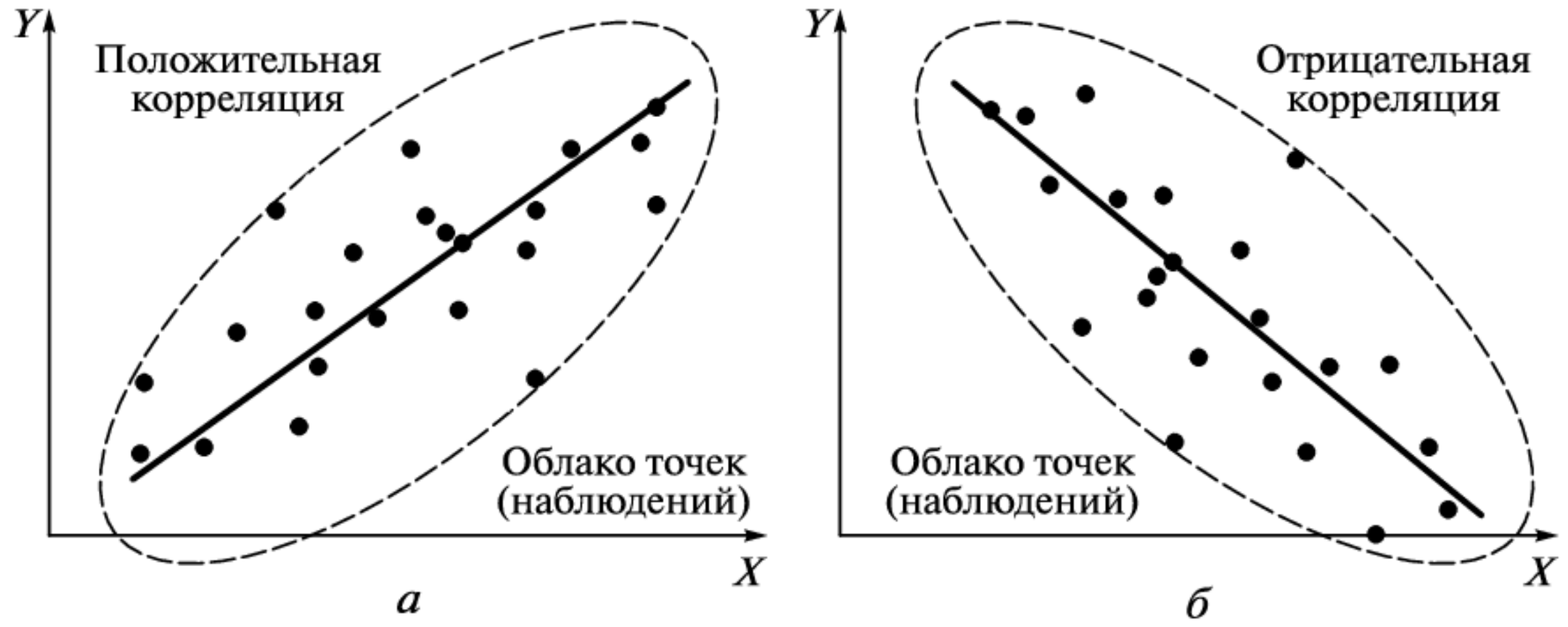


Рис. 3.2. Примеры прямой (*а*) и обратной (*б*) линейной связи

корреляции равен нулю: $\mathrm{Cor}(X;Y) = 0$. Если же в ходе исследования оказалось, что коэффициент корреляции для каких-то двух величин равен нулю, то из этого следует только то, что между показателями нет линейной связи.

• ***Во-вторых***, значение коэффициента корреляции по модулю не превосходит единицу: $|\mathbf{Cor}(X, Y)| \leq 1$. Это означает, что коэффициент корреляции может принимать значения в диапазоне ***от –1 до 0*** и ***от 0 до +1***. В случае если коэффициент корреляции меньше нуля, говорят о том, что связь между признаками отрицательная, рассматриваемые величины ***отрицательно коррелированы***. И, напротив, положительный коэффициент корреляции указывает на наличие положительной линейной зависимости, т. е. рассматриваемые величины ***положительно коррелированы***. Наличие прямой и обратной связи иллюстрируется на рис. 3.2.

• ***В-третьих***, в том случае, если коэффициент корреляции равен –1 или +1, то между рассматриваемыми величинами существует абсолютная линейная зависимость, которая носит детерминистский характер: *каждая из двух величин может быть представлена как* ***линейная функция*** *от другой величины. В первом случае* (–1) — *это* ***обратная*** *зависимость, во втором случае* (+1) — ***прямая*** *зависимость.*

Таким образом, коэффициент корреляции может служить мерой линейной взаимосвязи двух случайных (т. е. изменчивых) величин. Для вычисления коэффициента корреляции ***надо знать совместное распределение*** двух величин (показателей). Часто бывает, что на практике такого распределения мы достоверно не знаем (у нас может не быть достаточной информации, необходимых сведений). Однако можно провести измерения в ходе *выборочного* исследования и тогда для каждого объекта исследования (а если речь идет об опросе общественного мнения, то для каждого респондента) зафиксировать значения X_i и Y_i и по выборке получить сведения о совместном распределении X и Y. В результате будут получены пары наблюдений

$\begin{pmatrix} x_1 & x_2 & x_3 & x_4 & x_5 & \dots & x_n \\ y_1 & y_2 & y_3 & y_4 & y_5 & \dots & y_n \end{pmatrix}$, по которым можно ***оценить*** (т. е. найти ***примерное*** значение) теоретический (истинный, характерный для изучаемой генеральной совокупности) коэффициент корреляции. Выборочная оценка коэффициента линейной корреляции была предложена английским учёным *Карлом Пирсоном* (1857 — 1936). Как правило, выборочная оценка коэффициента корреляции обозначается буквой «*r*» или сокращённо «cor» и вычисляется по формуле

$$r_{\text{Пирсон}} = \frac{\text{cov}(X;Y)}{s_x s_y},$$

т. е. коэффициент корреляции равен отношению выборочной ковариации между X и Y к произведению выборочных оценок стандартных отклонений двух признаков, изучаемых в ходе исследования. Ковариация ***для выборки*** наблюдений вычисляется следующим образом:

$$\text{cov}(X;Y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n}.$$

Для каждой пары наблюдений (точек) рассчитывается *своя частная сопряжённая изменчивость* как $(x_i - \bar{x})(y_i - \bar{y})$. Далее сопряжённые изменчивости суммируются по всем парам наблюдений, а полученная общая сумма нормируется к объёму выборки (т. е. к общему числу пар наблюдений). С содержательной точки зрения ковариация показывает, чему равна ***в среднем*** сопряжённая изменчивость для одной (т. е. каждой) пары наблюдений$(x_i; y_i)$. Знак умножения между скобками $(x_i - \bar{x})(y_i - \bar{y})$ свидетельствует о том, что мы стремимся зафиксировать, как ***одновременно*** ведут себя значения признаков $(x_i; y_i)$по отношению к своим ***средним арифметическим*** $(\bar{x};\bar{y})$.

Истинные значения стандартных отклонений рассматриваемых показателей (т. е. $\sqrt{D(X)}$ и $\sqrt{D(Y)}$) нам достоверно не известны, поэтому оцениваем их примерно — на основе выборочных дисперсий (s_x^2 и s_y^2).

Однако сто́ит обратить внимание, что вовсе не обязательно рассчитывать отдельно числитель и знаменатель выражения $r_{\text{Пирсон}} = \frac{\text{cov}(X;Y)}{s_x s_y}$. Гораздо удобнее осуществить *некоторые преобразования*: мы видим, что при вычислениях ковариации и стандартных отклонений задействован объём выборки «*n*». Поэтому можно выборочную оценку ковариации и выборочные оценки стандартных отклонений свести вместе, вынеся за скобки объём выборки. Тогда

получим единую формулу для нахождения коэффициента корреляции «без участия» n:

$$r_{\text{Пирсон}} = \frac{\operatorname{cov}(X;Y)}{s_x s_y} = \frac{\dfrac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{n}}{\sqrt{\sum_{i=1}^{n}\dfrac{(x_i-\bar{x})^2}{n}}\sqrt{\sum_{i=1}^{n}\dfrac{(y_i-\bar{y})^2}{n}}} =$$

$$= \frac{\dfrac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{n}}{\dfrac{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}}{\sqrt{n}}\dfrac{\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}{\sqrt{n}}} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{i=1}^{n}(y_i-\bar{y})^2}},$$

т. е.

$$r_{\text{Пирсон}} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{i=1}^{n}(y_i-\bar{y})^2}}.$$

Интерпретация значений коэффициента корреляции Пирсона (т. е. ***выборочного***).

В практике исследований используются некоторые *условные* границы коэффициента корреляции, которые служат для интерпретации полученных на выборке значений:

- если значение коэффициента корреляции не превышает (по модулю) 0,4, т. е. $\boldsymbol{0 \le |r| < 0{,}4}$, то говорят, что теснота связи выражена *слабо*, либо вообще отсутствует;
- если значение коэффициента корреляции лежит (по модулю) в диапазоне $\boldsymbol{0{,}4 \le |r| < 0{,}7}$, то это свидетельствует об *умеренно* выраженной тесноте связи;
- если значение коэффициента корреляции находится (по модулю) в интервале $\boldsymbol{0{,}7 \le |r| < 1}$, то это интерпретируется как *высокая* теснота связи;
- если значение коэффициента корреляции ***в точности равно*** (по модулю) $\boldsymbol{1}$, то это интерпретируется как присутствие жестко детерминистской (однозначной) линейной зависимости;
- знак («+» или «–») указывает на наличие положительной (прямой) или отрицательной (обратной) зависимости между переменными X и Y соответственно.

Разберём конкретную практическую задачу.

Пример 3.1. «Безработица и преступность». Рассмотрим некоторые данные официальной статистики для 11 регионов Северо-Западного федерального округа РФ. Имеются сведения об уровнях безработицы и преступности за 2010 г. Безработица определяется как отношение численности незанятого населения к общей численности трудоспособного населения (%). Преступность измеряется как число преступлений, зарегистрированных на территории региона СЗФО, приходящееся на 100 тыс. жителей этого региона. В результате имеем таблицу:

Регион РФ в составе СЗФО	Уровень безработицы, %	Уровень преступности, число преступлений на 100 тыс. жителей
Архангельская обл.	7,2	2 124
Вологодская обл.	7,9	2 266
Калининградская обл.	10,9	2 108
Ленинградская обл.	7,2	1 978
Мурманская обл.	7,6	1 958
Ненецкий автономный округ	9,7	2 303
Новгородская обл.	6,4	2 187
Псковская обл.	11	1 901
Республика Карелия	10	2 299
Республика Коми	11,8	2 475
Санкт-Петербург	4,1	1 680

Можно ли считать две эти социальные характеристики — уровень безработицы и уровень преступности — ***коррелированными***?

Решение. Задачи на нахождение коэффициента линейной корреляции могут быть решены в несколько этапов (шагов):

1) строится диаграмма рассеивания точек (наблюдений). Это позволяет сделать наглядной имеющуюся информацию по изучаемым признакам и зафиксировать предварительный «интуитивный» прогноз относительно того, в какой мере рассматриваемые признаки линейно взаимосвязаны;

2) вычисляются средние значения признаков (отдельно для X и для Y);

3) вычисляются основные компоненты формулы: $\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})$, $\sum_{i=1}^{n}(x_i-\bar{x})^2$ и $\sum_{i=1}^{n}(y_i-\bar{y})^2$. Так, следует найти сумму произведений, показывающих сопряжённую изменчивость значений X и Y — в разрезе каждой пары наблюдений, а далее — суммы квадратов отклонений для X и для Y;

4) вычисляется коэффициент корреляции;

5) даётся интерпретация, содержательный вывод.

Последовательно реализуем этот алгоритм применительно к данной задаче.

Во-первых, строим диаграмму (облако) рассеивания, схематично рисуем линию тренда (рис. 3.3). Точками (объектами) являются 11 регионов РФ. Каждая точка характеризуется парой координат — значением уровня безработицы и значением уровня преступности. Видим, что точки весьма плотно лежат относительно некоторой прямой. Поэтому можно ожидать, что в расчётах получится оценка коэффициента корреляции Пирсона, свидетельствующая о том, что теснота связи выражена в достаточной степени, а также что уровень преступности и уровень безработицы положительно коррелированы (т. е. чем выше значения одного показателя, тем, как правило, выше значения и другого показателя). Далее произведём расчёты.

Во-вторых, находим средние арифметические:

$$\bar{x}=\frac{7{,}2\cdot 2+7{,}9+10{,}9+7{,}6+9{,}7+6{,}4+11+10+11{,}8+4{,}1}{11}\approx 8{,}5\ (\%)$$

— средний уровень безработицы в регионах СЗФО;

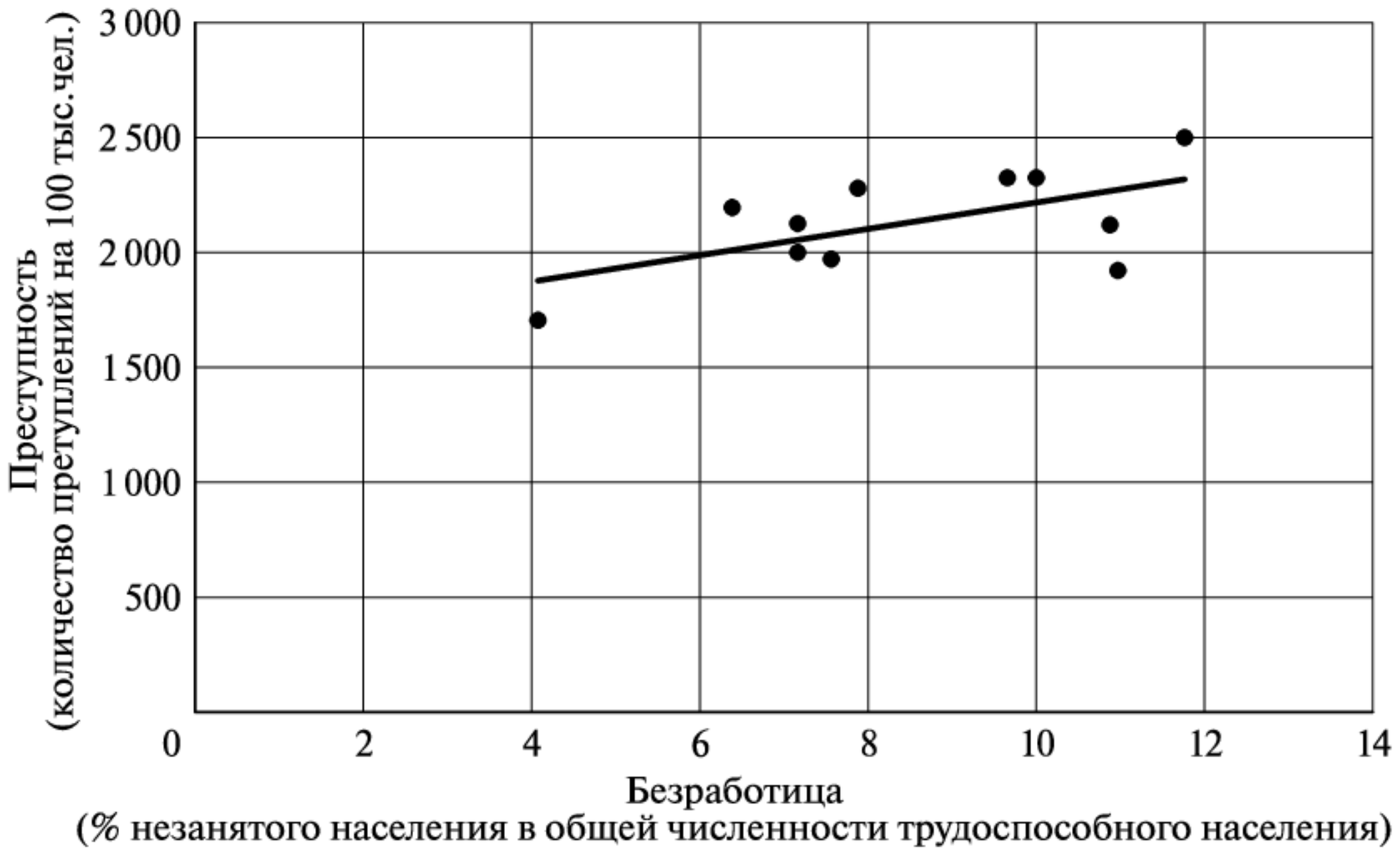


Рис. 3.3. Диаграмма рассеивания (см. пример 3.1)

$$\bar{y} = (2\,124 + 2\,266 + 2\,108 + 1\,978 + 1\,958 + 2\,303 + 2\,187 +$$
$$+ 1\,901 + 2\,299 + 2\,475 + 1\,680) / 11 \approx 2116 \text{ (преступлений)}$$

— среднее число преступлений, приходящееся на 100 тыс. жителей региона.

В-третьих, вычисляем основные компоненты формулы для определения коэффициента корреляции. Это можно сделать достаточно быстро, построив расчётную таблицу, в которой отразить ход вычислений таких величин, как $(x_i - \bar{x})$, $(x_i - \bar{x})^2$, $(y_i - \bar{y})$, $(y_i - \bar{y})^2$, $(x_i - \bar{x})(y_i - \bar{y})$, $\sum_{i=1}^{n}(x_i - \bar{x})^2$, $\sum_{i=1}^{n}(y_i - \bar{y})^2$ и $\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$. Получим:

x_i	y_i	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$
7,2	2 124	7,2 – 8,5 = –1,3	2 124 – 2 116 = 8	–1,3 · 8 = –10,4	1,69	64
7,9	2 266	7,9 – 8,5 = –0,6	2 266 – 2 116 = = 150	–0,6 · 150 = –90	0,36	22 500
10,9	2 108	10,9 – 8,5 = 2,4	2 108 – 2 116 = –8	2,4 (–8) = –19,2	5,76	64
7,2	1 978	7,2 – 8,5 = –1,3	1 978 – 2 116 = = –138	–1,3 (–138) = = 179,4	1,69	19 044
7,6	1 958	7,6 – 8,5 = –0,9	1 958 – 2 116 = = –158	–0,9 (–158) = = 142,2	0,81	24 964
9,7	2 303	9,7 – 8,5 = 1,2	2 303 – 2 116 = = 187	1,2 · 187 = 224,4	1,44	34 969
6,4	2 187	6,4 – 8,5 = –2,1	2 187 – 2 116 = 71	–2,1 · 71 = –149,1	4,41	5 041
11	1 901	11 – 8,5 = 2,5	1 901 – 2 116 = = –215	2,5 (–215) = = –537,5	6,25	46 225
10	2 299	10 – 8,5 = 1,5	2 299 – 2 116 = = 183	1,5 · 183 = 274,5	2,25	33 489
11,8	2 475	11,8 – 8,5 = 3,3	2 475 – 2 116 = = 359	3,3 · 359 = 1 184,7	10,89	128 881
4,1	1 680	4,1 – 8,5 = –4,4	1 680 – 2 116 = = –436	–4,4 (–436) = = 1 918,4	19,36	190 096
$\bar{x} = 8{,}5$	**$\bar{y} = 2116$**	—	—	**$\sum_{i=1}^{n=11}(x_i - \bar{x})(y_i - \bar{y}) = 3117{,}4$**	**$\sum_{i=1}^{n=11}(x_i - \bar{x})^2 = 54{,}91$**	**$\sum_{i=1}^{n=11}(y_i - \bar{y})^2 = 505\,337$**

Далее определяем коэффициент линейной корреляции:

$$r_{\text{Пирсон}} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{3\,117{,}4}{\sqrt{54{,}91}\sqrt{505\,337}} = \frac{3\,117{,}4}{7{,}41 \cdot 710{,}87} \approx 0{,}59.$$

Положительное значение коэффициента корреляции указывает на наличие ***прямой*** зависимости между уровнем безработицы и преступностью в рассмотренных регионах, которая выражена ***умеренно***. Вывод: да, можно считать две эти социальные характеристики взаимосвязанными (коррелированными).

§ 3.3. Коэффициент ранговой корреляции Спирмена

Для переменных, измеренных на основе *порядковой шкалы*, в математике разработаны различные коэффициенты ранговой корреляции[1]. Наиболее распространённым в исследовательской практике является коэффициент, предложенный английским психологом, статистиком Чарльзом Эдвардом Спирменом (1863 — 1945). Ключевым понятием является слово «ранг» — своего рода место в иерархии, занимаемое объектом. В основе вычисления этого коэффициента корреляции лежит оценка *степени согласованности рангов* (их *совпадения* или *не совпадения*). Все объекты выборки *ранжируются* (т. е. расставляются, выстраиваются *по порядку или рейтингу* — по убыванию или возрастанию) сначала по признаку *X*, а затем по признаку *Y*. В результате получаются *две ранжировки* — одна для *X*, другая соответствует переменной *Y*. Коэффициент корреляции Спирмена позволяет оценить степень *согласованности рангов в двух ранжировках*. Чем больше коэффициент стремится ***к значению «+1»***, тем в большей степени между рангами существует ***положительная зависимость*** (*положительная согласованность*). При стремлении коэффициента ***к значению «–1»*** очевиднее совершенно противоположный факт — ранги столь же *согласованы*, *но в обратном порядке* **(*отрицательно согласованы*)**. Чем ближе ***значение коэффициента к нулю***, тем более очевидным становится ***отсутствие какой-либо согласованности*** рангов. В основе вычислений лежит расчёт так называемой «***разности рангов***».

Формула для вычисления (содержательный смысл компонентов которой будет, безусловно, пояснён далее) выглядит так:

---

[1] При этом коэффициент ранговой корреляции может быть рассчитан наряду с коэффициентом линейной корреляции, когда речь идет о количественной шкале.

$$r_{\text{Спирмен}} = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)},$$

где d_i^2 — сумма квадратов разности рангов по первому и второму признакам (т. е. по признакам X и Y). Значение получаемого коэффициента корреляции будет лежать в диапазоне от –1 до +1. ***Значение «+1» соответствует полной взаимосвязи рангов, а значение «–1» — полной противоположности рангов***.

Пример 3.2. «Учёба». Прежде чем перейти к решению задачи, иллюстрирующей алгоритм вычислений коэффициента, попытаемся осмыслить словосочетание «ранжирование объектов», т. е. приписывание рангов наблюдениям в выборке. Пусть имеется совокупность, состоящая из шести объектов (студентов). Рассмотрим оценки этих студентов, полученные ими в зимнюю сессию по двум предметам — история социологии (X) и линейная алгебра (Y). Каждый объект характеризуется каким-то значением X и некоторым значением Y. Как X, так и Y измеряются в баллах по 10-балльной порядковой шкале:

Студент	Иван	Артём	Ольга	Мария	Пётр	Анна
X, оценка по истории социологии	5	7	9	6	10	8
Y, оценка по линейной алгебре	6	7	4	5	8	10

Проранжируем студентов по двум признакам — сначала по оценке X, затем — по оценке Y. «Проранжировать» — это значит определить «место в рейтинге», упорядочить: определить, кто на первом месте, кто на последнем. В нашем распоряжении имеется 6 «посадочных мест». И эти места нам нужно распределить между студентами. Условимся ранг «1» присваивать тому студенту, чей балл по предмету является наивысшим, тогда ранг «6» будет соответствовать студенту с наименьшим результатом. В итоге получим:

Студент	Иван	Артём	Ольга	Мария	Пётр	Анна
l_i, ранг по X	6	4	2	5	1	3
k_i, ранг по Y	4	3	6	5	2	1

Мы осуществили переход от непосредственно исходных баллов (оценок) к рангам, и получили две ранжировки. Узнать, в какой мере оценки по двум предметам согласуются, значит найти ***степень совпадения двух ранжировок***. Поэтому дальнейшим логичным действием является нахождение разности между этими рангами для

каждой пары значений. Это позволяет оценить, как *велико* или *мало расхождение* (несоответствие). Разность рангов — её принято обозначать d_i — вычисляется как $d_i = l_i - k_i$. Добавляем в последнюю таблицу ещё одну строку:

Студент	Иван	Артём	Ольга	Мария	Пётр	Анна
l_i, ранг по X	6	4	2	5	1	3
k_i, ранг по Y	4	3	6	5	2	1
d_i, разность рангов	+2	+1	–4	0	–1	+2

Для оценки общей величины соответствия рангов нужно взять сумму. Обратите внимание, что одна часть разностей принимает положительные значения, а другая — отрицательные, поэтому сложить «в лоб» эти значения нельзя, чтобы «+» и «–» не погасили друг друга. Для решения этой проблемы в математике идут на оправданную «хитрость» — вычисляют квадраты разностей рангов и суммируют их как общий итог. Так, в данном примере получаем таблицу с ещё одной дополнительной строкой и столбцом, резюмирующим расчёты:

$$\sum_{i=1}^{6} d_i^2 = 4 + 1 + 16 + 0 + 1 + 4 = 26.$$

Студент	Иван	Артём	Ольга	Мария	Пётр	Анна	Всего
l_i, ранг по X	6	4	2	5	1	3	
k_i, ранг по Y	4	3	6	5	2	1	
d_i, разность рангов	+2	+1	–4	0	–1	+2	
d_i^2, квадрат разности рангов	+4	+1	+16	0	+1	+4	26

Теперь можно подставить значение «26» в формулу для вычисления коэффициента ранговой корреляции:

$$r_{\text{Спирмен}} = 1 - \frac{6 \cdot 26}{6 \cdot (6^2 - 1)} = 1 - \frac{156}{210} = 0{,}26.$$

Согласованность (совпадение) оценок, полученных студентами по двум предметам, очень слабая. Нельзя сказать, что чем выше оценка по истории социологии, тем выше оценка по линейной алгебре. И наоборот. Точно также нельзя сказать, что чем выше оценка по одному из предметов, тем ниже оценка по другому предмету. Данные не обнаруживают какой-либо явной закономерности.

Ранги, образующие «связки». Рассмотрим ещё одну ситуацию, когда значения признака для некоторых объектов совпадают.

Вернёмся к *примеру 3.2* и предположим, что оценки, полученные студентами Артёмом и Марией по линейной алгебре, одинаковые и равны 6 баллам. Пусть имеется таблица:

Студент	Иван	Артём	Ольга	Мария	Пётр	Анна
X, оценка по истории социологии	5	6	9	6	10	8
Y, оценка по линейной алгебре	6	7	4	5	8	10

В таком случае Артём и Мария *как бы делят между собой* 4-е и 5-е места в рейтинге (ранжировке). Поскольку оценки у этих студентов одинаковые, у нас нет объективных оснований отдать предпочтение в присвоении рангов кому-то одному — допустим, Артёму присвоить ранг 5, а Марии — 4, или же Марии — 5, а Артёму — 4. В таких случаях прибегают к маленькой математической «хитрости», выражающейся в исследовательской беспристрастности, а именно, вычисляют среднее между рангами, на которые «претендуют» студенты, т. е. находят среднее арифметическое значений 4 и 5. Средний ранг для Артёма и Марии равен 4,5. Аналогичным образом поступают в случаях, когда в массиве данных три и более совпадающих значений. Если, например, совпадает три значения, то средний ранг рассчитывается как среднее арифметическое трёх рангов, *образующих «связку»*. В данном примере получим таблицу:

Студент	Иван	Артём	Ольга	Мария	Пётр	Анна
l_i, ранг по X	6	4,5	2	4,5	1	3
k_i, ранг по Y	4	3	6	5	2	1

Далее вычисление коэффициента ранговой корреляции производится по уже известному вам алгоритму[1].

Некоторые пояснения к формуле. Коснёмся небольшого сюжета о том, почему формула Спирмена выглядит именно так $\left(r_{\text{Спирмен}} = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)} \right)$ и как она «работает». Если представить себе,

[1] Отметим, что если в выборке присутствует много связных (парных) рангов, то в формуле для вычисления коэффициента корреляции Спирмена дополнительно используется некоторая ***поправка на связные ранги***, которая в данном учебнике не рассматривается. Эту поправку можно опустить в том случае, если доля (удельный вес) объектов со связными рангами невелика по сравнению с общим объёмом выборки. Поэтому в задачах поправкой на связные ранги можно пренебречь и использовать обычную формулу.

что между ранжировками существует полное соответствие, то без труда можно сказать, что сумма квадратов разностей рангов $\sum_{i=1}^{n} d_i^2$ будет равна нулю. Тогда дробь $\frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}$ также превратится в нуль, и коэффициент корреляции будет равен «1». Это и будет означать, что между рангами наблюдается полное соответствие (положительная согласованность). И, наоборот: если между ранжировками имеется полное *не*соответствие (они противоположны), то $\sum_{i=1}^{n} d_i^2 = \frac{n(n^2-1)}{3}$.

Для пояснений рассмотрим небольшой пример.

Пример 3.3. «Про формулу». Пусть в выборке имеются n объектов. Каждый из объектов характеризуется сочетанием (l_i, ранг по X; k_i, ранг по Y). Также пусть между значениями X и Y наблюдается абсолютная (полная) противоположность — ранги по двум переменным полностью противоположны друг другу. Тогда имеем:

Объект	№ 1	№ 2	№ 3	...	№ n
l_i, ранг по X	1	2	3	...	n
k_i, ранг по Y	n	$n-1$	$n-2$	...	1
d_i, разность рангов	$(n-1)$	$(n-3)$	$(n-5)$	...	$-(n-1)$
d_i^2, квадрат разности рангов	$(n-1)^2$	$(n-3)^2$	$(n-5)^2$	...	$(n-1)^2$

В нижней строке таблицы видим последовательность, состоящую из значений: $(n-1)^2$, $(n-3)^2$, $(n-5)^2$, ..., $(n-3)^2$, $(n-1)^2$. Математикам известно, чему равна сумма значений такого вида последовательностей. Она определяется как $\frac{n(n^2-1)}{3}$. Давайте убедимся в этом на таком примере: пусть у нас имеются всего три объекта. Ранги для них выглядят следующим образом:

Объект	№1	№2	№3	Всего
l_i, ранг по X	1	2	3	
k_i, ранг по Y	3	2	1	
d_i, разность рангов	2	0	–2	
d_i^2, квадрат разности рангов	4	0	4	8

Значение «8» можно получить и по формуле: $\frac{n(n^2-1)}{3}=\frac{3(3^2-1)}{3}$.

Знание о том, как можно вычислить сумму квадратов разностей для случая, когда две ранжировки полностью противоположны друг другу, и использовал Ч. Э. Спирмен при проектировании коэффициента ранговой корреляции и его «подгонки» под значения коэффициента корреляции Пирсона — от «–1» до «+1». Число «6» в числителе было добавлено Спирменом для того, чтобы формула «давала ответ: –1» в ситуации, когда наблюдается абсолютная противоположность между рангами значений по X и Y. Действительно, подставляя $\frac{n(n^2-1)}{3}$ вместо $\sum_{i=1}^{n} d_i^2$, получим в формуле:

$$r_{\text{Спирмен}} = 1 - \frac{6\frac{n(n^2-1)}{3}}{n(n^2-1)} = 1 - 2 = -1.$$

Резюмируя, опишем последовательность действий при вычислении коэффициента корреляции Спирмена:

- каждому значению признака следует противопоставить порядковый номер (ранг) — по убыванию или возрастанию;
- рассчитать для каждой пары значений свою разность двух рангов — по переменной X и переменной Y;
- каждую вычисленную разность рангов возвести в квадрат, чтобы «избавиться» от знака «+» или «–», и найти сумму этих квадратов;
- рассчитать коэффициент корреляции рангов по указанной формуле и сделать содержательный вывод о полученном значении.

ЗАДАЧИ ДЛЯ САМОСТОЯТЕЛЬНОГО РЕШЕНИЯ

Задача 1. «Шкалы»[1].

Для вопросов из анкет социологического опроса укажите тип шкалы, используемый при измерении признаков.

1. «Какие мысли и ассоциации приходят Вам в голову, когда Вы слышите слово «мода»? (выберите один вариант ответа)»:

1) вкус, чувство прекрасного;

2) современность, соответствие духу времени;

3) всемирно известные модные бренды, имена «законодателей моды»;

[1] В этом задании приводятся краткие фрагменты анкет из академических исследований студентов II курса ф-та социологии НИУ ВШЭ, 2010—2011 уч. г.: Алипов Дмитрий, Горелова Ольга, Подстрешная Евгения (проект «*Социальные функции потребления брендов модной одежды* (*на примере опроса мужчин среднего возраста*)»), Гостева Анна, Маркова Ольга (проект «*Представление о моде как элементе повседневной жизни*»). Научный руководитель — А. В. Пашкевич.

4) звезды шоу-бизнеса, «иконы стиля»;

5) телепередачи, журналы, посвящённые моде;

6) магазины, реклама;

7) другое (укажите, пожалуйста, что именно): _____.

2. «Где Вы, как правило, покупаете одежду для себя (укажите категорию магазина)? (выберите один вариант ответа)»:

1) в бутиках дизайнеров с мировым именем (Gianfranco Ferre, Dolce & Gabbana, Giorgio Armani, Christian Dior и пр.);

2) в магазинах модной одежды категории «премиум» (Lady & Gentlemen CITY, Massimo Dutti, Kenzo, Missoni, Zac Possen, Tommy Hilfiger и пр.);

3) в магазинах модной одежды среднеценового сегмента (Mango, Topshop/Topman, Zara, H & M и пр.);

4) в магазинах модной одежды по доступной цене (Incity, Sela, New Yorker и пр.);

5) в торговых точках, предлагающих одежду по демократичным ценам (рыночные торговые ряды, сток-центры, палатки и пр.);

6) в магазинах для представителей определённых субкультур;

7) другое (укажите, пожалуйста, что именно): _____.

3. «Стремитесь ли Вы к тому, чтобы Ваша одежда соответствовала модным тенденциям, или нет? (выберите один вариант ответа)»:

1) нет, не стремлюсь;

2) скорее не стремлюсь;

3) скорее стремлюсь;

4) да, стремлюсь.

4. «Если бы Вам предложили работу в модной индустрии, согласились бы Вы или нет? (выберите один вариант ответа)»:

1) нет, не согласился (-лась) бы;

2) да, согласился (-лась) бы.

5. «Скажите, пожалуйста, в какой мере каждое из утверждений соответствует Вам (оцените, пожалуйста, каждое суждение по шкале от 1 до 5, обведя подходящую цифру кружком)»:

№ п/п	Суждение	Полностью соответствует	Во многом соответствует	В равной степени и соответствует, и не соответствует	Мало соответствует	Совсем не соответствует
1	С помощью одежды, которая есть в моём гардеробе, я демонстрирую свой внутренний мир, тот образ, который я придумал (-а)	5	4	3	2	1

Окончание табл.

№ п/п	Суждение	Полностью соответствует	Во многом соответствует	В равной степени и соответствует, и не соответствует	Мало соответствует	Совсем не соответствует
2	С помощью различных предметов одежды я стремлюсь «примерить» различные образы, чтобы найти себя	5	4	3	2	1
3	Моя одежда выражает мою яркость и индивидуальность, помогает мне выделиться из толпы и быть непохожим на других	5	4	3	2	1
4	Я ношу вещи, которые актуальны среди светских персон, звёзд «шоу-бизнеса»	5	4	3	2	1
5	Покупка дорогой одежды известных производителей помогает мне казаться немного богаче в материальном отношении	5	4	3	2	1
6	Я хожу по магазинам модной одежды для того, чтобы снять напряжение после трудного дня, расслабиться, отдохнуть	5	4	3	2	1
7	Поход по магазинам помогает мне отвлечься от повседневных проблем, снять стресс, «убежать от реальности»	5	4	3	2	1
8	Я обновляю гардероб, потому что люблю всё новое, разнообразное	5	4	3	2	1

6. «Как часто Вы обновляете (дополняете) свой гардероб? (выберите один вариант ответа)»:

1) почти каждую неделю;

2) несколько раз в месяц;
3) один раз в месяц;
4) несколько раз за полгода;
5) один раз в полгода и реже;
6) другое (укажите, пожалуйста, что именно): _____.

7. «Пожалуйста, укажите, какую примерно сумму денег Вы обычно тратите в течение месяца на покупку одежды для себя?»: _____ руб.

8. «Что является источником Вашего дохода? (укажите все подходящие варианты ответа)»:
1) самостоятельный заработок;
2) помощь родителей и/или других родственников;
3) государственные выплаты, пособия;
4) помощь друзей;
5) другое (укажите, пожалуйста, что именно): _____.

9. «Имеете ли Вы высшее образование?»:
1) да;
2) нет.

10. «Укажите, пожалуйста, Ваш возраст: _____ лет (пожалуйста, укажите число полных лет)».

11. «Какую должность в настоящее время Вы занимаете в организации, в которой работаете? (выберите один вариант ответа)»:
1) руководитель высшего звена;
2) руководитель среднего звена;
3) специалист с высшим образованием;
4) служащий без высшего образования;
5) квалифицированный рабочий;
6) другое (укажите, пожалуйста, что именно): _____.

12. «Укажите, пожалуйста, какой примерно среднемесячной суммой денег Вы располагаете (ваш личный среднемесячный доход)», руб.:
1) менее 5 000;
2) 5000—10 000;
3) 10 001—15 000;
4) 15 001—30 000;
5) 30 001—60 000;
6) 60 001—90 000;
7) более 90 000.

Задача 2. «Без названия». Вычислите коэффициент корреляции Пирсона между двумя признаками *X* и *Y*:

X	2	6	7	5	9	7
Y	9	3	1	5	0	0

Задача 3. «Физика и химия». Десять школьников написали две экзаменационные работы — по физике и химии. Каждая работа включала 20 заданий. Для каждого учащегося зафиксировано число верно выполненных заданий по двум предметам:

Школьник	Миша	Петя	Оля	Рома	Ира	Нина	Катя	Лиза	Ваня	Коля
X: число верно выполненных заданий по физике	10	5	3	17	15	12	10	20	14	14
Y: число верно выполненных заданий по химии	18	12	20	10	15	15	20	5	8	7

Вычислите коэффициент линейной корреляции (Пирсона) между числом верно выполненных заданий по физике и числом верно выполненных заданий по химии. В какой мере они коррелируют?

Задача 4. «Разводы и среднедушевые доходы населения». Рассмотрим некоторые данные официальной статистики для 14 регионов Приволжского федерального округа РФ[1]. Имеются сведения о соотношении браков и разводов и информация о среднедушевых доходах за 2010 г. Первая переменная показывает число разводов, приходящихся на 1 000 зарегистрированных браков, вторая — среднедушевой доход, руб. Данные сведены в таблицу:

Регион РФ в составе ПФО	Число разводов на 1 000 браков	Среднедушевой доход, руб.
Республика Башкортостан	477	17 677
Республика Марий Эл	493	10 195
Республика Мордовия	507	11 055
Республика Татарстан	444	18 158
Удмуртская Республика	428	12 423
Чувашская Республика	463	10 885
Пермский край	459	19 422
Кировская обл.	523	13 385
Нижегородская обл.	587	16 358
Оренбургская обл.	527	13 398
Пензенская обл.	574	12 700
Самарская обл.	563	20 279

[1] Данные Федеральной службы государственной статистики РФ [http://www.gks.ru/bgd/regl/b11_14p/Main.htm] (проверено 15.06.2012).

Окончание табл.

Регион РФ в составе ПФО	Число разводов на 1 000 браков	Среднедушевой доход, руб.
Саратовская обл.	566	11 961
Ульяновская обл.	549	12 905

Можно ли считать две эти социальные характеристики коррелированными? Рассчитайте коэффициент линейной корреляции Пирсона. Найдите коэффициент ранговой корреляции Спирмена.

Задача 5. «Расходы на рекламу». Корпорацией А* денежные средства на рекламу продукции тратятся, а интегральная прибыль не увеличивается. Имеется выборка, представленная данными за шесть отчётных периодов:

Отчётный период	1	2	3	4	5	6
Затраты на рекламу, усл. ед.	220	170	200	270	240	120
Прибыль, усл. ед.	105	140	110	90	95	160

Есть ли линейная зависимость между двумя величинами? Насколько эта связь существенна?

Задача 6. «Миграция и уровень жизни». Рассмотрим некоторые данные официальной статистики для 18 регионов Центрального федерального округа РФ[1]. Имеются сведения о миграционном приросте в регионы России и данные о среднедушевых доходах за 2010 г. Миграционный прирост — это прирост населения за счёт мигрантов (на 10 тыс. чел. населения в регионе). Знак «—» означает убыль. Среднедушевой доход измеряется в руб. Данные сведены в таблицу:

Регион РФ в составе ЦФО	Миграционный прирост на 10 тыс. чел. населения	Среднедушевой доход, руб.
Белгородская обл.	39	16 839
Брянская обл.	–19	13 298
Владимирская обл.	5	12 424
Воронежская обл.	31	13 580
Ивановская обл.	5	10 980
Калужская обл.	6	15 342
Костромская обл.	–18	12 656
Курская обл.	7	14 694

[1] Данные Федеральной службы государственной статистики РФ [http://www.gks.ru/bgd/regl/b11_14p/Main.htm] (проверено 15.06.2012).

Окончание табл.

Регион РФ в составе ЦФО	Миграционный прирост на 10 тыс. чел. населения	Среднедушевой доход, руб.
Липецкая обл.	11	15 804
Московская обл.	108	22 324
Орловская обл.	–9	13 017
Рязанская обл.	2	13 663
Смоленская обл.	–20	14 770
Тамбовская обл.	–8	13 592
Тверская обл.	–3	13 925
Тульская обл.	3	15 358
Ярославская обл.	13	14 548
г. Москва	75	43 876

Можно ли считать две эти социальные характеристики коррелированными? Рассчитайте коэффициент линейной корреляции Пирсона. Найдите коэффициент ранговой корреляции Спирмена.

Задача 7. «Задачи двух типов». Семи учащимся начальных классов средней общеобразовательной школы (взята случайная выборка учащихся), были предложены задания, направленные на измерение скорости решения заданий по математике двух типов — задания типа «А» и типа «Б». Исследователями измерялось среднее время решения заданий (время фиксировалось в секундах). Исследователей интересует, существует ли взаимосвязь между временем решения этих двух типов задач. Переменная X обозначает среднее время решения задач первого типа, а переменная Y — среднее время решения заданий второго типа. Вычислите выборочный коэффициент линейной корреляции. Интерпретируйте полученный ответ.

Школьник	Маша	Митя	Катя	Коля	Варя	Ваня	Ася
X	20	30	30	45	30	35	40
Y	15	10	20	30	30	30	20

Задача 8. «Рождаемость и соотношение городского и сельского населения». Рассмотрим некоторые данные официальной статистики для 12 регионов Сибирского федерального округа РФ[1]. Имеются сведения о рождаемости и доле сельского населения в общей численности населения ре-

[1] Данные Федеральной службы государственной статистики РФ [http://www.gks.ru/bgd/regl/b11_14p/Main.htm] (проверено 15.06.2012).

гиона в 2010 г. Рождаемость измеряется как число родившихся на 1000 человек населения. Вторая переменная показывает удельный вес сельского населения в общей численности населения региона (%). Данные представлены в таблице:

Регион РФ в составе СФО	Рождаемость (число родившихся на 1 000 чел. населения)	Удельный вес сельского населения, %
Республика Алтай	20,5	72,4
Республика Бурятия	17,0	41,6
Республика Тыва	26,9	46,9
Республика Хакасия	15,1	32,7
Алтайский край	12,7	45,3
Забайкальский край	15,9	34,1
Красноярский край	13,6	23,7
Иркутская обл.	15,2	20,5
Кемеровская обл.	13,2	14,6
Новосибирская обл.	13,2	22,7
Омская обл.	13,2	28,5
Томская обл.	13,1	29,8

Можно ли считать две эти социальные характеристики коррелированными? Рассчитайте коэффициент линейной корреляции Пирсона. Найдите коэффициент ранговой корреляции Спирмена.

Задача 9. «Оценки и посещаемость занятий». Приведены результаты (фрагмент) успеваемости десяти студентов-социологов по предмету «Теория вероятностей и математическая статистика» в сессию 2012 г. и число занятий, которые они посетили (%).

Студент	1	2	3	4	5	6	7	8	9	10
Оценка по предмету по 10-балльной шкале	8	4	7	3	10	7	2	9	5	6
Число посещённых занятий, %	90	60	90	30	100	100	10	70	60	80

Можно ли считать оценку, полученную студентом по математической статистике, и число посещённых им занятий (%) взаимосвязанными? За-

дачу решите с помощью коэффициента ранговой корреляции Спирмена. Интерпретируйте полученный ответ.

Задача 10. «Социальная реклама». В ходе разработки проекта социальной рекламы о детях были приглашены эксперты для оценки восьми подготовленных видеороликов, предназначенных для транслирования по центральному телевидению. Оценка роликов осуществлялась по 7-балльной порядковой шкале (где 1 — «ролик совершенно не решает поставленные в проекте задачи», 7 — «ролик блестяще раскрывает идею проекта»). В результате экспертного опроса были получены следующие данные.

Телеролик	Эксперт 1	Эксперт 2
А	6	4
Б	3	2
В	3	3
Г	6	5
Д	7	7
Е	4	1
Ж	7	5
З	5	7

На основе данных, приведенных в таблице, сделайте выводы относительно согласованности (совпадения, взаимосвязи) мнений двух экспертов — в какой мере они единодушны? Используйте коэффициент ранговой корреляции Спирмена. При расчётах обратите внимание на наличие рангов, образующих «связки».

Задача 11. «Практиканты». Руководитель одного из отделов компании А* проводит аттестацию практикантов, подводя итоги их работы. В качестве критериев выбраны: 1) результат теста на выявление теоретических (академических) знаний молодого специалиста, 2) практические результаты, достигнутые за время производственной практики (оценка, выставленная на основе отзывов коллег о качестве исполненных обязанностей). В таблице результат тестирования указан как сумма баллов (использовалась 100-балльная интервальная/количественная шкала). Переменная «отзывы» измерялась по порядковой 10-балльной шкале. И в том, и в другом случае бóльший балл соответствует более высокому профессионализму стажёров.

Практикант	Дмитрий	Иван	Ирина	Павел	Алла	Ксения	Сергей	Михаил
Результат теста, балл	52	81	74	100	43	78	65	67
Оценка, балл	7	9	6	8	5	7	10	4

Определите коэффициент корреляции Спирмена. Что можно сказать о связи между этими двумя характеристиками? При расчётах обратите внимание на наличие рангов-«связок».

Задача 12. «Научные статьи». Два рецензента независимо друг от друга оценивают качество шести статей, присланных в редакцию научного журнала. Используется 15-балльная шкала. Бóльший балл соответствует более высокому качеству публикации. Результаты экспертизы выглядят следующим образом:

Статья	Оценка	
	Рецензент 1	Рецензент 2
№ 1	7	5
№ 2	15	15
№ 3	9	7
№ 4	12	14
№ 5	14	11
№ 6	10	9

В какой мере эксперты единодушны в своих оценках? Задачу решите с помощью коэффициента ранговой корреляции Спирмена. Полученный ответ интерпретируйте.

Глава 4

НОРМАЛЬНОЕ РАСПРЕДЕЛЕНИЕ И СТАНДАРТНОЕ НОРМАЛЬНОЕ Z-РАСПРЕДЕЛЕНИЕ

В этой главе вспомним основные черты (свойства) такого распределения случайных величин, которое принято характеризовать как *нормальное* распределение. Подробно нормальный закон распределения изучается студентами в курсах «Теория вероятностей» и большинство основных положений должно быть уже знакомо читателям. Напомним принципиально важные идеи и приведём ряд примеров, подчеркнув значимость (и необходимость) процедуры z-стандартизации, определяющей переход от произвольного нормального распределения к стандартному нормальному распределению. Изучение материала, представленного в данной главе, позволит в дальнейшем приступить к реализации проверки статистических гипотез и построению доверительных интервалов для разных параметров, которые рассчитываются в ходе исследований, и будут рассмотрены в двух последующих, наиболее важных, главах учебника.

§ 4.1. Нормальное распределение (распределение Гаусса)

Нормальный закон распределения был наиболее полно описан в XIX в. Карлом Фридрихом Гауссом (1777—1855) — немецким математиком и астрономом. Содержание закона может быть продемонстрировано рис. 4.1, на котором представлены различные плотности распределения вероятностей.

Плотность нормального распределения описывается выражением

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

Площадь под графиком плотности распределения показывает вероятность того, что случайная величина X примет значение, ле-

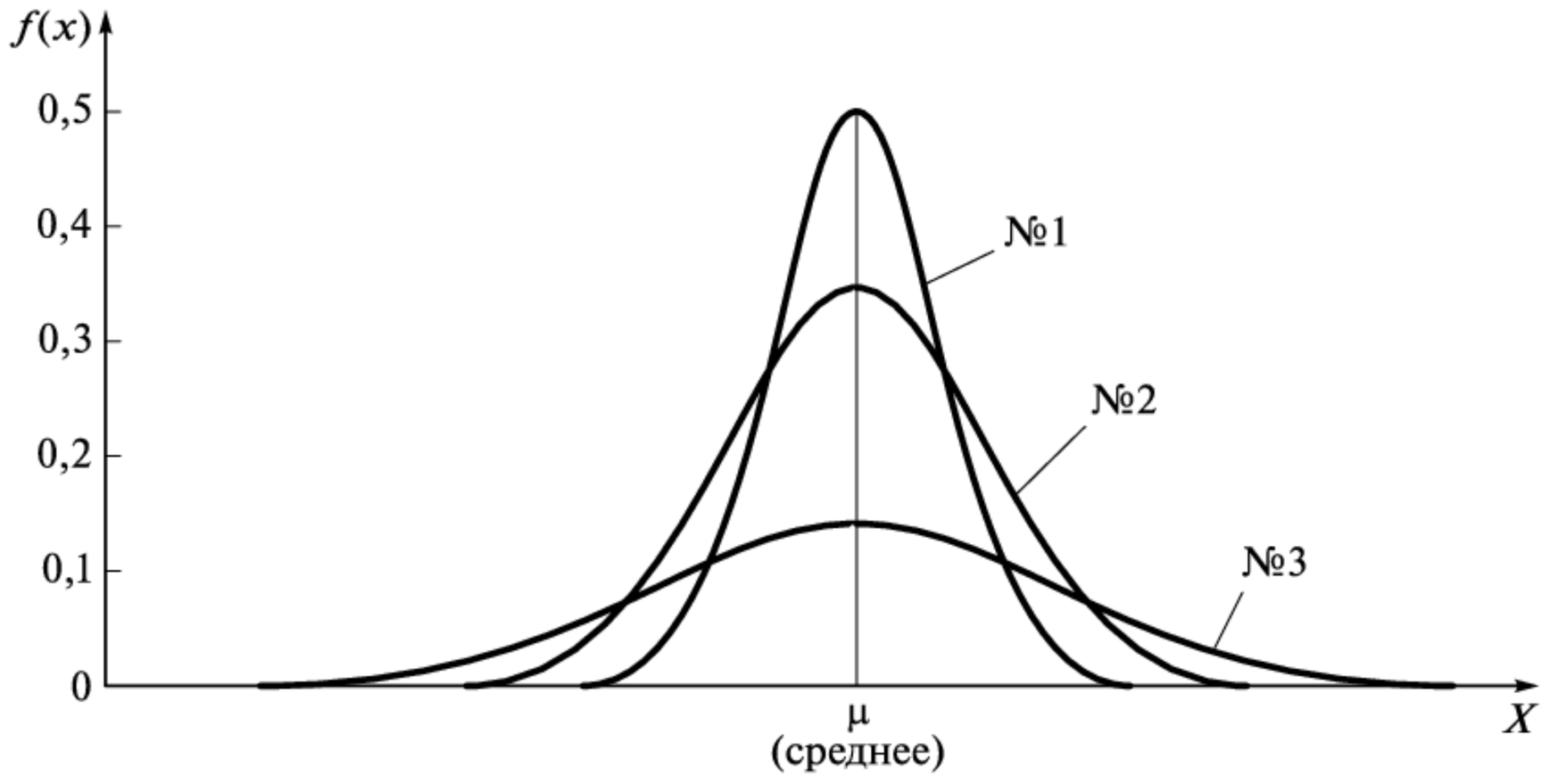


Рис. 4.1. Схематичный внешний вид нормального распределения

жащее в определённом интервале (на определённом отрезке, участке). Два ключевых параметра — *среднее* (μ) и *стандартное отклонение* (σ) — *полностью* определяют *положение* и *вид* кривой нормального распределения: среднее указывает на центр распределения и положение кривой *относительно среднего* значения *X*; стандартное отклонение определяет *форму* «*крыльев*» распределения — пологие (как бы прижаты к оси *X*) или крутые (вытянутые, заострённые). *Чем больше* стандартное отклонение, *тем более* «распластанным» («размазанным») по оси значений *X* будет распределение, так как вариация значений высока (см. схематичный вид функции *f*(*x*) для распределения № 3 на рис. 4.1). *Чем меньше* стандартное отклонение, *тем менее* разбросаны значения *X*, они более сконцентрированы вокруг среднего, а значит, и крылья распределения становятся «круче», заострённее. Это свидетельствует о том, что в рассматриваемой совокупности данных наблюдается высокая однородность (см. схематичный вид функции плотности *f*(*x*) для распределения № 1 на рис. 4.1). Нормальное распределение является математической *моделью*, позволяющей описывать распределения многих признаков, характеризующих какие-либо природные явления и/или социальные факты. Так, нормальным распределением описываются (моделируются) многие явления и социальные факты, поддающиеся числовому выражению (измерению).

Напомним, что площадь под кривой плотности распределения (по определению плотности распределения) равна 1 (или 100 %). Вероятность того, что *X* попадёт в тот или иной конкретный диапазон значений, распределяется между *всеми потенциально возможными значениями X* и охватывает их. Вероятность того, что величина (измеренная по интервальной шкале или шкале отношений) примет значения, бóльшие или равные *a*, но при этом меньшие или равные

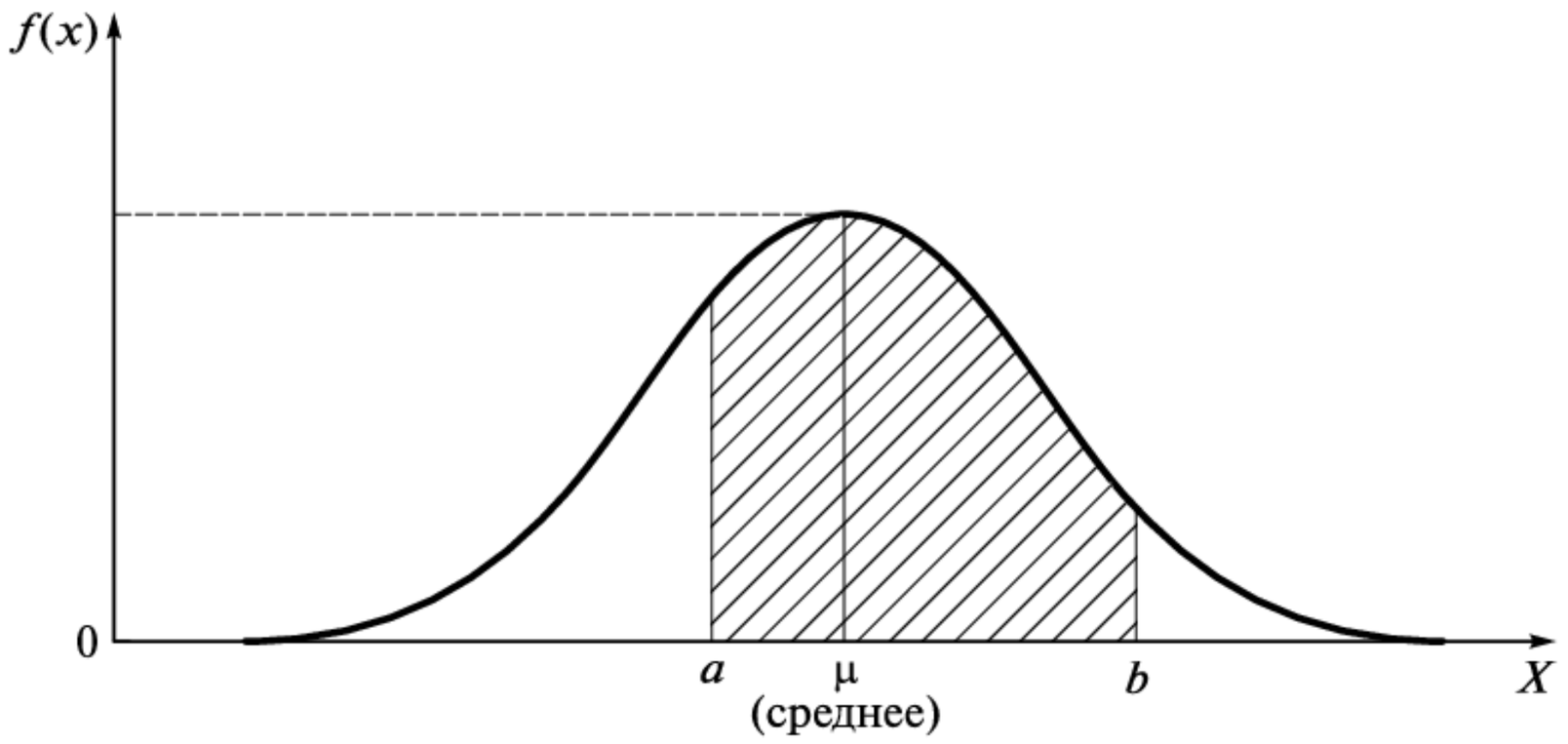


Рис. 4.2. Нормальное распределение и поиск вероятности на отрезке [a; b]

b, будет соответствовать площади под графиком кривой нормального распределения $f(x)$ на отрезке [a; b] (см. заштрихованный фрагмент рис. 4.2).

Абсолютно очевидно, что ***чем ближе*** значения X лежат относительно среднего (центра), *тем выше вероятность* их появления и тем ***более типичными (характе́рными)*** являются эти значения. И, наоборот, ***чем дальше*** от среднего располагаются значения X, тем ***более нетипичными (редкими, аномальными)*** они будут, а вероятность их встречаемости будет *невелика*. Для вычисления искомой вероятности того, что X примет значения, принадлежащие интервалу [a;b], достаточно найти соответствующий интеграл:

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

где μ и σ — конкретные числа. При этом обратите внимание, что в своих исследованиях мы имеем дело с самыми разными числовыми величинами, измеренными в различных единицах измерения (проценты, килограммы, рубли и т. д.). А это означает, что всякий раз, когда исследователю требовалось бы найти вероятность конкретного события (скажем, вероятность того, что X примет значения, принадлежащие интервалу [a; b]), то ***всякий раз*** приходилось бы вычислять значения интеграла. Для того чтобы искать ответ на поставленный вопрос (задачу) более рациональным способом, в математике существует ***процедура стандартизации*** (*унификации*) величин и переход *от произвольной* единицы измерения *к универсальной* (*единообразной*) с заданным диапазоном значений — преимущественно от «–3» до «+3» (вспомните так называемое «правило трёх сигм»). А вместе с ней и переход от нормального распределения к *стандартному* нормальному распределению.

§ 4.2. Процедура стандартизации. Стандартное нормальное *z*-распределение

Процедура стандартизации величин (признаков) напрямую связана с естественным желанием исследователей выразить их исходные значения ***в единицах стандартного отклонения***, поскольку это позволит дать ***относительную*** оценку того, *как слабо* или *как сильно* отклоняется каждое конкретное значение исследуемой совокупности от среднего по этой совокупности. Начнём с небольшого примера.

Пример 4.1. «Опрос автомобилистов». Представьте, что в опросе общественного мнения автомобилистов принял участие некоторый человек по имени Иван. Пусть в анкете содержались разные вопросы, два из которых касались величины его ***среднемесячных расходов на содержание личного автомобиля*** *X* (бензин, платные парковки, автосервис, покупка аксессуаров для автомобиля и пр.) и величины ***среднемесячных затрат времени на занятия спортом*** *Y* (утренняя зарядка, велосипедные прогулки, посещение бассейна, тренажёрного зала и пр.). Допустим, что и признак *X*, и признак *Y* подчиняются нормальному распределению (что вполне разумно). Предположим, что ответы Ивана были таковы: 5 920 руб. (на первый вопрос) и 14,5 ч (на второй вопрос). Мы хотим сопоставить эти значения между собой. Однако сопоставить «в лоб» их нельзя. Каждый признак измерялся в своих единицах — в рублях (руб.), с одной стороны, и часах (ч) — с другой. В этой ситуации ***применяется процедура стандартизации***, позволяющая привести ответы на разные вопросы (с использованием разных шкал и разных единиц измерения) к *одному знаменателю*, к так называемым ***безусловным единицам измерения***. Процедура стандартизации осмыслена только для величин (признаков), имеющих нормальное распределение, а не для любых распределений. Мы *совершенно не нарушим* структуру исходных данных для каждой переменной, если осуществим такую операцию как стандартизация и сформируем (т. е. сконструируем) *новую* переменную типа: $Z = \frac{X - \mu}{\sigma}$. Из *каждого* индивидуального возможного значения *X* вычтем среднее μ (т. е. от*центрируем*), а затем *выразим* полученное расхождение от этого среднего в *единицах стандартного отклонения*, *разделив на* σ.

Полученные таким образом индивидуальные значения будут выражены в отклонениях исходных («сырых») значений от среднего значения так, что измерены в единицах сигмы — как мало́ или как велико́ *отличие* (несовпадение) конкретного значения исследуемой величины от среднего (т. е. *типичного*) значения в совокупности, в расчёте на 1 (*одну*) сигму.

Возвращаясь к примеру, предположим, что *средняя величина ежемесячных расходов на содержание автомобиля среди автомобилистов* составляет 5 240 руб., а стандартное отклонение — 1 100 руб. Средняя величина ежемесячных затрат времени на занятия спортом среди автомобилистов составляет 17,3 ч, а стандартное отклонение — 3,1 ч. Тогда после применения процедуры стандартизации ответов Ивана получим для него следующие z-оценки (*z-scores*):

$$z_{\text{Автомобиль_Иван}} = \frac{5\,920 - 5\,240}{1\,100} = \frac{680}{1\,100} = +0{,}62 \approx +0{,}6$$

— величина среднемесячных расходов на содержание автомобиля у него *чуть выше*, чем среднее по совокупности (в сравнении с другими автомобилистами);

$$z_{\text{Спорт_Иван}} = \frac{14{,}5 - 17{,}3}{3{,}1} = \frac{-2{,}8}{3{,}1} = -0{,}903 \approx -0{,}9$$

— величина среднемесячных затрат времени на занятия спортом у него *несколько меньше*, чем среднее по совокупности (в сравнении с другими автомобилистами).

По сути, z-оценки в случае нормального распределения исследуемого признака выступают ***эталоном «типичности»*** или же, напротив, ***«уникальности»*** индивидуальных характеристик респондента (если речь идёт об опросах общественного мнения) в сравнении с остальными людьми. В какой мере его (или её) характеристики ***такие же, как у большинства других? Или они ощутимо другие?*** Совершенно очевидно, *чем сильнее* z-значение отклоняется от нуля в положительную или отрицательную сторону, *тем более уникальным* является индивидуальное значение x_i. А *чем ближе* z-значение *к нулю* (центру совокупности), *тем в большей степени* характеристика объекта является *типичной* (распространённой, характе́рной, т. е. такой же, как и у многих других в исследуемой совокупности). И в этом случае x_i незначительно отличается от среднего μ.

Всё это позволяет заключить: ***если исходные значения переменной, имеющей нормальное распределение, можно преобразовать в z-оценки, являющиеся относительной (безусловной) мерой отклонения от среднего, то тогда и нормальное распределение с параметрами*** $(\mu;\sigma)$ ***можно превратить в стандартное нормальное распределение с параметрами*** $(\mu = 0;\ \sigma = 1)$.

Существует специальная статистическая таблица, воспользовавшись которой можно без особого труда найти вероятность искомых событий, а именно: определить вероятность того, что некоторая величина X примет значения, бóльшие или равные a, но при этом меньшие или равные b. Искомая площадь под кривой плотности нормального распределения $f(x)$ на отрезке $[a;b]$ будет полностью

соответствовать (совпадать) с площадью под кривой стандартного нормального распределения $f(z)$ на отрезке $[z_a; z_b]$.

В табл. 4.1 (которая приводится далее и устроена просто) по строкам указаны z-значения с точностью до десятых долей, полученные при стандартизации переменных — эти значения охватывают диапазон *от 0,0 до 3,0* (*как три сигмы*). Столбцы же содержат *сотые* доли, которые могут потребоваться, если z-значение не целое и содержит в себе сотую часть после запятой. В этом случае нужно добавить («доклеить») к z-значению, указанному слева, этот «хвостик» после запятой. На пересечении соответствующих строки и столбца указана площадь под кривой стандартного нормального распределения на отрезке *от нуля до z*. Это и будет значением вероятности того, что Z принадлежит отрезку $[0; z]$. Например, пусть нас интересует, чему равно значение вероятности $P(0 \le Z \le 1{,}43)$ — площадь под кривой на отрезке от 0 до +1,43. По строкам слева находим значение 1,4 и добавляем к нему значение в соответствующем столбце («0,03») — так как до 1,43 нам не хватает 0,03. На пересечении строки и столбца получим искомый ответ: 0,4236 (ответ выделен в ячейке таблицы).

Обратите внимание, что в табл. 4.1 не приводятся отрицательные значения z. Почему? Значение площади между нулём и некоторым отрицательным значением будет соответствовать площади между нулем и положительным значением заданной величины. Ведь стандартное нормальное распределение абсолютно *симметрично* относительно нуля! Левая часть распределения (содержащая отрицательные значения величины) — зеркальное отражение правой (содержащей положительные значения величины). Так, площадь под кривой от –2,35 до 0 будет такой же, как и площадь под кривой от 0 до +2,35 и составит 0,4906 (ответ выделен в ячейке таблицы).

Пример 4.2. «Цветы». Муж каждую неделю дарит жене букет цветов. Средняя стоимость букета составляет 400 руб., а стандартное отклонение равно 100 руб. Какова вероятность того, что стоимость очередного букета составит не менее 500 руб.? (Если считать, что сумма денежных средств, затрачиваемых мужем на букет, описывается моделью нормального распределения с указанными параметрами).

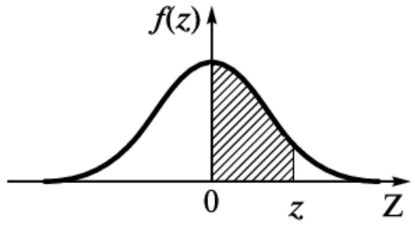
Решение. Нам следует вычислить вероятность того, что X (стоимость букета цветов) будет лежать в интервале $[500; +\infty]$, где 500 — нижняя граница интервала (a), $+\infty$ — верхняя граница интервала (b). Тогда:

$$P(500 \le X \le +\infty) = P\left(\frac{a-400}{100} \le Z \le \frac{b-400}{100}\right) =$$

$$= P\left(\frac{500-400}{100} \le Z \le \frac{+\infty-400}{100}\right) = P(1 \le Z \le +\infty) =$$

$$= P(-\infty \le Z \le +\infty) - P(-\infty \le Z \le 1).$$

Таблица 4.1. **Таблица стандартного нормального z-распределения (интеграл от 0 до *z*)**



Z	**0,00**	**0,01**	**0,02**	**0,03**	**0,04**	**0,05**	**0,06**	**0,07**	**0,08**	**0,09**
0,0	0,0000	0,0040	0,0080	0,0120	0,0160	0,0199	0,0239	0,0279	0,0319	0,0359
0,1	0,0398	0,0438	0,0478	0,0517	0,0557	0,0596	0,0636	0,0675	0,0714	0,0753
0,2	0,0793	0,0832	0,0871	0,0910	0,0948	0,0987	0,1026	0,1064	0,1103	0,1141
0,3	0,1179	0,1217	0,1255	0,1293	0,1331	0,1368	0,1406	0,1443	0,1480	0,1517
0,4	0,1554	0,1591	0,1628	0,1664	0,1700	0,1736	0,1772	0,1808	0,1844	0,1879
0,5	0,1915	0,1950	0,1985	0,2019	0,2054	0,2088	0,2123	0,2157	0,2190	0,2224
0,6	0,2257	0,2291	0,2324	0,2357	0,2389	0,2422	0,2454	0,2486	0,2517	0,2549
0,7	0,2580	0,2611	0,2642	0,2673	0,2704	0,2734	0,2764	0,2794	0,2823	0,2852
0,8	0,2881	0,2910	0,2939	0,2967	0,2995	0,3023	0,3051	0,3078	0,3106	0,3133
0,9	0,3159	0,3186	0,3212	0,3238	0,3264	0,3289	0,3315	0,3340	0,3365	0,3389
1,0	0,3413	0,3438	0,3461	0,3485	0,3508	0,3531	0,3554	0,3577	0,3599	0,3621
1,1	0,3643	0,3665	0,3686	0,3708	0,3729	0,3749	0,3770	0,3790	0,3810	0,3830

1,2	0,3849	0,3869	0,3888	0,3907	0,3925	0,3944	0,3962	0,3980	0,3997	0,4015
1,3	0,4032	0,4049	0,4066	0,4082	0,4099	0,4115	0,4131	0,4147	0,4162	0,4177
1,4	0,4192	0,4207	0,4222	0,4236	0,4251	0,4265	0,4279	0,4292	0,4306	0,4319
1,5	0,4332	0,4345	0,4357	0,4370	0,4382	0,4394	0,4406	0,4418	0,4429	0,4441
1,6	0,4452	0,4463	0,4474	0,4484	0,4495	0,4505	0,4515	0,4525	0,4535	0,4545
1,7	0,4554	0,4564	0,4573	0,4582	0,4591	0,4599	0,4608	0,4616	0,4625	0,4633
1,8	0,4641	0,4649	0,4656	0,4664	0,4671	0,4678	0,4686	0,4693	0,4699	0,4706
1,9	0,4713	0,4719	0,4726	0,4732	0,4738	0,4744	0,4750	0,4756	0,4761	0,4767
2,0	0,4772	0,4778	0,4783	0,4788	0,4793	0,4798	0,4803	0,4808	0,4812	0,4817
2,1	0,4821	0,4826	0,4830	0,4834	0,4838	0,4842	0,4846	0,4850	0,4854	0,4857
2,2	0,4861	0,4864	0,4868	0,4871	0,4875	0,4878	0,4881	0,4884	0,4887	0,4890
2,3	0,4893	0,4896	0,4898	0,4901	0,4904	0,4906	0,4909	0,4911	0,4913	0,4916
2,4	0,4918	0,4920	0,4922	0,4925	0,4927	0,4929	0,4931	0,4932	0,4934	0,4936
2,5	0,4938	0,4940	0,4941	0,4943	0,4945	0,4946	0,4948	0,4949	0,4951	0,4952
2,6	0,4953	0,4955	0,4956	0,4957	0,4959	0,4960	0,4961	0,4962	0,4963	0,4964
2,7	0,4965	0,4966	0,4967	0,4968	0,4969	0,4970	0,4971	0,4972	0,4973	0,4974
2,8	0,4974	0,4975	0,4976	0,4977	0,4977	0,4978	0,4979	0,4979	0,4980	0,4981
2,9	0,4981	0,4982	0,4982	0,4983	0,4984	0,4984	0,4985	0,4985	0,4986	0,4986
3,0	0,4987	0,4987	0,4987	0,4988	0,4988	0,4989	0,4989	0,4989	0,4990	0,4990

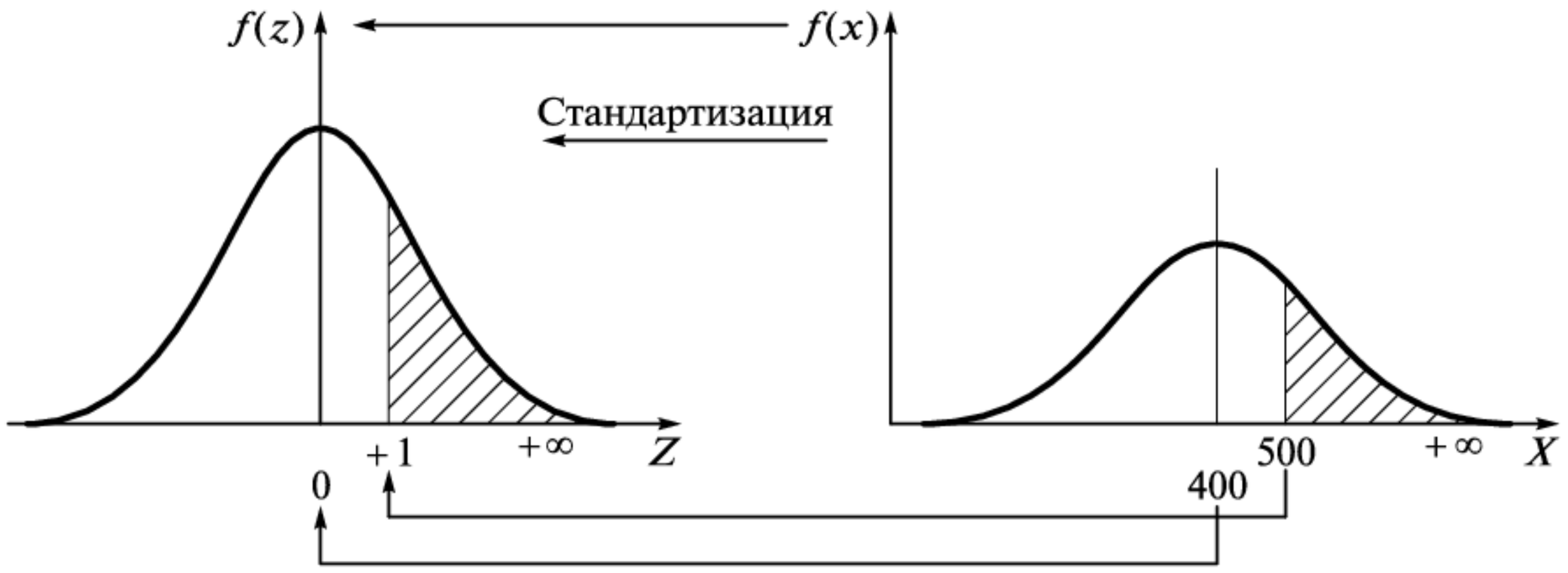


Рис. 4.3. Процедура стандартизации (см. пример 4.2 «Цветы»)

(Эту запись нужно прочитать так: «Из площади под кривой стандартного нормального распределения на отрезке от $-\infty$ до $+\infty$ следует вычесть площадь под кривой стандартного нормального распределения на отрезке от $-\infty$ до $+1$.) Результатом станет искомая площадь под кривой стандартного нормального распределения на требуемом отрезке $[z_a;z_b] = [1;+\infty]$» (рис. 4.3 графически поясняет смысл расчётов).

Запись $P(-\infty \le Z \le +\infty) - P(-\infty \le Z \le 1)$ можно преобразовать следующим образом: $P(-\infty \le Z \le +\infty) - P(-\infty \le Z \le 1) = (0{,}5 + P(0 \le Z \le +\infty)) - (0{,}5 + P(0 \le Z \le 1)) = P(0 \le Z \le +\infty) - P(0 \le Z \le 1)$, где $P(0 \le Z \le +\infty)$ — площадь под кривой стандартного нормального распределения от нуля до $+\infty$, а $P(0 \le Z \le 1)$ — площадь под кривой стандартного нормального распределения на отрезке от нуля до $+1$, а «0,5 (или 50 %)» — площадь под кривой на отрезке от $-\infty$ до 0. Значения «0,5» и «0,5» взаимно уничтожаются в выражении, если раскрыть скобки.

Выражение $\boldsymbol{P(0 \le Z \le +\infty) - P(0 \le Z \le 1)}$ имеет тот же смысл, что и графическое представление искомой вероятности, которой соответствует заштрихованная часть графика (рис. 4.3). И действительно, для того чтобы найти эту вероятность, нужно *из бóльшей* площади под кривой (от 0 до $+\infty$) вычесть *меньшую* площадь под кривой (от 0 до $+1$). Найденная разница и будет соответствовать площади под кривой распределения на отрезке от $+1$ до $+\infty$.

По таблице стандартного нормального z-распределения находим, что $P(0 \le Z \le +\infty) - P(0 \le Z \le 1) = 0{,}5 - 0{,}34 = 0{,}16$ (т. е. 16 %). Следовательно, вероятность того, что муж купит жене очередной букет, стоимость которого составит *не менее* 500 руб., равна 0,16. Соответствующие величины площади под кривой на отрезках от 0 до $+\infty$ и от 0 до $+1$ выделены в табл. 4.2.

Пример 4.3. «Без названия». Рассматривается некоторая случайная величина, которая может принимать значения в диапазоне от 0 до 100%. При этом случайная величина имеет нормальное распределение с математическим ожиданием (средним) 40 % и стандартным отклонением 10 %. С какой вероятностью следует ожидать того, что

Таблица 4.2. **Таблица стандартного нормального z-распределения (интеграл от 0 до *z*)**

Z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,0000	0,0040	0,0080	0,0120	0,0160	0,0199	0,0239	0,0279	0,0319	0,0359
0,1	0,0398	0,0438	0,0478	0,0517	0,0557	0,0596	0,0636	0,0675	0,0714	0,0753
0,2	0,0793	0,0832	0,0871	0,0910	0,0948	0,0987	0,1026	0,1064	0,1103	0,1141
0,3	0,1179	0,1217	0,1255	0,1293	0,1331	0,1368	0,1406	0,1443	0,1480	0,1517
0,4	0,1554	0,1591	0,1628	0,1664	0,1700	0,1736	0,1772	0,1808	0,1844	0,1879
0,5	0,1915	0,1950	0,1985	0,2019	0,2054	0,2088	0,2123	0,2157	0,2190	0,2224
0,6	0,2257	0,2291	0,2324	0,2357	0,2389	0,2422	0,2454	0,2486	0,2517	0,2549
0,7	0,2580	0,2611	0,2642	0,2673	0,2704	0,2734	0,2764	0,2794	0,2823	0,2852
0,8	0,2881	0,2910	0,2939	0,2967	0,2995	0,3023	0,3051	0,3078	0,3106	0,3133
0,9	0,3159	0,3186	0,3212	0,3238	0,3264	0,3289	0,3315	0,3340	0,3365	0,3389
1,0	0,3413	0,3438	0,3461	0,3485	0,3508	0,3531	0,3554	0,3577	0,3599	0,3621
1,1	0,3643	0,3665	0,3686	0,3708	0,3729	0,3749	0,3770	0,3790	0,3810	0,3830
1,2	0,3849	0,3869	0,3888	0,3907	0,3925	0,3944	0,3962	0,3980	0,3997	0,4015
1,3	0,4032	0,4049	0,4066	0,4082	0,4099	0,4115	0,4131	0,4147	0,4162	0,4177
1,4	0,4192	0,4207	0,4222	0,4236	0,4251	0,4265	0,4279	0,4292	0,4306	0,4319
1,5	0,4332	0,4345	0,4357	0,4370	0,4382	0,4394	0,4406	0,4418	0,4429	0,4441

Окончание табл. 4.2

Z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
1,6	0,4452	0,4463	0,4474	0,4484	0,4495	0,4505	0,4515	0,4525	0,4535	0,4545
1,7	0,4554	0,4564	0,4573	0,4582	0,4591	0,4599	0,4608	0,4616	0,4625	0,4633
1,8	0,4641	0,4649	0,4656	0,4664	0,4671	0,4678	0,4686	0,4693	0,4699	0,4706
1,9	0,4713	0,4719	0,4726	0,4732	0,4738	0,4744	0,4750	0,4756	0,4761	0,4767
2,0	0,4772	0,4778	0,4783	0,4788	0,4793	0,4798	0,4803	0,4808	0,4812	0,4817
2,1	0,4821	0,4826	0,4830	0,4834	0,4838	0,4842	0,4846	0,4850	0,4854	0,4857
2,2	0,4861	0,4864	0,4868	0,4871	0,4875	0,4878	0,4881	0,4884	0,4887	0,4890
2,3	0,4893	0,4896	0,4898	0,4901	0,4904	0,4906	0,4909	0,4911	0,4913	0,4916
2,4	0,4918	0,4920	0,4922	0,4925	0,4927	0,4929	0,4931	0,4932	0,4934	0,4936
2,5	0,4938	0,4940	0,4941	0,4943	0,4945	0,4946	0,4948	0,4949	0,4951	0,4952
2,6	0,4953	0,4955	0,4956	0,4957	0,4959	0,4960	0,4961	0,4962	0,4963	0,4964
2,7	0,4965	0,4966	0,4967	0,4968	0,4969	0,4970	0,4971	0,4972	0,4973	0,4974
2,8	0,4974	0,4975	0,4976	0,4977	0,4977	0,4978	0,4979	0,4979	0,4980	0,4981
2,9	0,4981	0,4982	0,4982	0,4983	0,4984	0,4984	0,4985	0,4985	0,4986	0,4986
3,0	0,4987	0,4987	0,4987	0,4988	0,4988	0,4989	0,4989	0,4989	0,4990	0,4990

в некотором случайном эксперименте (опыте) данная величина примет значение, не меньшее чем 35 %?

Решение. В этой задаче предстоит вычислить вероятность того, что X (%) будет лежать в интервале [35;100], где 35 % — нижняя граница интервала (a), 100 % — верхняя граница интервала (b).

$$P(35 \le X \le 100) = P\left(\frac{a-40}{10} \le Z \le \frac{b-40}{10}\right) =$$

$$= P\left(\frac{35-40}{10} \le Z \le \frac{100-40}{10}\right) = P(-0{,}5 \le Z \le +6) =$$

$$= P(-\infty \le Z \le +6) - P(-\infty \le Z \le -0{,}5) =$$

$$= P(0 \le Z \le +6) + P(0 \le Z \le +0{,}5).$$

(Эту запись можно прочитать так: «К площади под кривой стандартного нормального распределения на отрезке от 0 до +6 следует ***прибавить*** площадь под кривой стандартного нормального распределения на отрезке от 0 до +0,5».) Результатом станет искомая площадь под кривой стандартного нормального распределения на требуемом отрезке $[z_a;z_b] = [-0{,}5;+6]$. Выражение $P(0 \le Z \le +6) + P(0 \le Z \le +0{,}5)$ имеет тот же содержательный смысл, что и графическое представление искомой вероятности, которой соответствует заштрихованная часть графика (рис. 4.4).

И действительно, для того чтобы найти искомую вероятность интересующего события, нужно сложить *две* площади под кривой (на отрезке от 0 до +6 и на отрезке от 0 до +0,5), что и будет целиком соответствовать площади под кривой распределения на отрезке от –0,5 до +6. По таблице стандартного нормального z-распределения находим, что $P(0 \le Z \le +6) + P(0 \le Z \le +0{,}5) = 0{,}5 + 0{,}19 = 0{,}69$ (т. е. 69 %). Итак, вероятность того, что в некотором случайном эксперименте (опыте) рассматриваемая величина примет значение, не меньшее чем 35 %, равна 0,69. Соответствующие величины площади под кривой на отрезках от 0 до +6 и от 0 до +0,5 выделены в табл. 4.3.

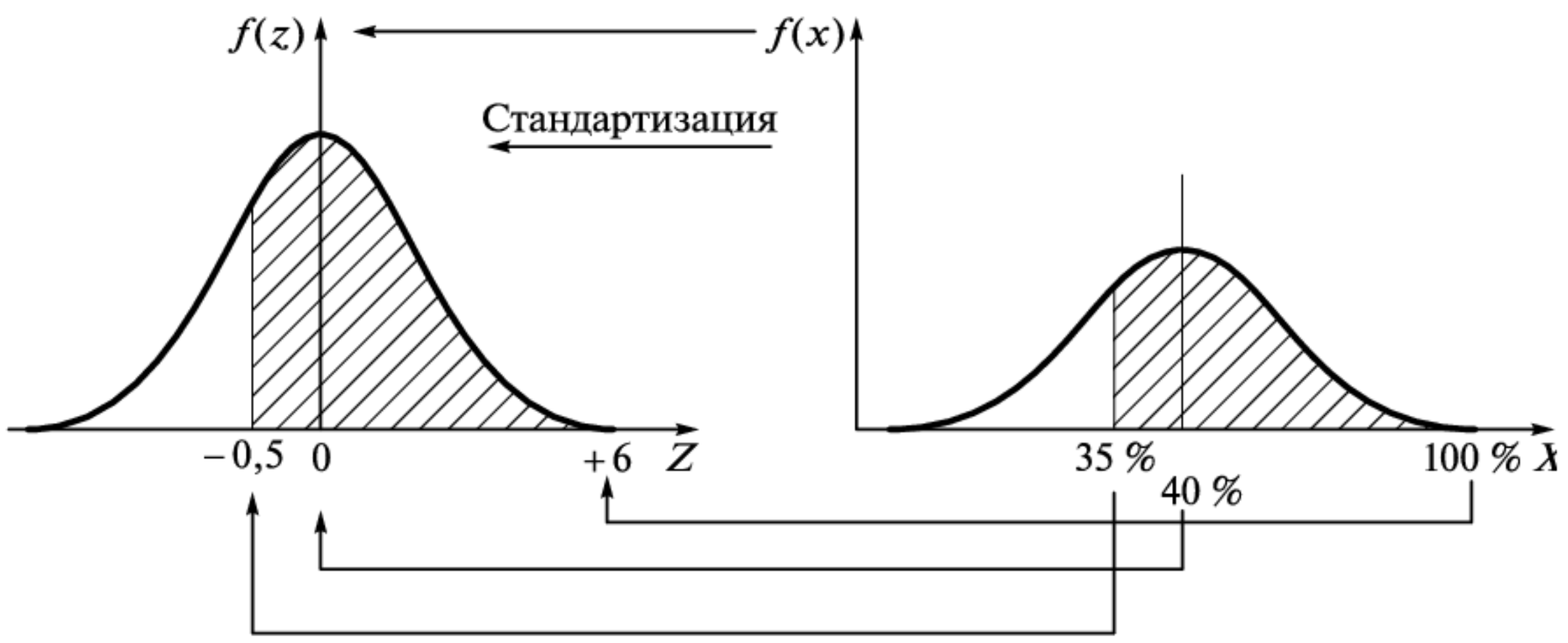


Рис. 4.4. Процедура стандартизации (см. пример 4.3 «Без названия»)

Таблица 4.3. **Таблица стандартного нормального z-распределения (интеграл от 0 до *z*)**

Z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,0000	0,0040	0,0080	0,0120	0,0160	0,0199	0,0239	0,0279	0,0319	0,0359
0,1	0,0398	0,0438	0,0478	0,0517	0,0557	0,0596	0,0636	0,0675	0,0714	0,0753
0,2	0,0793	0,0832	0,0871	0,0910	0,0948	0,0987	0,1026	0,1064	0,1103	0,1141
0,3	0,1179	0,1217	0,1255	0,1293	0,1331	0,1368	0,1406	0,1443	0,1480	0,1517
0,4	0,1554	0,1591	0,1628	0,1664	0,1700	0,1736	0,1772	0,1808	0,1844	0,1879
0,5	0,1915	0,1950	0,1985	0,2019	0,2054	0,2088	0,2123	0,2157	0,2190	0,2224
0,6	0,2257	0,2291	0,2324	0,2357	0,2389	0,2422	0,2454	0,2486	0,2517	0,2549
0,7	0,2580	0,2611	0,2642	0,2673	0,2704	0,2734	0,2764	0,2794	0,2823	0,2852
0,8	0,2881	0,2910	0,2939	0,2967	0,2995	0,3023	0,3051	0,3078	0,3106	0,3133
0,9	0,3159	0,3186	0,3212	0,3238	0,3264	0,3289	0,3315	0,3340	0,3365	0,3389
1,0	0,3413	0,3438	0,3461	0,3485	0,3508	0,3531	0,3554	0,3577	0,3599	0,3621
1,1	0,3643	0,3665	0,3686	0,3708	0,3729	0,3749	0,3770	0,3790	0,3810	0,3830
1,2	0,3849	0,3869	0,3888	0,3907	0,3925	0,3944	0,3962	0,3980	0,3997	0,4015
1,3	0,4032	0,4049	0,4066	0,4082	0,4099	0,4115	0,4131	0,4147	0,4162	0,4177
1,4	0,4192	0,4207	0,4222	0,4236	0,4251	0,4265	0,4279	0,4292	0,4306	0,4319

1,5	0,4332	0,4345	0,4357	0,4370	0,4382	0,4394	0,4406	0,4418	0,4429	0,4441
1,6	0,4452	0,4463	0,4474	0,4484	0,4495	0,4505	0,4515	0,4525	0,4535	0,4545
1,7	0,4554	0,4564	0,4573	0,4582	0,4591	0,4599	0,4608	0,4616	0,4625	0,4633
1,8	0,4641	0,4649	0,4656	0,4664	0,4671	0,4678	0,4686	0,4693	0,4699	0,4706
1,9	0,4713	0,4719	0,4726	0,4732	0,4738	0,4744	0,4750	0,4756	0,4761	0,4767
2,0	0,4772	0,4778	0,4783	0,4788	0,4793	0,4798	0,4803	0,4808	0,4812	0,4817
2,1	0,4821	0,4826	0,4830	0,4834	0,4838	0,4842	0,4846	0,4850	0,4854	0,4857
2,2	0,4861	0,4864	0,4868	0,4871	0,4875	0,4878	0,4881	0,4884	0,4887	0,4890
2,3	0,4893	0,4896	0,4898	0,4901	0,4904	0,4906	0,4909	0,4911	0,4913	0,4916
2,4	0,4918	0,4920	0,4922	0,4925	0,4927	0,4929	0,4931	0,4932	0,4934	0,4936
2,5	0,4938	0,4940	0,4941	0,4943	0,4945	0,4946	0,4948	0,4949	0,4951	0,4952
2,6	0,4953	0,4955	0,4956	0,4957	0,4959	0,4960	0,4961	0,4962	0,4963	0,4964
2,7	0,4965	0,4966	0,4967	0,4968	0,4969	0,4970	0,4971	0,4972	0,4973	0,4974
2,8	0,4974	0,4975	0,4976	0,4977	0,4977	0,4978	0,4979	0,4979	0,4980	0,4981
2,9	0,4981	0,4982	0,4982	0,4983	0,4984	0,4984	0,4985	0,4985	0,4986	0,4986
3,0	0,4987	0,4987	0,4987	0,4988	0,4988	0,4989	0,4989	0,4989	0,4990	0,4990

ЗАДАЧИ ДЛЯ САМОСТОЯТЕЛЬНОГО РЕШЕНИЯ

Задача 1. «Антикварный магазин». Средняя сумма чека в некотором антикварном магазине составляет 1070 усл. ед., а стандартное отклонение — 320 усл. ед. Какова вероятность того, что очередной покупатель совершит покупку на сумму от 600 до 900 усл. ед.? (если считать, что сумма покупки — случайная величина, которая описывается нормальным законом распределения с указанными параметрами).

Задача 2. «Мороженое». В тёплое время года на покупку мороженого семья тратит, в среднем, 520 руб. в неделю. Стандартное отклонение составляет 107 руб. Какова вероятность того, что в следующую неделю семья потратит на мороженое от 240 до 300 руб.? (если считать, что расходы семьи на мороженое описываются нормальным законом распределения с указанными параметрами). Схематично отразите искомую вероятность на графике нормального распределения и на графике стандартного нормального распределения.

Задача 3. «Кофейня». Средняя сумма чека в некоторой кофейне составляет 810 руб., а стандартное отклонение — 230 руб. Какова вероятность, что очередной посетитель кофейни сделает заказ на сумму от 700 до 1100 руб.? (если считать, что сумма чека — случайная величина, которая описывается нормальным законом распределения с указанными параметрами). Схематично отразите искомую вероятность на графике нормального распределения и на графике стандартного нормального распределения.

Задача 4. «Грусть». Один человек иногда грустит. Среднее число дней в месяц, когда он грустный, равно 12, а стандартное отклонение составляет 3 дня. Какова вероятность того, что в следующем месяце он будет грустить не более 8 дней? (если считать, что число «грустных» дней — случайная величина, которая описывается нормальным законом распределения с указанными параметрами). Схематично отразите искомую вероятность на графике нормального распределения и на графике стандартного нормального распределения.

Задача 5. «Салоны красоты». Средняя стоимость стрижки в парикмахерских города составляет 560 руб., а стандартное отклонение 150 руб. Какова вероятность, что в наугад выбранной парикмахерской стоимость стрижки составит от 300 до 400 руб.? (если считать, что стоимость стрижки — случайная величина, которая описывается нормальным законом распределения с указанными параметрами). Схематично отразите искомую вероятность на графике нормального распределения и на графике стандартного нормального распределения.

Задача 6. «Баранки». На покупку баранок семья тратит, в среднем, 200 руб. в неделю. Стандартное отклонение составляет 48 руб. Какова вероятность, что в следующую неделю семья потратит на баранки от 100 до 150 руб.? (если считать, что расходы семьи на баранки описываются нормальным законом распределения с указанными параметрами). Схематично отразите искомую вероятность на графике нормального распределения и на графике стандартного нормального распределения.

Задача 7. «Сертификат». Допустим, что по опыту предыдущих лет подмечено, что оценки за экзамен, который сдают люди на получение международного сертификата, подтверждающего средний уровень владения иностранным языком, подчиняются нормальному закону распределения с математическим ожиданием 56 баллов и дисперсией 64 (шкала 100-балльная). Чему равна вероятность того, что случайно выбранный человек, который придёт сдавать экзамен, получит оценку между 40 и 50 баллами? Схематично отразите искомую вероятность на графике нормального распределения и на графике стандартного нормального распределения.

Задача 8. «Покупка лекарств». На покупку лекарств в осенне-зимний период среднестатистическая семья тратит, в среднем, 370 усл. ед. в неделю на одного человека. Стандартное отклонение составляет 95 усл. ед. Какова вероятность того, что случайно выбранная семья тратит на лекарства от 600 до 800 усл. ед. на человека? (если считать, что расходы на лекарства описываются нормальным законом распределения с указанными параметрами). Схематично отразите искомую вероятность на графике нормального распределения и на графике стандартного нормального распределения.

Задача 9. «Погодные условия». Предположим, что по данным метеостанции летняя среднесуточная температура в регионе А* составляет 22 °С, а стандартное отклонение равно 3,4 °С. Какова вероятность того, что в случайно выбранный летний день эта температура составит от 28 до 32 °С? (если считать, что температура описывается нормальным законом распределения с указанными параметрами). Схематично отразите искомую вероятность на графике нормального распределения и на графике стандартного нормального распределения.

Глава 5

ТОЧЕЧНОЕ ОЦЕНИВАНИЕ ПАРАМЕТРОВ И ПОСТРОЕНИЕ ДОВЕРИТЕЛЬНЫХ ИНТЕРВАЛОВ

В этой главе мы приступаем к изучению правил (принципов) переноса и обобщения результатов исследования с выборки на генеральную совокупность. Научимся строить доверительные интервалы — такие интервалы значений, в пределах которых находятся *истинные* значения интересующих нас параметров. Рассмотрим, каким образом вычисляется объём выборки, требуемый для эмпирического исследования, обсудим понятие несмещённой выборочной оценки параметра и, в частности, понятие несмещённой выборочной оценки дисперсии.

§ 5.1. Выборка и понятие несмещённой выборочной оценки. Несмещённая оценка дисперсии

Расчёт *несмещённой* выборочной дисперсии отличается от той, которая уже известна вам (и подробно рассмотрена в гл. 2, § 2.3), лишь в *одном* формальном аспекте. Несмещённая дисперсия вычисляется как сумма квадратов отклонения значений выборки от *среднего* (как и в случае расчёта «обычной» дисперсии), но делённая *не на* «*n*» (!), а *на* «$(\boldsymbol{n-1})$». Это *принципиально важный* момент для математической статистики. Выясним, почему это так и какая *содержательная* модель, какие соображения лежат в основе вычислений.

Выборочная дисперсия является ***оценкой*** истинного значения дисперсии (в генеральной совокупности), которую мы не знаем. Оценками также являются и среднее арифметическое, и выборочная медиана — они суть оценки математического ожидания (т. е. истинного значения среднего в генеральной совокупности). Оценка (как функция от наблюдений) является величиной случайной. Это значит, что как любая случайная величина, она может принимать множество

значений, и на этом множестве тем или иным образом может быть задано распределение вероятностей.

Если оценка является дискретной случайной величиной, то такое распределение может быть задано с помощью ряда распределения (с указанием частот или относительных частот). Так, доля «успехов» в n испытаниях Бернулли может рассматриваться как оценка вероятности «успеха» (большое количество примеров приведено в гл. 1, § 1.2).

В случае если оценка является непрерывной случайной величиной, то её распределение вероятностей задаётся функцией плотности распределения этой случайной величины (например, среднее арифметическое как оценка математического ожидания в нормальной совокупности).

Заметим, что для оценивания одного и того же параметра можно использовать различные оценки (т. е. различные функции от наблюдений).

Так, в качестве оценок математического ожидания совокупности могут быть взяты или среднее арифметическое, или медиана, или взвешенное среднее или даже первое значение выборки. Важно понимать, что каждая из этих оценок может быть выбрана, но дисперсия у неё может быть очень большая или маленькая, а потому одни оценки более предпочтительны, другие — менее. Например, в практике исследований медиана реже используется как оценка математического ожидания, потому что медиана нормальной совокупности имеет бóльшую дисперсию, чем, скажем, среднее арифметическое. Хотя и медиана, и среднее арифметическое являются оценками несмещёнными.

Поскольку оценка является случайной величиной, то, конечно, хотелось бы, чтобы математическое ожидание этой оценки (т. е. её среднее значение) совпадало с тем значением параметра, которое мы оцениваем с помощью этой оценки.

Оцениваемые в ходе исследования статистические параметры (как правило, это параметры распределений в генеральной совокупности) являются величинами ***неслучайными***, они являются ***константами*** (потому что истинные, неколеблющиеся). Поэтому ***оценка практически никогда (!) не равна оцениваемому параметру***, ***а отклоняется от него в ту или иную сторону*** (*бóльшую или меньшую*). В этой ситуации, конечно, нам бы хотелось, чтобы *в среднем* оценка совпадала с оцениваемым параметром генеральной совокупности. В свою очередь это и будет означать, что ***её математическое ожидание совпадает с оцениваемым параметром генеральной совокупности***. Именно ***такие оценки называются несмещёнными***.

Пример 5.1. «Без названия». Выборочная доля является несмещённой оценкой вероятности. Покажем это. Для этого возьмём классический пример из теории вероятностей. Рассмотрим математическое ожидание выборочной доли:

$$E\left(\frac{S_n}{n}\right)=\frac{1}{n}E(S_n)=\frac{1}{n}np=p,$$

где E — математическое ожидание, т. е. среднее значение доли; S_n — число «успехов» в серии из n независимых испытаний Бернулли с вероятностью успеха p.

Так, если мы стремимся узнать, какова вероятность того, что житель страны, случайно выбранный «в толпе», скажет, что любит зиму, то *несмещённой оценкой этой вероятности будет именно выборочная доля*, равная отношению числа человек в выборке, которые ответят, что любят зиму, к общему числу n опрошенных респондентов. Поэтому если в ходе выборочного исследования было опрошено 810 человек и при этом 348 сказали, что любят зиму, то выборочная доля $\frac{348}{810}\approx 0{,}43$ является несмещённой оценкой истинного значения вероятности того, что жителю страны нравится зима.

Пример 5.2. «Без названия-2». Покажем, что среднее арифметическое $\bar{X}$ является несмещённой оценкой математического ожидания. Рассмотрим случайную величину $X: x_1, x_2, x_3, ..., x_n$, имеющую нормальное распределение с параметрами μ (математическим ожиданием), и σ^2 (истинным значением дисперсии в генеральной совокупности), т. е. $\sim N(\mu;\sigma^2)$. Тогда

$$E(\bar{X})=E\left(\frac{x_1+x_2+x_3+...+x_n}{n}\right)=\frac{1}{n}E(x_1+x_2+x_3+...+x_n)=$$

$$=\frac{E(x_1)+E(x_2)+E(x_3)+...+E(x_n)}{n}=\frac{\mu+\mu+\mu+...+\mu}{n}=\frac{\mu n}{n}=\mu.$$

Иногда на практике мы сталкиваемся со ***смещёнными оценками***. Например, в быту постоянное смещение дают нам постоянно спешащие или, напротив, постоянно отстающие часы: далеко не каждые часы являются абсолютно точным инструментом, показывающим время. Такие часы дают завышенную или заниженную оценку времени. И, определяя по ним время, мы получаем ***лишь примерную (приблизительную) оценку времени***. Приведём другие примеры.

Пример 5.3. «Без названия-3». Пусть требуется оценить дисперсию нормальной совокупности с помощью выборочной дисперсии, в которой сумма квадратов отклонений ***делится на n***: $s^2=\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n}$. Выборочная оценка дисперсии является смещённой оценкой параметра σ^2. Покажем это:

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

$$=E\left[\frac{(x_1-\mu-\bar{x}+\mu)^2+(x_2-\mu-\bar{x}+\mu)^2+\ldots+(x_n-\mu-\bar{x}+\mu)^2}{n}\right]=$$

$$=E\left[\frac{((x_1-\mu)-(\bar{x}-\mu))^2+((x_2-\mu)-(\bar{x}-\mu))^2+\ldots+((x_n-\mu)-(\bar{x}-\mu))^2}{n}\right]=$$

$$=\{\textit{можно сказать: }((x_1-\mu)-(\bar{x}-\mu))^2=(a-b)^2=a^2-2ab+b^2\}=$$

$$=E\left[\frac{1}{n}((x_1-\mu)^2-2(x_1-\mu)(\bar{x}-\mu)+(\bar{x}-\mu)^2+(x_2-\mu)^2-2(x_2-\mu)(\bar{x}-\mu)+\right.$$

$$\left.+(\bar{x}-\mu)^2+\ldots+(x_n-\mu)^2-2(x_n-\mu)(\bar{x}-\mu)+(\bar{x}-\mu)^2)\right]=$$

$$=\{\textit{преобразуем выражение}\}=$$

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

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

$$=\frac{1}{n}\left(n\sigma^2-\frac{2n\sigma^2}{n}+n\frac{\sigma^2}{n}\right)=\frac{1}{n}(n\sigma^2-\sigma^2)=\sigma^2-\frac{\sigma^2}{n}=$$

$$=\sigma^2\left(1-\frac{1}{n}\right)=\sigma^2\left(\frac{n-1}{n}\right)\neq\sigma^2.$$

Получается, что математическое ожидание выборочной оценки дисперсии *не равно в точности* значению дисперсии в генеральной совокупности:

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

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

Получили, что несмещённая выборочная оценка дисперсии отличается от смещённой выборочной оценки дисперсии поправкой

вида $\frac{n}{n-1}$. В результате формула для вычисления *несмещённой* выборочной оценки дисперсии выглядит следующим образом (в знаменателе фигурирует не «n», а «$n-1$»):

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

Несмещённая оценка дисперсии, рассчитываемая для выборки наблюдений, когда в знаменателе появляется $(n-1)$, существенно ближе к истинному значению оцениваемого параметра — дисперсии в генеральной совокупности. Очевидным (и напрашивающимся) является вывод о том, что чем *больше размер* выборки, тем *в большей степени* соотношение $\frac{n}{n-1}$ стремится к *единице*, а значит, исчезает разница между значениями двух оценок дисперсии — выборочной смещённой и выборочной несмещённой. Они практически полностью совпадают. Действительно, когда объём выборки крайне мал (например, число респондентов в исследовании составляет всего лишь 10 человек), тогда значение дисперсии, найденной по формуле $s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}$, должно быть увеличено на 11,1 %, чтобы получить *несмещённую* оценку. Эта поправка заметна и ощутима.

Когда объём выборки составляет, например, 30 наблюдений, поправка равна $\frac{30}{30-1} = 1{,}034$, т. е. соотношение несмещённой и смещённой оценок составляет только 3,4 %. А при объёмах выборки 100 наблюдений и более, это расхождение практически полностью исчезает и не превышает $\frac{100}{100-1} = 1{,}010$, что соответствует всего лишь 1 %. Следовательно, дифференциация оценок дисперсии, рассчитанной как $s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}$ или как $s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$, отсутствует. Заметьте, что при очень большом объёме выборки, равном, к примеру, 1000, $\frac{n}{n-1} = \frac{1\,000}{1\,000-1} = 1{,}0010$, т. е. всего 0,1 %. В таком случае различия в оценках дисперсии (с делением на n или на $n-1$) вовсе не заметны. Итак, видим, что смещение «проявляет себя» на выборках мало-

го размера. Тем не менее, когда речь идёт о выборочной дисперсии, принято рассчитывать *несмещённую* оценку дисперсии. Даже если «эффект масштаба» (размер выборки, превосходящий 100 единиц) устраняет несоответствие между выборочной смещённой и выборочной несмещённой оценками дисперсии.

Пример 5.4. «Без названия-4». Выборка представлена следующими значениями X: 8, 6, 4, 2, 2, 2. Вычислите ***смещённую*** и ***несмещённую выборочные оценки*** дисперсии и стандартного отклонения.

Решение. Сначала вычисляем среднее арифметическое:

$$\bar{X} = \frac{8+6+4+2+2+2}{6} = \frac{8+6+4+2\cdot 3}{6} = 4.$$

Смещённая и несмещённая дисперсии будут соответственно равны:

$$s^2 = \frac{(8-4)^2+(6-4)^2+(4-4)^2+(2-4)^2\cdot 3}{6} = \frac{16+4+0+12}{6} = \frac{32}{6} \approx 5{,}33;$$

$$s^2 = \frac{(8-4)^2+(6-4)^2+(4-4)^2+(2-4)^2\cdot 3}{6-1} = \frac{16+4+0+12}{6-1} = \frac{32}{5} = 6{,}4.$$

Тогда смещённая и несмещённая выборочные оценки стандартного отклонения составят: $s = \sqrt{5{,}33} \approx 2{,}31$, а $s = \sqrt{6{,}4} \approx 2{,}53$.

Пример 5.5. «Время на сон». В ходе небольшого опроса, проведённого на одном из занятий по математической статистике среди студентов факультета прикладной политологии НИУ «Высшая школа экономики», участникам семинара было предложено ответить на несколько коротких вопросов. Одним из таких вопросов был следующий: «Укажите, пожалуйста, сколько часов Вы тратите на сон в будни (напишите с точностью до получаса)». Объём выборки составил 19 человек. Распределение ответов представлено ниже:

Сколько часов Вы спите?	4	5	5,5	6	7	7,5	8
Количество человек	3	3	1	4	4	2	2

Найдите выборочные оценки среднего, дисперсии (как смещённую, так и несмещённую), оцените величину смещения. Рассчитайте несмещённую выборочную оценку стандартного отклонения.

Решение. Сначала вычисляем среднее арифметическое:

$$\bar{X} = \frac{4\cdot 3+5\cdot 3+5{,}5\cdot 1+6\cdot 4+7\cdot 4+7{,}5\cdot 2+8\cdot 2}{19} \approx 6{,}1$$

— в среднем, один студент затрачивает на сон в будни 6,1 (ч).

Смещённая и несмещённая дисперсии будут соответственно равны:

$$s^2 = \frac{(4-6{,}1)^2 \cdot 3 + (5-6{,}1)^2 \cdot 3 + (5{,}5-6{,}1)^2 + (6-6{,}1)^2 \cdot 4 + \ldots}{19} \approx 1{,}7 \text{ (ч)}^2;$$

$$s^2 = \frac{(4-6{,}1)^2 \cdot 3 + (5-6{,}1)^2 \cdot 3 + (5{,}5-6{,}1)^2 + (6-6{,}1)^2 \cdot 4 + \ldots}{19-1} \approx 1{,}8 \text{ (ч)}^2.$$

Величина смещения равна $\frac{19}{19-1} = \frac{1{,}8}{1{,}7} \approx 1{,}056 \approx 5{,}6\,\%$. Смещённая выборочная оценка дисперсии даёт значение меньшее (заниженное) на 5,6 %, в сравнении с несмещённой выборочной оценкой.

Несмещённая оценка стандартного отклонения равна: $s = \sqrt{1{,}8} \approx \approx 1{,}34$(ч).

Пример 5.6. «Рост». В ходе того же опроса, проведённого среди студентов факультета прикладной политологии НИУ «Высшая школа экономики», участникам семинара было предложено ответить на следующий вопрос: «Укажите, пожалуйста, Ваш рост (напишите с точностью до сантиметра)». Объём выборки юношей составил 8 человек, девушек — 11 человек:

Юноши: 169, 173, 181, 178, 172, 178, 178, 192.

Девушки: 172, 166, 174, 164, 176, 159, 161, 176, 175, 176, 167.

Для обеих выборок найдите выборочные оценки среднего, дисперсии (как смещённую, так и несмещённую), оцените величину смещения. Рассчитайте несмещённые выборочные оценки стандартного отклонения. Определите выборочные оценки коэффициента вариации.

Решение.

Юноши	Девушки
Среднее (см): $\bar{X} \approx 177{,}6$	Среднее (см): $\bar{Y} \approx 169{,}6$
Смещённая оценка дисперсии (см^2): $s^2 \approx 43{,}2$	Смещённая оценка дисперсии (см^2): $s^2 \approx 37{,}7$
Несмещённая оценка дисперсии (см^2): $s^2 \approx 49{,}4$	Несмещённая оценка дисперсии (см^2): $s^2 \approx 41{,}5$
Величина смещения: $\frac{8}{8-1} = \frac{49{,}4}{43{,}2} \approx 1{,}14 \approx 14\,\%$	Величина смещения: $\frac{11}{11-1} = \frac{41{,}5}{37{,}7} \approx 1{,}10 \approx 10\,\%$
Коэффициент вариации: $V = \frac{\sqrt{49{,}4}}{177{,}6} 100\,\% \approx 4\,\%$	Коэффициент вариации: $V = \frac{\sqrt{41{,}5}}{169{,}6} 100\,\% \approx 4\,\%$

Вывод: Наблюдаем крайне высокую степень однородности выборки юношей по росту. Разброс значений выражен слабо. Также крайне высока степень однородности выборки девушек по росту. Разброс значений тоже выражен слабо.

Обратите внимание, что величина смещения оценки дисперсии для выборки девушек меньше, чем для выборки юношей, так как объём выборки девушек больше, чем объём выборки юношей (11 и 8 соответственно).

§ 5.2. Точечная оценка и доверительный интервал. Ошибки выборки: стандартная и предельная

Мы не раз отмечали, что разумной альтернативой сплошным исследованиям выступают выборочные исследования. *Выборочное* исследование охватывает лишь *некоторую часть* генеральной совокупности (и, как правило, весьма небольшую). По определённым правилам и в соответствии с чёткими процедурами проектируется уменьшенная копия (т. е. уменьшенная модель, микрокосм) генеральной совокупности — репрезентативная выборка, сохраняющая внутреннюю структуру и способная «нести в себе» характеристики и закономерности всей совокупности. До настоящего момента рассматривались лишь оценки, полученные для *отдельной* выборки наблюдений. Ключевая же задача статистики заключается в том, чтобы «уйти» («устраниться») от имеющихся выборочных оценок и корректно перенести результаты, полученные для *одной* выборки конкретного исследования, на *всю генеральную совокупность* (*т. е. обобщить*). Статистический анализ данных (и обобщение результатов на генеральную совокупность) должен быть осуществлён с учётом *ошибки выборки*. Как это следует понимать?

Допустим, по выборке мы получили, что среднее число конфет, съедаемых в течение недели ребёнком дошкольного возраста, составляет 12 шт. Как эта выборочная оценка среднего соотносится с истинным значением среднего, которое нам не известно? Чему равно истинное значение среднего числа конфет, съедаемых обычным (среднестатистическим) ребенком? Именно с этой целью осуществляется переход от точечной оценки к получению интервальной оценки. На языке математики этот переход будет предполагать необходимость вычисления ***интервала (диапазона), в котором с высокой вероятностью будет гарантированно содержаться истинное значение интересующего параметра*** (например, *истинное значение среднего*, *истинное значение доли*, *истинное значение коэффициента корреляции и т. д.*). ***Построить (найти) доверительный интервал означает перейти от характеристик***

выборки к характеристикам генеральной совокупности. Ключевыми для нас словами должны стать понятия «***точечная оценка***» и «***интервальная оценка***». ***Точечной оценкой*** будем называть значение случайной величины (признака), вычисленное на основе имеющейся выборки наблюдений. Это значение является конкретным, дискретным, числовым. Точечной оценкой могут выступать среднее арифметическое *выборки* ($\bar{X}$), *выборочная* оценка дисперсии (s^2), медиана *выборки* (Me), *выборочная* оценка коэффициента корреляции для рассматриваемых пар наблюдений (r) и многие другие *выборочные оценки параметров* (*т.е. значения конкретной статистики, или показателя, полученные для выборки*). Слова «выборка», «выборочный» в данном контексте крайне *принципиальны*.

Представьте, что проводится опрос населения на тему «Есть ли у вас личный автомобиль?». Пусть в ходе выборочного исследования выяснилось, что автомобиль имеется у 38 % респондентов. Предположим, что спустя несколько дней мы вновь решили повторить процедуру опроса. Есть ли у нас ожидания и уверенность в том, что получится тот же результат — 38 %? Конечно, нет: едва ли такое совпадение может иметь место. Мы понимаем, что результат (с той или иной вероятностью) будет, определённо, колебаться вокруг цифры 38%, но вряд ли абсолютно повторять её! Могут быть получены значения и 37 %, и 40 %, и 41 %, и 36 %. Доля в данном случае изменчива, результат колеблется вокруг некоторого воображаемого истинного значения. Оценённая *по выборке*, эта доля *не есть абсолютно* истинная оценка. Выборочная доля — это лишь точечная оценка истинного значения доли в генеральной совокупности, которая нам достоверно не известна.

Вместе с тем исследователю хочется судить не просто о выборочных оценках, а о том, ***каковы же значения интересующих параметров в генеральной совокупности***. ***Интервальной оценкой*** станет такая оценка, которая будет представлена в виде ***числового диапазона*** (отрезка значений), внутри которого и содержится истинное значение оцениваемого параметра, характерного для всей генеральной совокупности. Интервал, содержащий с заданной вероятностью истинное значение параметра, носит название ***доверительного интервала***, а вероятность того, что истинное значение действительно принадлежит этому интервалу (т.е. попадает в него), называется ***доверительной вероятностью***.

При построении доверительного интервала будем использовать следующую схему-«конструктор», в которую входят такие компоненты:

- ***истинное значение параметра*** (в общем виде обозначим его буквой ***L***) — *это значение параметра* ***в генеральной совокупности***. Примерами таких параметров могут выступать: математическое ожидание, значение доли в генеральной совокупности, истинное значение дисперсии в генеральной совокупности, истинное значение

коэффициента корреляции в генеральной совокупности и др. Заметьте, заранее мы ***не знаем истинного*** значения параметра ***L***, но на основе имеющейся выборочной информации пытаемся найти границы интервала, который наверняка «охватывает» его;

• ***выборочная оценка параметра*** (*обозначим буквой l*) — её мы получаем по данным ***выборки***, она ***доступна*** нашему знанию;

• ***доверительная вероятность*** (*обозначим её греческой буквой* γ — «*гамма*») показывает, с какой вероятностью истинное значение параметра наверняка (т. е. гарантированно) попадает в найденные границы интервала;

• ***предельная ошибка (погрешность) выборки*** (*обозначим её греческой буквой* ε — «*ипсилон*») задает ***ширину диапазона*** относительно ***выборочной*** оценки параметра. Эта ошибка рассчитывается по данным выборки и с учётом задаваемой доверительной вероятности.

Для того чтобы построить доверительный интервал, сначала нужно из выборочной оценки параметра *вычесть* погрешность (предельную ошибку) выборки (в результате будет получена ***левая граница*** доверительного интервала), а далее к выборочной оценке параметра *прибавить* погрешность выборки (в результате будет получена ***правая граница*** доверительного интервала):

$$l - \varepsilon \le L \le l + \varepsilon.$$

«Читать» данное неравенство следует так: ***с вероятностью*** γ ***истинное значение параметра находится (содержится) в интервале*** $[l - \varepsilon; l + \varepsilon]$. Или иными словами: ***с вероятностью*** γ ***истинное значение параметра накрывается интервалом*** $[l - \varepsilon; l + \varepsilon]$. Именно так, с помощью доверительного интервала, мы стремимся оценить, насколько выборочная оценка близка к истинному значению.

В понятии «доверительный интервал» нам пока остаются не до конца прояснёнными два аспекта — доверительная вероятность и предельная ошибка выборки. Далее разберём, что они означают с содержательной точки зрения и как «работают» в вычислениях.

Ошибка выборки. Вполне естественно сказать, что ошибка — это всегда некоторое расхождение между двумя величинами. Так, *ошибка выборки* (или ее ещё называют *ошибкой репрезентативности*) — это ***величина расхождения*** (разницы, несовпадения, несоответствия) ***между оценкой показателя***, ***рассчитанного на выборке***, ***и его истинным значением в генеральной совокупности***. Чтобы понять идею вычисления ошибки выборки следует представить себе следующий мысленный эксперимент. Вообразим, что у нас есть генеральная совокупность размером N. Мы планируем исследование на выборке объёма n. Существует C_N^n способов извлечь случайным образом из генеральной совокупности размером N выборку размера n. Например,

если генеральная совокупность насчитывает всего *четыре человека* (например, Алина, Иван, Марина, Дмитрий), а нам для исследования требуется выборка размера $n = 2$ *человека*, то определим, каким тогда будет потенциально возможное количество выборок в ходе случайного отбора респондентов: $C_{N=4}^{n=2} = \frac{4!}{2! \cdot 2!} = \frac{2! \cdot 3 \cdot 4}{2! \cdot 1 \cdot 2} = 6$ — столько разных выборок (с разным составом участников) можно получить[1]. И, действительно, возможны следующие выборки:

- Алина и Иван (1);
- Алина и Марина (2);
- Алина и Дмитрий (3);
- Иван и Марина (4);
- Иван и Дмитрий (5);
- Марина и Дмитрий (6).

Всего имеем шесть возможных комбинаций (вариантов). Вероятность извлечь любую из них составляет $\frac{1}{C_{N=4}^{n=2}} = \frac{1}{6} \approx 0{,}17$. Теперь представим, что генеральная совокупность насчитывает *100 человек*, а нам для исследования требуется выборка размером *5 человек*. *Потенциально возможное* количество выборок в ходе случайного отбора респондентов будет тогда таким: $C_{N=100}^{n=5} = \frac{100!}{5! \cdot (100-5)!} = 75\,287\,520$ — такое огромное количество разных выборок (с разным составом участников) можно получить. Они гипотетически возможны. Мы же для своего исследования чаще всего случайно извлекаем из генеральной совокупности всего лишь *одну-единственную* выборку, именно она ложится в основу сбора данных, регистрации информации. Если речь идет об опросах общественного мнения, то в выборку случайно попадают совершенно конкретные люди. И вероятность извлечения *именно этой конкретной выборки* (из огромнейшего числа потенциально возможных выборок по составу участников) равна: $\frac{1}{C_{N=100}^{n=5}} = \frac{1}{75\,287\,520}$. Эта вероятность ничтожно мала: она стремится к нулю, но, тем не менее, ненулевая. Видим, что гипотетически (т. е. теоретически) мы обладаем возможностью извлечь колоссальное число выборок, различающихся по составу участников исследования, но вероятность извлечения каждой отдельной выборки может быть крайне малой.

А теперь мысленно представим, что многократно (раз за разом) из генеральной совокупности мы извлекали выборку фиксированного объёма и «перепробовали» большое число выборок равного

---

[1] Нами рассчитано количество комбинаций (сочетаний) для заданного числа элементов.

объёма (с бóльшим числом наблюдений) на этой (одной и той же) генеральной совокупности. Пусть всякий раз среди участников исследования мы проводили один и тот же опрос общественного мнения, одни и те же измерения. Допустим, у респондентов фиксировался их рост. Итак, если в нашем распоряжении были выборки X_1, X_2,X_3, ..., X_k, где k — номер выборки, то ***для каждой выборки*** мы имели возможность рассчитать ***выборочные оценки среднего*** роста людей и получить: $\bar{X}_1$, $\bar{X}_2$,$\bar{X}_3$, ..., $\bar{X}_k$. ***Каждое из этих значений среднего арифметического является оценкой математического ожидания***. И каждая выборка имеет (несёт в себе) ***свою ошибку репрезентативности***.

Имея сведения о выборочных оценках среднего, гипотетически можно построить график плотности распределения вероятностей для этих средних значений. В случае большого количества выборочных средних (*оценок* ровно столько, сколько и выборок), это распределение — ***нормальное.*** Центром распределения является ***математическое ожидание*** — истинное значение среднего в генеральной совокупности.

Основная часть выборочных оценок среднего будет сконцентрирована вокруг центра — своего математического ожидания. Это означает, что бóльшая часть выборок, извлечённых из генеральной совокупности, даёт схожий результат — оценки среднего роста в этих выборках совпадают или незначительно отличаются, концентрация же выборочных оценок среднего в «хвостах» распределения, напротив, будет невелика. Схематично построим это распределение (где по оси абсцисс откладываются выборочные оценки среднего, а по оси ординат указана соответствующая этим оценкам плотность вероятности).

Обратите внимание, что какие-то выборочные оценки среднего могут ***далеко*** отстоять от истинного значения среднего (μ), а какие-то, наоборот, быть ***крайне близки*** к нему. На рис. 5.1 показано, что для каждой конкретной выборочной оценки среднего вероятность

Рис. 5.1. Схематичное распределение выборочных оценок среднего

быть близко расположенной к истинному значению, гораздо выше, чем, например, вероятность попасть в «хвосты» распределения.

У каждой отдельной выборки есть своя ошибка. Ошибка каждой отдельной выборки определяется тем, какова ***вариация*** (***дисперсия***) признака X в этой выборочной совокупности, *с одной стороны*, и каков ***размер этой выборки*** — *с другой*. ***Дисперсия*** *самой* ***выборочной оценки среднего*** будет равна ***отношению выборочной оценки дисперсии к объёму выборки*** (т. е. будет равна ***дисперсии*** признака X, ***уменьшенной в n раз***). Говоря другими словами, дисперсия выборочной оценки среднего есть не что иное, как величина дисперсии признака (вариации, изменчивости), ***приходящаяся на одно наблюдение в выборке*** (т. е. *в среднем*). Чем она больше, тем больше ошибка; и, соответственно, наоборот — чем она меньше, тем и ошибка меньше. Математически дисперсия среднего записывается следующим образом:

$$s^2_{\bar{x}} = \frac{s^2_x}{n}.$$

Тогда ***стандартная ошибка*** выборочной оценки среднего будет равна корню квадратному из дисперсии среднего арифметического:

$$s_{\bar{x}} = \sqrt{\frac{s^2_x}{n}} = \frac{s_x}{\sqrt{n}}.$$

Стандартная ошибка является стандартным отклонением выборочной оценки и служит показателем точности рассчитанного на выборке параметра. У стандартной ошибки среднего имеются и другие синонимичные названия (они могут встречаться в учебной литературе): ***средняя ошибка выборки***, ***средняя ошибка среднего***, ***стандартная ошибка выборки***. Все они означают одно и то же: $s_{\bar{x}} = \sqrt{\frac{s^2_x}{n}}$. Стандартная ошибка среднего является своего рода ***мерой близости*** произвольно взятой выборочной оценки среднего к истинному значению среднего (к своему математическому ожиданию). Она показывает, ***как слабо или как сильно колеблется выборочная оценка среднего вокруг себя***. ***Чем меньше*** стандартная ошибка, ***тем точнее*** наши представления об истинном значении среднего. И наоборот: ***чем больше*** стандартная ошибка, т. е. *чем больше волатильность* (изменчивость) оценки, ***тем менее точными*** могут быть наши знания и прогнозы относительно истинного значения среднего в генеральной совокупности.

Интервал $\left[\bar{X}_1 - \frac{s_x}{\sqrt{n}}; \bar{X}_1 + \frac{s_x}{\sqrt{n}} \right]$ — это своего рода «коридор», внутри которого «крутится» (т. е. колеблется) значение оценки среднего для

отдельной выборки. Из выборочной оценки среднего вычитается и к ней же прибавляется одно стандартное отклонение, характеризующее изменчивость выборочной оценки среднего, т. е. ***вычитается и прибавляется одна единица стандартного отклонения для среднего арифметического***. А согласно правилу «*трёх сигм*»[1], интервал, охватывающий значения на участке $\mu \pm \sigma$, покрывает около 68 % всех возможных значений признака.

Поскольку в реальной жизни в ходе конкретного исследования мы располагаем данными, как правило, только одной-единственной выборки, на которой и проводим сбор данных, то мы ***никогда достоверно не знаем***, близка или далека наша выборочная оценка среднего ***относительно истинного значения*** **среднег**о. Поэтому возникает вполне разумное решение: а почему бы не «накрыть» (используя образное сравнение — «как одеялом») выборочную оценку среднего так, чтобы с высокой вероятностью утверждать, что истинное значение среднего оказалось охваченным (даже не имея сведений, чему оно равно в генеральной совокупности). Так возникает понятие ***предельной ошибки*** выборки, обозначаемой буквой ε («ипсилон»):

$$\varepsilon = zs_{\bar{x}} = z\frac{s_x}{\sqrt{n}}.$$

Предельная ошибка выборки — ***максимально возможная ошибка выборки при заданной вероятности — равна z-кратному числу стандартных ошибок среднего***, т. е. это максимально возможное (предельное) *расхождение* между фактической выборочной оценкой среднего и истинным значением среднего в генеральной совокупности при заданном уровне вероятности. Обратите внимание, что от стандартной ошибки среднего предельная ошибка отличается только наличием дополнительного множителя в формуле. ***Компонентом***, ***регулирующим меру нашей уверенности в выводах*** относительно того, чему же равно истинное значение параметра, выступает ***z-значение***. Можно сказать, что z показывает, во сколько раз следует увеличить стандартную ошибку выборки (и «растянуть» границы $\left[\bar{X}_1 - \frac{s_x}{\sqrt{n}}; \bar{X}_1 + \frac{s_x}{\sqrt{n}}\right]$), чтобы получить интервал, в который с вероятностью γ *гарантированно* «укладывается» (помещается) истинное значение параметра (рис. 5.2).

Как правило, в качестве доверительной вероятности (т. е. той вероятности, с которой можно гарантированно утверждать, чему равно истинное значение параметра в генеральной совокупности) выбираются значения ***90 %***, ***95 %***, ***99 %***. ***Доверительная вероятность служит мерой надёжности выводов*** относительно истинного значения, и является степенью нашей уверенности в результатах,

[1] Так называемое правило «трёх сигм» (или правило «68—95—99»).

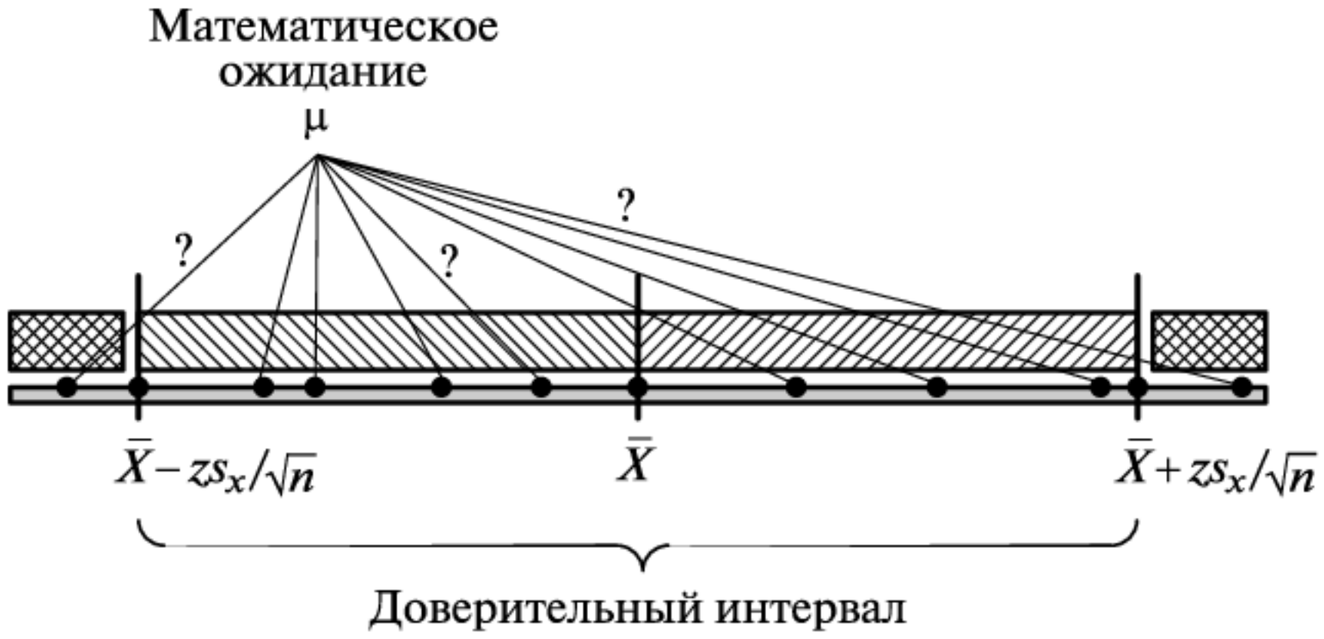


Рис. 5.2. Построение доверительного интервала для среднего

которые были получены в ходе исследований. Само истинное среднее — величина не случайная, но мы говорим о вероятности, так как случайными являются границы этого интервала: они определяются по имеющейся случайной выборке и при этом зависят от выбранного уровня доверия. Для достаточно больших выборок из нормального распределения можно указать простое правило, примерно задающее границы 95%-го доверительного интервала. Эти границы примерно рассчитываются как выборочное среднее «плюс» — «минус» два выборочных стандартных отклонения, деленных на $\sqrt{n}$ (или «плюс» — «минус» две стандартные ошибки среднего).

Если при нахождении доверительного интервала доверительная вероятность будет взята на уровне 0,95 (т.е. 95 %), то в результате мы получим возможность утверждать, что для 950 выборок из 1 000 выборочная оценка среднего будет принадлежать этому интервалу, т.е. интервал «охватит» 95 % всех потенциально возможных значений выборочной оценки среднего. Применительно к отдельной выборке ***интерпретация*** будет следующей: ***с вероятностью 95 % истинное значение среднего в генеральной совокупности лежит в диапазоне*** ±2 ***стандартные ошибки от выборочной оценки среднего***; 95% совпадений — это довольно много, а 5 % несовпадений (указывающих на вероятность ошибочности выводов) — весьма мало, поэтому пятью процентами можно, в разумном смысле слова, пренебречь.

Так, если мы говорим, что среднее число конфет, съедаемых ребёнком дошкольного возраста в течение недели, составляет 12±2,5 (конфет), то подразумеваем, что 12 — это средняя величина, полученная в выборке наблюдений, но размер выборки и дисперсия значений выборки были таковы, что предельная ошибка составила 2,5 конфеты.

§ 5.3. Построение доверительного интервала для среднего (математического ожидания)

Для того чтобы определить доверительный интервал для среднего значения, нам требуется найти следующие величины:

- *выборочную оценку среднего* ($\bar{X}$);
- *стандартную ошибку среднего* ($s_{\bar{x}}$);
- *z-значение для требуемого уровня доверительной вероятности*;
- *предельную ошибку* (*погрешность*) *выборки* (ε) с учётом задаваемой доверительной вероятности.

Результатом решения должен стать интервал вида

$$\bar{X} - \varepsilon \le \mu \le \bar{X} + \varepsilon$$

или

$$\bar{X} - z\frac{s_x}{\sqrt{n}} \le \mu \le \bar{X} + z\frac{s_x}{\sqrt{n}},$$

где μ («*мю*») — *истинное значение среднего в генеральной совокупности* (*т. е. математическое ожидание*), *которое оценивается*.

Ответ записывается так:

$$\mathrm{I} = \left[\bar{X} - z\frac{s_x}{\sqrt{n}}; \bar{X} + z\frac{s_x}{\sqrt{n}} \right]_{\gamma=(0{,}90;0{,}95;0{,}99)}$$

или

$$\mathrm{I} = [a;b]_{\gamma=(0{,}90;0{,}95;0{,}99)},$$

где *a — нижняя граница доверительного интервала*; *b — верхняя граница доверительного интервала*; γ — *назначенный уровень доверительной вероятности*.

Пример 5.7. «Страны». На одном из семинаров по математической статистике со студентами II курса был реализован блиц-опрос на тему «В скольких зарубежных странах мира Вы успели побывать к настоящему моменту своей жизни?». Объём выборки составил 48 человек. Ответы респондентов сведены в таблицу (где X — число стран, в которых побывал студент, во второй строке указано количество человек, давших такой ответ):

X	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	17	20
Респонденты	2	4	3	4	2	5	3	2	3	4	5	2	1	2	1	1	2	2

Найдём доверительный интервал для среднего числа стран, в которых побывал один студент. Уровень доверительной вероятности примем равным 0,90.

Решение. Нам требуется построить доверительный интервал вида: $\bar{X} - \varepsilon \le \mu \le \bar{X} + \varepsilon$.

Сначала вычисляем выборочную оценку среднего. Получим, что среднее арифметическое равно:

$$\bar{X} = \frac{0 \cdot 2 + 1 \cdot 4 + 2 \cdot 3 + 3 \cdot 4 + \ldots + 17 \cdot 2 + 20 \cdot 2}{48} = \frac{360}{48} = 7{,}5 \text{ (стран).}$$

Несмещённая выборочная оценка дисперсии будет равна:

$$s^2 = \frac{(0-7{,}5)^2 \cdot 2 + (1-7{,}5)^2 \cdot 4 + (2-7{,}5)^2 \cdot 3 + \ldots + (20-7{,}5)^2 \cdot 2}{48-1} =$$

$$= \frac{1\,254}{47} = 26{,}68 \text{ (стран)}^2.$$

Тогда несмещённая оценка стандартного отклонения составит: $s = \sqrt{26{,}68} \cong 5{,}2 (\text{стран})$ — это характеристика изменчивости ответов, данных студентами. Вычислим стандартную ошибку самого среднего арифметического:

$$s_{\bar{x}} = \frac{s_x}{\sqrt{n}} = \frac{5{,}2}{\sqrt{48}} = 0{,}75 (\text{стран})$$

— это характеристика изменчивости (вариации) выборочной оценки среднего (т. е. она равна почти ±1 страна). Это минимальная ошибка имеющейся выборки объёмом $n = 48$. Так, мы бы могли сказать, что оценили среднее число стран, в которых побывал студент, с точностью до ±0,75 (стран), и указать интервал $7{,}5 \pm 0{,}75$. Однако этого недостаточно. Согласно правилу «трёх сигм», в этом интервале мы уверены не более, чем на 68 %, а нам бы хотелось (с учётом условия задачи) быть уверенными в выводах на 90 %. Следовательно, стандартную (среднюю) ошибку выборки $s_{\bar{x}} = 0{,}75$ нам потребуется увеличить в некоторое количество раз, с тем чтобы оценить ошибку выборки для доверительной вероятности 0,9. Найдем, чему равно

Таблица 5.1. **Фрагмент таблицы стандартного нормального *z*-распределения (интеграл от 0 до *z*)**

Z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,0000	0,0040	0,0080	0,0120	0,0160	0,0199	0,0239	0,0279	0,0319	0,0359
0,1	0,0398	0,0438	0,0478	0,0517	0,0557	0,0596	0,0636	0,0675	0,0714	0,0753
...										
1,5	0,4332	0,4345	0,4357	0,4370	0,4382	0,4394	0,4406	0,4418	0,4429	0,4441
1,6	0,4452	0,4463	0,4474	0,4484	0,4495	0,4505	0,4515	0,4525	0,4535	0,4545
1,7	0,4554	0,4564	0,4573	0,4582	0,4591	0,4599	0,4608	0,4616	0,4625	0,4633
...	и т. д.									

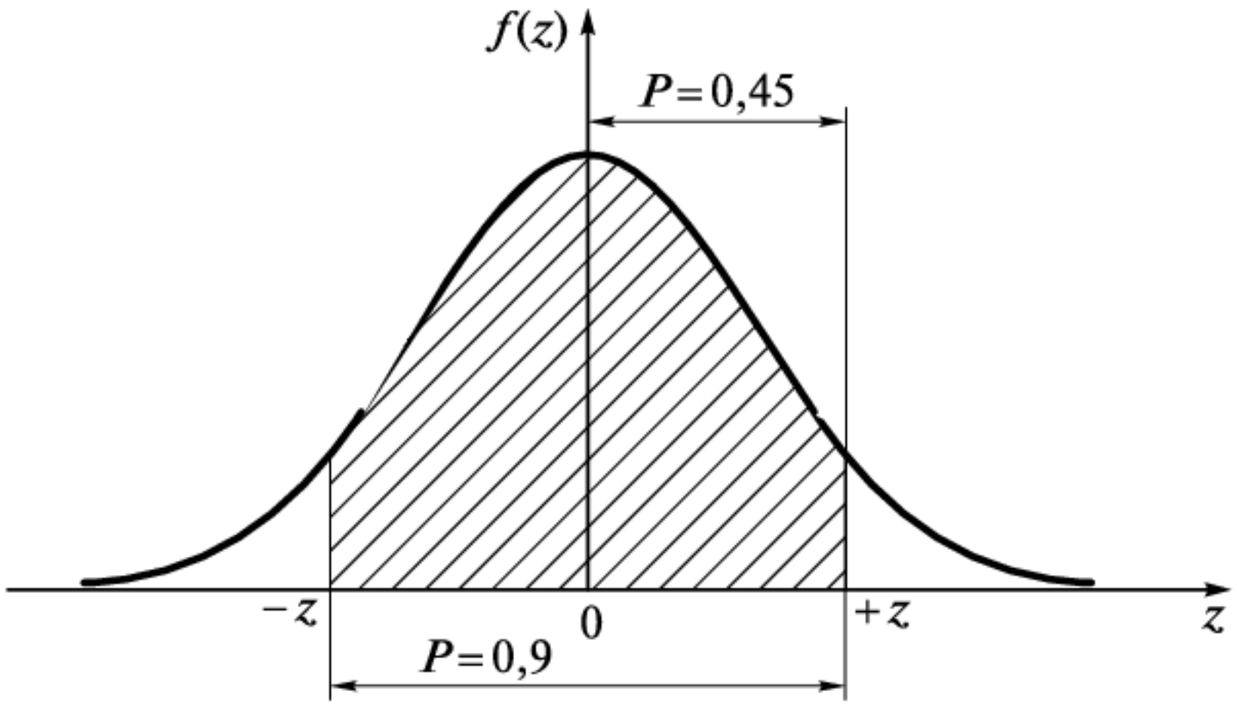


Рис. 5.3. Доверительная вероятность (см. пример 5.7 «Страны»)

z-значение для $\gamma = 0{,}9(90\,\%)$. Воспользуемся таблицей стандартного нормального распределения (табл. 5.1).

Напомним, что нам следует определить z-значение, при котором вероятность $P(-z \le Z \le +z) = 0{,}90$. Площадь под кривой стандартного нормального распределения составляет 0,9 на отрезке от $-z$ до $+z$. Поскольку стандартное нормальное распределение симметрично относительно нуля, то легко видеть, что площадь под кривой стандартного нормального распределения на отрезке значений от 0 до $+z$ составляет ровно половину, т. е. $P(0 \le Z \le +z) = 0{,}45$ (рис. 5.3).

По таблице стандартного нормального z-распределения находим, что z-значение, соответствующее величине интеграла 0,45, равно 1,65. Получаем, что для того чтобы быть уверенными в оценке истинного значения среднего на 90 %, нам потребуется увеличить стандартную ошибку среднего в 1,65 раза.

Тогда получим следующее значение предельной ошибки выборки:

$$\varepsilon = z\frac{s_x}{\sqrt{n}} = 1{,}65 \cdot 0{,}75 \approx 1{,}24 \text{ (стран)}.$$

Искомый доверительный интервал для среднего составит:

$$7{,}5 - 1{,}24 \le \mu \le 7{,}5 + 1{,}24,$$

или

$$6{,}26 \le \mu \le 8{,}74,$$

или

$$6{,}3 \le \mu \le 8{,}7,$$

$$I = [6{,}3;8{,}7]_{\gamma=0{,}90(90\ \%)}$$

Интерпретировать полученный доверительный интервал будем следующим образом: с вероятностью 90 % можно утверждать, что в генеральной совокупности истинное среднее значение числа стран, которые посетил один студент к настоящему моменту своей жизни, гарантированно содержится в диапазоне от 6,3 до 8,7.

Обратите внимание (!), что чем ***выше уровень доверительной вероятности*** (т. е. чем выше меру надёжности выводов мы задаем), тем ***шире будет доверительный интервал***. И, наоборот, чем ***с меньшей вероятностью*** мы стремимся дать оценку истинного значения среднего в генеральной совокупности, тем ***у́же*** *будет построенный* ***доверительный интервал*** для среднего.

Пример 5.8. «Дотации». Пусть в ходе выборочного исследования администрацией некоторого учебного заведения оцениваются средние еженедельные расходы на питание школьников. Эта оценка используется для дальнейшего определения размера дотаций нуждающимся семьям. Предположим, что для случайной выборки школьников средние расходы в неделю составляют 450 руб., а стандартная ошибка — 32 руб. Определим 95%-й доверительный интервал, в отношении которого справедливо говорить, что такова средняя величина еженедельных расходов на питание для *всех* школьников данного учебного заведения.

Решение. Требуется построить доверительный интервал вида $\bar{X} - \varepsilon \le \mu \le \bar{X} + \varepsilon$. Предельная ошибка выборки составит:

$$\varepsilon = 1{,}96\frac{s_x}{\sqrt{n}} = 1{,}96 \cdot 32 \approx 63 \text{ (руб.)}.$$

По таблице стандартного нормального z-распределения находим, что z-значение, соответствующее величине интеграла 0,475 $\left(\frac{0{,}95}{2}\right)$ на участке $[0;+z]$, равно 1,96. Получаем, что для того чтобы быть уверенными в знании истинного значения среднего на 95 %, нам необходимо увеличить стандартную ошибку в 1,96 раз.

Следовательно, искомый доверительный интервал для среднего составит:

$$450 - 63 \le \mu \le 450 + 63,$$

или

$$387 \le \mu \le 513,$$

$$\mathrm{I} = [387;513]_{\gamma=0{,}95(95\ \%)}.$$

С вероятностью 95 % можно утверждать, что для ***всех школьников*** данного учебного заведения *средняя* еженедельная величина расходов на питание лежит в интервале [387; 517] руб.

Пример 5.9. «Дорога». Среди студентов факультета социологии был проведен экспресс-опрос на тему: «Сколько времени Вы обычно затрачиваете на дорогу в будний день, чтобы добраться от дома до университетского кампуса, расположенного в Кочновском проезде[1] (т. е. «от порога до порога»)? Пожалуйста, укажите время в минутах».

[1] Речь идёт о социологическом факультете НИУ «Высшая школа экономики», здание которого расположено по адресу: г. Москва, Кочновский проезд, 3.

Выборка случайная, объём выборки составил 96 человек. Выборочная оценка среднего оказалась равной 81 мин, а несмещённая оценка стандартного отклонения составила 37 мин.

Оцените доверительный интервал, характеризующий средние затраты времени студента на дорогу от дома до университетского кампуса факультета социологии. Уровень доверительной вероятности принять равным 0,95.

Решение. Требуется построить доверительный интервал вида: $\bar{X} - \varepsilon \le \mu \le \bar{X} + \varepsilon$.

Вычислим стандартную ошибку для среднего арифметического:

$$s_{\bar{x}} = \frac{s_x}{\sqrt{n}} = \frac{37}{\sqrt{96}} = 3{,}8 \approx 4(\text{мин})$$

— это характеристика изменчивости (вариации) выборочной оценки среднего (т.е. она равна почти ±4 мин). Поскольку нам бы хотелось (с учетом условия задачи) быть уверенными в выводах на 95 %, то стандартную (среднюю) ошибку выборки следует увеличить в 1,96 раз, чтобы оценить предельную ошибку выборки для доверительной вероятности 0,95.

Тогда искомый доверительный интервал для среднего составит:

$$81 - 1{,}96 \cdot 3{,}8 \le \mu \le 81 + 1{,}96 \cdot 3{,}8,$$

или

$$81 - 7{,}4 \le \mu \le 81 + 7{,}4,$$

или

$$73{,}6 \le \mu \le 88{,}4,$$

$$I = [74;88]_{\gamma=0{,}95(95\ \%)}.$$

Итак, можно с уверенностью 95 % утверждать, что для всех студентов, приезжающих на занятия в университетский кампус, расположенный в Кочновском проезде, *среднее* количество времени, затрачиваемого на дорогу, лежит в интервале [74; 88] мин.

§ 5.4. Построение доверительного интервала для доли (альтернативного признака)

Предположим, что по результатам выборочного исследования (опроса общественного мнения) мы оценили, что доля людей, согласных пойти на президентские выборы в предстоящее воскресенье, составила 62 %. Что тогда можно сказать об истинном значении этой доли в генеральной совокупности, имея в виду всех граждан страны соответствующего возраста?

Если речь заходит о признаке, принимающем всего два возможных значения (две *альтернативы*), одно из которых классифицируется исследователем как «*успех*» (в нашем примере — «согласие участвовать в выборах»), другое — как «*неудача*» (в нашем примере — «отказ от участия в выборах»), пользуясь терминологией испытаний Бернулли, то доля случаев, оценённая по выборке, может быть обобщена на генеральную совокупность на основе построения доверительного интервала для генеральной доли.

Для того чтобы определить доверительный интервал для доли, потребуется найти следующие параметры:

- *выборочную оценку доли* (p);
- *стандартную ошибку доли* (s_p);
- *z-значение для требуемого уровня доверительной вероятности*;
- *предельную ошибку* (*погрешность*) *выборки* (ε) с учётом задаваемой доверительной вероятности.

Результатом решения должен стать интервал вида:

$$p - \varepsilon \le P \le p + \varepsilon,$$

или

$$p - z\frac{\sqrt{pq}}{\sqrt{n}} \le P \le p + z\frac{\sqrt{pq}}{\sqrt{n}},$$

где $\sqrt{pq}$ — мера рассеяния качественного (альтернативного) признака; P («*пэ большое*») — *истинное значение доли в генеральной совокупности*.

Ответ записывается так:

$$\mathrm{I} = \left[p - z\frac{\sqrt{pq}}{\sqrt{n}}; \quad p + z\frac{\sqrt{pq}}{\sqrt{n}} \right]_{\gamma=(0{,}90;0{,}95;0{,}99)},$$

или

$$\mathrm{I} = [a;b]_{\gamma=(0{,}90;0{,}95;0{,}99)},$$

где a — *нижняя граница доверительного интервала*; b — *верхняя граница доверительного интервала*; γ — *назначенный уровень доверительной вероятности*.

Почему стандартное отклонение альтернативного признака рассчитывается как $\sqrt{pq}$? Напомним, чему равна дисперсия случайной величины X, являющейся бинарной, т. е. принимающей только два значения — либо 1, либо 0. Пусть мы кодируем «успех» как «1» (интересующая нас характеристика присутствует), а «неудачу» — как «0» (интересующая нас характеристика отсутствует). Пусть вероятность успеха равна p, а вероятность неудачи равна $q = 1 - p$. Тогда получим закон распределения случайной величины X:

X	1	0
Вероятность	p	q

Среднее значение будет равно: $1 \cdot p + 0 \cdot q = p$.
Дисперсия значений альтернативного признака составит:

$$(1-p)^2 p + (0-p)^2 q = q^2 p + p^2 q = pq(q+p) = pq \cdot 1 = pq.$$

Стандартное отклонение X равно: $\sqrt{pq}$.

Тогда понятно, что *стандартная ошибка* доли качественного признака составит: $\sqrt{\frac{pq}{n}} = \frac{\sqrt{pq}}{\sqrt{n}}$.

Пример 5.10. «Зима». Предположим, что в ходе опроса жителям средней полосы России был задан вопрос «Любите ли Вы зиму?». Пусть случайная выборка из $n = 114$ человек, принявших участие в исследовании, показала, что 45 человек ответили утвердительно: «Да, я люблю зиму», остальные же 69 человек ответили: «Нет, я не люблю зиму». Найдите доверительный интервал для доли граждан, которые не любят зиму. Доверительную вероятность примите равной 90 %.

Решение. Нам требуется построить доверительный интервал вида: $p - \varepsilon \le P \le p + \varepsilon$, где P — истинное значение доли тех, кто зиму не любит, в генеральной совокупности.

«Успехом» будем называть ответ «нет, я не люблю зиму», а «неудачей» назовём противоположный ответ — «да, я люблю зиму».

Сначала вычислим выборочную оценку доли (долю «успехов» в выборке). Получим, что выборочная оценка доли тех, кто не любит зиму, составляет:

$$p = \frac{69}{114} = 0{,}605 = 60{,}5\ \%.$$

Выборочная оценка дисперсии будет равна:

$$s^2 = pq = 0{,}605 \cdot (1 - 0{,}605) = 0{,}605 \cdot 0{,}395 = 0{,}239.$$

Тогда оценка стандартного отклонения составит:

$$s = \sqrt{pq} = \sqrt{0{,}239} \approx 0{,}49.$$

Вычислим стандартную ошибку доли:

$$\sqrt{\frac{pq}{n}} = \frac{\sqrt{pq}}{\sqrt{n}} = \frac{0{,}49}{\sqrt{114}} = \frac{0{,}49}{10{,}68} = 0{,}046 = 4{,}6\ \%$$

— это характеристика изменчивости (вариации) выборочной оценки доли (т. е. она равна ±4,6 %). Поскольку нам бы хотелось (с учётом условия задачи) быть уверенными в выводах на 90 %, стандартную

ошибку доли следует увеличить в 1,65 раз, для того чтобы оценить предельную ошибку выборки для доверительной вероятности 0,90.

Тогда искомый доверительный интервал для истинного значения доли в генеральной совокупности составит:

$$60{,}5\,\% - 1{,}65 \cdot 4{,}6\,\% \le P \le 60{,}5\,\% + 1{,}65 \cdot 4{,}6\,\%,$$

или

$$60{,}5\,\% - 7{,}59\,\% \le P \le 60{,}5\,\% + 7{,}59\,\%,$$

или

$$52{,}91\,\% \le P \le 68{,}09\,\%,$$

или, округляя:

$$53\,\% \le P \le 68\,\%,$$

$$\mathrm{I} = [53;68]_{\gamma=0{,}90(90\,\%)}.$$

Можно с уверенностью (вероятностью) 90 % утверждать, что истинное значение доли людей (жителей средней полосы России), которые не любят зиму, в генеральной совокупности лежит в интервале [53 %; 68 %].

Пример 5.11. «Коррупция». Согласно результатам всероссийского исследования «Диагностика коррупции в России (2005 г.): опрос среди предпринимателей» респондентам (владельцам малого и среднего бизнеса) задавался следующий вопрос: «Предположим, что Вам рассказали про одного человека, который, узнав, что его начальник берёт взятки, написал про него в прокуратуру. Выберите одно из следующих слов (ассоциаций), которое, с Вашей точки зрения, наиболее точно характеризует этого человека»[1]. Распределение ответов на вопрос представлено так:

Слово (ассоциация)	Число респондентов	% ответивших
«Герой»	58	5,6
«Честный человек»	263	25,5
«Чудак»	221	21,5
«Завистник»	120	11,7
«Мстительный человек»	75	7,3
«Предатель»	41	4

[1] Данные Единого архива экономических и социологических данных [http://sophist.hse.ru/db/oprosy.shtml?ts=78&en=0] (проверено 06.07.2013).

Окончание табл.

Слово (ассоциация)	Число респондентов	% ответивших
Другое	136	13,2
Затрудняюсь ответить	116	11,2
Всего	$n = 1\,030$ человек	100 %

Из числа тех, кто не затруднился ответить на вопрос и не предложил свой собственный вариант ответа, определите доверительный интервал для доли предпринимателей, считающих, что такой поступок совершает «Чудак», «Завистник», «Мстительный человек» или «Предатель» вместе взятые. Доверительную вероятность примите равной 0,85 (85 %).

Решение. Из условия задачи следует, что нам требуется рассмотреть две категории ответов предпринимателей: ответы, которые отражают *осуждение* такого поступка (этой точке зрения соответствуют ассоциации «Чудак», «Завистник», «Мстительный человек», «Предатель»), с одной стороны, и ответы, которые свидетельствуют о порицании взяточничества (этой точке зрения соответствуют ассоциации «Герой», «Честный человек») — с другой. Категории ответов «Затрудняюсь ответить» и «Другое» исключим из анализа, как того требует условие задачи.

Перейдём к расчётам. «Успехом» (символом «1») будем считать ответы «Чудак», «Завистник», «Мстительный человек», «Предатель». «Неудачей» (символом «0») назовём ответы «Герой» и «Честный человек». Получим таблицу:

Слово (ассоциация)	Категория	Число респондентов	% ответивших
«Герой»	«Неудача»	58	7
«Честный человек»	«Неудача»	263	34
«Чудак»	«Успех»	221	29
«Завистник»	«Успех»	120	15
«Мстительный человек»	«Успех»	75	10
«Предатель»	«Успех»	41	5
Всего		$n^* = 778$ человек	100 %

Выборочная оценка доли тех, кто оправдывает взяточничество, составляет:

$$p = 29\,\% + 15\,\% + 10\,\% + 5\,\% = 59\,\%.$$

Выборочная оценка доли тех, кто разделяет альтернативную точку зрения (т. е. осуждает взяточничество, давая положительную характеристику человеку, обратившемуся в прокуратуру) будет равна:

$$q = 100\,\% - 59\,\% = 7\,\% + 34\,\% = 41\,\%\ (0{,}41).$$

Выборочная оценка дисперсии равна:

$$s^2 = pq = 0{,}59 \cdot 0{,}41 \approx 0{,}242.$$

Тогда оценка стандартного отклонения составит:

$$s = \sqrt{pq} = \sqrt{0{,}242} \approx 0{,}49.$$

Вычислим стандартную ошибку доли:

$$\sqrt{\frac{pq}{n^*}} = \frac{\sqrt{pq}}{\sqrt{n^*}} = \frac{0{,}49}{\sqrt{778}} = \frac{0{,}49}{27{,}9} \approx 0{,}0176 \approx 1{,}8\,\%$$

— это характеристика изменчивости (вариации) выборочной оценки доли (т. е. она равна ±1,8 %). Это минимальная ошибка рассматриваемой выборки объёмом 778. Поскольку нам бы хотелось (с учётом условия задачи) быть уверенными в выводах на 85 %, то определим, во сколько раз нам следует увеличить стандартную ошибку доли, для того чтобы найти предельную ошибку выборки для доверительной вероятности 0,85. По таблице стандартного нормального распределения найдём, чему равно z-значение для $\gamma = 0{,}85$ (85%).

Итак, нам следует определить z-значение, при котором вероятность $P\ (-z \le Z \le +z) = 0{,}85$. Площадь под кривой стандартного нормального распределения составляет 0,85 на отрезке от $-z$ до $+z$.

Таблица 5.2. **Фрагмент таблицы стандартного нормального *z*-распределения (интеграл от 0 до *z*)**

Z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,0000	0,0040	0,0080	0,0120	0,0160	0,0199	0,0239	0,0279	0,0319	0,0359
0,1	0,0398	0,0438	0,0478	0,0517	0,0557	0,0596	0,0636	0,0675	0,0714	0,0753
...										
1,4	0,4192	0,4207	0,4222	0,4236	0,4251	0,4265	0,4279	0,4292	0,4306	0,4319
1,5	0,4332	0,4345	0,4357	0,4370	0,4382	0,4394	0,4406	0,4418	0,4429	0,4441
1,6	0,4452	0,4463	0,4474	0,4484	0,4495	0,4505	0,4515	0,4525	0,4535	0,4545
...	и т. д.									

Площадь под кривой стандартного нормального распределения на отрезке значений от 0 до $+z$ составляет ровно половину, т. е. 0,425.

По таблице стандартного нормального z-распределения (табл. 5.2) находим, что z-значение, соответствующее величине интеграла 0,425 на участке $[0;+z]$, равно 1,44. Получаем, что для того чтобы быть уверенными в оценке истинного значения доли на 85 %, нам потребуется увеличить стандартную ошибку в 1,44 раза. Тогда получим искомый доверительный интервал для значения доли в генеральной совокупности:

$$59\,\% - 1{,}44 \cdot 1{,}8\,\% \le P \le 59\,\% + 1{,}44 \cdot 1{,}8\,\%,$$

или

$$59\,\% - 2{,}59\,\% \le P \le 59\,\% + 2{,}59\,\%,$$

или

$$56{,}41\,\% \le P \le 61{,}59\,\%,$$

или

$$56{,}4\,\% \le P \le 61{,}6\,\%,$$

или, округляя:

$$56\,\% \le P \le 62\,\%,$$

$$\mathrm{I} = [56;62]_{\gamma=0{,}85(85\,\%)}.$$

Можно с уверенностью (вероятностью) 85 % утверждать, что *истинное значение* доли предпринимателей в генеральной совокупности, *оправдывающих взяточничество* (за исключением тех, кто затрудняется с ответом или предлагает свой собственный вариант ответа), лежит в интервале [56 %; 62 %].

Пример 5.12. «Парашюты». В ходе одного из исследований случайная выборка 96 студентов факультетов социологии и прикладной политологии, показала, что 68 из них хотели бы совершить прыжок с парашютом, остальные — нет. Постройте 95%-й доверительный интервал доли юношей и девушек, которые хотели бы прыгнуть с парашютом.

Решение. Нам требуется построить доверительный интервал вида: $p - \varepsilon \le P \le p + \varepsilon$, где P — истинное (т. е. неизвестное) значение доли студентов в генеральной совокупности рассматриваемых факультетов, кто хотел бы прыгнуть с парашютом.

«Успехом» будем называть ответ: «Да, хочу», а «неудачей» — «Нет, не хочу». Сначала вычисляем выборочную оценку доли (долю «успехов» в выборке). Получим, что выборочная оценка доли тех, кто хотел бы прыгнуть с парашютом, составляет:

$$p = \frac{68}{96} = 0{,}708 = 70{,}8\,\%.$$

Выборочная оценка дисперсии будет равна:

$$s^2 = pq = 0{,}708 \cdot (1 - 0{,}708) = 0{,}708 \cdot 0{,}292 \approx 0{,}207.$$

Вычислим стандартную ошибку доли:

$$\sqrt{\frac{pq}{n}} = \frac{\sqrt{pq}}{\sqrt{n}} = \frac{\sqrt{0{,}207}}{\sqrt{96}} \approx \frac{0{,}45}{9{,}8} = 0{,}046 = 4{,}6\,\%$$

— это характеристика изменчивости (вариации) выборочной оценки доли (т. е. она равна ±4,6 %). Искомый же доверительный интервал для генеральной доли составит:

$$70{,}8\,\% - 1{,}96 \cdot 4{,}6\,\% \le P \le 70{,}8\,\% + 1{,}96 \cdot 4{,}6\,\%,$$

или

$$70{,}8\,\% - 9{,}02\,\% \le P \le 70{,}8\,\% + 9{,}02\,\%,$$

или

$$61{,}78\,\% \le P \le 79{,}82\,\%,$$

или, округляя:

$$62\,\% \le P \le 80\,\%,$$

$$\mathrm{I} = [62;80]_{\gamma=0{,}95(95\,\%)}.$$

Можно с уверенностью (вероятностью) 95 % утверждать, что *истинное значение доли* юношей и девушек в рассматриваемой генеральной совокупности, которые хотят прыгнуть с парашютом, содержится в интервале [62 %; 80 %].

Пример 5.13. «Велосипеды». Предположим, что 800 жителям студенческого городка исследователи задали следующий вопрос: «Воспользуетесь ли Вы в период с мая по сентябрь арендой велосипеда в нашем университетском городке?» Пусть число положительных ответов оказалось равным 280. Оценим долю потенциальных потребителей (т. е. долю целевой аудитории) во всей генеральной совокупности студентов и построим доверительный интервал для доли студентов, заинтересованных в возможности аренды велосипеда, если степень доверия результатам не должна быть меньше 95 %.

Решение. Нам требуется определить доверительный интервал вида:

$$p - \varepsilon \le P \le p + \varepsilon,$$

где P — *истинное* (*неизвестное*) *значение доли целевой аудитории в генеральной совокупности студентов*.

Положительный ответ студента будем классифицировать как «успех», отказ — как «неудачу». Тогда выборочная оценка доли (доли

«успехов» в выборке) будет соответствовать доле тех, кто хотел бы воспользоваться услугой аренды:

$$p = \frac{280}{800} = 0{,}35 = 35\,\%.$$

Выборочная оценка дисперсии будет равна:

$$s^2 = pq = 0{,}35(1 - 0{,}35) = 0{,}35 \cdot 0{,}65 \approx 0{,}228.$$

Вычислим стандартную ошибку доли:

$$\sqrt{\frac{pq}{n}} = \frac{\sqrt{pq}}{\sqrt{n}} = \frac{\sqrt{0{,}228}}{\sqrt{800}} = \frac{0{,}48}{28{,}28} \approx 0{,}017 = 1{,}7\,\%$$

— это характеристика изменчивости (вариации) выборочной оценки доли (т. е. она равна ±1,7 %).

Искомый доверительный интервал для доли составит:

$$35\,\% - 1{,}96 \cdot 1{,}7\,\% \le P \le 35\,\% + 1{,}96 \cdot 1{,}7\,\%,$$

или

$$35\,\% - 3{,}3\,\% \le P \le 35\,\% + 3{,}3\,\%,$$

или, округляя:

$$32\,\% \le P \le 38\,\%,$$

$$\mathrm{I} = [32;38]_{\gamma=0{,}95(95\,\%)}.$$

Можно с уверенностью (вероятностью) 95 % утверждать, что истинное значение доли юношей и девушек, проживающих в студенческом городке и заинтересованных в аренде велосипеда, в генеральной совокупности содержится в интервале [32 %; 38 %]. Эта и есть интервальная оценка истинной доли целевой аудитории данного вида услуги.

§ 5.5. Расчёт объёма выборки в исследовании

В данном параграфе рассмотрим формулы для расчёта объёма выборки, необходимого для проведения исследования. Выведем их из уже известных формул для нахождения доверительных интервалов. Обратите внимание, что размер выборки n фигурирует в этих формулах в таком компоненте, как стандартная ошибка. Выпишем уже известное нам выражение для предельной ошибки выборки и выразим объём выборки через остальные элементы формулы:

• предельная ошибка выборки равна: $\varepsilon = z\frac{s_x}{\sqrt{n}}$. Возведём обе части уравнения в квадрат: $\varepsilon^2 = z^2\frac{s_x^2}{n}$. Отсюда выразим размер выборки: $n = z^2\frac{s_x^2}{\varepsilon^2} = \frac{z^2 s_x^2}{\varepsilon^2}$;

• аналогичным образом поступим для доли (для случая с альтернативным признаком). Размер выборки равен: $n = \frac{z^2 pq}{\varepsilon^2}$.

Из полученных формул видим, что для расчёта требуемого объёма выборки (а сколько наблюдений нужно охватить исследованием?), необходимо иметь следующие данные:

• ***z-значение***. Эту величину может директивно задать (назначить) сам исследователь, исходя из своих притязаний относительно надёжности выводов, которые будут получены на выходе исследования. Обычно мера доверия будущим результатам, как было указано ранее, устанавливается на уровне 90, 95 и 99 % в зависимости от того, насколько «строгим» будет исследование. Чем выше требования к надёжности будущих данных, чем более щепетильным по своему дизайну должно быть исследование, тем строже и требования к доверительной вероятности, а, значит, тем выше и z-значение, и, следовательно, размер выборки. И чем меньше уровень доверительной вероятности, тем менее строгими являются притязания и тем меньший объём выборки потребуется в исследовании;

• ***предельная ошибка выборки*** (т. е. точность выборки). При планировании объёма выборки предстоящего исследования точность также может задать исследователь. Он сам назначает, какую точность хочет обеспечить данным, которые получит по итогам исследования. Обратите внимание, что предельная ошибка выборки и объём выборки находятся в ***обратной*** зависимости друг от друга. Если для исследования сгодится относительно большая погрешность выборки (т. е. величина ε велика), то занижаются и требования к объёму выборки. И наоборот: если исследователь не готов мириться с большой погрешностью выборки, то требуемый объём выборки следует увеличить;

• ***дисперсия***. Имеется принципиально важный вопрос: где же взять информацию о дисперсии изучаемого признака, если исследование еще не проведено и оценка дисперсии пока не известна? В таком случае прибегают к использованию сведений из трёх возможных источников информации о дисперсии:

а) обращение *к **опыту предыдущих исследований*** (банку проектов по схожей тематике), в которых что-то подобное уже изучалось;

б) обращение *к **данным официальной статистики***, содержащей информацию по ряду важных социально-экономических показателей, которые могут пересекаться с содержательными задачами текущего (планируемого) исследования;

в) ***пилотажное (разведывательное) исследование***, которое под силу провести исследователю или его команде. Как правило, любое исследование полного цикла включает в себя реализацию пилотажа (мини-, пробного исследования). Пилотаж — крайне важный этап исследования, в ходе которого проверяются понятность, однозначность терминологии анкеты, логика анкетных вопросов, пригодность гайда интервью, корректность карточки наблюдения, исправляются стратегические и тактические ошибки вопросника, связанные с формулировкой вопросов и полнотой охвата вариантов ответа. Пилотаж призван улучшить качество исследовательского инструментария. *Наряду с этим* в процессе пилотажа можно не только оценить качество и пригодность методических характеристик исследовательского инструментария, но и, проводя интервью, *собрать реальные эмпирические данные* — конкретную, фактическую информацию. Все эти сведения могут пригодиться *для предварительной оценки дисперсии*, чтобы учесть её при расчёте необходимого объёма выборки. Но здесь ***важно помнить об одном важном обстоятельстве — процедуре отбора участников для пилотажного исследования***. Этот отбор не должен быть поверхностным. Имеет смысл уделять пристальное внимание составу участников, чтобы выборка разведывательного исследования во многом повторяла дизайн будущей выборки и отвечала определению эмпирического объекта в этом исследовании.

Пример 5.14. «Майонез». Администратор университетского кафе хотел бы оценить долю посетителей, которых возмущает большое содержание майонеза в блюдах. Выборка какого объёма потребуется в этом исследовании, если надёжность выводов должна составить не менее 90 %, а точность выборки не менее 5 %? В ходе пилотажного исследования (выборка которого случайна) администратор опросил 32 человека, при этом 20 человек ответили, что не удовлетворены этим «кулинарным обстоятельством».

Решение. В рассматриваемом исследовании администрация стремится оценить долю посетителей кафе, недовольных большим содержанием майонеза в предлагаемом меню. В задаче имеем дело с альтернативным признаком: удовлетворенность присутствует, удовлетворенность отсутствует. Пусть «успехом» будем считать факт недовольства, а «неудачей» — противоположный факт. Выпишем основную формулу, которая потребуется для расчёта необходимого объёма выборки, а также формализуем те сведения, которые даны в условии задачи. Применим следующую формулу: $n = \frac{z^2 pq}{\varepsilon^2}$.

По условию задачи уровень доверительной вероятности в исследовании должен составить 90 %, т. е. $\gamma = 0{,}90$. Тогда по таблице стандартного нормального распределения находим, что $z = 1{,}65$. Точность выборки должна составить ±5 %. Это означает, что предельная ошибка нашей выборки не будет превышать 5 %. Остается понять, чему равна дисперсия? Мы имеем дело с данными разведывательного исследования, которое показало, что из 32 респондентов, принявших участие в пилотаже, 20 выразили своё возмущение. Следовательно, мы вполне можем *примерно* оценить по данным пилотажа долю p и q.

Получим: $p = \frac{20}{32} = 0{,}625 = 62{,}5\,\%$, $q = \frac{32-20}{32} = 1 - 0{,}625 = 0{,}375 = 37{,}5\,\%$.

Примерная оценка дисперсии будет равна: $pq = 0{,}625 \cdot 0{,}375 = 0{,}234$. Тогда объём выборки, удовлетворяющий требованиям исследования, должен составить:

$$n = \frac{1{,}65^2 \cdot 0{,}234}{0{,}05^2} = 254{,}83 \approx 255.$$

Менеджеру университетского кафе в ходе своего исследования потребуется опросить 255 человек, для того чтобы обеспечить выполнение требований, предъявленных к качеству выборки.

Пример 5.15. «Товары детского творчества». Торговая компания, реализующая развивающие товары для детей — игрушки, книги, наборы для детского творчества, — решила открыть магазин в новом развивающемся микрорайоне города. Это преимущественно новые дома, где бóльшая часть жителей — семейные пары с маленькими детьми. Корпорация хотела бы оценить планируемые ежемесячные расходы родителей на покупку таких товаров для детей. Для этого решено провести опрос семейных пар с маленькими детьми. Выборка какого объёма необходима исследователям, если предъявляются следующие требования к качеству выборки: доверительная вероятность 0,95, предельная ошибка выборки — не более 300 руб.? Пилотажное (пробное, разведывательное) мини-исследование показало, что оценка стандартного отклонения планируемой суммы расходов составляет около 2 000 руб.

Решение. В этом исследовании нас интересует такая характеристика, как планируемая родителями сумма ежемесячных расходов на покупку товаров рассматриваемой категории. По условию задачи известно, что $\gamma = 0{,}95$. Тогда по таблице стандартного нормального распределения находим, что $z = 1{,}96$. Истинное значение средней суммы расходов, которую планирует семья, должна быть оценена с точностью до ±300 руб. Это означает, что предельная ошибка выборки (ε) не должна превысить 300 руб. Также дана информация о стандартном отклонении, значение которого оценено на основе ответов, которые дали респонденты в пилотажном исследовании: оно

составляет $s_x = 2\,000$ руб. Итак, теперь можно найти необходимый объём выборки:

$$n = \frac{z^2 s_x^2}{\varepsilon^2} = \frac{1{,}96^2 \cdot 2\,000^2}{300^2} = \frac{15\,366\,400}{90\,000} = 170{,}7 \approx 171,$$

т. е. в ходе данного исследования потребуется опросить 171 семейную пару (респондентов).

§ 5.6. Распределение Стьюдента (*t*-распределение). Число степеней свободы

До настоящего момента мы руководствовались знаниями о нормальном распределении. В этом параграфе рассмотрим распределение, которое получило название «*t*-распределение Стьюдента». Это распределение было предложено английским инженером Уильямом Госсетом (1876 — 1937), Стьюдент (Student) — это его псевдоним. Стьюдент был знаком с Карлом Пирсоном, обсуждал с ним вопросы, касающиеся ограничений математической статистики относительно экстраполяции (обобщения) результатов выборки на исследуемую генеральную совокупность. Стьюдент пришел к выводу о необходимости уделять в исследованиях пристальное внимание стандартным ошибкам выборок, размер которых довольно мал.

Распределение Стьюдента будет использоваться нами в ряде случаев для построения доверительных интервалов и определения границ *истинного* значения параметров (чаще всего распределение Стьюдента используется в задачах, которые связаны с оценкой математического ожидания); для нахождения необходимого объёма выборки и при проверке некоторых статистических гипотез. Введение в проверку статистических гипотез изложено в гл. 6. Разберём суть этого распределения и поймём, где и в чём оно может нам понадобиться.

У. Госсет (он же Стьюдент) работал в составе команды исследователей на ирландских пивоваренных заводах, которые принадлежали бренду (как бы мы выразились в настоящее время) Guinness. Он участвовал в проведении научных исследований в области изучения урожайности зерновых (в первую очередь ячменя как основного компонента в пивоварении), а также в реализации выборочного контроля качества продукта в научно-исследовательской лаборатории завода. Особенностью большинства проводимых исследований были выборки ***сравнительно малого объёма*** (!), в то время как методы и инструменты математической статистики, которые были доступны для аналитиков той эпохи, стремительно развивались и совершенствовались преимущественно для выборок с ***большим*** числом

наблюдений. Всё это создавало для исследовательских работ Госсета определенные трудности научного характера: как же в ситуации выборок малого объёма применять, например, стандартное нормальное распределение при построении доверительных интервалов и обобщении выборочных оценок на всю генеральную совокупность? Правомерно ли говорить, что выборочные средние, рассчитанные для множества малых выборок, извлечённых из одной генеральной совокупности, будут подчиняться нормальному распределению? Или будут описываться каким-то другим (особым) распределением?.. Ответ не был очевидным.

Ранее было отмечено, что когда объёмы k выборок, извлеченных ***из одной и той же*** генеральной совокупности, ***велики***, то распределение средних значений, рассчитанных ***для множества*** этих выборок, описывается ***нормальным*** распределением, т. е. согласуется с ним. А как обстоит дело в ситуации, когда ***объёмы выборок малы***? Ранее мы успели затронуть важный аспект статистических исследований — *количество наблюдений* в выборке. Если мы имеем выборки малого объёма (а это выборки, размер которых не превосходит 30 наблюдений), то оценки среднего ($\bar{X}$) и дисперсии (s^2), получаемые при вычислениях, ***неустойчивы***: они могут сильно меняться от одной выборки к другой, на них сложно полагаться, а к получаемому знанию хочется относиться с ощутимой долей сомнения.

Все эти обстоятельства позволили признать наличие существенных ограничений в области методологии статистических исследований и поставили вопрос о корректности использования методов статистических исследований, предназначенных для больших выборок, применительно к выборкам малого размера. Поэтому ключевой темой математических интересов Госсета стал ***поиск распределения вероятностей средних (стандартных) ошибок для выборок малого объёма***. Результатом его прикладных исследований стала научная теория стандартных ошибок для выборочных оценок, полученных на ***малом*** числе наблюдений.

Госсет рассмотрел для малых выборок такую величину как

$$t = \frac{\bar{x} - \mu}{\frac{s_x}{\sqrt{n}}},$$

где $\bar{x}$ — выборочная оценка среднего; μ — истинное значение среднего в генеральной совокупности (т. е. математическое ожидание), а $\frac{s_x}{\sqrt{n}}$ — стандартная ошибка среднего. Эта ***величина*** $\frac{\bar{x} - \mu}{\frac{s_x}{\sqrt{n}}}$ (для удобства запишем её как $\frac{\sqrt{n}(\bar{x} - \mu)}{s_x}$) ***получила название «t-статистика»,***

или «t-статистика Стьюдента»: она является ***нормированным значением отклонения выборочной оценки среднего от истинного значения среднего***. Поясним изложенное. В числителе $(\bar{x} - \mu)$ в абсолютном выражении указана степень расхождения (несовпадения) между оценкой среднего, которую даёт выборка малого объёма, от истинного значения среднего в генеральной совокупности. Нормированное значение этой разности может быть получено, если выразить её в ***единицах стандартной ошибки среднего***, т. е. разделить на $\frac{s_x}{\sqrt{n}}$ (на своего рода «масштабирующий» элемент, дающий возможность «уйти» от абсолютных единиц измерения и перейти к относительным). Что и сделал Уильям Госсет. Всё это, кстати, очень похоже на z-стандартизацию и сильно напоминает знакомую процедуру вычисления z-оценок: $Z = \frac{X - \mu}{\sigma}$.

Видим, что у t-статистики, полученной для выборки малого объёма, ***два источника погрешности*** (вариации): 1) сама выборочная оценка среднего $\bar{x}$, значение которой очень чувствительно к малому числу наблюдений в выборке; 2) выборочная оценка стандартного отклонения s_x (она «завязана» на оценку среднего, так как рассчитывается только после того, как найдено среднее). В случае выборки малого объёма оценка дисперсии неустойчива: её значения могут *заметно* меняться от одной выборки к другой, подвержены заметным колебаниям. Именно исходя из этих соображений нельзя быть уверенными, что распределение вероятностей t-статистики будет описываться нормальным распределением (кривой Гаусса). Госсет в своих исследовательских работах показал, что это распределение t-статистики ($t = \frac{\sqrt{n}(\bar{x} - \mu)}{s_x}$) описывается ***несколько иным распределением***, но по внешнему виду очень напоминающим форму «колокольчика» (форма t-распределения Стьюдента действительно очень похожа на форму нормального распределения, однако t-распределение более пологое, более «вытянутое» вдоль оси абсцисс, в сравнении с классическим нормальным распределением Гаусса). «Хвосты» распределения Стьюдента *медленнее* стремятся к нулю, чем «хвосты» нормального распределения (рис. 5.4).

На рисунке, иллюстрирующем внешний вид распределений, видно, что t-распределение более «прижато» к оси абсцисс, в сравнении с традиционным стандартным нормальным распределением, и имеет более широкие крылья.

Теперь остановимся на важном параметре распределения Стьюдента: t-распределение Стьюдента относится к *группе непрерывных*

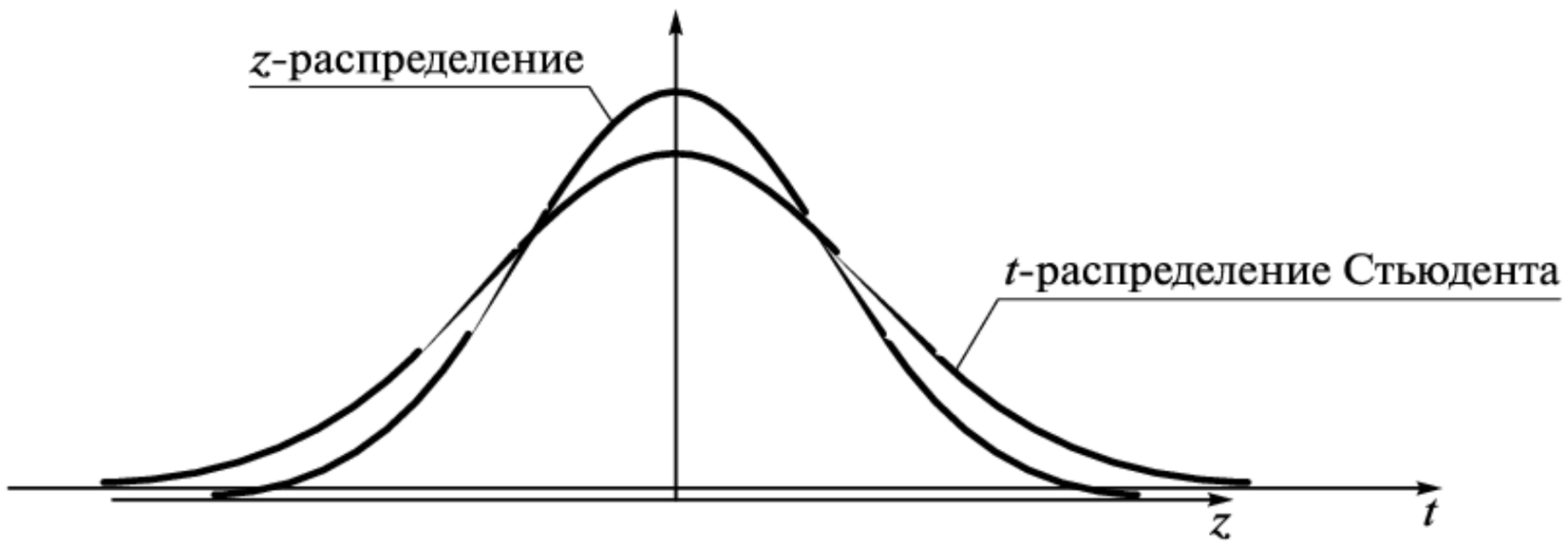


Рис. 5.4. Внешний вид *z*-распределения и *t*-распределения Стьюдента

одномерных распределений (*это целое семейство кривых*![1]), у которого имеется только *один параметр — число степеней свободы* (*degree of freedom*, *df*)[2].

Математически уравнение этого распределения выглядит довольно сложно — оно включает в себя так называемую гамма-функцию. Наличие гамма-функции означает, что ***форма распределения является разной в зависимости от объёма выборки*** (и, следовательно, от *df*). На страницах данной книги ограничимся лишь словесным описанием *t*-распределения, его содержательным смыслом и интерпретацией (этого будет достаточно для умения пользоваться распределением на практике).

При изменении объёма выборки (*n*), т. е. при увеличении или уменьшении *n*, *t*-распределению удовлетворяют разные по форме кривые: либо кривая ещё более островершинная, либо, наоборот, более плосковершинная соответственно. Более остроконечным (как бы «вытягивающимся вверх») *t*-распределение становится *по мере роста числа наблюдений в выборке* (т. е. с увеличением *n*), постепенно сходясь к нормальному распределению (рис. 5.5). Чем больше число степеней свободы, тем ближе *t*-распределение стремится к нормальному.

Разберём новое и очень важное для нас понятие — ***число степеней свободы***. Непростое понятие «число степеней свободы» часто используется в лексике многих наук. Чаще всего под этим выражением в естественных науках (например, в физике, химии, биологии) понимают *наименьшее* число факторов, *зная которые*, можно безошибочно *предугадать* (определить) *конфигурацию* — строение и/или положение в пространстве — интересующего объекта. Поясним это на ***простом примере***.

---

[1] Для выборок фиксированного малого объёма существуют свои кривые Стьюдента (для $n = 2, 3, 4, \ldots, 15, 16, \ldots$ объёма наблюдений характерен ***свой уникальный внешний «дизайн»*** распределения Стьюдента).

[2] Это новый для нас термин, который будет уточнён чуть далее в данном параграфе.

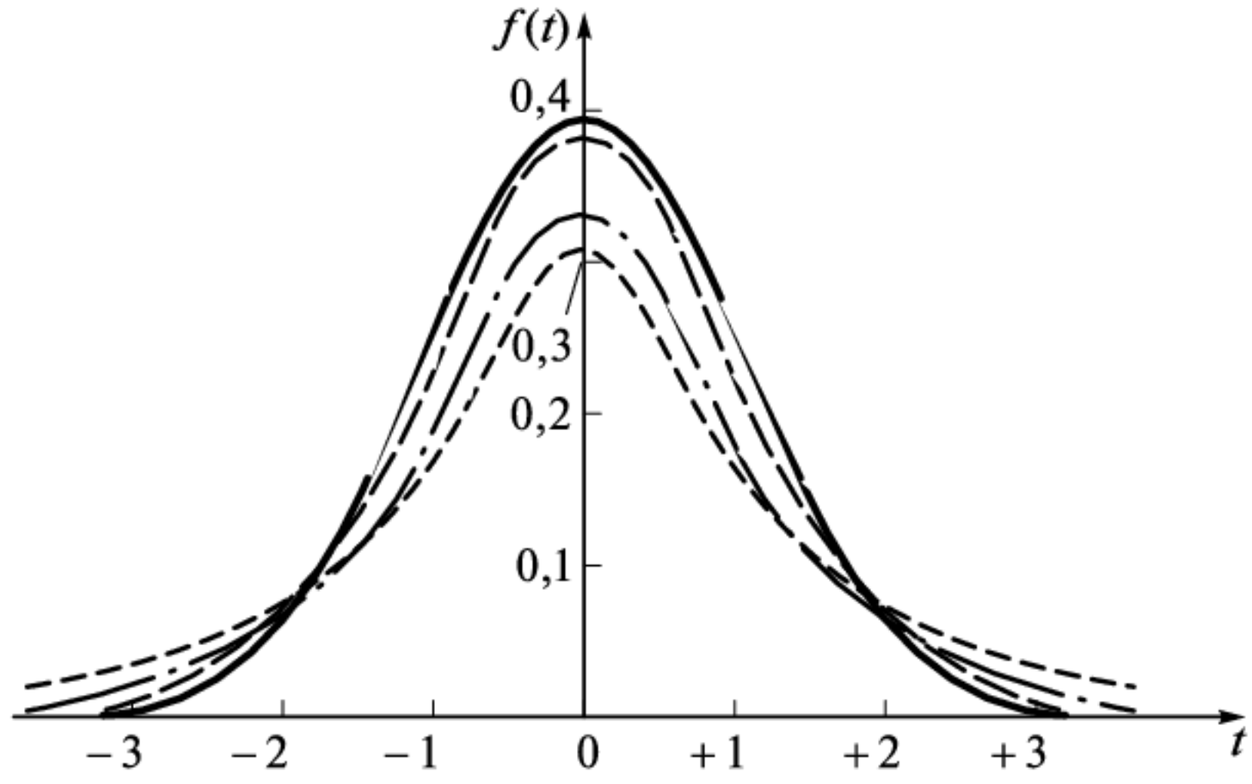


Рис. 5.5. График *t*-распределения Стьюдента для разного числа степеней свободы (*df*)[1]

Пример 5.16. «***Конфеты***». Предположим, в моём кармане лежит 10 конфет, и я хочу угостить пятерых своих друзей — раздать все конфеты без остатка Марине, Ирине, Ивану, Дмитрию, Михаилу. Пусть Марине я отдала две конфеты, Ирине — тоже две, Дмитрию — одну и Ивану — одну. Сколько же теперь конфет я должна отдать Михаилу, если не хочу оставлять у себя конфеты? Конечно, четыре (10 – 2 – – 2 – 1 – 1 = 4). Как только я раздала конфеты *первым четверым* из пяти своих друзей, у меня просто нет другого выбора, кроме как отдать Михаилу 4 конфеты. Я не вольна́ в выборе иного числа конфет, которые могу ему отдать, при том условии, что в моём кармане не должно остаться ни одной конфеты. Распределяя конфеты между пятью своими друзьями, я обладаю ***только четырьмя степенями свободы*** (не пятью!)[2]. Возвращаясь к понятию числа степеней свободы, можно уверенно говорить о том, что в этом примере я обладаю (n – 1) числом степеней свободы, т. е. 5 – 1 = 4.

Итак, применительно к статистике ***числом степеней свободы будем называть число наблюдений выборки***, ***которые остались «свободными» в нашем распоряжении после того***, ***как часть информации о них (наблюдениях выборки) мы использовали (т. е. как бы «израсходовали», «потратили») в других предшествующих видах анализа***. *Числом степеней свободы* станет *число наблюдений выборки*, *уменьшенное на число параметров*, *которые оцениваются исследователем до того*, *как вычисляется текущий*

---

[1] Макет графиков плотности *t*-распределения Стьюдента, представленных на рис. 5.5, взят из открытых источников Интернета.

[2] Интернет пример (иллюстрирующий понятие *df*) про распределение менеджером бюджета для бизнес-проектов приводится в кн.: *Ниворожкина Л. И.*, *Морозова З. А.* Теория вероятностей и математическая статистика. — М. : Эксмо, 2008. — С. 242 — 243.

параметр (*характеристика*). *Это словно «чистый» объём выборки, на котором рассчитана интересующая нас характеристика.*

Для выборок малого размера актуально следующее: оценка дисперсии *зависит* от того, сколько параметров вычисляется *до неё*. А это только *один* (!) параметр — среднее арифметическое. Неустойчивость оценки среднего *сказывается* на расчёте оценки дисперсии. Чем *меньше число степеней свободы*, тем *меньше доверия* к результатам (оценке), которые получаем на выходе исследования, больше сомнения. И наоборот: чем *больше число степеней свободы*, тем *более уверены* мы в наших данных и выборочных оценках параметров, которые рассчитываем, которыми располагаем. ***Число степеней свободы напрямую связано с точностью оценивания, с точностью выборки***. В t-распределении Стьюдента ***число степеней свободы будет равно***: $df = (n - 1)$ для каждой выборки, задействованной в анализе. Специалистами составлена справочная таблица для оценивания значений t-статистики (аналогично таблице стандартного нормального распределения), с помощью которой определяют площадь под кривой плотности t-распределения на отрезке интересующих значений $[a; b]$. Таблица t-распределения составлена для целого семейства распределений, так как числа степеней свободы меняются от одной выборки к другой и могут составлять 1, 2, 3, 4, 5, ... и т. д. С помощью таблицы решается та же задача, что и в случае стандартного нормального распределения: ***находится вероятность того, что случайная величина принимает значения в диапазоне от a до b***.

Статистическая таблица для t-распределения устроена просто: по строкам указаны числа степеней свободы (от 1 до 120), по столбцам — *наиболее часто используемые уровни доверительной вероятности*, которые могут интересовать исследователя (90, 95, 98, 99 и 99,9 %). Уровень доверительной вероятности соответствует величине площади под кривой плотности t-распределения Стьюдента ***на отрезке от –t до +t***. На пересечении соответствующих строки и столбца таблицы указано требуемое значение t-статистики. Это и будет тем искомым значением, которое удовлетворяет условию $P(-t \le T \le +t) = \gamma$. Буквой α («*альфа*») обозначается площадь под кривой плотности в обоих «хвостах» t-распределения Стьюдента, эти участки ***не охватываются*** отрезком $[-t; +t]$ (рис. 5.6). Иными словами, вероятность того, что значение t-статистики Стьюдента будет принадлежать одному из «хвостов» распределения, соответствует вероятности $P(-\infty \le T < -t) = \frac{\alpha}{2}$ и $P(+t < T \le +\infty) = \frac{\alpha}{2}$.

Например, значение t-статистики, которое соответствует доверительной вероятности 0,95 и числу степеней свободы 8 (т. е. для выборки объёмом $n = 9$), составляет $\mathbf{2{,}306} \approx 2{,}31$. А значение t-статистики, которое получится для доверительной вероятности 0,95 и числа степеней свободы 100 (т. е. для выборки размером $n = 101$), составляет

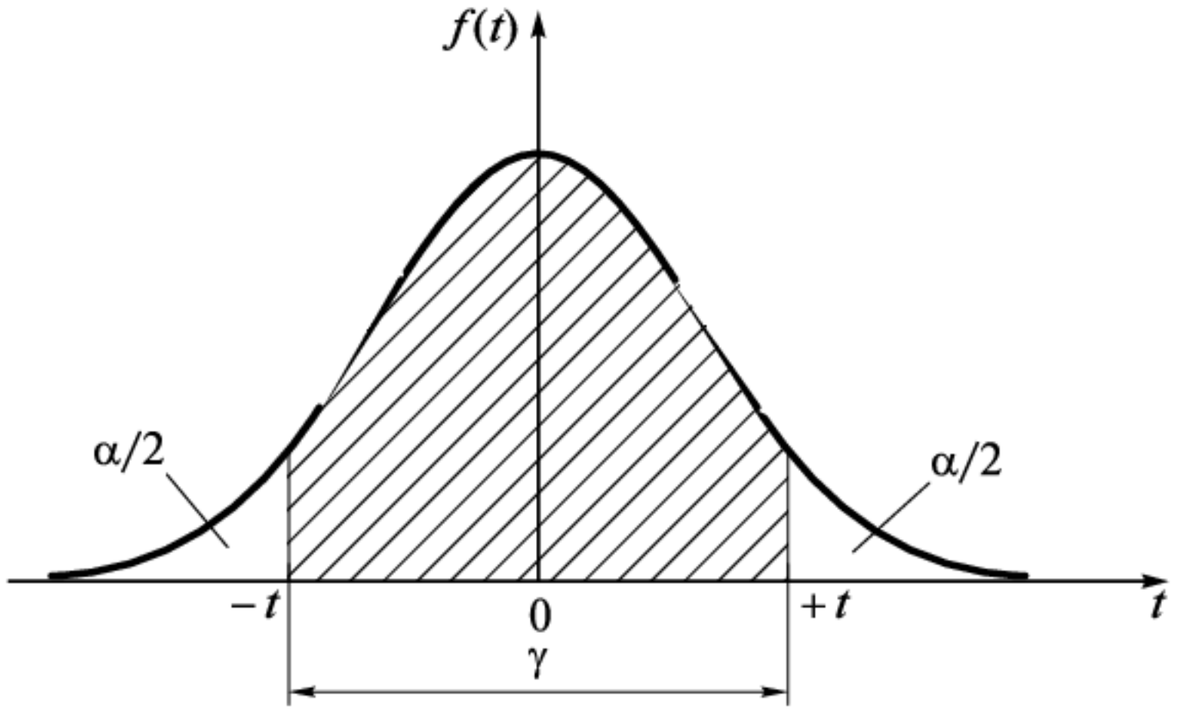
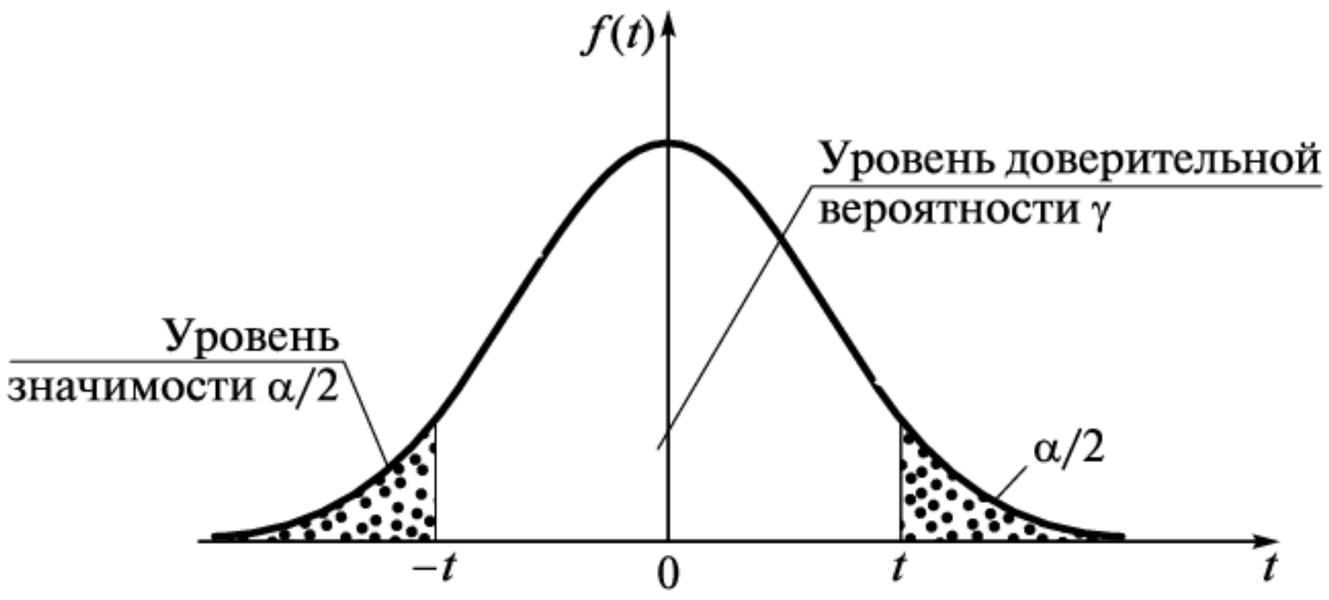


Рис. 5.6. График *t*-распределения Стьюдента и доверительная вероятность γ

1,984 (почти равно *z*-значению в стандартном нормальном распределении для аналогичного уровня доверительной вероятности — **1,96**). Таблица 5.3 иллюстрирует, что ***расхождение*** (несовпадение) значений *t*-статистики и *z*-статистики ***увеличивается с уменьшением df*** (различия между двумя распределениями нарастают) и ***уменьшается с увеличением df*** (различия между двумя распределениями постепенно исчезают вовсе, *t* приближается к *z*).

Таблица 5. 3. **Таблица для *t*-распределения Стьюдента (двустороннее значение вероятности)**



df	Доверительная вероятность и уровень α				
	γ=0,9 (90 %)	γ=0,95 (95 %)	γ=0,98 (98 %)	γ=0,99 (99 %)	γ=0,999 (99,9 %)
	α=0,1 (т. е. α/2=0,05)	α=0,05 (т. е. α/2=0,025)	α=0,02 (т. е. α/2=0,01)	α=0,01 (т. е. α/2=0,005)	α=0,001 (т. е. α/2=0,0005)
1	6,314	12,706	31,821	63,657	636,619
2	2,920	4,303	6,965	9,925	31,599

Продолжение табл. 5.3

df	Доверительная вероятность и уровень α				
	γ=0,9 (90 %)	γ=0,95 (95 %)	γ=0,98 (98 %)	γ=0,99 (99 %)	γ=0,999 (99,9 %)
	α=0,1 (т. е. α/2=0,05)	α=0,05 (т. е. α/2=0,025)	α=0,02 (т. е. α/2=0,01)	α=0,01 (т. е. α/2=0,005)	α=0,001 (т. е. α/2=0,0005)
3	2,353	3,182	4,541	5,841	12,924
4	2,132	2,776	3,747	4,604	8,610
5	2,015	2,571	3,365	4,032	6,869
6	1,943	2,447	3,143	3,707	5,959
7	1,895	2,365	2,998	3,500	5,408
8	1,860	2,306	2,897	3,355	5,041
9	1,833	2,262	2,821	3,250	4,781
10	1,813	2,228	2,764	3,169	4,587
11	1,796	2,201	2,718	3,106	4,437
12	1,782	2,179	2,681	3,055	4,318
13	1,771	2,160	2,650	3,012	4,221
14	1,761	2,145	2,625	2,977	4,141
15	1,753	2,131	2,603	2,947	4,073
16	1,746	2,120	2,584	2,921	4,015
17	1,740	2,110	2,567	2,898	3,965
18	1,734	2,101	2,552	2,878	3,922
19	1,729	2,093	2,540	2,861	3,883
20	1,725	2,086	2,528	2,845	3,850
21	1,721	2,080	2,518	2,831	3,819
22	1,717	2,074	2,508	2,819	3,792
23	1,714	2,069	2,500	2,807	3,768
24	1,711	2,064	2,492	2,797	3,745
25	1,708	2,060	2,485	2,787	3,725
26	1,706	2,056	2,479	2,779	3,707

Окончание табл. 5.3

df	Доверительная вероятность и уровень α				
	γ=0,9 (90 %)	γ=0,95 (95 %)	γ=0,98 (98 %)	γ=0,99 (99 %)	γ=0,999 (99,9 %)
	α=0,1 (т. е. α/2=0,05)	α=0,05 (т. е. α/2=0,025)	α=0,02 (т. е. α/2=0,01)	α=0,01 (т. е. α/2=0,005)	α=0,001 (т. е. α/2=0,0005)
27	1,703	2,052	2,473	2,771	3,690
28	1,701	2,048	2,467	2,763	3,674
29	1,699	2,045	2,462	2,756	3,659
30	1,697	2,042	2,457	2,750	3,646
35	1,690	2,030	2,438	2,724	3,591
40	1,684	2,021	2,423	2,705	3,551
45	1,679	2,014	2,412	2,690	3,520
50	1,676	2,009	2,403	2,678	3,496
55	1,673	2,004	2,396	2,668	3,476
60	1,671	2,000	2,390	2,660	3,460
70	1,667	1,994	2,381	2,648	3,435
80	1,664	1,990	2,374	2,639	3,416
90	1,662	1,987	2,369	2,632	3,402
100	1,660	1,984	2,364	2,626	3,391
110	1,659	1,982	2,361	2,621	3,381
120	1,658	1,980	2,358	2,617	3,374
∞	1,645	1,960	2,326	2,576	3,291

Обратите особое внимание на тот факт, что приближение *t*-распределения к *z*-распределению становится очень быстрым и очевидным, ***начиная с числа степеней свободы df = 30*** и далее. Расхождение (несовпадение) между значениями *z* и *t* наблюдается только для тех случаев, когда объём выборки ***мал***. Начиная с *df* = 30, оба распределения — и стандартное нормальное, и *t*-распределение — могут рассматриваться как почти идентичные. Поэтому ***«30 наблюдений» является в статистике условной границей (рубежом) между понятиями «большая выборка» и «малая выборка»***.

$$\bar{X} = \frac{12+0+3+8+6+2+4+0+1+4}{10} = \frac{40}{10} = 4 \text{ (посещений).}$$

Несмещённая выборочная оценка дисперсии равна:

$$s^2 = \frac{(12-4)^2 + (0-4)^2 + (3-4)^2 + \ldots + (4-4)^2}{10-1} = \frac{130}{9} = 14{,}4 \text{ (посещений)}^2.$$

Для доверительной вероятности $\gamma = 95\,\%$ значение t-статистики по табл. 5.3 составляет $\{df = 10-1 = 9,\ \gamma = 95\,\%\} = 2{,}262$. Тогда требуемый объём выборки составит: $n = \frac{2{,}262^2 \cdot 14{,}4}{1^2} = 73{,}68 \approx 74$ (человек).

Итак, для того чтобы с точностью ± 1 оценить, каково среднее число посещений театра жителями города, в исследовании потребуется выборка объёмом 74 человек (респондентов).

Пример 5.17. «ДТП». Страховая компания оценивает среднее число дорожно-транспортных происшествий, в которых водители побывали за последний год. Случайная выборка, состоящая из 11 человек, представлена следующими данными о числе ДТП: **0, 1, 3, 2, 3, 0, 6, 3, 2, 0, 2.** На основании имеющейся информации постройте доверительный интервал для среднего числа аварий в генеральной совокупности водителей. Уровень доверительной вероятности примите равным **0,95**. Дайте содержательную интерпретацию найденному доверительному интервалу.

Решение. Объём выборки имеющегося исследования мал, так как составил всего 11 наблюдений. Для построения доверительного интервала воспользуемся выражением: $\bar{X} - t\frac{s_x}{\sqrt{n}} \le \mu \le \bar{X} + t\frac{s_x}{\sqrt{n}}$ (рассматриваемая величина — число ДТП; в основе измерения — количественная шкала).

Рассчитаем выборочные оценки среднего и несмещённой дисперсии для числа ДТП:

$$\bar{X} = \frac{0+1+3+2+3+0+6+3+2+0+2}{11} = \frac{22}{11} = 2 \text{ (ДТП).}$$

Несмещённая выборочная оценка дисперсии равна:

$$s^2 = \frac{3\cdot(0-2)^2 + (1-2)^2 + 3\cdot(2-2)^2 + 3\cdot(3-2)^2 + (6-2)^2}{11-1} = \frac{32}{10} = 3{,}2 (\text{ДТП})^2.$$

Несмещённая выборочная оценка стандартного отклонения составит:

$$s = \sqrt{3{,}2} \approx 1{,}79 \text{ (ДТП).}$$

Резюмируя, сделаем практические выводы из изученного материала и решим задачу.

Итак, если объём выборки исследования мал (т.е. не превосходит 30 наблюдений, а более строго — не превышает 100 наблюдений), то при нахождении доверительных интервалов для истинного значения среднего или значения доли следует использовать t-статистику взамен z-статистики. Тогда математическая запись доверительных интервалов примет вид:

- ***доверительный интервал для среднего***:

$$\bar{X} - t\frac{s_x}{\sqrt{n}} \le \mu \le \bar{X} + t\frac{s_x}{\sqrt{n}};$$

- ***доверительный интервал для доли***:

$$p - t\frac{\sqrt{pq}}{\sqrt{n}} \le P \le p + t\frac{\sqrt{pq}}{\sqrt{n}};$$

- ***необходимый объём выборки***:

$$n = \frac{t^2 s_x^2}{\varepsilon^2} \text{ или } n = \frac{t^2 pq}{\varepsilon^2}.$$

Пример 5.16. «Театр». В исследовании, которое проводится в городе А*, оценивается среднее количество посещений театра в год. Из пилотажа (выборка которого случайна) известно, что распределение ответов 10 респондентов на вопрос исследователей: «Сколько раз за минувший год Вы посетили театр?» выглядит следующим образом:

№ интервью	1	2	3	4	5	6	7	8	9	10
Количество посещений	12	0	3	8	6	2	4	0	1	4

Найдите объём выборки, требуемый для проведения полноценного исследования, чтобы точность выборки составила *не менее* ±1 посещений, а уровень надёжности выводов — *не менее* 95 %. Обратите внимание на малый объём выборки в пилотажном исследовании, выбирая соответствующую статистику (z или t?).

Решение. Необходимый объём выборки рассчитаем по формуле

$$n = \frac{t^2 s_x^2}{\varepsilon^2}.$$

Объём выборки в пилотаже составил 10 человек. Пилотажное исследование позволяет оценить среднее и дисперсию числа посещений:

Вычислим стандартную ошибку для среднего арифметического:

$$s_{\bar{x}} = \frac{s_x}{\sqrt{n}} = \frac{1,79}{\sqrt{11}} = 0,54 \approx 0,5 (\text{ДТП})$$

— это характеристика изменчивости (вариации) выборочной оценки среднего.

Поскольку при обобщении результатов исследования на генеральную совокупность нам бы хотелось (с учётом условия задачи) быть уверенными в выводах на 95 %, то стандартную ошибку следует увеличить в ***t*** раз.

Для доверительной вероятности $\gamma = 95\,\%$ значение t-статистики по таблице Стьюдента составляет $\{df = 11-1 = 10,\ \gamma = 95\,\%\} = 2,228$.

Важное примечание. Значение 2,228 больше значения 1,96 (к которому мы привыкли в случае стандартного нормального распределения при аналогичном уровне доверия, равном 0,95). Это означает, что доверительный интервал в рассматриваемом примере будет несколько шире, чем при использовании z-статистики. А, следовательно, наши представления о значении среднего в генеральной совокупности будут менее однозначными (так как доверительный интервал получится относительно более широким, более «размытым»). Причина тому — малый объём выборки: именно небольшое число наблюдений выборки, на основе которой осуществлена оценка параметров (среднего и дисперсии), вынуждает нас с бóльшим недоверием относиться к полученной информации, а потому, делая прогноз относительно истинного значения среднего, мы словно «боимся» ошибиться в своих выводах и опираемся на более широкий доверительный интервал. Так, мы уверены, что истинное значение оцениваемого параметра (в нашем случае — среднего) наверняка «укладывается» в этот найденный интервал.

Итак, искомый доверительный интервал для среднего составит:

$$2 - 2,23 \cdot 0,54 \le \mu \le 2 + 2,23 \cdot 0,54,$$

или

$$2 - 1,2 \le \mu \le 2 + 1,2,$$

или

$$0,8 \le \mu \le 3,2,$$

$$I = [0,8; 3,2]_{\gamma = 0,95 (95\,\%)}.$$

На основе имеющихся данных можно с уверенностью 95 % утверждать, что для водителя, имеющего страховой полис в рассматриваемой компании (имеется в виду генеральная совокупность водителей), среднее число аварий в год лежит в интервале [0,8; 3,2]. Вывод можно сформулировать и так: с вероятностью 0,95 можно

утверждать, что среднее число дорожно-транспортных происшествий (в расчёте на 10 водителей) охватывается интервалом [8; 32].

ЗАДАЧИ ДЛЯ САМОСТОЯТЕЛЬНОГО РЕШЕНИЯ

Задача 1. «Без названия». Выборка представлена следующими значениями X: 10, 2, 6, 6, 4, 2.

Вычислите несмещённую выборочную оценку дисперсии и стандартного отклонения.

Задача 2. «Спорт». В ходе небольшого опроса, проведённого на одном из занятий по математической статистике (в 2013 г.) среди студентов факультета прикладной политологии НИУ «Высшая школа экономики», участникам семинара было предложено ответить на несколько коротких вопросов. Одним из таких вопросов был следующий: «Укажите, пожалуйста, сколько примерно часов в неделю Вы тратите на спорт и различные физические упражнения (включая бег, бассейн, утреннюю зарядку, ролики, танцы и прочие виды активности). Напишите с точностью до получаса». Объём выборки юношей составил 8 человек, девушек — 11 человек. Ответы распределились следующим образом:

Юноши: 2; 5,5; 6,5; 10; 2,5; 5; 7; 7.

Девушки: 7; 4; 2; 5; 4; 0; 0; 2; 3; 2; 5.

Для обеих выборок найдите выборочные оценки среднего, дисперсии (как смещённую, так и несмещённую), оцените величину смещения. Рассчитайте несмещённые выборочные оценки стандартного отклонения.

Задача 3. «Агентство недвижимости». Случайная выборка исследования из $n = 225$ человек, обратившихся в агентство недвижимости, показала, что 174 из них нашли себе подходящее съёмное жильё с его помощью. Постройте 92%-й доверительный интервал доли людей, которые успешно находят себе жилье через агентство недвижимости.

Задача 4. «Кошки». На выставке экзотических пород кошек случайной выборке владельцев этих животных задали вопрос: «Сколько экзотических кошек живёт у Вас дома?» Получены следующие результаты опроса:

Сколько кошек живёт у Вас дома?	1	2	3	4	5	6	Всего
Количество респондентов	10	20	15	12	5	2	$n = 64$

Для ответа на вопрос, сколько экзотических кошек, в среднем, живёт у одного владельца, постройте 95%-й доверительный интервал. Также найдите коэффициент вариации. Интерпретируйте его.

Задача 5. «Португальский язык». Десять случайно отобранных в некоторый вуз абитуриентов получили следующие баллы на вступительных экзаменах по португальскому языку: 5, 2, 10, 7, 3, 7, 5, 5, 6, 8. Всего португальский язык сдавало несколько сотен человек. Каков интервал, в котором в генеральной совокупности с вероятностью 0,95 лежит математическое

ожидание баллов (т.е. истинное среднее), полученных всеми поступающими в вуз абитуриентами? Каким будет интервал, если уровень доверительной вероятности задать равным 0,90; 0,99?

Задача 6. «О праздниках». По данным ВЦИОМ (Всероссийского центра изучения общественного мнения) от 18.02.2011[1], существенная часть жителей России отмечает православный праздник День Петра и Февронии, который с 2008 г. на государственном уровне называется как День семьи, любви и верности. Опрос общественного мнения, проведенный ВЦИОМ, показывает, что многие россияне не представляют семейных отношений и любви без верности: «Независимо от того, состоят ли в официальном браке, живут ли вместе вне брака или просто встречаются, — пишет ВЦИОМ — измену второй половины не простят ни при каких обстоятельствах 71 % россиян. Даже флирт, правда, с оговоркой «постоянный», не потерпят 59 % опрошенных». Зная, что объём случайной выборки составил 1 600 человек, постройте 95%-й доверительный интервал для доли людей, которые говорят о верности в отношениях.

Задача 7. «Ректорат». Ректорат университета хотел бы знать мнение студентов о том, нравится ли им, как построено обучение в университете. Из 500 опрошенных студентов 103 ответили, что обучение им не нравится. Оцените долю студентов, довольных обучением в университете. Постройте 93%-й доверительный интервал.

Задача 8. «Подростковая преступность». В ходе одного из всероссийских исследований гражданам задавался такой вопрос: «Какое действие, по Вашему мнению, должна предпринять милиция и суд по отношению к подросткам, которые жестоко избили бомжа? Если это совершил 15-летний подросток в компании с другими подростками, но он был не зачинщиком, а участником?»[2]. Распределение ответов представлено в таблице:

№ п/п	Действие (принимаемое взыскание)	Количество респондентов
1	Провести разъяснительную беседу с подростком и родителями	166
2	Оштрафовать родителей, обязать их возместить ущерб	148
3	Поставить на учет в милицию	680
4	Поместить в спецшколу (коррекционную школу)	243
5	Наказать лишением свободы условно (до следующего правонарушения)	206

[1] Подробно о результатах этого исследования читайте на сайте [http://wciom.ru/index.php?id=269&uid=111374] (проверено 28.06.2012).

[2] Данные Единого архива экономических и социологических данных (проверено 28.06.2012).

Окончание табл.

№ п/п	Действие (принимаемое взыскание)	Количество респондентов
6	Наказать лишением свободы с отбыванием срока в колонии	69
7	Другие меры	11
	Всего	1 523

Постройте доверительный интервал для доли граждан в России, считающих, что в такой ситуации можно ограничиться проведением беседы или штрафом родителей (т. е. категории ответов № 1 и № 2). Уровень доверительной вероятности примите равным 0,95.

Задача 9. «Благосостояние». Предположим, что при исследовании благосостояния населения необходимо обеспечить интервальную оценку дохода (с помощью среднего арифметического) на уровне доверительной вероятности 0,95, которая будет находиться в пределах ±1 000 руб. Какой объём выборки потребуется для этого исследования, если из опыта предыдущих обследований населения известно, что стандартное отклонение составляет 12 200 руб.?

Задача 10. «Электоральные предпочтения россиян». По результатам опроса «ФОМнибус» от 19—20 ноября 2011 г. Фонд «Общественное мнение» опубликовал распределение ответов россиян на вопрос: «Если в декабре 2011 г. Вы примите участие в выборах Государственной думы, то как, за какую партию Вы, скорее всего, проголосуете?» (закрытый вопрос, карточка, один ответ). Опрошено 3 000 человек в 204 населенных пунктах, расположенных в 64 субъектах РФ. Получены следующие результаты:

Партия	Процент голосов поддержки (от общего числа респондентов)
Единая Россия	39
КПРФ	12
ЛДПР	10
Справедливая Россия	9
Яблоко	1
Патриоты России	1
Правое дело	< 1
Затрудняюсь ответить	14
Не пойду на выборы	13

Окончание табл.

Партия	Процент голосов поддержки (от общего числа респондентов)
Испорчу бюллетень	1
Всего	100

Найдите доверительный интервал:

а) для доли граждан, которые не готовы идти на предстоящие выборы (т. е. категория ответа «Не пойду на выборы»);

б) для доли граждан, которые готовы поддержать партию «Единая Россия»;

в) для доли граждан, которые пока окончательно не определились в своём выборе (т. е. категория ответа «Затрудняюсь ответить»);

г) исключите из рассмотрения такие категории ответов, как «Испорчу бюллетень», «Не пойду на выборы», «Затрудняюсь ответить». И снова постройте доверительный интервал для доли тех, кто готов поддержать на выборах партию «Единая Россия» (из того числа граждан, кто уже определился, за какую из предложенных партий проголосует на предстоящих выборах).

Задача 11. «Праздник 23 февраля». Приурочив исследование к празднованию 23 февраля, ВЦИОМ в 2010 г. провел блиц-опрос жителей России на тему: «Что для Вас значит 23 февраля?»[1]. Получены следующие результаты:

Вариант ответа	Количество респондентов
Праздник всех российских мужчин	656
День всех российских военнослужащих	480
День рождения Красной Армии	336
Обычный выходной	96
Затрудняюсь ответить	16
Всего	$n = 1\,600$ человек

Для ответа на вопрос, какая доля жителей РФ воспринимает этот день как обычный выходной, постройте 94%-й доверительный интервал для доли, обобщив выборочную оценку на генеральную совокупность. Категорию ответа «Затрудняюсь ответить» не исключайте из анализа.

Задача 12. «Экология планеты». Предположим, что респонденты некоторой выборочной совокупности были опрошены по установочной шкале

[1] Подробно о результатах этого исследования см.: [http://wciom.ru/index.php?id=459&uid=13195] (проверено 28.06.2012).

Лайкерта. Было выяснено, насколько они встревожены экологической проблемой глобального потепления на Земле. Результаты измеряемой установки (в виде индекса сумматорного типа) у 17 респондентов были следующими: 10, 3, 18, 15, 20, 14, 5, 19, 20, 21, 8, 30, 24, 27, 27, 22, 20 (баллов по шкале). Индекс изменялся в диапазоне от 0 до 30 баллов, где большее значение соответствует большей озабоченности и тревожности за экологическое будущее планеты. Найдите 95%-й доверительный интервал для математического ожидания индекса (т. е. истинного значения среднего). Обратите внимание на малый объём выборки при выборе статистики (z или t?). Определите также коэффициент вариации. О чём свидетельствует его значение?

Задача 13. «Пуговицы». С производственной линии, производящей фурнитуру для одежды, было отобрано 900 пуговиц, 45 из которых оказались бракованными. Оцените долю дефектных пуговиц во всей совокупности (т. е. во всей партии), определив доверительный интервал, если мера надёжности выводов должна составить не менее 90 %.

Задача 14. «Спорт». Если 14 % респондентов в возрасте 18 — 45 лет, опрошенных в ходе исследования, выборка которого была случайной и состоящей из 1 310 человек, говорят, что для них посещение бассейна (плавание) является основным видом физической нагрузки и занятий спортом, чему равно значение этой доли в генеральной совокупности? Используйте уровень доверия 80 %. Проинтерпретируйте полученный результат.

Задача 15. «Серебряная медаль Е. Плющенко». Радиостанция «Радио 7» после завершения выступлений спортсменов в мужском одиночном катании на Олимпийских играх в Ванкувере провела опрос: «Кто “виноват” в серебряной медали Евгения Плющенко?» Результаты голосования таковы:

Вариант ответа	Количество респондентов
Засудили	92
Катание не тянуло на золото	44
Искать виноватых нет смысла	50
Я не знаю/ трудно сказать	12
Всего	$n = 198$ человек

Предположим, что мы имеем дело со случайной выборкой[1]. Для ответа на вопрос, какая доля людей считает, что «искать виноватых нет смысла», постройте 95%-й доверительный интервал для определения истинного значения доли.

Задача 16. «Пульс». Постройте доверительный интервал для математического ожидания с доверительной вероятностью 0,99, если вы изучаете, как увеличивается частота пульса у людей среднего возраста после утренней пробежки и физических упражнений. Представьте, что вы располагаете

[1] Хотя это, конечно, может быть сказано с большой оговоркой, так как в основе подобных опросов чаще всего лежат так называемые выборки добровольцев.

следующими данными случайной выборки (10 наблюдений): 15, 13, 12, 7, 8, 6, 8, 14, 10, 12 дополнительных ударов в минуту. Как, в среднем, меняется пульс у одного человека, если обобщать данные на генеральную совокупность? Обратите внимание на малый объём выборки при выборе статистики (z или t?).

Задача 17. «Библиотеки». Администрация университета решила определить, сколько студентов пользуется ресурсами других библиотек для подготовки к занятиям. При пилотаже выяснилось, что из 25 студентов только 4 ходят в другие библиотеки. Выборку какого объёма необходимо взять, чтобы доверительный интервал имел ширину не более 4 %, а доверительная вероятность была равна 95 %? Обратите внимание на малый объём выборки в пилотажном исследовании при выборе статистики (z или t?).

Задача 18. «Ювелирный магазин». Фирма, решив открыть ювелирный магазин в новом районе города, хотела бы оценить долю людей, заинтересованных в продукции магазина. Компания хотела бы знать оценку генеральной доли с предельной ошибкой ±5 % и доверительной вероятностью 90 %. Для этого предполагается провести опрос жителей района. Стандартное отклонение примите равным 0,5 (т. е. 50 %). Чему равен необходимый объём выборки для исследования?

Задача 19. «Театр-2». В ходе опроса студентов факультета социологии было зафиксировано, как часто юноши и девушки посещают театр. В случайной выборке представлены ответы 54 девушек:

Количество посещений	0	1	2	3	4	5	6	7	8	9	10	20	25	Всего
Количество респондентов	10	10	14	7	3	1	2	1	1	1	1	2	1	$n = 54$

По методу Дж. Тьюки проанализируйте данные на предмет наличия в выборке нехарактерных значений. По итогам этого анализа исключите из рассмотрения статистические выбросы. Далее обобщите результаты выборки на генеральную совокупность девушек факультета социологии, построив доверительный интервал для среднего числа посещений театра.

Задача 20. «Книги». Предположим, что в результате опроса по случайной выборке 100 жителей города выяснилось, что в среднем они прочитывают 4,5 книги по художественной литературе в год. С какой максимальной вероятностью можно утверждать, что жители города читают менее 5 книг в год? Стандартное отклонение равно 2,3 книги, согласно данным этого опроса. Иными словами, требуется найти уровень доверительной вероятности.

Задача 21. «Голосование». В некотором городе запланировано проведение социологического исследования с целью определить, сколько человек пойдёт на выборы. По статистическим данным и опыту предыдущих лет подмечено, что в этом городе на выборы ходит примерно каждый четвёртый. Найдите требуемый объём выборки при предельной ошибке ±3 % и доверительной вероятности 80 % для проведения опроса общественного мнения.

Задача 22. «Мужчины, которые курят». На заводе работают 1 470 мужчин, которые курят. Генеральным директором, провозглашающим здоровый образ жизни, оценивается среднее количество выкуриваемых в день сигарет сотрудниками мужского пола. Из пилотажного исследования, выборка которого составляет 8 человек, известно:

№ интервью	1	2	3	4	5	6	7	8
Количество выкуриваемых сигарет в день, шт.	20	16	18	25	13	10	5	10

Вычислите необходимый объём выборки для исследования, чтобы точность выборки составила не менее ±2 сигареты с доверительной вероятностью 0,90.

Задача 23. «Электрички». Оценить работу общественного транспорта (электричек) социологи попросили высказать своё мнение 200 человек, из них 55 человек высказали крайне негативное мнение. Найдите доверительный интервал для доли жителей, возмущённых работой общественного транспорта. Уровень доверительной вероятности примите равным 99 %.

Задача 24. «Обед на улице». Компания ГфК-Русь в 2007 г. провела всероссийское исследование «Типология потребителя» (омнибус)[1]. Респондентам было предложено выразить степень своего согласия со следующим суждением: «Из-за отсутствия достаточного количества времени мне часто приходится посещать кафе-закусочные быстрого питания (фаст-фуд)». Распределение ответов:

«Мне часто приходится посещать фаст-фуды»	Количество респондентов
Совершенно не согласен	999
Скорее не согласен	518
Ни согласен, ни не согласен	354
Скорее согласен	240
Полностью согласен	54

Дайте выборочную оценку доли российских граждан, которые нередко пользуются фаст-фудами (т. е. категории ответов «Скорее согласен» + «Полностью согласен»), и обобщите эту оценку на генеральную совокупность, построив доверительный интервал.

---

[1] Данные Единого архива экономических и социологических данных [http://sophist.hse.ru/db/oprosy.shtml?ts=120&en=0] (проверено 28.06.2012).

Задача 25. «Отношение к рекламе». Компания ГфК-Русь в 2007 г. провела всероссийское исследование «Типология потребителя» (омнибус)[1]. Респондентам было предложено выразить степень своего согласия со следующим суждением: «Реклама мне настолько надоела, что я стараюсь её не смотреть». Распределение ответов:

«Реклама мне надоела»	Количество респондентов
Совершенно не согласен	113
Скорее не согласен	185
Ни согласен, ни не согласен	480
Скорее согласен	566
Полностью согласен	818

Дайте выборочную оценку доли российских граждан, которые положительно относятся к просмотру рекламы (т. е. категории ответов «Скорее не согласен» + «Совершенно не согласен»), и обобщите эту оценку на генеральную совокупность, построив доверительный интервал.

Задача 26. «Кухня других стран». Компания ГфК-Русь в 2007 г. провела всероссийское исследование «Типология потребителя» (омнибус)[2]. Респондентам было предложено выразить степень своего согласия со следующим суждением: «Я люблю кухню других стран мира». Распределение ответов:

«Я люблю кухню других стран мира»	Количество респондентов
Совершенно не согласен	287
Скорее не согласен	353
Ни согласен, ни не согласен	619
Скорее согласен	590
Полностью согласен	321

Дайте выборочную оценку доли российских граждан, которые положительно относятся к потреблению блюд кухни других стран мира (т. е. категории ответов «Скорее согласен» + «Полностью согласен»), и обобщите эту

[1] Данные Единого архива экономических и социологических данных [http://sophist.hse.ru/db/oprosy.shtml?ts=120&en=0] (проверено 28.06.2012).

[2] Там же.

оценку на генеральную совокупность, построив доверительный интервал для доли.

Задача 27. «Качество питьевой воды». Компания ВЦИОМ опубликовала данные всероссийского исследования «Население России о качестве питьевой воды». В частности, жителям Москвы в декабре 2007 г. был задан вопрос: «Какую воду в качестве питьевой Вы чаще всего используете (по месту постоянного жительства)?». Распределение ответов:

«Какую воду в качестве питьевой Вы чаще всего используете?»	Количество респондентов
Водопроводную некипяченую	144
Водопроводную кипяченую	420
Водопроводную, но предварительно очищенную бытовым фильтром	480
Из природных источников	12
Бутилированную	144
Затрудняюсь ответить	0
Всего	1 200

Дайте выборочную оценку доли жителей Москвы, которые предпочитают покупать бутилированную воду, и обобщите эту оценку на генеральную совокупность, построив доверительный интервал для доли.

Задача 28. «Без названия». Для каждого из представленных уровней доверительной вероятности определите соответствующее значение *t*-статистики Стьюдента:

Уровень доверительной вероятности	Объём выборки	*t*-статистика Стьюдента
0,90	31	?
0,98	24	?
0,99	12	?
0,95	42	?
0,95	7	?

Задача 29. «Покупка кухни». Используя данные случайной выборки заказов, сделанных клиентами мебельной компании, производящей кухонные интерьеры и аксессуары к ним, найдите 90%-й доверительный интервал

для средней суммы расходов покупателей на приобретение кухонной мебели у данного производителя:

№ заказа (покупатель)	1	2	3	4	5	6	7	8	9	10
Сумма заказа, тыс. руб.	142	23	41	10	17	93	17	174	220	13

Указание. Проверкой на наличие резко выделяющихся значений (нетипичных значений признака) пренебречь.

Задача 30. «Просмотры телепередач». Специалисты по маркетингу желают изучить привычки телезрителей в небольшом городке. Для этого они спроектировали случайную выборку пилотажного (предварительного) исследования, состоящую из 28 человек. Каждого участника попросили сделать подробные записи о телепередачах, которые он (она) просмотрел(-а) в течение недели. Оказалось, что среднее время просмотра телепередач равно 15,3 ч, а стандартное отклонение — 3,8 ч. Постройте 95%-й доверительный интервал, содержащий математическое ожидание продолжительности просмотра телепередач за неделю.

Решите «обратную» задачу». Выборку какого объёма следует взять исследователям, чтобы уровень надёжности статистического вывода составил 90 %, а среднее время продолжительности просмотра телепередач было бы оценено с точностью ±2 ч? (Оценку стандартного отклонения примите равной 3,8 ч.)

Глава 6

ПРОВЕРКА СТАТИСТИЧЕСКИХ ГИПОТЕЗ

В данной главе мы познакомимся с ещё одним способом переноса результатов выборочного исследования на генеральную совокупность — проверкой статистических гипотез. При этом будут введены новые понятия и изучены новые средства анализа имеющейся информации.

§ 6.1. Основные понятия проверки статистических гипотез. Глоссарий

Неоднократно подчёркивалось, что при изучении различных социальных явлений истинное положение дел часто сокрыто от нас, недоступно нашему знанию или опыту. Мы судим о вещах «в общем», часто опираясь лишь на фрагмент данных об окружающем мире[1]. Представления о чём-то нередко складываются на основе наблюдений, сделанных в ходе *выборочных* исследований, охватывающих *некоторую небольшую часть* генеральной совокупности. Однако полученным эмпирическим фактам по итогам исследования мы всё же стремимся придать характер закономерностей, справедливых для генеральной совокупности в целом.

В этом параграфе рассмотрим ещё один способ обобщения выборочных данных на генеральную совокупность, введём важное для нас понятие — понятие статистической гипотезы, договоримся об основной терминологии, которой нам предстоит пользоваться, и раскроем общие принципы проверки гипотез.

Статистической гипотезой будем называть некоторое *предположение* (догадку, суждение) о том, каково истинное распределение вероятностей, характерное для конкретной выборки наблюдений из генеральной совокупности:

---

[1] Одним из таких фрагментов, в частности, является выборка.

• *во-первых*, каков вид этого предполагаемого (т. е. неизвестного нам достоверно) распределения;

• *во-вторых*, каковы предполагаемые величины параметров распределения (выбранного (гипотетического) распределения), каковы свойства этих параметров.

Например, в качестве статистической гипотезы могут выступить следующие суждения.

1. ***Генеральная совокупность данных имеет нормальное распределение.*** *Э*то предположение/догадка о том, каков может быть истинный закон распределения данных (случайной величины). ***Содержательный пример:*** В качестве гипотезы (предположения) может выступить утверждение «***Стоимость 1 кг яблок определённого сорта, продаваемых в торговых палатках города, описывается нормальным распределением***». В этом примере случайной величиной является стоимость (цена) 1 кг яблок. Закон, которым, как мы полагаем, описывается распределение вероятностей случайной величины, является нормальное распределение. Генеральная совокупность представлена всеми торговыми палатками города, продающими яблоки данного сорта. Выборочная совокупность — только теми торговыми палатками, которые попали в выборку нашего исследования, и стоимость яблок в которых была изучена, зафиксирована в определённый момент времени. Охватить исследованием всю генеральную совокупность не представляется возможным (по разумным соображениям). Поэтому, судя по данным выборки, мы и стремимся понять, насколько справедливой является гипотеза в реальности. Каковы её шансы на достоверность.

2. ***Математические ожидания обеих выборок, взятых из нормально распределённых генеральных совокупностей, равны друг другу.*** Это предположение/догадка о том, каковы параметры известного закона распределения — в данном случае нормального; параметрами (т. е. характеристиками этих распределений) здесь выступают математические ожидания, т. е. истинные средние значения. ***Содержательный пример:*** рассмотрим случайную величину — количество электрической энергии, расходуемой домохозяйством в течение одних суток (в расчёте на одного жильца). Каждое из действий того или иного жильца (члена семьи/домохозяйства) порождает потребление электроэнергии: периодическое включение электроплиты, компьютера, настольной лампы, телевизора, фена, стиральной машины и других электроприборов. Заметим при этом, что одними приборами люди могут пользоваться постоянно и с одинаковой (или разной) длительностью, а какими-то — лишь время от времени, изредка. Поэтому расход электроэнергии в течение суток для каждого домохозяйства — величина случайная (колеблющаяся, нефиксированная). Следовательно, и суточный расход электроэнергии в расчёте на одного человека в домохозяйстве — также. Значение этой случайной величины для конкретных суток

является результатом независимой работы различных электроприборов в течение этих суток (арифметической суммой всех расходов электроэнергии, затраченной домохозяйством на функционирование каждого из электроприборов)[1]. Поэтому есть разумные основания считать, что распределение ежедневных расходов на потребление электроэнергии (в расчёте на одного члена домохозяйства) адекватно описывается моделью нормального распределения[2]. Напомним, что у нормального распределения есть два ключевых параметра, которые *полностью* определяют его положение и форму — это математическое ожидание (среднее) и дисперсия[3].

Пусть мы стремимся понять, одинаковое ли количество электроэнергии расходуют, в среднем, жители крупных городов (численностью населения 500 — 1 000 тыс. человек) и городов средней величины (с численностью населения около 250 — 500 тыс. человек) — одинаковые ли у тех, и у других нормы потребления электроэнергии. Пусть на основе имеющейся случайной выборки данных мы намерены понять, можно ли считать равными между собой средние значения этой величины для обеих генеральных совокупностей? Статистическая гипотеза будет сформулирована так: «***Средний ежедневный расход электроэнергии (в расчёте на одного члена домохозяйства) у жителей крупных и средних городов одинаков***» (т. е. нормы (запросы) в потреблении равны).

3. ***В генеральной совокупности дисперсии обеих выборок, взятых из нормально распределённых генеральных совокупностей, равны, одинаковы.*** Это также предположение о том, каковы параметры известного закона распределения — в данном случае нормального; параметрами здесь выступают дисперсии. ***Содержательный пример:*** предположим, что два спортсмена, занимающихся стрельбой из пневматической винтовки, показали на протяжении многократных тренировок перед Олимпийскими играми свои результаты — проценты точных выстрелов по мишени. Пусть известно, что процент точных выстрелов — случайная величина, имеющая нормальное распределение[4]. Аттестационная комиссия стремится понять, демонстрируют ли оба спортсмена одинаковое постоянство/стабильность (т. е. одинаковость вариации/изменчивости) в своих результатах. Статистическая гипотеза примет вид: «***Дисперсия процента точных выстрелов первого спортсмена такая же, как и дисперсия процента точных попаданий у второго спортсмена***» (у обоих спортсменов результаты одинаково стабильны).

---

[1] Значение (и учёт) этой величины в течение суток фиксируется счётчиком.

[2] Как распределение суммы независимых случайных величин.

[3] Чтобы указать нормальное распределение исследуемой случайной величины, в первую очередь требуется задать именно эти два параметра, оценки которых можно взять из выборочных обследований.

[4] Если выстрелов достаточно много, то это вполне допустимое предположение, основанное на результатах теоремы Муавра — Лапласа.

4. ***Математическое ожидание выборки, взятой из нормально распределенной генеральной совокупности, равно некоторой константе (числу).*** Наряду с приведёнными ранее формулировками это предположение о том, каково истинное значение параметра известного закона распределения — в данном случае нормального; параметром здесь выступает истинное среднее. ***Содержательный пример***: допустим, менеджер магазина модной обуви стремится понять, каково среднее время пребывания посетителей в торговом зале магазина. В течение дня организовав выборочное наблюдение за несколькими посетителями магазина (объём случайной выборки 25 человек), менеджером было зафиксировано среднее время их пребывания в магазине (допустим, по выборке было получено, что среднее равно 34 мин, а дисперсия составила 65 (мин)^2). Менеджер считает работу своего торгового персонала эффективной в том случае, если посетитель находится в зале, в среднем не менее 45 мин (это означает, что его удаётся заинтересовать коллекцией, предложить отвечающие его предпочтениям и потребностям модели обуви, а, следовательно, пригласить к примерке и подбору подходящего ассортимента). На основании этих разумных предположений, менеджер стремится оценить, а можно ли, тем не менее, считать это время (оценённое по выборке как 34 мин) равным 45 мин (в генеральной совокупности[1]). Или оно на самом-то деле меньше указанного значения 45 мин? ***«Математическое ожидание времени нахождения посетителя в торговом зале магазина обуви равно 45 мин»*** — пример формулировки статистической гипотезы.

Важно понимать, что формулировки гипотез, не приводящие к предположению о форме и/или величинах параметров распределений, не относятся к статистическим гипотезам. Так, распространённое в астрономии суждение «Началу Вселенной послужил Большой взрыв» нельзя считать статической гипотезой, так как указанная формулировка не отвечает определению статистической гипотезы.

Словосочетание ***«проверка статистической гипотезы»*** предполагает процесс принятия решения о том, в какой мере статистическая гипотеза выглядит правдоподобной (истинной) для имеющейся выборки наблюдений. Решением станет вывод о том, в какой степени гипотеза и имеющиеся данные соотносятся между собой и не противоречат ли друг другу; насколько выдвинутое предположение согласуется с теми эмпирическими данными, которые имеются в распоряжении по результатам выборочного исследования.

Простая статистическая гипотеза — это такая гипотеза, которая однозначно задаёт распределение рассматриваемых данных. Например: «Истинная доля людей, которые умеют кататься на велосипеде, равна 0,48».

---

[1] Под генеральной совокупностью здесь могут подразумеваться все потенциальные посетители этого магазина.

Сложная статистическая гипотеза — это такая гипотеза, которая задаёт распределение рассматриваемых данных *с точностью до некоторого множества* параметров этого распределения, и может включать в себя большое число (в ряде случаев бесконечное) простых гипотез. Например: «Истинная доля людей, которые умеют кататься на велосипеде, больше 0,48». В этом случае может быть осуществлена декомпозиция («расщепление») сложной гипотезы на множество простых. В данном примере этими простыми гипотезами могут стать предположения: «Истинная доля равна 0,49», «Истинная доля равна 0,51», «Истинная доля равна 0,78», «Истинная доля равна 0,785» и т. д. — любое число, превышающее 0,48, но не превышающее 1 (100 %).

Нулевой гипотезой (или *нуль-гипотезой*) будем называть *основное* (главное) *предположение*, выдвинутое для статистической проверки: проверке на соответствие имеющихся эмпирических данных выборки содержащемуся (указанному) в гипотезе суждению.

Нулевая гипотеза обозначается H_0 и произносится: «Аш ноль». *Нулевые гипотезы чаще всего являются простыми статистическими гипотезами*. Нулевая гипотеза может быть как истинной, так и ложной. При формулировании нулевой гипотезы нам заранее не известно, верна ли она в действительности или является заблуждением. Мы соотносим эмпирические данные с этой гипотезой и пытаемся дать статистическую оценку вероятности того, что нулевая гипотеза в реальности верна. Так, достоверность нулевой гипотезы проверяется с помощью методов математической статистики.

Альтернативной (конкурирующей, противоположной) гипотезой будем называть *утверждение*, *противоречащее* нулевой гипотезе. Конкурирующая гипотеза обозначается как H_1 и читается как «Аш один». Альтернативная гипотеза статистической проверке ***никогда*** *не подвергается*: она либо принимается, либо отвергается в зависимости от того, каким окажется результат статистической проверки нулевой (основной) гипотезы. Чаще всего нулевые гипотезы непосредственно указывают предполагаемое распределение данных. Но бывают случаи, когда нулевая гипотеза формулируется в более общем виде (табл. 6.1).

Таблица 6.1. **Различные формулировки нулевых статистических гипотез и альтернативных им**

Примеры формулировки нулевой гипотезы	Нулевой гипотезе можно противопоставить одну из таких формулировок альтернативной гипотезы
• «Коэффициент линейной корреляции между уровнем преступности и уровнем безработицы ***равен 0,6***»	• «Коэффициент линейной корреляции между уровнем преступности и уровнем безработицы ***не равен 0,6***»

Продолжение табл. 6.1

Примеры формулировки нулевой гипотезы	Нулевой гипотезе можно противопоставить одну из таких формулировок альтернативной гипотезы
	• «Коэффициент линейной корреляции между уровнем преступности и уровнем безработицы ***строго больше 0,6***» • «Коэффициент линейной корреляции между уровнем преступности и уровнем безработицы ***строго меньше 0,6***»
• «***Нет линейной зависимости*** между посещаемостью лекций и суммой баллов, набранных студентом II курса факультета социологии на экзамене по математической статистике»	• «Существует ***положительная линейная зависимость*** между посещаемостью лекций и суммой баллов, набранных студентом II курса факультета социологии на экзамене по математической статистике» • «Существует ***отличная от нуля линейная зависимость*** между посещаемостью лекций и суммой баллов, набранных студентом II курса факультета социологии на экзамене по математической статистике»
• «Средняя стоимость[1] (математическое ожидание) классической мужской стрижки в парикмахерских города ***составляет 450 руб.***»	• «Средняя стоимость (математическое ожидание) классической мужской стрижки в парикмахерских города ***не равна 450 руб.***» • «Средняя стоимость (математическое ожидание) классической мужской стрижки в парикмахерских города ***меньше 450 руб.***» • «Средняя стоимость (математическое ожидание) классической мужской стрижки в парикмахерских города ***больше 450 руб.***»
• «Успешность игры на фортепиано ***не имеет*** заметных ***различий*** у тех учеников, кто начал обучение в музыкальной школе в 6—7 лет, и тех, кто начал обучение в 8—9 лет»	• «Успешность игры на фортепиано заметно ***различается*** у тех учеников, кто начал обучение в музыкальной школе в 6—7 лет, и тех, кто начал обучение в 8—9 лет»

[1] Стоимость классической мужской стрижки — случайная величина, имеющая нормальное распределение с заданным средним (математическим ожиданием) 450 руб. и неизвестной дисперсией.

Окончание табл. 6.1

Примеры формулировки нулевой гипотезы	Нулевой гипотезе можно противопоставить одну из таких формулировок альтернативной гипотезы
	• «Успешность игры на фортепиано ***выше*** у тех учеников, кто начал обучение в музыкальной школе в 6—7 лет, чем у тех, кто начал обучение в 8—9 лет» • «Успешность игры на фортепиано ***ниже*** у тех учеников, кто начал обучение в музыкальной школе в 6—7 лет, чем у тех, кто начал обучение в 8—9 лет»
• «***Зависимости*** между видом применённой методики лечения[1] (*A*, *B* или *C*) некоторого заболевания и появлением улучшений в состоянии здоровья пациента ***нет*** (***не наблюдается***)» [1] Применённая методика — это лечение заболевания одним из медицинских препаратов.	• ***«Зависимость*** между видом применённой методики лечения (*A*, *B* или *C*) некоторого заболевания и появлением улучшений в состоянии здоровья пациента ***есть*** (***наблюдается***)»
• «Мужчины и женщины в возрасте 25—35 лет тратят на занятия спортом /физкультурой[2] в среднем ***одинаковое количество часов*** в неделю (средние показатели затрат времени на спорт у них равны)» [2] Затраты времени на спорт в течение недели у мужчин и женщин — случайная величина, имеющая нормальное распределение с заданным средним и дисперсией, которые оценены по выборке.	• «Мужчины и женщины в возрасте 25—35 лет тратят на занятия спортом/физкультурой в среднем ***разное количество часов*** в неделю» • «Мужчины в возрасте 25—35 лет тратят на занятия спортом/физкультурой в среднем ***большее количество часов*** в неделю, нежели женщины того же возраста (средний показатель затрат времени у первых больше, чем у вторых)» • «Мужчины в возрасте 25—35 лет тратят на занятия спортом/физкультурой в среднем ***меньшее количество часов*** в неделю, нежели женщины аналогичного возраста (средний показатель затрат времени у вторых больше, чем у первых)»

Статистическим критерием проверки нулевой гипотезы выступает определенное *правило* (т. е. алгоритм), согласно которому с высокой долей вероятности можно подтвердить или, наоборот, опровергнуть проверяемую нулевую гипотезу. В основе этого правила обычно лежит функция от наблюдений (её называют *статистикой критерия*). Статистика, как функция от наблюдений, сама является величиной ***случайной***, потому что в различных выборочных исследованиях она может принимать разные (меняющиеся от одной выборки к другой) значения, заранее неизвестные и не поддающиеся однозначной оценке. ***Распределение статистики критерия*** (т. е. распределение вероятностей всех заведомо возможных значений, которые она может принять) ***задается нулевой гипотезой***.

При проверке нулевой гипотезы мы сталкиваемся с тем, что можем допустить два типа ошибок (два типа заблуждений):

- ***ошибку I рода***: она возникает в том случае, когда по итогам статистической проверки нулевая гипотеза объявляется ложной (в противовес тому, что она-то как раз и верна в реальности). Признавая ошибочность нулевой гипотезы, мы обманываемся, а наше знание будет ложным;
- ***ошибку II рода***: она заключается в том, что по итогам статистической проверки нулевая гипотеза объявляется верной, справедливой (хотя в действительности не верна). Признавая истинность нулевой гипотезы, мы вновь обманываемся, и наше знание также будет ложным.

Эти два типа ошибок ***разные*** — за ними стоят ***разные*** потенциальные риски и последствия от ошибочно сделанных выводов. Рассмотрим пример, иллюстрирующий эти два типа статистической ошибки.

Пример 6.1. «Экзаменационный тест». Пусть в экзаменационном тесте по некоторому предмету предложено 10 вопросов с четырьмя вариантами ответа на каждый. Пусть тест выполняет некий студент и нам не известно (нет информации)[1], кто перед нами — сильный студент или же студент, выбирающий ответы наугад (т. е. не знает предмет, не готовился[2]). Мы хотим понять, разумно ли считать границей удовлетворительной (положительной) оценки за тест пять верно выполненных тестовых заданий? Пусть по опыту нам известно, что сильный студент даёт правильный ответ в подобном тесте с вероятностью 0,75 (следовательно, с вероятностью 0,25

[1] В ряде случаев бывает, что в экзаменационной работе не указываются персональные данные студента (Ф.И.О), а присутствует лишь некоторый уникальный идентификационный номер. Это делается с той целью, чтобы процесс проверки работ был беспристрастным и свободным от субъективных решений экзаменатора.

[2] Тогда он с одинаковой вероятностью выбирает один из четырех предложенных вариантов ответа в каждом тестовом задании.

при ответе на вопрос он ошибается). Если студент не знает предмет и отвечает наугад, то вероятность верного ответа на каждый из вопросов теста равна $^1/_4 = 0{,}25$. Вероятность же неверного ответа составит $^3/_4 = 0{,}75$.

Чтобы принять решение о границе положительной оценки (а, следовательно, и одобрительного решения в аттестации знаний), можно проверить нулевую гипотезу следующего вида: **H_0: «*Перед нами студент, выбирающий правильный ответ наудачу* (*с вероятностью верного ответа в тестовом задании* $P = 0{,}25$). *Не готовился к экзамену*».** Альтернативная гипотеза, которая может быть противопоставлена нулевой гипотезе, будет сформулирована так: **H_1: «*Перед нами сильный студент, выбирающий правильный ответ обдуманно* (*с вероятностью верного ответа в тестовом задании* $P = 0{,}75$). *Готовился к экзамену*»**. В качестве статистики критерия для проверки нулевой гипотезы рассмотрим ***число правильных ответов***. Напомним, что статистикой критерия выступает некоторая функция от наблюдений (т. е. от исходных данных), справедливая для проверяемой гипотезы H_0.

Ответ на каждый из тестовых заданий является величиной бинарной, т. е. принимающей только два значения, — ответ либо верен, либо нет. Выполнение каждого задания завершается либо «успехом», либо «неудачей», а сам процесс выполнения экзаменационного теста превращается для студента в последовательность (серию) испытаний Бернулли, состоящую из 10 независимых «опытов» (так как в тесте 10 вопросов). Найдём закон биномиального распределения случайной величины ***k* — *число верно выполненных тестовых заданий*** (для каждой из двух ситуаций)[1]. Для этого рассчитаем вероятность для каждого из возможных значений числа правильно выполненных заданий в тесте:

1) исходя из предположения, что верна гипотеза H_0,

2) исходя из предположения, что верна гипотеза H_1.

Если верна гипотеза H_0, то получим следующее распределение вероятностей числа верно выполненных заданий в тесте:

k	0	1	2	3	4	5	6	7	8	9	10
Вероятность	0,056	0,188	0,282	0,250	0,146	0,058	0,016	0,003	0,0004	0,00003	0,00...

Если верна гипотеза H_1, имеем следующее распределение вероятностей числа верно выполненных заданий в тесте:

[1] Воспользуемся формулой Бернулли для вычисления вероятности события, когда число успехов равно k: $P(k) = C_n^k p^k q^{n-k}$, где n — число опытов (испытаний) — в данном случае число тестовых заданий; p — вероятность «успеха» (вероятность верного ответа на тестовое задание); q — вероятность «неудачи» (вероятность ошибочного ответа на тестовое задание); C_n^k — общее число способов дать верный ответ ровно на k тестовых заданий.

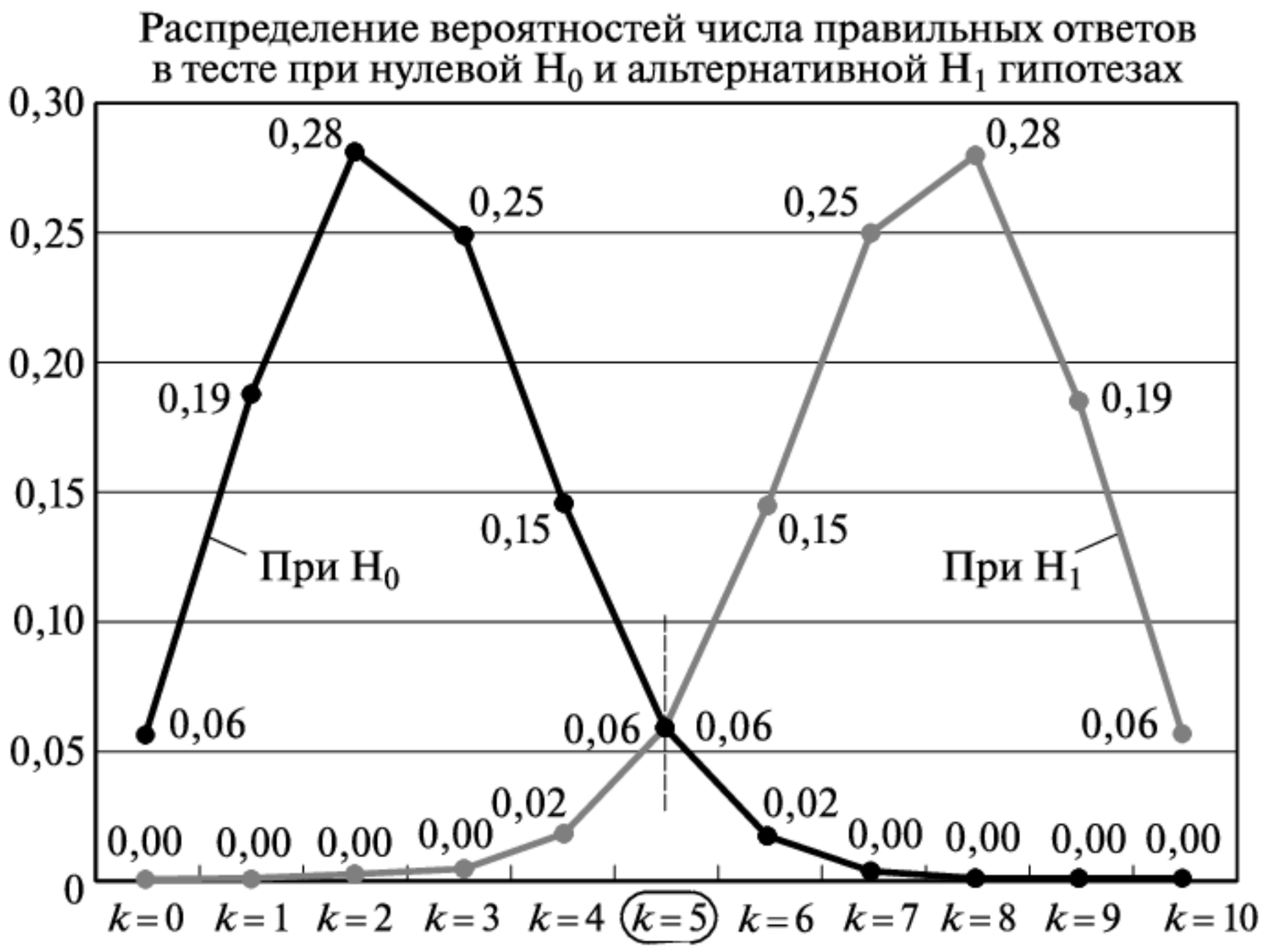


Рис. 6.1. Распределение вероятностей при гипотезах H_0 и H_1 (см. пример 6.1 «Экзаменационный тест»)

k	0	1	2	3	4	5	6	7	8	9	10
Вероятность	0,00...	0,00003	0,0004	0,003	0,016	0,058	0,146	0,250	0,282	0,188	0,056

Представим графически найденные законы распределения вероятностей для числа ***k*** (рис. 6.1).

Важно понимать, что процесс принятия преподавателем решения о том, а какое число (*как минимум*) верно выполненных заданий в тесте можно считать удовлетворительным (положительным) результатом, должен стать далеко *не поверхностным*. Посмотрим, каковы могут быть риски («плата») за неверно (или поспешно) принятое решение. На первый взгляд кажется, что границей удовлетворительной оценки вполне может стать число верных ответов, равное пяти. Однако принятие преподавателем решения о границе удовлетворительной оценки сопряжено с двумя рисками (ошибками): они совершенно различны по своим последствиям. Практическая значимость этих ошибок разная. Посмотрим, что может произойти вследствие некорректно принятого решения относительно H_0.

Если провести границу на уровне пяти правильно выполненных заданий, то вероятность выполнить пять заданий при совершенно разных гипотезах: а) студент отвечает наудачу и б) студент знает предмет — одинаковая (!) и составляет около 6 %.

Если в реальности гипотеза H_0 *истинна*, то проводя границу на уровне пяти (как «зачтено»), мы рискуем «пропустить» человека, который на самом деле предмет не знает. Это решение будет сопря-

жено со следующими последствиями: мы волей-неволей позволяем случиться тому, что будет аттестована часть студентов, в реальности не имеющих должной квалификации и профессиональных навыков (они просто угадали пять ответов!). Тогда, не «вынуждая» впоследствии этих студентов пересдать предмет, мы в ряде случаев (хотим того или нет) «поощряем» такой исход: деквалификацию в образовании. В ряде случаев можно понести риски от такого решения. Интуитивно каждый из нас ждёт друг от друга профессионализма: мы хотим, чтобы на прилавках магазинов продавались свежие продукты, в больницах предлагалась качественная медицинская помощь, в университетах имелись хорошие библиотеки, на улицах города — чистота и безопасность[1]. И чтобы работа выполнялась «на совесть», и чтобы каждый на своём месте демонстрировал профессионализм. Конечно, рано или поздно время и фактические результаты работы многое расставляют на свои места, но, всё же, какой-то части людей можно нанести урон из-за этой ошибки. И речь здесь идет не только о репутации.

В данном контексте, отклоняя нулевую гипотезу (о том, что тест выполнен студентом, который не знает предмет) в пользу альтернативы (о том, что эта работа на самом-то деле выполнена сильным студентом), **мы совершаем ошибку I рода.**

Если же провести границу на уровне шести верно выполненных заданий (т. е. поднимем планку и уровень требований), чтобы заметно сократить вероятность ошибки I рода (чтобы зачёт не получили студенты, не имеющие на самом деле должной квалификации, т. е. отвечающие наугад), то какую еще ошибку можно совершить в данном случае? В данном случае возрастает риск, что не будет поставлен зачёт сильному студенту. Если представить, что перед нами на самом деле подготовленный студент, ответственно и добросовестно подходивший к изучению предмета, но по каким-то причинам плохо написавший этот тест (например, верно выполнивший только пять вместо минимальных для зачёта шести заданий). Причины эти могут быть разными: плохое самочувствие и проблемы со здоровьем, пережитый накануне экзамена сильный стресс, эмоциональное «выгорание».., то каковы последствия такого решения? — принятия нулевой гипотезы, которая в реальности не верна (мы ошибочно признаём H_0 справедливой). Риски следующие: незачётом можно спровоцировать нервный срыв у студента, «отбить» желание учиться и совершенствоваться, сильно огорчить; а в ряде случаев (в некоторых университетах), поставив незачёт, «отнять» у студента скидку на оплату обучения (например, при наличии неудовлетворительной оценки по одному из предметов). Невозможность финансовой компенсации дальнейшего обучения со стороны семьи/близких может вообще привести к отчислению из университета весьма подготов-

---

[1] Это касается ***всех без исключения профессий***.

ленного (ориентированного на получение знаний) студента. И на карту оказывается поставленной дальнейшая судьба и жизненная траектория человека.

В данном контексте, принимая нулевую гипотезу как истинную (т.е. принимая работу сильного студента за работу неподготовленного, и производя такую ложную «подмену»), ***мы совершаем ошибку II рода***.

Подведём важные итоги рассмотренного примера, с помощью которого мы стремились проиллюстрировать два типа статистической ошибки. Чтобы преподавателю провести границу удовлетворительной (положительной) оценки за экзаменационный тест, ему прежде всего нужно для себя принять решение о том, ***какая из ошибок (I или II рода)*** для него ***менее желательна***. И этот вопрос в математической статистике является принципиальным, очень важным. От этого и будет зависеть, каким станет итог проверки нулевой гипотезы — будет ли она принята или отклонена в пользу альтернативы. Необходимо ответить на вопросы: чем боимся поступиться в большей степени, какого последствия опасаемся больше? Совершение какой ошибки вызывает у нас бóльший страх, бóльшие опасения, бóльшее попечение, какова так называемая «цена вопроса», что имеем на чаше весов? Ответы на эти вопросы и диктуют аналитику, *какой* уровень значимости выбрать при проверке статистической нулевой гипотезы. О понятии «уровень значимости» пойдёт речь далее более подробно (уже на основе рассмотренного ранее примера).

Говоря об ошибке I рода, мы чаще пользуемся словосочетанием «Уровень значимости». Как было отмечено ранее, это вероятность отвергнуть нулевую гипотезу, по итогам статистической проверки объявив её ложной, которая на самом деле верна. Другими словами, это ***вероятность ошибочно отклонить верную нуль-гипотезу***, ***риск принять неверное решение.*** Это вероятность нашего ошибочного аналитического заключения (заблуждения) о том, что нулевая гипотеза в реальности ложна. Такая ошибка возникает вследствие того, что аналитик, опираясь на данные выборочного исследования, усомнится в правдоподобности нулевой гипотезы (и, как следствие, отвергнет её), хотя именно она и является истинной (верной) в генеральной совокупности (но представление о которой попросту недоступно его знанию). Может сложиться так, что данные выборки будут таковы, что не убедят аналитика в необходимости считать истинной нулевую гипотезу. В результате чего им и будет отвергнута верная нулевая гипотеза. Такое случается редко, но всё же может иметь место (например, в тех случаях, когда методология исследования имеет существенные ограничения: под вопросом оказывается качество используемых методик изучения, или когда в выборке представлено малое число наблюдений и она не отличается высокой репрезентативностью, т.е. погрешность выборки высока).

Обозначается уровень значимости буквой α («альфа») и равен: $(1-\gamma)$, где γ — уровень доверительной вероятности, который выбирается как мера надёжности (достоверности) выводов, делаемых в ходе статистической проверки. Буквой γ («гамма») обозначена вероятность принятия (подтверждения) нулевой гипотезы, при условии, что она и в действительности на самом-то деле верна. Чаще всего в практике исследований используются три уровня значимости: 5 % (для доверительной вероятности 95 %), 1 % (для доверительной вероятности 99 %) и 0,1 % (для доверительной вероятности 99,9 %). Таким образом, уровень значимости — это та *степень ошибки, которую мы считаем* ***допустимой*** **(*приемлемой*)** *при статистической проверке нулевой гипотезы.*

Область принятия и область отклонения нулевой гипотезы. Статистика критерия, как и любая случайная величина, может попадать в те или иные зоны своих значений. При этом одни зоны этих значений могут иметь большую вероятность (т. е. статистика критерия должна попадать в них часто или очень часто), а другие зоны этих значений — малую вероятность (т. е. статистика критерия должна попадать в них редко или крайне редко). Например, если статистика критерия имеет стандартное нормальное распределение, то нам известно, что зона значений стандартного нормального распределения, сконцентрированная около 0 (нуля), пусть, на отрезке [–2; +2] имеет высокую вероятность (около 95 %[1]), а зоны значений на отрезках [–∞;–2] и [2;+∞] имеют малую вероятность при нулевой гипотезе. Исходя из сказанного, это позволяет разбить всё множество значений статистики на две области — ***область принятия*** нулевой гипотезы и область её отклонения. Область принятия нулевой гипотезы (её ещё называют ***областью характерных значений***) состоит из такого множества значений статистики критерия, которые имеют *близкую к единице* вероятность встречаемости. ***Область отклонения*** нуль-гипотезы носит специальное название — ***критическая область***. По своему смыслу критическая область — множество (зона) тех значений статистики, вероятность которых при нулевой гипотезе мала (меньше выбранного уровня значимости или равна ему). Критическая область охватывает такие значения статистики критерия, которые являются ***маловероятными***, ***нетипичными***, ***нехарактерными***, ***редкими***. Критическая область — это интервал или несколько (обычно два) интервалов, если статистика критерия является непрерывной случайной величиной. В случае, когда статистика критерия — дискретная случайная величина — критическая область состоит из некоторого набора отдельных значений этой случайной величины. Для любой случайной величины (как дискретной, так и непрерывной) можно указать много маловероятных зон значений. Какие же тогда из этих зон мы стремимся отнести к критической об-

[1] Так называемое правило «трёх сигм» (или правило «68 — 95 — 99»).

ласти? А ровно те, вероятность попадания в которые ***с точки зрения альтернативной гипотезы*** велика.

Чтобы прояснить эту идею, ***вернёмся к примеру 6.1 «Экзаменационный тест»***, где обсуждался вопрос о том, на уровне какого значения правильно выполненных тестовых заданий целесообразно провести границу удовлетворительной (положительной) оценки. При истинности H_0 мы будем стремиться «отодвинуть» эту границу как можно дальше/правее (чтобы постараться заметно сократить вероятность удовлетворительной аттестации студента, на самом деле не знающего предмет). Обратите внимание, что в случае, если эта граница пройдёт на уровне пяти верно выполненных заданий, то вероятность будет равна сумме вероятностей того, что студент верно выполнит 5, 6, 7, 8, 9 или даже 10 заданий. Вероятность такого события будет равна: 0,058 + 0,016 + 0,003 + 0,0004 + 0,00003 + + 0,0000.. = 0,07743 (примерно ***0,077***, ***т.е. около 8 %***). В том случае, если границу удовлетворительной оценки провести на уровне шести верно выполненных тестовых заданий, то эта вероятность будет заметно меньше: 0,016 + 0,003 + 0,0004 + 0,00003 + 0,0000.. = 0,01943 (примерно ***0,019***, ***т.е. около 2 %***). В первом случае критическую область составят значения статистики критерия, принадлежащие отрезку [5;10], во втором — отрезку [6;10]. Поэтому, если у нас имеются весомые «подозрения» («ощущения», основания) что в реальности верна нулевая гипотеза, то при проверке статистической гипотезы разумно выбрать такую критическую область, которую составят целые (дискретные) значения статистики критерия, принадлежащие отрезку [6;10], а не [5;10]. И при этом видим, что к критической области мы стремимся отнести такие значения статистики критерия, вероятность попадания в которые ***с точки зрения альтернативной гипотезы*** H_1 велика. Для сильного студента вероятность выполнить не менее шести тестовых заданий (т.е. 6, 7, 8, 9 или 10) по-прежнему (как и в случае выполнения не менее пяти заданий) крайне высока и составляет: 0,146 + 0,250 + 0,282 + 0,188 + 0,056 = 0,922 (примерно ***0,92***, ***т.е. около 92 %***, рис. 6.2).

Для нахождения критической области (границ интервалов) бывают полезны статистические таблицы, в которых указано, с какой вероятностью интересующая нас статистика критерия попадает в ту или иную область значений. Однако совсем не обязательно заранее строить критическую область. Можно сначала по имеющейся выборке рассчитать наблюдаемое значение статистики критерия, а потом уже по этому значению вычислить так называемый ***«минимальный уровень значимости»*** (подробнее см. далее) и принять дальнейшее решение.

Если наблюдаемое значение статистики критерия находится в критической области, то истинность (правдоподобность, убедительность) нулевой гипотезы оказывается под сомнением и нулевую гипотезу не принимают, объявляя её ложной. ***Например***, ***если некоторый***

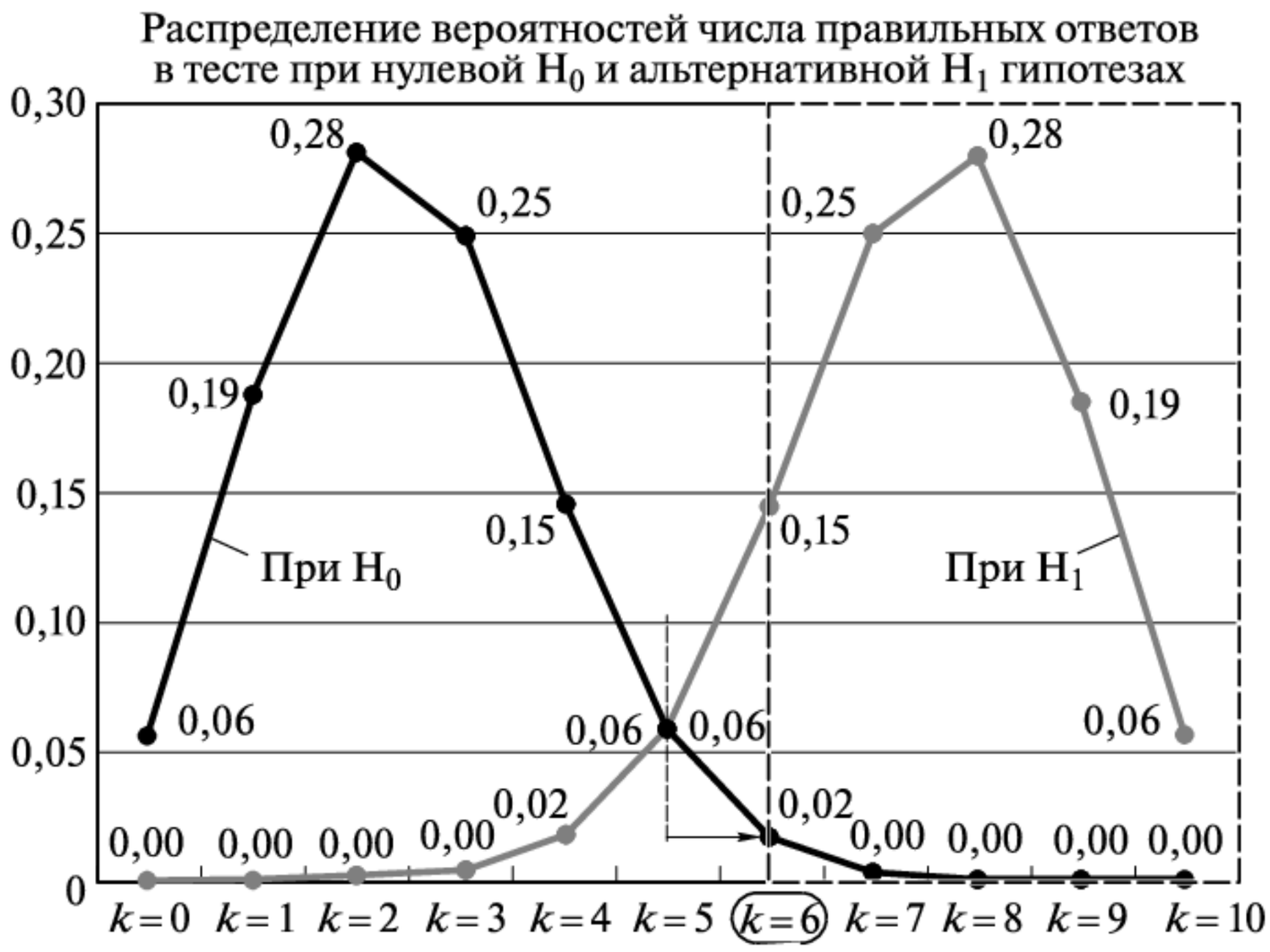


Рис. 6.2. Выбор критической точки (см. пример 6.1 «Экзаменационный тест»)

студент при написании экзаменационного теста правильно выполнил семь заданий, мы скорее поверим в то, что перед нами сильный студент, нежели поверим в то, что студент отвечал на вопросы теста наугад. Другими словами, нулевая гипотеза не покажется нам убедительной при таких данных. И мы будем в большей мере склонны её отвергнуть в пользу альтернативы, т. е. не принимать.

Точками, которые разделяют всё множество значений статистики критерия на части — область принятия нулевой гипотезы и область её отклонения — называют ***критическими (пограничными) точками***.

Принято выделять ***несколько типов критической области***: ***левостороннюю***, ***правостороннюю и двустороннюю***. Левосторонняя критическая область включает в себя такой диапазон значений статистики критерия, которые меньше его критической (пограничной) точки; правосторонняя критическая область охватывает такие значения статистики критерия, которые превосходят его критическую (пограничную) точку. Двусторонняя критическая область состоит из двух критических областей — левосторонней и правосторонней одновременно (рис. 6.3).

Так, при проверке основной (нулевой) гипотезы аналитиком *оценивается вероятность* (*шанс*) *попадания значения статистики критерия* (*наблюдаемой*, *рассчитанной на имеющейся выборке данных*) *в определенный промежуток на графике плотности распределения вероятностей этой статистики критерия*. И далее определяется, в какой мере полученное значение критерия является характе́рным (типичным) для данного распределения, соответству-

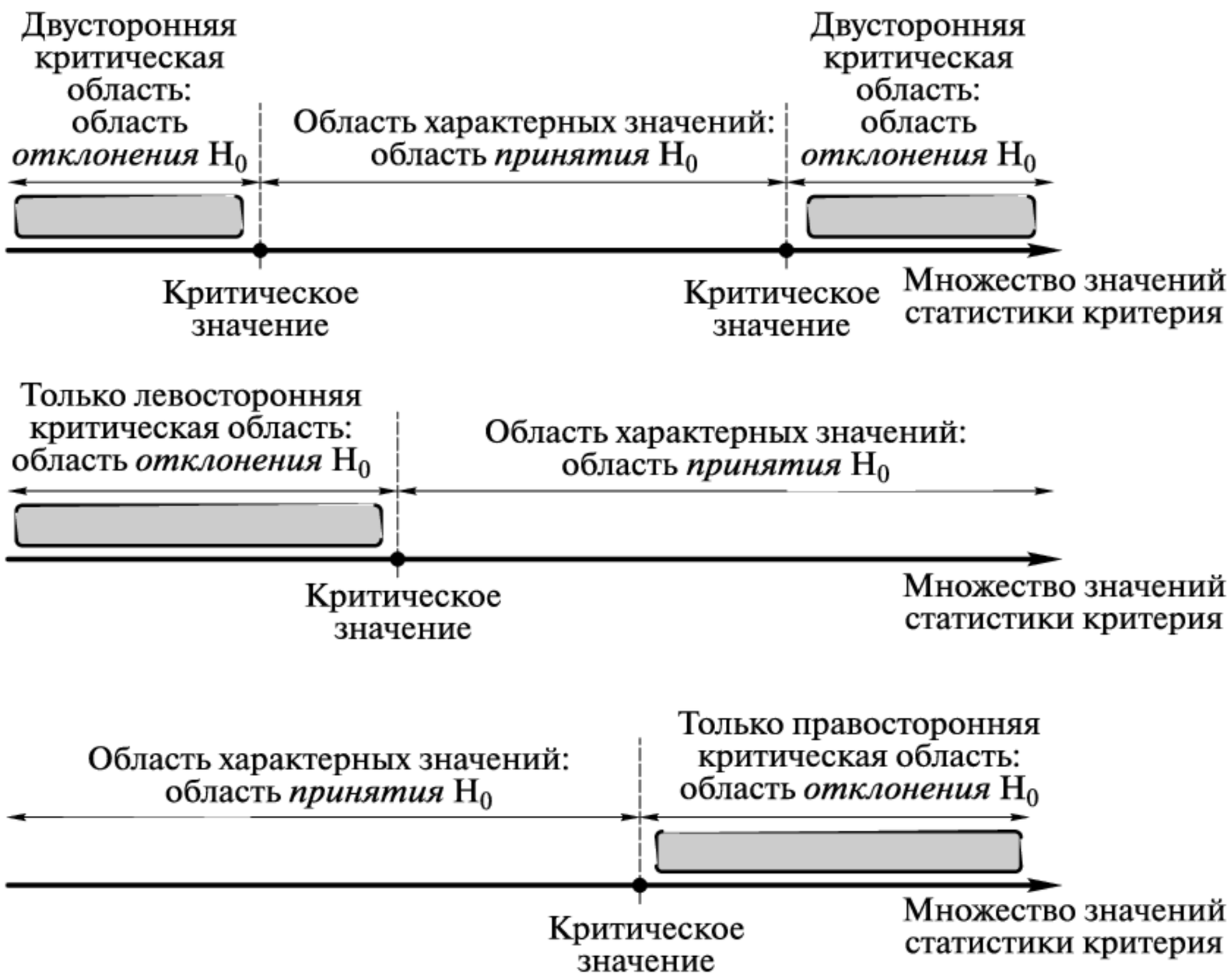


Рис. 6.3. Двусторонние и односторонние критические области

ющего нулевой гипотезе. Исходя из этого, делается финальное заключение об истинности или ложности нулевой гипотезы.

Вообще ***при тестировании статистических гипотез*** *выделяют* ***два подхода*** (два алгоритма): 1) *классический* (традиционный), основанный на выборе *фиксированного уровня значимости*; 2) основанный на *расчёте минимального уровня значимости* (*p-значения*, *p-value*). Второй подход чаще реализуется компьютерными статистическими программами. Разберём особенности этих двух процедур проверки статистических гипотез.

Логика проверки нулевой гипотезы включает в себя несколько последовательных шагов. Сначала изложим алгоритм ***с фиксированным уровнем значимости***:

1. Под конкретную содержательную задачу (позже будут рассмотрены виды этих задач и соответствующие им критерии) *выбирается подходящая функция от наблюдений — статистика критерия.* Эта функция выбирается по целому ряду соображений, одним из главных является то, что при нулевой гипотезе можно указать вероятностное распределение статистики, что позволит в дальнейшем построить критическую область.

2. *По выборке* данных, которыми располагает аналитик, *рассчитывается наблюдаемое* (*т. е. фактическое*, *эмпирическое*) *значение статистики этого критерия.*

3. На основе выбранного (фиксированного) уровня значимости (и уровня доверительной вероятности), а также исходя из возможных альтернативных гипотез, по соответствующей статистической таблице *ищется критическая точка* и *строится критическая область* для выбранной статистики (функции от наблюдений).

4. *Выясняется, в какую из областей своих значений попало значение статистики критерия, рассчитанное по выборке* (*т. е. по имеющимся данным*). Если наблюдаемое значение статистики критерия принадлежит области принятия нулевой гипотезы, то нулевая гипотеза не отвергается (нет оснований её отклонить). Если же наблюдаемое значение статистики критерия принадлежит критической области, то нулевая гипотеза отвергается (так как есть основания усомниться в её правдивости) и истинной объявляется альтернативная гипотеза.

Теперь рассмотрим понятие «минимальный уровень значимости», которое мы уже успели упомянуть.

Проверять статистические гипотезы можно и без построения критической области. В практике исследований чаще применяется подход, основанный на расчёте так называемого ***предельного уровня значимости*** (или *минимального уровня значимости*). Его обычно обозначают ***p-value***, ***значение вероятности*** (*или* ***sig. — significance***, т. е. ***значимость***). Это ***самая меньшая вероятность ошибки I рода*** (для полученного на выборке значения статистики критерия) в ситуации, когда отвергается (т. е. ***не*** принимается) ***верная*** нулевая гипотеза. Значение p-value вычисляют уже после того, как рассчитаны наблюдаемые параметры выборки (чаще всего с помощью специальных компьютерных программ, в которых осуществляется статистический анализ данных). Определение минимального уровня значимости зависит от типа альтернативных гипотез. Обозначим (в общем виде) статистику критерия (как случайную величину) через U, а некоторое её конкретное наблюдаемое значение — через u, и приведём формальные определения.

Минимальный уровень значимости (***p-value***): ***односторонний.***

• При $u > 0$ под односторонним уровнем значимости $P(U \geq u)$ будем понимать вероятность того, что значение U-статистики на выборке может принять ещё бóльшие положительные значения, чем наблюдаемое u. Так, минимальный уровень значимости показывает, с какой вероятностью статистика критерия может принять значение не меньшее, чем наблюдаемое, т.е. ровно *такое же* или *большее*, чем наблюдаемое u. ***Здесь рассматривается правый «хвост» плотности распределения вероятностей статистики критерия.***

• При $u < 0$ под односторонним уровнем значимости $P(U \leq u)$ будем понимать вероятность того, что значение U-статистики на выборке может принять ещё бóльшие отрицательные значения, чем наблюдаемое u. Так, минимальный уровень значимости показывает, с какой вероятностью статистика критерия примет значение не боль-

шее, чем наблюдаемое, т. е. ровно *такое же* или *меньшее*, чем наблюдаемое *u*. ***Здесь рассматривается левый «хвост» плотности распределения вероятностей статистики критерия.***

Минимальный уровень значимости **(*p-value*)**: ***двусторонний***.

• Под двусторонним уровнем значимости $P(|U| \geq u) = P(U \leq -u) + P(U \geq +u)$ будем понимать вероятность того, что значение *U*-статистики на выборке может принять значение, не меньшее (***по модулю***), чем наблюдаемое. Другими словами, двусторонний минимальный уровень значимости характеризует степень наших ожиданий относительно *вероятности получения такого значения статистики критерия*, *которое* ***в абсолютном выражении*** *не меньше наблюдаемого на выборке значения статистики критерия*. ***Здесь рассматриваются оба «хвоста» плотности распределения вероятностей статистики критерия — и правый, и левый.***

Приведём алгоритм проверки нулевой гипотезы с использованием минимального уровня значимости.

1. Под конкретную содержательную задачу *выбирается подходящая функция от наблюдений — статистика критерия*, так как если предположить, что нулевая гипотеза в реальности справедлива (истинна), то не составляет труда задать закон распределения вероятностей случайной величины (статистики критерия) при этой нулевой гипотезе.

2. *По имеющейся выборке* данных *рассчитывается наблюдаемое* (*т. е. фактическое*, *эмпирическое*) *значение статистики этого критерия*;

3. С помощью специальных таблиц (или компьютерных программ) для полученного наблюдаемого значения статистики критерия определяется *минимальный уровень значимости* (*p-value*) — как односторонний (в случае если нулевая гипотеза проверяется против направленной альтернативы), так и двусторонний (в случае если нулевая гипотеза проверяется против *не*направленной альтернативы). Напомним, что аналитик всегда исходит из предположения, что выдвинутая им для статистической проверки нулевая гипотеза в реальности справедлива. И разумным становится стремление оценить ***шансы*** **(*т. е. вероятность*)** ***получить на выборке данных именно такое значение статистики критерия***, ***которое равно как минимум наблюдаемому значению*** **(*взятому по модулю*, *со знаком*** «+»**).** И, таким образом, оценить типичность (характе́рность) наблюдаемого значения статистики критерия при нулевой гипотезе, и, следовательно, оправданность выдвижения самой нулевой гипотезы, её убедительность.

4. *Задаётся* (т. е. фиксируется, *назначается*) обычный уровень значимости **α** (для случая с односторонним *p-value*) или **α/2** (для случая с двусторонним *p-value*). Уровень значимости **α** или **α/2** отражает тот риск, на который аналитик готов пойти в случае ошибочного отклонения нулевой гипотезы. (Значение **α** или **α/2** в дальнейшем

и становится главным основанием для принятия решения — считать ли нулевую гипотезу справедливой или же нет). Естественно, что аналитик самостоятельно устанавливает (решает для себя), при каком уровне надёжности статистического вывода нулевую гипотезу следует принять, а при каком — отклонить. Поэтому процесс тестирования статистических гипотез ***никогда*** не происходит на уровне **100%**-й уверенности в результате (в выводах).

5. Чтобы принять решение относительно истинности нулевой гипотезы, ***сравниваются два значения***: ***заданный*** аналитиком ***уровень значимости*** (**α**), ***с одной стороны***, и полученный ***минимальный уровень значимости*** (***p-value***) — ***с другой***. *P-value* позволяет понять, в какой мере *реально* (т.е. правдоподобно), или же, напротив, маловероятно, получить наблюдаемое значение статистики критерия при нулевой гипотезе. Другими словами, *p-value* позволяет понять, какова потенциальная «жизнеспособность» проверяемой нулевой гипотезы. Так, ***если p-value оказалось меньше выбранного уровня значимости***, то аналитик отклоняет нулевую гипотезу в пользу конкурирующей (нулевая гипотеза не может быть принята при таких данных): в этом случае шансы на истинность нулевой гипотезы в реальности ничтожно малы (имеющаяся выборка данных противоречит нулевой гипотезе). ***Если p-value оказалось больше выбранного уровня значимости***, то у аналитика нет серьёзных оснований отклонить нулевую гипотезу, и она принимается (имеющаяся выборка данных согласуется с нулевой гипотезой). Ведь если эта вероятность *p-value* не столь мала (т.е. больше, чем **α**), то нет и оснований ставить под сомнение правдивость нулевой гипотезы.

Статистический вывод проверки гипотезы: по итогам осуществленной проверки нулевой гипотезы следует сделать заключение о том, можно ли считать верной нулевую гипотезу. Вывод формулируется с использованием устоявшихся словосочетаний, клише и специальной терминологии:

- ***если нулевая гипотеза принимается***, то делаем вывод: «При фиксированном уровне значимости (указывается, при каком именно) на основе имеющихся *выборочных* данных и по результатам проверки статистической гипотезы у нас *нет оснований отвергнуть* H_0. Гипотеза *не противоречит* нашим данным, она *не отклоняется*»;
- ***если нулевая гипотеза отклоняется***, то говорим: «При фиксированном уровне значимости (указывается, при каком именно) на основе имеющихся *выборочных* данных и по результатам проверки гипотезы у нас имеются *значимые основания отвергнуть* H_0 *в пользу альтернативы*[1] H_1 Гипотеза *противоречит* нашим данным, она *не принимается*».

---

[1] В пользу направленной или ненаправленной альтернативы — в зависимости от постановки задачи, — когда рассматривается односторонняя или двусторонняя критическая область соответственно.

Наши выборочные данные ***не позволяют на 100% делать достоверные выводы***. Эмпирические данные, которые получены в ходе выборочного обследования, могут быть искажены (в той или иной степени и в силу разных обстоятельств). Поэтому мы и говорим: «***При имеющихся данных нет оснований*** отклонить H_0» или «***При имеющихся данных есть значимые основания*** отвергнуть H_0 в пользу альтернативы H_1».

В завершении этого параграфа введём ещё два важных понятия, касающихся выборки исследования, которые будут нам полезны при проверке статистических гипотез:

Независимыми (несвязными) выборками называют выборки, полученные одним из двух способов:

- извлечённые из генеральной совокупности совершенно автономно (т. е. изолированно) друг от друга. Процедура отбора участников исследования в одной из них *никак не влияет* на процедуру отбора в другой. Случайным образом (произвольно) и независимо друг от друга выбираются единицы для исследования (объекты, респонденты, участники) внутри каждой совокупности. Таким образом, отбор в обеих совокупностях осуществляется независимо. Кроме того, процедура измерения интересующих исследователя характеристик тоже автономна.
- полученные из одной общей выборки, которая поделена на части по какому-либо критерию. Например: в нашем распоряжении имеется выборка населения России. Пусть мы хотим сопоставить мнение мужчин и женщин, тогда можно разбить выборку на две независимые выборки мужчин и женщин, или разбить на четыре независимые выборки по типу населенных пунктов, в которых проживают участники исследования: жители Москвы и Санкт-Петербурга (1), жители крупных городов (2), жители малых городов (3), жители сельской местности (4). Полученные в результате такого деления выборки (группы) являются непересекающимися и также автономными, независимыми.

Парные (связные) выборки (*в частных случаях их ещё называют выборками типа «до-и-после»*), напротив, не являются автономными.

Во-первых, связные выборки могут возникнуть, когда у одного и того же объекта исследования один и тот же параметр (одно и то же качество, характеристика, свойство) измеряется в разные моменты времени — как минимум дважды. Так, если доктора интересует, как меняется самочувствие людей *до* назначенной им терапии и *после* неё, то, очевидно, что первая выборка — это выборка пациентов, которым адресована терапия, а вторая — это выборка, состоящая *из тех же* пациентов, что и в первом случае. Таким образом, вторая выборка — это не любая другая случайная выборка пациентов со схожим заболеванием, с которой нужно сравнить результаты, а всё та же выборка с *фиксированным* составом участников.

Во-вторых, связными выборками могут быть такие выборки, в которых процедура отбора участников исследования и/или процедура измерения интересующих исследователя свойств (качеств, характеристик) *оказывают влияние друг на друга*. Например, одна из студенческих групп в 2012—2013 учебном году проводила социологическое исследование на тему: «Характер конфликтов в молодой семье, их причины и способы преодоления»[1]. Выборка данного исследования была связной, поскольку авторам данного исследования было важно зафиксировать не столько мнение мужчин и женщин в отдельности — что те и другие думают о конфликтах в семье, а сравнить эти мнения *в контексте каждой* конкретной супружеской пары, т. е. рассмотреть мнение мужчины и женщины «в сочетании», «в связке» друг с другом.

§ 6.2. Анализ таблиц сопряжённости. Критерий хи-квадрат

Рассмотрим класс задач, касающихся вопроса о том, существует или отсутствует статистически выраженная связь между двумя явлениями (признаками), измеренными в номинальной шкале. Ранее вы уже познакомились с такими средствами проверки существования зависимости, как расчёт коэффициентов корреляции — коэффициента линейной корреляции Пирсона и коэффициента ранговой корреляции Спирмена. Тогда речь шла о переменных, измеренных с помощью количественных и порядковых шкал соответственно, для которых и предусмотрены эти коэффициенты. А как быть в тех случаях, когда мы имеем дело с переменными, в основе измерения которых лежат сугубо номинальные шкалы?[2] На помощь приходят так называемые ***таблицы сопряжённости*** (их ещё называют таблицами перекрестной классификации) и специально разработанный критерий для изучения независимости двух признаков — статистический ***критерий хи-квадрат***. Об этом новом для нас критерии и пойдёт речь дальше.

---

[1] Студенты II курса факультета социологии НИУ ВШЭ: Полина Попова, Павел Голоднюк, Ярослав Раков (научный руководитель — А. В. Пашкевич).

[2] В социологических исследованиях классическим примером признака, измеряемого в номинальной шкале, является, например, пол респондента (у этого признака две градации: мужчина и женщина). Другим примером может служить материальное благосостояние респондента (уровень дохода). У этого признака можно выделить много градаций. Например, такие: «стеснённые в деньгах» (с уровнем дохода менее 1 000 усл. ед.), «с приемлемым достатком» (с уровнем дохода от 1 001 до 2 000 усл. ед.), «обеспеченные» (с уровнем дохода от 2 001 до 3 000 усл. ед.), «свободные в деньгах» (с уровнем дохода более 3 000 усл. ед.) и так далее.

Таблицы сопряжённости представляют собой ***двумерное*** (***совместное***) частотное эмпирическое распределение для двух признаков. По строкам таблицы откладываются градации одного признака, по столбцам таблицы — градации другого признака. На пересечении соответствующей строки и столбца указывается число наблюдений, для которых характерно сочетание конкретных значений признаков. Рассмотрим пример, который позволит нам проиллюстрировать постановку задачи.

Пример 6.2. «Шоколад». В ходе одной из лекций был проведён экспресс-опрос среди студентов: юношам и девушкам было предложено ответить на вопрос о том, какой шоколад они больше любят — горький или молочный. Кроме того, фиксировался пол участников этого исследования. В анкете (письменно) каждый из респондентов указал свои ответы на два поставленных вопроса. Это позволило построить двумерное распределение ответов, которые были сведены в следующую таблицу:

Признаки		Предпочтения в шоколаде		Всего, человек
		1. Горький шоколад	2. Молочный шоколад	
Пол	1. Юноши	**15** (1-я ячейка)	**6** (2-я ячейка)	21
	2. Девушки	**23** (3-я ячейка)	**41** (4-я ячейка)	64
Всего, человек		38	47	85

Примечание. Студентов, которые вообще не едят шоколад, в выборке не оказалось.

Зададимся вопросом: можно ли утверждать, что между полом человека и «шоколадными» предпочтениями есть определённая выраженная зависимость? Можно ли утверждать, что эти предпочтения различны у юношей и девушек? Такова содержательная постановка задачи. При этом отметим, что *формулировка нулевой статистической гипотезы* будет звучать *иначе*, нежели описанное предположение (см. далее).

В анализе присутствуют ***две*** характеристики респондентов:

1. *Пол* (который имеет две градации): 1) мужской, 2) женский;

2. *Предпочтения в шоколаде* (которые в данном случае тоже имеют только две градации, хотя их могло бы быть и больше): 1) горький шоколад, 2) молочный шоколад.

Оба признака — пол и предпочтения в шоколаде — измерены в ***номинальной шкале***. Их совместное эмпирическое распределение образует таблицу — таблицу сопряжённости, из которой следует

четыре различных сочетания (комбинации) градаций наших признаков:

3) *юноши*, которые предпочитают *горький* шоколад;

4) *юноши*, которые предпочитают *молочный* шоколад;

5) *девушки*, которые предпочитают *горький* шоколад;

6) *девушки*, которые предпочитают *молочный* шоколад.

Эмпирические частоты, указанные в ячейках, соответствуют числу респондентов, давших те или иные категории ответов. Обозначим эти частоты буквой n_{kl} (для матриц эти обозначения привычны и ассоциируются с понятием «объём выборки», «число наблюдений»), имеющей два нижних индекса — k и l, где ***k — порядковый номер строки*** (соответствующий определённой градации первого признака); ***l — порядковый номер столбца*** (соответствующий определённой градации второго признака). Так как в нашем примере признаки имеют по две градации каждая, то $k = 1, 2$ и $l = 1, 2$. Тогда имеем: $n_{11} = 15$, $n_{12} = 6$, $n_{21} = 23$, $n_{22} = 41$. Эти частоты называются в статистике ***наблюдаемыми частотами (observed frequencies)*** — *эмпирическими*, *фактическими* частотами совместного распределения признаков.

Частоты, характеризующие суммы по строкам и столбцам, принято называть ***маргинальными частотами*** (по краям таблицы, на её полях) — ими являются суммы частот по каждой строке либо по каждому столбцу, показывающие общее число наблюдений в отдельных столбцах и строках таблицы. Они также обозначаются буквой $n_{k.}$ или $n_{.l}$ с нижним индексом, указывающим, по какой строке или столбцу происходит суммирование наблюдений в таблице:

1) *если суммирование производится по строке*, то пишут «$n_{k.}$», где k — номер строки, по которой осуществляется сложение частот, а значком «**.**» (точка) указывается, что в операции суммы задействованы все столбцы;

2) *если суммирование производится по столбцу*, то пишут «$n_{.l}$», где l — номер столбца, по которому осуществляется сложение частот. Аналогичным образом значок «**.**» (точка) в нижнем индексе указывает на то, что в операции суммы задействованы все строки.

Для рассматриваемого примера имеем следующие маргинальные частоты: $n_{1.} = 21$, $n_{2.} = 64$, $n_{.1} = 38$, $n_{.2} = 47$. Иными словами, 21 человек — это общее число юношей в выборке (безотносительно к шоколаду), 64 человека — общее число девушек в выборке (безотносительно к шоколаду). Тогда 38 человек — общее число тех, кто любит горький шоколад (будь то юноши или девушки, т. е. безотносительно к полу); 47 человек — общее число юношей и девушек, которые любят молочный шоколад; 85 человек — совокупный объём выборки исследования. Он обозначается буквой $n_{..}$ или просто буквой n.

Итак, мы ввели основные обозначения. Теперь перейдём к обсуждению вопроса, прозвучавшего в постановке задачи: ***можно ли сказать, что между признаками обнаруживается статисти-***

чески устойчивая зависимость? *Вглядываясь* в частоты таблицы, можно *подметить*, что прослеживается зависимость между полом и предпочтениями в выборе шоколада — на первый взгляд кажется, что юноши любят скорее горький, а девушки, напротив, молочный шоколад. Чтобы подтвердить эти догадки (или, напротив, опровергнуть) о том, что в генеральной совокупности (а не просто в имеющейся выборке наблюдений) *статистически устойчива* подобная закономерность, нужны специальные методы. Метод «пристального взгляда» не пригоден для внимательного и строгого научного анализа. Наши интуитивные представления, сформированные опытом выборочного исследования, — только гипотезы, требующие проверки. *Необходимо проверить, можно ли действительно считать, что эта тенденция, подмеченная на случайной выборке наблюдений (ответов респондентов), справедлива для всей генеральной совокупности. Можно ли наши эмпирические данные уверенно обобщить на всю генеральную совокупность? И может ли гипотеза приобрести статус истинного (справедливого) утверждения, раскрывающего определённую социальную закономерность?*

Проверить гипотезу о взаимосвязи довольно сложно, ведь при этом придётся еще уточнять силу этой взаимосвязи, которая нам совершенно не известна. Поэтому гораздо проще проверить противоположное утверждение: ***«Связь между признаками отсутствует»***. Именно это предположение будет выбрано ***в качестве нулевой статистической гипотезы***. Эта гипотеза является простой. Она будет однозначно задавать теоретические (ожидаемые) частоты в таблице совместного распределения признаков. Альтернативная (конкурирующая) статистическая гипотеза, которая находится «в запасе», звучит так: H_1: ***«Зависимость между признаками есть»***. Её формулировка указывает на то, что расхождение в ответах юношей и девушек не<u></u>случайно, оно не является следствием погрешностей выборки, а свидетельствует об устойчиво выраженной зависимости между признаками, о достоверных, заметных, ощутимых различиях в выборе шоколада.

Для того чтобы проверить свои предположения, требуется воспользоваться определённым формальным правилом (критерием), с помощью которого с высокой долей уверенности (вероятности) можно подтвердить H_0 или опровергнуть её имеющимися результатами выборочного исследования. Этот критерий для таблиц сопряжённости носит специальное название: критерий хи-квадрат. С его помощью и проверяется стартовое предположение об отсутствии связи между признаками. Чтобы понять, как он устроен, нам потребуется разобрать ещё один важный вопрос.

Итак, в нашем распоряжении эмпирические частоты (данные нам выборкой исследования). Чтобы сказать, есть ли зависимость между признаками, естественно предположить, что эти частоты нужно сравнить с каким-то эталоном (своего рода нормативом), ***характе-***

ризующим независимость. Этим эталоном станут так называемые ***ожидаемые частоты* (*expected frequencies*)** — их также называют *теоретические*, *прогнозируемые* частоты, или частоты *модели*. Это такие частоты, которые ***наблюдались бы в выборке***, ***если бы в действительности между признаками зависимости в сущности никакой бы не было***. Это означает, что нам следует смоделировать (предугадать, рассчитать, спрогнозировать), какими ***оказались бы*** частоты в ячейках таблицы, если предположить, что в действительности между признаками *нет абсолютно никакой зависимости*?! Ответ на этот вопрос подсказывает теория вероятностей, когда обсуждается вопрос о независимости двух событий. Вспомните, когда мы считаем два события независимыми? Именно тогда, когда факт наступления одного события никак не сказывается на наступлении или ненаступлении другого. В таких случаях мы говорим, что событие B не определяет событие A, и наоборот — событие A не определяет B. Они не влияют друг на друга. Они независимы. Тогда вероятность их совместного (одновременного) наступления рассчитывается как ***произведение вероятностей отдельных событий*** — A и B, а ***условная вероятность*** наступления одного из этих событий равна безусловной вероятности этого события: $\boldsymbol{P(AB) = P(A|B)P(B) = P(A)P(B)}$. Теперь воспользуемся этим знанием применительно к нашей таблице сопряжённости. У нас есть два признака — «пол» и «предпочтения в шоколаде». Будем считать событием A={Респондент является юношей}, а событием B={Респондент любит молочный шоколад}. Чему тогда ***приблизительно*** равны вероятности $P(A)$ и $P(B)$? Очевидно, что $P(A) = \frac{n_{1\bullet}}{n}$, а $P(B) = \frac{n_{\bullet 2}}{n}$. Ведь частоты *дают нам оценку вероятностей*. Если рассматривать события A и B независимыми (никак не связанными между собой), то чему равна вероятность их совместного наступления? Чему равна вероятность события AB = {Респондент является юношей, который любит молочный шоколад}? Ясно, что эта вероятность ***примерно*** *равна* $P(AB) = \frac{n_{12}}{n}$ (обозначения см., пожалуйста, ниже в таблице).

Признаки		Сорт шоколада		Всего
		1. Горький шоколад	2. Молочный шоколад	
Пол	1. Юноши	15	$n_{12} = \mathbf{6}$	$n_{1\bullet} = 21$
	2. Девушки	23	41	64
Всего		38	$n_{\bullet 2} = 47$	$n = 85$

Тогда в случае независимости событий A и B должно обязательно выполняться следующее равенство[1]:

$$P(AB) \approx P(A) \cdot P(B).$$

Подставив вместо буквенных обозначений имеющиеся расчётные значения этих вероятностей, получим: $\frac{n_{12}}{n} \approx \frac{n_{1.}}{n} \cdot \frac{n_{.2}}{n}$, откуда $n_{12} \approx \frac{n_{1.} n_{.2}}{n}$.

Подставим значения маргинальных частот в полученное выражение и определим, чему *должна быть равна ожидаемая* (*теоретическая*) *частота* в ячейке (1;2) — первая строка, второй столбец — чтобы между полом и выбором сорта шоколада имеющиеся данные не обнаруживали бы абсолютно никакой взаимосвязи:

$$n_{12} \approx \frac{n_{1.} n_{.2}}{n} = \frac{21 \cdot 47}{85} = 11{,}6 \text{ (человек)}.$$

А по факту в таблице 6 человек! Таким же образом рассчитаем ожидаемые частоты для остальных ячеек таблицы. В общем виде запишем, как вычисляются теоретические частоты для произвольно взятой комбинации значений градаций признаков:

$$n_{kl} \approx \frac{n_{k.} n_{.l}}{n},$$

где в числителе стоит произведение маргинальных частот соответствующей строки и соответствующего столбца, пересечение которых и образует интересующее нас сочетание градаций признаков; *n — общий объём выборки исследования*.

В результате такого несложного расчёта каждой эмпирической (наблюдаемой) частоте будет противопоставлена своя теоретическая (ожидаемая, прогнозная) частота (которая должна *была бы быть* в ячейке, *если бы* данные свидетельствовали о полном отсутствии зависимости между признаками). Посмотрим, что получается для рассматриваемого примера. Укажем четыре ожидаемые частоты для ячеек таблицы (округлять значения до целого не обязательно). Теоретически мы ожидаем, что частоты в ячейках таблицы распределятся между юношами и девушками следующим образом:

$$n_{11}^{\text{ожидаемая}} \approx \frac{n_{1.} n_{.1}}{n} = \frac{21 \cdot 38}{85} = 9{,}4 \text{ (человек)};$$

$$n_{12}^{\text{ожидаемая}} \approx \frac{n_{1.} n_{.2}}{n} = \frac{21 \cdot 47}{85} = 11{,}6 \text{ (человек)};$$

---

[1] Используем знак приблизительного равенства (≈), так как в примере речь идёт о примерной оценке *вероятностей* (на основе *частот* таблицы).

$$n_{21}^{\text{ожидаемая}} \approx \frac{n_{2.} n_{.1}}{n} = \frac{64 \cdot 38}{85} = 28{,}6 \text{ (человек)};$$

$$n_{22}^{\text{ожидаемая}} \approx \frac{n_{2.} n_{.2}}{n} = \frac{64 \cdot 47}{85} = 35{,}4 \text{ (человек)}.$$

Обратите внимание, что, вычислив *только одну* *из ожидаемых частот*, можно легко определить ожидаемые частоты в *остальных трёх* ячейках. Например, вычислив $n_{11}^{\text{ожидаемая}}$, можно без труда «восстановить» ожидаемые частоты в остальных трёх ячейках несколько иным способом: из соответствующих маргинальных частот вычесть это значение 9,4. Тогда имеем:

$$n_{12}^{\text{ожидаемая}} \approx 21 - 9{,}4 = 11{,}6;$$

$$n_{21}^{\text{ожидаемая}} \approx 38 - 9{,}4 = 28{,}6;$$

$$n_{22}^{\text{ожидаемая}} \approx 47 - (21 - 9{,}4) = 47 - 11{,}6 = 35{,}4;$$

$$(\text{или } n_{22}^{\text{ожидаемая}} \approx 64 - (38 - 9{,}4) = 64 - 28{,}6 = 35{,}4).$$

Естественно, что ответы получатся такие же, как если бы мы рассчитывали частоты по формуле $n_{kl} \approx \frac{n_{k.} n_{.l}}{n}$. Говорят, что в таблице сопряжённости размером 2×2 мы обладаем «***одной степенью свободы***».

Число степеней свободы (df) для таблицы сопряжённости вычисляется как произведение $df = (K-1)(L-1)$, где K — количество возможных градаций, которые может принимать первый признак (т. е. число строк в таблице), L — количество градаций, которые может принимать второй признак (т. е. число столбцов в таблице). Обратите внимание, что таблица сопряжённости состоит из (KL) ячеек, образованных сочетанием значений признаков. Тогда, если известны частоты как *минимум* в $(K-1)(L-1)$ ячейках, можно «восстановить» недостающие сведения — каковы частоты в остальных ячейках. Поэтому $(K-1)(L-1)$ — это ***минимальное количество ячеек, частоты которых должны быть известны, чтобы можно было «восстановить» информацию о частотах в ячейках, которые не известны***. В некотором смысле в нашем распоряжении только $(K-1)(L-1)$ «свободных» ячеек, так как знание о частотах в остальных ячейках всегда можно получить путём несложных вычислений.

В примере 6.2 («Шоколад») ячеек ровно 4, всего ровно 4 возможных сочетания (комбинации), образованных градациями признаков. Число степеней свободы для такой таблицы сопряжённости равно $(K-1)(L-1) = (2-1) \cdot (2-1) = 1 \cdot 1 = 1$. Итак, мы располагаем *одной* сте-

пенью свободы. Если нам известна частота лишь для одной из четырёх ячеек, для остальных — нет, то, зная маргинальные частоты, *всегда* можно определить (вычислить) остальные три. Только одна ячейка является «свободной» (независимой от остальных), остальные три — «несвободные».

Итак, запишем ***примерные*** значения ожидаемых частот в таблицу, приписав каждой $n_{kl}^{\text{наблюдаемая}}$ соответствующую $n_{kl}^{\text{ожидаемая}}$. Получим:

Признаки		Сорт шоколада		Всего
		1. Горький шоколад	2. Молочный шоколад	
Пол	1. Юноши	15/**9,4**	6/**11,6**	21
	2. Девушки	23/**28,6**	41/**35,4**	64
Всего		38	47	85

Заметьте при этом, что маргинальные частоты остаются неизменными — мы перераспределяем объём выборки (число наблюдений) между ячейками так, чтобы частоты, которые в них содержатся, указывали на независимость двух признаков. Теперь зададимся вопросом: почему мы говорим, что частоты $n_{11}^{\text{ожидаемая}}$, $n_{12}^{\text{ожидаемая}}$, $n_{21}^{\text{ожидаемая}}$, $n_{22}^{\text{ожидаемая}}$ свидетельствуют о том, что связи между признаками никакой нет?

Чтобы ответить на вопрос, какова ***примерная*** вероятность того, что человек любит горький шоколад, ***при условии***, что перед нами человек мужского пола, нужно вычислить соответствующую долю: $\frac{n_{11}^{\text{ожидаемая}}}{n_{1.}} = \frac{9,4}{21} = 0,45 = 45\,\%$. А какова тогда вероятность того, что он, напротив, любит молочный шоколад? Эта примерная вероятность составит: $\frac{n_{12}^{\text{ожидаемая}}}{n_{1.}} = \frac{11,6}{21} = 1 - 0,45 = 0,55 = 55\,\%$. Теперь рассчитаем *условные* вероятности аналогичных событий для девушек. Какова вероятность того, что случайно выбранная девушка любит горький шоколад? Эта вероятность *примерно* равна отношению: $\frac{n_{21}^{\text{ожидаемая}}}{n_{2.}} = \frac{28,6}{64} = 0,45 = 45\,\%$. Вероятность же того, что перед нами девушка, которой нравится молочный шоколад, соответствует доле: $\frac{n_{22}^{\text{ожидаемая}}}{n_{2.}} = \frac{35,4}{64} = 1 - 0,45 = 0,55 = 55\,\%$. Видим, что доли тех, кто ценит молочный шоколад, среди юношей и девушек одинаковые ($55\,\% = 55\,\%$)! То же самое можно сказать и про горький шоколад ($45\,\% = 45\,\%$).

В том случае, если бы ответы респондентов в выборке распределились согласно теоретическим (ожидаемым) частотам, которые получили, ***смоделировав ситуацию полной (100%-й) независимости***, мы бы зафиксировали, что предпочтения юношей и девушек в отношении выбора шоколада ***схожие***, ***различий нет***, а значит, ***нет и связи*** между полом и выбираемым сортом шоколада.

Итак, после расчёта (прогноза) ожидаемых частот, появилась ***возможность сравнить*** фактические данные с теоретической моделью полной (стопроцентной) независимости двух переменных. Основанием для такого сравнения служит ***разность между наблюдаемыми и ожидаемыми частотами***. В основе вычисления статистики критерия хи-квадрат лежит соотнесение наблюдаемой и ожидаемой частот в каждой ячейке таблицы сопряжённости. Для каждой ячейки рассчитывается разность вида: $(n_{kl}^{\text{наблюдаемая}} - n_{kl}^{\text{ожидаемая}})$. Поскольку эта разность может быть как положительной (со знаком «+»), так и отрицательной (со знаком «–»), то «избавляются» от знака путем возведения этой разности в квадрат: $(n_{kl}^{\text{наблюдаемая}} - n_{kl}^{\text{ожидаемая}})^2$. Эту разность для каждой ячейки «нормируют» («масштабируют») путём деления на свою ожидаемую частоту, получая тем самым значение квадрата разности, приходящееся на одну единицу ожидаемой частоты: $\frac{(n_{kl}^{\text{наблюдаемая}} - n_{kl}^{\text{ожидаемая}})^2}{n_{kl}^{\text{ожидаемая}}}$. Далее значения по всем ячейкам таблицы сопряжённости складывают, оценивая ***общим итогом квадрат нормированной разности между наблюдаемыми и ожидаемыми частотами***, приходящийся на ожидаемые частоты модели. В результате получают так называемое *наблюдаемое значение статистики критерия хи-квадрат*:

$$\chi^2 = \sum_{k,l}^{K,L} \frac{(n_{kl}^{\text{наблюдаемая}} - n_{kl}^{\text{ожидаемая}})^2}{n_{kl}^{\text{ожидаемая}}}.$$

Из формулы видно, что статистика критерий хи-квадрат всегда принимает положительное значение или же равна нулю.

Возвращаясь к рассматриваемому примеру, рассчитаем наблюдаемое значение статистики критерия хи-квадрат для имеющихся данных (слагаемых будет ровно столько, сколько сочетаний признаков, образованных их градациями):

$$\chi^2_{\text{наблюдаемое}} = \sum_{k,l}^{K=2,L=2} \frac{(n_{kl}^{\text{наблюдаемая}} - n_{kl}^{\text{ожидаемая}})^2}{n_{kl}^{\text{ожидаемая}}} =$$

$$= \frac{(n_{11}^{\text{наблюдаемая}} - n_{11}^{\text{ожидаемая}})^2}{n_{11}^{ожидаемая}} + \frac{(n_{12}^{\text{наблюдаемая}} - n_{12}^{\text{ожидаемая}})^2}{n_{12}^{ожидаемая}} +$$

$$+\frac{(n_{21}^{\text{наблюдаемая}}-n_{21}^{\text{ожидаемая}})^2}{n_{21}^{\text{ожидаемая}}}+\frac{(n_{22}^{\text{наблюдаемая}}-n_{22}^{\text{ожидаемая}})^2}{n_{22}^{\text{ожидаемая}}}=$$

$$=\frac{(15-9{,}4)^2}{9{,}4}+\frac{(6-11{,}6)^2}{11{,}6}+\frac{(23-28{,}6)^2}{28{,}6}+\frac{(41-35{,}4)^2}{35{,}4}=$$

$$=3{,}34+2{,}70+1{,}10+0{,}88=8{,}02.$$

Наблюдаемое значение статистики критерия хи-квадрат служит мерой оценивания того, насколько велико (или, наоборот, мало) разногласие (несовпадение) между частотами — эмпирическими (которые даны по факту) и теоретическими (предсказываемые моделью полной независимости). *Чем ближе значение* χ^2 *к нулю*, тем в большей степени совпадают наблюдаемые частоты с ожидаемыми частотами модели: это и понятно, каждое из слагаемых в таком случае стремится к нулю и, следовательно, их сумма также стремится к нулю. А значит, и *растёт наша убеждённость в том, что связи между признаками действительно никакой нет*. И наоборот: *чем больше эта рассогласованность* (*дистанция*) *между фактическими и ожидаемыми частотами*, тем *бóльшим* получается наблюдаемое значение χ^2 и тем *сильнее растёт наша уверенность в истинности противоположного утверждения*: *связь между признаками в действительности есть!*

Следующий важный вопрос, который нужно затронуть: ***а каким должно быть*** **значение** χ^2, ***достигнув которого можно уверенно сказать, что расхождение между наблюдаемыми и ожидаемыми частотами уже столь ощутимо, что следует считать этот факт неслучайным***? А, значит, поверить в справедливость альтернативного утверждения о том, что связь между признаками в реальности есть. Напомним, что мы получили оценку статистики критерия хи-квадрат на основе информации, собранной в ходе совершенно *конкретного выборочного* исследования, и справедливой *именно для этой конкретной выборки*. Однако вместе с тем наша главная цель — от результатов анализа данных конкретной выборки перейти к выводам, касающимся *всей генеральной совокупности*. В рассматриваемом примере фактическое значение оказалось равным 8,02. Само по себе это значение пока не говорит нам о том, а много ли это, или мало. Для того чтобы интерпретировать его, ***следует задать критическую область или использовать понятие минимального уровня значимости*** (подробнее см. материал § 6.1).

Вспомним, что статистика критерия — как функция от наблюдений — величина *случайная*. При нулевой гипотезе «Зависимости между признаками нет» эта случайная величина (т. е. статистика критерия) имеет определённое распределение вероятностей — в данном случае распределение хи-квадрат с числом степеней свободы *df*. Статистика критерия χ^2 — это *непрерывная* случайная величина,

принимающая значения на положительном континууме (значения χ^2 всегда неотрицательны). Плотность распределения χ^2 выглядит по-разному в зависимости от того, каково число степеней свободы в таблице сопряжённости. Чаще всего по своей форме распределение χ^2 является асимметричным, скошенным. Однако по мере возрастания числа наблюдений и числа степеней свободы асимметрия исчезает, плотность распределения приближается к нормальному (рис. 6.4).

В специальной статистической таблице содержатся значения статистики критерия хи-квадрат для различных *степеней свободы таблицы сопряжённости* (*df*), с одной стороны, и для различных *уровней значимости* — с другой.

Заметим, что если нулевая гипотеза об отсутствии связи между рассматриваемыми признаками справедлива, то мы ожидаем, что наблюдаемое на имеющейся выборке данных значение статистики хи-квадрат будет *близко к нулю*, что свидетельствует о том, *что наблюдаемые и ожидаемые частоты в таблице сопряжённости различаются незначительно*. Напомним, что *вероятность встречаемости больших* (*нетипичных*) *значений статистики критерия хи-квадрат крайне мала для ситуации, когда* нулевая гипотеза истинна (так как такие большие значения почти невозможны). Если же нулевая гипотеза на самом-то деле неверна, то мы ожидаем, что фактическое значение статистики критерия хи-квадрат, наоборот, будет довольно большим (в силу того, что будут большими рас-

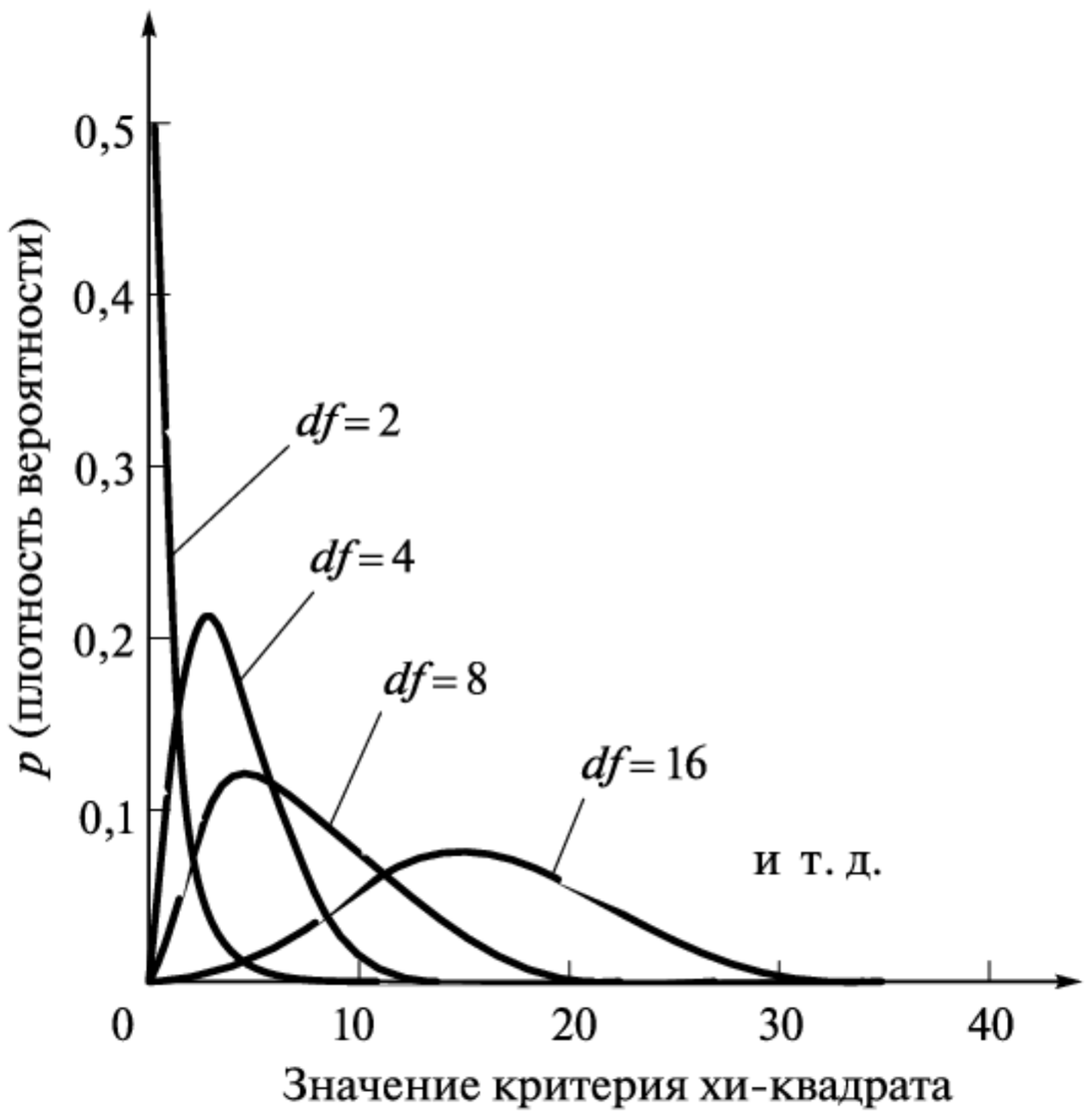


Рис. 6.4. Плотность распределения хи-квадрат при различных степенях свободы (*df*)

хождения между наблюдаемыми и ожидаемыми частотами в таблице сопряжённости).

Научимся по специальной статистической таблице определять критическое значение статистики критерия хи-квадрат и задавать критическую область (т. е. область отклонения нулевой гипотезы). Для того чтобы найти *критическое* (*пограничное*) *значение* статистики хи-квадрат, потребуется вычислить *число степеней свободы* для имеющейся таблицы сопряжённости: это позволит выделить из множества различных плотностей распределения хи-квадрат *одно единственное* (*конкретное*) *распределение*. Таблицы сопряжённости могут иметь разный размер, а потому разным будет и число степеней свободы. Минимальное число степеней свободы равно 1 (оно получается для таблицы сопряжённости размером 2×2). *Чем выше число сочетаний, образованных градациями двух признаков, тем больше и число степеней свободы таблицы сопряжённости*. Таблица 6.2 устроена просто. По строкам указаны числа степеней свободы: от 1 и выше. Столбцы содержат *наиболее часто используемые на практике фиксированные уровни значимости*: **0,1; 0,05; 0,01 и 0,001**. Искомая пограничная точка указана в таблице на пересечении соответствующих строки и столбца.

Таблица 6.2. **Таблица критических точек для χ^2-распределения**

df	Доверительная вероятность и уровень α			
	$\gamma = 0{,}9$ (90 %)	$\gamma = 0{,}95$ (95 %)	$\gamma = 0{,}99$ (99 %)	$\gamma = 0{,}999$ (99,9 %)
	$\alpha = 0{,}1$	$\alpha = 0{,}05$	$\alpha = 0{,}01$	$\alpha = 0{,}001$
1	2,71	3,84	6,63	10,83
2	4,61	5,99	9,21	13,82
3	6,25	7,81	11,34	16,27
4	7,78	9,49	13,28	18,47
5	9,24	11,07	15,09	20,52
6	10,64	12,59	16,81	22,46
7	12,02	14,07	18,48	24,32
8	13,36	15,51	20,09	26,12
9	14,68	16,92	21,67	27,88
10	15,99	18,31	23,21	29,59

Окончание табл. 6.2

df	Доверительная вероятность и уровень α			
	$\gamma = 0{,}9$ (90 %)	$\gamma = 0{,}95$ (95 %)	$\gamma = 0{,}99$ (99 %)	$\gamma = 0{,}999$ (99,9 %)
	$\alpha = 0{,}1$	$\alpha = 0{,}05$	$\alpha = 0{,}01$	$\alpha = 0{,}001$
11	17,28	19,68	24,72	31,26
12	18,55	21,03	26,22	32,91
13	19,81	22,36	27,69	34,53
14	21,06	23,68	29,14	36,12
15	22,31	25,00	30,58	37,70
16	23,54	26,30	32,00	39,25
17	24,77	27,59	33,41	40,79
18	25,99	28,87	34,81	42,31
19	27,20	30,14	36,19	43,82
20	28,41	31,41	37,57	45,31
21	29,62	32,67	38,93	46,80
22	30,81	33,92	40,29	48,27
23	32,01	35,17	41,64	49,73
24	33,20	36,42	42,98	51,18
25	34,38	37,65	44,31	52,62
и т. д.	...	...	...	...

Итак, будет найдено критическое (рубежное, пограничное) значение, которое удовлетворяет условию $\boldsymbol{P(0 \le \chi^2 < \chi^2_{\text{критическое}}) = \gamma}$. Буквой α («альфа») обозначается площадь под кривой χ^2-распределения (площадь «хвоста» распределения справа), которая не охватывается отрезком значений $[0; \chi^2_{\text{критическое}})$. По сути, это вероятность редких событий: *вероятность того, что при условии независимости двух признаков по результатам выборочного исследования может быть получено наблюдаемое значение статистики критерия хи-квадрат, превышающее его критическое значение либо равное ему* (рис. 6.5).

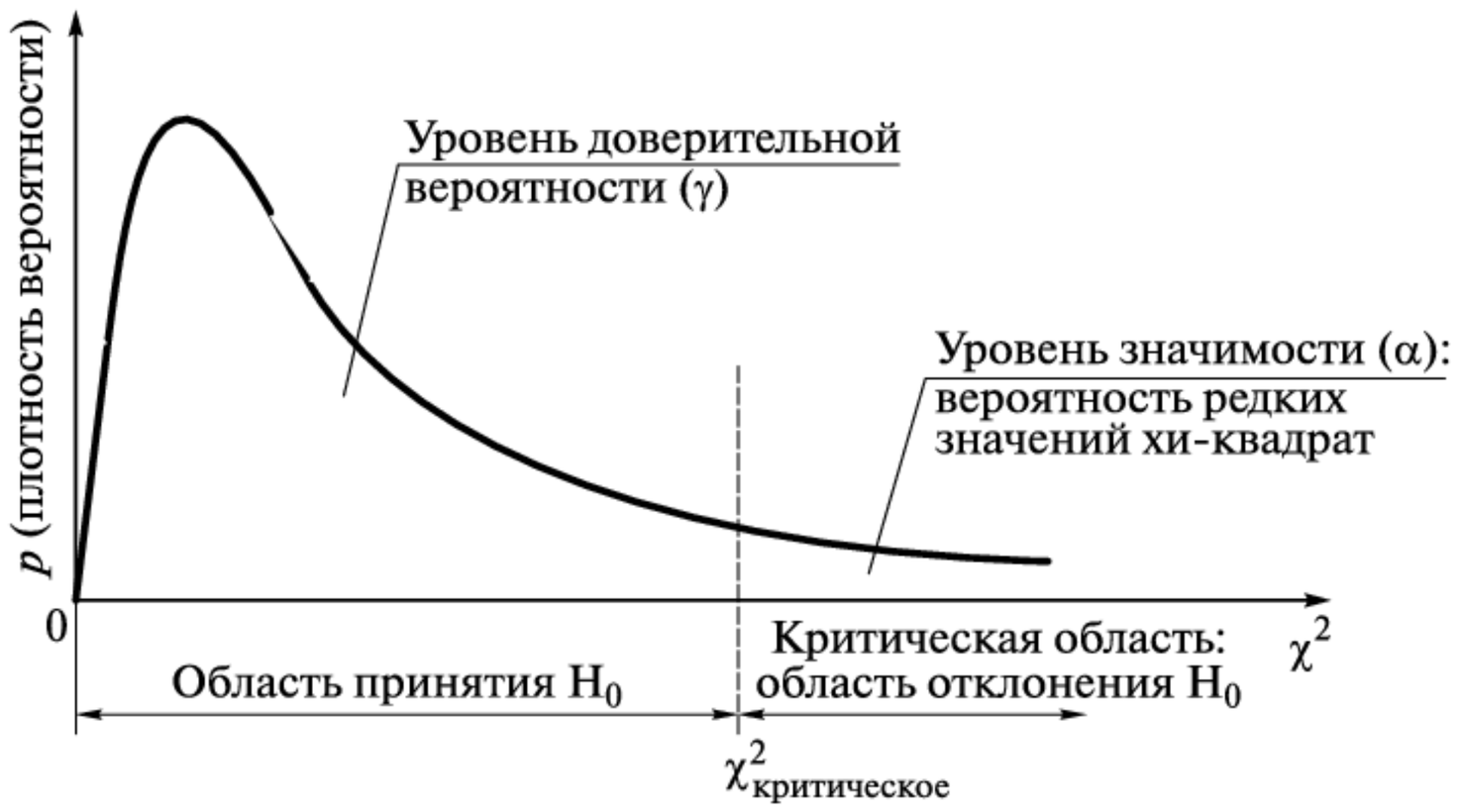


Рис. 6.5. Плотность распределения хи-квадрат и построение критической области

Принцип проверки статистической гипотезы об отсутствии зависимости заключается в следующем:

• *если наблюдаемое значение статистики критерия хи-квадрат не превышает критического значения* (т.е. лежит *левее его*), *то нулевая гипотеза об отсутствии связи не отвергается*;

• *если наблюдаемое значение статистики критерия хи-квадрат превосходит критическое* значение (т.е. лежит *правее* его) либо *равно* ему, то имеются основания *отклонить нулевую гипотезу в пользу альтернативной о том, что связь между признаками на самом деле есть*. Почему? Ведь если случилось так, что наблюдаемое значение статистики критерия превысило критическое, то это свидетельствует о том, что произошло событие, вероятность которого (при условии, что признаки в генеральной совокупности независимы) *ничтожно мала*! Эта вероятность равна $\alpha = 1 - \gamma$. Ведь если признаки независимы, то мы этого события не ожидаем, а предполагаем обратное: что оно едва ли может случиться. А если всё-таки случилось, то это порождает *сомнения* в справедливости H_0 и *укрепляет убеждённость* в H_1.

Вывод: ***нулевая гипотеза будет нами отвергнута в том случае, если наблюдаемое значение χ^2 будет не меньше своего теоретического (критического) значения для заданного уровня доверительной вероятности (γ) и заданного таблицей сопряжённости числа степеней свободы (df).***

Теперь вернёмся к рассмотренной в этом параграфе задаче про шоколад и завершим её решение. Итак, $\chi^2_{\text{наблюдаемое}} = 8{,}02$, число степеней свободы в таблице сопряжённости $df = 1$. Пусть мы хотим быть уверенными в выводах, распространяемых на генеральную совокупность, на 99%. Тогда по таблице находим, что $\chi^2_{\text{критическое}} = 6{,}63$. Итак, $8{,}02 > 6{,}63$. ***Случилось событие (которое состоит в том, что***

значение статистики критерия хи-квадрат по факту оказалось равным 8,02), ***т. е. попало в ту область*** (***критическую***), ***вероятность которой мала.*** Но это событие случилось! Поэтому при имеющихся данных мы отклоняем нулевую гипотезу об отсутствии связи и принимаем конкурирующую гипотезу как истинную. Зависимость есть. Значит, мы утверждаем, что в генеральной совокупности между полом человека и предпочтениями в выборе шоколада есть определённая выраженная зависимость. Эта ***связь статистически обоснована***. Действительно, можно утверждать, что у юношей и девушек ***предпочтения различны***.

Приведём алгоритм решения задач на проверку статистической гипотезы об отсутствии связи между двумя признаками:

1) строится таблица сопряжённости. Находятся маргинальные частоты таблицы — суммы частот по строкам и столбцам;

2) формулируются нулевая и альтернативная статистические гипотезы: $\mathbf{H_0}$: «Зависимости между признаками ***нет***» и $\mathbf{H_1}$: «Зависимость между признаками ***есть***»;

3) вычисляются ожидаемые (прогнозируемые) частоты для каждой ячейки;

4) рассчитывается наблюдаемое значение статистики критерия $\chi^2_{\text{наблюдаемое}}$;

5) определяется число степеней свободы *df*;

6) при заданном уровне доверия по таблице находится критическое значение (критическая точка) $\chi^2_{\text{критическое}}$;

7) сравниваются два значения: если $\chi^2_{\text{наблюдаемое}} < \chi^2_{\text{критическое}}$, то принимается H_0: «Связи между признаками нет». Если $\chi^2_{\text{наблюдаемое}} \geq \geq \chi^2_{\text{критическое}}$, то отвергается H_0 и принимается H_1: «Связь между признаками есть»;

8) формулируется содержательный вывод найденного решения.

Важное примечание. Критерий хи-квадрат используется нередко и пригоден для многих исследовательских задач, за исключением ситуации, когда таблица сопряжённости содержит *немалое количество ячеек*, в которых наблюдений почти нет. *Малонаполненными* (*«карликовыми»*) ячейками обычно считаются те, в которых частоты не превышают 5 (т. е. $n_{kl} < 5$). Это обстоятельство может быть следствием *слишком дробного разбиения* уровней (градаций) признаков в таблице сопряжённости, или же *редкой встречаемостью* того или иного сочетания (комбинации) градаций рассматриваемых признаков (что само по себе, конечно, возможно), или же следствием небольшого (*малого*) *объема выборки* вообще и др. В результате довольно неопределёнными могут стать *оценки* ожидаемых частот и, как следствие, сама оценка статистики критерия χ^2 становится неоднозначной, не поддается точной оценке. Существует некоторый выход из положения: объединить те или иные строки или столбцы

между собой (т. е. «склеить» их, присоединить ячейки друг к другу) или же исключить из анализа те или иные градации признаков. Руководствоваться в данном случае нужно содержательными соображениями, не нарушая содержательную постановку задачи. В результате полученная таблица сопряжённости меньшего размера будет основой для дальнейшей проверки статистической гипотезы об отсутствии связи. Число степеней свободы при этом будет определено для того количества сочетаний (комбинаций) уровней двух признаков, которые будут образованы после объединения ячеек либо исключения «карликовых» вообще.

Пример 6.3. «Чужое». Во всероссийском исследовании ценностей, проведённом в 2007 г. Центром социального прогнозирования под руководством Ф. Э. Шереги, двум категориям граждан (молодому и старшему поколениям — людям в возрасте 17 — 26 лет и 40 — 60 лет) был адресован вопрос: «Как, на Ваш взгляд, следует относиться к такому поступку, как присвоение найденных вещей, денег?»[1]. Объёмы выборок соответствующие: 1796 человек (молодёжь), 655 человек (старшее поколение).

Присвоение найденных вещей, денег	Молодёжь, 17 — 26 лет	Старшее поколение, 40 — 60 лет
Никогда не может быть оправдано	491	265
Иногда это допустимо	938	294
К этому следует относиться снисходительно	363	95
Затрудняюсь ответить	4	1
Всего, человек	1796	655

Исследователи полагают, что отношение к такому поступку, как присвоение найденных вещей или денег связано с возрастом человека. Проверьте статистическую гипотезу о том, что зависимости между указанными признаками не существует. Уровень доверительной вероятности примите равным 0,95. Ответы «Затрудняюсь ответить» исключите из анализа.

Решение. В анализе имеются два признака — отношение человека к присвоению чужого — вещей или денег (пусть это будет первым признаком), и возрастная группа, к которой принадлежит респондент

[1] Данные Единого архива экономических и социологических данных [http://sophist.hse.ru/db/oprosy.shtml?ts=121&en=0] (проверено 15.06.2012).

(пусть это будет вторым признаком). Совместное распределение этих признаков образует таблицу сопряжённости, включающую шесть различных сочетаний (комбинаций) уровней этих признаков.

Для решения используем приведённый алгоритм.

Во-первых, строим таблицу сопряжённости, исключая категорию ответа «Затрудняюсь ответить» и добавляя маргинальные частоты по строкам и столбцам. Получим:

Присвоение найденных вещей, денег	Молодёжь, 17 — 26 лет	Старшее поколение, 40 — 60 лет	Всего, человек
Никогда не может быть оправдано	**491**	**265**	756
Иногда это допустимо	**938**	**294**	1 232
К этому следует относиться снисходительно	**363**	**95**	458
Всего, человек	1 792	654	2 446

Во-вторых, формулируем нулевую и альтернативную статистические гипотезы:

H_0: «Зависимости между признаками нет»;

H_1: «Зависимость между признаками есть».

В-третьих, вычисляем ожидаемые (прогнозируемые) частоты для каждой ячейки (их всего 6):

$$n_{11}^{\text{ожидаемая}} = \frac{n_{1.} n_{.1}}{n} = \frac{756 \cdot 1\,792}{2\,446} = 554 \text{ (человек)};$$

$$n_{12}^{\text{ожидаемая}} = \frac{n_{1.} n_{.2}}{n} = \frac{756 \cdot 654}{2\,446} = 202 \text{ (человек)};$$

$$n_{21}^{\text{ожидаемая}} = \frac{n_{2.} n_{.1}}{n} = \frac{1\,232 \cdot 1\,792}{2\,446} = 903 \text{ (человек)};$$

$$n_{22}^{\text{ожидаемая}} = \frac{n_{2.} n_{.2}}{n} = \frac{1\,232 \cdot 654}{2\,446} = 329 \text{ (человек)};$$

$$n_{31}^{\text{ожидаемая}} = \frac{n_{3.} n_{.1}}{n} = \frac{458 \cdot 1\,792}{2\,446} = 335 \text{ (человек)};$$

$$n_{32}^{\text{ожидаемая}} = \frac{n_{3.} n_{.2}}{n} = \frac{458 \cdot 654}{2\,446} = 123 \text{ (человек)}.$$

Подставим эти значения в таблицу сопряжённости:

Присвоение найденных вещей, денег	Молодёжь, 17—26 лет	Старшее поколение, 40—60 лет	Всего, человек
Никогда не может быть оправдано	**491/554**	**265/202**	756
Иногда это допустимо	**938/903**	**294/329**	1 232
К этому следует относиться снисходительно	**363/335**	**95/123**	458
Всего, человек	1 792	654	2 446

В-четвертых, рассчитываем наблюдаемое значение статистики критерия χ^2:

$$\chi^2_{\text{наблюдаемое}} = \sum_{k,l}^{K=3,L=2} \frac{(n_{kl}^{\text{наблюдаемая}} - n_{kl}^{\text{ожидаемая}})^2}{n_{kl}^{\text{ожидаемая}}} =$$

$$= \frac{(n_{11}^{\text{наблюдаемая}} - n_{11}^{\text{ожидаемая}})^2}{n_{11}^{\text{ожидаемая}}} + \frac{(n_{12}^{\text{наблюдаемая}} - n_{12}^{\text{ожидаемая}})^2}{n_{12}^{\text{ожидаемая}}} +$$

$$+ \frac{(n_{21}^{\text{наблюдаемая}} - n_{21}^{\text{ожидаемая}})^2}{n_{21}^{\text{ожидаемая}}} + \frac{(n_{22}^{\text{наблюдаемая}} - n_{22}^{\text{ожидаемая}})^2}{n_{22}^{\text{ожидаемая}}} +$$

$$+ \frac{(n_{31}^{\text{наблюдаемая}} - n_{31}^{\text{ожидаемая}})^2}{n_{31}^{\text{ожидаемая}}} + \frac{(n_{32}^{\text{наблюдаемая}} - n_{32}^{\text{ожидаемая}})^2}{n_{32}^{\text{ожидаемая}}} =$$

$$= \frac{(491-554)^2}{554} + \frac{(265-202)^2}{202} + \frac{(938-903)^2}{903} +$$

$$+ \frac{(294-329)^2}{329} + \frac{(363-335)^2}{335} + \frac{(95-123)^2}{123} =$$

$$= 7{,}16 + 19{,}65 + 1{,}36 + 3{,}72 + 2{,}34 + 4{,}06 = 38{,}29.$$

В-пятых, определяем число степеней свободы для имеющейся таблицы сопряжённости: $df = (3 - 1)(2 - 1) = 2$.

В-шестых, по таблице (ниже приведен её фрагмент) находим критическое значение (критическую точку) — это значение указано на пересечении столбца, соответствующего доверительной вероятности 0,95 и уровню значимости 0,05 и строки, соответствующей числу степеней свободы 2. Получим: $\chi^2_{\text{критическое}} = 5{,}99$ (табл. 6.3).

Таблица 6.3. **Фрагмент таблицы критических точек для χ^2-распределения**

df	Доверительная вероятность и уровень α			
	$\gamma = 0{,}9$ (90 %)	$\gamma = 0{,}95$ (95 %)	$\gamma = 0{,}99$ (99 %)	$\gamma = 0{,}999$ (99,9 %)
	$\alpha = 0{,}1$	$\alpha = 0{,}05$	$\alpha = 0{,}01$	$\alpha = 0{,}001$
1	2,71	3,84	6,63	10,83
2	4,61	**5,99**	9,21	13,82
3	6,25	7,81	11,34	16,27
4	7,78	9,49	13,28	18,47
5	9,24	11,07	15,09	20,52

И наконец, сопоставляем два значения: $\chi^2_{\text{наблюдаемое}} > \chi^2_{\text{критическое}}$: $38{,}29 >> 5{,}99$, поэтому при заданном уровне значимости 5 % H_0 отвергается в пользу H_1: «Связь между признаками есть». При имеющихся данных можно говорить, что ***расхождение между эмпирическими и теоретическими частотами носит <u>не</u>случайный характер***, ***это расхождение ощутимо велико***, ***статистически достоверно***, ***статистически значимо.*** Следовательно, гипотеза об отсутствии связи не кажется правдоподобной и убедительной. Поэтому на уровне значимости (на уровне ошибки) $\alpha = 0{,}05$ мы её отклоняем и принимает альтернативную. Итак, H_0 не принимается.

Формулируем содержательный вывод решения: отношение к присвоению найденных вещей или денег *связано* с возрастом человека. Молодёжь и старшее поколение РФ относятся к этому поступку *по-разному*. Возраст и рассматриваемая социальная установка являются взаимосвязанными.

§ 6.3. Проверка гипотезы о равенстве дисперсий; *F*-критерий Фишера

Ранее отмечалось, что понятие дисперсии в ряде случаев является для социологов гораздо более важным понятием (с точки зрения глубины содержательного смысла, который в нем содержится), чем, например, среднее значение. Дисперсия крайне важна потому, что позволяет оценивать, в какой мере изучаемая совокупность харак-

теризуется ***стабильностью***, ***устойчивостью***, ***постоянством***, ***однородностью или же***, ***напротив***, ***их отсутствием.*** Вариация (изменчивость, рассеяние значений) напрямую указывает на *выраженность социальных различий*, *социальную однородность или*, *наоборот*, *неоднородность*, *дифференциацию*, которая (в свою очередь) требует от исследователей дальнейшего *объяснения причин* наблюдаемой неоднородности. Сравнивая между собой какие-либо общности людей по тем или иным показателям их жизнедеятельности, важно научиться уделять внимание сопоставлению дисперсий, характеризующих изменчивость интересующих нас показателей в этих социумах. В проверке гипотезы о равенстве дисперсий речь будет идти о том, чтобы выяснить, *в какой мере* схожи (т. е. одинаковы) однородности двух групп (общностей), рассматриваемых в исследовании.

Постановка задачи может быть сформулирована следующим образом: рассматриваются ***две независимые выборки*** (объёмами n_1 и n_2 и извлечённые из двух генеральных совокупностей), в каждой из которых зафиксированы значения интересующего признака, измеренного в интервальной (числовой) шкале. ***Примерами*** таких признаков могут быть:

- стоимость 1 м^2 жилья экономкласса в регионе А (*первая* совокупность) и в регионе Б (*вторая* совокупность);
- среднесуточная температура зимой в городе И (*первая* совокупность) и в городе К (*вторая* совокупность);
- число аварий (дорожно-транспортных происшествий), в которых побывали за последний год водители — мужчины (*первая* совокупность) и женщины (*вторая* совокупность);
- число дорожно-транспортных происшествий, в которых побывали за последний год водители: имеющие стаж вождения не более трёх лет (*первая* совокупность) и имеющие стаж более трёх лет (*вторая* совокупность);
- объём продаж в выходные дни в обувном магазине № 1 (*первая* совокупность) и в обувном магазине № 2 (*вторая* совокупность);

Цель исследования состоит ***в сравнении дисперсий*** в двух рассматриваемых совокупностях. Можно ли считать, что для обеих генеральных совокупностей характерна одинаковая изменчивость (т. е. равная дисперсия, равная вариация, рассеивание) значений исследуемого признака? Основная статистическая гипотеза формулируется как H_0: «$\sigma_1^2 = \sigma_2^2$ (дисперсии *равны*)»[1]. Конкурирующей гипотезой может стать как ***ненаправленная***, так и ***направленная альтернативы*** (когда основной гипотезе противопоставляется гипотеза о *превалировании* какой-либо из двух дисперсий): H_1: «$\sigma_1^2 \neq \sigma_2^2$ (дисперсии не *равны*)»; H_1: «$\sigma_1^2 > \sigma_2^2$ (дисперсия в первой

[1] σ^2 (сигма в квадрате) — истинное значение дисперсии в генеральной совокупности.

совокупности *больше*, чем дисперсия во второй совокупности)» или H_1: «$\sigma_2^2 > \sigma_1^2$ (дисперсия во второй генеральной совокупности *больше*, чем дисперсия в первой генеральной совокупности)». В зависимости от альтернативы выбирается ***двусторонняя*** или ***односторонняя критическая область*** (т. е. уровень значимости в «хвостах» распределения будет $\alpha/2$ или просто α).

Важно! Направленные и ненаправленные альтернативы. Когда основной гипотезе противостоит альтернативное утверждение о том, что дисперсии <u>не</u> равны (т. е. ей противопоставляется *<u>не</u>*направленная конкурирующая гипотеза), то с содержательной точки зрения это означает следующее. Мы в равной степени допускаем (т. е. верим), что *одинаково возможны* два случая: первая дисперсия больше, чем вторая (в генеральной совокупности), и вторая дисперсия больше, чем первая (в генеральной совокупности). При этом не отдаем чёткого приоритета одной из этих двух возможных ситуаций. В случае же когда альтернативой выступает направленная конкурирующая гипотеза, мы полагаем (убеждены), что *только одна из* дисперсий (либо *только первая* либо *только вторая*) может быть больше (или, наоборот, меньше) в сравнении с другой. ***Характер выдвигаемых альтернатив*** (***направленных или ненаправленных***) ***продиктован***, ***как правило***, ***соображениями здравого***, ***содержательного смысла рассматриваемой в исследовании конкретной ситуации.*** Ход этих размышлений абсолютно аналогичен и для проверки статистических гипотез о других параметрах (математических ожиданий, долей, коэффициентов корреляции и пр.).

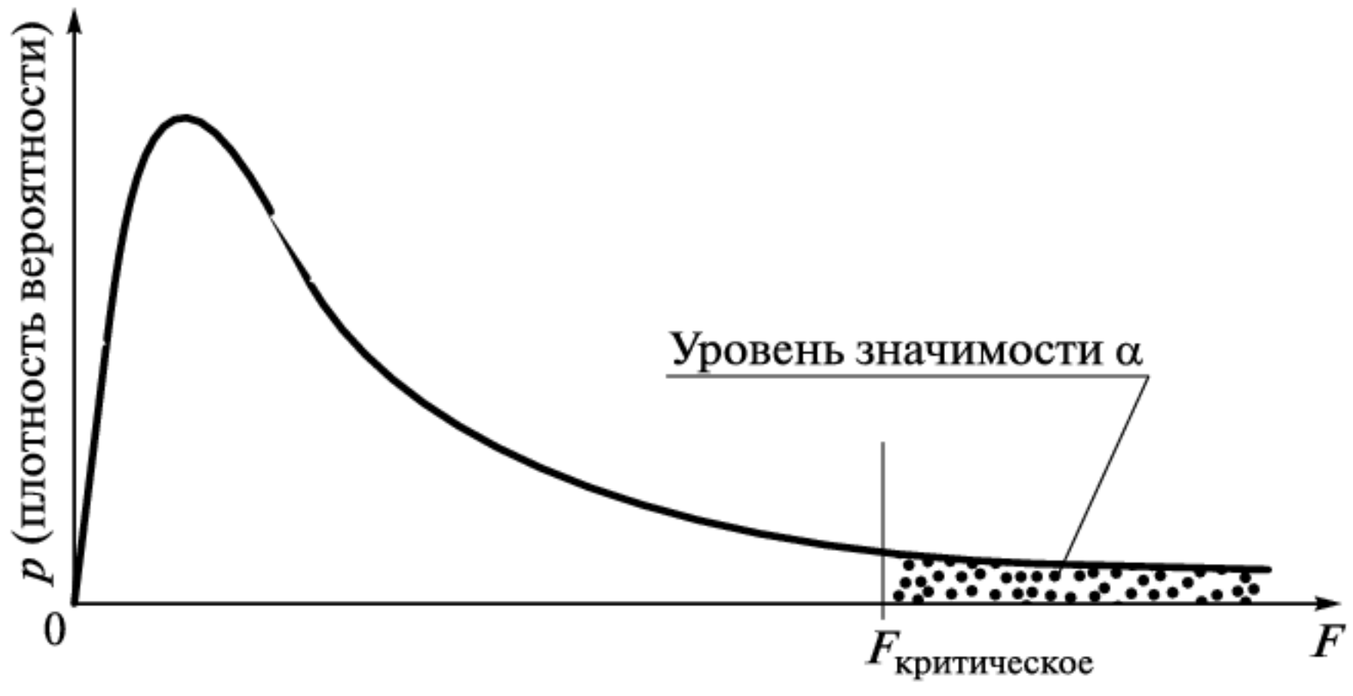
В основе статистического критерия, предложенного для проверки нулевой гипотезы о равенстве дисперсий, лежит функция от наблюдений — статистика критерия, опирающаяся на выборочные оценки дисперсий в каждой из двух совокупностей. Эта функция вычисляется как отношение выборочных оценок дисперсий — одной и другой выборки: $\frac{s_1^2}{s_2^2}$, которое позволяет определить, ***во сколько раз*** одна выборочная оценка дисперсии отличается от другой (в частности, *превосходит* её). ***В числитель обычно ставят бóльшую дисперсию***, ***в знаменатель — меньшую***[1]. Так устроен ***статистический F-критерий Фишера***, предложенный английским математиком Рональдом Фишером (1890 — 1962). *F*-критерий определяется отношением двух случайных величин: ***отношением двух выборочных оценок дисперсий***. Распределение вероятностей значений этой статистики критерия при нулевой гипотезе описывается *F-распределением Фишера* (*распределение дисперсионного отношения*), которое относится к группе непрерывных распределений с **двумя параметрами** — *числом степеней свободы первой выборки*

---

[1] Но это не обязательно.

$\left(df_1 = n_1 - 1 \text{ для числителя дроби } \frac{s_1^2}{s_2^2}\right)$ и *числом степеней свободы другой выборки* $\left(df_2 = n_2 - 1 \text{ для знаменателя дроби } \frac{s_1^2}{s_2^2}\right)$. По форме F-распределение является асимметричным (степень асимметрии диктуется соотношением двух степеней свободы — первой и второй

Таблица 6.4. **Фрагмент таблицы критических точек для *F*-распределения Фишера (для уровня значимости $\alpha = 0{,}05$ (5 %))**



df		Число степеней свободы для числителя дроби (выборка №1): df_1										
		1	2	3	4	5	6	7	8	9	10	...
Число степеней свободы для знаменателя дроби (выборка № 2): df_2	1	161,5	199,5	215,7	224,6	230,2	234	236,8	238,9	240,5	241,9	и т. д.
	2	18,51	19,00	19,16	19,25	19,30	19,33	19,35	19,37	19,38	19,40	
	3	10,13	9,55	9,28	9,12	9,01	8,94	8,87	8,85	8,81	8,79	
	4	7,71	6,94	6,59	6,38	6,26	6,16	6,09	6,04	6,00	5,96	
	5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,77	4,74	
	6	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10	4,06	
	7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,64	
	8	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39	3,35	
	9	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,14	
	10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,98	
	...	и т. д.										

выборки). Значения F-критерия всегда ≥ 0, так как ни одна из дисперсий не может быть отрицательной. Дисперсия всегда величина неотрицательная.

Проверка нулевой гипотезы осуществляется на основе выбранного аналитиком уровня доверия (γ), указывающего на степень надёжности статистического вывода, и построенной критической области. Критическая (пограничная) точка определяется по специальной таблице для F-распределения Фишера. Составлены статистические таблицы для критических точек — в зависимости от того, какой уровень значимости α требуется при проверке статистической гипотезы. В частности, наиболее распространенными являются таблицы для случаев, когда требуемый уровень значимости равен **0,1; 0,05; 0,025** или **0,01**[1]. Поэтому прежде чем определить критическую точку, нужно выбрать нужную таблицу с требуемым уровнем значимости.

В столбцах табл. 6.4 указывается число степеней свободы для числителя дроби $\frac{s_1^2}{s_2^2}$ (т. е. число ограничений свободы вариации для одной выборки), в строках — число степеней свободы для знаменателя дроби (т. е. число ограничений свободы вариации для другой выборки). На пересечении соответствующих строки и столбца определяется искомое критическое значение F-статистики — $F_{\text{критическое}}$.

Обратите внимание, что отношение выборочных оценок дисперсий будет больше 1, если $s_1^2 > s_2^2$, и меньше 1, если $s_1^2 < s_2^2$. Поэтому для «удобства» выборочные оценки дисперсий можно «поменять местами» — при вычислении статистики F-критерия ***в числитель дроби*** подставляется выборочная оценка дисперсии той выборки, где эта ***дисперсия больше***, а ***в знаменатель*** — соответственно оценка дисперсии той выборки, где ***дисперсия меньше***. В результате значение F-статистики ***оказывается не меньшим 1***. Однако подчеркнём, что деление большей оценки дисперсии на ме́ньшую оправдано лишь для односторонних альтернатив, а для двусторонних этого делать не обязательно. В качестве критической (пограничной) точки выбирается либо $F_{\text{критическое}}$ (для случая, если нулевая гипотеза проверяется против направленной альтернативы, и тогда строится односторонняя критическая область), либо $F_{\text{критическое}}$ и $\frac{1}{F_{\text{критическое}}}$ (для случая, если нулевая гипотеза проверяется против ***не***направленной альтернативы, и тогда строится двусторонняя критическая область).

Итак, когда альтернативная гипотеза является ***ненаправленной*** (говорят, что имеется ***двусторонняя критическая область***), типичными, т. е. характе́рными, для распределения являются значения

---

[1] Перечень таблиц для F-распределения (для $\alpha = 2{,}5\,\%;\ 5\,\%$) см. в прил. 2.

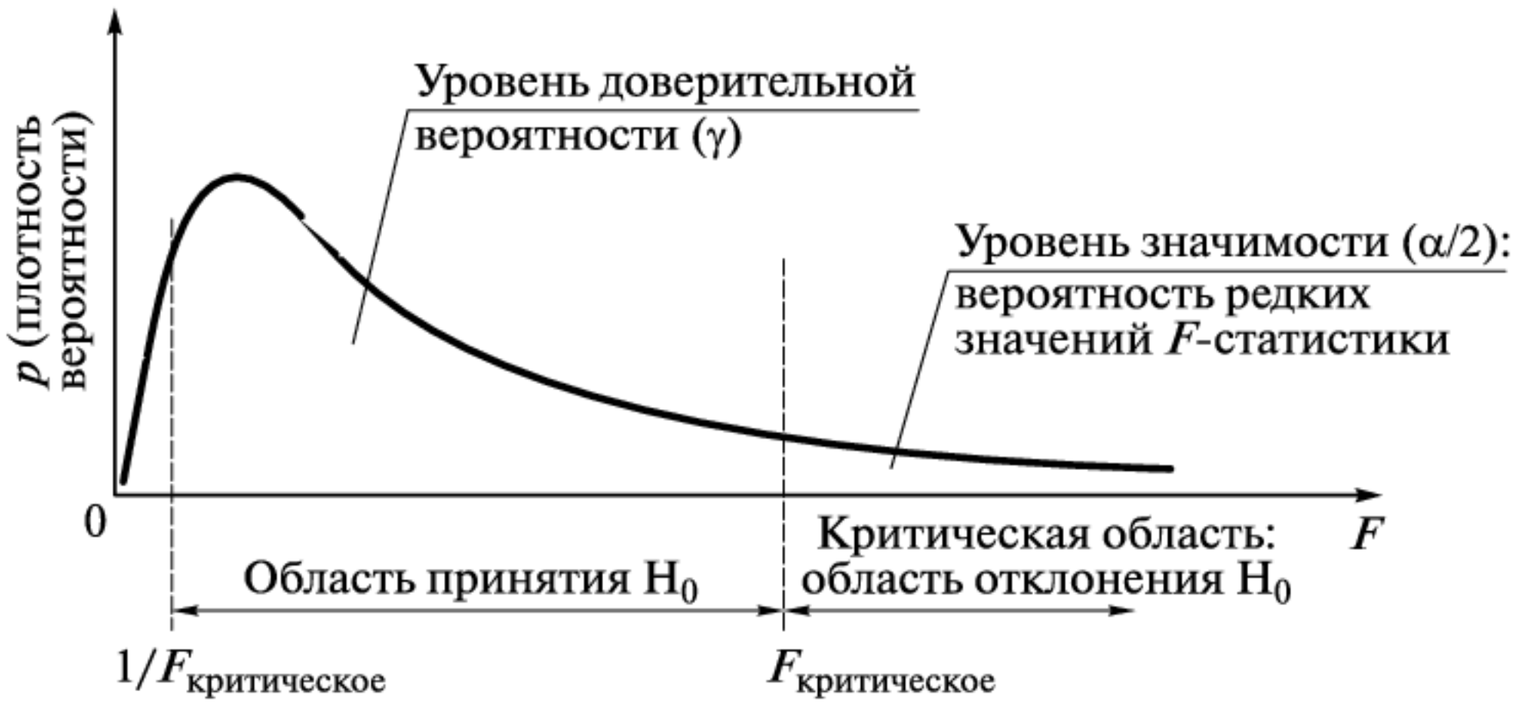


Рис. 6.6. *F*-распределение Фишера и построение двусторонней критической области

F-критерия, которые лежат в диапазоне $\left(\frac{1}{F_{\text{критическое}}};\ F_{\text{критическое}}\right)$ (рис. 6.6). Критическую область (т. е. область отклонения нулевой гипотезы) составляют такие значения, которые не охватываются этим диапазоном. Поэтому в тех ситуациях, когда наблюдаемое значение *F*-критерия «помещается» в этот интервал, считается, что *нет оснований* отклонять нулевую гипотезу. Так, гипотеза о равенстве дисперсий, а, следовательно, и предположение о случайности расхождений двух выборочных оценок дисперсий, считаются верными, и наоборот. Если наблюдаемое значение статистики *F*-критерия выходит за пределы обозначенного интервала, основную гипотезу следует отклонить и принять альтернативную.

Для случая, когда альтернативная гипотеза является ***направленной*** (говорят, что имеется ***односторонняя критическая область***), наблюдаемое значение *F*-критерия нужно сравнивать с табличным значением $F_{\text{критическое}}$ (рис. 6.7).

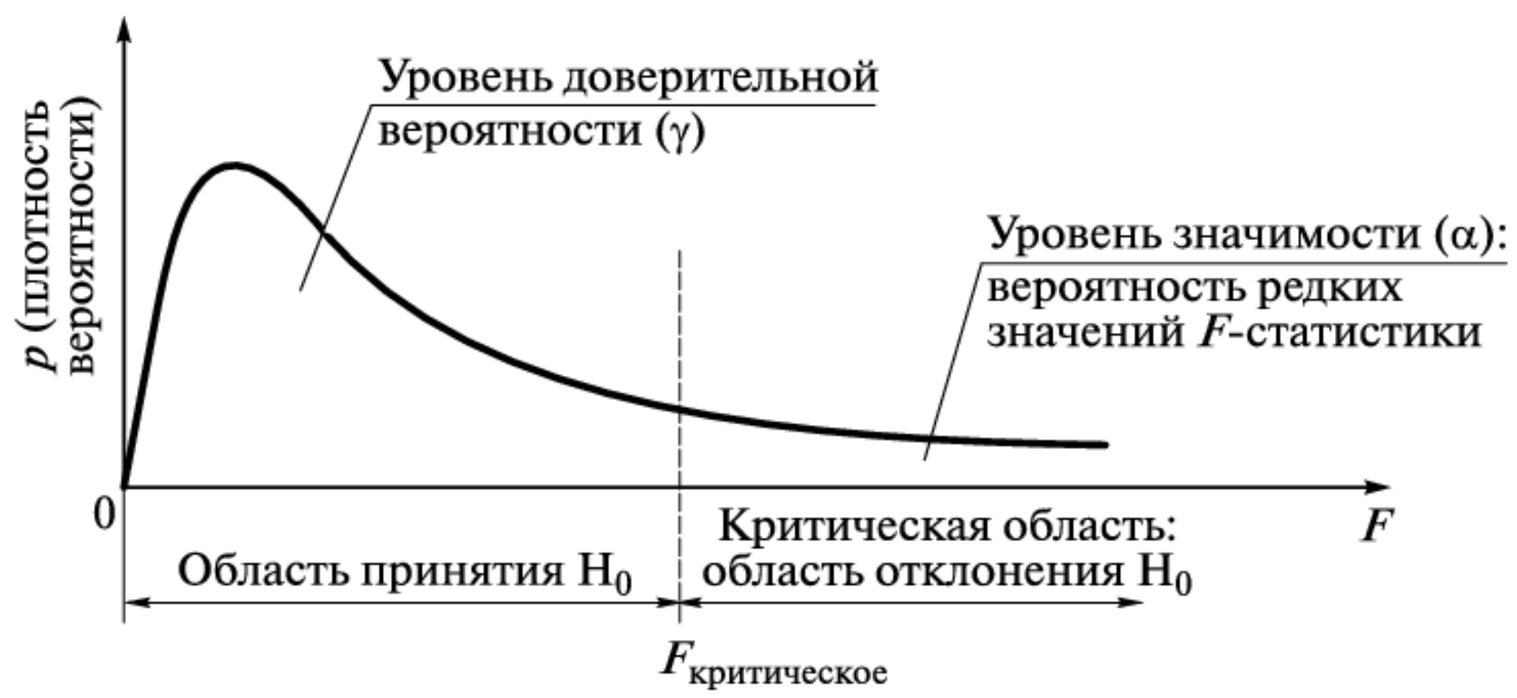


Рис. 6.7. *F*-распределение Фишера и построение односторонней критической области

Пример 6.4. «Ниже прожиточного минимума». Дифференциация (неоднородность) доходов населения является важным социальным показателем. В этом примере представлена доля населения с денежными доходами ниже прожиточного минимума (в процентах от общей численности населения субъекта) в регионах двух федеральных округов Российской Федерации — Сибирском и Северо-Западном[1] (данные приведены за 2010 г.). Сравним два федеральных округа между собой: насколько они отличаются с точки зрения дифференциации величин прожиточного минимума (табл. 6.5).

В качестве одного из показателей, который мы рассмотрим, будет величина *разброса* (*дисперсии*) доли населения, с доходами ниже прожиточного минимума, в регионах двух рассматриваемых федеральных округов. Величина разброса будет указывать на *степень однородности* ситуации, наблюдавшейся в регионах *каждого из* федеральных округов в 2010 г. Правильным ли будет утверждение, что исследуемые федеральные округа не отличаются друг от друга с точки зрения одинаковой вариации (дисперсии) доли населения с доходами ниже величины прожиточного минимума? Можно ли утверждать, что оба округа характеризуются одинаковой степенью *однородности* социально-экономического расслоения в регионах?

Решение. Исследование затрагивает вопрос единообразия в имущественном расслоении[2] в двух субъектах РФ: изучаются дисперсии доли людей (%) с доходами ниже прожиточного минимума. Есть ли у нас основания утверждать, что дисперсия этого показателя в регионах Северо-Западного федерального округа ниже, чем в регионах Сибирского Федерального округа (т. е. в регионах Сибирского ФО этот показатель менее стабилен, имеет заметно бóльшую вариацию, колебания, нежели в регионах Северо-Западного ФО за тот же период)? Действительно ли регионы Сибирского ФО менее однородны и имеют бóльшую дифференциацию в сравнении друг с другом?

Проверку соответствующей гипотезы реализуем на уровне доверия 95%.

Зафиксируем нулевую и конкурирующую гипотезы в задаче:

H_0: «$\sigma_1^2 = \sigma_2^2$ (дисперсии *равны*)»; H_1: «$\sigma_1^2 > \sigma_2^2$ (дисперсия в первой совокупности *больше*, чем дисперсия во второй совокупности)».

Вычислим средние значения, а далее и значения выборочных дисперсий в двух округах:

---

[1] С данными можно ознакомиться в публикациях Федеральной службы государственной статистики в разделе «Регионы России. Социально-экономические показатели», см.: [http://www.gks.ru/bgd/regl/b11_14p/IssWWW.exe/Stg/d01/05-11.htm] (проверено 03.08.2012).

Приведённые данные получены на основе выборочного обследования: используются сведения о бюджетах домашних хозяйств и показателях денежных доходов населения.

[2] Имеется в виду имущественное расслоение, естественно, в упрощённом виде.

Таблица 6.5. **Дифференциация доходов населения для СФО и СЗФО**

Регионы Сибирского федерального округа	Доля (%) населения с доходами ниже прожиточного минимума (X_1)	Регионы Северо-Западного федерального округа	Доля (%) населения с доходами ниже прожиточного минимума (X_2)
Республика Алтай	17,9	Архангельская обл.	14,3
Республика Бурятия	19,8	Вологодская обл.	17,0
Республика Тыва	30,0	Калининградская обл.	12,6
Республика Хакасия	18,5	Ленинградская обл.	12,8
Алтайский край	24,3	Мурманская обл.	13,0
Забайкальский край	19,3	Новгородская обл.	14,9
Красноярский край	18,4	Псковская обл.	15,8
Иркутская обл.	18,4	Республика Карелия	15,2
Кемеровская обл.	10,9	Республика Коми	14,7
Новосибирская обл.	16,7	Санкт-Петербург	8,7
Омская обл.	14,3		
Томская обл.	17,4		
Всего	$n_1 = 12$ наблюдений	Всего	$n_2 = 10$ наблюдений

$$\bar{X}_1=\frac{17,9+19,8+30,0+...+14,3+17,4}{12}\approx 18,8\ (\%)$$

— такова в среднем доля (%) населения с доходами ниже прожиточного минимума в каждом из рассматриваемых регионов Сибирского ФО;

$$\bar{X}_2=\frac{14,3+17,0+12,6+...+14,7+8,7}{10}=13,9\ (\%)$$

— такова в среднем доля (%) населения с доходами ниже прожиточного минимума в каждом из рассматриваемых регионов Северо-Западного ФО.

Находим несмещённые выборочные оценки дисперсий и стандартных отклонений:

$$s_1^2=\frac{(17,9-18,8)^2+(19,8-18,8)^2+(30,0-18,8)^2+...+(17,4-18,8)^2}{12-1}\approx 22,5(\%)^2$$

(тогда стандартное отклонение составит: $s_1=\sqrt{22,5}\approx 4,74(\%)$);

$$s_2^2=\frac{(14,3-13,9)^2+(17,0-13,9)^2+(12,6-13,9)^2+...+(8,7-13,9)^2}{10-1}\approx 5,27\ (\%)^2$$

(тогда стандартное отклонение составит: $s_2=\sqrt{5,27}\cong 2,3\ (\%)$).

Найдем наблюдаемое значение статистики *F*-критерия:

$$F_{\text{наблюдаемое}}=\frac{s_1^2}{s_2^2}=\frac{22,5}{5,27}\approx 4,27.$$

Значение 4,27 показывает, что выборочная оценка дисперсии исследуемого показателя в Сибирском ФО в 4,27 раз превышает выборочную оценку аналогичного показателя в Северо-Западном ФО. А много это или мало? Как определить, являются ли эти различия в дисперсиях случайными, несущественными? Или же эти наблюдаемые различия, напротив, статистически достоверны (значимы, весомы, ощутимы)? Чтобы ответить на поставленный вопрос, следует определить, попадает ли наблюдаемое значение статистики *F*-критерия в критическую область или нет. Чтобы задать критическую область, найдём по статистической таблице критическое (пограничное) значение: $F_{\text{критическое}}=\{df_1=n_1-1=12-1=11,\ df_2=n_2-1=10-1=9,\ \alpha=0,05\}=\{$в столбцах таблицы для $\alpha=5\,\%$ находим число степеней свободы для дисперсии первой выборки, т.е. Сибирского ФО, равное 11, а в строках — число степеней свободы для дисперсии второй выборки, т.е. Северо-Западного ФО, равное $9\}=3,1$. Обратите внимание, что в качестве уровня значимости выбирается именно 5 % (как 1 – 0,95), а не 2,5 % ((1 – 0,95)/2), так как имеем направленную альтернативу H_1: «$\sigma_1^2>\sigma_2^2$ (а следовательно, и рассматриваем таблицу *F*-критерия, соответствующую $\alpha=5\,\%$). От-

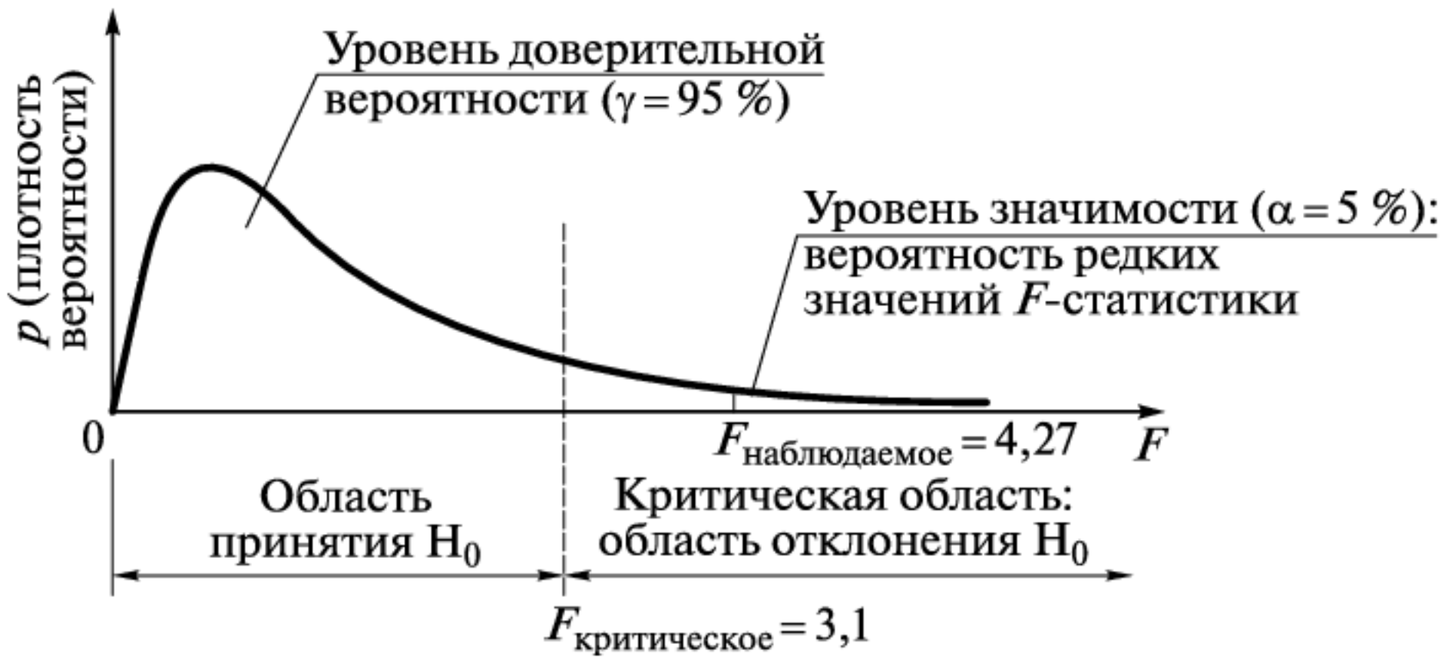


Рис. 6.8. Построение односторонней критической области (см. пример 6.4 «Ниже прожиточного минимума»)

метим наблюдаемое и критическое значения на схематичном графике плотности распределения вероятностей F-статистики (рис. 6.8).

Итак, по результатам имеющихся расчётов видим, что значение 4,27 принадлежит критической области, так как $4{,}27 > 3{,}1$: $F_{\text{наблюдаемое}} > F_{\text{критическое}}$, значит, можно заключить, что есть основания отвергнуть гипотезу H_0. При имеющихся данных различия в дисперсиях статистически достоверны. На уровне значимости 5 % можно утверждать, что *разброс* показателя доли населения с доходами ниже прожиточного минимума в регионах двух федеральных округов — Сибирском и Северо-Западном — различается. Федеральные округа не являются одинаково однородными по этому показателю. Регионы Северо-Западного ФО *более однородны* (похожи между собой) по рассмотренному показателю социально-экономического расслоения, нежели регионы Сибирского ФО — они более разнородны, т. е. менее одинаковы.

Пример 6.5. «Лояльность». Рассмотрим один из таких важных показателей работы магазинов фирменной сети, продающей коллекцию обуви торговой марки А*, как лояльность (преданность) клиентов этой торговой марке. Пусть по двум магазинам фирменной сети, расположенным в двух разных городах, у нас имеется динамика данных (за несколько полугодий) о доле лояльных покупателей. Будет ли верным утверждение, что исследуемые магазины (№ 1 и № 2) не отличаются друг от друга с точки зрения стабильности (т. е. постоянства вариации) доли лояльных клиентов? Имеются две выборки данных, описывающих работу двух магазинов за некоторый период времени:

Магазин	Полугодие						
	1-е	2-е	3-е	4-е	5-е	6-е	7-е
№1	0,40	0,35	0,38	0,42	0,38	0,45	0,52
№2	0,32	0,28	0,20	0,25	0,38	0,40	0,47

Можно ли сказать, что доля лояльных покупателей, характерная для магазина №1, более постоянна (неизменна, стабильна), в сравнении с долей лояльных покупателей в магазине № 2? Проверку соответствующей гипотезы осуществите на уровне доверительной вероятности 0,95 против направленной альтернативы.

Решение. Выпишем нулевую и конкурирующую гипотезы в задаче:

H_0: «$\sigma_1^2 = \sigma_2^2$ (дисперсии *равны*)»;

H_1: «$\sigma_2^2 > \sigma_1^2$»(дисперсия в первой совокупности *меньше*, чем дисперсия во второй совокупности; иными словами, дисперсия второй совокупности превосходит дисперсию первой)».

Вычислим средние значения, а далее и значения выборочных оценок дисперсий доли лояльных покупателей в двух магазинах:

$$\bar{X}_1 = \frac{0,40 + 0,35 + 0,38 + 0,42 + \ldots + 0,52}{7} \approx 0,41\ (41\,\%)$$

— средняя доля лояльных потребителей в магазине № 1, приходящаяся на одно полугодие;

$$\bar{X}_2 = \frac{0,32 + 0,28 + 0,20 + 0,25 \ldots + 0,47}{7} \approx 0,33\ (33\,\%)$$

— средняя доля лояльных потребителей в магазине № 2, приходящаяся на одно полугодие.

Находим несмещенные выборочные оценки дисперсий и стандартных отклонений:

$$s_1^2 = \frac{(0,40-0,41)^2 + (0,35-0,41)^2 + (0,38-0,41)^2 + \ldots + (0,52-0,41)^2}{7-1} = 0,003;$$

$$s_2^2 = \frac{(0,32-0,33)^2 + (0,28-0,33)^2 + (0,20-0,33)^2 + \ldots + (0,47-0,33)^2}{7-1} \approx 0,009$$

(тогда стандартные отклонения составят: $s_1 = \sqrt{0,003} = 0,055$ (т.е. 5,5 %); $s_2 = \sqrt{0,009} = 0,095$ (т.е. 9,5 %)).

Найдем наблюдаемое значение статистики *F*-критерия, равное отношению двух выборочных оценок дисперсии:

$$F_{\text{наблюдаемое}} = \frac{s_2^2}{s_1^2} = \frac{0,009}{0,003} = 3 \text{ (раз)}.$$

Далее следует задать критическую точку: $F_{\text{критическое}}$. По статистической таблице Фишера находим, что $F_{\text{критическое}} = \{df_2 = n_2 - 1 = 7 - 1 = 6,\ df_1 = n_1 - 1 = 7 - 1 = 6,\ \alpha = 0,05\}$ = {в столбцах таблицы для $\alpha = 5\,\%$ находим число степеней свободы для дисперсии ***второй*** выборки, т.е. магазина № 2, равное 6; а в строках — число степеней свободы для дисперсии первой выборки, т.е. магазина № 1, также

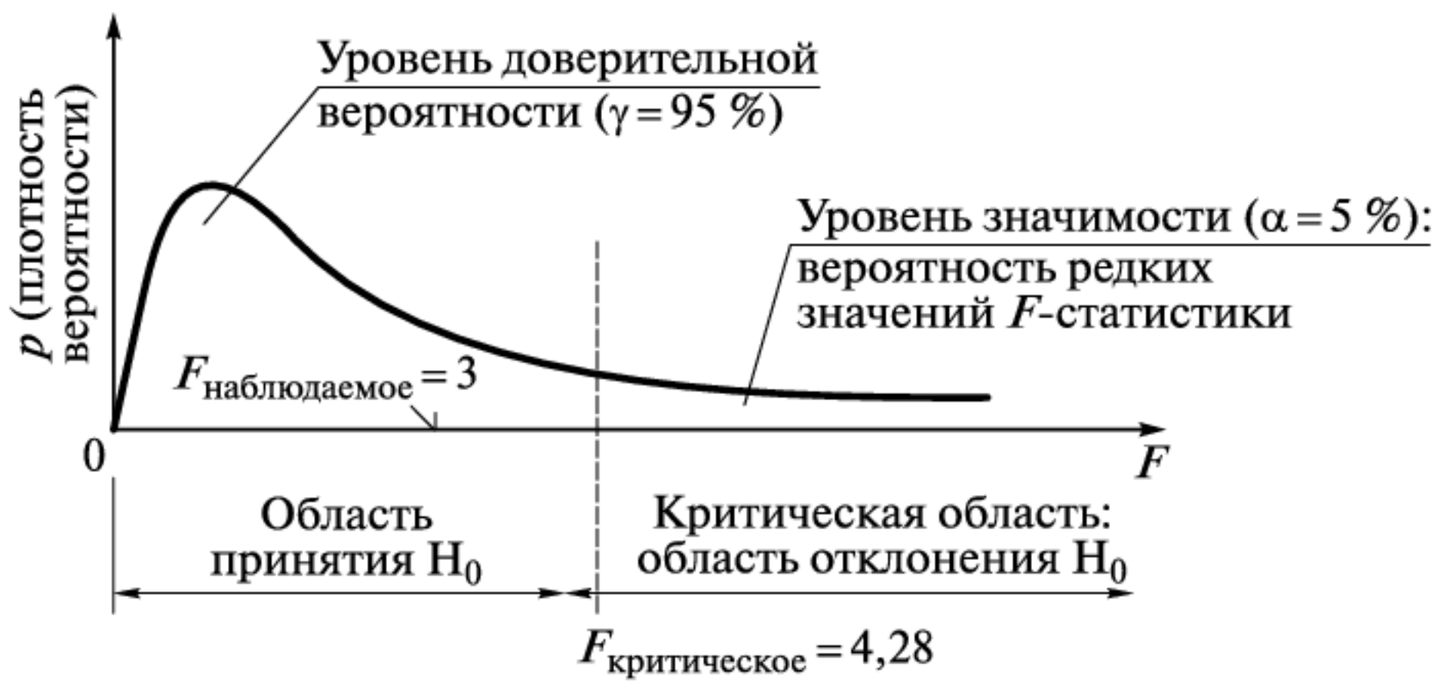


Рис. 6.9. Построение односторонней критической области (см. пример 6.5 «Лояльность»)

равное 6, поскольку дисперсии мы поменяли местами при вычислении дисперсионного отношения, чтобы получить положительное значение статистики критерия} = 4,28. Обратите внимание, что в качестве уровня значимости выбирается именно 5 % (как 1 – 0,95), а не 2,5 % (как (1 – 0,95)/2), так как имеем направленную конкурирующую гипотезу о превосходстве второй дисперсии над первой (а следовательно, рассматриваем таблицу F-критерия, соответствующую именно $\alpha = 5\,\%$).

Итак, по результатам имеющихся расчётов видим, что значение 3 не попадает в критическую область, так как $3 < 4{,}28$: $F_{\text{наблюдаемое}} < F_{\text{критическое}}$, значит, можно заключить, что нет оснований отвергать гипотезу H_0. При имеющихся данных нет оснований считать различия в выборочных оценках дисперсии ощутимыми, т. е. статистически достоверными. Отметим наблюдаемое и критическое (пограничное) значения на графике распределения вероятностей F-статистики (рис. 6.9).

На уровне доверия 95 % можно утверждать, что *разброс* показателей доли лояльных покупателей в двух рассмотренных магазинах одинаковый; вариация не различается. Действительно, на уровне доверия 95 % будет верным утверждение, что результаты работы исследуемых магазинов (№ 1 и № 2) не отличаются друг от друга с точки зрения стабильности (т. е. постоянства вариации) доли лояльных клиентов от одного полугодия к другому (при имеющихся данных).

§ 6.4. Проверка гипотезы о равенстве средних. Независимые выборки и *t*-критерий Стьюдента

В данном параграфе изучим алгоритм проверки ещё одной важной статистической гипотезы, а именно: гипотезы *о равенстве средних величин*. Постановку задачи начнём с небольшого *примера*:

предположим, что на основе эмпирических данных *выборочного* обследования бюджетов домашних хозяйств оценка средних показателей денежных доходов, приходящихся на одного члена семьи, в регионе А* составила 11 140 руб., а в регионе Б* — 12 080 руб. Можно ли утверждать, что уровень материального благополучия во втором регионе (Б*) действительно несколько выше, чем в первом (А*), ведь 12 080 руб. > 11 140 руб.? Ответ на вопрос задачи потребует вычисления статистической погрешности сделанных в ходе исследования измерений. Следует посмотреть на две эти группы (регион А* и регион Б*) как на две разные генеральные совокупности. Данные, которые получены на выходе двух исследований, — это *выборочные* оценки. Вместе с тем напомним, что наша основная задача — перейти от оценок для выборки к выводам, справедливым для всей генеральной совокупности. Поэтому для генеральной совокупности мы должны проверить истинность предположения ***о равенстве средних величин на самом деле. Можем ли мы считать, что в генеральной совокупности действительно нет различий?*** Или же различия статистически выражены и потому в реальности присутствуют?

Один из способов найти ответ на этот вопрос вам уже известен — это *построение доверительных интервалов и сравнение их между собой*. В том случае, если интервалы идут *сильно* «внахлёст» по отношению друг к другу или *заметно* пересекаются, то можно с определенной долей вероятности (как правило, высокой) утверждать, что различий на самом деле нет (относительно рассматриваемого примера это будет означать, что истинные средние показатели в обоих регионах ***одинаковые***). Если же доверительные интервалы *не* пересекаются, то напрашивается противоположный вывод — истинные средние значения среднедушевых доходов в этих регионах действительно ***разные***. При этом не до конца ясно, какой вывод можно сделать, когда доверительные интервалы для двух средних значений хотя и пересекаются, но в целом заметно не совпадают. Общий, универсальный способ решения этой задачи даёт ***методика проверки статистической гипотезы о равенстве средних с помощью специального статистического критерия***:

- проверяется *нулевая* (*т. е. основная*) *гипотеза* вида: H_0: «$\mu_1 = \mu_2$ (средние *равны*)» (читается как «мю один *равно* мю два»), где μ_1 — *истинное значение среднего* (математическое ожидание) в *первой* генеральной совокупности (в 1-й группе), а μ_2 — *истинное значение среднего* (математическое ожидание) во *второй* генеральной совокупности (во 2-й группе). Нулевой гипотезе противопоставляется альтернативная гипотеза, которая *может быть как направленной, так и ненаправленной*;
- *конкурирующая гипотеза* (т. е. противоречащая нулевой гипотезе) записывается следующим образом: H_1: «$\mu_1 \neq \mu_2$ (средние *не* равны)» и читается как «мю один *не равно* мю два». В этом случае она будет называться *ненаправленной* (в её формулировке специально

не подчёркивается, какое именно из двух средних больше в генеральной совокупности — первое или второе). Если же конкурирующая гипотеза является направленной, то она выглядит так: H_1: «$\mu_1 > \mu_2$ (первое среднее *больше* второго среднего)» и читается как «мю один *больше* мю два». Или альтернативная направленная гипотеза выглядит так: H_1: «$\mu_2 > \mu_1$ (второе среднее *больше* первого среднего)» и читается как «мю два *больше* мю один»;

- от того, какая альтернативная гипотеза выбирается — направленная или ненаправленная — зависит, какой окажется критическая область — односторонней или двусторонней соответственно. Этот тезис будет пояснён в ходе решения задач;
- в основе проверки нулевой гипотезы лежит статистика критерия следующего вида: $t = \left(\bar{X}_1 - \bar{X}_2 \right) \Big/ \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$, либо некоторые модификации статистики этого критерия[1].

С содержательной точки зрения критерий устроен просто: в числителе стоит арифметическая разность двух выборочных средних (разность оценок). Эта разность фиксирует различие между двумя средними величинами в абсолютном выражении (в тех единицах, в которых они измеряются — например, рубли, баллы, штуки, килограммы, километры и т. д.) и отвечает на вопрос: *на сколько?* одно среднее больше второго. Далее разность нормируется на своё стандартное отклонение: $\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$ (см. знаменатель дроби). Процедура деления разности двух средних на стандартное отклонение позволяет оценить в *относительном* выражении, *в какой именно мере* (*как сильно*) различаются две выборочные оценки среднего между собой, т. е. помогает оценить *интенсивность* (*выраженность*) присутствующих различий. Таким образом, исследователем рассматривается ***разность между двумя средними***, ***выраженная в единицах стандартного отклонения этой разности***. Выражение $t = \left(\bar{X}_1 - \bar{X}_2 \right) \Big/ \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$ является ***случайной величиной***, распределённой по закону Стьюдента[2].

---

[1] Эти модификации определяются тем, какие дополнительные предположения делаются о равенстве дисперсий (об этом см. далее).

[2] Распределённой по закону Стьюдента с числом степеней свободы: $df = \left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} \right)^2 \Big/ \left[\frac{\left(\frac{s_1^2}{n_1} \right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2} \right)^2}{n_2 - 1} \right]$.

Ход рассуждений аналогичен рассмотренным ранее примерам: очевидно, что выражение $(\bar{X}_1 - \bar{X}_2)$ является случайной величиной — оно меняется от одной пары выборок к другой; ***разность двух средних оценок — величина колеблющаяся***. Если представить, что мы многократно (бесконечное число раз) извлекаем из двух генеральных совокупностей две выборки, для каждой из которых считаем оценки $\bar{X}_1$ и $\bar{X}_2$, вычисляем несмещённые оценки дисперсии s_1^2, s_2^2 и рассчитываем наблюдаемое значение величины $\left(\bar{X}_1 - \bar{X}_2\right) \Big/ \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$, то в результате такого эксперимента всем найденным значениям этой величины оказывается возможным противопоставить вероятность их встречаемости. Тем самым можно найти ***закон распределения вероятности этих значений*** (поскольку разности $(\bar{X}_1 - \bar{X}_2)$ начинают быть распределёнными по определённому закону). Математикам известно, что распределение вероятности *всех потенциально возможных значений* $\frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$, которые можно получить в ходе сравнения двух средних, будет прекрасно описываться t-распределением Стьюдента, т. е. подчиняться ему. А зная закон распределения случайной величины, несложно произвести проверку статистической гипотезы о равенстве математических ожиданий.

Теперь давайте внимательнее посмотрим на знаменатель дроби: $\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$. Поскольку мы говорим о том, что разность двух выборочных оценок — величина случайная, значения которой подвержены колебаниям (ведь мы имеем дело с выборками), то чему равна тогда дисперсия этой разности (предполагая, что **две *выборки независимые***)? Для этого вспомним фрагмент курса *Теории вероятностей* о том, что ***дисперсия разности двух независимых случайных величин есть сумма дисперсий этих величин***. Так, $D(X - Y) = D(X) + D(Y)$. Тогда дисперсия разности двух выборочных оценок среднего будет равна: $D(\bar{X}_1 - \bar{X}_2) = D(\bar{X}_1) + D(\bar{X}_2)$. А дисперсию выборочной оценки среднего мы умеем вычислять, и знаем, что она равна $\frac{s^2}{n}$. Дисперсия выборочной оценки среднего в первой выборке составляет $\frac{s_1^2}{n_1}$, а во второй она вычисляется как $\frac{s_2^2}{n_2}$. Тогда дисперсия разности двух средних, равная сумме их дисперсий, должна быть рассчитана как $D(\bar{X}_1 - \bar{X}_2) = D(\bar{X}_1) + D(\bar{X}_2) = \frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}$.

А значит, стандартное отклонение разности будет равно корню квадратному из этой дисперсии: $\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}$. Именно поэтому в знаменателе критерия фигурирует такая запись.

Теперь рассмотрим, как «работает» критерий проверки статистической гипотезы о равенстве средних. Задумаемся, в каких случаях мы получим большое значение статистики t-критерия, лежащее в «хвостах» t-распределения? Очевидно, именно в тех случаях, *когда числитель дроби будет большим* или *знаменатель будет малым*: тогда и вся дробь $\frac{\bar{X}_1-\bar{X}_2}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}}$ будет стремиться к большому числу. Рассуждения о проверке статистической гипотезы базируются на здравом смысле и устроены так: ***чтобы мы смогли отвергнуть нулевую гипотезу***, ***в основе которой лежит предположение о равенстве истинных средних***, ***должны быть близки к нулю дисперсии выборочных оценок этих средних***. Это будет означать, что дисперсии почти нет. В содержательном смысле это означает, что обе выборки (группы) внутри себя *однородны*. Следовательно, каждая из выборочных оценок среднего *типична* (*характерна*) для своей совокупности, а генеральные совокупности — так скажем, *безразбросные*. В этом случае *выборочные оценки* средних с высокой точностью отражают *истинные значения* средних. Поэтому ***если дисперсии малые, то мы очень высоко доверяем оценкам средних в обеих выборках***. А значит, ***верим и различию (т. е. разности) между ними***: $(\bar{X}_1-\bar{X}_2)$. И эта разность становится для нас очень показательной, заметной. А это в свою очередь повышает уверенность в том, что нулевая гипотеза должна быть отклонена: ведь выборочные оценки средних, измеренные с высокой точностью (т. е. малой погрешностью), указывают на то, что истинные значения средних не пересекаются, не совпадают, между ними дистанция, они действительно различны.

Тогда в каких же случаях мы, напротив, получим маленькое наблюдаемое значение статистики t-критерия, сконцентрированное в центре распределения этой случайной величины и/или в области типичных (характерных) для неё значений? Ясно, что именно в случае, ***когда числитель дроби будет малым*** **или же** ***знаменатель получится большим***: тогда и вся дробь $\frac{\bar{X}_1-\bar{X}_2}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}}$ будет стремиться к малому числу. Содержательные рассуждения здесь противоположны: выборочные оценки средних столь схожи, что их разность

$(\bar{X}_1 - \bar{X}_2)$ не позволяет дифференцировать две генеральные совокупности между собой, а значит, требует подтверждения истинности нулевой гипотезы.

Итак, подведём промежуточные итоги: если две выборки являются независимыми, то для проверки статистической гипотезы о равенстве истинных средних (средних в генеральной совокупности) используется *t*-критерий Стьюдента. Наблюдаемое значение статистики критерия рассчитывается по имеющимся данным выборочных обследований. Критическое же значение статистики *t*-критерия определяется по известной вам таблице критических точек *t*-распределения Стьюдента.

Статистический t-критерий для проверки гипотезы о равенстве средних имеет модификации, в зависимости от того, каковы, во-первых, ***размеры двух выборок***, и, во-вторых, от того, ***можно ли считать выборочные оценки дисперсии в двух выборках одинаковыми***.

- ***Если число степеней свободы обеих выборок велико*** ($df_1 \geq 30$ и $df_2 \geq 30$), то статистический критерий имеет вид: $t \to z = \dfrac{\bar{X}_1 - \bar{X}_2}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}}$.

Поскольку речь идет о выборках большого объёма, то вспомним, что при $n > 30$ *t*-распределение Стьюдента мало отличается от стандартного нормального *z*-распределения. Распределение вероятности *всех потенциально возможных значений* величины $\dfrac{\bar{X}_1 - \bar{X}_2}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}}$ будет в достаточной степени хорошо описываться *z*-распределением, т. е. подчиняться ему. Поэтому для построения критической области можно воспользоваться статистической таблицей критических точек *z*-распределения.

- ***Если число степеней свободы хотя бы одной из двух выборок мало*** ($df_1 < 30$ и/или $df_2 < 30$), то для проверки нулевой гипотезы используется ***одна из двух*** модификаций *t*-критерия:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}} \text{ или } t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}\sqrt{\dfrac{s_1^2(n_1 - 1) + s_2^2(n_2 - 1)}{n_1 + n_2 - 2}}}.$$

Какую же из предложенных модификаций критерия следует выбрать и почему вторая запись критерия выглядит столь громоздко и осложнена дополнениями в знаменателе?

Попытаемся дать разъяснения. Для этого вернёмся к двум выборкам, объём хотя бы одной из которых (или обеих) мал. Что значит это обстоятельство для выборочных оценок дисперсии? Это означа-

ет, что хотя бы одна из дисперсий оценена с меньшей точностью (ведь объём выборки мал!), и это, в свою очередь, формирует скептичное к ней отношение: мы недоверчиво относимся к этой выборочной оценке, осознавая содержащуюся в ней бóльшую погрешность. К сожалению, истинные значения дисперсий в обеих генеральных совокупностях нам чаще всего неизвестны. Мы судим о них по выборочным оценкам дисперсии. ***Имеет принципиальное значение, одинаковые ли дисперсии в обеих генеральных совокупностях***. Это важно по следующей причине: *если* представить себе, что *истинные значения дисперсии в обеих группах разные* (т.е. одна совокупность отличается бóльшим разбросом, а вторая — мéньшим; одна совокупность менее однородна, а вторая — соответственно, более однородна), *то тогда* ***будет разным и доверие выборочным оценкам среднего*** $\bar{X}_1$ и $\bar{X}_2$. *Ведь одна оценка более точно характеризует среднюю тенденцию в своей совокупности, а вторая характеризует истинное значение среднего в своей совокупности несколько хуже* (*менее точно*, *более поверхностно*). Поэтому *если дисперсии* в двух совокупностях *разные*, то будет использоваться критерий вида $t = \dfrac{\bar{X}_1 - \bar{X}_2}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}}$. Число степеней свободы критерия вычисляется довольно трудоёмко: $df = \dfrac{\left(\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}\right)^2}{\dfrac{\left(\dfrac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \dfrac{\left(\dfrac{s_2^2}{n_2}\right)^2}{n_2 - 1}}$. При определении числа степеней свободы учитываются объёмы каждой из выборок в отдельности: в той выборке, объём которой больше, среднее и дисперсия оценены с большей точностью (т.е. погрешность этих оценок ниже), а в той выборке, объём которой меньше, среднее и дисперсия оценены с меньшей точностью (т.е. погрешность этих оценок выше). Поэтому в первом случае мы доверяем оценкам среднего и выборочной дисперсии больше, чем во втором (так как во втором случае параметры оценены с меньшей точностью) . Это обстоятельство мы и учитываем при расчёте числа степеней свободы[1].

Теперь представим себе, что ***истинные значения дисперсии в обеих группах на самом-то деле одинаковые*** (т.е. обе совокупности характеризуются одинаковым разбросом), то чему тогда должна быть равна дисперсия разности $(\bar{X}_1 - \bar{X}_2)$? Рассчитаем её:

---

[1] Здесь мы не будем останавливаться на том, каково математическое «происхождение» такого громоздкого выражения для *df*, оставим его без доказательства (за рамками текста).

$\mathrm{D}(\bar{X}_1 - \bar{X}_2) = \mathrm{D}(\bar{X}_1) + \mathrm{D}(\bar{X}_2) = \frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}$. Если обе дисперсии в генеральной совокупности одинаковые, то обозначим их одной буквой s^2(без нижнего индекса, поскольку в этом теперь нет необходимости) и заменим s_1^2 *и* s_2^2 *на* s^2. Тогда получим:

$$\mathrm{D}(\bar{X}_1 - \bar{X}_2) = \frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} = \frac{s^2}{n_1} + \frac{s^2}{n_2} = s^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right).$$

Чему же будет равна *общая для обеих выборок* оценка дисперсии s^2? Определим её как *среднее значение из двух несмещённых выборочных оценок дисперсий, взвешенных* на своё число наблюдений (число степеней свободы):

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

Подставим это выражение в вычисления дисперсии разности для двух средних арифметических и получим:

$$\mathrm{D}(\bar{X}_1 - \bar{X}_2) = s^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right) = \frac{s_1^2(n_1 - 1) + s_2^2(n_2 - 1)}{n_1 + n_2 - 2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right).$$

Следовательно, стандартное отклонение разности $(\bar{X}_1 - \bar{X}_2)$ будет равно корню квадратному из дисперсии:

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

Именно поэтому в знаменателе t-критерия присутствует эта запись, кажущаяся на первый взгляд сложной. Но только на первый взгляд. Поэтому *если дисперсии* в двух совокупностях *одинаковые*, то будет использоваться следующий критерий вида: $t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\sqrt{\frac{s_1^2(n_1 - 1) + s_2^2(n_2 - 1)}{n_1 + n_2 - 2}}}$. Число степеней свободы критерия вычисляется как $df = (n_1 + n_2 - 2)$.

Осталось коснуться вопроса о том, как определить, *равны ли истинные значения дисперсии в обеих группах.* Следует ли считать значения дисперсий одинаковыми или разными в обеих генеральных совокупностях? Для ответа на этот вопрос проверим статистическую гипотезу о равенстве истинных значений дисперсий (т. е. таких значений дисперсий, которые имеют место в генеральных совокупностях) с помощью F-критерия Фишера. Поэтому прежде чем выбрать соответствующую модификацию критерия для проверки гипотезы

о равенстве средних, сначала следует проверить промежуточную гипотезу о равенстве двух дисперсий. И по итогам этой проверки определиться с дальнейшим ходом решения.

В заключение приведём памятку-путеводитель, в компактном виде содержащую алгоритм выбора статистического критерия для проверки гипотезы о равенстве средних для двух независимых выборок (рис. 6.10). Схема наглядно указывает, что ***выбор подходящей модификации t-критерия Стьюдента для проверки статистической гипотезы о равенстве средних зависит от сочетания четырех параметров***: (df_1 и s_1^2), с одной стороны, и (df_2 и s_2^2) — другой.

Пример 6.6. «Выпускники». Последние годы существует распространённая практика, когда студенты старших курсов (выпускники

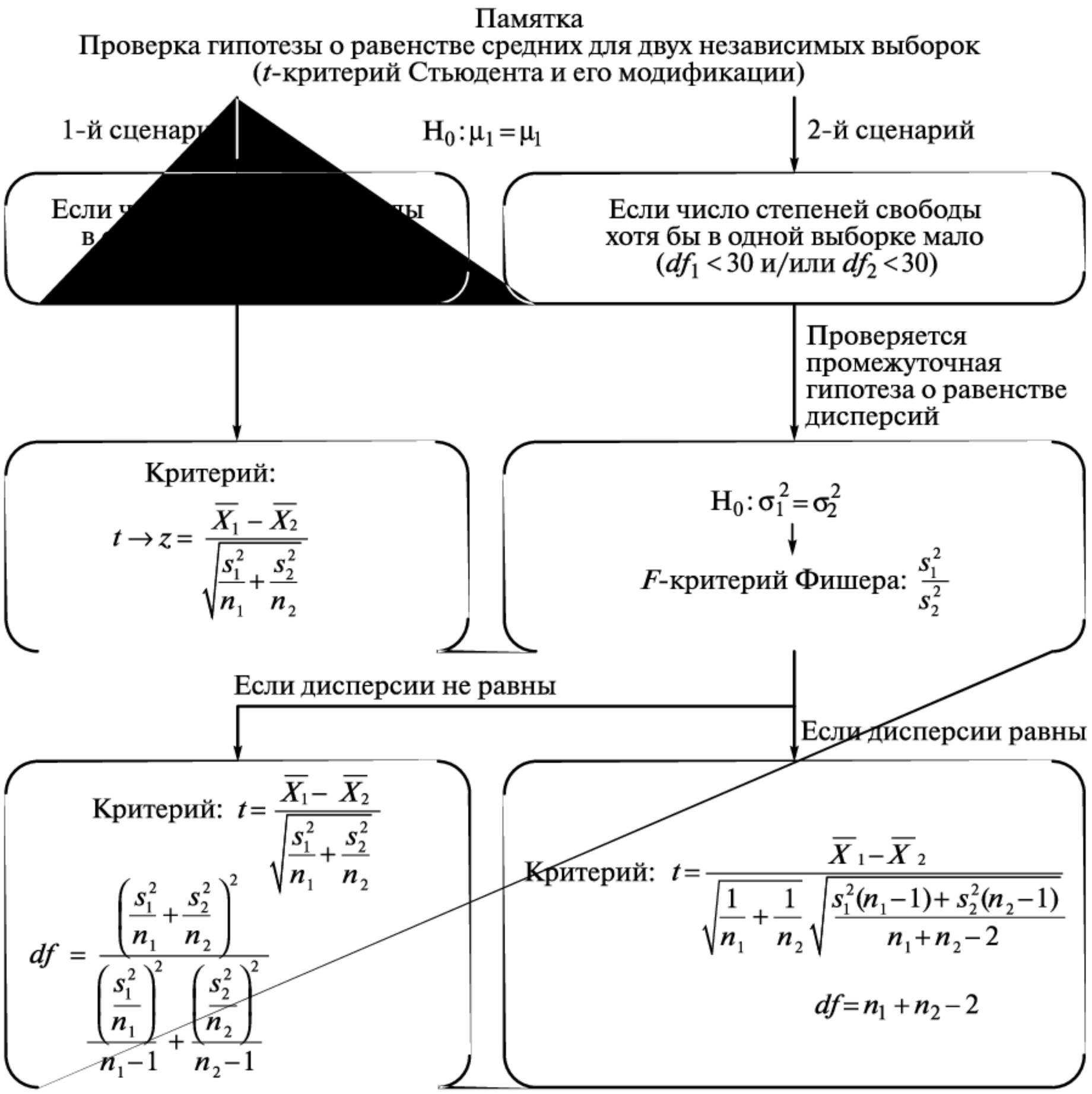


Рис. 6.10. Алгоритм проверки гипотезы о равенстве средних для двух независимых выборок

бакалавриата) подыскивают себе работу и в результате совмещают учёбу с профессиональной деятельностью на постоянной основе. В ходе одного из опросов[1], проведённого среди студентов IV курса, совмещающих учёбу и работу, исследователи задавали респондентам такие вопросы: «Собираетесь ли Вы после окончания бакалавриата продолжать дальнейшее обучение в магистратуре?», «Укажите, пожалуйста, каким примерно является Ваш доход в течение месяца от трудовой деятельности, которой Вы занимаетесь, совмещая учёбу и работу?». В результате были получены следующие данные: среднемесячный доход студента IV курса, желающего продолжать обучение в магистратуре, равен 12 860 руб., в то время как среднемесячный доход выпускника, не желающего продолжать обучение в магистратуре, на момент окончания вуза составил 18 220 руб. Выборочные оценки стандартного отклонения (на основе данных проведённого опроса) соответственно равны 3 360 и 5 290 руб. Из опрошенных собираются пойти в магистратуру 165 человек и не собираются 34 человека. Можно ли считать *средние показатели дохода по двум категориям студентов равными*? Уровень доверительной вероятности примите равным 95 %. Проверку соответствующей гипотезы осуществите против направленной альтернативы: «Средний показатель дохода в первой группе меньше, чем во второй».

Решение. Рассматриваются две группы: работающие студенты, которые планируют продолжить академическое обучение после окончания IV курса и работающие студенты, которые планируют завершить обучение, отказавшись от поступления в магистратуру. Объёмы выборок соответственно равны: $n_1 = 165$ человек и $n_2 = 34$ человека. Выборочные оценки среднего в группах составляют: $\bar{X}_1 = 12\,860$ руб. и $\bar{X}_2 = 18\,220$ руб. Выборочные оценки дисперсий соответственно равны: $s_1^2 = 3\,360^2$ и $s_2^2 = 5\,290^2$.

Требуется проверить статистическую гипотезу о равенстве средних: H_0: «$\mu_1 = \mu_2$ (средние *равны*)». Нулевой гипотезе противопоставляется направленная альтернативная гипотеза: H_1: «$\mu_1 < \mu_2$ (первое среднее *меньше* второго среднего)». Альтернатива является левосторонней (знак «<»). Уровень значимости α равен $0{,}05 = 1 - 0{,}95$. Выберем критерий для проверки статистической гипотезы. Число степеней свободы в обеих выборках высоко ($df_1 = 164$, $df_2 = 33$), поэтому критерием проверки станет: $t \to z = \dfrac{\bar{X}_1 - \bar{X}_2}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}}$. Рассчитаем наблюдаемое значение статистики z-критерия:

[1] Опрос проведен студентами II курса НИУ «Высшая школа экономики» в рамках курсовой работы в 2011 г. (*Тотров М.*, *Новикова Е.* «Учиться или работать: образовательные и карьерные стратегии студентов старших курсов». Научный руководитель — А.В. Пашкевич).

$$z_{\text{наблюдаемое}} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{12\ 860 - 18\ 220}{\sqrt{\frac{3\ 360^2}{165} + \frac{5\ 290^2}{34}}} = \frac{12\ 860 - 18\ 220}{\sqrt{\frac{3\ 360^2}{165} + \frac{5\ 290^2}{34}}} =$$

$$= \frac{-5\ 360}{\sqrt{68\ 422 + 823\ 062}} = \frac{-5\ 360}{\sqrt{891\ 484}} = \frac{-5\ 360}{944} \approx -5{,}68.$$

Полученное фактическое значение статистики критерия нужно сравнить с табличным (критическим) значением. Найдём его по таблице стандартного нормального распределения: $z_{\text{критическое}} = -1{,}65$ (табл. 6.6). Так как в данном примере нулевая гипотеза проверяется против направленной конкурирующей гипотезы H_1: $\mu_1 < \mu_2$, то критическая область будет односторонней. Более того, так как наблюдаемое значение z-статистики получилось меньше нуля (отрицательное) ввиду того, что из меньшего значения среднего $\bar{X}_1$ вычиталось большее значение среднего $\bar{X}_2$, то построение критической области будет основано на области отрицательных значений, она будет левосторонней. Отметим наблюдаемое и критическое значения на графике плоскости распределения вероятностей z-критерия (рис. 6.11).

Сопоставляем два значения: $z_{\text{наблюдаемое}} << z_{\text{критическое}}$ ($-5{,}68 << -1{,}65$), значение –5,68 принадлежит критической области, поэтому на уровне значимости $\alpha = 5\,\%$ у нас есть основания отвергнуть H_0 в пользу H_1: $\mu_1 < \mu_2$ (первое среднее *меньше* второго среднего)». Расхождение между двумя выборочными оценками среднего носит <u>не</u>случайный характер, ведь статистика критерия попала в область своих маловероятных значений в предположении о справедливости нулевой гипотезы. Следовательно, гипотеза о равенстве двух средних не кажет-

Таблица 6.6. **Фрагмент таблицы стандартного нормального *z*-распределения (интеграл от 0 до *z*)**

Z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,0000	0,0040	0,0080	0,0120	0,0160	0,0199	0,0239	0,0279	0,0319	0,0359
0,1	0,0398	0,0438	0,0478	0,0517	0,0557	0,0596	0,0636	0,0675	0,0714	0,0753
…										
1,5	0,4332	0,4345	0,4357	0,4370	0,4382	0,4394	0,4406	0,4418	0,4429	0,4441
1,6	0,4452	0,4463	0,4474	0,4484	0,4495	0,4505	0,4515	0,4525	0,4535	0,4545
1,7	0,4554	0,4564	0,4573	0,4582	0,4591	0,4599	0,4608	0,4616	0,4625	0,4633
…	и т. д.									

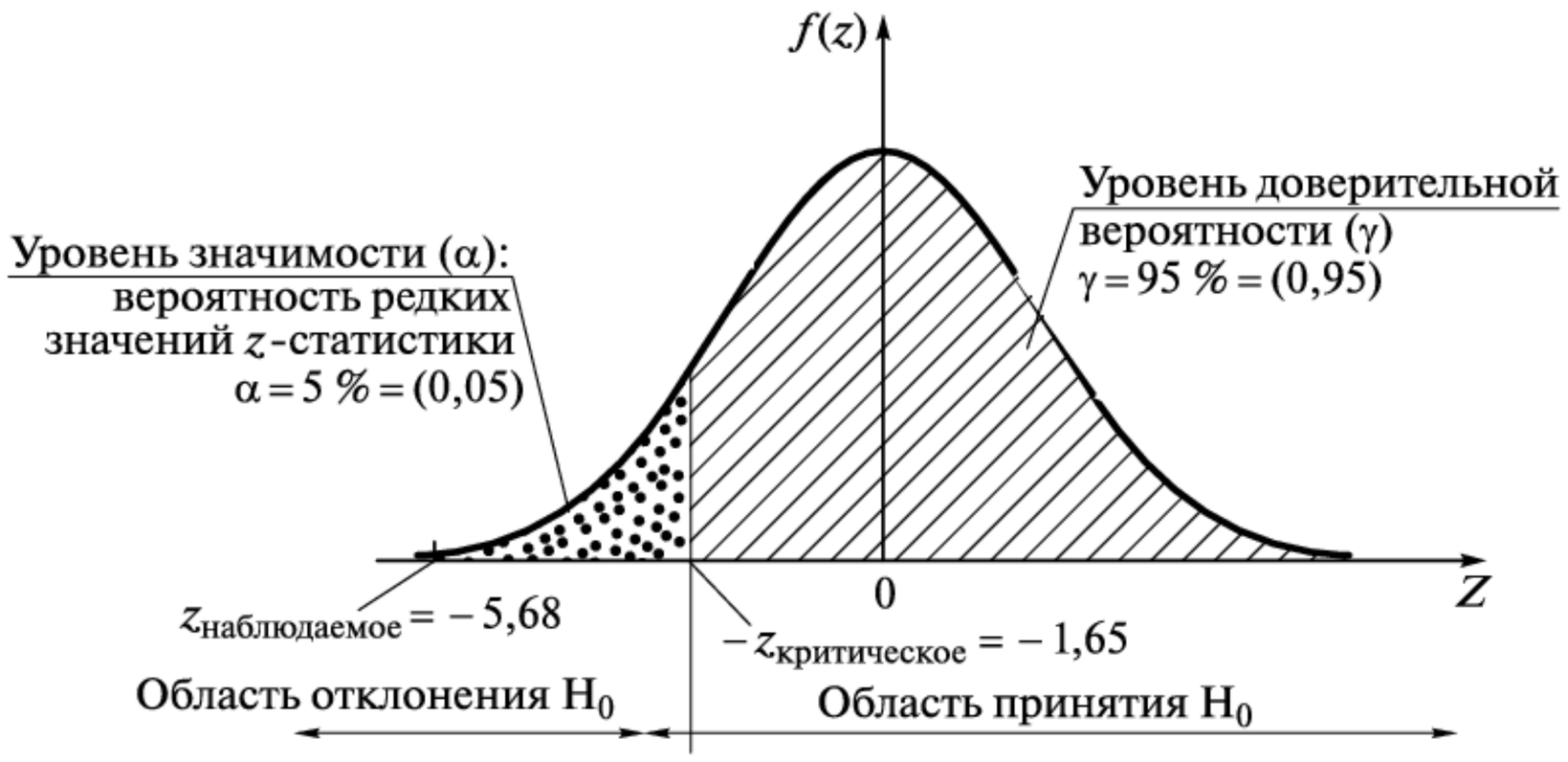


Рис. 6.11. Построение односторонней (левосторонней) критической области (см. пример 6.6 «Выпускники»)

ся убедительной. Поэтому на уровне значимости (на уровне ошибки) $\alpha = 0{,}05$ у нас есть основания её отклонить в пользу альтернативной гипотезы.

В завершение формулируем содержательный вывод найденного решения: на основании имеющихся данных исследования нельзя считать средние показатели ежемесячного дохода по двум категориям студентов равными. Статистически обосновано, что у студентов, стремящихся продолжать обучение в магистратуре, средний показатель дохода ниже, чем у тех студентов, которые после окончания бакалавриата не стремятся к продолжению учёбы в магистратуре. Наблюдаемые на выборках различия статистически значимы.

Пример 6.7. «Сигареты». Приведены результаты наблюдений по двум выборкам (7 мужчин и 6 женщин) о числе сигарет, выкуриваемых в день:

X_1(мужчины): 5, 10, 4, 15, 10, 3, 2 (сигарет);

X_2 (женщины): 3, 6, 1, 4, 11, 5 (сигарет).

С помощью t-критерия Стьюдента нужно проверить нулевую гипотезу о равенстве средних значений в рассматриваемых группах для 95%-го уровня доверия против *ненаправленной* конкурирующей гипотезы. (Для выбора подходящей модификации t-критерия дополнительно требуется проверить промежуточную гипотезу о равенстве дисперсий двух выборок).

Решение. В задаче рассматриваются две группы людей — курящие мужчины и курящие женщины. Объёмы выборок соответственно равны: $n_1 = 7$ человек и $n_2 = 6$ человек. Анализируется среднее число сигарет, выкуриваемых в день одним человеком (мужчиной или женщиной). Найдем выборочные оценки среднего в каждой группе:

$$\bar{X}_1 = \frac{5+10+4+15+10+3+2}{7} = \frac{49}{7} = 7$$

— выборочная оценка среднего количества сигарет, выкуриваемых одним мужчиной;

$$\bar{X}_2 = \frac{3+6+1+4+11+5}{6} = \frac{30}{6} = 5$$

— выборочная оценка среднего количества сигарет, выкуриваемых одной женщиной.

Рассчитаем несмещенные выборочные оценки дисперсий:

$$s_1^2 = \frac{(5-7)^2 + (10-7)^2 \cdot 2 + \ldots + (2-7)^2}{7-1} =$$

$$= \frac{4+18+9+64+16+25}{6} = \frac{136}{6} \approx 22{,}7 \text{ (сигарет)}^2,$$

стандартное отклонение составляет $\sqrt{22{,}7} \approx 4{,}8$(сигарет);

$$s_2^2 = \frac{(3-5)^2 + (6-5)^2 + \ldots + (5-5)^2}{6-1} =$$

$$= \frac{4+1+16+1+36+0}{5} = \frac{58}{5} = 11{,}6 \text{ (сигарет)}^2,$$

стандартное отклонение составляет $\sqrt{11{,}6} \approx 3{,}4$(сигарет).

Требуется проверить статистическую гипотезу о равенстве средних:

H_0: «$\mu_1 = \mu_2$ (средние *равны*)» (наблюдаемые на выборках различия в средних являются случайными). Нулевой гипотезе противопоставляется ненаправленная альтернативная гипотеза:

H_1: «$\mu_1 \neq \mu_2$ (средние не равны)» (наблюдаемые на выборках различия в средних никак НЕ являются случайными).

Критическая область значений статистики критерия будет разделена на две части (два «хвоста» распределения): от $-\infty$ до $-t_{\text{критическое}}$, с одной стороны, и от $+t_{\text{критическое}}$ до $+\infty$, с другой. Уровень значимости (статистическая ошибка) делится поровну между двумя «хвостами»: $\alpha/2 = 0{,}05:2 = (1 - 0{,}95):2 = 0{,}025 = 2{,}5\,\%$. Выберем критерий для проверки статистической гипотезы. Поскольку число степеней свободы обеих выборок не превышает 30, их объёмы маленькие ($df_1 = 6$, $df_2 = 5$), следует проверить промежуточную гипотезу о равенстве дисперсий: H_0: «$\sigma_1^2 = \sigma_2^2$». Конкурирующей гипотезой служит ненаправленная альтернатива: H_1: «$\sigma_1^2 \neq \sigma_2^2$». Проверку гипотезы о равенстве дисперсий также осуществим на уровне значимости 5 %. Тогда $\alpha/2 = 0{,}025 = 2{,}5\,\%$. Воспользуемся F-критерием Фишера, вычислив его наблюдаемое значение: $F_{\text{наблюдаемое}} = \frac{s_1^2}{s_2^2} = \frac{22{,}7}{11{,}6} = 1{,}96$.

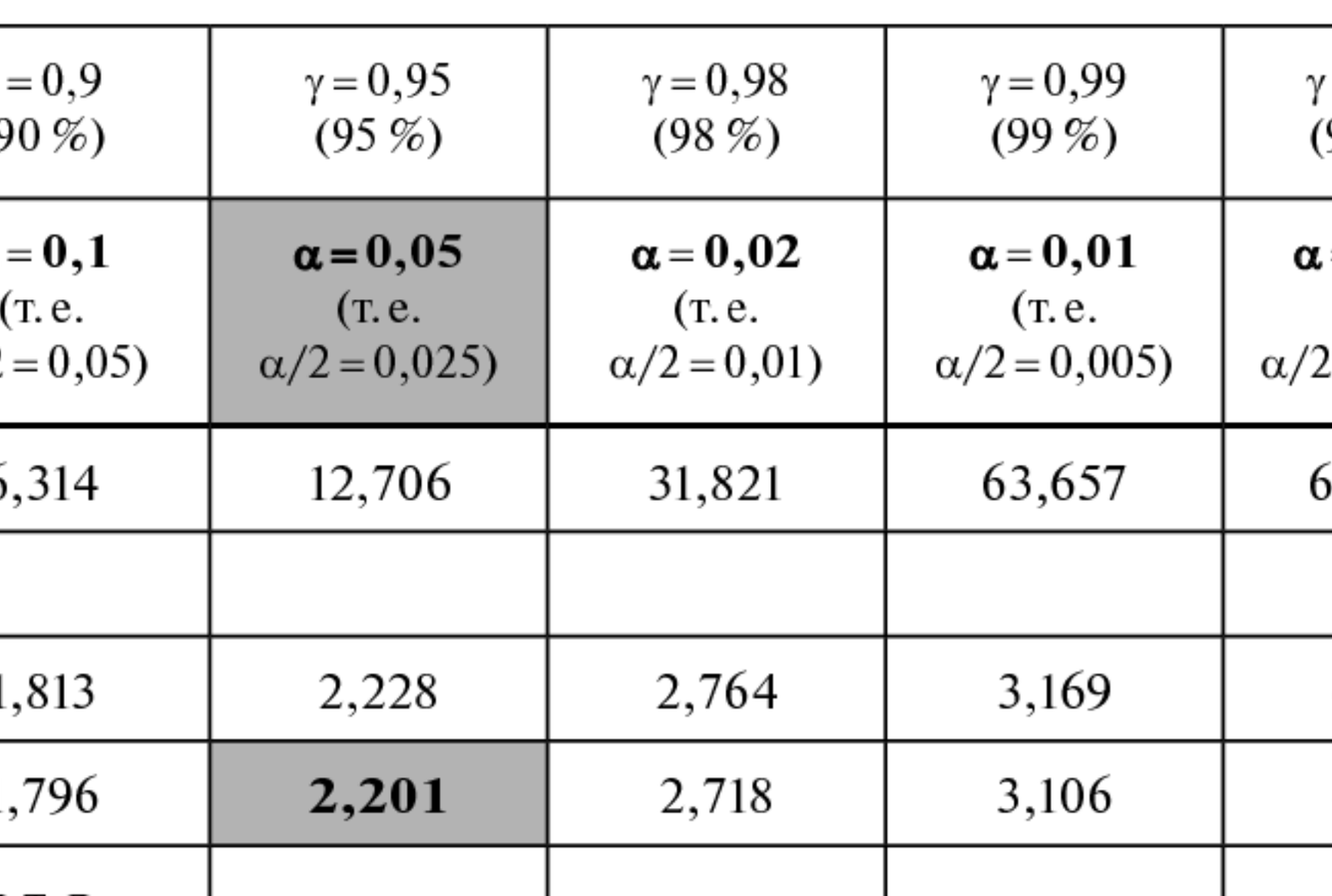
По таблице критических значений F-распределения находим, что $F_{\text{критическое}} = \left\{ df_1 = n_1 - 1 = 6; df_2 = n_2 - 1 = 5; \frac{\alpha}{2} = 0{,}025 \right\} = 6{,}98$. На уровне

доверительной вероятности 95 % нет оснований отвергать гипотезу о равенстве дисперсий. Нет статистических оснований считать дисперсии разными. Наблюдаемое на выборках различие в оценках дисперсий (указывающее на то, что дисперсия количества сигарет, выкуриваемых мужчинами, почти в два раза больше дисперсии количества сигарет, выкуриваемых женщинами) считается *случайным*. Различие дисперсий статистически *незначимо*.

Согласно результатам проверки гипотезы о равенстве дисперсий, гипотеза о равенстве математических ожиданий будет проверена с помощью следующей модификации t-критерия Стюдента:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \sqrt{\frac{s_1^2(n_1 - 1) + s_2^2(n_2 - 1)}{n_1 + n_2 - 2}}}.$$

Таблица 6.7. **Фрагмент таблицы для *t*-распределения Стьюдента (двустороннее значение вероятности)**



df	Доверительная вероятность и уровень α				
	γ = 0,9 (90 %)	γ = 0,95 (95 %)	γ = 0,98 (98 %)	γ = 0,99 (99 %)	γ = 0,999 (99,9 %)
	α = 0,1 (т. е. α/2 = 0,05)	**α = 0,05** (т. е. α/2 = 0,025)	**α = 0,02** (т. е. α/2 = 0,01)	**α = 0,01** (т. е. α/2 = 0,005)	**α = 0,001** (т. е. α/2 = 0,0005)
1	6,314	12,706	31,821	63,657	636,619
…					
10	1,813	2,228	2,764	3,169	4,587
11	1,796	**2,201**	2,718	3,106	4,437
…	и т. д.				

Рассчитаем наблюдаемое значение статистики t-критерия:

$$t_{\text{наблюдаемое}} = \frac{7-5}{\sqrt{\frac{1}{7}+\frac{1}{6}}\sqrt{\frac{22{,}7\cdot(7-1)+11{,}6\cdot(6-1)}{7+6-2}}} = \frac{7-5}{\sqrt{0{,}31}\sqrt{17{,}65}} =$$

$$= \frac{2}{\sqrt{0{,}31}\sqrt{17{,}65}} = \frac{2}{2{,}34} = +0{,}85.$$

Полученное фактическое значение статистики критерия нужно сравнить с табличным (критическим) значением. Чтобы найти его по таблице, требуется еще указать число степеней свободы: $df = (n_1 + n_2 - 2) = 7 + 6 - 2 = 11$. Теперь найдем критическое значение t-статистики по таблице t-распределения Стьюдента (табл. 6.7): $t_{\text{критическое}} = 2{,}201$. Отметим наблюдаемое и критическое значения на схематичном графике плотности распределения вероятностей рассматриваемой статистики t-критерия (рис. 6.12).

Наблюдаемое значение статистики критерия (0,85) не попало в критическую область, поэтому на уровне доверительной вероятности $\gamma = 95\ \%$ нет оснований отвергать H_0: «$\mu_1 = \mu_2$ (средние равны)». Расхождение между двумя выборочными оценками среднего считается случайным — в пределах статистической погрешности.

Итак, сформулируем содержательный вывод найденного решения: по имеющимся данным выборочного исследования нет оснований считать, что мужчины и женщины в генеральной совокупности в среднем за день выкуривают разное число сигарет.

Пример 6.8. «География». Учащиеся средней школы выполняли контрольную работу по географии, состоящую из 10 заданий. Фиксировалось среднее число правильно выполненных заданий (имеем две выборки).

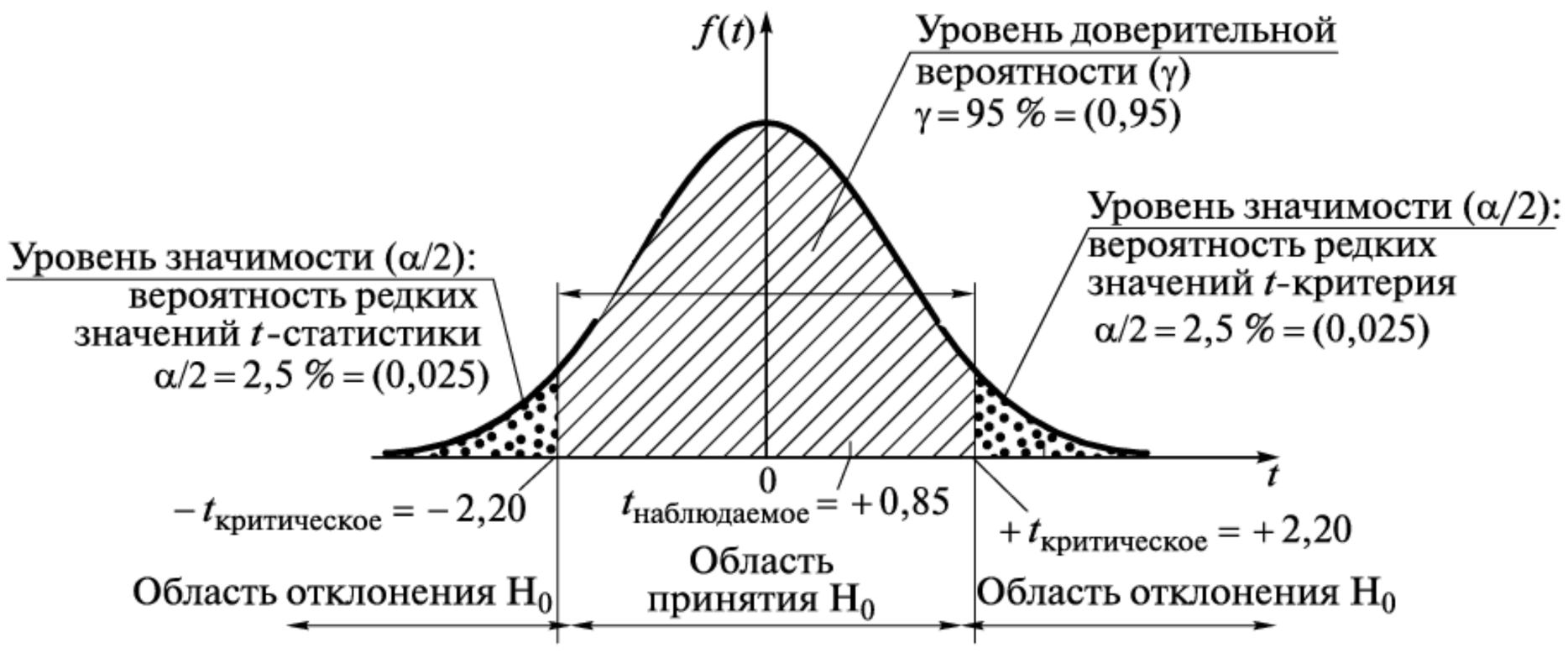


Рис. 6.12. Построение двусторонней критической области (см. пример 6.7 «Сигареты»)

Группа учащихся	Объёмы выборки, человек	Среднее число правильно выполненных заданий	Оценка дисперсии
Мальчики	$n_1 = 10$	8	4
Девочки	$n_2 = 12$	6	3

С помощью критерия Стьюдента нужно проверить гипотезу о равенстве средних значений в двух группах (*считая их дисперсии равными*) при 95%-м уровне доверия против направленной альтернативы: «мальчики написали контрольную работу лучше, чем девочки».

Решение. Рассматриваются две группы учащихся — мальчики и девочки. Объёмы выборок соответственно равны: $n_1 = 10$ человек и $n_2 = 12$ человек. Выборочные оценки среднего в группах составляют: $\bar{X}_1 = 8$ заданий и $\bar{X}_2 = 6$ заданий. Выборочные оценки дисперсий соответственно равны: $s_1^2 = 4$ и $s_2^2 = 3$.

Необходимо проверить статистическую гипотезу о равенстве средних: H_0: «$\mu_1 = \mu_2$ (средние *равны*)». Нулевой гипотезе противопоставляется направленная альтернативная гипотеза: H_1: «$\mu_1 > \mu_2$ (первое среднее *больше* второго среднего, т. е. среднее число правильно выполненных заданий у мальчиков выше, чем у девочек)». Фиксированный уровень значимости $\alpha = 0{,}05 = 1 - 0{,}95 = 5\,\%$. Обе выборки малые (число степеней свободы в них $df_1 = 9$, $df_2 = 11$). В условии задачи сказано о том, что следует считать дисперсии равными[1], поэтому выпишем критерий, пригодный для проверки выдвинутой гипотезы: $t = \dfrac{\bar{X}_1 - \bar{X}_2}{\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}\sqrt{\dfrac{s_1^2(n_1 - 1) + s_2^2(n_2 - 1)}{n_1 + n_2 - 2}}}$.

Рассчитаем наблюдаемое значение t-статистики:

$$t_{\text{наблюдаемое}} = \frac{8 - 6}{\sqrt{\dfrac{1}{10} + \dfrac{1}{12}}\sqrt{\dfrac{4 \cdot (10 - 1) + 3 \cdot (12 - 1)}{10 + 12 - 2}}} =$$

$$= \frac{2}{\sqrt{0{,}183}\sqrt{3{,}45}} = \frac{2}{0{,}79} = +2{,}53.$$

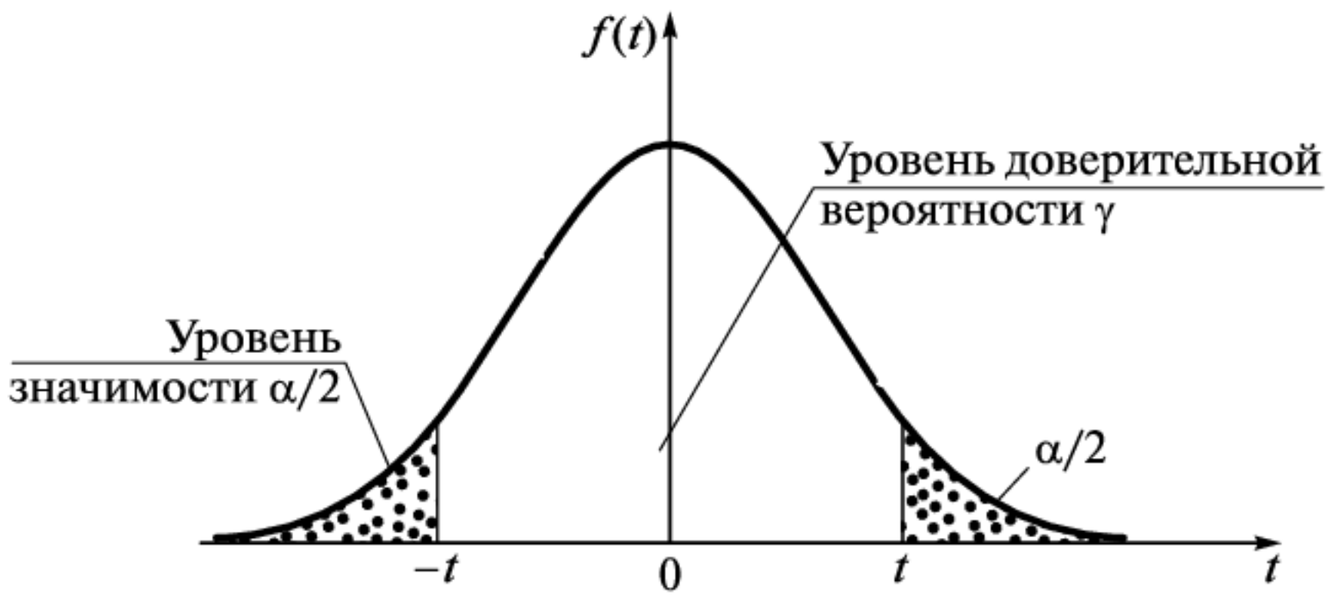
Полученное фактическое значение статистики критерия нужно сравнить с табличным (критическим) значением. Чтобы найти его по та-

[1] Вы можете убедиться в этом, если проверите промежуточную статистическую гипотезу о равенстве дисперсий с помощью F-критерия Фишера.

блице, нужно уточнить число степеней свободы: $df = (n_1 + n_2 - 2) = 10 + 12 - 2 = 20$. Теперь найдём критическое значение t-статистики по таблице t-распределения Стьюдента: $t_{\text{критическое}} = +1{,}725$ (табл. 6.8).

Обратите внимание, как определяется критическое значение в этой задаче. В используемой нами таблице приведены значения критических точек для двусторонней области: когда альтернативная гипотеза *не*направленная и уровень значимости α делится пополам ($\alpha/2$), т. е. распределяется поровну между левым и правым «хвостами» распределения. Поэтому если в задаче требуется рассмотреть одностороннюю критическую область, то вероятность *в одном «хвосте»* распределения должна составить α (*целое*, *без деления на 2*). Так, в нашем примере уровень значимости $\alpha = 0{,}05$ (как $1 - 0{,}95$). Поэтому искомую критическую точку следует посмотреть именно в том столбце таблицы, где отмечено, что площадь под кривой графика плотности t-распределения в одном «хвосте» составляет 0,05 (а не

Таблица 6.8. **Фрагмент таблицы для *t*-распределения Стьюдента (двустороннее значение вероятности)**



df	Доверительная вероятность и уровень α				
	γ = 0,9 (90 %)	γ = 0,95 (95 %)	γ = 0,98 (98 %)	γ = 0,99 (99 %)	γ = 0,999 (99,9 %)
	α = 0,1 (т. е. α/2 = 0,05)	α = 0,05 (т. е. α/2 = 0,025)	α = 0,02 (т. е. α/2 = 0,01)	α = 0,01 (т. е. α/2 = 0,005)	α = 0,001 (т. е. α/2 = 0,0005)
1	6,314	12,706	31,821	63,657	636,619
…					
19	1,729	2,093	2,540	2,861	3,883
20	**1,725**	2,086	2,528	2,845	3,850
…	и т. д.				

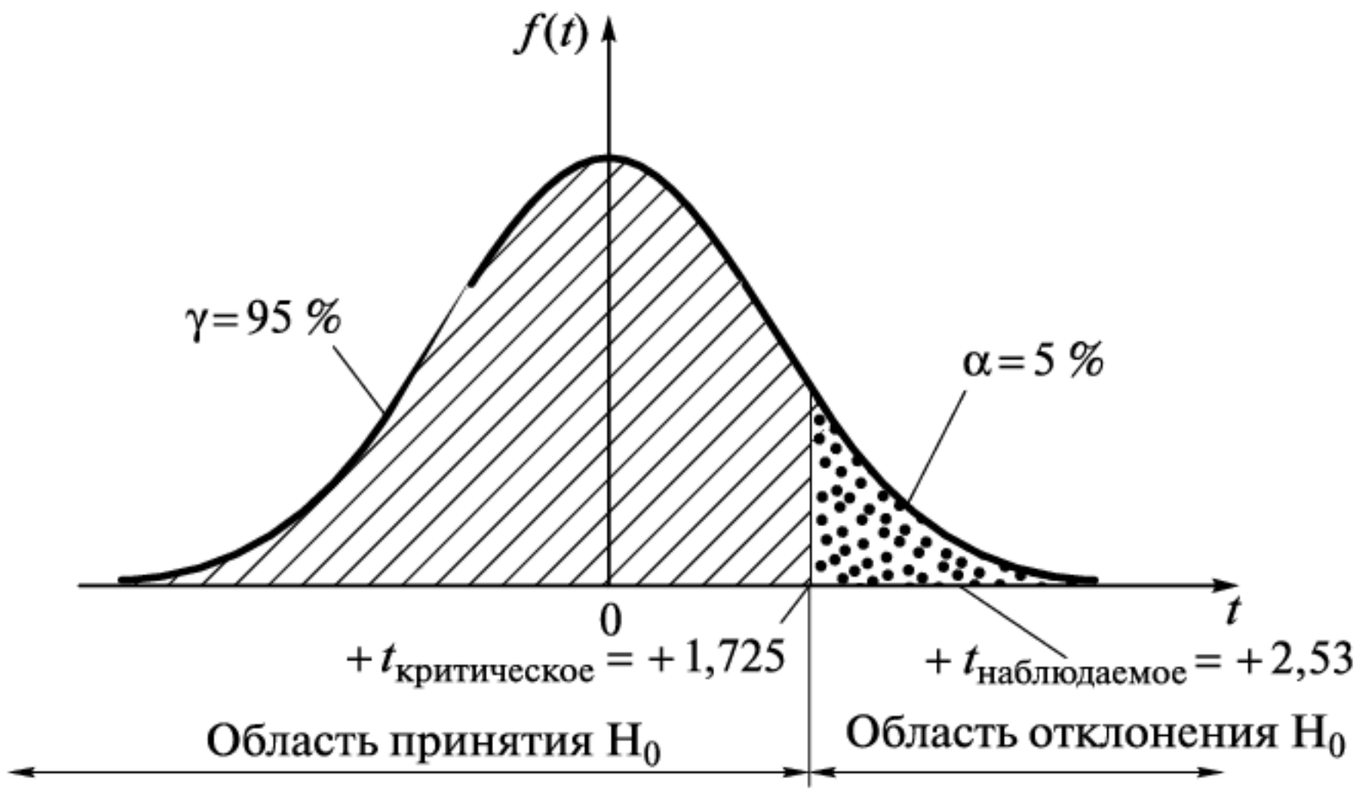


Рис. 6.13. Построение односторонней (правосторонней) критической области (см. пример 6.8 «География»)

0,025). В таблице этот столбец выделен серым цветом. Собственно, это и будет означать, что как раз половина от 10 % (т. е. 5 %) и приходится на один «хвост» распределения. Получим: $t_{\text{критическое}} = +1{,}725$.

Отметим наблюдаемое и критическое значения на схематичном графике плотности распределения вероятностей рассматриваемой статистики, построив критическую область (рис. 6.13).

Сопоставляем два значения: $t_{\text{наблюдаемое}} > t_{\text{критическое}} : +2{,}53 > +1{,}725$, поэтому на уровне значимости $\alpha = 5\,\%$ H_0 отвергается (не принимается) и принимается H_1: «$\mu_1 > \mu_2$ (первое среднее *больше* второго среднего)». Расхождение между двумя выборочными оценками среднего носит *не*случайный характер. Видим, что вероятность попадания в критическую область невелика. Но по факту случилось именно это — наблюдаемое значение статистики критерия попало в критическую область! Следовательно, гипотеза о равенстве двух средних не может быть принята при имеющихся данных — слишком заметны и ощутимы различия в выборочных средних. Поэтому на уровне значимости (на уровне ошибки) $\alpha = 0{,}05$ мы её отклоняем в пользу конкурирующей.

Итак, сформулируем содержательный вывод найденного решения: На основе имеющихся данных выборочного исследования нельзя считать средние результаты написания контрольной работы одинаковыми для мальчиков и девочек. Статистически обосновано, что среднее число правильно выполненных заданий у мальчиков выше, чем у девочек. Наблюдаемые на выборках *различия статистически значимы*.

Пример 6.9. «8 марта». Продавец цветочного магазина — в День 8 марта решила (из любознательности) выяснить, существует ли разница в том, сколько, в среднем, на букет тратят молодые мужчины (в возрасте ≤ 35 лет) и мужчины более старшего возраста (> 35 лет). Организовав случайное выборочное наблюдение за 10

покупателями в возрасте «до 35», она получила, что средняя стоимость букета составила 670 руб., а выборочная дисперсия оказалась равной 14 400 (руб.)^2 (т. е. стандартное отклонение равно 120 руб.). Выборка же 13 покупателей старше 35 лет выявила, что, в среднем, такой мужчина тратит 850 руб., при этом выборочная дисперсия равна 22 500 (руб.)^2 (т. е. стандартное отклонение равно 150 руб.). На уровне доверительной вероятности 95 % определите, существует ли разница в средней стоимости букета, приобретаемого молодыми мужчинами и мужчинами более зрелого возраста. Проверку статистической гипотезы осуществите против направленной альтернативы.

Решение. Изучаются две группы людей — мужчины до 35 лет и старше 35 лет. Объёмы выборок соответственно равны: $n_1 = 10$ человек и $n_2 = 13$ человек. Выборочные оценки среднего в группах составляют: $\bar{X}_1 = 670$ руб. и $\bar{X}_2 = 850$ руб. Выборочные оценки дисперсий соответственно равны: $s_1^2 = 120^2 = 14\,400$ и $s_2^2 = 150^2 = 22\,500$.

Требуется проверить статистическую гипотезу о равенстве средних: H_0: «$\mu_1 = \mu_2$ (средние *равны*)». Нулевой гипотезе противопоставляется направленная конкурирующая гипотеза: H_1: «$\mu_2 > \mu_1$ (второе среднее *больше* первого среднего)». Уровень значимости α равен $0{,}05 = 1 - 0{,}95$. Обе выборки малые, число степеней свободы в них невелико. Для выбора соответствующей модификации t-критерия нам дополнительно потребуется проверить промежуточную гипотезу о равенстве дисперсий с помощью F-статистики Фишера. Рассматривается промежуточная гипотеза вида: H_0: «$\sigma_1^2 = \sigma_2^2$». Конкурирующей гипотезой служит ненаправленная альтернатива: H_1: «$\sigma_1^2 \neq \sigma_2^2$». Проверку гипотезы о равенстве дисперсий также осуществим на уровне значимости 5 %. Тогда $\alpha/2 = 0{,}025$ (2,5 %). Прибегнем к критерию Фишера, вычислив его наблюдаемое значение: $F_{\text{наблюдаемое}} = \dfrac{s_2^2}{s_1^2} = \dfrac{22\,500}{14\,400} = 1{,}56$ (раз). В 1,56 раз одна выборочная оценка дисперсии больше другой. Является ли эта разница в оценках двух дисперсий статистически значимой, можно судить по таблице критических значений F-распределения. Найдём, что $F_{\text{критическое}} = \left\{ df_1 = n_1 - 1 = 9;\ df_2 = n_2 - 1 = 12;\ \dfrac{\alpha}{2} = 0{,}025 \right\} = 3{,}87$. На уровне доверия 95 % *гипотезу о равенстве дисперсий следует принять.* Нет оснований считать дисперсии разными. Наблюдаемое на двух выборках различие в оценках дисперсий (которое указывает на то, что дисперсия стоимости букета у мужчин старше 35 лет в 1,56 раз больше дисперсии стоимости букета у мужчин до 35 лет) считается *случайным*.

Различие выборочных оценок двух дисперсий статистически *незначимо* при имеющихся данных.

Теперь выпишем критерий для проверки главной статистической гипотезы о равенстве средних:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\sqrt{\frac{s_1^2(n_1 - 1) + s_2^2(n_2 - 1)}{n_1 + n_2 - 2}}}.$$

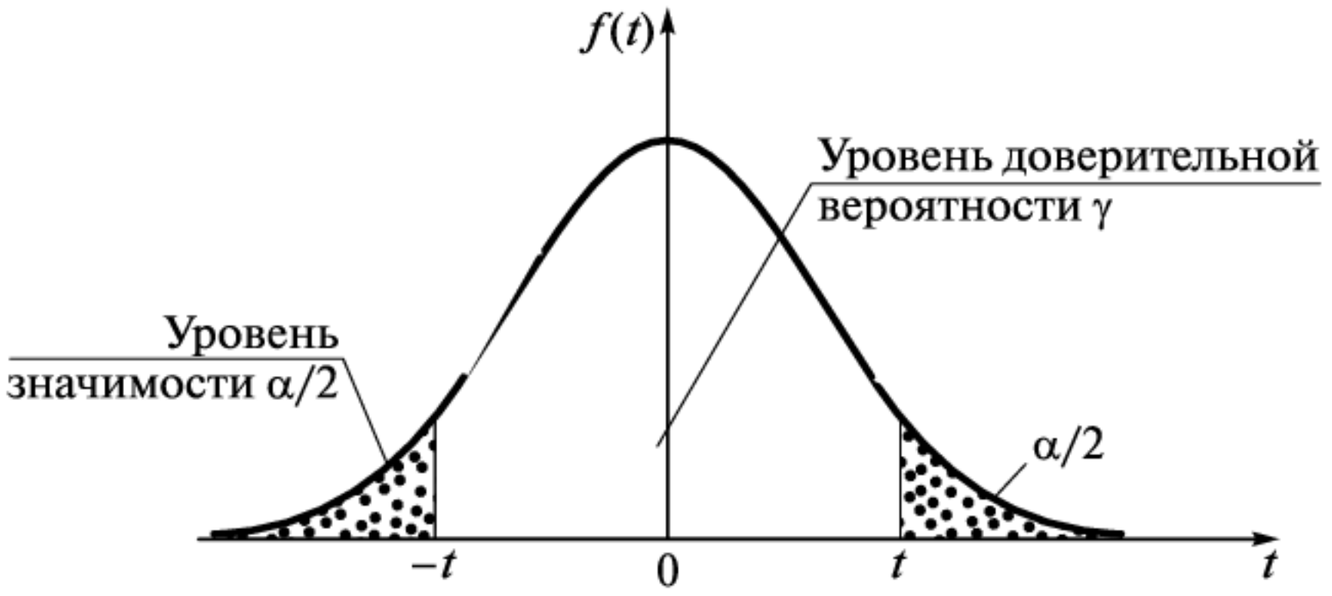
Рассчитаем наблюдаемое значение статистики t-критерия:

$$t_{\text{наблюдаемое}} = \frac{670 - 850}{\sqrt{\frac{1}{10} + \frac{1}{13}}\sqrt{\frac{14\,400 \cdot (10 - 1) + 22\,500 \cdot (13 - 1)}{10 + 13 - 2}}} =$$

$$= \frac{-180}{\sqrt{0{,}421}\sqrt{19\,029}} = \frac{-180}{0{,}649 \cdot 138} = \frac{-180}{58} = -3{,}10.$$

Далее построим критическую область (и полученное фактическое значение статистики критерия сравним с табличным (критическим) значением). Чтобы найти по таблице критическую точку, требуется

Таблица 6.9. **Фрагмент таблицы для *t*-распределения Стьюдента (двустороннее значение вероятности)**



df	Доверительная вероятность и уровень α				
	γ = 0,9 (90 %)	γ = 0,95 (95 %)	γ = 0,98 (98 %)	γ = 0,99 (99 %)	γ = 0,999 (99,9 %)
	α = 0,1 (т. е. α/2 = 0,05)	α = 0,05 (т. е. α/2 = 0,025)	α = 0,02 (т. е. α/2 = 0,01)	α = 0,01 (т. е. α/2 = 0,005)	α = 0,001 (т. е. α/2 = 0,0005)
1	6,314	12,706	31,821	63,657	636,619
…					
21	**1,721**	2,080	2,518	2,831	3,819
…	и т. д.				

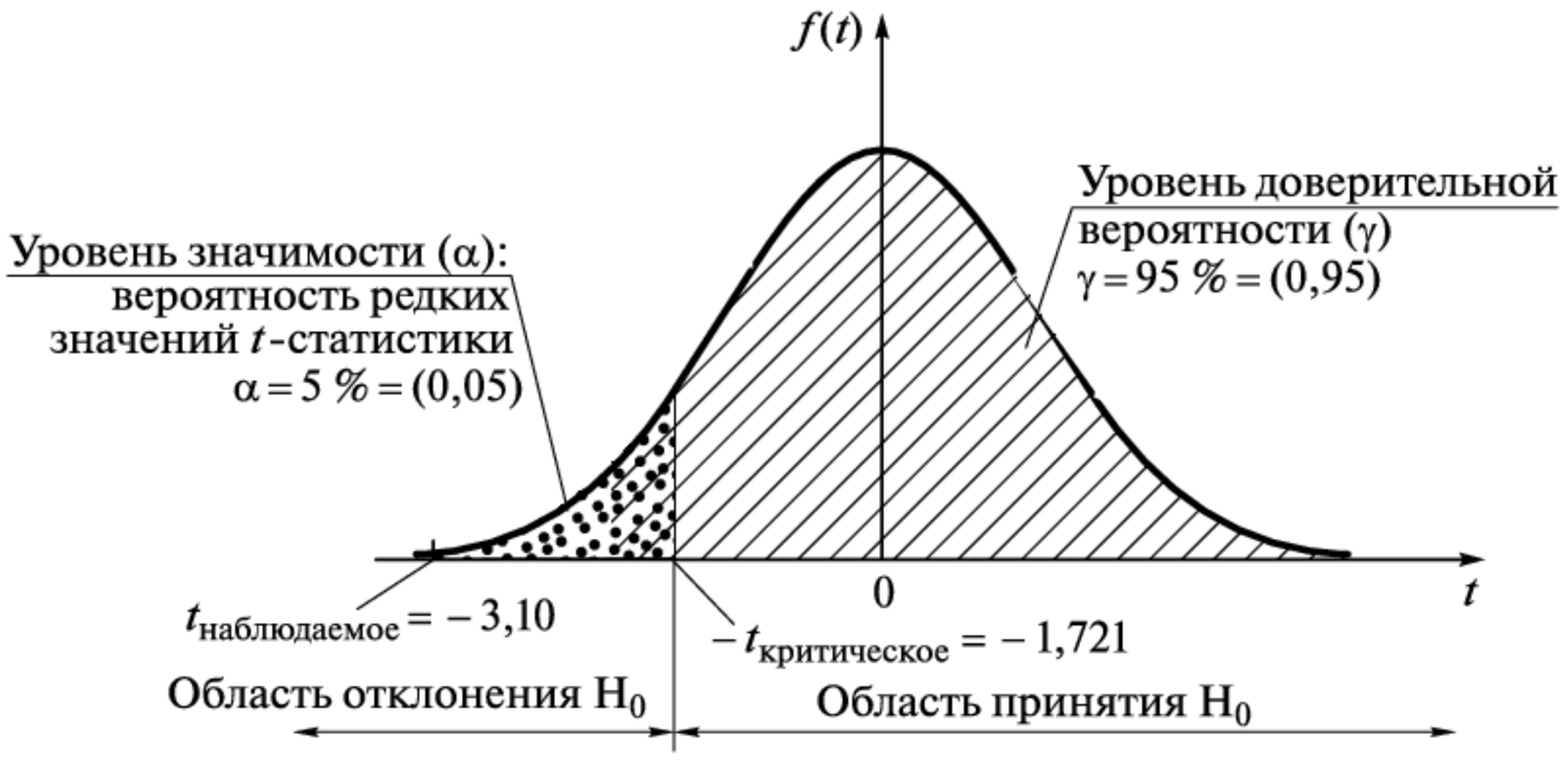


Рис. 6.14. Построение односторонней (левосторонней) критической области (см. пример 6.9 «8 марта»)

ещё указать число степеней свободы: $df = (n_1 + n_2 - 2) = 10 + 13 - 2 = 21$. Теперь найдём критическое значение t-статистики по таблице t-распределения Стьюдента.

Обратите внимание, как находится критическое значение в этой задаче. В табл. 6.9 приведены значения критических точек для двусторонней области: когда альтернативная гипотеза *не*направленная и уровень значимости α делится пополам ($\alpha/2$), т. е. распределяется поровну, между левым и правым «хвостами». Поэтому если в задаче требуется рассмотреть одностороннюю критическую область, то вероятность *в одном «хвосте»* распределения должна составить α (*целое*, *без деления на 2*). В примере уровень значимости $\alpha = 0{,}05$ (1 – 0,95). Поэтому искомую критическую точку следует посмотреть именно в том столбце таблицы, где отмечено, что площадь под кривой функции плотности t-распределения в одном «хвосте» составляет 0,05 (а не 0,025). В таблице этот столбец выделен серым цветом. Собственно, это и будет означать, что как раз половина от 10 % (т. е. 5 %) и приходится на один «хвост» распределения. Получим: $t_{\text{критическое}} = -1{,}721$.

Отметим наблюдаемое и критическое значения на схематичном графике плотности распределения вероятностей рассматриваемой статистики, построив критическую область (рис. 6.14).

Сопоставляем два значения: $t_{\text{наблюдаемое}} < t_{\text{критическое}} (-3{,}10 < -1{,}721)$. Против нулевой гипотезы свидетельствуют бóльшие по модулю отрицательные значения статистики. Поэтому на уровне значимости $\alpha = 5\,\%$ есть основания отклонить H_0 в пользу альтернативы. Расхождение между двумя выборочными оценками среднего считается неслучайным. Если бы истинные значения средних в этих двух группах мужчин были одинаковы, то вероятность попасть в критическое множество мала. Это было бы почти невозможно. Но по факту случилось именно это! Очень заметны и ощутимы различия

в выборочных средних. Поэтому на уровне значимости (на уровне ошибки) $\alpha = 0{,}05$ мы отклоняем нулевую гипотезу в пользу конкурирующей.

Итак, сформулируем содержательный вывод найденного решения: статистически обосновано, что мужчины старше 35 лет тратят на букет больше, чем мужчины моложе. Наблюдаемые на выборках *различия статистически значимы*. При имеющихся данных исследования можно считать, что в генеральной совокупности их «щедрость» действительно выше.

Пример 6.10. «Спортивный инвентарь». Маркетологов компании, продающей спортивный инвентарь, интересует, определяет ли вид полки (стеллажа), на которой размещена продукция магазина, объём продаж. Чтобы выяснить это, для исследовательского эксперимента была взята случайная выборка из 26 магазинов, а затем случайным образом эту выборку разделили пополам, выделив тем самым 2 группы — контрольную и экспериментальную. В 13 магазинах первой группы интересующий товар размещали на обычном торговом оборудовании среди других спортивных товаров, а в 13 магазинах второй группы — на специализированных стеллажах с рекламой и сопутствующими промо-материалами. Спустя некоторое время были зафиксированы коммерческие показатели продаж (в сопоставимых единицах):

Вид стеллажа/полки торгового оборудования

Обычный	Специализированный
Среднее = 48 (усл. ед.) Дисперсия = 450 (усл. ед.)^2	Среднее = 65 (усл. ед.) Дисперсия = 120 (усл. ед.)^2

Совпадают ли средние объёмы продаж спортивных товаров, размещённых на обычных стеллажах и специализированных? На уровне доверительной вероятности 95 % определите, существуют ли статистически значимые (выраженные) различия. Проверку статистической гипотезы осуществите против направленной альтернативы: «Объём продаж товара, размещённого на специализированных стеллажах, выше».

Решение. Изучаются две группы магазинов с разной концепцией презентации товара — обычной и специализированной. Объёмы выборок соответственно равны: $n_1 = 13$ магазинов и $n_2 = 13$ магазинов. Выборочные оценки среднего показателя продаж в группах составляют: $\bar{X}_1 = 48$ усл. ед. и $\bar{X}_2 = 65$ усл. ед. Выборочные оценки дисперсий соответственно равны: $s_1^2 = 450$ и $s_2^2 = 120$.

Требуется проверить статистическую гипотезу о равенстве математических ожиданий: H_0: «$\mu_1 = \mu_2$ (средние *равны*)». Нулевой ги-

потезе противопоставляется направленная конкурирующая гипотеза: H_1: «$\mu_2 > \mu_1$ (второе среднее *больше* первого среднего)». Уровень значимости $\alpha = 0{,}05$ (1 – 0,95). Обе выборки малые, число степеней свободы в них невелико. Для выбора соответствующей модификации *t*-критерия нам дополнительно потребуется проверить промежуточную гипотезу о равенстве дисперсий с помощью *F*-статистики Фишера. Рассматривается промежуточная гипотеза вида: H_0: «$\sigma_1^2 = \sigma_2^2$». Конкурирующей гипотезой служит ненаправленная альтернатива: H_1: «$\sigma_1^2 \neq \sigma_2^2$». Проверку гипотезы о равенстве дисперсий тоже осуществим на уровне значимости 5 %. Тогда $\alpha/2 = 0{,}025$ (2,5 %). Прибегнем к критерию Фишера, вычислив его наблюдаемое значение: $F_{\text{наблюдаемое}} = \dfrac{s_1^2}{s_2^2} = \dfrac{450}{120} = 3{,}75$ (раз) — в 3,75 раза одна выборочная оценка дисперсии больше, чем вторая. Является ли эта разница в оценках двух дисперсий статистически значимой (достоверной), мы сможем судить по таблице критических значений *F*-распределения.

Найдём, что $F_{\text{критическое}} = \left\{ df_1 = n_1 - 1 = 12;\ df_2 = n_2 - 1 = 12;\ \dfrac{\alpha}{2} = 0{,}025 \right\} = 3{,}28$. На уровне статистической ошибки 5% *гипотезу о равенстве дисперсий следует отклонить*. Нет оснований считать дисперсии равными. Наблюдаемое на двух выборках различие в оценках дисперсий (которое указывает на то, что дисперсия показателя продаж в первой группе магазинов в 3,75 раза больше дисперсии показателя продаж во второй группе магазинов) считается *<u>не</u>случайным*. Различие выборочных оценок двух дисперсий статистически *значимо*.

Теперь выпишем критерий для проверки главной статистической гипотезы о равенстве средних показателей продаж: $t = \dfrac{\bar{X}_1 - \bar{X}_2}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}}$. Рассчитаем наблюдаемое значение статистики *t*-критерия:

$$t_{\text{наблюдаемое}} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}} = \frac{48 - 65}{\sqrt{\dfrac{450}{13} + \dfrac{120}{13}}} = \frac{-17}{\sqrt{43{,}85}} = \frac{-17}{6{,}62} \approx -2{,}57.$$

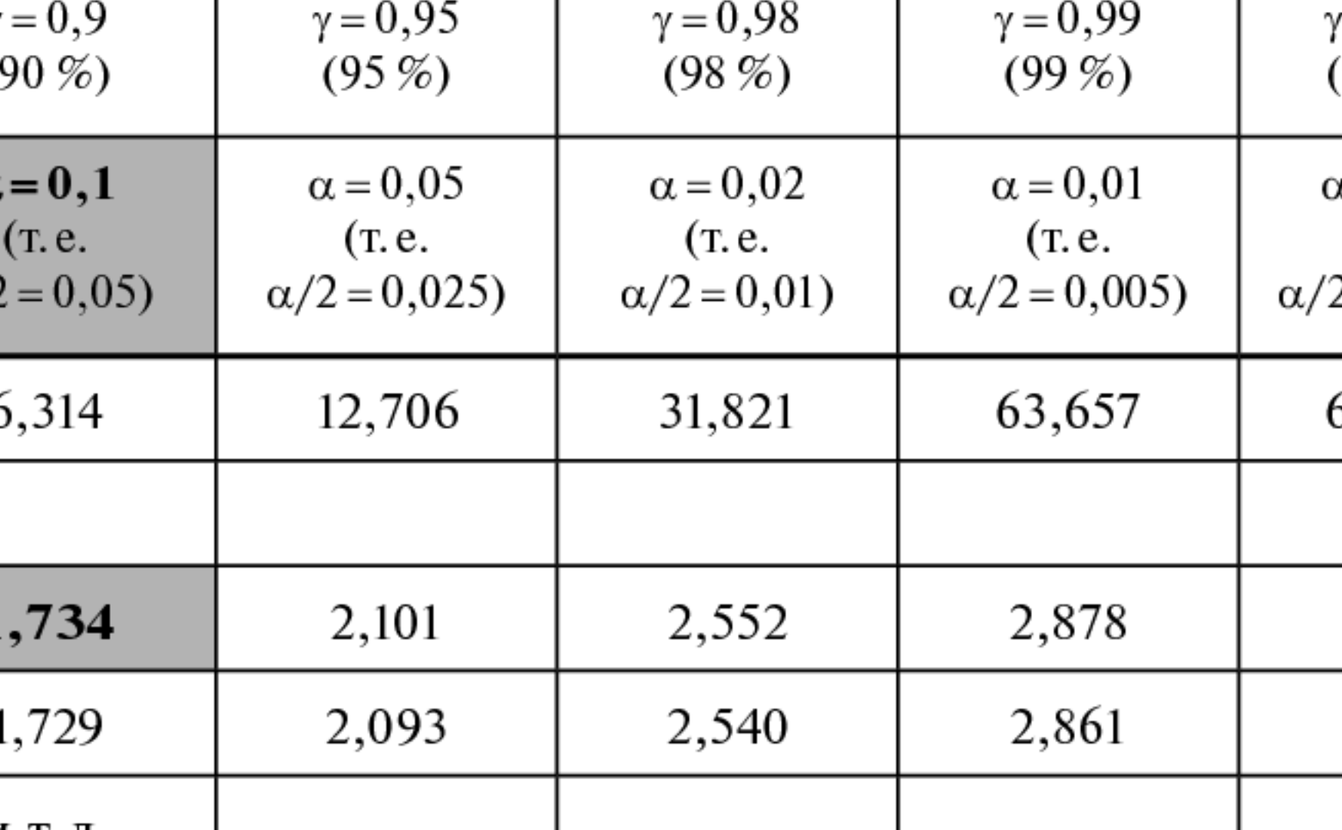
Далее построим критическую область (и полученное фактическое значение статистики критерия сравним с табличным (критическим) значением). Чтобы найти по таблице критическую точку, требуется указать число степеней свободы. Число степеней свободы в данном случае вычисляется довольно трудоёмко:

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}} = \frac{\left(\frac{450}{13} + \frac{120}{13}\right)^2}{\frac{\left(\frac{450}{13}\right)^2}{13-1} + \frac{\left(\frac{120}{13}\right)^2}{13-1}} = 17{,}975 \approx 18.$$

Результат вычисления числа степеней свободы для случая, когда дисперсии различны, редко бывает целым значением, обычно получаемое число — дробное. Поэтому, чтобы иметь возможность определить *t*-статистику Стьюдента по таблице критических точек, значение *df* округляют до целой величины. В данном случае — до 18.

Теперь найдем критическое значение *t*-статистики по таблице *t*-распределения Стьюдента (табл. 6.10). Так, в нашем примере уровень значимости $\alpha = 0{,}05$ (1 – 0,95). Нулевая гипотеза проверяется против

Таблица 6.10. **Фрагмент таблицы для *t*-распределения Стьюдента (двустороннее значение вероятности)**



df	Доверительная вероятность и уровень α				
	γ = 0,9 (90 %)	γ = 0,95 (95 %)	γ = 0,98 (98 %)	γ = 0,99 (99 %)	γ = 0,999 (99,9 %)
	α = 0,1 (т. е. α/2 = 0,05)	α = 0,05 (т. е. α/2 = 0,025)	α = 0,02 (т. е. α/2 = 0,01)	α = 0,01 (т. е. α/2 = 0,005)	α = 0,001 (т. е. α/2 = 0,0005)
1	6,314	12,706	31,821	63,657	636,619
…					
18	**1,734**	2,101	2,552	2,878	3,922
19	1,729	2,093	2,540	2,861	3,883
…	и т. д.				

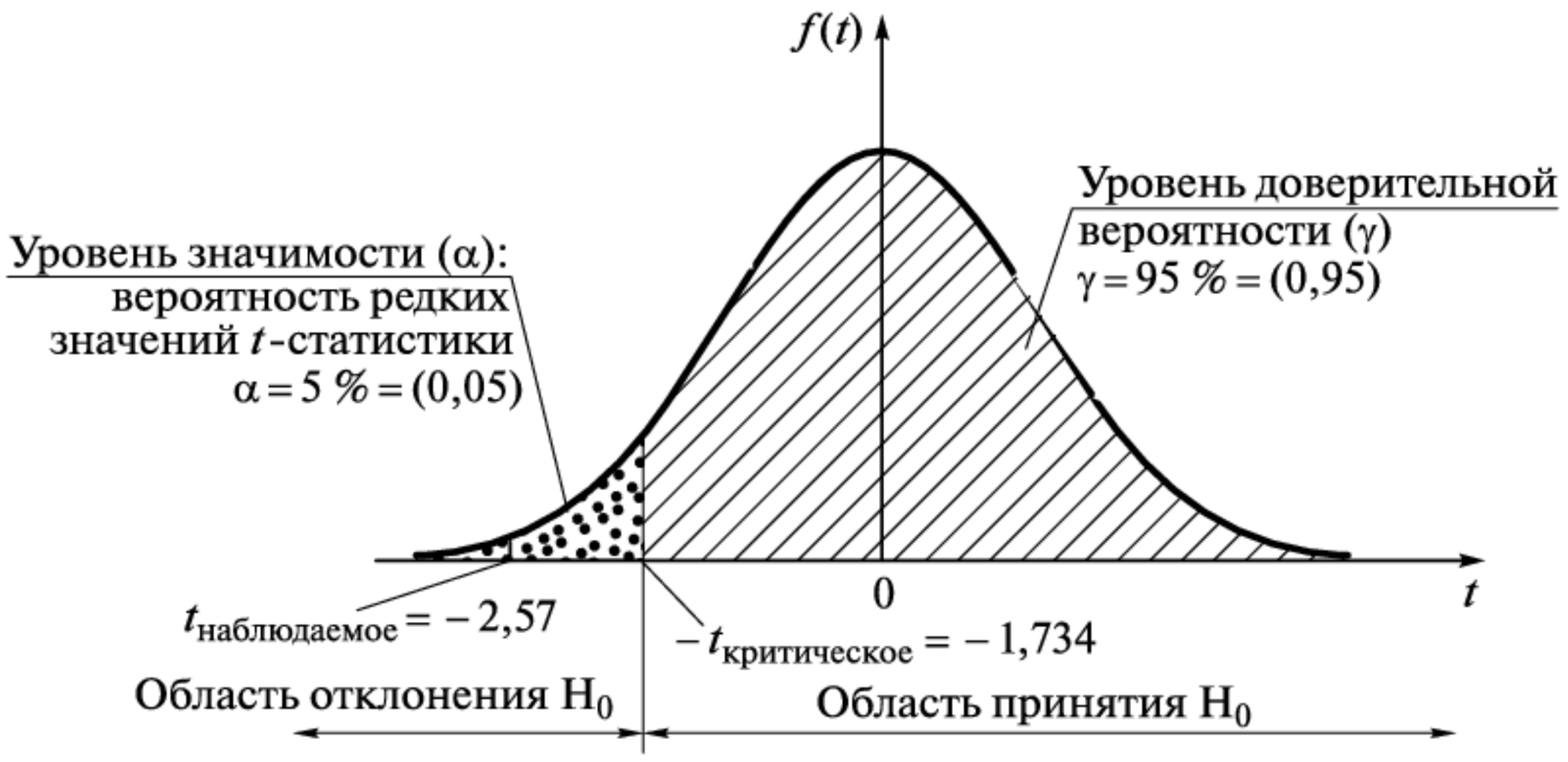


Рис. 6.15. Построение односторонней (левосторонней) критической области (см. пример 6.15 «Спортивный инвентарь»)

направленной альтернативы. Поэтому искомую критическую точку следует посмотреть именно в том столбце таблицы, где отмечено, что площадь под кривой t-распределения в одном «хвосте» составляет 0,05 (а не 0,025). В таблице этот столбец выделен серым цветом. Собственно, это и будет означать, что как раз половина от 10 % (т. е. 5 %) и приходится на один «хвост» распределения. Получим: $t_{\text{критическое}} = -1{,}734$. Отметим наблюдаемое и критическое значения на схематичном графике плотности распределения вероятностей рассматриваемой статистики, построив критическую область (рис. 6.15).

Судя по полученным результатам, H_0 не может быть принята. На уровне значимости $\alpha = 5\ \%$ есть основания отклонить H_0 в пользу направленной альтернативы. Расхождение между двумя выборочными оценками среднего считается неслучайным. Итак, сформулируем содержательный вывод найденного решения: при имеющихся данных выборочного исследования можно говорить, что вид полки (стеллажа), на которой размещена продукция спортивного магазина, определяет объём продаж — объём продаж товара, размещённого на специализированных стеллажах, выше, в сравнении с обычными стеллажами.

§ 6.5. Проверка гипотезы об отсутствии среднего эффекта воздействия для связных (парных) выборок; *t*-критерий Стьюдента

Приведём алгоритм проверки статистической гипотезы ***об отсутствии среднего эффекта воздействия*** *на взаимосвязанных выборках*[1] (*для парных наблюдений*). В § 6.1 было указано о том, что взаимосвязанными выборками называют парные выборки, в каждой

[1] *Related samples* (англ.).

из которых отбор единиц происходит ***НЕ*** независимо друг от друга. Есть целый ряд примеров исследований, в которых применяются именно парные выборки: как правило, это эксперименты, проведение которых основано на изучении свойств исследуемой совокупности *до* и *после* какого-либо воздействия. Например:

- изучается самочувствие пациентов: температура тела *до* и *после* применения некоторого лекарства. В данном случае не сравнивают температуры разных пациентов, а сравнивают *каждого* пациента *с самим собой* — во-первых, его температуру тела в случае когда лекарство не принималось и, во-вторых, его же температуру тела *после* приёма лекарства. С исследовательской точки зрения, сначала нас интересует, существует ли эффект воздействия от применения лекарства относительно самочувствия *каждого* отдельного человека. А далее, обобщая полученные результаты, мы стремимся понять, а существует ли ***средний*** эффект воздействия;
- изучается масса тела *до* и *после* применения диетотерапии;
- изучается уровень социальной фрустрации, замкнутости, агрессии детей *до* и *после* психологического тренинга;
- изучаются остаточные знания студентов по некоторому предмету *в период сессии* и *спустя месяц после* неё.

Примеров такого рода, где тестирование (измерение) исследуемого признака проходит *дважды*, может быть довольно много. Все они касаются сопоставления показателей (характеристик) ***двух групп (срезов)***, ***в основе которых лежит одна и та же (по составу участников) выборка***, ***одна и та же исследуемая совокупность***. Важно отметить, что связные данные могут и не быть однородными выборками (по составу участников), а наоборот — могут состоять из объектов, весьма не похожих между собой.

Можно привести ещё ряд других содержательных примеров парных наблюдений (связных выборок):

- изучая семейные ценности двух разных поколений — детей и родителей, важно не просто сопоставить мнение двух возрастных групп (двух поколений — родителей и детей), а сопоставить их, ***рассмотрев совпадение или несовпадение мнений попарно*** — в разрезе каждой семьи[1];

---

[1] Так, тема курсовой работы одного из студентов НИУ «Высшая школа экономики» звучала как «Идеальная семья в представлениях детей и родителей (на примере опроса в г. Снежинске)». Опрос проводился среди выборки учащихся старших классов в средней школе. Родители этих школьников также стали участниками исследования. В данном исследовании автором (Екатериной Герасимовой) анализировались именно парные наблюдения — ответы детей и ответы их собственных родителей (либо отца, либо матери). Далее оценивалось, а как вообще ***в среднем*** соотносятся представления детей об идеальной модели семьи и представления родителей — *в какой мере они совпадают* (*т. е. одинаковы*) *или различаются* (*т. е. между мнениями есть разрывы*, *преломления*). Кроме того, важно отметить, что выборка опроса не была однородной — напротив, она состояла из семей с самой различной «биографией» (научный руководитель — А. В. Пашкевич).

• когда речь идёт об исследованиях матримониального поведения, можно сравнивать не просто показатели по двум независимым выборкам мужчин и женщин, а сравнивать мнение внутри конкретной супружеской пары: мужчины и женщины;

• при исследованиях организационных конфликтов, или механизмов функционирования корпоративной культуры, в организациях можно (и нужно) рассматривать не просто две независимые выборки данных: например, мнение служащих, с одной стороны, и мнение руководителей — с другой, а «привязать» эти мнения к одному функциональному подразделению.

Всё это крайне важно, поскольку ***вариация (дисперсия, изменчивость показателей) может объясняться индивидуальными особенностями каждой пары наблюдений***. И эти индивидуальные различия можно зафиксировать и учесть (это делает анализ более деликатным), чего нельзя выполнить, например, для независимых выборок.

Рассмотрим статистический критерий, применяемый для проверки *статистической гипотезы об отсутствии среднего эффекта воздействия для связных выборок*. Математически речь будет идти о том, что мы введем новую случайную величину, отражающую этот «эффект воздействия», имеющийся в данных. Начнём с небольшого примера (задачи).

Пример 6.11. «Остаточные знания». При проверке остаточных знаний по курсу «Теория вероятностей и математическая статистика» случайной выборке студентов было предложено решить экзаменационные задачи по математике *в период* сессии и *спустя месяц* после неё. Проверим гипотезу об отсутствии среднего эффекта воздействия (воздействия фактора «времени»). Можно ли считать, что по прошествии некоторого времени уровень знаний остался прежним, а студенты столь же подготовлены и не утратили навыки, не забыли пройденный материал? В таблице представлены результаты: количество *верно* выполненных заданий (из 10 предложенных).

Студент	Ваня	Катя	Оля	Дима	Петя	Ира
Количество верно выполненных заданий (в сессию)	7	7	9	6	8	9
Количество верно выполненных заданий (спустя месяц)	5	8	10	4	8	5

При проверке уровень доверительной вероятности примем равным 0,95. Альтернативную гипотезу примем как направленную:

«Среднее количество верно выполненных заданий спустя месяц после сессии стало ниже, чем количество верно выполненных заданий в период сессии».

Решение. Приступая к решению задачи, подчеркнём, что имеем дело с двумя связными выборками (парными наблюдениями), поскольку этот эксперимент реализован на сравнении двух групп (двух срезов), основанных на одной и той же совокупности субъектов (студентов), но в разные моменты времени.

Каждая выборка (срез) состоит из шести наблюдений (человек). Каждый студент, принявший участие в этом эксперименте, характеризуется парой значений: количеством верно выполненных заданий в сессию (обозначим как X) и количеством верно выполненных заданий спустя месяц после сессии (обозначим как Y). Следовательно, каждому участнику исследования соответствует своя пара (X_i, Y_i). Статистически проверяется гипотеза вида: H_0: «$\bar{U} = 0$» («Средний эффект воздействия равен нулю»). Нулевой гипотезе противопоставляется направленная конкурирующая гипотеза: H_1: «$\bar{U} < 0$» («Средний эффект воздействия меньше нуля, т. е. результат (количество верно выполненных заданий) спустя месяц ниже, чем аналогичный показатель в период сессии»). Уровень значимости $\alpha = 0{,}05 =$ $= (1 - 0{,}95 = 5\,\%)$. Какой же здесь будет статистика критерия? Для начала рассмотрим так называемые ***попарные разности*** (обозначим их как u_i). Для каждого участника разность u_i будет своя и равна: $u_i = (Y_i - X_i)$. Вычтем значения «в период сессии» из значений «спустя месяц», и в дальнейшем будем анализировать эти парные разности. Добавим к имеющейся таблице ещё одну строку и получим:

Студент	Ваня	Катя	Оля	Дима	Петя	Ира
Количество верно выполненных заданий (в сессию)	7	7	9	6	8	9
Количество верно выполненных заданий (спустя месяц)	5	8	10	4	8	5
Разность u_i	$5 - 7 = -2$	$8 - 7 = 1$	$10 - 9 = 1$	$4 - 6 = -2$	$8 - 8 = 0$	$5 - 9 = -4$

Мы сконструировали новую переменную (новую величину) — «***Разность***». Теперь будем работать с этой переменной, рассмотрев ***среднюю*** *разность между показателями*, чтобы охарактеризовать, а как *в среднем* (т. е. в расчёте на одного студента) изменились знания. Для этого вычислим выборочную оценку среднего арифметического:

$$\bar{u} = \frac{\sum_{i=1}^{n} u_i}{n} = \frac{-2+1+1-2+0-4}{6} = \frac{-6}{6} = -1 \text{ (задание)}$$

— ***в среднем*** количество верно решённых задач спустя месяц после сессии *у каждого студента* стало *на единицу ниже*, *чем* в период сессии (*см. знак* «–»). Судя по данной выборочной оценке среднего, спустя месяц студенты в среднем демонстрируют более *низкий* результат («эффект воздействия» отрицательный, имеем знак «–»).

Охарактеризуем вариацию (разброс, изменчивость) этой разности, поскольку у одних студентов (см. исходную таблицу данных) число выполненных заданий изменилось *существенно* (например, на 4 задания у Иры), а у других осталось прежним, без изменений (например, у Пети). Тогда несмещённая оценка дисперсии будет вычислена так:

$$s_u^2 = \frac{\sum_{i=1}^{n} (u_i - \bar{u})^2}{n-1} =$$

$$= \frac{(-2-(-1))^2 + (1-(-1))^2 + (1-(-1))^2 + (-2-(-1))^2 + (0-(-1))^2 + (-4-(-1))^2}{6-1} =$$

$$= \frac{1+4+4+1+1+9}{5} = \frac{20}{5} = 4 \text{ (заданий)}^2.$$

Стандартное отклонение разности составляет $s_u = \sqrt{4} = 2$, т. е. ±2 задания. Следовательно, стандартная ошибка разности (т. е. точность, с которой она нам известна) будет вычислена как $\frac{s_u}{\sqrt{n}}$.

Далее критерий устроен просто: среднее значение разности $\bar{u}$ сопоставляется со своим значением стандартной ошибки $\frac{s_u}{\sqrt{n}}$, т. е. с той точностью, с которой оценена на имеющихся парных выборках средняя разность. Получается выражение $\frac{\bar{u}}{\frac{s_u}{\sqrt{n}}}$. Дробь позволяет определить, *во сколько раз* ***среднее разности*** больше ***своей ошибки***, т. е. выразить среднее значение разности в *единицах стандартной ошибки* этой самой разности!

Итак, статистика критерия имеет вид: $t = \frac{\bar{u}}{\frac{s_u}{\sqrt{n}}}$. Рассматриваемая случайная величина (а она случайная; её наблюдаемые значения могут меняться от одного исследования к другому — если бы в эксперименте приняли участие другие шесть студентов, выбранные случайно,

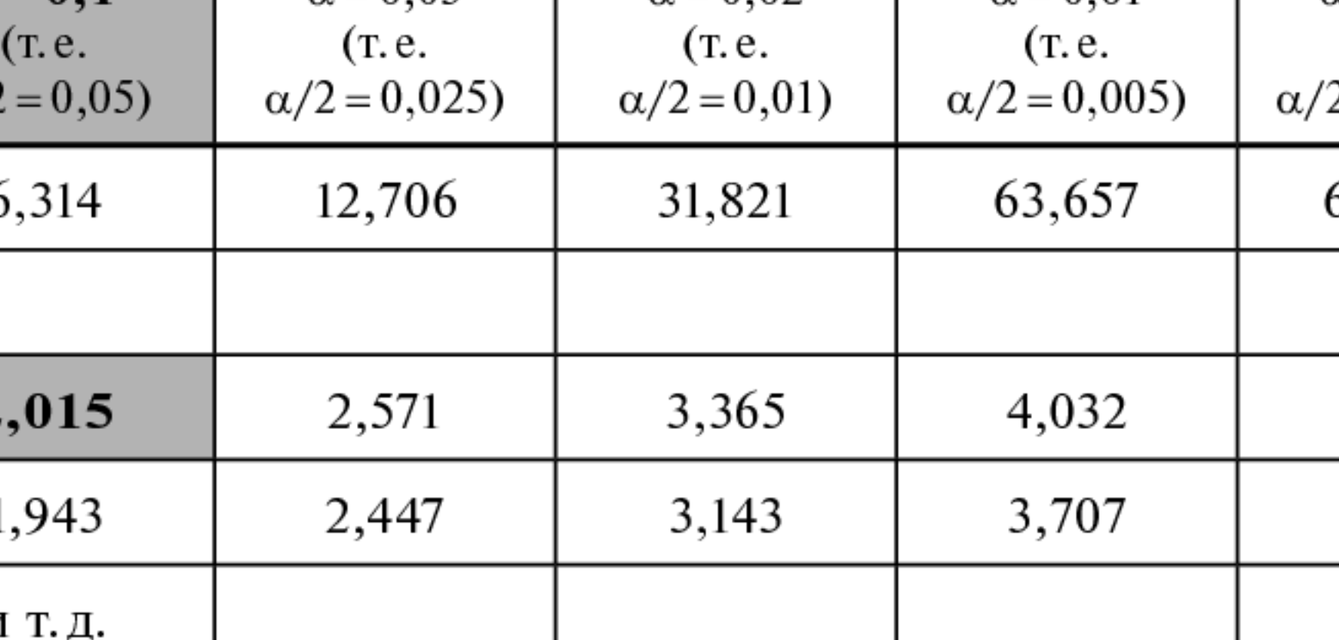
наблюдаемые выборочные оценки показателей были бы, скорее всего, отличны от текущих). Эта величина хорошо согласуется с *t*-распределением Стьюдента, поэтому критическое (максимально допустимое значение статистики критерия при заданном уровне доверительной вероятности) будем искать по таблице Стьюдента (табл. 6.11). Рассчитаем последовательно стандартную ошибку разности и найдём наблюдаемое значение рассматриваемой статистики:

$$\frac{s_u}{\sqrt{n}} = \frac{2}{\sqrt{6}} = \frac{2}{2{,}45} = +0{,}816 \text{ (заданий)}.$$

Так мы оценили погрешность измерений для средней разности $\bar{u}$ (по сути, это стандартное отклонение среднего арифметического разностей u_i):

$$t_{\text{наблюдаемое}} = \frac{\bar{u}}{\frac{s_u}{\sqrt{n}}} = \frac{-1}{+0{,}816} = -1{,}26.$$

Таблица 6.11. **Фрагмент таблицы для *t*-распределения Стьюдента (двустороннее значение вероятности)**



df	Доверительная вероятность и уровень α				
	γ = 0,9 (90 %)	γ = 0,95 (95 %)	γ = 0,98 (98 %)	γ = 0,99 (99 %)	γ = 0,999 (99,9 %)
	α = **0,1** (т. е. α/2 = 0,05)	α = 0,05 (т. е. α/2 = 0,025)	α = 0,02 (т. е. α/2 = 0,01)	α = 0,01 (т. е. α/2 = 0,005)	α = 0,001 (т. е. α/2 = 0,0005)
1	6,314	12,706	31,821	63,657	636,619
…					
5	**2,015**	2,571	3,365	4,032	6,869
6	1,943	2,447	3,143	3,707	5,959
…	и т. д.				

Чтобы сказать, попадает ли это значение (–1,26) в критическую область для ситуации, когда средний «эффект воздействия» равен нулю (отсутствует), воспользуемся таблицей *t*-распределения (*число ограничений свободы вариации для критерия равно* $df = n - 1$).

Обратите внимание на то обстоятельство, что альтернативная гипотеза является направленной (и при этом *лево*сторонней): H_1: «$\bar{U} < 0$» («средний эффект воздействия меньше нуля, т. е. результат (количество верно выполненных заданий) спустя месяц в среднем стал ниже, чем результат в период сессии»).

Также обратите внимание, как находится критическое значение в данном примере. Здесь требуется рассмотреть одностороннюю критическую область, поэтому вероятность *в одном «хвосте»* распределения должна составить α (*целое*, *без деления на 2*), т. е. $\alpha = 0{,}05$ (1 – 0,95). Поэтому искомую критическую точку следует искать именно в том столбце таблицы, где отмечено, что площадь под кривой *t*-распределения в одном «хвосте» составляет 0,05 (а не 0,025). В таблице этот столбец выделен серым цветом.

Собственно, это и будет означать, что как раз половина от 10 % (т. е. 5 %) и приходится на один «хвост» распределения. Получим: $t_{\text{критическое}} = -2{,}015$.

Отметим наблюдаемое и критическое значения на схематичном графике плотности распределения вероятностей рассматриваемой статистики (рис. 6.16).

Наблюдаемое значение статистики критерия не попало в критическую область, оно не принадлежит области редких значений, поэтому (при имеющихся данных исследования) *нет очевидных оснований* отклонить нулевую гипотезу. Итак, на уровне доверительной вероятности $\gamma = 95\ \%$ H_0 не отклоняется (принимается).

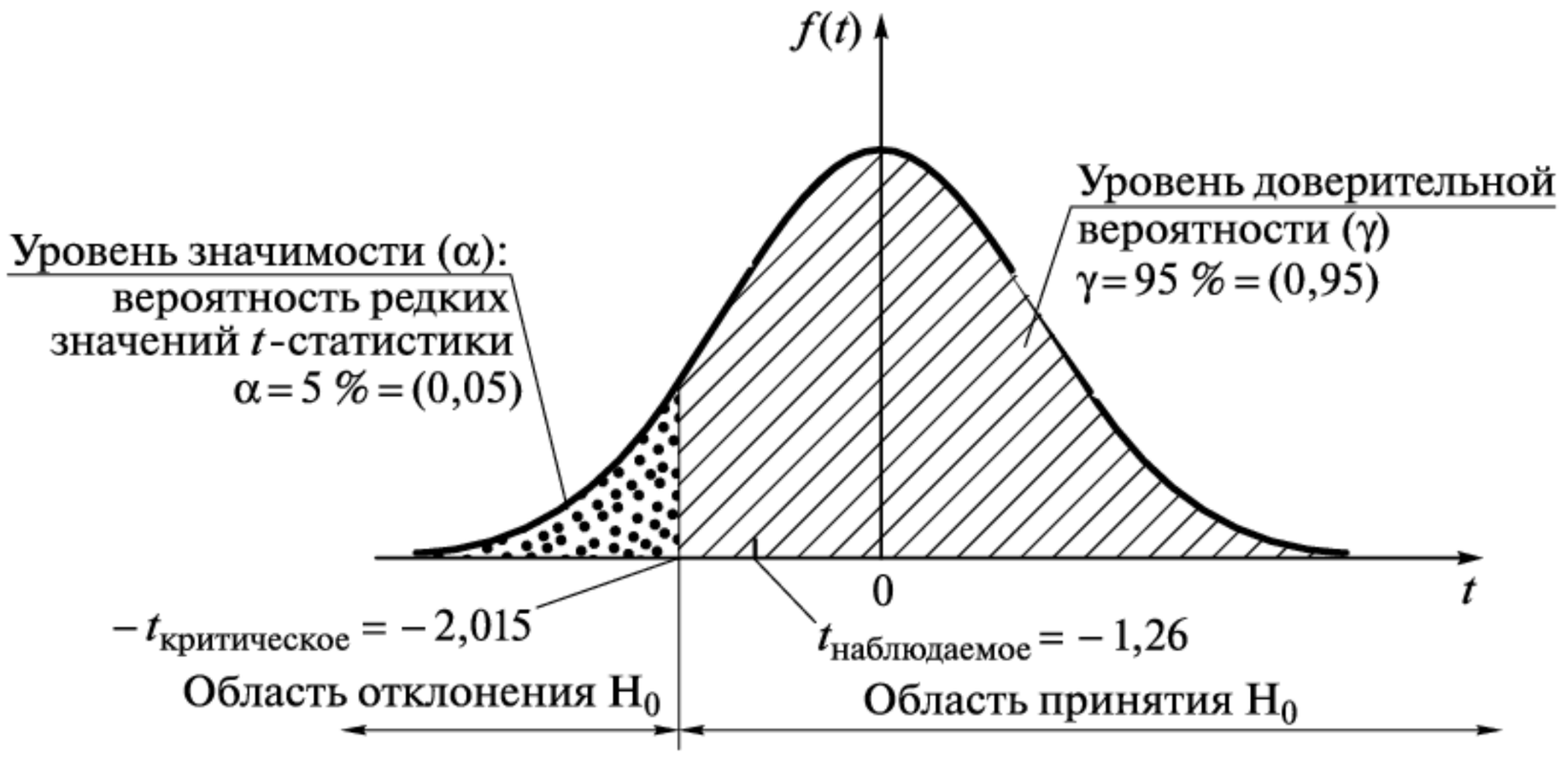


Рис. 6.16. Построение односторонней (левосторонней) критической области (см. пример 6.11 «Остаточные знания»)

Расхождение между успешностью выполнения заданий в период сессии и после неё можно считать абсолютно случайным. Средний «эффект» отсутствует, т. е. равен нулю.

Сформулируем содержательный вывод найденного решения: обосновано, что различия в знаниях «во время сессии и после неё» статистически незначимы. Уровень подготовки/знаний одинаков. Наблюдаемые на выборках *различия* (*колебания*) *статистически незначимы*. Можно считать (обобщая результаты исследования на генеральную совокупность), что остаточные знания остались на прежнем уровне. При имеющихся данных.

Итак, подведём итог: в случае наличия в исследовании связных выборок (или парных наблюдений) для проверки гипотезы о равенстве нулю среднего «эффекта воздействия» используется t-критерий Стьюдента следующего вида: $t = \dfrac{\bar{u}}{\dfrac{s_u}{\sqrt{n}}}$. В основе его вычисления лежит рассмотрение попарных разностей u_i, расчёт средней разности между показателями на частном (индивидуальном) уровне $\bar{u}$ и характеристика разброса значений этой разности s_u.

Выборочная оценка средней разности «нормируется» на свою стандартную ошибку $\dfrac{s_u}{\sqrt{n}}$. Принципиально важным в задачах такого типа является вычисление частных (индивидуальных) разностей, т. е. попарно, «в связке».

§ 6.6. Проверка гипотезы о равенстве долей; *z*-критерий

Крайне важной для практики социологических исследований является задача сопоставления доле́й некоторого изучаемого признака в двух генеральных совокупностях, в двух общностях. Перейдём к изучению алгоритма проверки статистической гипотезы о равенстве долей в двух исследуемых совокупностях. Содержательную постановку задачи проиллюстрируем на небольшом примере.

Пример 6.12. «Кому выносить мусор?». Пусть в исследовании по разным показателям сравнивается структура распределения домашних обязанностей между мужем и женой в двух разных обществах («западного» и «восточного» типов[1]). В частности, мужчинам был задан вопрос о том, кто чаще выносит мусор — он или жена. Пред-

[1] В контексте этой задачи примем эти названия условно, как метафоры. Можно прибегнуть к использованию и других клише: «современного» и «традиционного» типов.

положим, что было опрошено 130 мужчин (представителей общества «западного» типа), из которых 80 сказали, что чаще выносят мусор именно они. Из 110 опрошенных мужчин (представителей общества «восточного» типа) 75 сказали, что они чаще выносят мусор. Можно ли утверждать, что распределение ролей в семье между мужем и женой в таком «деликатном» вопросе, как *кому выносить мусор*, одинаково в обществах «западного» и «восточного» типа? Проверим статистическую гипотезу о равенстве долей против направленной альтернативы на уровне доверия результатам 95 %.

Решение. Изучаются две группы мужчин — представители «западных» обществ и представители «восточных» обществ. Это эмпирический объект исследования. В фокусе исследования — распределение ролей между супругами при ведении домашнего хозяйства. Объёмы выборок соответственно равны: $n_1 = 130$ человек и $n_2 = 110$ человек. Число мужчин, которые обладают интересующим нас признаком: «мусор чаще выношу именно я, а не супруга», обозначим буквой k. Тогда по выборкам имеем: $k_1 = 80$ человек и $k_2 = 75$ человек.

В задаче потребуется проверить статистическую гипотезу о равенстве долей: H_0: «$P_1 = P_2$ (обе доли *равны*)». Нулевой гипотезе противопоставляется направленная конкурирующая гипотеза: H_1: «$P_2 > P_1$ (доля во второй совокупности больше, чем доля в первой; т. е. доля мужчин, которые выносят мусор чаще, чем их жены, выше в обществах «восточного» типа)». Уровень значимости $\alpha = 0{,}05$ (1 – 0,95). Критерий для проверки гипотезы устроен аналогично критерию, рассмотренному для проверки гипотезы о равенстве математических ожиданий. Вспомним, что в общем виде он выглядел как $t = \dfrac{\bar{X}_1 - \bar{X}_2}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}}$.

В рассматриваемом случае имеем дело со схемой Бернулли (долей случаев, указывающих на присутствие определённой характеристики, и долей случаев, указывающих на её отсутствие). «Успехом» в задаче считаем ответ «я выношу мусор чаще», «неудачей» — альтернативный ответ респондента. Поскольку объёмы выборок чаще всего большие, то критерий примет вид:

$$t \to z = \frac{\bar{p}_1 - \bar{p}_2}{\sqrt{\dfrac{\bar{p}\bar{q}}{n_1} + \dfrac{\bar{p}\bar{q}}{n_2}}} = \frac{\bar{p}_1 - \bar{p}_2}{\sqrt{\bar{p}\bar{q}\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}},$$

где $\bar{p} = \dfrac{k_1 + k_2}{n_1 + n_2}$ — ***средняя доля*** мужчин в обеих выборках, которые выносят мусор; $\bar{q} = 1 - \bar{p} = 1 - \dfrac{k_1 + k_2}{n_1 + n_2}$ — ***средняя доля*** мужчин в обеих

выборках, которые мусор ***не*** выносят (а выносят их жены). Здесь k_1 — количество наблюдений в первой выборке (группе), рассматриваемых как «успех»; k_2 — количество наблюдений во второй выборке (группе), рассматриваемых как «успех»; $p_1 = \frac{k_1}{n_1}$ — выборочная ***оценка доли*** в *первой* выборке; $p_2 = \frac{k_2}{n_2}$ — выборочная ***оценка доли*** во *второй* выборке. В знаменателе указано стандартное отклонение (т. е. стандартная ошибка) разности долей, представленной в числителе дроби. Поясним это. Если предположить, что дисперсии *одинаковые* в обеих выборках, то каждую из них можно заменить на ***среднее значение дисперсий***: $\bar{p}\bar{q}$. Иными словами, если действительно доли равны, то дисперсия разности долей может быть записана (представлена) как

$$\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2} = \frac{\bar{p}\bar{q}}{n_1} + \frac{\bar{p}\bar{q}}{n_2},$$

где $\bar{p}$ — *средняя* доля, *общая* для *обеих* выборок (т. е. *безотносительно* к первой или второй). Это выражение и представлено в знаменателе.

Реализуем решение задачи, найдя соответствующие рассмотренному критерию компоненты:

$$p_1 = \frac{k_1}{n_1} = \frac{80}{130} = 0{,}615 = 61{,}5\,\%;\ p_2 = \frac{k_2}{n_2} = \frac{75}{110} = 0{,}682 = 68{,}2\,\%;$$

$$\bar{p} = \frac{k_1 + k_2}{n_1 + n_2} = \frac{80 + 75}{130 + 110} = \frac{155}{240} = 0{,}646 = 64{,}6\,\%;$$

$$\bar{q} = 1 - \bar{p} = 1 - \frac{k_1 + k_2}{n_1 + n_2} = 1 - 0{,}646 = 0{,}354 = 35{,}4\,\%;$$

$$z_{\text{наблюдаемое}} = \frac{p_1 - p_2}{\sqrt{\bar{p}\bar{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{61{,}5\,\% - 68{,}2\,\%}{\sqrt{64{,}6\,\% \cdot 35{,}4\,\% \cdot \left(\frac{1}{130} + \frac{1}{110}\right)}} =$$

$$= \frac{-6{,}7\,\%}{\sqrt{38(\%)^2}} = \frac{-6{,}7\,\%}{6{,}20\,\%} = -1{,}08.$$

Теперь требуется определить (задать) критическую область. Полученное фактическое значение статистики критерия сравним с табличным (критическим) значением. Найдём его по таблице стандартного нормального распределения: $z_{\text{критическое}} = -1{,}65$ (табл. 6.12). Отметим наблюдаемое и критическое значения на схематичном графике плотности распределения вероятностей рассматриваемой статистики (рис. 6.17).

Сопоставляем два значения: $z_{\text{наблюдаемое}} > z_{\text{критическое}}$ (–1,08>–1,65). Наблюдаемое значение (–1,08) не принадлежит критической области, а принадлежит области принятия нулевой гипотезы, поэтому нет статистических оснований отклонить нулевую гипотезу о равенстве долей. На уровне доверительной вероятности $\gamma = 95\ \%$ H_0 принимается. Расхождение между выборочными долями считаются случайными. Получено вполне типичное (характерное для ситуации равенства долей) наблюдаемое значение z-статистики.

Сформулируем содержательный вывод найденного решения: обосновано, что различия в долях (наблюдаемые на имеющихся

Таблица 6.12. **Фрагмент таблицы стандартного нормального *z*-распределения (интеграл от 0 до *z*)**

Z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,0000	0,0040	0,0080	0,0120	0,0160	0,0199	0,0239	0,0279	0,0319	0,0359
0,1	0,0398	0,0438	0,0478	0,0517	0,0557	0,0596	0,0636	0,0675	0,0714	0,0753
...										
1,5	0,4332	0,4345	0,4357	0,4370	0,4382	0,4394	0,4406	0,4418	0,4429	0,4441
1,6	0,4452	0,4463	0,4474	0,4484	0,4495	0,4505	0,4515	0,4525	0,4535	0,4545
1,7	0,4554	0,4564	0,4573	0,4582	0,4591	0,4599	0,4608	0,4616	0,4625	0,4633
1,8	0,4641	0,4649	0,4656	0,4664	0,4671	0,4678	0,4686	0,4693	0,4699	0,4706
...	и т. д.									

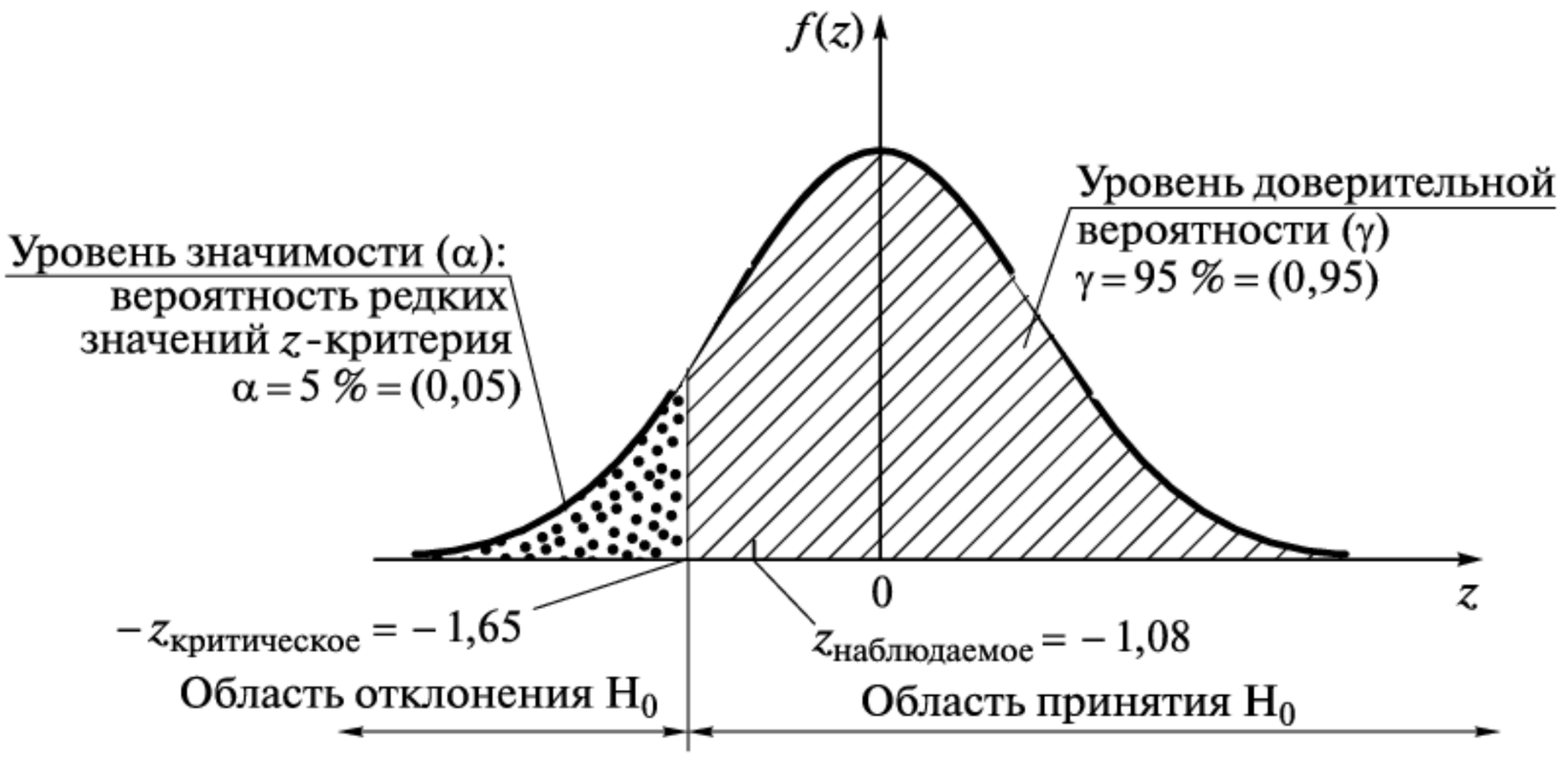


Рис. 6.17. Построение односторонней (левосторонней) критической области (см. пример 6.12 «Кому выносить мусор?»)

выборках) статистически незначимы. Можно утверждать, что распределение ролей между мужем и женой в таком вопросе, как *кому выносить мусор*, действительно одинаково в семьях «западного» и «восточного» типа. Можно считать (обобщая результаты исследования на генеральную совокупность), что доли равны. При имеющихся данных.

Пример 6.13. «О судах и судьях». В период Еженедельного опроса «ФОМнибус»[1] от 23—24 июня 2012 г., проведённого в 43 субъектах РФ (в 100 населенных пунктах) жителям России был задан вопрос: «Как Вы в целом оцениваете деятельность российских судов и судей — положительно или отрицательно?». Распределение ответов жителей Москвы и сельских поселений выглядит следующим образом:

Как Вы в целом оцениваете деятельность российских судов и судей — положительно или отрицательно?	Жители	
	Москвы	сельских поселений
Положительно (+)	15	105
Отрицательно (–)	62	129
Затрудняюсь ответить (?)	58	156
Всего, чел.	135	390

Можно ли считать, что доли людей, негативно оценивающих деятельность российских судов и судей, одинаковы у жителей Москвы и сельских поселений? Уровень доверительной вероятности принять равным 0,99. Проверку нулевой гипотезы осуществить против ненаправленной альтернативы.

Решение. Необходимо проверить статистическую гипотезу о равенстве долей: H_0: «$P_1 = P_2$ (обе доли *равны*)». Нулевой гипотезе противопоставляется ненаправленная конкурирующая гипотеза: H_1: «$P_1 \neq P_2$» (доли не равны)». Уровень значимости $\alpha = 1\,\%$, а $\alpha/2$ (т. е. вероятность попадания в «хвосты» распределения статистики критерия) составляет $0{,}01 : 2 = 0{,}005 = 0{,}5\,\%$ (полпроцента).

Критерием проверки гипотезы будет служить z-критерий:

$$z = \frac{p_1 - p_2}{\sqrt{\bar{p}\bar{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}.$$

[1] Исходные данные опроса ФОМ размещены в разделе «Безопасность и право» см.: [http://fom.ru/Bezopasnost-i-pravo/10551] (проверено 01.08.2012).

Здесь $\bar{p} = \frac{k_1 + k_2}{n_1 + n_2}$ — средняя доля жителей (Москвы и сельских поселений), которые отрицательно оценивают деятельность российских судов и судей; $\bar{q} = 1 - \bar{p} = 1 - \frac{k_1 + k_2}{n_1 + n_2}$ — средняя доля жителей, чья позиция иная — они оценивают деятельность судов и судей положительно или затрудняются дать однозначный ответ; k_1 — количество наблюдений в первой выборке (жители Москвы), рассматриваемых как «успех»; k_2 — количество наблюдений во второй выборке (жители сел), рассматриваемых как «успех»; $p_1 = \frac{k_1}{n_1}$ — выборочная оценка доли в первой выборке; $p_2 = \frac{k_2}{n_2}$ — выборочная оценка доли во второй выборке. Реализуем решение, найдя соответствующие рассмотренному критерию компоненты:

$$p_1 = \frac{k_1}{n_1} = \frac{62}{135} = 0{,}46 = 46\,\%;\ p_2 = \frac{k_2}{n_2} = \frac{129}{390} = 0{,}33 = 33\,\%;$$

$$\bar{p} = \frac{k_1 + k_2}{n_1 + n_2} = \frac{62 + 129}{135 + 390} = \frac{191}{525} = 0{,}36 = 36\,\%;$$

$$\bar{q} = 1 - \bar{p} = 1 - \frac{k_1 + k_2}{n_1 + n_2} = 1 - 0{,}36 = 0{,}64 = 64\,\%;$$

$$z_{\text{наблюдаемое}} = \frac{p_1 - p_2}{\sqrt{\bar{p}\bar{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{46\,\% - 33\,\%}{\sqrt{36\,\% \cdot 64\,\% \cdot \left(\frac{1}{135} + \frac{1}{390}\right)}} =$$

$$= \frac{13\,\%}{\sqrt{22{,}97(\%)^2}} = \frac{13\,\%}{4{,}79\,\%} = +2{,}71.$$

Полученное фактическое значение статистики критерия нужно сравнить с табличным (критическим) значением. Найдем его по таблице стандартного нормального распределения (табл. 6.13): $z_{\text{критическое}} = +2{,}58$ (так как площадь под кривой стандартного нормального распределения от нуля до искомого значения $z_{\text{критическое}}$ должна составлять $50\,\% - 0{,}5\,\% = 49{,}5\,\% = 0{,}495$). Отметим наблюдаемое и критическое значения на схематичном графике плотности распределения вероятностей рассматриваемой статистики (рис. 6.18).

Сравниваем два значения: $z_{\text{наблюдаемое}} > z_{\text{критическое}}$ $(+2{,}71 > +2{,}58)$ (значение +2,71 попадает в критическую область), поэтому есть все статистические основания отклонить нулевую гипотезу о равенстве долей. На уровне значимости $\alpha = 1\,\%$ нулевая гипотеза (H_0) откло-

Таблица 6.13. **Фрагмент таблицы стандартного нормального *z*-распределения (интеграл от 0 до *z*)**

Z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
...										
2,3	0,4893	0,4896	0,4898	0,4901	0,4904	0,4906	0,4909	0,4911	0,4913	0,4916
2,4	0,4918	0,4920	0,4922	0,4925	0,4927	0,4929	0,4931	0,4932	0,4934	0,4936
2,5	0,4938	0,4940	0,4941	0,4943	0,4945	0,4946	0,4948	0,4949	0,4951	0,4952
2,6	0,4953	0,4955	0,4956	0,4957	0,4959	0,4960	0,4961	0,4962	0,4963	0,4964
...	и т. д.									

няется и принимается альтернативная (H_1). Расхождение между выборочными долями считаются *не*случайными. Мы наблюдаем совершенно *не*типичное (*не*характе́рное) значение z-статистики для ситуации, если бы в реальности существовало равенство долей.

Сформулируем содержательный вывод найденного решения: при имеющихся данных обоснованно говорить, что различия в долях статистически значимы. Обобщая результаты выборочного опроса на всю генеральную совокупность жителей страны, на уровне ошибки 1 % можно утверждать, что доли населения Москвы и сельских поселений, отрицательно оценивающих деятельность судов и судей, действительно разные.

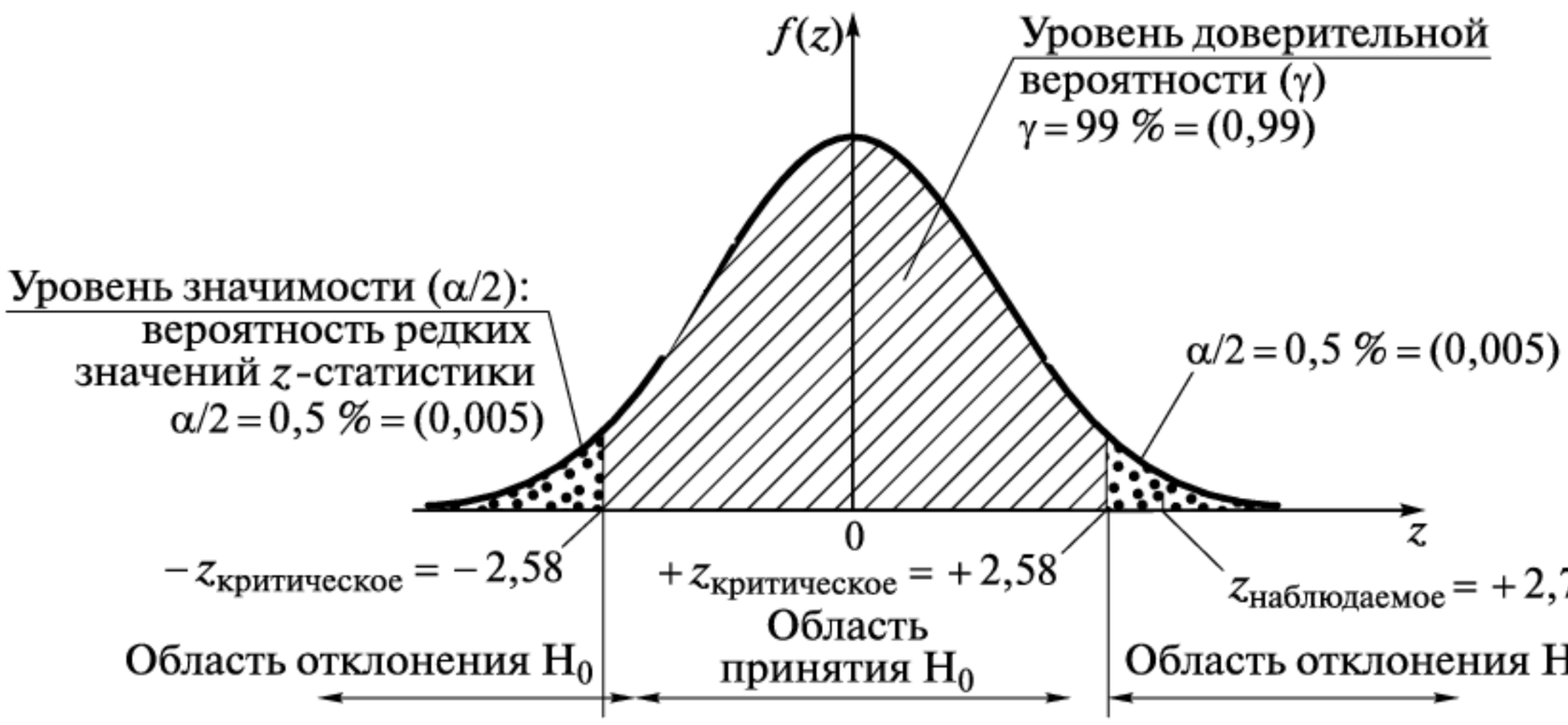


Рис. 6.18. Построение двусторонней критической области (см. пример 6.13 «О судах и судьях»)

§ 6.7. Проверка гипотезы о статистической незначимости коэффициентов корреляции Пирсона и Спирмена

В гл. 3, посвященной коэффициентам связи, мы научились вычислять выборочные оценки двух коэффициентов корреляции — коэффициента линейной корреляции К. Пирсона и коэффициента ранговой корреляции Ч. Спирмена. Для того чтобы проверить статистическую значимость этих коэффициентов и иметь возможность обобщить результат выборки на всю генеральную совокупность, следует рассмотреть специальные статистические критерии (правила), позволяющие осуществить такой переход.

Предположим, что по выборке, состоящей из 15 пар наблюдений, получена оценка коэффициента линейной корреляции Пирсона, равная 0,33. Теснота связи выражена слабо, но она есть. Или её всё-таки нет? Есть ли линейная связь на самом деле в генеральной совокупности? Действительно ли зависимость присутствует, хотя и выражена слабо? Или 0,33 — случайный результат выборочного исследования?

Для того чтобы проверить, значим ли коэффициент корреляции Пирсона, рассчитанный по выборке, в математике проверяется статистическая гипотеза вида: «$\rho = 0$» (читается: «Ро равно нулю»). Буквой «ρ» («ро») обозначается истинное значение коэффициента корреляции в генеральной совокупности, которое нам не известно. Стартовое предположение, из которого мы исходим, таково: коэффициент корреляции на самом-то деле равен нулю (т. е. теснота линейной связи не выражена вовсе). Так, нулевая гипотеза записывается в виде: H_0: «$\rho = 0$». Альтернативой ей служат ненаправленная гипотеза H_1: «$\rho \neq 0$» или направленные гипотезы вида: H_1: «$\rho > 0$» или H_1: «$\rho < 0$». Уровень значимости определяется исходя из того, односторонний или двусторонний критерий используется при решении.

Статистическая значимость коэффициента корреляции Пирсона. Статистический *t*-критерий, применяемый для проверки гипотезы о равенстве нулю коэффициента корреляции, выглядит следующим образом: $t = r\sqrt{\frac{n-2}{1-r^2}}$, где r — выборочная оценка коэффициента линейной корреляции Пирсона (как её вычислять, см. материал, изложенный в гл. 3), а $(n - 2)$ — число ограничений свободы вариации. Почему, кстати, $(n - 2)$? Мы говорили о том, что число степеней свободы — это число наблюдений, оставшихся *как бы* «нетронутыми» после того, как часть информации о них (наблюдениях) мы «израсходовали» в других предшествующих видах анализа. Расчёт коэффициента корреляции ссылается (опирается) на вы-

числение двух выборочных дисперсий[1] — по переменной X и переменной Y. Числом степеней свободы для коэффициента корреляции Пирсона будет число пар наблюдений выборки, уменьшенное на число параметров, которые оцениваются исследователем до того, как вычисляется этот текущий коэффициент. А этих параметров ровно два — *дисперсия X* и *дисперсия Y*. Поэтому число ограничений свободы вариации, которым мы располагаем при вычислении коэффициента корреляции, составляет $(n - 2)$.

Рассмотрим, как «работает» t-критерий Стьюдента для проверки гипотезы о равенстве нулю коэффициента корреляции Пирсона. Для того чтобы отвергнуть нулевую гипотезу и принять конкурирующую, наблюдаемое значение статистики t-критерия должно быть довольно большим. Когда мы получим относительно большое значение статистики критерия $t = r\sqrt{\frac{n-2}{1-r^2}}$?

Во-первых, когда квадрат выборочной оценки коэффициента корреляции будет стремиться к единице, если и ***сам коэффициент*** (по модулю) стремится к значению «1», т. е. к своему максимально возможному значению. Что в этом случае происходит с дробью t-критерия? Значение $(1 - r^2)$ стремится к нулю, а дробь $\frac{n-2}{1-r^2}$ стремится к бесконечности $(+\infty)$. Тогда и всё выражение $r\sqrt{\frac{n-2}{1-r^2}}$ стремится к большому числу $(+\infty)$. Так, шансы на то, что истинное значение коэффициента корреляции равно нулю, становятся ничтожно малы, а уверенность в том, что коэффициент корреляции статистически значим (достоверен), напротив, растёт.

Во-вторых, немаловажную роль играет ***объём выборки***, на которой произведены вычисления коэффициента корреляции. Если $n \to +\infty$, то и числитель дроби $\frac{n-2}{1-r^2}$ и сама дробь стремится к $+\infty$, а значит, и величина статистики t-критерия стремится к большому значению. Большой объём выборки свидетельствует о том, что выборочная оценка коэффициента сделана с высокой точностью. Поэтому мы верим, что выборочная оценка не сильно отличается от истинного значения в генеральной совокупности. В случае же если всё происходит наоборот, а именно: размер выборки мал и/или выборочная оценка коэффициента корреляции крайне низка $(\to 0)$,

---

[1] Точнее: на расчёт ***суммы квадратов отклонений*** по X и по Y — в знаменателе формулы для вычисления выборочной оценки коэффициента линейной корреляции присутствуют величины $\sum_{i=1}^{n}(x_i - \bar{x})^2$ и $\sum_{i=1}^{n}(y_i - \bar{y})^2$.

то и наблюдаемое значение *t*-статистики стремится к нулю, что в свою очередь означает, что с высокой вероятностью нулевая гипотеза об отсутствии линейной зависимости между признаками будет считаться истинной.

Пример 6.14. «Зрители театров и безработица в РФ». Рассмотрим некоторые данные официальной статистики за 2010 г. для восьми федеральных округов РФ. Имеются сведения о численности зрителей театров на 1 000 человек населения и средних уровнях безработицы в этих округах[1]. Безработица определяется как отношение численности незанятого населения к общей численности трудоспособного населения (%). Численность зрителей определяется отношением численности зрителей к среднегодовой численности населения федерального округа. В результате располагаем таблицей:

Федеральный округ	Численность зрителей театров на 1 000 человек населения (X)	Уровень безработицы, % (Y)
Центральный	262	4,7
Северо-Западный	279	6,2
Южный	123	7,7
Северо-Кавказский	90	16,9
Приволжский	210	7,6
Уральский	209	8,0
Сибирский	239	8,7
Дальневосточный	185	8,7

Можно ли считать, что две эти социальные характеристики коррелируют? Рассчитайте коэффициент линейной корреляции Пирсона и проверьте его статистическую достоверность на уровне значимости 5 %. Проверку осуществите против направленной альтернативы: «Коэффициент корреляции меньше нуля» (т. е. суще-

[1] Подробно с материалами Федеральной службы государственной статистики «Регионы России. Социально-экономические показатели-2011» можно ознакомиться по ссылке [http://www.gks.ru/bgd/regl/b11_14p/Main.htm] (проверено 02.08.2012).

ствует *обратная линейная* зависимость между рассматриваемыми признаками: *чем выше* уровень безработицы, *тем ниже* численность зрителей театров на 1 000 человек населения — люди ходят в театр реже и/или в меньшем количестве. И наоборот — чем *ниже* уровень безработицы, тем в *большей* степени люди заинтересованы в посещении театров).

Решение. Напомним, что задачи на нахождение выборочной оценки коэффициента линейной корреляции решаются в несколько этапов (шагов): Строится диаграмма рассеяния точек. Это позволяет сделать наглядной имеющуюся информацию об изучаемых признаках. Вычисляются средние значения для обеих переменных. Вычисляются основные компоненты формулы: находится сумма произведений, показывающих сопряжённую изменчивость значений X и Y — в разрезе каждой пары наблюдений: $\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$, а также $\sum_{i=1}^{n}(x_i - \bar{x})^2$ и $\sum_{i=1}^{n}(y_i - \bar{y})^2$ — суммы квадратов отклонений для X и для Y. Далее определяется значение коэффициента корреляции для выборки. Только после этого (выполненных вычислений) требуется проверить нулевую гипотезу о равенстве нулю коэффициента корреляции. По итогам статистической проверки даётся заключительный содержательный вывод.

Последовательно реализуем этот алгоритм решения применительно к данной задаче. Сначала построим диаграмму (облако) рассеяния, затем схематично нарисуем линию тренда (рис. 6.19). Точками являются восемь федеральных округов РФ. Каждая точка характеризуется парой координат — значением численности зрителей театров и значением уровня безработицы.

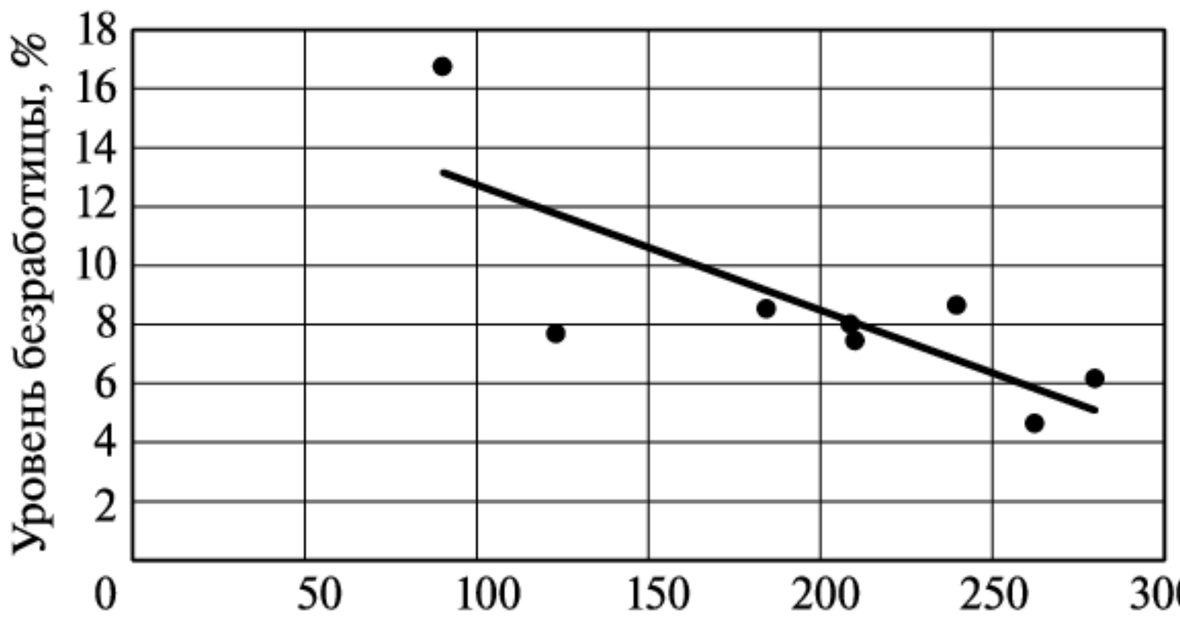


Рис. 6.19. Диаграмма рассеяния (см. пример 6.14 «Зрители театров и безработица в РФ»)

Далее считаем средние арифметические:

$$\bar{x} = \frac{262 + 279 + 123 + 90 + \ldots + 239 + 185}{8} = 199{,}6 \approx 200 \text{ (человек)}$$

— средняя численность зрителей театров на 1 000 жителей в федеральном округе;

$$\bar{y} = \frac{4{,}7 + 6{,}2 + 7{,}7 + 16{,}9 + \ldots + 8{,}7 \cdot 2}{8} \approx 8{,}6 \ (\%)$$

— средний уровень безработицы в федеральном округе.

Находим основные компоненты формулы для вычисления выборочной оценки коэффициента корреляции. Для этого построим расчётную таблицу, в которой укажем ход вычислений таких величин, как $(x_i - \bar{x})$, $(x_i - \bar{x})^2$, $(y_i - \bar{y})$, $(y_i - \bar{y})^2$, $(x_i - \bar{x})(y_i - \bar{y})$, $\sum_{i=1}^{n}(x_i - \bar{x})^2$, $\sum_{i=1}^{n}(y_i - \bar{y})^2$ и $\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$. В результате получим:

№ пары	x_i	y_i	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$
1	262	4,7	262 – 200 = 62	4,7 – 8,6 = –3,9	–241,8	3 844	15,21
2	279	6,2	279 – 200 = 79	6,2 – 8,6 = –2,4	–189,6	6 241	5,76
3	123	7,7	123 – 200 = –77	7,7 – 8,6 = –0,9	69,3	5 929	0,81
4	90	16,9	90 – 200 = –110	16,9 – 8,6 = 8,3	–913	12 100	68,89
5	210	7,6	210 – 200 = 10	7,6 – 8,6 = –1	–10	100	1
6	209	8,0	209 – 200 = 9	8,0 – 8,6 = –0,6	–5,4	81	0,36
7	239	8,7	239 – 200 = 39	8,7 – 8,6 = 0,1	3,9	1 521	0,01
8	185	8,7	185 – 200 = –15	8,7 – 8,6 = 0,1	–1,5	225	0,01
—	$\bar{x} = \mathbf{200}$	$\bar{y} = \mathbf{8{,}6}$	—	—	$\sum_{i=1}^{n=8}(x_i - \bar{x})(y_i - \bar{y}) = -\mathbf{1\,288{,}1}$	$\sum_{i=1}^{n=8}(x_i - \bar{x})^2 = \mathbf{30\,041}$	$\sum_{i=1}^{n=8}(y_i - \bar{y})^2 = \mathbf{92{,}05}$

Далее определим коэффициент линейной корреляции для выборки:

$$r_{\text{Пирсон}} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{-1\,288{,}1}{\sqrt{30\,041}\sqrt{92{,}05}} = \frac{-1\,288{,}1}{1\,662{,}9} \approx -0{,}77.$$

Выборочная оценка коэффициента линейной корреляции указывает на наличие ***обратной*** зависимости между уровнем безработицы и численностью зрителей театров, которая выражена ***весьма высоко***.

Проверим ***статистическую значимость*** коэффициента корреляции. Приступим к проверке нулевой гипотезы: H_0: «$\rho = 0$». Альтернативная гипотеза является направленной H_1: «$\rho < 0$». Значит, критическая область будет левосторонней. Уровень значимости $\alpha = 0{,}05$ (по условию задачи).

Выпишем статистический *t*-критерий, применяемый для проверки гипотезы о равенстве нулю коэффициента корреляции: $t = r\sqrt{\frac{n-2}{1-r^2}}$.

Найдём наблюдаемое значение статистики критерия:

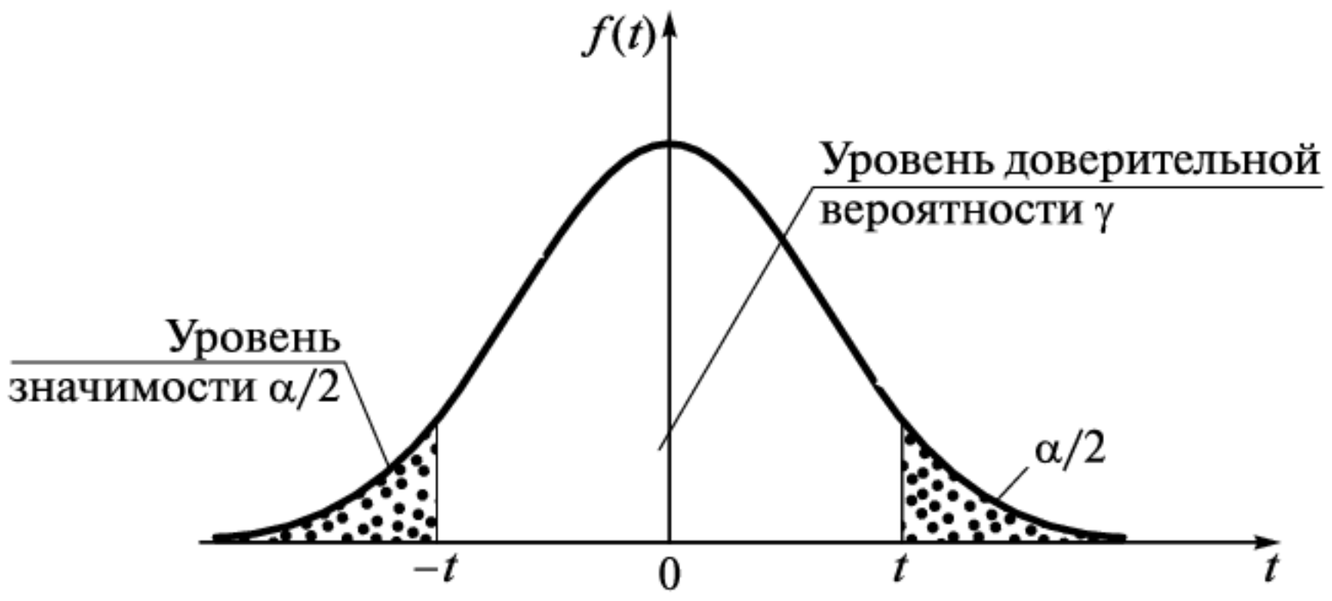
$$t_{\text{наблюдаемое}} = r\sqrt{\frac{n-2}{1-r^2}} = -0{,}77\sqrt{\frac{8-2}{1-(-0{,}77)^2}} =$$

$$= -0{,}77\sqrt{\frac{6}{1-0{,}59}} = -0{,}77 \cdot 3{,}83 = -2{,}95.$$

Полученное наблюдаемое значение критерия нужно сравнить с его табличным (критическим) значением. Найдём его по таблице *t*-распределения Стьюдента (табл. 6.14). Снова (как и в ряде других задач, рассмотренных ранее) обращаем ваше внимание на то, как следует найти критическое значение. В используемой нами таблице приведены значения критических точек для двусторонней области: когда альтернативная гипотеза *не*направленная и уровень значимости α делится пополам ($\alpha/2$), т. е. распределяется поровну между левым и правым «хвостами» распределения. Поэтому если в задаче требуется рассмотреть одностороннюю критическую область, то вероятность *в одном «хвосте»* распределения должна составить α (*целое*, *без деления на 2*). Так, в нашем примере уровень значимости $\alpha = 0{,}05$ (1 – 0,95). Поэтому искомую критическую точку следует посмотреть именно в том столбце таблицы, где отмечено, что площадь под кривой *t*-распределения в одном «хвосте» составляет 0,05 (а не 0,025). В таблице этот столбец выделен серым цветом. Собственно, это и будет означать, что половина от 10 % (т. е. 5 %) и приходится на один «хвост» распределения. Получим: $t_{\text{критическое}} = -1{,}943$.

Отметим наблюдаемое и критическое значения на схематичном графике плотности распределения вероятностей рассматриваемой статистики (рис. 6.20).

Таблица 6.14. **Фрагмент таблицы для *t*-распределения Стьюдента (двустороннее значение вероятности)**



df	Доверительная вероятность и уровень α				
	γ = 0,9 (90 %)	γ = 0,95 (95 %)	γ = 0,98 (98 %)	γ = 0,99 (99 %)	γ = 0,999 (99,9 %)
	α = 0,1 (т. е. α/2 = 0,05)	α = 0,05 (т. е. α/2 = 0,025)	α = 0,02 (т. е. α/2 = 0,01)	α = 0,01 (т. е. α/2 = 0,005)	α = 0,001 (т. е. α/2 = 0,0005)
1	6,314	12,706	31,821	63,657	636,619
…					
5	2,015	2,571	3,365	4,032	6,869
6	**1,943**	2,447	3,143	3,707	5,959
…	и т. д.				

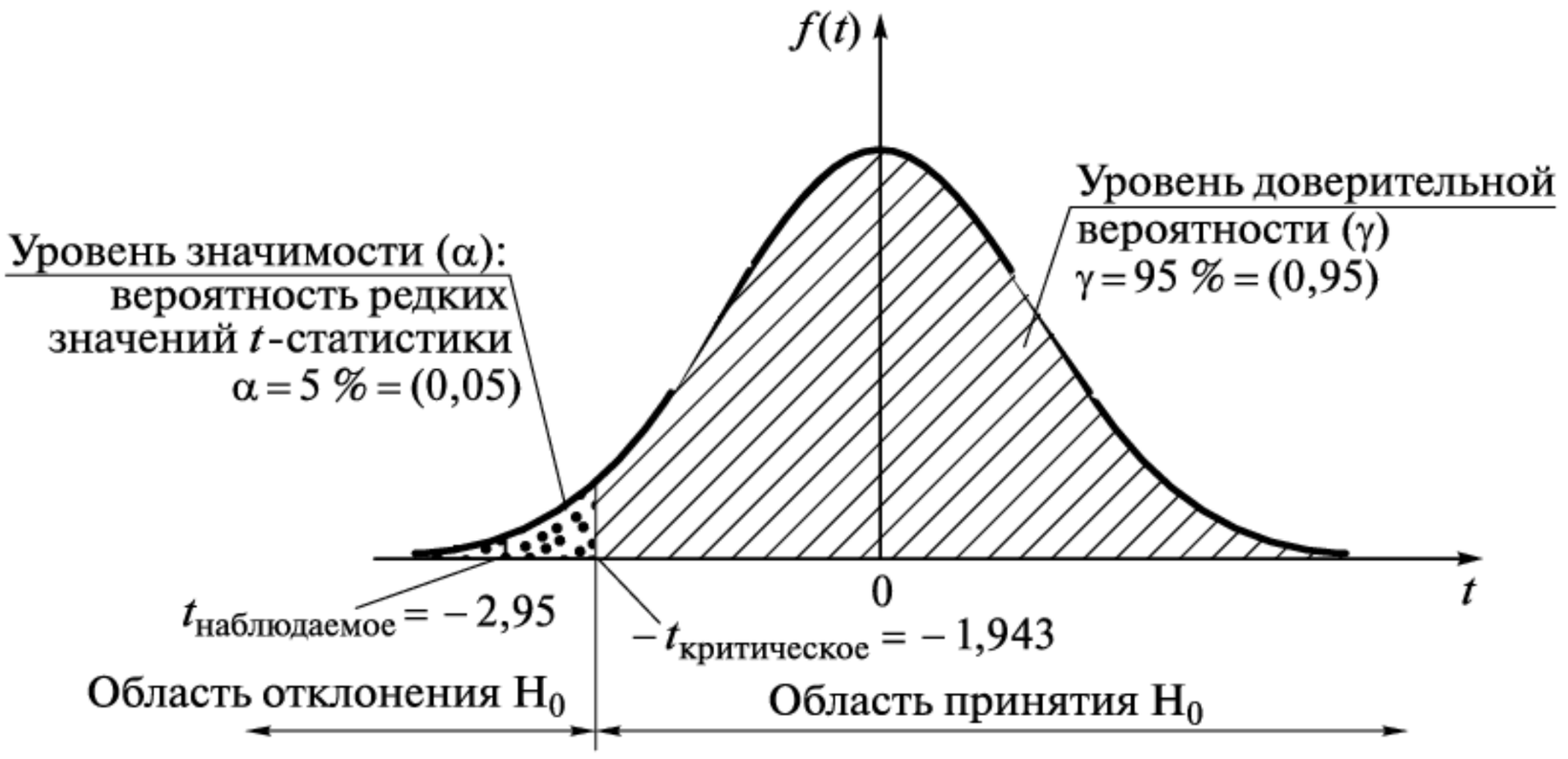


Рис. 6.20. Построение односторонней (левосторонней) критической области (см. пример 6.14 «Зрители театров и безработица в РФ»)

Сопоставляем два значения: $t_{\text{наблюдаемое}} > t_{\text{критическое}}$ (–2,95 > –1,943). Против нулевой гипотезы свидетельствуют бóльшие по модулю отрицательные значения статистики критерия. Поэтому есть все очевидные основания отклонить нулевую гипотезу о равенстве нулю коэффициента корреляции. На уровне значимости $\alpha = 5\ \%$ H_0 не принимается (т. е. отвергается). И принимается альтернативная гипотеза H_1: «$\rho < 0$» (истинное значение коэффициента корреляции меньше нуля). Значит, выборочная оценка коэффициента корреляции *статистически значима. Выборка не даёт оснований сомневаться в наличии линейной связи между переменными X и Y.* Итак, сформулируем содержательный вывод найденного решения: При имеющихся данных выборочного исследования можно считать (обобщая результаты исследования на генеральную совокупность), что такие социальные показатели, как уровень безработицы и численность зрителей театров являются коррелированными (связанными): между ними существует обратная линейная зависимость, которая выражена сильно.

Статистическая значимость коэффициента корреляции Спирмена. Статистический z-критерий, применяемый для проверки гипотезы о равенстве нулю коэффициента корреляции Спирмена, выглядит следующим образом: $z \approx r\sqrt{n-1}$, где r — выборочная оценка коэффициента ранговой корреляции Спирмена (методологию и технику вычисления коэффициента ранговой корреляции см. в гл. 3). Величина $r\sqrt{n-1}$ *асимптотически* имеет нормальное распределение. Принципы проверки гипотезы о статистической незначимости коэффициента ранговой корреляции полностью аналогичны принципам проверки гипотезы о статистической незначимости коэффициента линейной корреляции.

Пример 6.15. «Окружающая среда». Рассмотрим некоторые данные официальных статистических публикаций международных организаций и национальных статистических служб зарубежных стран Европы — «Выбросы в атмосферу загрязняющих веществ»[1]. В таблице представлена выборка 14 стран. Приведены сведения о выбросах диоксида серы и оксидов азота (единица измерения показателей — килограммы на душу населения) в 2006 г.

Страна Европы	Выбросы (на душу населения, кг)	
	диоксида серы (X)	оксидов азота (Y)
Австрия	3,4	27,2
Бельгия	10,7	22,0
Германия	6,8	16,9

[1] Подробно с материалами Федеральной службы государственной статистики «Россия и страны мира» можно ознакомиться по ссылке [http://www.gks.ru/bgd/regl/b10_39/Main.htm] (проверено 02.08.2012).

Окончание табл.

Страна Европы	Выбросы (на душу населения, кг)	
	диоксида серы (X)	оксидов азота (Y)
Греция	48,3	28,5
Дания	4,6	35,2
Ирландия	14,0	27,9
Испания	26,8	33,6
Италия	6,6	18,0
Норвегия	4,5	40,8
Португалия	18,0	23,6
Великобритания	11,2	26,3
Финляндия	16,0	36,7
Франция	8,4	22,7
Швеция	4,3	19,2

Можно ли считать две эти социальные характеристики статистически связанными (коррелированными)? Решите задачу с помощью коэффициента ранговой корреляции Спирмена и проверьте его статистическую значимость на уровне доверия 95 %. Проверку осуществите против ненаправленной альтернативы.

Решение. Напомним, что задачи на нахождение выборочной оценки коэффициента ранговой корреляции решаются в несколько этапов (шагов): Каждому значению признака следует противопоставить порядковый номер (ранг) — по убыванию или возрастанию[1]; рассчитать для каждой пары значений свою разность двух рангов — по переменной X и переменной Y; каждую вычисленную разность рангов возвести в квадрат, чтобы «избавиться» от знака — плюса или минуса — и найти сумму этих квадратов; рассчитать выборочную оценку коэффициента корреляции рангов по указанной формуле и сделать содержательный вывод о полученном значении. Далее необходимо проверить *нулевую гипотезу о равенстве нулю коэффициента корреляции*. По итогам статистической проверки подвести итог задачи, сделав содержательный вывод.

Последовательно реализуем этот алгоритм вычислений.

[1] Важно помнить, что характер упорядочивания (т. е. принцип, согласно которому исходным значениям двух признаков присваивается наименьший ранг или наибольший) в обоих случаях должен быть ***одинаков***.

Сначала проранжируем страны по двум показателям — выбросам диоксида серы и выбросам оксидов азота. Ранг «1» будет присваивать стране, чей показатель наибольший, ранг «14» — чей показатель является наименьшим в выборке. Получим таблицу:

Страна Европы	Выбросы диоксида серы (X)	Ранг по X	Выбросы оксидов азота (Y)	Ранг по Y
Австрия	3,4	**14**	27,2	**7**
Бельгия	10,7	**7**	22,0	**11**
Германия	6,8	**9**	16,9	**14**
Греция	48,3	**1**	28,5	**5**
Дания	4,6	**11**	35,2	**3**
Ирландия	14,0	**5**	27,9	**6**
Испания	26,8	**2**	33,6	**4**
Италия	6,6	**10**	18,0	**13**
Норвегия	4,5	**12**	40,8	**1**
Португалия	18,0	**3**	23,6	**9**
Великобритания	11,2	**6**	26,3	**8**
Финляндия	16,0	**4**	36,7	**2**
Франция	8,4	**8**	22,7	**10**
Швеция	4,3	**13**	19,2	**12**

Теперь рассчитаем для каждой пары значений разность двух рангов — по переменной X и переменной Y, и найдём сумму квадратов разности рангов. Получим d_i (разности рангов) и d_i^2 (квадраты разности рангов):

Страна Европы	Ранг по X	Ранг по Y	Разность рангов	Квадрат разности рангов
Австрия	14	7	**14 – 7 = 7**	49
Бельгия	7	11	**7 – 11 = –4**	16
Германия	9	14	**9 – 14 = –5**	25
Греция	1	5	**1 – 5 = –4**	16

273

Окончание табл.

Страна Европы	Ранг по X	Ранг по Y	Разность рангов	Квадрат разности рангов
Дания	11	3	**11 – 3 = 8**	64
Ирландия	5	6	**5 – 6 = –1**	1
Испания	2	4	**2 – 4 = –2**	4
Италия	10	13	**10 – 13 = –3**	9
Норвегия	12	1	**12 – 1 = 11**	121
Португалия	3	9	**3 – 9 = –6**	36
Великобритания	6	8	**6 – 8 = –2**	4
Финляндия	4	2	**4 – 2 = 2**	4
Франция	8	10	**8 – 10 = –2**	4
Швеция	13	12	**13 – 12 = 1**	1
				Всего = 354

Теперь можем подставить значение «354» в формулу для вычисления коэффициента ранговой корреляции и получим ответ:

$$r_{\text{Спирмен}} = 1 - \frac{6 \cdot 354}{14 \cdot (14^2 - 1)} = 1 - \frac{2\,124}{2\,730} = 1 - 0,78 = 0,22.$$

На основе выборочных данных видим, что согласованность рангов, полученных для двух показателей, характеризующих загрязнение окружающей среды (диоксида серы и оксидов азота), ***слабая***. Можно ли сказать, что положительная связь, хоть и слабая, но всё-таки действительно в реальности имеется? Чтобы ответить на этот вопрос, проверим статистическую значимость коэффициента ранговой корреляции. Приступим к проверке нулевой гипотезы: H_0: «$\rho = 0$». Альтернативная гипотеза ***не***направленная H_1: «$\rho \neq 0$». Критическая область будет двусторонней. Уровень значимости составит $\alpha/2 = 0,05/2 = 0,025$.

Выпишем статистический z-критерий, применяемый для проверки гипотезы о равенстве нулю коэффициента ранговой корреляции: $z \approx r\sqrt{n-1}$.

Найдём наблюдаемое значение статистики критерия:

$$z_{\text{наблюдаемое}} \approx r\sqrt{n-1} = 0,22 \cdot \sqrt{14-1} = 0,22 \cdot 3,61 = 0,79.$$

Таблица 6.15. **Фрагмент таблицы стандартного нормального *z*-распределения (интеграл от 0 до *z*)**

Z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
...										
1,7	0,4554	0,4564	0,4573	0,4582	0,4591	0,4599	0,4608	0,4616	0,4625	0,4633
1,8	0,4641	0,4649	0,4656	0,4664	0,4671	0,4678	0,4686	0,4693	0,4699	0,4706
1,9	0,4713	0,4719	0,4726	0,4732	0,4738	0,4744	0,4750	0,4756	0,4761	0,4767
2,0	0,4772	0,4778	0,4783	0,4788	0,4793	0,4798	0,4803	0,4808	0,4812	0,4817
2,1	0,4821	0,4826	0,4830	0,4834	0,4838	0,4842	0,4846	0,4850	0,4854	0,4857
...	и т.д.									

Полученное наблюдаемое значение статистики критерия следует сравнить с табличным (критическим) значением. Найдём его по таблице стандартного нормального z-распределения: $z_{\text{критическое}} = 1{,}96$ (табл. 6.15). Укажем наблюдаемое и критическое значения на схематичном графике плотности распределения вероятностей рассматриваемой статистики (рис. 6.21).

Сопоставляем два значения: $z_{\text{наблюдаемое}} < z_{\text{критическое}} (+0{,}79 < +1{,}96)$ (наблюдаемое значение статистики критерия не попадает в критическую область), проверяемую гипотезу о равенстве нулю коэффициента корреляции следует принять. На уровне доверительной вероятности $\gamma = 95\ \%$ при имеющихся данных H_0 считается истинной, нет оснований её отклонять. Следовательно, выборочная оценка коэф-

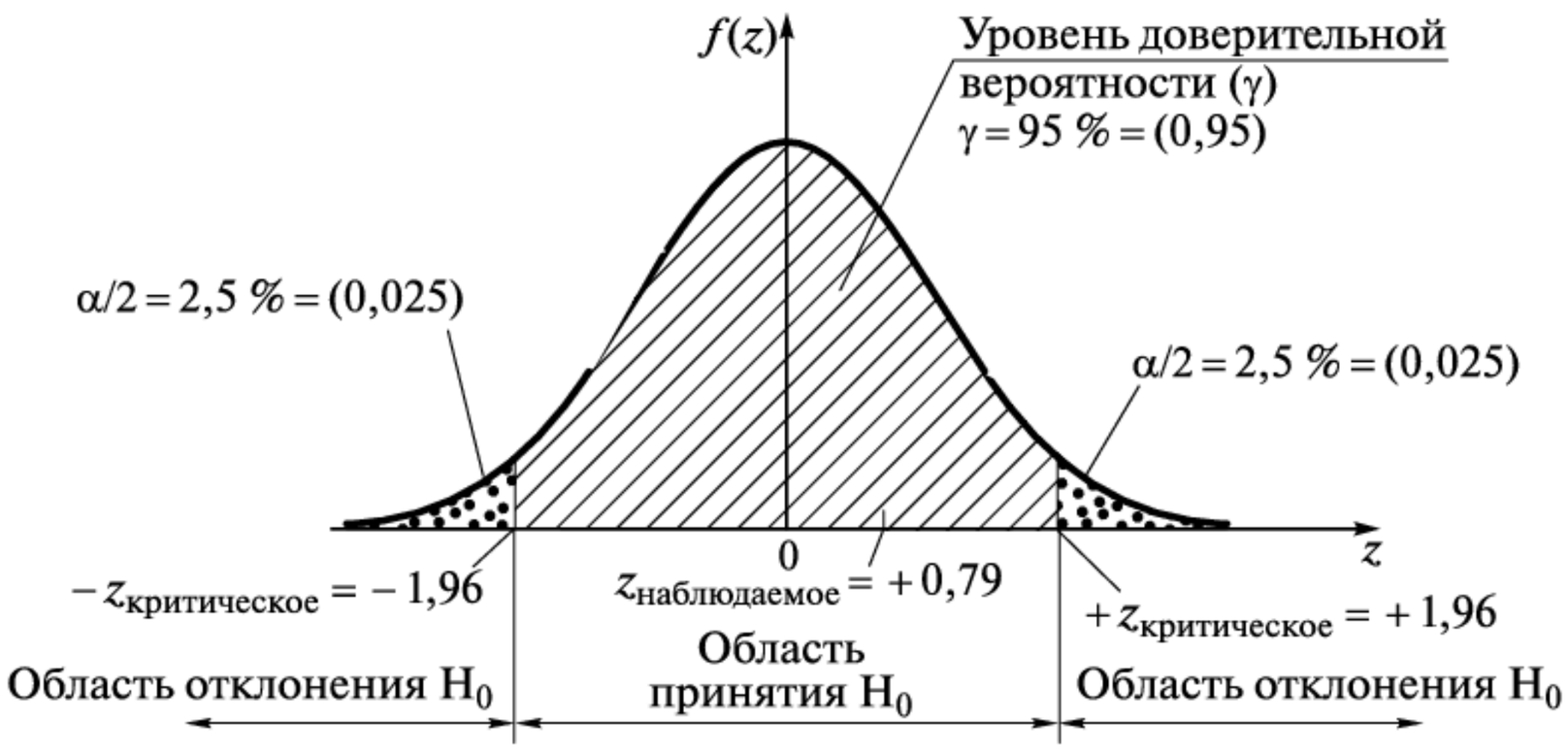


Рис. 6.21. Построение двусторонней критической области (см. пример 6.15 «Окружающая среда»)

фициента ранговой корреляции, равная 0,22, *статистически незначима*. Выборка *заставляет сомневаться* в наличии связи, хотя и слабой, между рассматриваемыми признаками. Итак, сформулируем содержательный вывод найденного решения: при имеющихся данных выборочного исследования нельзя считать (обобщая результаты исследования на генеральную совокупность), что такие социальные показатели, как выбросы в атмосферу диоксида серы и оксидов азота, взаимосвязаны: между ними не существует монотонной зависимости, связь совершенно не выражена.

ЗАДАЧИ ДЛЯ САМОСТОЯТЕЛЬНОГО РЕШЕНИЯ

Задача 1. «Безбилетный проезд». Во всероссийском исследовании ценностей «Молодёжь новой России» двум категориям граждан (молодому и старшему поколениям) был адресован вопрос: «Как, на Ваш взгляд, следует относиться к такому поступку, как *безбилетный проезд в общественном транспорте*?»[1]. Объёмы выборок соответствующие: $n_1 = 1796$ респондентов (молодёжь), $n_2 = 655$ респондентов (старшее поколение):

Безбилетный проезд в общественном транспорте	Молодёжь	Старшее поколение
«Никогда не может быть оправдан»	280	148
Всего, человек	1796	655

Проверьте гипотезу о равенстве долей в ответах молодёжи и старшего поколения, считающих этот поступок *крайне недопустимым* (выбор категории ответа «Никогда не может быть оправдан»). Проверку гипотезы осуществите против альтернативы «Доля среди старшего поколения больше, чем среди молодёжи». Уровень доверительной вероятности примите равным 0,98.

Задача 2. «Швейная фабрика». Руководитель швейной фабрики, работающей в две смены, решил выяснить, существуют ли различия в производительности труда швей дневной и ночной смены. Случайно взятая выборка, состоящая из 10 человек дневной смены, показала, что средний выпуск продукции составил 7 костюмов на работника, а выборочная дисперсия 4,2 (шт.)^2. Вторая выборка, состоящая из 13 швей ночной смены, позволила зафиксировать, что средний выпуск изделий равнялся 6 при дисперсии 6,3 (шт.)^2. На уровне доверительной вероятности 95% определите, существуют ли различия в производительности труда швей дневной и ночной смены. Проверку осуществите против альтернативы: «Производительность работников дневной смены выше, чем ночной».

Задача 3. «Учиться или работать?». На семинаре по математической статистике в 2011 г. среди студентов II курса был проведен блиц-опрос.

[1] Данные Единого архива экономических и социологических данных [http://sophist.hse.ru/db/oprosy.shtml?ts=121&en=0] (проверено 02.08.2012).

В частности, один вопрос был таким: «Есть ли у Вас в настоящее время работа/трудовая занятость помимо учёбы? Или такой занятости нет?». Распределение ответов среди юношей и девушек выглядит следующим образом (в ячейках указано количество респондентов):

Есть ли у Вас в настоящее время работа?	Да, работа есть	Нет, работы нет
Юноши	9	12
Девушки	18	49

Можно ли считать два признака статистически связанными (уровень доверительной вероятности примите равным 0,95)? Задачу решить с помощью критерия хи-квадрат.

Задача 4. «Математика». Предположим, что в двух профильных школах города по математике Министерство образования оценивало знания учащихся выпускных классов. Школьникам было предложено письменное задание, успешность выполнения которого оценивалась по 100-балльной шкале. В случайной выборке из 32 выпускников первой школы средний балл составил 75, при стандартном отклонении 5 баллов. Для случайной выборки выпускников второй школы (объём выборки составил 35 человек) средний балл оказался равным 68 при стандартном отклонении 9 баллов. Можно ли считать, что обе профильные школы одинаково эффективно обучают учащихся? Уровень доверительной вероятности выводов примите равным 95 %. Проверку соответствующей гипотезы осуществите против ненаправленной альтернативы.

Задача 5. «Без названия-1». В исследовании две выборки наблюдений. Вот их характеристики:

Выборочная оценка дисперсий	Объёмы выборок
Дисперсия$_1$ = 9	$n_1 = 9$
Дисперсия$_2$ = 5	$n_2 = 11$

Проверьте гипотезу о равенстве дисперсий против альтернативы: «Дисперсия первой совокупности больше дисперсии второй совокупности». Уровень доверительной вероятности примите равным 0,95.

Задача 6. «Без названия-2». В исследовании оценка коэффициента корреляции Пирсона была вычислена по 26 парам наблюдений и составила $r = 0{,}32$. Проверьте нулевую гипотезу о равенстве коэффициента корреляции нулю (т. е. его статистической незначимости) для доверительной вероятности 0,95 против альтернативы «коэффициент корреляции не равен нулю».

Задача 7. «Профессия социолога». На одном из семинаров по математической статистике в 2012 г. среди студентов II курса было проведено мини-анкетирование. В частности, один вопрос был таким: «Было ли в Ваших планах поступить на социологический факультет после окончания школы

(т. е. целенаправленно ли Вы поступали на социологический факультет университета или «по воле случая»?)». Распределение ответов среди юношей и девушек, проживающих в Москве и приехавших из других городов России, выглядит так (в ячейках — количество респондентов):

Целенаправленно ли Вы поступали на социологический факультет университета?	Юноши и девушки	
	проживающие в Москве	приехавшие из других городов
Да, целенаправленно	20	9
Нет, «по воле случая»	20	19

Можно ли считать два признака статистически связанными (уровень доверительной вероятности примите равным 0,95)? Задачу решите двумя способами — с помощью критерия хи-квадрат и с помощью проверки статистической гипотезы о равенстве долей.

Задача 8. «Без названия-3». Пусть нужно проверить гипотезу о равенстве долей для двух выборок против ненаправленной альтернативной гипотезы. Уровень доверительной вероятности равен 0,99. Допустим, что в ходе анализа было получено наблюдаемое значение $z_{\text{наблюдаемое}} = 2{,}3$. Что нужно сделать с нулевой гипотезой — принять или отклонить? Ответ обоснуйте.

Задача 9. «Законопослушность». Во всероссийском исследовании ценностей «Молодёжь новой России», проведённом в 2007 г. Центром социального прогнозирования под руководством Ф. Э. Шереги, двум категориям граждан (молодому и старшему поколениям) был адресован такой вопрос: «*Как, на Ваш взгляд, надо относиться к закону?*». Объёмы выборок соответствующие: 1796 человек (молодёжь), 655 человек (старшее поколение). Категорию ответа «Всегда и во всём следует соблюдать букву закона, даже если закон устарел и не вполне соответствует сегодняшним реальностям» среди молодёжи выбрали 288 человек, а среди старшего поколения — 143. Можно ли считать, что доли людей, разделяющих такое утверждение, одинаковы у представителей молодого и старшего поколений? Уровень доверительной вероятности примите равным 0,99. Проверку нулевой гипотезы осуществите против ненаправленной альтернативы.

Задача 10. «Контрольная работа». Студенты двух факультетов — А и Б — писали итоговую контрольную работу по курсу «Математическая статистика». Результат оценивался по 30-балльной шкале (где минимум — 0 баллов, максимум — 30 баллов). Преподаватель поставил себе задачу выяснить, знания учащихся какого факультета — А или Б — более однородны. Были получены следующие результаты (см. данные двух случайных выборок):

Факультет	Результат (оценка, баллов):
А	26, 19, 28, 20, 15, 17, 22, 30, 29, 24, 16
Б	25, 23, 29, 19, 21, 18, 17, 18, 26, 14

Можно ли сказать, что знания учащихся факультета Б более однородны, чем знания студентов факультета А? Проверку соответствующей гипотезы осуществите на уровне доверия 95 % против направленной альтернативы.

Задача 11. «Мерчендайзинг». В маркетинге мерчендайзингом называется искусство (или культура) презентации товара, комфорт и эстетика его подачи для покупателя. Пусть компания, продающая коллекцию одежды в сети фирменных магазинов, наблюдала невысокие (в сравнении с её ожиданиями) финансовые результаты продаж в течение последней недели. Было принято решение скорректировать презентацию коллекции в торговых залах магазинов. В конце следующей недели в очередной раз были зафиксированы коммерческие показатели работы магазинов — *объём продаж*, *приходящийся на 1 м² торговой площади магазина* (*усл. ед.*). Получены следующие данные по 10 магазинам компании:

Финансовый результат	Магазин									
	1	2	3	4	5	6	7	8	9	10
До проведения коррекции в презентации товара	140	160	180	200	210	190	210	250	150	190
После проведения коррекции в презентации товара	270	170	170	200	190	230	220	250	140	220

Считая выборки связными (парными), проверьте гипотезу о равенстве среднего эффекта воздействия нулю (т.е. отсутствии различий в финансовых результатах компании до и после введения новой концепции мерчендайзинга).

При проверке уровень доверительной вероятности примите равным 0,95. Альтернативная гипотеза направленная: «Средний показатель финансового результата стал лучше (выше)». Имеется ли эффект воздействия? Оказались ли принятые меры эффективными (принесли ли положительный результат)?

Задача 12. «Магистратура». В таблице представлен фрагмент данных опроса выпускников бакалавриата. Выпускникам, работающим в российских и зарубежных компаниях, социологами был задан вопрос: «Планируете ли Вы продолжить обучение в магистратуре?». В ячейках — количество респондентов. Можно ли считать два признака статистически связанными? Уровень доверительной вероятности примите равным 0,95. Задачу решите с помощью критерия хи-квадрат.

Планируете ли Вы продолжить обучение в магистратуре?	Место работы респондента	
	в российской компании	в иностранной компании
Да, планирую	42	13
Нет, не планирую	128	37

Задача 13. «Уклонение от уплаты налогов». Во всероссийском исследовании ценностей «Молодёжь новой России», проведённом в 2007 г. Центром социального прогнозирования под руководством Ф. Э. Шереги, двум категориям граждан (молодому и старшему поколениям — людям 17 — 26 лет и 40 — 60 лет) был адресован вопрос: «Как, на Ваш взгляд, следует относиться к такому поступку, как уклонение от уплаты налогов?»[1]. Объёмы выборок соответствующие: 1796 человек (молодёжь), 655 человек (старшее поколение).

Уклонение от уплаты налогов	Молодёжь, 17 — 26 лет	Старшее поколение, 40 — 60 лет
Никогда не может быть оправдано	713	333
Иногда это допустимо	831	258
К этому следует относиться снисходительно	248	63
Затрудняюсь ответить	4	1

Проверьте гипотезу о равенстве долей в ответах молодёжи и старшего поколения, считающих, что «К этому следует относиться снисходительно». Уровень доверительной вероятности примите равным 0,99. Ответы «Затрудняюсь ответить» исключите из анализа.

Задача 14. «Игорные заведения». В 2006 и 2009 гг. ВЦИОМ провёл опрос населения на тему, в какой мере россияне увлекаются азартными играми и что они думают о выведении игорных заведений за пределы городов и посёлков. Вот небольшой фрагмент данных динамики, полученных по одному из вопросов этого исследования: «В целом, на Ваш взгляд, выведение игорных заведений за пределы городов и посёлков в четыре специально созданные зоны, принесёт больше пользы или вреда? (закрытый вопрос, один ответ)»:

Ответ	Данные	
	2006 г.	2009 г.
Это принесёт больше пользы, чем вреда	1 040	1 120
Это принесёт больше вреда, чем пользы	304	224
Затрудняюсь ответить	256	256
Объёмы выборки, человек	1 600	1 600

[1] Данные Единого архива экономических и социологических данных.

Можно ли считать, что доля граждан, считающих, что это принесет больше вреда, чем пользы, в 2009 г. действительно стала меньше? Ответы «Затрудняюсь ответить» исключите из рассмотрения. Доверительную вероятность примите равной 0,95.

Задача 15. «Служба в армии». Во всероссийском исследовании ценностей «Молодёжь новой России», проведённом в 2007 г. Центром социального прогнозирования под руководством Ф. Э. Шереги, двум категориям граждан (молодому и старшему поколениям — людям 17 — 26 лет и 40 — 60 лет) был адресован вопрос: «Как, на Ваш взгляд, следует относиться к такому поступку, как уклонение от службы в армии?»[1]. Объёмы выборок соответствующие: 1796 человек (молодёжь), 655 человек (старшее поколение):

Уклонение от службы в армии	Молодёжь, 17 — 26 лет	Старшее поколение, 40 — 60 лет
Никогда не может быть оправдано	438	204
Иногда это допустимо	943	348
К этому следует относиться снисходительно	413	102
Затрудняюсь ответить	2	1

Проверьте гипотезу о равенстве долей в ответах молодёжи и старшего поколения, считающих, что уклонение от службы в армии — приемлемый поступок (т. е. категории ответов «Иногда это допустимо» + «К этому следует относиться снисходительно»). Уровень доверительной вероятности примите равным 0,95. Ответы «Затрудняюсь ответить» исключите из анализа.

Задача 16. «Физика и химия». Пусть имеется следующая информация (выборочная) относительно оценок, полученных школьниками на выпускных экзаменах ЕГЭ (100-балльная система) по двум предметам:

Предмет	Школьник									
	1	2	3	4	5	6	7	8	9	10
Физика	56	46	47	93	81	49	65	51	74	60
Химия	64	41	52	87	71	57	67	52	70	70

Справедливо ли утверждение о том, что, в среднем, уровень подготовки выпускника по химии не отличается от уровня подготовки по физике? Задачу решите, используя критерий для связных (парных) выборок. Уровень доверительной вероятности примите равным 0,95.

[1] Данные Единого архива экономических и социологических данных.

Задача 17. «Лекарственное средство». Предположим, что врачи-исследователи утверждают, что эффективность действия нового разрабатываемого лекарственного препарата связана с тем, из какого вещества сделана оболочка капсулы — А, Б или В. Проверьте это утверждение по следующим данным, имеющимся в результате тестирования этого лекарства среди пациентов, принявших добровольное участие в медицинском исследовательском проекте (на пересечении строк и столбцов указано количество человек):

Форма выпуска лекарственного препарата	Эффективность действия препарата	
	Высокая	Низкая
Капсула, покрытая оболочкой А	13	23
То же, оболочкой Б	20	26
То же, оболочкой В	19	22

Можно ли считать два признака статистически связанными (уровень доверительной вероятности примите равным 0,99)? Задачу решите с помощью критерия хи-квадрат.

Задача 18. «Диета». Доктор рекомендует своим пациентам, находящимся на лечении и считающих, что они имеют избыточную массу тела, придерживаться специальной молочной диеты для коррекции фигуры. При этом он фиксирует начальную массу восьми пациентов и массу спустя некоторое время (после завершения курса). Данные таковы:

Масса, кг	Пациент							
	1	2	3	4	5	6	7	8
До диеты	76	71	57	95	69	69	78	82
После диеты	81	85	52	95	63	61	70	81

Диетолог не помнит математической статистики. Пытается всмотреться в цифры. Помогите ему сделать вывод об эффективности молочной диеты. Уменьшает ли такая методика избыточную массу? Уровень доверительной вероятности примите равным 95 %. Конкурирующую гипотезу считайте направленной.

Задача 19. «Проявление неприязни». Во всероссийском исследовании ценностей «Молодёжь новой России», проведённом в 2007 г. Центром социального прогнозирования под руководством Ф. Э. Шереги, двум категориям граждан (молодому и старшему поколениям — людям 17 — 26 лет и 40 — 60 лет) был адресован вопрос: «Как, на Ваш взгляд, следует относиться к такому поступку, как публичное проявление неприязни к представителям других национальностей?» Объёмы выборок соответствующие: 1796 человек (молодёжь), 655 человек (старшее поколение).

Публичное проявление неприязни	Молодёжь, 17 — 26 лет	Старшее поколение, 40 — 60 лет
Никогда не может быть оправдано	1 113	505
Иногда это допустимо	523	109
К этому следует относиться снисходительно	156	41
Затрудняюсь ответить	4	0

Проверьте гипотезу о равенстве долей в ответах молодёжи и старшего поколения, считающих проявление неприязни к представителям других национальностей приемлемым поступком (т. е. категории ответов «Иногда это допустимо» + «К этому следует относиться снисходительно»). Уровень доверительной вероятности примите равным 0,90. Ответы «Затрудняюсь ответить» исключите из рассмотрения.

Задача 20. «Досуг и развлечения». На одном из семинаров по математической статистике в 2013 г. в ходе мини-анкетирования студентам II курса был адресован вопрос: «Какую сумму денег Вы обычно тратите ежемесячно на досуг и развлечения? Пожалуйста, укажите в рублях». Было подсчитано, что среди юношей средняя сумма расходов составляет 4 120 руб., стандартное отклонение равно 2 570 руб. Среди девушек средняя величина бюджета, выделяемого на досуг и развлечения, составляет 3 210 руб., стандартное отклонение равно 3 000 руб. Объёмы выборок соответственно равны: 13 и 50 человек. На уровне доверительной вероятности 95 % определите, существуют ли статистически значимые различия в средних ежемесячных расходах на досуг и развлечения у юношей и девушек. Проверку нулевой гипотезы осуществите против альтернативы: «Юноши тратят на досуг и развлечения в среднем больше денежных средств, чем девушки».

Задача 21. «Фрустрация». Предположим, что для сравнения показатели уровня фрустрации (проявляющейся в таких эмоциональных состояниях, как тревога, раздражение, уныние, разочарование, отчаяние и др.) до и после психологической беседы с девятью подростками из детского дома были измерены дважды с помощью специальной методики (в баллах), при этом больший балл соответствует большему уровню фрустрации. Результаты представлены ниже. Проверьте, имеется ли положительный результат от этой психологической беседы или она оказалась безрезультативной?

Фрустрация, баллы	Подросток								
	1	2	3	4	5	6	7	8	9
До психологической беседы	31	35	38	34	39	40	27	27	32
После психологической беседы	33	35	35	30	36	36	38	29	40

Уровень доверительной вероятности примите равным 90 %. Конкурирующую гипотезу считайте направленной (исходя из данных).

Задача 22. «Автомобили». Предположим, что два автомобильных салона, расположенных в двух разных городах, продают машины одного и того же класса. В течение определённого времени в первом салоне зафиксировано 47 сделок (из них 20 — количество проданных автомобилей марки А*), во втором салоне зафиксировано 59 сделок (из них 29 — количество проданных автомобилей марки А*). Итак, выяснили, что покупки распределились следующим образом:

Показатели	Салон	
	№ 1	№ 2
Продано автомобилей марки А*	20	29
Всего сделок, шт.	47	59

Свидетельствует ли эта ситуация о том, что автомобили марки А* пользуются одинаковой популярностью у потребителей в первом и втором салонах? Проверку соответствующей гипотезы осуществите на уровне надёжности 95 % против ненаправленной альтернативы.

Задача 23. «Макароны». Социологическое выборочное исследование, проведённое менеджером университетского кафе, показало, что из 40 опрошенных студентов 17 любят макароны, предлагаемые в меню. Случайная же выборка 30 преподавателей показала, что макароны, предлагаемые в меню, нравятся 11 человекам. Используя 95%-ю доверительную вероятность, определите, одинаковы ли предпочтения студентов и преподавателей в этом кулинарном аспекте. Проверку соответствующей статистической гипотезы осуществите против ненаправленной альтернативы.

Задача 24. «Текучесть кадров». Совет директоров компании А* уделяет пристальное внимание текучести кадров персонала (продавцов-консультантов), работающих в торговых залах магазинов в первые 6 месяцев их работы в компании. По данным предыдущего года зафиксировано, что из 40 сотрудников, принятых на работу в должности «продавец-консультант», уволились 15 человек, не проработав полгода. Компания поменяла стратегию мотивации персонала (систему вознаграждения, организационную культуру, механизмы вовлечения/интеграции человека в коллектив, поддержания лояльности и лидерства в коллективе и пр.). Так, в конце текущего года было зафиксировано, что из 55 принятых на работу сотрудников уволились 12 человек. Можно ли считать, что корпорация добилась успеха в управлении торговым персоналом, сократив текучесть кадров принятыми действиями? Проверку соответствующей статистической гипотезы осуществите против направленной альтернативы на уровне значимости 5 %.

Задача 25. «Хоккей». Пусть физическая подготовка восьми юных спортсменов была проверена при поступлении в детскую спортивную школу в секцию «Хоккей», а затем спустя месяц тренировок. Итоги аттестации (в баллах) оказались следующими (в первой строке указано число баллов,

полученных каждым спортсменом на вступительных испытаниях в школу, во второй — после месяца обучения и тренировок):

Подготовка, баллы	Хоккеист							
	Митя	Коля	Саша	Андрей	Ваня	Олег	Юра	Слава
До тренировок	69	65	70	59	76	91	57	60
После тренировок	74	53	70	62	79	95	55	62

Установите, улучшилась ли подготовка ребят или осталась без изменений.

Уровень доверительной вероятности примите равным 95 %. Конкурирующую гипотезу считайте направленной.

Задача 26. «Электорат партий о судах и судьях». В период Еженедельного опроса «ФОМнибус»[1] от 23 — 24 июня 2012 г., проведённого в 43 субъектах РФ (в 100 населенных пунктах) жителям России был задан вопрос: «Как Вы в целом оцениваете деятельность российских судов и судей — положительно или отрицательно?». Распределение ответов среди электората выглядит следующим образом:

Как Вы в целом оцениваете деятельность российских судов и судей — положительно или отрицательно?	Электораты партий			
	«Единая Россия»	КПРФ	ЛДПР	«Справедливая Россия»
Положительно (+)	206	25	38	20
Отрицательно (–)	200	79	62	43
Затрудняюсь ответить (?)	239	61	35	27
Всего, человек	645	165	135	90

Можно ли считать два признака (отношение к деятельности судов и судей, с одной стороны, и принадлежность к определённому электорату — с другой) статистически связанными (уровень доверительной вероятности примите равным 0,99)? Задачу решите с помощью критерия хи-квадрат.

Задача 27. «Отношение к повышению цен на водку». В период Еженедельного опроса «ФОМнибус»[2] от 9 — 10 июня 2012 г., проведённого в 43 субъектах РФ (в 100 населенных пунктах) жителям России был задан вопрос: «С первого июля в России цены на водку будут повышены почти на треть. Вы одобряете или не одобряете повышение цен на водку?». Распределение

[1] Исходные данные опроса ФОМ размещены в разделе «Безопасность и право» по ссылке [http://fom.ru/Bezopasnost-i-pravo/10551] (проверено 01.08.2012).

[2] Исходные данные опроса ФОМ размещены в разделе «Общество» по ссылке [http://fom.ru/obshchestvo/10569] (проверено 01.08.2012).

ответов среди людей с разным уровнем материального благополучия выглядит следующим образом:

Одобряете ли Вы повышение цен на водку?	Ежемесячный доход, руб.			
	≤ 4 000	4 001—9 000	9 001—20 000	> 20 000
Одобряю (+)	86	94	170	70
Не одобряю (—)	80	202	263	67
Затрудняюсь ответить (?)	59	94	152	43
Всего, чел.	225	390	585	180

Можно ли считать два признака (отношение к повышению цен на водку, с одной стороны, и уровень ежемесячного дохода — с другой) статистически связанными (уровень доверительной вероятности примите равным 0,95)? Задачу решите с помощью критерия хи-квадрат.

Россиянам в ходе этого исследования также был предложен вопрос: «Как Вы полагаете, повышение цен на водку способно или не способно снизить уровень пьянства в России?». Распределение ответов среди людей с разным уровнем образования представлено ниже:

Как Вы полагаете, способно ли повышение цен на водку снизить уровень пьянства в России?	Образование		
	Среднее общее и ниже	Среднее специальное	Высшее
Способно (+)	95	92	53
Не способно (–)	454	400	244
Затрудняюсь ответить (?)	81	48	33
Всего, человек	630	540	330

Можно ли считать два признака (оценка перспектив борьбы с пьянством с помощью меры повышения цен на водку, с одной стороны, и уровень образования — с другой) статистически связанными (уровень доверительной вероятности примите равным 0,95)? Задачу решите с помощью критерия хи-квадрат.

Задача 28. «К задачам гл. 3 — К. Пирсон». В задачах № 3—10, предложенных для самостоятельного решения к гл. 3, проверьте статистическую значимость коэффициентов линейной корреляции Пирсона. Проверьте нулевую гипотезу о равенстве коэффициента корреляции нулю (т. е. его статистической незначимости) для доверительной вероятности 0,95 против альтернативы «коэффициент корреляции не равен нулю».

Задача 29. «К задачам гл. 3 — Ч. Спирмен». В задачах № 11 — 13, предложенных для самостоятельного решения к гл. 3, проверьте статистическую значимость коэффициентов ранговой корреляции Спирмена. Проверьте нулевую гипотезу о равенстве коэффициента корреляции нулю (т. е. его статистической незначимости) для доверительной вероятности 0,95 против альтернативы «коэффициент корреляции не равен нулю».

ПРИЛОЖЕНИЯ

Приложение 1

Образцы контрольной экзаменационной работы

Контрольная работа. Вариант № 1

Ф. И. О. _____ № группы _____

Задача 1. (*4 балла*)

В автосервисе проводится опрос клиентов на тему: «Сколько раз в месяц Вы моете у нас свой автомобиль?». Ответы первых 14 водителей представлены ниже:

X (число раз): **2, 1, 14, 2, 4, 1, 4, 8, 5, 2, 1, 1, 4, 10.**

Найдите:

а) медиану выборки;

б) нижнюю и верхнюю квартили выборки (по методу Дж. Тьюки).

в) Определите по методу Дж. Тьюки, есть ли в этой выборке нетипичные (нехарактерные) значения наблюдений — «подозрительные» и/или «выбросы»

Ответ:

Медиана равна _____

Нижняя квартиль равна _____

Верхняя квартиль равна _____

Если есть «подозрительные» наблюдения, укажите их значения: _____

_____

Если есть «выбросы», укажите их значения: _____

_____

Задача 2. (*3 балла*)

Студентам университета был задан вопрос: «Сколько примерно раз в месяц Вы ходите в спортзал?». В опросе участвовало 10 студентов. Выборка представлена следующими ответами:

2, 0, 1, 3, 5, 0, 1, 3, 3, 2.

Вычислите *несмещённую выборочную оценку* дисперсии и стандартного отклонения; стандартную ошибку среднего.

Ответ:

Оценка дисперсии равна _____

Оценка стандартного отклонения равна _____

Стандартная ошибка среднего равна _____

Задача 3. (*4 балла*)

Юношам и девушкам в возрасте 17 — 20 лет и проживающим в России был адресован вопрос: «Употребляете ли Вы в пищу картофельные чипсы?» Вот распределение ответов:

Ответ	Количество респондентов
1. Нет, я не ем картофельные чипсы	460
2. Да, я ем картофельные чипсы	1 520
Всего респондентов	1 980

Постройте *доверительный интервал для доли* граждан в возрасте 17—20 лет, которые едят чипсы (т.е. категория ответа «Да, я ем картофельные чипсы»). Уровень доверительной вероятности примите равным 0,80.

Ответ: _____

Интерпретация: _____

_____

Задача 4. (*5 баллов*)

Два искусствоведа оценивают эстетические характеристики шести картин, представленных на выставке. Бóльшая оценка соответствует лучшим, по их мнению, эстетическим качествам этих произведений:

Оценки	Картина					
	А	Б	В	Г	Д	Е
Искусствовед 1	5	12	8	10	6	15
Искусствовед 2	3	11	8	14	9	12

В какой мере оценки (мнения) двух искусствоведов согласуются между собой?

Вычислите коэффициент ранговой корреляции Спирмена и дайте интерпретацию.

Ответ:

Коэффициент корреляции Спирмена равен _____

Интерпретация полученного значения: _____

_____

Задача 5. (*5 баллов*)

Исследователей интересовало, являются ли взаимосвязанными такие социальные показатели, как уровень безработицы в регионах РФ и процент голосов поддержки, отданных на минувших выборах за партию А*. Выборка представлена данными по шести регионам страны. Попарно приведены эти значения:

Регион РФ	№ 1	№ 2	№ 3	№ 4	№ 5	№ 6
Безработица, %	8	6	2	4	2	2
Процент голосов за партию А*	6	2	4	2	6	10

Рассчитайте коэффициент линейной корреляции Пирсона и дайте содержательную интерпретацию полученного значения.

Ответ:

Коэффициент корреляции Пирсона равен _____

Интерпретация полученного значения: _____

_____

_____

Задача 6. (*6 баллов*)

Компания провела исследование для получения типологии потребителей. Респондентам был задан вопрос: «Каким был доход Вашей семьи в прошлом месяце в расчёте на одного члена семьи (включая детей)?». Вот фрагмент данных этого опроса (данные округлены):

Доход, тыс. усл. ед.	0 — 2	2 — 4	4 — 6	6 — 8	8 — 10
Количество респондентов	10	70	40	20	10

а) Оцените среднедушевой доход в семье с помощью среднего арифметического.

б) Охарактеризуйте рассеяние (разброс) значений с помощью стандартного отклонения.

в) Рассчитайте коэффициент вариации.

г) Постройте 90%-й доверительный интервал для среднего. Интерпретируйте полученный ответ.

Ответ:

Среднее: _____ Стандартное отклонение: _____

Коэффициент вариации: _____

Доверительный интервал: _____

Интерпретация доверительного интервала: _____

_____

_____

_____

Итоги контрольной работы

№ задачи	1	2	3	4	5	6	Всего
Максимум баллов	4	3	4	5	5	6	27
Фактически баллов							

Контрольная работа. Вариант № 2

Ф. И. О._____ № группы _____

Задача 1. (*4 балла*)

Компания, продающая канцелярские товары, изучает, каковы ежемесячные расходы её постоянных клиентов на покупку ручек, блокнотов, карандашей и пр. Выборка, состоящая из 11 элементов, представлена следующими значениями:

X (руб.): **420**, **840**, **710**, **530**, **600**, **2200**, **3700**, **980**, **830**, **700**, **590**.

Найдите:

а) медиану выборки;

б) нижнюю и верхнюю квартили выборки (по методу Дж. Тьюки).

в) Определите по методу Дж. Тьюки, есть ли в этой выборке нетипичные (нехарактерные) значения наблюдений — «подозрительные» и/или «выбросы».

Ответ:

Медиана равна _____

Нижняя квартиль равна _____

Верхняя квартиль равна _____

Если есть «подозрительные» наблюдения, укажите их значения: _____

Если есть «выбросы», укажите их значения: _____

Задача 2. (*3 балла*)

Страховая компания оценивает число ДТП, в которых водители побывали за последний год. В исследовании участвовало 11 водителей. Выборка представлена следующими их ответами о числе ДТП:

0, **1**, **3**, **2**, **3**, **0**, **6**, **3**, **2**, **0**, **2**.

Вычислите *несмещённую выборочную оценку* дисперсии и стандартного отклонения; стандартную ошибку среднего.

Ответ:

Оценка дисперсии равна _____

Оценка стандартного отклонения равна _____

Стандартная ошибка среднего равна _____

Задача 3. (*4 балла*)

Жителям мегаполисов был адресован вопрос: «Используете ли Вы дома бытовые фильтры для очистки водопроводной воды?». Вот распределение ответов:

Ответ	Количество респондентов
1. Нет, не использую	1 410
2. Да, использую	580
Количество респондентов	1 990

Постройте *доверительный интервал для доли* граждан, проживающих в мегаполисах, не использующих бытовые фильтры для очистки воды (т. е. категория ответа «Нет, не использую»). Уровень доверительной вероятности примите равным 0,84.

Ответ: _____
Интерпретация: _____
_____

Задача 4. (*5 баллов*)

На кинофестивале состоялась премьера семи фильмов. Конкурсное жюри оценивает качество их реализации по 10-балльной шкале. Более высокие оценки соответствуют более высокому качеству их экранизации:

Фильм	№ 1	№ 2	№ 3	№ 4	№ 5	№ 6	№ 7
Эксперт 1	7	4	6	9	6	10	8
Эксперт 2	8	5	4	7	8	6	10

В какой мере оценки (мнения) двух экспертов согласуются между собой?

Вычислите коэффициент ранговой корреляции Спирмена и дайте интерпретацию.

Ответ:

Коэффициент корреляции Спирмена равен _____
Интерпретация полученного значения: _____
_____

Задача 5. (*5 баллов*)

Маркетологов интересовало, являются ли взаимосвязанными такие финансовые показатели деятельности фирмы, как расходы на рекламу и объём продаж. Выборка представлена данными за шесть отчётных периодов. Попарно приведены эти значения:

Временной период	t_1	t_2	t_3	t_4	t_5	t_6
Расходы на рекламу, усл. ед.	2	4	2	1	0	3
Объём продаж, усл. ед.	2	2	6	2	4	2

Рассчитайте коэффициент линейной корреляции Пирсона и дайте содержательную интерпретацию полученного значения.

Ответ:

Коэффициент корреляции Пирсона равен _____
Интерпретация полученного значения: _____
_____

Задача 6. (*6 баллов*)

Среди студентов был проведен опрос. Вопрос был таким: «Какую, примерно, сумму денег Вы тратите каждый месяц на покупку одежды для себя?». Распределение ответов среди девушек выглядит так (данные округлены):

Расходы на покупку одежды, тыс. руб.	0 — 2	2 — 4	4 — 6	6 — 8
Количество респондентов	20	20	15	10

а) Рассчитайте, чему равны в среднем ежемесячные расходы на покупку одежды.

б) Охарактеризуйте разброс значений с помощью стандартного отклонения.

в) Рассчитайте коэффициент вариации.

г) Постройте 86%-й доверительный интервал для среднего. Интерпретируйте полученный ответ.

Ответ:

Среднее: _____ Стандартное отклонение: _____

Коэффициент вариации: _____

Доверительный интервал: _____

Интерпретация доверительного интервала: _____

_____

_____

Итоги контрольной работы

№ задачи	1	2	3	4	5	6	Всего
Максимум баллов	4	3	4	5	5	6	27
Фактически баллов							

Образцы итоговой экзаменационной работы

Экзаменационная работа. Вариант № 1

Фамилия, имя _____ № группы _____

Задача 1. (*2 балла*)

Менеджер отдела бытовой техники оценивает, сколько кофеварок продается за день. Он знает, что число кофеварок, продаваемых за день, является случайной величиной с законом распределения:

X, шт.	0	1	2	3
p (вероятность)	0,4	0,3	0,15	?

Найдите:

а) среднее значение случайной величины X;

б) дисперсию случайной величины X;

в) стандартное отклонение случайной величины X,

где X — количество кофеварок, продаваемых за день в отделе бытовой техники.

Ответ:

Среднее = Дисперсия = Стандартное отклонение =

Задача 2. (*4 балла*)

Средняя стоимость автомобиля класса «Премиум» в салонах города составляет 1 000 усл. ед., а дисперсия равна 10 000 (усл. ед.)^2. Какова вероятность того, что в выбранном наугад автосалоне стоимость такого автомобиля составит от 950 до 1 100 усл. ед.? (если считать, что стоимость автомобиля — случайная величина, которая описывается нормальным законом распределения с указанными параметрами). Схематично отразите искомую вероятность на графике нормального распределения и на графике стандартного нормального распределения.

Ответ: _____

Задача 3. (*4 балла*)

В турагентстве опрашивают своих клиентов на тему: «В каком количестве зарубежных стран Вы успели побывать к настоящему моменту своей жизни? Назовите, пожалуйста, число». Ответы первых 11 клиентов представлены ниже:

X (стран): **11, 6, 1, 4, 2, 6, 0, 7, 6, 3, 17.**

Найдите:

а) медиану выборки;

б) нижнюю и верхнюю квартили выборки (по методу Дж. Тьюки).

в) Определите по методу Дж. Тьюки, есть ли в этой выборке нетипичные, нехарактерные значения наблюдений — «подозрительные» и/или «выбросы».

Ответ:

Медиана равна _____ Нижняя квартиль равна _____

Верхняя квартиль равна _____

Если есть «подозрительные» наблюдения, укажите их значения: _____

Если есть «выбросы», укажите их значения: _____

Задача 4. (*3 балла*)

Планируется опрос жителей в крупных городах с целью узнать их мнение о миграционной политике, регулирующей численность трудовых мигрантов. Планируется оценить долю жителей, которые считают, что нужно полностью запретить въезд граждан, приезжающих из других стран на заработки. Какой объём выборки потребуется исследователям, чтобы обеспечить выполнение следующих требований: уровень надёжности выводов будущего исследования должен составить 99 %, а доля жителей, придерживающихся жёстких мер миграционной политики, должна быть оценена с точностью ±5 %? По предварительным оценкам экспертов, доля оценивается в 15 %.

Ответ: _____

Задача 5. (*4 балла*)

Проводится дегустация семи новых сортов кофе. Два дегустатора оценивают их вкусовые качества по 10-балльной шкале. Бóльшая оценка соответствует лучшим вкусовым свойствам кофе:

Оценки	Сорт кофе						
	А	Б	В	Г	Д	Е	Ж
Дегустатор 1	9	4	7	6	8	3	1
Дегустатор 2	6	6	10	5	4	3	7

В какой мере оценки (мнения) двух дегустаторов согласуются между собой? Вычислите коэффициент ранговой корреляции Спирмена и дайте его содержательную интерпретацию.

Ответ:

Коэффициент корреляции Спирмена равен _____

Содержательная интерпретация: _____

_____

Задача 6. (*4 балла*)

Торгово-производственная компания продаёт обувь. В ходе разработки маркетинговой политики по стимулированию продаж фирмы исследовалась связь между уровнем материального положения покупателей и их отношением к наличию скидок (уценке) на товар в конце сезона. Данные сведены в таблицу. Можно ли утверждать, что существует связь между указанными переменными (признаками)? Проверку соответствующей гипотезы проведите на уровне доверительной вероятности 0,99.

Материальное положение	Отношение к наличию скидок (уценке)	
	«Скидки — это странно»	«Скидки — это хорошо»
Высокое	35	25
Среднее	30	60

Ответ:

Статистический вывод проверки гипотезы: _____

_____

Содержательный вывод проверки гипотезы: _____

_____

Задача 7. (*4 балла*)

Проведено исследование, в ходе которого, в частности, ставился вопрос о том, кто больше времени уделяет спорту — мужчины или женщины? Было опрошено 35 мужчин в возрасте 25—35 лет и 40 женщин того же возраста. Выяснили, что мужчины в среднем тратят на спорт 4,2 ч в неделю, женщины — 3,1 часов. Оценки стандартного отклонения составили соответственно — 1,3 для мужчин и 0,9 для женщин. Можно ли на основе полученной

информации подтвердить суждение о том, что мужчины более активны в занятиях спортом? Проверку соответствующей гипотезы реализовать на 95%-м уровне доверительной вероятности против направленной альтернативы: «мужчины тратят на спорт больше времени, чем женщины».

Ответ:

Статистический вывод проверки гипотезы: _____

_____

Содержательный вывод проверки гипотезы: _____

_____

Итого баллов (заполняется преподавателем): _____ ***баллов***

Экзаменационная работа. Вариант № 2

Фамилия, имя _____ № группы _____

Задача 1. (*2 балла*)

В клубе начинающих парашютистов у всех участников исследователи поинтересовались, сколько раз он (она) успел(-а) прыгнуть с парашютом к настоящему времени. Число прыжков является случайной величиной с законом распределения:

X, прыжков	0	1	2	3
p (вероятность)	0,2	?	0,3	0,1

Найдите:

а) среднее значение случайной величины X;

б) дисперсию случайной величины X;

в) стандартное отклонение случайной величины X,

где X — число прыжков, совершенных парашютистом.

Ответ:

Среднее = Дисперсия = Стандартное отклонение =

Задача 2. (*4 балла*)

Средний возраст посетителей некоторого бильярдного клуба составляет 37,5 лет. Стандартное отклонение составляет 4,6 лет. Какова вероятность того, что возраст случайного посетителя будет равен от 40 до 45 лет? (если считать, что возраст посетителей бильярдного клуба описывается нормальным законом распределения с указанными параметрами). Схематично отразите искомую вероятность на графике нормального распределения и на графике стандартного нормального распределения.

Ответ: _____

Задача 3. (*4 балла*)

На основе данных опроса «ФОМнибус» («Репетиторство — мать учения?»), проведённого среди жителей России летом 2012 г., располагаем следующей информацией об ответах людей (с высшим образованием и без

высшего образования) на вопрос: «Если говорить в целом, в таком явлении, как репетиторство, больше хорошего или больше плохого?» (число ответивших составляет 1 035 человек):

В репетиторстве больше хорошего или плохого?	Образование	
	Есть высшее	Нет высшего
Больше хорошего	656	204
Больше плохого	151	24

В ячейках — количество респондентов. Можно ли считать два признака статистически связанными? Проверку соответствующей гипотезы осуществите на уровне доверительной вероятности 0,99.

Ответ:

Статистический вывод проверки гипотезы: _____

_____

Содержательный вывод проверки гипотезы: _____

_____

Задача 4. (*3 балла*)

Выборочное обследование 3 000 человек, проведённое Фондом «Общественное мнение» в ноябре 2011 г. перед выборами Государственной думы, обнаружило, что категорически не хотят идти на выборы 13 % избирателей. Основываясь на этих данных, с вероятностью 0,92 сделайте прогноз относительно доли тех, кто, напротив, планирует принять участие в голосовании.

Ответ: _____

Задача 5. (*6 баллов*)

Имеется следующая статистика по ресторанам-закусочным быстрого обслуживания (в выборку включены рестораны, рассчитанные на большое число посетителей и имеющие просторные обеденные залы), расположенным на автомобильной магистрали:

Ресторан-закусочная	№ 1	№ 2	№ 3	№ 4	№ 5	№ 6
Средняя цена тарелки супа, руб.	40	30	120	50	140	100
Процент занятых столиков в дневное время	90	90	70	85	65	80

Вычислите коэффициент вариации для показателя «средняя цена тарелки супа». Интерпретируйте его. Вычислите коэффициент линейной корреляции Пирсона между указанными двумя показателями и дайте его содержательную интерпретацию.

Ответ:

Коэффициент вариации равен _____

Содержательная интерпретация: _____
_____
Коэффициент корреляции Пирсона равен: _____
Содержательная интерпретация: _____
_____

Задача 6*. (*3 балла*)

На уровне значимости 0,05 проверьте статистическую достоверность коэффициента корреляции, рассчитанного в предыдущей задаче № 6. Проверку осуществите против направленной альтернативы. Сделайте статистический и содержательный выводы.

Ответ:

Статистический вывод проверки гипотезы: _____
_____
Содержательный вывод проверки гипотезы: _____
_____

Задача 7. (*4 балла*)

Исследователи выяснили, что среди 120 опрошенных ими студентов филологических факультетов различных вузов оказалось 13 % юношей. Среди 140 опрошенных студентов социологических факультетов этот показатель составил 21 %. Обе выборки вероятностные. Является ли неслучайным наблюдаемое различие в выборочных оценках долей — на социологических факультетах действительно учится относительно больше юношей, чем на филологических факультетах? Проверку соответствующей гипотезы реализуйте против направленной альтернативы. Уровень доверительной вероятности примите равным 0,99.

Ответ:

Статистический вывод проверки гипотезы: _____
_____
Содержательный вывод проверки гипотезы: _____
_____

Итого баллов* (*заполняется преподавателем*): _____ *баллов

Экзаменационная работа. Вариант № 3

Фамилия, имя _____ № группы _____

Задача 1. (*2 балла*)

Продавец оценивает, сколько зонтиков продается у него за день. Он знает, что число зонтиков, требуемых за день, является случайной величиной с законом распределения:

X, шт.	0	1	2	3
p (вероятность)	0,3	0,3	?	0,2

Найдите:
а) Среднее значение случайной величины X
б) Дисперсию случайной величины X
в) Стандартное отклонение случайной величины X,
где X — количество зонтиков, продаваемых в магазине за день.
Ответ:
Среднее = Дисперсия = Стандартное отклонение =

Задача 2. (*4 балла*)

Средняя продолжительность нахождения покупателя в некотором магазине одежды составляет 33 мин. Стандартное отклонение равно 8 мин. Какова вероятность, что продолжительность пребывания в магазине для случайно выбранного покупателя окажется больше 20 мин и меньше 36 мин (если считать, что продолжительность пребывания в магазине одежды описывается нормальным законом распределения с указанными параметрами)? Схематично отразите искомую вероятность на графике нормального распределения и на графике стандартного нормального распределения.

Ответ: _____

Задача 3. (*4 балла*)

В ходе опроса людей, которые курят, респондентам был адресован вопрос: «Сколько примерно сигарет Вы выкуриваете в течение дня?». Ответы первых 10 респондентов представлены ниже:

X (число сигарет): **20**, **7**, **9**, **9**, **5**, **4**, **2**, **1**, **11**, **10**.

Найдите:

а) медиану выборки,

б) нижнюю и верхнюю квартили выборки (по методу Дж. Тьюки).

в) Определите по методу Дж. Тьюки, есть ли в этой выборке нетипичные, нехарактерные значения наблюдений — «подозрительные» и/или «выбросы».

Ответ:

Медиана равна _____ Нижняя квартиль равна _____

Верхняя квартиль равна _____

Если есть «подозрительные» наблюдения, укажите их значения: _____

Если есть «выбросы», укажите их значения: _____

Задача 4. (*3 балла*)

Обследование (в котором объём случайной выборки составил 350 семейных пар с маленькими детьми) показало, что в среднем они посещают цирк 1,4 раз в год, дисперсия оказалась равной 1,44 (посещений)^2. Можно ли утверждать на уровне значимости 0,1, что в среднем се́мьи, в которых есть маленькие дети, посещают цирковые представления чаще одного раза в год?

Ответ: _____

Задача 5. (*4 балла*)

Компания предложила двум маркетологам (экспертам) оценить по 10-балльной шкале эстетическое восприятие нового кулинарного журнала,

выполненного в 8-и разных обложках. Более высокие оценки соответствуют более приятному эстетическому восприятию журнала.

Обложка журнала	Эксперт 1, оценки	Эксперт 2, оценки
А	8	6
Б	5	9
В	6	7
Г	7	5
Д	4	4
Е	10	8
Ж	9	10
З	3	4

В какой мере оценки (мнения) двух экспертов согласуются между собой? Вычислите коэффициент ранговой корреляции Спирмена и дайте его содержательную интерпретацию.

Ответ:

Коэффициент корреляции Спирмена равен: _____

Содержательная интерпретация: _____

_____

Задача 6. (*4 баллов*)

Проводилось исследование, в рамках которого сравнивались академические результаты выпускников средних школ в двух российских городах — А и Б. В частности, изучался вопрос, насколько результаты ЕГЭ однородны у выпускников из городов А и Б. Имеются следующие результаты выборки:

Город	Средний балл, полученный на выпускных экзаменах (ЕГЭ)
А	40, 50, 40, 70 (объём выборки = 4 чел.)
Б	30, 90, 60, 70, 50 (объём выборки = 5 чел.)

Можно ли утверждать, что степень *однородности* знаний, продемонстрированных выпускниками на финальной аттестации (ЕГЭ) одинакова в городах А и Б? Проверку соответствующей статистической гипотезы осуществите на уровне доверия 0,95 против ненаправленной альтернативы.

Ответ:

Статистический вывод проверки гипотезы: _____

_____

Содержательный вывод проверки гипотезы: _____

_____

Задача 7. (*4 балла*)

На основе данных опроса «ФОМнибус» («О поведении в экстренных ситуациях»), проведённого среди жителей России в июле 2012 г., располагаем следующей информацией об ответах на вопрос «Скажите, пожалуйста, Вы знаете или не знаете, где находится ближайшее бомбоубежище?»:

Знаете ли Вы, где находится ближайшее бомбоубежище?	Пол	
	Мужчины	Женщины
Да, знаю	182	148
Нет, не знаю	493	677

В ячейках — количество респондентов. Можно ли считать два признака статистически связанными? Проверку соответствующей гипотезы осуществите на уровне доверительной вероятности 0,90.

Ответ:

Статистический вывод проверки гипотезы: _____

_____

Содержательный вывод проверки гипотезы: _____

_____

Итого баллов* (*заполняется преподавателем*): _____ *баллов

Экзаменационная работа. Вариант № 4

Фамилия, имя _____ № группы _____

Задача 1. (*2 балла*)

Художник размышляет, сколько заказов (с просьбой нарисовать портрет) поступает к нему в течение месяца. Он знает, что число таких заказов является случайной величиной с законом распределения:

X, заказов	0	1	2	3
p (вероятность)	0,1	0,1	0,1	?

Найдите:

а) среднее значение случайной величины X;

б) дисперсию случайной величины X;

в) стандартное отклонение случайной величины X,

где X — число заказов, поступающих художнику в течение месяца.

Ответ:

Среднее = Дисперсия = Стандартное отклонение =

Задача 2. (*4 балла*)

Средняя стоимость 1 кг арбуза в торговых палатках города составляет 30 руб., а дисперсия 16. Какова вероятность, что в наугад выбранной торго-

вой палатке стоимость 1 кг арбуза составит менее 20 руб.? (если считать, что стоимость 1 кг арбуза — случайная величина, которая описывается нормальным законом распределения с указанными параметрами)? Схематично отразите искомую вероятность на графике нормального распределения и на графике стандартного нормального распределения.

Ответ: _____

Задача 3. (*4 балла*)

В ходе исследования в различных странах мира изучался уровень безработицы (измеренный как процент людей, не имеющих работу, от общей численности экономически активного населения). Ниже представлена выборка из 12 стран:

X (уровень безработицы, %): **7**, **2**, **1**, **10**, **12**, **10**, **17**, **15**, **12**, **27**, **9**, **13**.

Найдите:

а) медиану выборки;

б) нижнюю и верхнюю квартили выборки (по методу Дж. Тьюки).

в) Определите по методу Дж. Тьюки, есть ли в этой выборке нетипичные, нехарактерные значения наблюдений — «подозрительные» и/или «выбросы».

Ответ:

Медиана равна _____ Нижняя квартиль равна _____

Верхняя квартиль равна _____

Если есть «подозрительные» наблюдения, укажите их значения: _____

Если есть «выбросы», укажите их значения: _____

Задача 4. (*3 балла*)

В ходе одного из выборочных обследований Фонда «Общественное мнение» были опрошены 1 500 человек. В опросе оценивалась доля жителей, которые считают, что Россия в настоящее время развивается в правильном направлении; 465 человек заявили о том, что страна развивается в верном направлении. Какой вывод следует сделать о доле тех граждан, кто оценивает траекторию развития страны как правильную? Доверительную вероятность примите равной 0,90.

Ответ: _____

Задача 5. (*4 балла*)

В школе по фигурному катанию два тренера оценивали по 12-балльной шкале профессиональную подготовку своих спортсменов в парном катании (шесть танцевальных пар). Больший балл соответствует более высокому профессионализму спортсменов. Вот оценки, которые были выставлены фигуристам:

Оценка	Номер танцевальной пары					
	1	2	3	4	5	6
Тренер 1	9	6	11	11	5	8
Тренер 2	12	10	8	10	7	6

Можно ли утверждать, что мнения (оценки) двух тренеров согласуются между собой? Задачу решите с помощью коэффициента ранговой корреляции Спирмена. Интерпретируйте полученный ответ.

Ответ:

Коэффициент корреляции Спирмена равен: _____

Содержательная интерпретация: _____

_____

Задача 6. (*4 балла*)

Команде врачей-исследователей предстоит оценить эффективность новой лечебной гимнастики (специального набора упражнений), предложенной пациентам, страдающим заболеванием А*. Состояние здоровья шести человек, принявших участие в эксперименте, было зафиксировано до начала проекта и спустя три месяца после его завершения (оно измерялось по специально разработанной 20-балльной медицинской шкале). Бóльший балл соответствует более хорошему самочувствию. Считая выборки связными (парными), проверьте соответствующую гипотезу об отсутствии среднего эффекта воздействия. Можно ли считать гимнастику эффективным способом лечения? Конкурирующая гипотеза направленная: «Самочувствие пациентов, в среднем, стало лучше, чем было». Уровень доверия принять 0,99.

Самочувствие, баллы по шкале	Пациент					
	A	B	C	D	E	F
До лечебной гимнастики	10	9	10	15	12	13
После лечебной гимнастики	12	15	10	13	13	15

Ответ:

Статистический вывод проверки гипотезы: _____

_____

Содержательный вывод проверки гипотезы: _____

_____

Задача 7. (*4 балла*)

В исследовательском проекте «Доминанты» (опрос «ФОМнибус» по теме: «Новый тренер сборной России по футболу»), проведённом в июле 2012 г., жителям России был задан вопрос: «Как Вы относитесь к тому, что новым тренером сборной России по футболу стал итальянский тренер Фабио Капелло — положительно или отрицательно?». Распределение ответов среди мужчин и женщин представлено в таблице (число ответивших составляет 1 351 человек):

Как Вы относитесь — положительно или отрицательно?	Пол	
	Мужчины	Женщины
Положительно	128	83

Окончание табл.

Как Вы относитесь — положительно или отрицательно?	Пол	
	Мужчины	Женщины
Отрицательно	189	116
Мне это безразлично	290	545
Всего, чел.	608	743

Сравните доли ответов «Мне это безразлично» среди мужчин и женщин. Можно ли на основе полученной информации подтвердить довольно распространённое суждение о том, что женщины в большей степени безразличны к «футбольным» делам российской сборной, нежели мужчины? Проверку соответствующей гипотезы реализовать на 95%-м уровне доверительной вероятности против направленной альтернативы, исходя из данных.

Ответ:

Статистический вывод проверки гипотезы: _____

Содержательный вывод проверки гипотезы: _____

_____

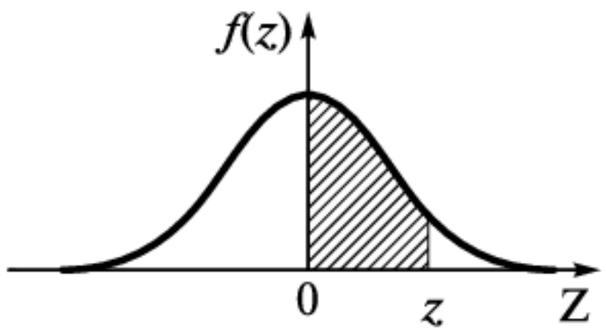
Итого баллов* (*заполняется преподавателем*): _____ *баллов

Статистические таблицы распределений

Примечание. Таблицы составлены с помощью операторов функций в среде Microsoft Excel.

Таблица П1. **Стандартное нормальное *z*-распределение**

Интеграл от 0 до *z*



Z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,0000	0,0040	0,0080	0,0120	0,0160	0,0199	0,0239	0,0279	0,0319	0,0359
0,1	0,0398	0,0438	0,0478	0,0517	0,0557	0,0596	0,0636	0,0675	0,0714	0,0753
0,2	0,0793	0,0832	0,0871	0,0910	0,0948	0,0987	0,1026	0,1064	0,1103	0,1141
0,3	0,1179	0,1217	0,1255	0,1293	0,1331	0,1368	0,1406	0,1443	0,1480	0,1517
0,4	0,1554	0,1591	0,1628	0,1664	0,1700	0,1736	0,1772	0,1808	0,1844	0,1879
0,5	0,1915	0,1950	0,1985	0,2019	0,2054	0,2088	0,2123	0,2157	0,2190	0,2224
0,6	0,2257	0,2291	0,2324	0,2357	0,2389	0,2422	0,2454	0,2486	0,2517	0,2549
0,7	0,2580	0,2611	0,2642	0,2673	0,2704	0,2734	0,2764	0,2794	0,2823	0,2852
0,8	0,2881	0,2910	0,2939	0,2967	0,2995	0,3023	0,3051	0,3078	0,3106	0,3133
0,9	0,3159	0,3186	0,3212	0,3238	0,3264	0,3289	0,3315	0,3340	0,3365	0,3389
1,0	0,3413	0,3438	0,3461	0,3485	0,3508	0,3531	0,3554	0,3577	0,3599	0,3621
1,1	0,3643	0,3665	0,3686	0,3708	0,3729	0,3749	0,3770	0,3790	0,3810	0,3830
1,2	0,3849	0,3869	0,3888	0,3907	0,3925	0,3944	0,3962	0,3980	0,3997	0,4015
1,3	0,4032	0,4049	0,4066	0,4082	0,4099	0,4115	0,4131	0,4147	0,4162	0,4177
1,4	0,4192	0,4207	0,4222	0,4236	0,4251	0,4265	0,4279	0,4292	0,4306	0,4319
1,5	0,4332	0,4345	0,4357	0,4370	0,4382	0,4394	0,4406	0,4418	0,4429	0,4441
1,6	0,4452	0,4463	0,4474	0,4484	0,4495	0,4505	0,4515	0,4525	0,4535	0,4545
1,7	0,4554	0,4564	0,4573	0,4582	0,4591	0,4599	0,4608	0,4616	0,4625	0,4633

Окончание табл. П1

Z	**0,00**	**0,01**	**0,02**	**0,03**	**0,04**	**0,05**	**0,06**	**0,07**	**0,08**	**0,09**
1,8	0,4641	0,4649	0,4656	0,4664	0,4671	0,4678	0,4686	0,4693	0,4699	0,4706
1,9	0,4713	0,4719	0,4726	0,4732	0,4738	0,4744	0,4750	0,4756	0,4761	0,4767
2,0	0,4772	0,4778	0,4783	0,4788	0,4793	0,4798	0,4803	0,4808	0,4812	0,4817
2,1	0,4821	0,4826	0,4830	0,4834	0,4838	0,4842	0,4846	0,4850	0,4854	0,4857
2,2	0,4861	0,4864	0,4868	0,4871	0,4875	0,4878	0,4881	0,4884	0,4887	0,4890
2,3	0,4893	0,4896	0,4898	0,4901	0,4904	0,4906	0,4909	0,4911	0,4913	0,4916
2,4	0,4918	0,4920	0,4922	0,4925	0,4927	0,4929	0,4931	0,4932	0,4934	0,4936
2,5	0,4938	0,4940	0,4941	0,4943	0,4945	0,4946	0,4948	0,4949	0,4951	0,4952
2,6	0,4953	0,4955	0,4956	0,4957	0,4959	0,4960	0,4961	0,4962	0,4963	0,4964
2,7	0,4965	0,4966	0,4967	0,4968	0,4969	0,4970	0,4971	0,4972	0,4973	0,4974
2,8	0,4974	0,4975	0,4976	0,4977	0,4977	0,4978	0,4979	0,4979	0,4980	0,4981
2,9	0,4981	0,4982	0,4982	0,4983	0,4984	0,4984	0,4985	0,4985	0,4986	0,4986
3,0	0,4987	0,4987	0,4987	0,4988	0,4988	0,4989	0,4989	0,4989	0,4990	0,4990

Таблица П2. *t*-распределение Стьюдента

df	Доверительная вероятность и уровень α				
	$\gamma = 0{,}9$ (90 %)	$\gamma = 0{,}95$ (95 %)	$\gamma = 0{,}98$ (98 %)	$\gamma = 0{,}99$ (99 %)	$\gamma = 0{,}999$ (99,9 %)
	$\alpha = 0{,}1$ $(\alpha/2 = 0{,}05)$	$\alpha = 0{,}05$ $(\alpha/2 = 0{,}025)$	$\alpha = 0{,}02$ $(\alpha/2 = 0{,}01)$	$\alpha = 0{,}01$ $(\alpha/2 = 0{,}005)$	$\alpha = 0{,}001$ $(\alpha/2 = 0{,}0005)$
1	6,314	12,706	31,821	63,657	636,619

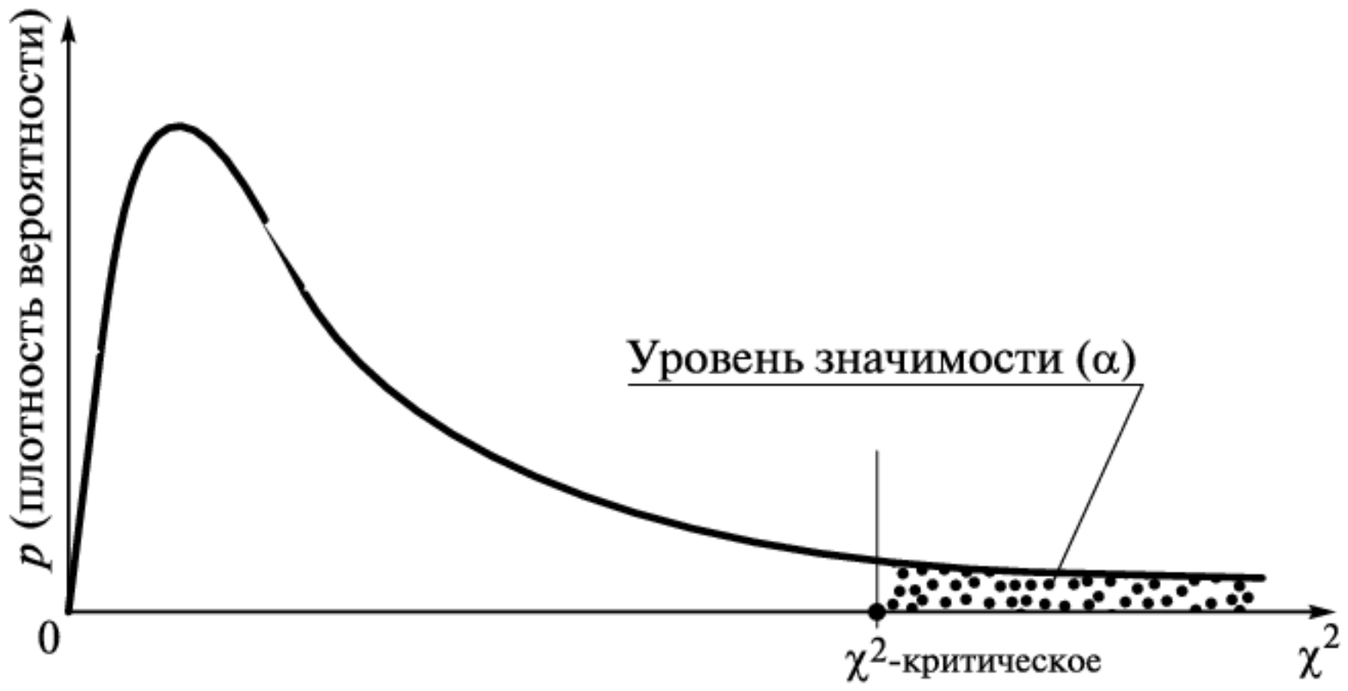
Продолжение табл. П2

df	Доверительная вероятность и уровень α				
	$\gamma = 0{,}9$ (90 %)	$\gamma = 0{,}95$ (95 %)	$\gamma = 0{,}98$ (98 %)	$\gamma = 0{,}99$ (99 %)	$\gamma = 0{,}999$ (99,9 %)
	$\alpha = 0{,}1$ ($\alpha/2 = 0{,}05$)	$\alpha = 0{,}05$ ($\alpha/2 = 0{,}025$)	$\alpha = 0{,}02$ ($\alpha/2 = 0{,}01$)	$\alpha = 0{,}01$ ($\alpha/2 = 0{,}005$)	$\alpha = 0{,}001$ ($\alpha/2 = 0{,}0005$)
2	2,920	4,303	6,965	9,925	31,599
3	2,353	3,182	4,541	5,841	12,924
4	2,132	2,776	3,747	4,604	8,610
5	2,015	2,571	3,365	4,032	6,869
6	1,943	2,447	3,143	3,707	5,959
7	1,895	2,365	2,998	3,500	5,408
8	1,860	2,306	2,897	3,355	5,041
9	1,833	2,262	2,821	3,250	4,781
10	1,813	2,228	2,764	3,169	4,587
11	1,796	2,201	2,718	3,106	4,437
12	1,782	2,179	2,681	3,055	4,318
13	1,771	2,160	2,650	3,012	4,221
14	1,761	2,145	2,625	2,977	4,141
15	1,753	2,131	2,603	2,947	4,073
16	1,746	2,120	2,584	2,921	4,015
17	1,740	2,110	2,567	2,898	3,965
18	1,734	2,101	2,552	2,878	3,922
19	1,729	2,093	2,540	2,861	3,883
20	1,725	2,086	2,528	2,845	3,850
21	1,721	2,080	2,518	2,831	3,819
22	1,717	2,074	2,508	2,819	3,792
23	1,714	2,069	2,500	2,807	3,768

Окончание табл. П2

df	Доверительная вероятность и уровень α				
	$\gamma = 0{,}9$ (90 %)	$\gamma = 0{,}95$ (95 %)	$\gamma = 0{,}98$ (98 %)	$\gamma = 0{,}99$ (99 %)	$\gamma = 0{,}999$ (99,9 %)
	$\alpha = 0{,}1$ ($\alpha/2 = 0{,}05$)	$\alpha = 0{,}05$ ($\alpha/2 = 0{,}025$)	$\alpha = 0{,}02$ ($\alpha/2 = 0{,}01$)	$\alpha = 0{,}01$ ($\alpha/2 = 0{,}005$)	$\alpha = 0{,}001$ ($\alpha/2 = 0{,}0005$)
24	1,711	2,064	2,492	2,797	3,745
25	1,708	2,060	2,485	2,787	3,725
26	1,706	2,056	2,479	2,779	3,707
27	1,703	2,052	2,473	2,771	3,690
28	1,701	2,048	2,467	2,763	3,674
29	1,699	2,045	2,462	2,756	3,659
30	1,697	2,042	2,457	2,750	3,646
35	1,690	2,030	2,438	2,724	3,591
40	1,684	2,021	2,423	2,705	3,551
45	1,679	2,014	2,412	2,690	3,520
50	1,676	2,009	2,403	2,678	3,496
55	1,673	2,004	2,396	2,668	3,476
60	1,671	2,000	2,390	2,660	3,460
70	1,667	1,994	2,381	2,648	3,435
80	1,664	1,990	2,374	2,639	3,416
90	1,662	1,987	2,369	2,632	3,402
100	1,660	1,984	2,364	2,626	3,391
110	1,659	1,982	2,361	2,621	3,381
120	1,658	1,980	2,358	2,617	3,374
∞	1,645	1,960	2,326	2,576	3,291

Таблица П3. **χ^2-распределение Пирсона**

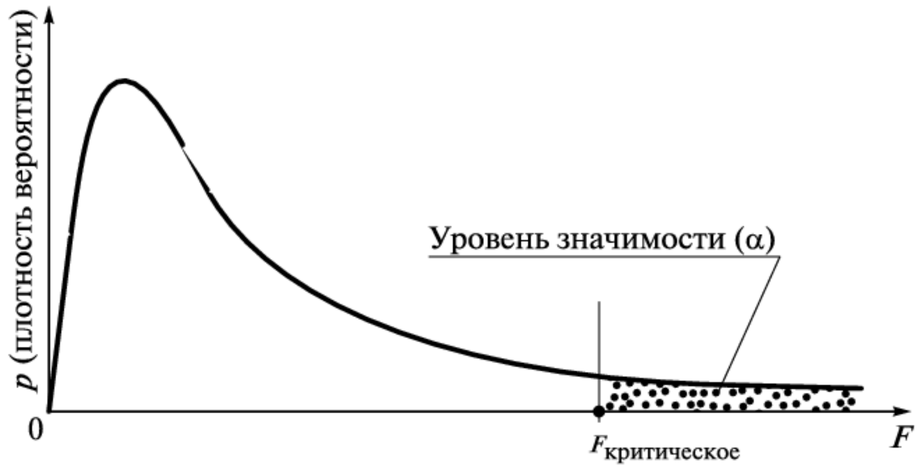


df	Доверительная вероятность и уровень α			
	$\gamma = 0{,}9$ (90 %)	$\gamma = 0{,}95$ (95 %)	$\gamma = 0{,}99$ (99 %)	$\gamma = 0{,}999$ (99,9 %)
	$\alpha = 0{,}1$	$\alpha = 0{,}05$	$\alpha = 0{,}01$	$\alpha = 0{,}001$
1	2,71	3,84	6,63	10,83
2	4,61	5,99	9,21	13,82
3	6,25	7,81	11,34	16,27
4	7,78	9,49	13,28	18,47
5	9,24	11,07	15,09	20,52
6	10,64	12,59	16,81	22,46
7	12,02	14,07	18,48	24,32
8	13,36	15,51	20,09	26,12
9	14,68	16,92	21,67	27,88
10	15,99	18,31	23,21	29,59
11	17,28	19,68	24,72	31,26
12	18,55	21,03	26,22	32,91
13	19,81	22,36	27,69	34,53
14	21,06	23,68	29,14	36,12
15	22,31	25,00	30,58	37,70
16	23,54	26,30	32,00	39,25

Окончание табл. П3

df	Доверительная вероятность и уровень α			
	$\gamma = 0{,}9$ (90 %)	$\gamma = 0{,}95$ (95 %)	$\gamma = 0{,}99$ (99 %)	$\gamma = 0{,}999$ (99,9 %)
	$\alpha = 0{,}1$	$\alpha = 0{,}05$	$\alpha = 0{,}01$	$\alpha = 0{,}001$
17	24,77	27,59	33,41	40,79
18	25,99	28,87	34,81	42,31
19	27,20	30,14	36,19	43,82
20	28,41	31,41	37,57	45,31
21	29,62	32,67	38,93	46,80
22	30,81	33,92	40,29	48,27
23	32,01	35,17	41,64	49,73
24	33,20	36,42	42,98	51,18
25	34,38	37,65	44,31	52,62
и т. д.	...	...	...	...

Таблица П4. ***F*-распределение Фишера. Уровень значимости α = 0,05 (5 %)**

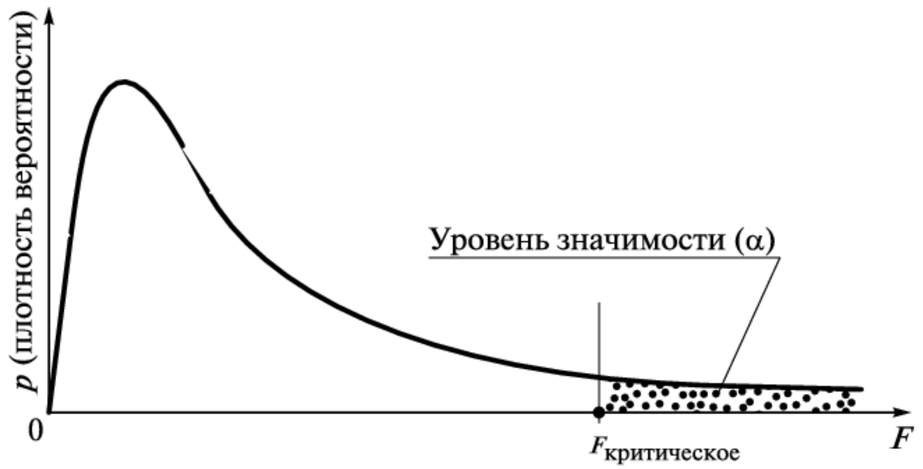


df		Число степеней свободы числителя дроби (выборка № 1): df_1																
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**12**	**15**	**20**	**30**	**40**	**60**	+∞
Число степеней свободы знаменателя дроби (выборка № 2): df_2	**1**	161,5	199,5	215,7	224,6	230,2	234	236,8	238,9	240,5	241,9	243,9	246	248	250,1	251,1	252,2	254,3
	2	18,51	19,00	19,16	19,25	19,30	19,33	19,35	19,37	19,38	19,40	19,41	19,43	19,44	19,46	19,47	19,48	19,50
	3	10,13	9,55	9,28	9,12	9,01	8,94	8,87	8,85	8,81	8,79	8,74	8,71	8,66	8,62	8,59	8,57	8,53
	4	7,71	6,94	6,59	6,38	6,26	6,16	6,09	6,04	6,00	5,96	5,91	5,86	5,80	5,75	5,72	5,69	5,63
	5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,77	4,74	4,68	4,62	4,56	4,50	4,46	4,43	4,37
	6	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10	4,06	4,00	3,94	3,87	3,81	3,77	3,74	3,67
	7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,64	3,57	3,51	3,44	3,38	3,34	3,30	3,23

Окончание табл. П4

df		Число степеней свободы числителя дроби (выборка № 1): df_1																
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**12**	**15**	**20**	**30**	**40**	**60**	$+\infty$
Число степеней свободы знаменателя дроби (выборка № 2): df_2	**8**	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39	3,35	3,28	3,22	3,15	3,08	3,04	3,01	2,93
	9	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,14	3,07	3,01	2,94	2,86	2,83	2,79	2,71
	10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,98	2,91	2,85	2,77	2,70	2,66	2,62	2,54
	12	4,75	3,89	3,49	3,26	3,11	3,00	2,91	2,85	2,80	2,75	2,69	2,62	2,54	2,47	2,43	2,38	2,30
	15	4,54	3,68	3,29	3,06	2,90	2,79	2,71	2,64	2,59	2,54	2,48	2,40	2,33	2,25	2,20	2,16	2,07
	20	4,35	3,49	3,10	2,87	2,71	2,60	2,51	2,45	2,39	2,35	2,28	2,20	2,12	2,04	1,99	1,95	1,84
	30	4,17	3,32	2,92	2,69	2,53	2,42	2,33	2,27	2,21	2,16	2,09	2,01	1,93	1,84	1,79	1,74	1,62
	40	4,08	3,23	2,84	2,61	2,45	2,34	2,25	2,18	2,12	2,08	2,00	1,92	1,84	1,74	1,69	1,64	1,51
	60	4,00	3,15	2,76	2,53	2,37	2,25	2,17	2,10	2,04	1,99	1,92	1,84	1,75	1,65	1,59	1,53	1,39
	∞	3,84	3,00	2,60	2,37	2,21	2,10	2,01	1,94	1,88	1,83	1,75	1,67	1,57	1,46	1,39	1,32	1,00

Таблица П5. ***F*-распределение Фишера. Уровень значимости α = 0,025 (2,5 %)**



df		Число степеней свободы числителя дроби (выборка № 1): df_1																
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**12**	**15**	**20**	**30**	**40**	**60**	+∞
Число степеней свободы знаменателя дроби (выборка № 2): df_2	**1**	647,8	799,5	864,2	599,6	921,9	937,1	948,2	956,7	963,3	968,6	976,7	984,9	993,1	1 001	1 006	1 010	1 018
	2	38,51	39,00	39,17	39,25	39,30	39,33	39,36	39,37	39,39	39,40	39,41	39,43	39,45	39,47	39,47	39,48	39,50
	3	17,44	16,04	15,44	15,10	14,88	14,73	14,62	14,54	14,47	14,42	14,34	14,25	14,17	14,08	14,04	13,99	13,90
	4	12,22	10,65	9,98	9,60	9,36	9,20	9,07	8,98	8,90	8,84	8,75	8,66	8,56	8,46	8,41	8,36	8,26
	5	10,01	8,43	7,76	7,39	7,15	6,98	6,85	6,76	6,68	6,62	6,52	6,43	6,33	6,23	6,18	6,12	6,02
	6	8,81	7,26	6,60	6,23	5,99	5,82	5,70	5,60	5,52	5,46	5,37	5,27	5,17	5,07	5,01	4,96	4,85
	7	8,07	6,54	5,89	5,52	5,29	5,12	4,99	4,90	4,82	4,76	4,67	4,57	4,47	4,36	4,31	4,25	4,14

Окончание табл. П5

df		Число степеней свободы числителя дроби (выборка № 1): df_1																
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**12**	**15**	**20**	**30**	**40**	**60**	$+\infty$
Число степеней свободы знаменателя дроби (выборка № 2): df_2	**8**	7,57	6,06	5,42	5,05	4,82	4,65	4,53	4,43	4,36	4,30	4,20	4,10	4,00	3,89	3,84	3,78	3,67
	9	7,21	5,71	5,08	4,72	4,48	4,32	4,20	4,10	4,03	3,96	3,87	3,77	3,67	3,56	3,51	3,45	3,33
	10	6,94	5,46	4,83	4,47	4,24	4,07	3,95	3,85	3,78	3,72	3,62	3,52	3,42	3,31	3,26	3,20	3,08
	12	6,55	5,10	4,47	4,12	3,89	3,73	3,61	3,51	3,44	3,37	3,28	3,18	3,07	2,96	2,91	2,85	2,73
	15	6,20	4,77	4,15	3,80	3,58	3,41	3,29	3,20	3,12	3,06	2,96	2,86	2,76	2,64	2,59	2,52	2,40
	20	5,87	4,46	3,86	3,51	3,29	3,13	3,01	2,91	2,84	2,77	2,68	2,57	2,46	2,35	2,29	2,22	2,09
	30	5,57	4,18	3,59	3,25	3,03	2,87	2,75	2,65	2,57	2,51	2,41	2,31	2,20	2,07	2,01	1,94	1,79
	40	5,42	4,05	3,46	3,13	2,90	2,74	2,62	2,53	2,45	2,39	2,29	2,18	2,07	1,94	1,88	1,80	1,64
	60	5,29	3,93	3,34	3,01	2,79	2,63	2,51	2,41	2,33	2,27	2,17	2,06	1,94	1,82	1,74	1,67	1,48
	+∞	5,02	3,69	3,12	2,79	2,57	2,41	2,29	2,19	2,11	2,05	1,94	1,83	1,71	1,57	1,48	1,39	1,00

Греческий алфавит

Название буквы	Прописная	Строчная	Название буквы	Прописная	Строчная
Альфа	A	α	Ню	N	ν
Бета	B	β	Кси	Ξ	ξ
Гамма	Γ	γ	Омикрон	O	o
Дельта	Δ	δ	Пи	Π	π
Эпсилон	E	ε	Ро	P	ρ
Дзета	Z	ζ	Сигма	Σ	σ
Эта	H	η	Тау	T	τ
Тета	Θ	θ	Ипсилон	Y	υ
Йота	I	ι	Фи	Φ	φ
Каппа	K	κ	Хи	X	χ
Лямбда	Λ	λ	Пси	Ψ	ψ
Мю	M	μ	Омега	Ω	ω

Синтаксис статистических функций в Microsoft Excel

Реализация расчётов может быть выполнена с помощью следующих команд вызова статистических функций в среде Microsoft Excel

Оцениваемый показатель	Команда вызова функции в Excel. Синтаксис	Краткие пояснения
	Визуализация данных	
Полигон распределения	Вставка → Вставка диаграммы → График	Строит полигон распределения для вариационного ряда
Гистограмма распределения	Вставка → Вставка диаграммы → Гистограмма	Строит гистограмму распределения для вариационного ряда
	Описательная статистика	
Сумма	=сумм(число1; число2; ...)	Суммирует значения в имеющемся ряду данных
Минимум	=мин(число1; число2;...)	Выявляет наименьшее (минимальное) значение в ряду данных
Максимум	=макс(число1; число2; ...)	Выявляет наибольшее (максимальное) значение в ряду данных
Мода	=мода(число1; число2; ...)	Определяет значение моды в ряду данных
Медиана	=медиана(число1; число2; ...)	Определяет значение медианы в ряду данных

Среднее	=срзнач(число1; число2; ...)	Рассчитывает среднее арифметическое для имеющегося ряда данных
Сумма квадратов отклонений	=квадроткл(число1; число2; ...)	Вычисляет общую сумму квадратов отклонений значений выборки от среднего арифметического этой выборки
Дисперсия генеральная	=диспр(число1; число2; ...)	Вычисляет смещенную или генеральную дисперсию значений ряда (в знаменателе дроби присутствует «n», а не «$n - 1$»)
Дисперсия выборочная (несмещённая)	=дисп(число1; число2; ...)	Вычисляет несмещённую (выборочную) дисперсию значений ряда (в знаменателе дроби присутствует «$n - 1$»)
Стандартное отклонение генеральное	=стандотклон(число1; число2; ...)	Вычисляет смещенное или генеральное стандартное отклонение значений ряда (равняется квадратному корню из дисперсии генеральной)
Стандартное отклонение выборочное (несмещенное)	=стандотклон(число1; число2; ...)	Вычисляет несмещенное стандартное отклонение значений ряда (равняется квадратному корню из дисперсии выборочной)
Нижняя квартиль (Q_1)	=квартиль(массив; часть)	Определяет значение нижней (первой) квартили в ряду данных. Чтобы задать массив, нужно указать (выделить) интересующий диапазон значений, для которого требуется определить первую квартиль. Слово «часть» заменить на «1» (это и будет означать, что рассчитывается нижняя квартиль)

Продолжение табл.

Оцениваемый показатель	Команда вызова функции в Excel. Синтаксис	Краткие пояснения
Верхняя квартиль (Q_3)	=квартиль(массив; часть)	Определяет значение верхней (третьей) квартили в ряду данных. Чтобы задать массив, нужно указать (выделить) интересующий диапазон значений, для которого требуется определить третью квартиль. Слово «часть» заменить на «3» (это и будет означать, что рассчитывается верхняя квартиль)
Ковариация	=ковар(массив1; массив2)	Рассчитывает ковариацию для имеющихся пар точек в двух выборках (рядах). Массив 1 соответствует диапазону значений по одной переменной (X). Массив 2 соответствует диапазону значений по второй переменной (Y)
Коэффициент корреляции Пирсона	= коррел(массив1; массив2)	Вычисляет коэффициент линейной корреляции Пирсона для двух множеств (X и Y). Массив 1 соответствует диапазону значений по одной переменной (X). Массив 2 соответствует диапазону значений по второй переменной (Y)
Доверительные интервалы		
Z-стандартизация	=нормализация(x; среднее; стандартное_откл) Пример: =нормализация(50; 20; 10) = 3 = нормализация(17; 25; 5) = –1,6	Осуществляет z-стандартизацию значения X на основе входной информации, задаваемой исследователем: числа X, для которого требуется вычислить z-оценку, значения среднего арифметического, значения стандартного отклонения

Предельная ошибка выборки	=доверит(альфа; станд_откл; размер) Пример: =доверит(0,05; 30; 120) = 5,37 Для доли: =доверит(0,05; 0,25; 150) = 0,04 = = 4 %	Вычисляет предельную ошибку выборки (т. е. ее точность) на основе входной информации, задаваемой исследователем: уровня значимости α, оценки стандартного отклонения s или σ, размера выборки n. Определенное таким образом значение ошибки должно быть прибавлено и отнято от среднего арифметического выборки. Это позволит получить доверительный интервал, в котором с заданной вероятностью содержится истинное значение среднего (математическое ожидание)
Функции распределений		
Критическое значение χ^2	=хи2обр(вероятность; степени_свободы) Пример: =хи2обр(0,05; 1) = 3,84	Определяет критическое значение хи-квадрат для заданных уровня значимости α и числа степеней свободы *df*
Критическое значение *F*-статистики Фишера	=Fраспобр(вероятность; степени_свободы1; степени свободы2) Пример: = Fраспобр(0,05; 5; 6) = 4,39	Определяет критическое значение *F*-статистики для заданных одностороннего уровня значимости α и чисел степеней свободы df_1 и df_2. Степени свободы 1 соответствуют дисперсии 1 (стоящей в числителе дроби), а степени свободы 2 соответствуют дисперсии 2 (стоящей в знаменателе дроби)
Критическое значение *t*-статистики Стьюдента	=Стьюдраспобр(вероятность; степени_свободы)	Определяет критическое значение *t*-статистики Стьюдента для заданных двустороннего уровня значимости α и числа степеней свободы *df*. Об-

Окончание табл.

Оцениваемый показатель	Команда вызова функции в Excel. Синтаксис	Краткие пояснения
	Пример: =Стьюдраспобр(0,05; 18) = 2,10 =Стьюдраспобр(0,1; 18) = 1,73	ратите внимание, что вычисляется значение *t*-критерия, соответствующее двусторонней критической области. Если $\alpha = 0{,}05$, то это означает, что на «хвосты» *t*-распределения приходятся вероятности 0,025 и 0,025. Поэтому если вам требуется вычислить критическое значение *t*-статистики для односторонней области для того же уровня значимости 0,05, то в синтаксисе необходимо ввести не 0,05, а $0{,}05 \cdot 2 = 0{,}10$. И будет получено значение 1,73

Примечание. Составлено А. В. Пашкевич

Приложение 5

Задачи для самоподготовки к экзамену

Задача 1. «Медицинские карточки». В ходе анализа медицинских карточек людей пожилого возраста было выяснено, сколько раз в течение полугода они посещали врачей/обращались с жалобами в поликлинику. Данные выборки о 14 пациентах представлены ниже:

X (число обращений с жалобами): **8, 25, 1, 0, 13, 10, 3, 9, 12, 14, 13, 30, 20, 9**.

Найдите:

а) медиану выборки;

б) нижнюю и верхнюю квартили выборки (по методу Дж. Тьюки), интерпретируйте нижнюю.

в) Определите по методу Дж. Тьюки диапазон характерных значений, есть ли в этой выборке нетипичные (нехарактерные) значения наблюдений — «подозрительные» и/или «выбросы».

Задача 2. «Авиапутешествия». В ходе опроса людей, которые пользуются авиаперелётами, респондентам был адресован вопрос: «Сколько авиапутешествий Вы совершили за минувший год (количество деловых поездок, поездок/туров с целью отдыха и пр.)? Под авиапутешествием понимается перелёт из точки А (точки отправления) в точку Б (точку назначения), не учитывая пересадки между рейсами. Пожалуйста, укажите число таких поездок». Ответы случайной выборки респондентов (состоящей из 15 человек) представлены ниже:

X (число поездок): **10, 9, 7, 9, 20, 7, 12, 9, 5, 4, 2, 1, 27, 12, 10**.

Найдите:

а) медиану выборки;

б) нижнюю и верхнюю квартили выборки (по методу Дж.Тьюки), интерпретируйте верхнюю.

в) Определите по методу Дж. Тьюки, есть ли в этой выборке нетипичные (нехарактерные) значения наблюдений — «подозрительные» и/или «выбросы».

Задача 3. «Радиостанции»[1]. Представленная таблица содержит данные по объёмам аудитории радиостанций в период с июля по сентябрь 2013 г. Информация получена по результатам исследования «Radio Index»:

Радиостанция	Reach Dly*
Авторадио	1 196

[1] Сюжет этой задачи предложен Светланой Чернышевой, студенткой III курса факультета социологии НИУ «Высшая школа экономики» в 2013/14 уч. г. В задаче используются результаты исследований TNS.

Окончание табл.

Радиостанция	Reach Dly*
Говорит Москва	111
Европа Плюс	1 098
Радио Шансон	1 074
Ретро FM	1 024
Эхо Москвы	1 023
Love радио	565
Радио дача	848
Юмор FM	844
Радио Romantika	279
Милицейская волна	671
Радио спорт	375

* Reach Dly – накопленное суточное количество слушателей радиостанции (тыс. чел.).

Найдите: среднее количество слушателей радиостанции (тыс. чел.), выборочную медиану, квартили (нижнюю и верхнюю), квартильный размах.

Задача 4. «Страны Латинской Америки». В ходе исследования, посвящённого изучению устойчивости политических режимов в государствах мира, было зафиксировано, как долго (т.е. сколько лет) существует нынешний политический режим в той или иной стране Латинской Америки. Случайная выборка девяти стран представлена следующими значениями: **18, 9, 12, 18, 10, 24, 21, 11, 12.**

Вычислите: *несмещённую* выборочную оценку дисперсии и стандартного отклонения; коэффициент вариации (дайте интерпретацию коэффициенту); стандартную ошибку среднего.

Задача 5. «Летние и зимние школы». Студентам старших курсов университета был задан вопрос: «В скольких студенческих конференциях и выездных школах-семинарах (летних и/или зимних) Вы приняли участие в прошлом учебном году? (укажите числом)». Случайная выборка 11 студентов представлена следующими ответами: **1, 0, 4, 0, 2, 6, 1, 2, 1, 4, 1.** Вычислите: *несмещённую* выборочную оценку дисперсии и стандартного отклонения; коэффициент вариации (дайте интерпретацию коэффициенту); стандартную ошибку среднего.

Используя све́дения этой задачи, постройте доверительный интервал, позволяющий обобщить результаты выборочного исследования на генеральную совокупность студентов университета. Уровень надёжности статистического вывода примите равным 0,90. Интерпретируйте полученный ответ.

Задача 6. «Горные лыжи». Любителям горнолыжных курортов в ходе исследования был задан вопрос: «Сколько лет Вы катаетесь на горных лыжах? (укажите в годах)». Случайная выборка десяти человек представлена следующими ответами: **5, 2, 7, 6, 4, 6, 10, 4, 9, 7.**

Вычислите: *несмещённую* выборочную оценку дисперсии и стандартного отклонения; коэффициент вариации (дайте интерпретацию коэффициенту); стандартную ошибку среднего. Используя сведения этой задачи, постройте доверительный интервал, позволяющий обобщить результаты выборочного исследования на генеральную совокупность людей, посещающих горнолыжные курорты. Уровень надёжности статистического вывода примите равным 0,95. Интерпретируйте полученный ответ.

Задача 7. «Самолёты». В ходе опроса авиапассажиров, выборка которого случайна, на вопрос исследователей о том, боитесь ли вы летать на самолёте, утвердительно ответили 25 из 146 респондентов. Постройте доверительный интервал для доли людей, которые боятся летать на самолёте. Уровень надёжности статистического вывода примите равным 0,95. Интерпретируйте полученный ответ.

Задача 8. «Страны Европы». Пусть имеется случайная выборка 31 государства Европы, для которых известно среднее значение индекса Джини за определённый год. Индекс Джини показывает уровень имущественного неравенства в стране (измеряется в шкале от 0 до 100, где 100 означает наибольшую степень имущественного неравенства, а 0 — наименьшую). Выборочная оценка среднего оказалась равной 30,5. Стандартное отклонение — 5,26.

Постройте доверительный интервал для среднего, используя уровень надёжности статистического вывода, равный 0,95. Интерпретируйте полученный ответ.

Задача 9. «Кафетерии и стоимость обеда». Некоторое маркетинговое агентство периодически публикует информацию о стоимости обедов (в расчёте на одного посетителя) в кафетериях Москвы, придерживающихся «демократичных» (доступных) цен. Для этого компания опирается на данные выборочных исследований.

Определите требуемый объём выборки исследования, если агентство стремится оценить среднюю стоимость обеда с точностью не ниже 100 усл. ед., а требуемый уровень надёжности запланированного исследования (доверие будущим результатам) составляет 90 %. Сколько ресторанов/кафе рассматриваемого сегмента следует охватить данным исследованием, если по предварительным экспертным оценкам стандартное отклонение средней стоимости обеда в них равно 340 усл. ед.?

Задача 10. «Индекс Джини и продолжительность жизни». В ходе исследования, посвящённого изучению имущественного неравенства в странах мира, выбраны два социальных показателя — индекс Джини (измеренный в шкале от 0 до 100, где 100 означает наибольшую степень имущественного неравенства, а 0 — наименьшую) и ожидаемая продолжительность жизни. В таблице представлены данные случайной выборки девяти стран.

Можно ли считать социальные признаки коррелированными (связанными)? Постройте диаграмму рассеяния. Вычислите коэффициент ранговой корреляции Спирмена между указанными двумя показателями и дайте содержательную интерпретацию. Является ли полученное значение коэффициента корреляции статистически значимым? Проверьте соответствующую статистическую гипотезу на уровне доверия 0,90. Конкурирующую гипотезу сформулируйте как направленную (исходя из имеющихся данных).

Страна	А	Б	В	Г	Д	Е	Ж	З	И
Индекс Джини (шкала от 0 до 100)	32	38	52	34	29	30	34	55	43
Ожидаемая продолжительность жизни, лет	79	67	70	62	74	67	71	63	68

Задача 11. «Преступность и миграция». В ходе исследования, посвящённого изучению российских регионов, выбраны два социальных показателя — уровень преступности (как число преступлений, приходящееся на 100 тыс. жителей региона в 2011 г.) и число мигрантов, прибывших в регион из-за пределов Российской Федерации (% от общего числа прибывших в регион в 2011 г.). В таблице представлены данные случайной выборки десяти регионов.

Постройте диаграмму рассеяния. Можно ли считать эти социальные признаки коррелированными (связанными)? Вычислите коэффициент ранговой корреляции Спирмена между двумя показателями и дайте содержательную интерпретацию. Проверьте статистическую значимость полученного значения коэффициента. Уровень доверительной вероятности примите равным 0,95. Конкурирующую гипотезу сформулируйте как направленную (исходя из имеющихся данных).

Регион	№1	№2	№3	№4	№5	№6	№7	№8	№9	№10
Число преступлений	1 628	1 247	1 538	1 683	1 441	1 302	2 216	2 372	1 511	1 662
Прибывших из-за пределов РФ	8,8	24,5	9,7	3,5	5,4	10,1	9	2,3	12,2	30,4

Задача 12. «Безработица и продолжительность жизни». Исследователей интересовало, являются ли взаимосвязанными такие социальные показатели, как уровень безработицы (в %) в различных государствах мира и ожидаемая продолжительность жизни в этих странах (показатели взяты за определённый год). Получены следующие данные:

Страна	1	2	3	4	5	6
Ожидаемая продолжительность жизни, лет	65	73	70	82	60	76
Безработица, %	7	2	3	2	6	10

Постройте диаграмму рассеяния. Рассчитайте коэффициент линейной корреляции Пирсона и дайте содержательную интерпретацию полученного значения. Проверьте статистическую значимость полученного значения коэффициента на уровне доверия 0,95. Конкурирующую гипотезу сформулируйте как ненаправленную.

Задача 13. «Фитнес-клуб». Предположим, что средний возраст посетителей (клиентов) некоторого фитнес-клуба в городе составляет 36 лет. Дисперсия равна 26. Какова вероятность того, что возраст посетителя (клиента), выбранного случайно, окажется равным от 34 до 43 лет? (Если считать, что возраст посетителей данного фитнес-клуба описывается нормальным законом распределения с указанными параметрами.) *Схематично отразите искомую вероятность на графике нормального распределения и на графике стандартного нормального распределения.*

Задача 14. «Кадровые агентства». В ходе исследования изучалась эффективность работы кадровых агентств по подбору персонала для бизнес-клиентов. Так оценивалась средняя продолжительность времени по подбору персонала (количество дней, в среднем затрачиваемых на подбор подходящей кандидатуры на соответствующую вакансию/должность). Подготовлены *случайная* выборка из семи наблюдений (вакансий) в агентстве № 1 и выборка из шести наблюдений (вакансий) в агентстве № 2.

Агентство	Среднее количество дней, затрачиваемых на подбор соответствующей кандидатуры на вакансию/должность, дней	Стандартное отклонение, дней
№ 1	8	2,5
№ 2	12	2,2

Существует ли значимое различие между рассматриваемыми кадровыми агентствами с точки зрения средней скорости подбора персонала? Проверку соответствующей статистической гипотезы осуществите на 95%-м уровне доверительной вероятности против *ненаправленной* альтернативы.

Задача 15. «Семейный бюджет на продукты питания»[1]. В ходе некоторого исследования молодых семейных пар (выборка *случайна*) изучались структура и характер расходов денежных средств на продукты питания. Были опрошены два типа супружеских пар: *первую группу* составили семьи, в которых работают оба супруга (45 супружеских пар в выборке), а *вторую группу* — те, в которых работает только один из супругов (таких в выборке оказалось 27 семейных пар).

В результате было выявлено, что в тех семьях, где работают и муж, и жена, на продукты питания в течение месяца расходуется в среднем 12 350 руб. (в расчёте на одного человека). В семьях, где работает только один супруг, такие расходы составляют 9 600 руб. (в расчёте на человека).

Выборочная оценка стандартного отклонения в первом случае составляет 4 100 руб., а во втором — 3 000 руб.

Справедливо ли утверждение, что семейные пары, в которых работает только один из супругов, склонны расходовать меньшую сумму денег на продукты питания (т.е. жить более экономно в этом аспекте)? Проверку соответствующей гипотезы реализуйте на уровне доверительной вероятности 95 % против *направленной* конкурирующей гипотезы.

Задача 16. «Монархии и республики». Имеется выборка восьми государств Европы, являющихся монархиями, для которых зафиксировано среднее значение ожидаемой продолжительности жизни 78,4 лет, выборочная оценка стандартного отклонения — 0,92 лет. Имеется выборка 34 государств Европы, являющихся республиками, для которых среднее значение ожидаемой продолжительности жизни составило 73,8 лет, выборочная оценка стандартного отклонения — 4,11 лет.

Ответьте на вопрос о том, различается ли средняя продолжительность жизни в странах Европы в зависимости от формы государственного правления. Проверку соответствующей гипотезы осуществите на уровне доверительной вероятности 0,95 против *направленной* альтернативы (исходя из имеющихся данных).

Задача 17. «Бюджет времени на отдых»[2]. В ходе небольшого исследования, посвящённого теме бюджета времени (выборка которого случайна), в 2012 г. приняли участие пять девушек — студенток II курса (которые в течение определённого периода заполняли специальные дневники, фиксируя информацию о том, сколько времени они выделяют на те или иные виды деятельности). В итоге было зафиксировано, сколько обычно часов в неделю они расходуют на отдых и развлечения.

Спустя год, в 2013 г. исследование повторили (с тем же составом участников и по той же схеме). Были получены следующие результаты (в ячейках указаны данные о еженедельных расходах времени на отдых и развлечения):

---

[1] Сюжет этой задачи предложен Полиной Поповой, студенткой III курса факультета социологии НИУ «Высшая школа экономики» в 2013/14 уч. г.

[2] Сюжет этой задачи предложен Ксенией Сырчиковой, студенткой III курса факультета социологии НИУ «Высшая школа экономики» в 2013/14 уч. г.

Бюджет времени на отдых, ч	Катерина	Ольга	Алла	Луиза	Марина
Данные 2012 г. (когда учились на II курсе)	30	37	45	34	19
Данные 2013 г. (когда перешли на III курс)	21	35	40	34	20

Можно ли сказать (при имеющихся данных), что с возрастом (по мере перехода на более старший курс обучения в университете) девушки склонны в меньшей степени тратить время на отдых и развлечения? Проверку соответствующей статистической гипотезы реализуйте на уровне доверительной вероятности 0,95 против *направленной* альтернативной гипотезы.

Задача 18. «Уровень жизни». В рамках исследования сравнивался уровень жизни людей в разных населённых пунктах. В качестве одного из показателей уровня жизни рассматривался среднедушевой ежемесячный доход жителей соответствующих поселений. Имеется следующая выборочная статистика:

Поселение	Объем выборки (количество респондентов)	Среднее, усл. ед.	Стандартное отклонение, усл.ед.
№ 1	350	10 250	2 990
№ 2	410	12 430	4 030

Можно ли утверждать, что исследуемые населённые пункты не отличаются друг от друга с точки зрения среднего показателя уровня жизни. Проверку соответствующей статистической гипотезы осуществите против *ненаправленной* альтернативы (исходя из данных). Уровень доверительной вероятности примите равным а) 0,95; б) 0,99.

Задача 19. «Кулинария». Могут ли домашние хозяйки сэкономить семейный бюджет, покупая популярные книги по кулинарии в интернет-магазине? Для ответа на этот вопрос была сформирована случайная выборка из десяти книг, содержащих кулинарные рецепты. В таблице приведены средние цены на эти книги в обычных магазинах и средние цены в интернет-магазинах (с учётом доставки):

Кулинарная книга	Стоимость, усл. ед.	
	обычный магазин	интернет-магазин
А	550	495
Б	371	350
В	100	106

Окончание табл.

Кулинарная книга	Стоимость, усл. ед.	
	обычный магазин	интернет-магазин
Г	367	268
Д	120	161
Е	1 371	963
Ж	1 200	1 095
З	147	160
И	350	342
К	269	185

Существуют ли основания утверждать, что средняя стоимость кулинарных книг в обычном магазине и интернет-магазине отличаются друг от друга? Считая выборки связными (парными), осуществите проверку основной статистической гипотезы на уровне надёжности 0,99 против направленной конкурирующей гипотезы: «Средняя стоимость покупки книги в обычном магазине выше».

Задача 20. «Против смертной казни». Против введения смертной казни за особо тяжкие преступления высказались 40 % мужчин (из опрошенных 150) и 56 % женщин (из 100 опрошенных). Обе выборки случайные.

Определите, существуют ли при имеющихся данных основания полагать, что мужчины в вопросе наказания проявляют жёсткость и безаппеляционность в большей мере, нежели женщины. Уровень надежности статистического вывода примите равным 0,95.

Задача 21. «Глобальное потепление». Из 120 учёных, опрошенных в ходе научных конференций, 21 % ответили, что испытывают острое чувство тревоги по поводу глобального потепления. А из 243 человек, не имеющих прямого отношения к науке, в ходе аналогичного опроса такой ответ дали 15 %. Обе выборки случайны.

Существует ли различие в долях на уровне значимости $\alpha = 0{,}05$? Проверку статистической гипотезы реализуйте против направленной конкурирущей гипотезы (исходя из данных).

Задача 22. «Превышение скорости на дорогах». В выборке, состоящей из 200 мужчин-автомобилистов, 64 человека сказали, что испытывают чувство стыда за превышение скорости. В выборке, состоящей из 300 женщин-автомобилистов, 129 человек отметили, что испытывают чувство стыда за превышение скорости.

На уровне надежности 0,99 проверьте утверждение, что мужчины и женщины одинаково «совестливы» в этом аспекте поведения на дорогах. Проверку нулевой статистической гипотезы реализуйте против направленной конкурирущей гипотезы (исходя из данных).

СПИСОК ЛИТЕРАТУРЫ

Учебники-референты по дисциплинам «Теория вероятностей и математическая статистика», «Методология и методы социологических исследований», которые рекомендуются студентам для изучения:

1. *Agresti A.* Statistical Methods for the Social Sciences / A. Agresti, B. Finley. — 4th ed. — Pearson Prentice Hall, 2009.

2. *Bluman A.* Elementary statistics. A Step by Step Approach. McGraw-Hill, 2008.

3. *Девятко И. Ф.* Методы социологического исследования: учеб. пособие. — 5-е изд. испр. и доп. — М.: Книжный дом «Университет», 2009.

4. *Иванов О. В.* Статистика. Ч. 1. Описательная статистка. Теоретико-вероятностные основания статистического вывода. — М., 2005.

5. *Иванов О. В.* Статистика: Учебный курс для социологов и менеджеров. Ч. 2. Доверительные интервалы. Проверка гипотез. Методы и их применение. — М., 2005.

6. *Крыштановский А. О.* Анализ социологических данных с помощью пакета SPSS: учеб. пособие для вузов. — М.: Издат. дом ГУ-ВШЭ, 2007.

7. *Математическая* статистика для социологов. Задачник. — М.: Издат. дом ГУ-ВШЭ, 2010.

8. *Татарова Г. Г.* Методология анализа данных в социологии: учебник для вузов. — М.: NOTA BENE, 1999.

9. *Толстова Ю. Н.* Анализ социологических данных: методология, дескриптивная статистика, изучение связей между номинальными признаками: учеб. пособие. — М.: Научный мир, 2000.

10. *Толстова Ю. Н.* Математико-статистические модели в социологии: учеб. пособие. — М.: Издат. дом ГУ-ВШЭ, 2007.

11. *Тюрин Ю. Н.* Анализ данных на компьютере: учебное пособие / Ю.Н. Тюрин, А.А. Макаров. — 4-е изд., перераб. — М.: Издат. дом «ФОРУМ», 2008.

12. *Тюрин Ю. Н.* Теория вероятностей: учебник для экономических и гуманитарных специальностей / Ю.Н. Тюрин, А.А. Макаров, Г. И. Симонова. — М.: МЦНМО, 2009.

13. *Чуриков А. В.* Основы проектирования выборки. — М.: Издат. дом ГУ-ВШЭ, 2005.

Базы данных, информационно-справочные и поисковые системы:

1. Единый архив экономических и социологических данных: <http://sophist.hse.ru/>

2. Информационный портал Фонда «Общественное мнение»: <http://fom.ru/>

3. Информационный портал Всероссийского центра изучения общественного мнения: <http://wciom.ru/>

4. Информационный портал Федеральной службы государственной статистики: <http://gks.ru>

5. Архив баз данных Всемирного исследования ценностей (World Values Survey): <http://worldvaluessurvey.org>

6. Федеральный информационный портал «Экономика. Социология. Менеджмент» (Разделы: Книги и статьи; Учебные программы; Журнальный зал).

ОГЛАВЛЕНИЕ

Предисловие автора 3
Предисловие научного редактора 5

Глава 1. Введение в математическую статистику 8

§ 1.1. История становления математической социологии 8
§ 1.2. Объект и предмет математической статистики 11
§ 1.3. Выборка и генеральная совокупность 15
§ 1.4. Выборка: немного истории 25

Глава 2. Методы описательной статистики 29

§ 2.1. Вариационный ряд. Графическое представление данных 29
§ 2.2. Меры центральной тенденции 38
§ 2.3. Выборочные квартúли и децúли 49
§ 2.4. Выборочные меры разброса (вариации) 54
§ 2.5. Коробчатая диаграмма («Ящик с усами») 64

Глава 3. Коэффициенты связи 77

§ 3.1. Шкалы 77
§ 3.2. Коэффициент линейной корреляции Пирсона 83
§ 3.3. Коэффициент ранговой корреляции Спирмена 93

Глава 4. Нормальное распределение и стандартное нормальное *z*-распределение 108

§ 4.1. Нормальное распределение (распределение Гаусса) 108
§ 4.2. Процедура стандартизации. Стандартное нормальное *z*-распределение 111

Глава 5. Точечное оценивание параметров и построение доверительных интервалов 124

§ 5.1. Выборка и понятие несмещённой выборочной оценки. Несмещённая оценка дисперсии 124
§ 5.2. Точечная оценка и доверительный интервал. Ошибки выборки: стандартная и предельная 131

§ 5.3. Построение доверительного интервала для среднего (математического ожидания) 138
§ 5.4. Построение доверительного интервала для доли (альтернативного признака) 143
§ 5.5. Расчёт объёма выборки в исследовании 151
§ 5.6. Распределение Стьюдента (*t*-распределение). Число степеней свободы 155

Глава 6. Проверка статистических гипотез 177

§ 6.1. Основные понятия проверки статистических гипотез. Глоссарий 177
§ 6.2. Анализ таблиц сопряжённости. Критерий хи-квадрат 197
§ 6.3. Проверка гипотезы о равенстве дисперсий; *F*-критерий Фишера 215
§ 6.4. Проверка гипотезы о равенстве средних. Независимые выборки и *t*-критерий Стьюдента 226
§ 6.5. Проверка гипотезы об отсутствии среднего эффекта воздействия для связных (парных) выборок; *t*-критерий Стьюдента 250
§ 6.6. Проверка гипотезы о равенстве долей; *z*-критерий 257
§ 6.7. Проверка гипотезы о статистической незначимости коэффициентов корреляции Пирсона и Спирмена 264

Приложения 288
Приложение 1. Образцы контрольной экзаменационной работы 288
Приложение 2. Статистические таблицы распределений 305
Приложение 3. Греческий алфавит 315
Приложение 4. Синтаксис статистических функций в Microsoft Excel 316
Приложение 5. Задачи для самоподготовки к экзамену 321
Список литературы 329

Учебное издание

Пашкевич Анна Валерьевна

Теория вероятностей и математическая статистика для социологов и менеджеров

Учебник

Редактор *Л. В. Честная*
Технический редактор *Н. Л. Ананьева*
Компьютерная верстка: *Г. Ю. Никитина*
Корректор *Г. Н. Петрова*

Изд. № 101115943. Подписано в печать 25.06.2014. Формат 60×90/16.
Гарнитура «Ньютон». Печать офсетная. Бумага офс. № 1. Усл. печ. л. 21,0.
Тираж 700 экз. Заказ №

ООО «Издательский центр «Академия». www.academia-moscow.ru
129085, Москва, пр-т Мира, 101 В, стр. 1.
Тел./факс: (495)648-0507, 616-0029.
Санитарно-эпидемиологическое заключение № РОСС RU. АЕ51. Н 16592 от 29.04.2014.

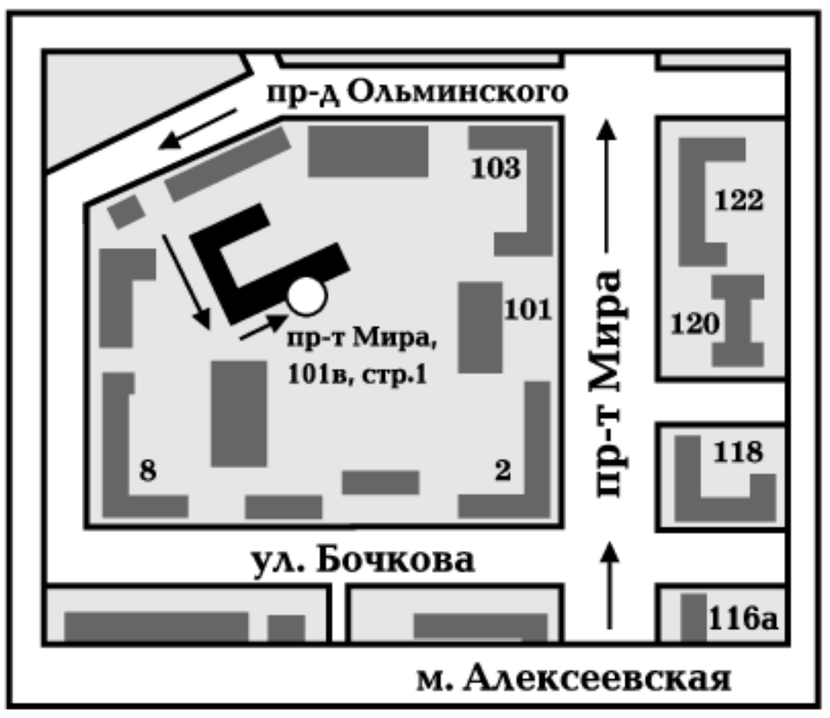
Отпечатано с электронных носителей, предоставленных издательством,
в ОАО «Первая образцовая типография».

Издательский центр «Академия»

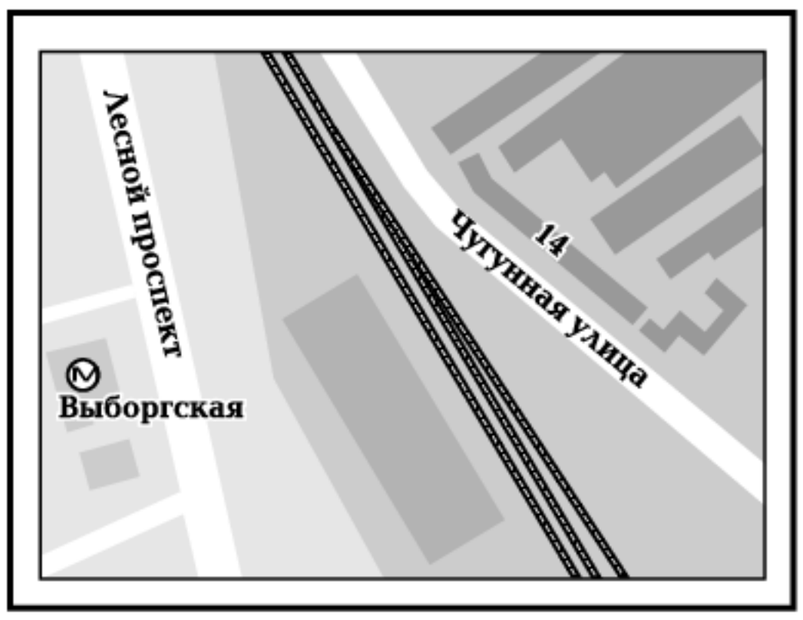
Учебная литература для профессионального образования

Наши книги можно приобрести (оптом и в розницу)

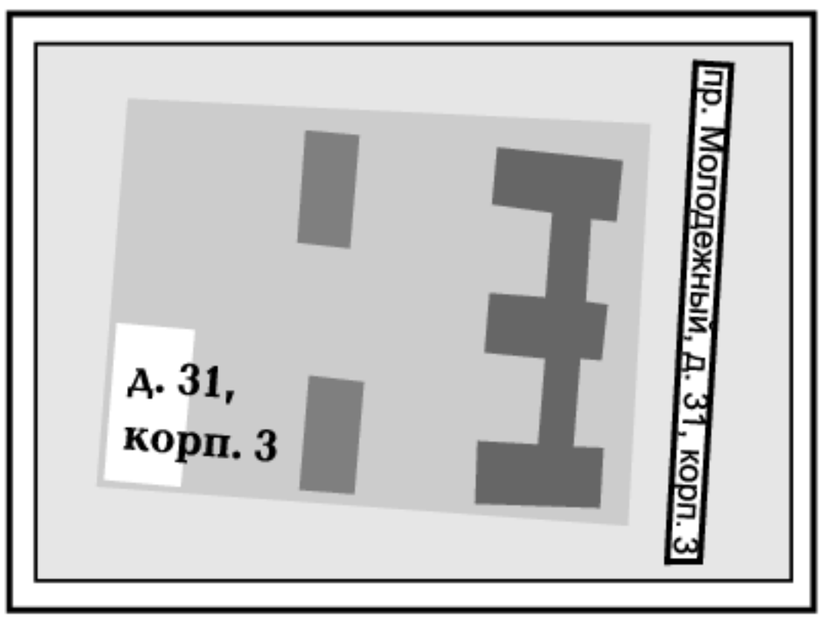
Москва
129085, Москва, пр-т Мира, д.101в, стр. 1
(м. Алексеевская)
Тел.: (495) 648-0507, факс: (495) 616-0029
E-mail: sale@academia-moscow.ru



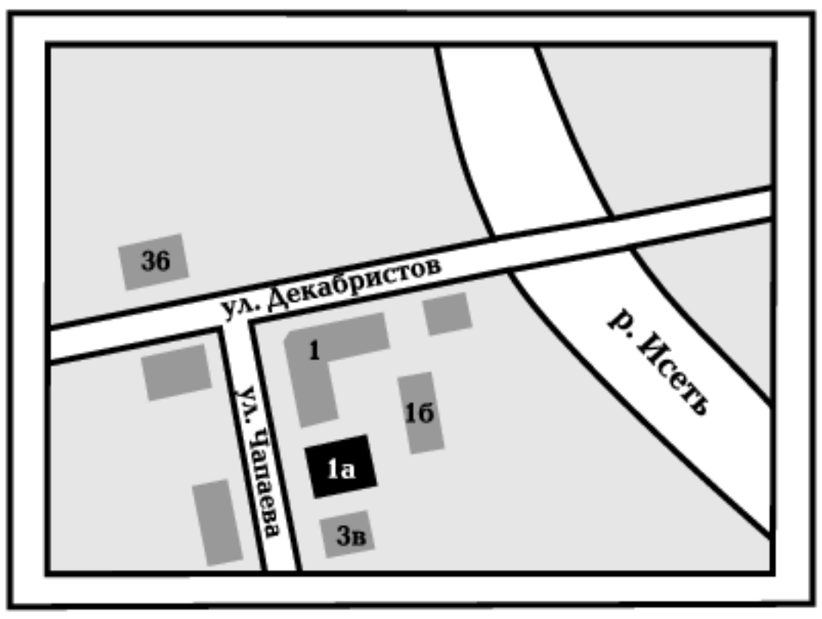
Филиалы:



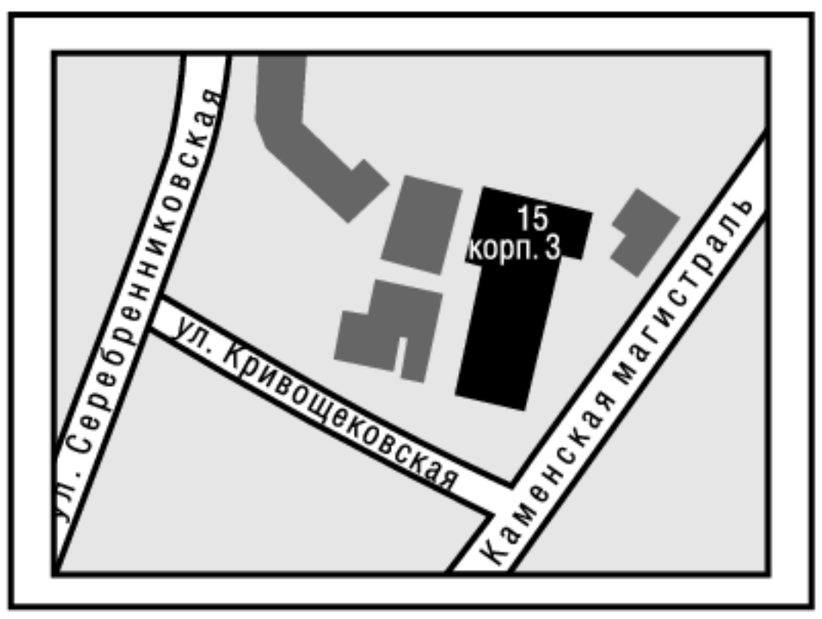
Северо-Западный
194044, Санкт-Петербург,
ул. Чугунная, д. 14, оф. 319
Тел./факс: (812) 244-92-53
E-mail: spboffice@acadizdat.ru



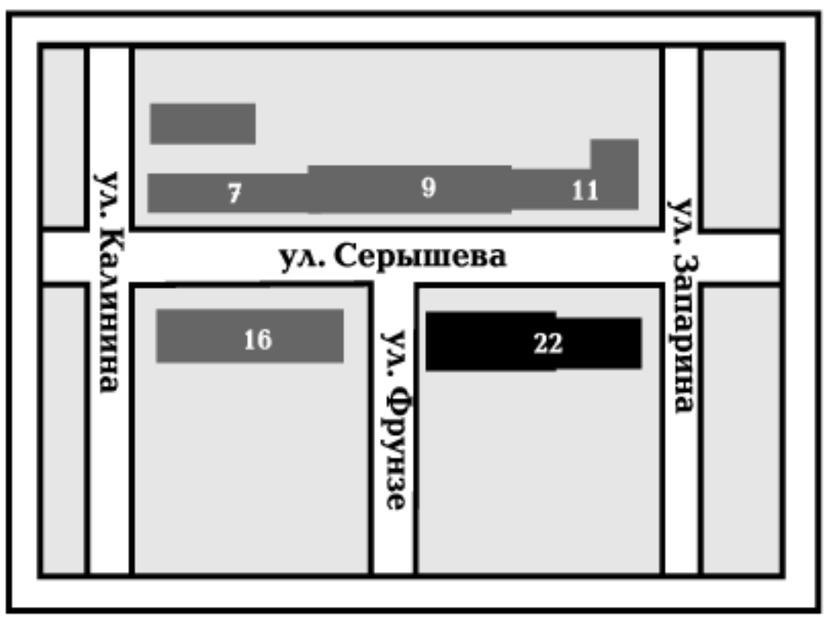
Приволжский
603101, Нижний Новгород,
пр. Молодежный, д. 31, корп. 3
Тел./факс: (831) 259-7431,
259-7432, 259-7433
E-mail: pf-academia@bk.ru



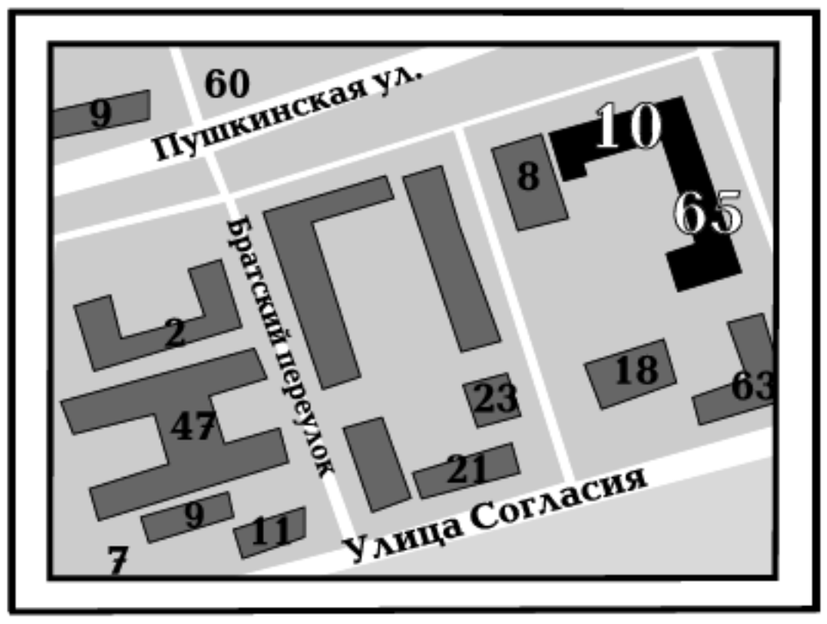
Уральский
620142, Екатеринбург, ул. Чапаева,
д. 1а, оф. 12а
Тел.: (343) 257-1006
Факс: (343) 257-3473
E-mail: academia-ural@mail.ru



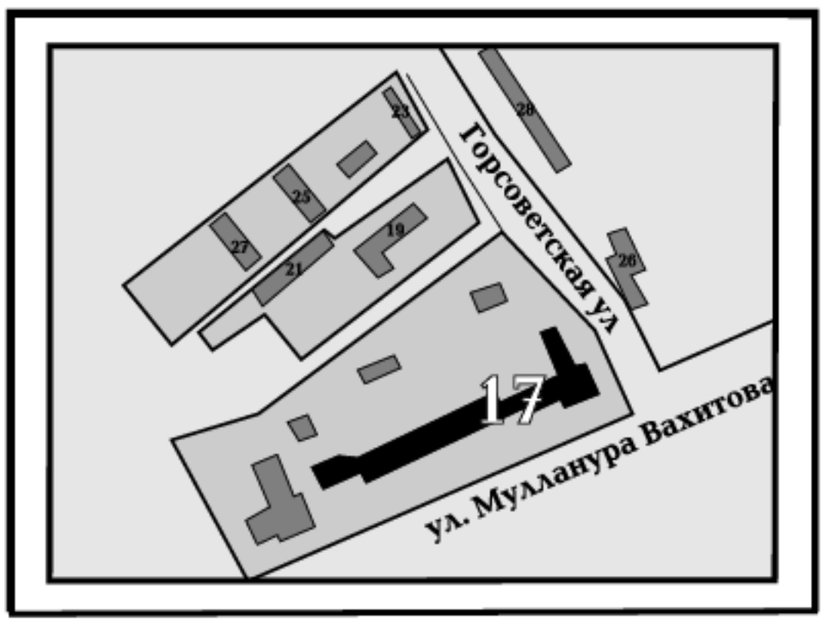
Сибирский
630007, Новосибирск,
ул. Кривощёковская, д. 15, корп. 3,
Тел./факс: (383) 362-2145, 362-2146
E-mail: academia_sibir@mail.ru



Дальневосточный
680038, Хабаровск, ул. Серышева,
д. 22, оф. 519, 520, 523
Тел./факс: (4212) 56-8810
E-mail: filialdv-academia@yandex.ru



Южный
344082, Ростов-на-Дону,
ул. Пушкинская, д.10/65
Тел.: (863) 203-5512
Факс: (863) 269-5365
E-mail: academia-UG@mail.ru



Представительства:

в Республике Татарстан
420034, Казань, ул. Горсоветская,
д. 17/1, офис 36
Тел./факс: (843) 562-1045
E-mail: academia-kazan@mail.ru

в Республике Дагестан
Тел.: 8-928-982-9248

www. academia-moscow. ru

Предлагаем вашему вниманию следующие книги

И. Н. СЕРГЕЕВ

ДИФФЕРЕНЦИАЛЬНЫЕ УРАВНЕНИЯ

Объем 288 с.

Учебник создан в соответствии с Федеральным государственным образовательным стандартом по направлениям подготовки «Математика», «Математика и компьютерные науки», «Механика и математическое моделирование», «Прикладная математика и информатика», «Фундаментальная информатика и информационные технологии» (квалификация «бакалавр»).

Материал учебника знакомит с геометрической интерпретацией уравнения первого порядка, с первыми интегралами, особыми точками и предельными циклами автономных систем, с теорией линейных уравнений и систем, в том числе с постоянными и периодическими коэффициентами, с вопросами существования, единственности и продолжаемости решений, их непрерывности и дифференцируемости по параметру, устойчивости по Ляпунову, а также с вопросами существования и единственности решения задачи Коши для уравнения с частными производными первого порядка. Даны точные определения, аккуратно сформулированы и доказаны утверждения, строго обоснованы наиболее важные методы решения задач. Приведены все необходимые теоретические сведения, сопутствующие понятия и факты из смежных разделов математики. Предложены задачи для самостоятельного решения, позволяющие глубже проникнуть в прочитанный материал.

Для студентов учреждений высшего профессионального образования.

В. И. СТЕПАНОВ, А. Ф. ТЕРПУГОВ

ЭКОНОМИКО-МАТЕМАТИЧЕСКОЕ МОДЕЛИРОВАНИЕ

Объем 112 с.

Учебное пособие посвящено математическим моделям микроэкономики. В нем рассматриваются основы теории спроса, элементы теории ценообразования; основы теории фирмы, конкурентное равновесие на микроэкономическом уровне, модель межотраслевого баланса Леонтьева, динамическая модель экономики Неймана, теорема Моришимы о магистрали.

Для студентов учреждений высшего профессионального образования.

www.academia-moscow.ru