Борисова Е.В.

• • • • •

# ПРИКЛАДНЫЕ СТАТИСТИЧЕСКИЕ МОДЕЛИ И МЕТОДЫ В СОЦИОЛОГИИ

## (уровень бакалавриата)

*Учебное пособие*
*Издание первое*

Издательство «АНАЛИТИКА РОДИС»
Московская обл., г. Ногинск
2016

УДК 316
ББК 60.5
Б82

Рецензенты:

Солодова Е.А. – профессор 1 кафедры (военной акмеологии и кибернетики) Военной академии РВСН имени Петра Великого, заслуженный работник высшей школы, доктор педагогических наук, профессор;

Евстифеева Е.А. – заведующая кафедрой философии и психологии с курсами биоэтики и истории отечества Тверского государственного медицинского университета, доктор философских наук, профессор.

Б82 Борисова Е.В. Прикладные статистические модели и методы в социологии (уровень бакалавриата): Учебное пособие. 1-е изд. – Московская обл., Ногинск: АНАЛИТИКА РОДИС, 2016. – 254 с.

В пособии рассмотрены вопросы использования статистических методов прикладного исследования в области социологии. Особое внимание уделено базовым моделям статистического анализа данных, представленных в различных типах шкал измерений и содержательной интерпретации полученных результатов.

Предназначено для обучающихся по направлению подготовки бакалавров 39.03.01 «Социология», профиля – общая социология, а также студентов других гуманитарных направлений, занимающихся сбором и обработкой статистических данных. Может быть использовано для самообразования.

ISBN 978-5-905277-73-3

ББК 60.5

# Содержание

# Предисловие

*Математика – это язык, на*
*котором написана книга природы*
*Галилео Галилей*

С 2011 года в Российской Федерации введено двухуровневое высшее профессиональное образование. Первый его уровень – бакалавриат. Нормативный срок по очной форме обучения для получения степени «бакалавр» составляет четыре года. Нововведение связано с тем, что в 2003 году Россия присоединилась к Болонскому процессу, направленному на формирование единого европейского образовательного пространства. В настоящее время, когда технологии и знания обновляются очень быстро, нецелесообразно готовить «узких» специалистов в стенах вуза, начиная с первого курса, в течение пяти-шести лет. Поэтому введение широкой бакалаврской программы с последующей специализацией в магистратуре или на производстве больше соответствует быстро меняющемуся рынку труда. Студенты, которые учатся по программе бакалавриата, получают обычное высшее образование, просто иным путем, в более сокращенные сроки. Уровневое высшее профессиональное образование позволяет на первых курсах обучения по широкому спектру осознанно выбрать профиль программы, который обычно реализуется на старших курсах. Выбор профиля это важное личное решение, которое дает возможность после изучения общих профессиональных дисциплин изучить профильные дисциплины, что позволит развить компетенции, а так же

проявить способности в понимании дальнейшей профессиональной деятельности, начиная с элементарного уровня и до самого сложного.

В соответствии с федеральным государственным образовательным стандартом высшего образования по направлению подготовки бакалавров – 39.03.01 «Социология», профиля – общая социология и видам деятельности – организационно-управленческая, научно-исследовательская – определено овладение основными методами прикладной статистики, наиболее востребованными и интенсивно применяемыми в социологии. Дисциплина «Методы прикладной статистики для социологов» относится к вариативной части Б1 ОПВО – дисциплины (модули) основной профессиональной образовательной программы (бакалавриат). Изучение направлено на приобретение навыков обработки и анализа основных статистических показателей, умения использовать их в профессиональной деятельности. Базируется на результатах изучения курсов «Математика», «Информатика», «Введение в социологию», «Теория вероятностей и математическая статистика», «Методология и методика социологических исследований», «Количественный метод социологических исследований». Является интегративной, обеспечивающей взаимосвязь естественнонаучных и социальных дисциплин. Приобретенные знания и владения необходимы в дальнейшем при изучении курсов «Анализ данных в социологии», «Социологический практикум» и других специальных дисциплин. Выполнению заданий по курсовому проектированию и учебной практики, подготовки выпускной квалификационной работы.

*Предметная область дисциплины* включает формирование научного представления о прикладных статистических методах исследования социальных явлений и способность использовать в профессиональной деятельности приобретенную совокупность знаний, умений и навыков. Формирования характера мышления и ценностных ориентаций, при которых вопросы использования математических методов для подготовки аналитических решений, экспертных заключения и рекомендаций, рассматриваются в качестве приоритета.

*Объектами изучения* дисциплины являются основы современных методов прикладной статистики используемых в социологии и возможностями их практического приложения для анализа данных используемых в эмпирических социологических исследованиях, база методов измерения в социологии для их адекватного использования в профессиональной деятельности.

*Основной целью* изучения является формирование комплекса знаний и владений применения основных статистических процедур, универсальный характер которых обеспечивает их успешное применение в различных предметных областях, в том числе для подготовки экспертных заключений и рекомендаций в профессиональной сфере.

*Задачами* дисциплины являются:

формирование представления о принципах изучения массовых явлений, об изменениях в социальных процессах;

изучение методов построения и анализа основных статистических показателей и умения использовать в профессиональной деятельности основные методы обработки и анализа статистических данных;

формирование умений содержательно интерпретировать полученные результаты.

*Планируемые результаты обучения*. Формирование и развитие общепрофессиональной компетенции: способность использовать основные законы естественнонаучных дисциплин в профессиональной деятельности, применять методы математического анализа и моделирования, теоретического и экспериментального исследования. Данная компетенция в своей основе содержит:

знания методов измерения данных в социологии, основ статистики экспериментальных данных, методов статистического оценивания и проверки гипотез; методов прикладной статистики для анализа и моделирования социальных явлений и процессов;

умения обрабатывать эмпирические данные; использовать математический язык и математическую символику при построении моделей; осваивать и анализировать информацию о тенденциях развития прикладной статистики в социологических исследованиях

владение статистическими и количественными методами решения типовых задач.

Указанная компетенция и сопровождающие знания, умения и владения формируются в процессе: проведения лекционных занятий, практических занятий в интерактивных формах; выполнении практических расчетных работ, в том числе с использование пакетов прикладных программ. Последовательный контроль уровня формирования компетенции выполняется с целью поддержания и распространения положительных результа-

тов, а при негативных проявлениях их своевременного распознавания и устранения причин этих изменений. Анализ факторов влияющих на уровень формируемой компетенции определяет воздействующие мероприятии (индивидуальные консультации, личностно-ориентированные задания, самостоятельное изучение отдельных вопросов дисциплины и подготовка презентации с последующим коллективным обсуждением в студенческих группах).

Самостоятельная работа заключается в изучении отдельных тем курса по заданию преподавателя по рекомендуемой учебной литературе, в подготовке к практическим занятиям, к текущему контролю успеваемости. В рамках изучения дисциплины выполняются индивидуальные задания, которые защищаются по окончании изучения отдельных разделов модуля. Формирование способностей к самостоятельному познанию и обучению, поиску литературы, обобщению, оформлению и представлению полученных результатов, их критическому анализу, поиску новых и неординарных решений, аргументированному отстаиванию своих предложений, умений подготовки выступлений и ведения дискуссий является залогом высокого уровня требуемой компетенции.

Жизнь человека, общества, цивилизации складывается из случайных явлений. Чтобы события были предсказуемыми, важно научиться оценивать, анализировать и прогнозировать случайности. Принятие решения о том, как следует использовать полученные знания – процесс, носящий субъективный характер. Современные задачи планирования, управления, прогнозирования социальных процессов невозможно решать, не распо-

лагая достоверными статистическими данными. Грамотно используемые статистические методы существенно расширяют возможности научного исследования. Теория в узко научном плане является высшей формой организации научного знания. Вместе с тем важно подчеркнуть, что в социологических исследованиях, прежде всего речь идет о теории описательного типа, главным образом решающей задачи упорядочения и описания эмпирического материала. Рост математической культуры специалистов в соответствующих областях приводит к тому, что изучение методов вычислений уже не встречает серьезных трудностей. Вместе с тем на практике оказывается, что одних математических познаний недостаточно для решения той или иной прикладной задачи – необходимо получить навыки в переводе исходной формулировки задачи на математический язык.

Бакалавриат, как вид квалификации, отвечает международным стандартам и понятен работодателям за рубежом. В Европу часто приглашают бакалавров, даже не уточная специальность по диплому, поскольку нужен просто образованный человек, умеющий работать с информацией, с людьми, способный готовить всевозможные документы. Программа подготовки бакалавра построена таким образом, что позволяет, при необходимости, легко сменить профессию. В соответствии с государственным образовательным стандартом, программы подготовки бакалавров по направлениям построены так, что позволяют за год перейти к одной из целого «веера» совместимых профессий. Данные обстоятельства определяют цели, задачи и педагогические технологии образовательного процесса на уровне бакалавриата.

Авторский курс по дисциплине «Методы прикладной статистики для социологов» адаптирован под требования подготовки бакалавров, согласно нормативно-методическим документам Министерства образования и науки Российский Федерации.

Основной целью учебного пособия является ознакомление студентов с базовыми методами статистической обработки и анализа данных, встречающихся в социологических исследованиях и представленных в различных типах шкал измерений. Особое внимание уделено содержательной интерпретации полученных результатов.

# Введение

Понимание содержания учебного пособия предполагает наличие у студента знания элементов высшей математики, теории вероятностей и математической статистики, методики проведения социологического исследования, в объеме программ, соответствующих обязательным курсам, читаемым студентам-социологам.

В данном учебном пособии преследуется две цели: изложить овладевающему профессией социолога студенту основы и приемы измерительных процедур и обработки данных в прикладной социологии и приобщить его к современному состоянию теории предмета. Поэтому в каждом разделе рассматриваются и теоретико-методологические проблемы социологического измерения, и практические техники соответствующих подходов. Немаловажный аспект – совмещение формализма измерения и обработки данных с пониманием сложности социального объекта. Здесь требуется и знание, и воображение, без которого невозможна профессия социолога. Как это совместить? Приходится искать оптимальное решение.

Подчеркнем, что измерение – это всегда моделирование. Цель такого моделирования – обеспечение возможности использования математики для решения социологических задач. Предполагаем, что студенты второго курса имеют представление об эмпирическом социологическом исследовании; знают, что такое анкета, из чего она обычно состоит; слышали об использовании в социологии шкал разных типов, признают наличие проблемы

интерпретации данных. В понятии интерпретации данных по существу отражается «стыковка» эмпирического и теоретического. В это понятие каждый исследователь включает нечто свое, определяемое его априорным видением изучаемых объектов и явлений. Тем не менее, существуют некоторые такие аспекты интерпретации, которые фигурируют практически в любом исследовании. Можно сказать, что *интерпретация данных* – это наше видение того, что за этими данными стоит, наше понимание смысла чисел полученных в результате измерения. Например, ранжируя объекты, респонденты руководствуются каким-то набором характеристик последних (одним и тем же для всех респондентов), т. е. мыслят эти объекты в виде точек некоторого признакового пространства. У каждого респондента имеется некоторое представление об «идеальном» объекте. И один объект кажется ему лучше, чем другой, если первый – ближе к этому идеальному объекту.

Ряд не формализуемых моментов интерпретации данных связан с определением типа используемых шкал. Имеется в виду не то, что, скажем, в «числах», полученных по номинальной шкале, видно лишь их сходство и различие, и не замечается того, что одни из них больше, другие меньше и т. д.

Тип фактически использующихся шкал тесно связан с показателем или признаком. Значит, анализ какого-либо рассматриваемого признака тоже следует включить в процесс интерпретации исходных данных. В социологии нельзя строить эффективные алгоритмы измерения, не имея в голове плана, включающего в себя все те моменты, о которых говорилось выше, т. е. не определяя принципы интерпретации результатов измерения.

В пособии большое внимание уделено проблемам измерения. Книги, рассчитанные на технарей, естественно, не учитывают типов тех шкал, по которым получаются значения рассматриваемых случайных величин. В западных учебниках, рассчитанных на читателя-гуманитария, коротко говорится о том, что исходные данные могут быть получены по номинальным, порядковым и интервальным шкалам, а выбор метода анализа данных зависит от типа данных. Круг рассматриваемых методов анализа данных при этом ограничивается самыми простейшими операциями типа вычисления показателей средней тенденции. Но от студента требуется знакомство с понятием допустимых преобразований шкалы. Это дает возможность четкого понимания того, почему один метод пригоден для той или иной шкалы, а другой – нет (имеется в виду определение формальной адекватности метода). Студент, имеющий соответствующие представления, может творчески подходить к решению вопроса о выборе метода в конкретной ситуации. Более того, процесс определения типа шкалы в социологическом исследовании зачастую не поддается формализации и определяется самим исследователем. Признаки (случайные величины) зачастую интересуют социолога не сами по себе, а как инструменты, позволяющие оценить некие латентные (не поддающиеся явному измерению) переменные. И в таком случае фактически использующийся тип шкалы может не совпадать с тем, который использовался при получении исходных данных.

Заметим, что ядром процесса формализации всегда служит выделение каких-то сторон реальности. Уточняя свойства выделенного фрагмента, мы фактически пользуемся моделями. По-

этому термин «модель» будет активно использоваться. Модель (в широком смысле) – аналог, условный образ какого – либо процесса или события, приближено воссоздающий оригинал. По количеству включаемых факторов модели делятся на однофакторные и многофакторные. Наиболее разработанной в теории статистики является методология парной корреляции – однофакторный корреляционный и регрессионный анализ. Построение и анализ двумерной модели является основой для изучения многофакторных связей. Важнейшим этапом построения модели (уравнения регрессии) является установление исходной информации.

Для практического использования моделей регрессии большое значение имеет их адекватность, т. е. соответствие фактическим статистическим данным. Проверка адекватности регрессионной модели может быть дополнена корреляционным анализом. Важнейшим этапом построения модели является установление в анализе исходной информации о математической функции. Так при анализе прямолинейной зависимости применяется уравнение однофакторной (парной) линейной корреляционной связи. Уравнение связи показывает среднее значение изменения результативного признака при изменении факторного признака на одну единицу его измерения. Знак коэффициента регрессии указывает направление этого изменения. Параметры уравнения находят методом наименьших квадратов, в основу которого положено требование минимальности сумм квадратов отклонений эмпирических данных от модельных. *Задачи регрессионного анализа* – выбор типа модели (формы связи), установление степени влияния независимых переменных на зависимую

и определение расчетных значений зависимой переменной (функции регрессии). *Задачи корреляционного анализа* сводятся к измерению тесноты известной связи между варьирующими признаками, определению неизвестных причинных связей (причинный характер которых, должен быть выяснен с помощью теоретического анализа) и оценке факторов, оказывающих наибольшее влияние на результативный признак.

Как известно, явления общественной жизни складываются под воздействием не одного, а целого ряда факторов. Между факторами существуют сложные взаимосвязи, поэтому их влияние комплексное и его нельзя рассматривать как простую сумму изолированных влияний. *Многофакторный анализ* позволяет оценить меру влияния на исследуемый результативный показатель каждого из включенных в модель (уравнение) факторов при фиксированном положении (на среднем уровне) остальных факторов, а также при любых возможных сочетаниях факторов с определенной степенью точности найти теоретическое значение этого показателя (важным условием является отсутствие между факторами функциональной связи).

Будучи как бы материальным воплощением научных представлений, достигнутых рассматриваемой ветвью науки к определенному моменту, формализация, несомненно, играет положительную роль. Она дает возможность четко обрисовать круг уже достигнутых результатов, выявить совокупность нерешенных задач, сформировать представления о возможных направлениях их решения и т. д. Формализация понятия измерения в социологии олицетворяется в так называемой репрезентационной теории измерений, что дало возможность, с одной стороны,

решить ряд стоящих перед социологической практикой задач, а с другой – выявить минусы этих представлений, определить, какие социологические ситуации остались не учтенными формализмом.

Для повышения культуры измерения исследователю полезно не просто четко формулировать используемые социально-психологические предположения, но и максимально использовать при этом математический язык. Практика показывает, что это имеет принципиальное значение: именно математика дает возможность не только четко проанализировать суть используемых моделей, но и добиться их адекватности. Всякий формализм, каким бы адекватным реальности он ни был, не может полностью, раз и навсегда удовлетворить исследователя уже в силу самой своей сути, поскольку любая формальная конструкция отражает лишь какую-то часть реальности. Прогнозы, получающиеся на основе соответствующего моделирования, оказываются более или менее оправдывающимися (это и служит проверкой качества модели). Но в какой-то момент становится ясно, что с формальной точки зрения учтено не все.

В сознании студента – социолога к сожалению, до сих пор существует своего рода психологическая «заслонка», мешающая воспринимать любой текст, написанный с использованием математического языка. Настроенные на гуманитарный лад, они с удивлением узнают, что им придется слушать довольно много математических курсов, полагая, что «истинный» социолог должен, прежде всего, понимать людей. Любой математический символ встречается «в штыки», и все, что следует за его введением, заведомо отторгается, студент уже не слышит, о чем идет речь.

Жизнь отдельного человека и всей цивилизации складывается из случайностей. Чтобы общество было устойчивым, а жизнь предсказуемой важно не давать случаю слишком большой воли (хотя исключить случайности полностью невозможно). История науки показывает, что само рождение основных положений математической статистики не в последнюю очередь было связано именно с потребностями обществознания, государственного управления и общественного развития.

# Раздел 1.
# Основные положения и история становления прикладной статистики в социологии

Потребности в статистических данных исторически определяются развитием государства. С помощью системы разнообразных показателей статистика стремиться получать характеристики различных сторон жизни общества. Статистика изучает случайные явления, которые, по своей сути, не поддаются однозначному описанию и прогнозированию. Например, нельзя абсолютно точно предсказать, сколько человек родится или умрет в стране за данный промежуток времени. Так же нельзя, с точностью до копейки, определить доход некоторой семьи (можно выиграть в лотерею, получить неожиданное наследство или потерять часть денег из-за болезни, кризиса, санкций). Современные задачи планирования, управления, прогнозирования невозможно решать, не располагая достоверными статистическими данными и не используя статистические методы обработки и анализа этих данных. Не всегда социальная проблема может быть решена в пределах имеющегося знания. В этих случаях необходимо проведение определенных теоретических и прикладных исследований. Непосредственным поводом к проведению социологического исследования служит реально возникшее противоречие в развитии социального организма.

Так, речь может идти о противоречии между социальной и профессиональной ориентацией молодежи, закончившей среднюю школу, и потребностями общества. При рассмотрении, например, проблемы текучести кадров социологическое исследование будет ориентировано как на изучение социальных факторов, порождающих это явление, так и на разработку системы мер для устранения или изменения направленности действия этих факторов. В первом случае в центре внимания научный аспект проблемы, во втором – практический. Единство научного и практического подходов к исследованию находит свое воплощение в постановке проблемы. Например, проблема текучести кадров, в сущности, сводится к двум вопросам: что заставляет людей менять место работы и как избежать излишней текучести кадров? Решение социальной проблемы направлено на улучшение условий труда и быта коллектива предприятия (организации, учреждения), населения региона, области, города, села. Человеческое общество осознало необходимость сбора статистических данных о различных сторонах жизни общества значительно раньше появления сопутствующего развитого математического аппарата. Серьезные математические методы стали использоваться для анализа статистических наблюдений сравнительно недавно.

Термин «статистика», от латинского «status», означает состояние, положение вещей. Первоначально статистика давала словесное описание «достопримечательностей» государства и только с 19 века появились количественные сведения. Сегодня под этим термином понимают отрасль практической деятельности, задачами которой являются сбор, обработка и анализ стати-

стических данных с целью решения задач прогнозирования и управления в социальных системах.

*Статистика* – наука, характеризующая количественную сторону качественно определенных массовых явлений в конкретных условиях места и времени.

Статистика при описании случайных явлений использует математику. Реальные ситуации анализируются методами теории вероятностей.

*Математическая статистика* – раздел математики, разрабатывающий методы регистрации, описания и анализа данных наблюдений и экспериментов с целью построения вероятностных моделей массовых случайных явлений.

*Прикладная статистика* – часть математической статистики, посвященная методам обработки реальных (в предметной области) статистических данных.

Заметим, что чисто математические задачи не включаются в прикладную статистику. Статистические методы следует применять во всех случаях, когда по результатам ограниченного числа наблюдений требуется установить причины улучшения или ухудшения ситуации, например, общественного явления. В определении статистики, как науки, прослеживаются ее основные особенности. Во-первых, количественную сторону явлений нельзя рассматривать в отрыве от их качественной определенности. Это своеобразие определяет особенности статистического анализа, заключающиеся в том, что статистические методы исследования органично сочетаются с методами науки, предмет которой изучается.

Во-вторых, статистические показатели не относятся к отдельному случаю, а всегда представляют результата обобщения

по массе случаев, которая называется статистической совокупностью. Стоит отметить, что значимой чертой статистической совокупности является вариативность признаков, на основании которой статистика может дать количественную характеристику исследуемой закономерности. Для статистической совокупности характерно свойство устойчивости, иначе, ее характеристики стабильны в течение длительного промежутка времени.

К специфическим приемам статистического исследования относятся:

массовое статистическое наблюдение;

сводка и группировка первичных данных;

определение обобщающих показателей;

анализ и интерпретация результатов.

Статистические методы и модели весьма эффективны в социологических, социально-экономических, управленческих областях знания. Не сложно понять, что специалисты различных «предметных областей» могут использовать одни и те же инструменты. Основное научное событие последних тридцати пяти лет – это становление научно-практической дисциплины «прикладная статистика», посвященной разработке и применению статистических методов и моделей.

В России термин «прикладная статистика» вошел в широкое употребление в 1981 г. после выхода массовым тиражом сборника «Современные проблемы кибернетики (прикладная статистика)». В этом сборнике обосновывалась трехкомпонентная структура прикладной статистики. Во-первых, в нее входят ориентированные на прикладную деятельность статистические методы анализа данных (эту область можно назвать прикладной

математической статисткой). Во-вторых, методология организации статистического исследования: как планировать исследование, как собирать данные, как подготавливать их к обработке, как представлять результаты. В-третьих, организация обработки данных, в том числе разработка и использование баз данных и электронных таблиц, статистических программных продуктов. При применении методов прикладной статистики к конкретным областям знаний и отраслям получаем научно-практические дисциплины типа «статистика в медицине», «статистика образования» и др.

Математическая статистика играет роль математического фундамента для прикладной статистики. Прикладная статистика нацелена на решение реальных задач. Поэтому в ней возникают новые постановки математических задач анализа статистических данных, развиваются и обосновываются новые методы. Сразу после возникновения теории вероятностей (Паскаль, Ферма, 17 век) вероятностные модели стали использоваться при обработке статистических данных. Например, изучалась частота рождения мальчиков и девочек, было установлено, что на 100 рожденных девочек обычно приходится 106 мальчиков. Анализировались причины того, что в парижских приютах эта вероятность отличается, от данных по всему Парижу. В 1794 г. (по другим данным – в 1795 г.) К. Гаусс разработал метод наименьших квадратов, один из наиболее популярных ныне статистических методов. В 19 веке заметный вклад в развитие практической статистики внес бельгиец А. Кетле, на основе анализа большого числа реальных данных показавший устойчивость относительных статистических показателей, которая проявляется

лишь в массе явлений. В трудах А. Кетле дано общее определение предмета статистики и сформулированы задачи статистического познания. Он активно участвовал в разработке правил проведения переписи населения и рекомендовал ее десятилетнюю периодичность. Современный этап развития прикладной статистики можно отсчитывать с 1900 г., когда англичанин К. Пирсон основан журнал «Biometrika». Изучались методы, основанные на анализе данных из параметрических семейств распределений. Наиболее популярным было нормальное (гауссово) распределение. Для проверки гипотез использовались критерии Пирсона, Стьюдента, Фишера. Были предложены метод максимального правдоподобия, дисперсионный анализ, сформулированы основные идеи планирования эксперимента.

Выборочные исследования – один из основных инструментов социологов. Разработанную в первой трети XX в. теорию называют параметрической статистикой, поскольку ее основной объект изучения – это выборки из распределений, описываемых одним или небольшим числом параметров. Наиболее общим является семейство кривых Пирсона, задаваемых четырьмя параметрами. Как правило, нельзя указать каких-либо веских причин, по которым конкретное распределение результатов наблюдений должно входить в то или иное параметрическое семейство. Исключения хорошо известны: если вероятностная модель предусматривает суммирование независимых случайных величин, то сумму естественно описывать нормальным распределением; если же в модели рассматривается произведение таких величин, то итог, видимо, приближается логарифмически нормальным распределением. Одновременно с параметрической

статистикой, в работах Спирмена и Кендалла появились первые непараметрические методы, основанные на коэффициентах ранговой корреляции, носящих ныне имена этих статистиков. В 30-е годы появились работы А.Н. Колмогорова и Н.В. Смирнова, предложивших и изучивших статистические критерии, носящие в настоящее время их имена. Эти критерии основаны на использовании, так называемого эмпирического процесса, разности между эмпирической и теоретической функциями распределения. В работе А.Н. Колмогорова (1933 г.) изучено предельное распределение супремума модуля эмпирического процесса, называемого сейчас критерием Колмогорова. Затем Н.В. Смирнов исследовал супремум и инфимум эмпирического процесса, а также интеграл (по теоретической функции распределения) квадрата эмпирического процесса. Следует отметить, что встречающееся в литературе словосочетание «критерий Колмогорова-Смирнова», не вполне корректно, поскольку эти два статистика никогда не печатались вместе и не изучали один и тот же критерий. Корректно выражение «критерий типа Колмогорова-Смирнова», применяемое для обозначения критериев, основанных на использовании супремума функций от эмпирического процесса. После второй мировой войны большую роль сыграли работы Вилкоксона и его школы. К настоящему времени с помощью непараметрических методов можно решать практически тот же круг статистических задач, что и с помощью параметрических.

Согласно общепринятой в настоящее время классификации статистических методов прикладная статистика делится на четыре области:

статистика (числовых) случайных величин;

многомерный статистический анализ;

статистика временных рядов и случайных процессов;

статистика объектов нечисловой природы.

Первые три из этих областей являются классическими. Четвертая, сравнительно нова. Ее именуют также статистикой нечисловых данных или попросту нечисловой статистикой. В классической математической статистике элементы выборки – это числа. В многомерном статистическом анализе – вектора. А в нечисловой статистике элементы выборки – это объекты нечисловой природы, которые нельзя складывать и умножать на числа. Другими словами, объекты нечисловой природы лежат в пространствах, не имеющих векторной структуры. Выделение статистики объектов нечисловой природы в качестве самостоятельного направления в прикладной статистике, определяет методы статистического анализа данных произвольной природы.

Примерами объектов нечисловой природы являются:

значения качественных признаков, т. е. результаты кодировки объектов (например, ответов на вопросы социологической анкеты) с помощью заданного перечня категорий (градаций);

упорядочение (ранжировка), классификации, результаты парных сравнений;

множества (обычные или нечеткие);

слова, предложения, тексты;

вектора, координаты которых – совокупность значений разнотипных признаков, часть из них носит качественный характер, а часть – количественный;

ответы на вопросы анкеты, часть из которых носит количественный характер (возможно, интервальный), часть сводится к выбору одной из нескольких подсказок, а часть представляет собой тексты.

Перспективное и быстро развивающееся направление последних лет – математическая статистика интервальных данных, которая идейно связана с интервальной математикой, где в роли чисел выступают интервалы. Это направление математики является дальнейшим развитием всем известных правил приближенных вычислений, посвященных выражению погрешностей суммы, разности, произведения, частного через погрешности тех чисел, над которыми осуществляются перечисленные операции. В частности, изучены проблемы регрессионного анализа, планирования эксперимента, сравнения альтернатив и принятия решений в условиях интервальной неопределенности. Она применима к оцениванию математического ожидания, дисперсии, коэффициента вариации, параметров гамма-распределения и характеристик аддитивных статистик, при проверке гипотез о параметрах нормального распределения, в том числе с помощью критерия Стьюдента, а также гипотезы однородности с помощью критерия Смирнова. В области асимптотической математической статистики интервальных данных российская наука имеет мировой приоритет. Развертывание работ в этом направлении позволит закрепить этот приоритет, получить теоретические результаты, основополагающие в новой области математической статистики и необходимые для обоснованного статистического анализа почти всех типов данных.

Классической задачей статистики является изучение изменений анализируемых показателей во времени. Эта задачи решается при помощи анализа рядов динамики (временных рядов), то есть расположенных в хронологической последовательности числовых значений статистического показателя, характеризующих изменение общественных явлений во времени. В каждом ряду динамики имеются два основных элемента: время ($t$) и конкретное значение показателя – уровень ряда ($y$). Уровни в динамическом ряду могут быть представлены абсолютными, средними или относительными величинами. Построение и анализ рядов динамики позволяют выявить и измерить закономерности развития явлений во времени. Эти закономерности не проявляются четко на каждом конкретном уровне, а лишь в тенденции, в достаточно длительной динамике. На основную закономерность динамики накладываются другие, прежде всего случайные влияния. Выявление основной тенденции в изменении уровней, именуемой трендом, является одной из главных задач анализа рядов динамики. Интервальным (периодическим) рядом динамики называется такой ряд, уровни которого характеризуют размер явлений за конкретный период времени (год, квартал, месяц). Значения уровней интервального ряда не содержатся в предыдущих или последующих показателях, их можно просуммировать, что позволяет получать ряды динамики более укрупненных периодов. Этим достигается суммарное обобщение результата развития изучаемого явления с начала отчетного периода.

*Объект социологического исследования* – деятельность людей, занимающих определенное социальное положение, и условия, при которых эта деятельность осуществляется. При опи-

сании объекта могут учитываться его различные характеристики: профессиональная (или отраслевая) принадлежность; пространственная ограниченность (регион, город, село); функциональная направленность (производственная, политическая, бытовая); временные границы и др.

Наличие отраслевых границ позволяет сосредоточиться на наиболее важных и определяющих чертах функционирования данной социальной системы, пространственные границы конкретизируют объект с точки зрения его производственно-территориальной общности. Временные границы конкретизируют сроки проведения исследования. Если объектами социологического исследования являются деятельность людей и ее условия, то единицами наблюдения, как правило, выступают носители этой деятельности – люди. Единицей наблюдения в социологии называется тот элемент исследуемой совокупности, в отношении которого непосредственно ведется сбор социальной информации. При этом важно различать единицы наблюдения, о которых собирается информация (единицы анализа), и единицы наблюдения, от которых поступает информация (единицы сбора). Социологическое исследование может включать различные по своим характеристикам единицы наблюдения. В исследовании текучести рабочей силы на промышленном предприятии можно выделить такие единицы наблюдения, как выбывшие рабочие, вновь поступившие рабочие, работающие в настоящее время на предприятии. Изучение таких данных позволит углубить представление о факторах, влияющих на текучесть рабочей силы, и выявить так называемую потенциальную текучесть.

Предметом исследования принято считать ту из сторон объекта, которая непосредственно подлежит изучению. Одному и тому же социальному объекту может соответствовать несколько различных предметов исследования, каждый из которых по содержанию определяется тем, какую именно сторону объекта он отражает, с какой целью, а главное, для решения какой социальной проблемы выбран. Объект и предмет совпадают, когда перед исследователем стоит задача познания всей совокупности закономерностей функционирования и развития конкретного социального явления, процесса. Когда же речь идет об изучении каких-либо отдельных характеристик объекта исследования, предметом становятся те стороны, которые содержат эти характеристики.

При исследовании миграционных процессов объектом исследования является население различных территориальных единиц: республики, области, района, отдельного населенного пункта. Предметом является миграция – переселение людей из одного места проживания в другое. Цель исследования (то, ради чего проводится исследование) – оптимизация миграционных процессов в некотором районе. Задача исследования – нахождение наилучших путей этой оптимизации (для практически ориентированного исследования) и установление закономерностей миграции населения (для теоретически ориентированного исследования). Один и тот же объект можно описать по-разному. Решение этого вопроса в каждом частном случае определяется социальной проблемой и целями социологического исследования. А от того, какие будут выделены элементы и связи, зависит выбор средств их фиксации (методы сбора и анализа данных).

Под *статистическими данными* понимают числовые или нечисловые значения контролируемых параметров исследуемых объектов, которые получены в результате наблюдений (измерений, анализов, испытаний, опытов и т. д.) определенного числа признаков, у каждой единицы, вошедшей в исследование. Способы получения статистических данных и объемы выборок устанавливают, исходя из постановок конкретной прикладной задачи на основе методов математической теории планирования эксперимента. Результаты наблюдений обрабатывают с помощью методов прикладной статистики, соответствующих поставленной задаче. Используют, как правило, аналитические методы, т. е. методы, основанные на численных расчетах (объекты нечисловой природы при этом описывают с помощью чисел). Также допустимо применение графических методов (визуального анализа).

Основной задачей социальной статистики является разработка и анализ системы показателей для характеристики качества и уровня жизни населения, отражающей демографическую ситуацию и различные аспекты социальных отношений. Проведение подобной работы является необходимым условием обеспечения всех уровней управления необходимой и обоснованной информацией для принятия управленческих решений по регулированию и прогнозированию общественного развития.

# Раздел 2.
# Генеральная совокупность и выборка

## 2.1. Виды выборочной совокупности

Понятия генеральной совокупности и выборки из нее являются основополагающими в статистике. Строгие определения пришли из теории вероятностей, хотя терминология математической статистики отличается от терминологии теории вероятностей, что объясняет трудности использования студентами-социологами знаний из курса теории вероятностей в практике изучения прикладной статистики. Так, например, в качестве случайных событий рассматриваются события, каждое их которых состоит в том, что респондент обладает определенным сочетанием значений рассматриваемых признаков. Сами признаки служат примерами случайных величин (вместо вероятностей фигурируют относительные частоты). Вместо случайной величины $X$ из теории вероятностей, в математической статистике говорят о генеральной совокупности $X$. Таким образом, понятие генеральной совокупности тождественно понятию случайной величины, т. е. включает в себя описание области определения (пространства элементарных исходов), множества значений, функциональной зависимости, закона распределения. Вместо эксперимента, в результате которого случайная величина $X$ приняла значение $x$ (в теории вероятностей), в математической статистике говорят о случайной выборке из генеральной совокупности $X$ значения $x$. Вместо $n$ независимых экспериментов,

в результате которых случайная величина $X$ приняла значения $x_1, x_2, ..., x_n$ (как принято в теории вероятностей), в математической статистике говорят о случайной выборке объема $n$ значений $x_1, x_2, ..., x_n$ из генеральной совокупности $X$. Например, социолог, изучающий мнение избирателей, под генеральной совокупностью понимает множество всех избирателей страны, а под выборкой объема $n$ – множество из $n$ человек, которых он опросил. Таким образом, генеральная совокупность – это все представители какой-либо группы, носители какого-либо важного признака, например:

все российские избиратели;

все потенциальные потребители пива, проживающие в Перми;

все подростки (12-16 лет) Поволжского региона;

все учителя физики и химии, работающие в средних школах;

все домохозяйства, имеющие доход от 500 до 1 500 долл. в месяц;

все компании, занимающиеся розничной торговлей в Москве, и т. д.

*Генеральной совокупностью* называют всю подлежащую изучению совокупность объектов, относительно некоторого признака, характеризующего эти объекты.

*Выборочной совокупностью (выборкой)* называют ту часть объектов генеральной совокупности, которую отобрали для непосредственного изучения признака.

Параметр – фиксированное, но неизвестное число. Обычно в распоряжении исследователя имеются лишь данные выборки, например значения количественного признака $x_1, x_2, ..., x_n$,

полученные в результате $n$ наблюдений. Через эти данные и выражают оцениваемый параметр.

*Параметр генеральной совокупности (параметр)* – показатель, вычисленный для всей генеральной совокупности (например, среднее квадратичное отклонение случайной величины).

*Параметр выборки (выборочный параметр, статистика)* – некоторый показатель, вычисленный на основе данных выборки.

Под *статистической оценкой* параметров генеральной совокупности понимают методы, позволяющие делать научно обоснованные выводы о значениях числовых параметров генеральной совокупности по случайной выборке из нее. Построение и анализ выборочного распределения является основным математическим способом исследования реальной случайной величины. Как уже отмечалось выборка – это множество данных, взятых с помощью определенных процедур из генеральной совокупности для исследовательского анализа.

*Репрезентативность выборки* – это показатель, заключающийся в том, что выборка полно и достоверно отображает признаки той совокупности, частью которой она является. Репрезентативность также можно определять, как свойство выборки наиболее полно представлять характеристики генеральной совокупности, существенные с точки зрения цели исследования.

Допустим, что генеральная совокупность – все ученики школы (900 человек из 30 классов, по 30 человек в каждом классе). Объект исследования – отношение школьников к курению. Выборочная совокупность, состоящая из 90 учащихся только старших классов, намного хуже представит всю совокупность,

чем выборка из тех же 90 учеников, куда вошли бы из каждого класса по 3 ученика. Главная причина – неравное распределение по возрастам. Таким образом, в первом случае репрезентативность выборки будет низкой. Во втором случае – высокой.

В качестве примера нерепрезентативной выборки можно привести классический случай, произошедший в 1936 году в США во время президентских выборов. Журнал «Литэрари дайджест», который до этого весьма успешно прогнозировал результаты предыдущих выборов, на этот раз ошибся в своих прогнозах, хотя разослал несколько миллионов письменных вопросов подписчикам, а также респондентам, которых они выбрали из телефонных книг и из списков регистрации автомобилей. В 1/4 бюллетеней, которые вернулись заполненные обратно, голоса распределились следующим образом: 57% отдали первенство кандидату от республиканцев по имени Альф Лэндон, и только 41% отдали предпочтение действующему президенту – демократу Франклину Рузвельту. В действительности, на выборах победил Ф. Рузвельт, который набрал почти 60% голосов. Ошибка «Литэрари дайджест» была в следующем. Они захотели увеличить репрезентативность выборки. А так как знали, что большинство их подписчиков относят себя к республиканцам, то решили расширить выборку за счет респондентов, выбранных ими из телефонных книг и автомобильных регистрационных списков. Но не учли существующих реалий и фактически отобрали еще больше сторонников республиканцев, потому что во времена Великой депрессии иметь автомобили и телефоны мог позволить себе средний и высший класс. А это и были по большей части республиканцы, а не демократы.

Существуют различные виды выборок: простая случайная, серийная, типическая, механическая и комбинированная.

*Простая* случайная выборка состоит в отборе из всей совокупности изучаемых единиц наугад без какой-либо системы.

*Механическую* выборку применяют тогда, когда в генеральной совокупности есть упорядоченность, например, имеется некая последовательность единиц (регистрационные номера работников, избирательные списки, номера телефонов респондентов, номера квартир и домов и другое).

*Типический* отбор используется в тех случаях, когда всю совокупность можно разделить на группы по типам. При работе с населением такими могут быть, например, образовательные, возрастные, социальные группы, при исследовании предприятий – отрасль или отдельная организация и др.

*Серийный* отбор удобен тогда, когда единицы объединены в небольшие серии или группы. Такой серией могут быть партии, школьные классы, трудовые коллективы и другие группы.

*Комбинированная* выборка предполагает использование всех предыдущих видов выборки в той или иной комбинации.

Кроме того, выделяют качественную и количественную репрезентативность. Случайность, гарантирующая *качественную (структурную) репрезентативность* статистических исследований, достигается выполнением ряда условий формирования выборочных групп (совокупностей):

1. Каждый член генеральной совокупности должен иметь равную вероятность попасть в выборку.

2. Отбор единиц наблюдения из генеральной совокупности необходимо проводить независимо от изучаемого признака. Ес-

ли отбор проводится целенаправленно, то и при этом необходимо соблюдать условия независимости распределения изучаемого признака.

3. Отбор должен проводиться из однородных групп.

Соблюдение условий, гарантирующих максимальную близость выборочной и генеральной совокупностей, обеспечивается специальными способами отбора. В зависимости от способа формирования различают следующие выборки:

1. Выборки, не требующие разделения генеральной совокупности на части (собственно, случайная повторная или бесповторная выборка).

2. Выборки, требующие разбиения генеральной совокупности на части (механическая, типическая или типологическая выборки, когортная, парно-сопряженная выборки).

Собственно, случайная выборка формируется случайным отбором – наудачу. В основе случайного отбора лежит перемешивание. Например: выбор шара в спортлото после перемешивания всех шаров, выбор выигрышных номеров, случайный выбор карточек больных для исследования и т. п. Иногда используют случайные числа, получаемые из таблиц случайных чисел или с помощью генераторов случайных чисел. Согласно этим числам из заранее пронумерованного массива генеральной совокупности выбираются единицы наблюдения с номерами, соответствующими выпавшим случайным числам. При составлении случайной выборки после того, как объект выбран, и все необходимые данные о нем зарегистрированы, можно поступать двояко: объект можно вернуть, или не вернуть в генеральную совокупность. В соответствии с этим выборку называют по-

вторной (объект возвращается в генеральную совокупность) или бесповторной (объект не возвращается в генеральную совокупность). Поскольку в большинстве статистических исследований разница между повторной и бесповторной выборками практически отсутствует, то априорно принимается условие, что выборка повторная.

## 2.2. Методы репрезентативного отбора

Для того чтобы выборочная совокупность была *количественно репрезентативной* по отношению к генеральной, необходимо первоначально оценить количество данных, которое требуется включить в выборочную совокупность. Достаточно ли чтобы относительно небольшая выборка (от нескольких сотен до нескольких тысяч представителей) репрезентировала (выразила) мнение генеральной совокупности? Как такое возможно? На каком основании можно распространять данные, полученные от небольшой группы людей, на существенно (в десятки и сотни раз) большую группу? Это происходит на основании гипотезы о том, что большинство представителей четко определенной социально-демографический группы будут реагировать сходным образом. Нет никакой необходимости опрашивать всех представителей этой группы, поскольку их мнение (с допустимой погрешностью) может представить (репрезентировать) небольшая выборка из ее представителей.

Существуют две группы методов построения выборки, в той или иной степени реализующих репрезентацию мнений и позиций генеральной совокупности: вероятностные и детерми-

нированные. Первая группа методов (*вероятностные*) базируется на использовании теории вероятности. В основе ее применения лежит постулат, что репрезентация будет достигнута в случае, если каждой единице генеральной совокупности обеспечено равновероятное попадание в выборку. Например, если генеральной совокупностью является все взрослое (16-85 лет) население города (200 тыс. человек), то каждому жителю должна быть обеспечена вероятность стать участником исследования (попасть в выборку), равная 1/200000. В противном случае выборка будет не случайной, а смещенной, т. е. менее репрезентативной. Реализовать это можно в случае, если все элементы генеральной совокупности могут быть тем или иным образом пронумерованы, а затем эти номера будут выбраны в определенной последовательности – «по воле случая». Например, в Москве около 2500 средних школ, каждая из которых имеет свой номер. Мы могли бы выбрать наугад 100 номеров и провести опрос 100 директоров (завучей, учителей физики, классных руководителей 11-х классов) в этих школах. Эти 100 номеров можно выбрать с помощью таблицы или «генератора случайных чисел». Такие способы построения выборки называются «*простой случайной выборкой*». Каждый ее элемент отбирается независимо и имеет равную вероятность попасть в выборку.

Можно выбрать наугад любое число от 1 до 25, например – 12, а затем взять в выборку школы с номерами: 12, 37, 62, 87, 112, 137 и т. д. Такой метод построения называемся «*систематической выборкой*», первый элемент которой выбирается произвольно, а затем выбирают каждый *i*-й элемент. Или сначала разделить эти школы на несколько страт (возможно, и пересе-

кающихся), например, на школы физико-математические, спортивные, лингвистические и гуманитарные, а затем произвести случайную или систематическую выборку (по 20-30 школ) из каждой страты. Такой метод построения называется «*стратифицированной выборкой*». Разновидностью стратифицированной выборки является «маршрутная выборка», суть реализации которой состоит в следующем. Город делится на 20-40 «секторов» по числу интервьюеров, задействованных и исследовании. Каждый интервьюер получает один сектор, маршрут обследования «своего» сектора и инструкцию по реализации простой случайной выборки. Например, «Начать обход с улицы Баумана, с дома № 2, третьего подъезда, второго этажа сверху, первой квартиры слева. Затем – дом № 4, второй подъезд, третий этаж, вторая квартира справа…, потом – переулок Строителей, нечетная сторона … и т. д. Наконец, можно разделить генеральную совокупность на непересекающиеся кластеры, к примеру, по муниципальным районам (их в Москве 125, и в каждом в среднем по 20 школ). Затем случайным образом выбрать пять районов и произвести обследование всех школ данного муниципального района. Такой метод построения называется «*кластерной выборкой*».

Тем не менее, у вероятностных методов построения выборки есть один весьма существенный недостаток. Каждый из них исходит из предположения о том, что все элементы генеральной совокупности являются равнодоступными: и в «техническом» смысле (у всех есть телефон для телефонного опроса или доступ в Интернет), и в «психологическом», т. е. все респонденты с примерно равной вероятностью согласятся или от-

кажутся принимать участие в исследовании. Однако это не так. Граждане с относительно высокими доходами менее доступны для исследователей, чем те, чьи доходы невысоки. И нет возможности, что бы заставить этих людей отвечать вопросы социологов.

Поэтому все выборки всегда смещены в сторону средне- и малообеспеченных групп населения. Во всех без исключения странах мира. Также, менее образованные граждане идут на контакт с социологами не охотно, в противоположность лицам с высшим образованием. Поэтому в большинстве выборок доля хорошо образованных граждан, как правило, существенно выше, чем в генеральной совокупности. Никто из сотрудников исследовательских компаний не желает общаться с бомжами, алкоголиками, наркоманами, психо- и социопатами и прочими маргиналами. У руководителя исследования нет решительно никаких возможностей заставить своих сотрудников делать это, а между прочим, к этим группам в России по взвешенным оценкам относится от 12 до 15% жителей. Следовательно, любая выборка смещена в сторону «вменяемых» граждан. Некоторые граждане боятся отвечать на вопросы, даже самые невинные. Таких людей немного, но они есть. А вот способов заставить их участвовать в опросе нет. Наконец, есть люди, которые просто не желают участвовать в исследовании. У них есть время, они ничего не боятся, они все понимают, но на вопросы отвечать отказываются. И точка. Таким образом, все выборки в социологии являются смещенными в сторону средне- и малообеспеченных, более образованных, контактных и вменяемых граждан. Они и репрезентируют общее мнение генеральной совокупности.

Преодолеть изложенные выше проблемы можно с помощью метода «квот», относящегося к *детерминированным методам*, при котором априори обеспечивается пропорциональное представительство носителей существенных признаков (пол, возраст, доход, образование и т. п.) генеральной совокупности в выборке. Это наиболее эффективный метод проведения массовых опросов. При его использовании существенно облегчается задача поиска корреляционных связей, сравнения различных типов (групп) потребителей между собой и экстраполяции выявленных закономерностей на генеральную совокупность.

Единственная, но весьма существенная трудность при реализации него метода состоит в том, что не всегда доподлинно известно распределение всех важных параметров в самой генеральной совокупности. В этом случае исследователь или консультант исследовательского проекта должен взять на себя смелость распределить квоты по своему усмотрению, в соответствии со своим видением и пониманием поставленной задачи.

Задача достижения строгой репрезентативности не всегда является обязательной. Иногда целесообразно воспользоваться существенно более простыми в реализации детерминированными методами:

*произвольным* – когда опрашивают того, кто «попался под руку» интервьюеру и согласился участвовать в опросе. Естественно, этот метод дает крайне ненадежные результаты. А вдруг под руку попадется рота солдат или команда волейболисток. Однако его использование допустимо в исследованиях, носящих поисковый характер, не требующих большой точности, при про-

ведении «пилотажа» анкеты. «Произвольность» можно компенсировать большим объемом выборки, из которой затем можно будет попробовать отобрать необходимое число «подходящих» анкет и составить уже из них репрезентативную в каких-то отношениях выборку;

*поверхностным* – когда отбор осуществляется по самым общим признакам, задаваемым исследователем интервьюерам в виде не очень строгого задания;

«*воронки*» – когда сначала отбираются наиболее «контактные», а затем среди них – наиболее «компетентные», подходящие респонденты;

«*концентрации*» – на представителях отдельных, сопоставимых сегментов, среди которых проводят «сплошной» опрос. Например, школьный 11«А» класс может представлять всех старшеклассников школы или даже города как «обычный», «типичный класс»;

«*снежного кома*» – когда начальная группа подбирается случайным образом, а дальнейший отбор ведется из кандидатов, указанных первыми респондентами.

Важным свойством выборки является ее достоверность. *Достоверность* – показатель вероятности того, что истинное значение изучаемого параметра генеральной совокупности попадет в доверительный интервал. Чем выше задаваемый уровень достоверности, тем больше должна быть выборка. Например, общероссийская городская выборка (14-65 лет) из 1200 респондентов имеет доверительный интервал 4% пунктов. При ее проведении 15% участников опроса заявили, что за последние три месяца были в кинотеатре хотя бы один раз. Эти

данные позволяют нам утверждать с заданным уровнем достоверности, что от 11 до 19% жителей российских городов в возрасте от 14 до 65 лет были в кинотеатре хотя бы один раз за последние три месяца. Иными словами, можно сказать, что все значения между 11 и 19% в данном случае находятся в пределах «допустимой статистической погрешности». Если бы мы хотели задать доверительный интервал в 2% пункта, то выборку (при прочих равных условиях) пришлось бы увеличить примерно в четыре раза.

Размер выборки практически не зависит от размера генеральной совокупности. И в мегаполисе с населением более миллиона человек, и в уездном городе с населением в 35 тыс. человек для построения выборки, репрезентативной по одинаковому числу параметров, потребуется опросить одинаковое число респондентов. От чего действительно зависит размер выборки – так это от числа параметров, по которым мы желаем добиться репрезентативности.

Если нас устраивает репрезентативность только по полу и возрасту, то выборки в 400 человек в одном населенном пункте будет более чем достаточно. Если параметров три, количество респондентов придется увеличить, например, до 600. Добиться репрезентативности выборки одновременно по пяти параметрам: полу, возрасту, доходу, образованию, сфере профессиональной деятельности – можно лишь на выборке из 1000-1200 человек в одном населенном пункте.

Задача прикладной статистики: описать закон распределения генеральной совокупности; подобрать значения параметров этого закона, оценить числовые характеристики генеральной со-

вокупности. Если имеется несколько выборок, извлеченных из разных генеральных совокупностей, определить, одинаково распределены эти генеральные совокупности или нет; одинаковы ли определенные числовые характеристики этих генеральных совокупностей или нет. Перечисленные вопросы сформулированы на языке теории вероятностей. От статистики требуют ответы и на другие вопросы: Какой будет численность населения страны в следующем году? Существует ли связь между уровнем заработной платы учителя и количеством выпускников, сдавших единый государственный экзамен с высокими баллами? Чтобы ответы на подобные вопросы соответствовали действительности, нужно уметь строить подходящие вероятностные модели для реальных ситуаций. А для этого нужно уметь представить выборку в подходящем для изучения виде. Возникает задача описания и представления выборки. Рассмотрим этот аспект на примере.

Пример. Измерим уровень образования каждого человека в списке. Получим неупорядоченный ряд результатов отдельных наблюдений: количество оконченных классов – 11, 4, 7, 9, 11, 8, 11, 11.

Если отдельные наблюдения расположить в порядке возрастания указанных выше значений признака, то получим вариационный ряд: 4, 7, 8, 9, 11, 11, 11, 11. По вариационному ряду количественного признака можно подсчитать, как часто каждое значение этого признака встречается в совокупности. В результате получим *частотное распределение* для данного признака. Определим значения относительных и накопленных частот при измерении уровня образования респондентов. Для

вышеприведенного примера с объемом выборки $n = 8$ человек частотное распределение выглядит так:

| Отдельные значения признака ($x_i$) | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
|---|---|---|---|---|---|---|---|---|
| Частота($n_i$) | 1 | 0 | 0 | 1 | 1 | 1 | 0 | 4 |

Абсолютное число, показывающее, сколько раз встречается то или иное значение признака $x_i$, называется *частотой* и обозначается соответственно $n_1, n_2, \ldots, n_k$.

*Относительной частотой* называется доля значений признака в общем числе наблюдений и обозначается $m_1, m_2, \ldots, m_k$.

Например, для приведенного частотного ряда относительная частота наибольшего значения признака = 50%.

Таблица относительных частот напоминает таблицу вероятностей дискретной случайной величины, только вместо значений случайной величины записывают варианты выборки, а роль вероятностей исполняют относительные частоты:

| Отдельные значения признака ($x_i$) | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
|---|---|---|---|---|---|---|---|---|
| Относительная частота ($m_i$) | 0,125 | 0 | 0 | 0,125 | 0,125 | 0,125 | 0 | 0,5 |

Накопленной частотой называется число вариант выборки, меньших данного числа $x$. Относительной накопленной частотой называется отношение накопленной частоты ко всему объему выборки. Составим ряды накопленных и относительно накопленных частот вариант выборки для рассматриваемого примера:

| Отдельные значения признака ($x_i$) | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
|---|---|---|---|---|---|---|---|---|
| Накопленная частота | 0 | 0 | 0 | 1 | 2 | 3 | 3 | 7 |
| Относительная накопленная частота | 0 | 0 | 0 | 0,125 | 0,25 | 0,375 | 0,5 | 1 |

Если выборка извлечена из непрерывно распределенной генеральной совокупности, причем ее объем $n$ достаточно велик, то такую выборку неразумно представлять в виде таблицы частот. Поэтому достаточно большую выборку, извлеченную из непрерывно распределенной генеральной совокупности, представляют сгруппированными данными.

## 2.3. Группировка данных

*Сгруппированные данные* – отдельные значения признаков, объединенные в группы (интервалы). Весь диапазон значений вариант разбивают на разумное число интервалов, как правило, одинаковой ширины $h$. В этом случае частоты соотносят уже не с каждым отдельным значением признака, как это делалось в предыдущем примере, а с рядом значений, попадающих в определенный интервал. Во избежание недоразумений при подсчете числа вариантов выборки, попавших в каждый интервал, левая граница интервала считается закрытой, а правая – открытой.

Частотой $i$-го интервала называется число, равное количеству вариант выборки, попавших в этот интервал. Аналогично вычисляют накопленные и относительные накопленные частоты для правых границ интервалов.

Например, распределение уровня образования в вышеприведенном примере может быть представлено в виде интервального ряда следующим образом:

| Образование (классы) | 4-7 | 8-9 | 10-11 |
|---|---|---|---|
| Частота | 2 | 2 | 4 |

При построении интервальных рядов большое значение имеет выбор размеров интервалов. Общее требование к этому выбору состоит в том, что группировка должна наиболее полно выявлять существенные свойства рядов распределения. Как правило, приходится делать выбор между двумя крайностями: слишком крупные интервалы для данного объема выборки скрадывают многие нюансы в описании явления, а слишком дробные ведут к статистически незначимым частотам внутри интервала. Интервальные ряды распределения могут строиться с равными и неравными интервалами. Неравные интервалы применяются при неравномерном распределении частот значений группировочного признака для выделения качественно отличных типов явления.

Если у исследователя нет предварительной информации о характере распределения по тому или иному признаку, то следует задавать равные интервалы. Равные интервалы также наиболее удобны при использовании методов математической статистики, при этом по каждому из признаков не следует брать более 20 группировочных интервалов. При образовании интервалов необходимо точно обозначить количественные границы группы, избегая таких обозначений границ интервалов, при которых отдельные единицы совокупности могут быть отнесены в две соседние группы.

Пример. Для 50 учеников составили тестовое задание, результаты выполнения оценили с точностью до сотых. Полученные данные представлены выборкой:

3,7 3,85 3,7 3,78 3,6 4,45 4,2 3,87 3,33 3,76 3,75 4,03 3,75 4,18 3,8 **4,75** 3,25 4,1 3,55 3,35 3,38 3,3 4,15 3,95 3,5 3,88 3,71 3,15

4,15 3,8 4,22 3,75 3,58 3,55 4,08 4,03 3,24 4,05 3,56 **3,05** 3,58 3,98 3,88 3,78 4,05 3,4 3,8 3,06 4,38 4,2

*Решение*. Для группировки этой выборки выделены наименьшее и наибольшее значения среди данных, которые соответственно равны 3,05 и 4,75. Выборка «упакована» в границы от 3-4,8. Выберем ширину интервала $h = 0{,}3$ и получим 6 интервалов.

Вычислим накопленные частоты для правых границ и составим таблицу 2.1.

**Таблица 2.1. Интервальная таблица накопленных частот**

| | **[3-3,3)** | **[3,3-3,6)** | **[3,6-3,9)** | **[3,9-4,2)** | **[4,2-4,5)** | **[4,5-4,8)** |
|---|---|---|---|---|---|---|
| Частоты | 5 | 11 | 17 | 11 | 5 | 1 |
| Относительные частоты | 0,1 | 0,22 | 0,34 | 0,22 | 0,1 | 0,02 |
| Накопленные частоты | 5 | 16 | 33 | 44 | 49 | 50 |
| Относительные накопленные частоты | 0,1 | 0,32 | 0,66 | 0,88 | 0,98 | 1,0 |

*Статистические таблицы* – форма рационального и наглядного представления цифровых характеристик исследуемых явлений, дающая возможность характеризовать их размеры, структуру и динамику. Таблицы бывают простые, групповые и комбинационные.

*Простая таблица* – в ней содержатся сводные показатели, относящиеся к перечню единиц наблюдения, хронологических дат или территориальных подразделений. Соответственно таблицы могут называться простыми перечневыми, хронологическими или территориальными. Приведем пример простой таблицы (табл. 2.2).

**Таблица 2.2. Динамика уровня безработицы в Российской Федерации за 1994-1998 гг.**

| | 1994 | 1995 | 1996 | 1997 | 1998 |
|---|---|---|---|---|---|
| Уровень безработицы, рассчитанный по методологии международной организации труда (МОТ) | 7,4 | 8,8 | 9,9 | 11,2 | 13,3 |
| Уровень официальной безработицы | 2,2 | 3,2 | 3,4 | 2,8 | 2,6 |

Различные значения уровня безработицы получаются из различий в методологии расчета. По методике, используемой службой занятости населения (уровень официальной безработицы), в основе расчетов лежат вопросы по финансированию мероприятий, связанных с профессиональной переподготовкой и обучением, выплатой пособий по безработице и т. д. Методология МОТ нацелена на определение реальной нагрузки на рынок труда со стороны предложения рабочей силы.

*Групповые таблицы* – в них статистическая совокупность разделяется на отдельные группы по какому-либо признаку, причем каждая из групп может быть охарактеризована рядом показателей (табл. 2.3).

Данные табл. 2.3 свидетельствуют о том, что если до 1997 г. в России наибольшая численность населения была занята в государственном секторе, то в 1998 г. 41,9% общей численности занятых приходится на частный сектор.

*Комбинационные таблицы* – в них статистическая совокупность разбита на группы не по одному, а по нескольким признакам. Идея их построения состоит в том, что каждую из групп в таблице разбивают на подгруппы по какому-либо при-

знаку, которые в свою очередь дальше могут разделяться по следующему признаку.

**Таблица 2.3. Распределение занятого населения России по секторам экономики**

| Занятое население | млн. чел. | | в % к итогу | |
|---|---|---|---|---|
| | 1997 г. | 1998 г. | 1997 г. | 1998 г. |
| В экономике, всего | 64,6 | 63,6 | 100 | 100 |
| В том числе по секторам: государственные и муниципальные предприятия и организации | 25,9 | 24,4 | 40,1 | 38,3 |
| Частный сектор | 25,8 | 26,6 | 39,9 | 41,9 |
| Общественные организации, фонды | 0,4 | 0,4 | 0,6 | 0,6 |
| Совместные предприятия и др. | 12,5 | 12,2 | 19,4 | 19,2 |

Результаты комбинационной группировки по большому количеству признаков даже при небольшом числе интервалов становятся труднообозримыми. Поэтому нецелесообразно составлять комбинационные таблицы по сочетанию более чем трех признаков и более четырех интервалов.

Группировка, осуществляемая одновременно по комплексу признаков, называется *многомерной*. Например, для характеристики технического уровня развития предприятий могут быть использованы различные показатели.

В табл. 2.4 приведены результаты группировки предприятий по уровню технического развития и производительности труда. При выделении однородных по техническому уровню групп предприятий применимы, например, методы кластерного анализа по восьми показателям технического уровня развития. Анализ данных табл. 2.4 позволяет выделить группы предпри-

ятий, добившихся наибольшего эффекта в своей деятельности, и группы предприятий, располагающих резервами роста производства за счет лучшего использования технического потенциала. Это те шесть предприятий, которые имеют техническую оснащенность выше средней по отрасли, но уровень производительности труда здесь ниже среднего по отрасли.

**Таблица 2.4. Распределение предприятий по уровням технической оснащенности и производительности труда**

| **Группы предприятий по уровню технической оснащенности** | **Группы предприятий по уровню производительности труда** | | | Итого |
|---|---|---|---|---|
| | Ниже среднего по отрасли уровня | Среднего по отрасли уровня | Выше среднего по отрасли уровня | |
| Ниже среднего по отрасли уровня | 9 | 8 | 8 | 25 |
| Среднего по отрасли уровня | 5 | 6 | 1 | 12 |
| Выше среднего по отрасли уровня | 6 | 3 | 7 | 16 |
| Итого | 20 | 17 | 16 | 53 |

Построение таблицы подчинено определенным правилам. Основное ее содержание должно быть отражено в названии (круг рассматриваемых вопросов, географические границы статистической совокупности, время, единицы измерения). Все таблицы должны быть последовательно пронумерованы: сквозная нумерация – табл. 1, табл. 2 – для небольших работ; индексационная поглавная нумерация – табл. 3.1, табл. 3.2 – для работ с несколькими пронумерованными частями. Если таблицы даются наряду с графиками, схемами и другим иллюстративным

материалом, то обычно они нумеруются отдельно. В случае если данные охватывают определенный период времени, его границы следует включить в заголовок. Если таблица частично или полностью составлена по сведениям другого источника, сразу под ней следует дать ссылку.

Хорошо сконструированная таблица позволяет исследователю более четко представить и описать смысл и сущность изучаемого им социального явления. Таким образом, метод группировки и представление материала в виде статистических таблиц уже дают определенные возможности для изучения социологических данных. С другой стороны, они являются совершенно необходимым средством для дальнейшего анализа и применения более тонких статистических методов.

# Раздел 3.
# Теоретическое отступление в математическую статистику

## 3.1. Основные понятия статистической оценки параметров выборки

Отметим методологические моменты, характеризующие процесс использования математического аппарата в социологических исследованиях.

*Причины, приводящие к использованию математики*

Стремление к четкой формулировке того или иного положения практически всегда приводит к возможности выражения его на математическом языке. Собственно, математика начинается там, где можно достаточно четко выделить в реальности интересующие аспекты, абстрагируясь от всех остальных проявлений изучаемого явления. Это было всегда. По меткому выражению Уайтхеда «Человек, заметивший аналогию между семью рыбами и семью днями, осуществил значительный сдвиг в истории мышления. Он был первым, кто ввел понятие, относящееся к науке чистой математики».

*Обеспечение адекватности формализма содержанию*

За каждым математическим методом стоит определенная модель того явления, которое с помощью этого метода изучается. Применяя метод, социолог четко должен представлять себе сущность, «содержательный» смысл этой модели. Он должен

давать себе отчет в том, в силу самого применения выбранного метода, что считает истинным, от чего абстрагируется, какие ограничения на реальность накладывает и т. д. Иначе метод перестает играть роль «орудия труда» исследователя.

*Неоднозначность математических моделей*

Сложность социальных явлений приводит к тому, что удачная формализация любого фрагмента интересующей социолога реальности, как правило, не бывает однозначной. Практически для любого явления оказывается возможной разработка целого ряда моделей, каждая из которых является естественной, но отражает лишь какой-то один его аспект. Складывается своеобразная ситуация объектов – если задачу в принципе оказывается возможным решить с помощью какого-либо формального (чаще всего математического) аппарата, то соответствующих решений, как правило, бывает несколько. И ни одно из них не может считаться «главным». Каждое отвечает определенной стороне реальности. С проблемой выбора метода постоянно имеет дело социолог, занимающийся анализом данных и использованием формального аппарата в процессе измерения.

*Взаимодействие социологии и математики в процедурах измерения*

Анализ развития многих методов измерения показывает, что имеется естественный логический процесс взаимодействия математики и реальности. Можно привести целый ряд примеров возникновения мощных ветвей прикладной статистики (метод парных сравнений, латентно-структурный анализ, многомерное шкалирование и др.). Стало быть, естественным является ожидание того, что абстрактным положениям отвечает нечто в ре-

альности, что социолог не сможет увидеть «невооруженным» (математикой) взглядом.

Основная цель статистики – получить знания объективным способом на основе наблюдений и анализа реальности. Методы математической статистики помогают обоснованно выбрать из возможных моделей ту, которая наилучшим образом соответствует имеющимся статистическим данным, характеризующим исследуемую систему. Задачи математической статистики – оценка неизвестных параметров генеральной совокупности и проверка статистических гипотез. В математической статистике обосновывается, что выводы, полученные путем анализа данных выборки наблюдений, можно распространить на всю исследуемую совокупность, а методы теории вероятностей позволяют оценить надежность этих выводов.

*Параметр выборки* (*выборочный параметр, статистика*) – некоторый показатель, вычисленный на основе данных выборки (например, среднее квадратичное отклонение случайной величины).

Если удалось установить вид распределения, к которому относится наблюдаемый или контролируемый признак, то возникает задача нахождения значений всех параметров, характеризующих данное распределение. Например, если известно, что изучаемый признак имеет распределение Пуассона (однопараметрическое), то необходимо оценить параметр $a$, которым это распределение определяется.

Пусть $X_1, X_2, ..., X_n$ – выборка, полученная в результате выборки наблюдений из некоторой генеральной совокупности $X$. По выборкам можно получить только приближенные значе-

ния неизвестного параметра $\theta$, которые служат его оценкой. Оценки, как правило, меняются от одной выборки к другой.

Оценкой $\theta^*$ (*статистической оценкой, оценочной функцией*) неизвестного параметра $\theta$ называют произвольную функцию $\theta^* = \theta^*(x_1, x_2, ..., x_n)$, зависящую от выборки $x_1,\ x_2,\ ..., x_n$. Оценка $\theta^*$, как функция от случайной выборки, является случайной величиной. Для одного и того же параметра можно построить различные оценочные функции. Основная задача статистического оценивания неизвестных параметров $\theta$ по выборке состоит в построении такой функции от имеющихся данных статистических наблюдений, которая давала бы наиболее точные приближенные значения реальных, но неизвестных исследователю значений этих параметров. В этой связи возникает необходимость в оценке приближенного значения параметра. Построение и анализ выборочного распределения является основным математическим способом исследования реальной случайной величины. Теоретическую основу применимости выборочного метода определяет *закон больших чисел*, согласно которому при неограниченном увеличении объема выборки случайные выборочные характеристики сколь угодно приближаются (сходятся по вероятности) к определенным параметрам генеральной совокупности.

Так, оценкой математического ожидания служит функция:

$$\theta^* = \frac{x_1 + x_2 + ... + x_n}{n}.$$

Пример. Пусть произведено $n$ испытаний в схеме Бернулли с неизвестной вероятностью успеха $\theta$. $X$ – число успехов в одном испытании. В результате наблюдений получена выборка

$x_1, x_2, ..., x_n$, где $x_i$ – число успехов в *i*-м испытании. Ряд распределения наблюдаемой величины $X$ имеет вид:

| $x$ | 0 | 1 |
|---|---|---|
| $P$ | $1-\theta$ | $\theta$ |

$k = x_1 + x_2 + ... + x_n$ – суммарное число успехов в $n$ испытаниях Бернулли. В качестве оценки $\theta^*$ рассмотрим функцию $\theta^* = \frac{k}{n}$. Параметр выборки $k$ распределен по биномиальному закону с параметром $\theta$. В таком случае, оценка $\theta^*$ имеет следующий ряд распределения:

| $\theta^*$ | 0 | $1/n$ | … | $m/n$ | … | 1 |
|---|---|---|---|---|---|---|
| $P$ | $(1-\theta)^n$ | $n\theta(1-\theta)^{n-1}$ | … | $C_n^m \theta^m (1-\theta)^{n-m}$ | … | $\theta^n$ |

Статистические оценки дают достоверные представления об оцениваемых параметрах, если последние обладают определенными свойствами. Укажем эти свойства.

1. Оценка $\theta^*$ неизвестного параметра $\theta$ называется *несмещенной*, если при любом объеме выборки $n$ ее математическое ожидание равно оцениваемому параметру $\theta$, т. е. $M(\theta^*) = \theta$. В противном случае оценка называется *смещенной*.

2. Разность $M(\theta^*) - \theta$ называют *смещением* (*систематической ошибкой* оценивания). Для несмещенных оценок систематическая ошибка равна нулю. Чтобы избежать систематических ошибок в сторону увеличения или уменьшения, необходимо для приближенного неизвестного параметра брать несмещенные оценки. Несмещенная оценка не всегда дает «хорошее» приближение параметра. Получаемые по различным выборкам оценки могут быть сильно рассеяны вокруг своего сред-

него значения, т. е. дисперсия оценки может быть велика. В этом случае оценка, полученная по выборке, будет сильно отличаться от оцениваемого параметра.

3. Оценка $\theta^*$ неизвестного параметра $\theta$ называется *состоятельной*, если с ростом объема выборки она сходится по вероятности к оцениваемому параметру $\theta$, т. е. для любого положительного $\varepsilon$ выполняется соотношение размер $\lim\limits_{n\to\infty} P\left\{\left|\theta^* - \theta\right| > \varepsilon\right\} = 0$.

Выполнение условия состоятельности гарантирует, что при достаточно больших выборках ($n$) не будет грубых ошибок в оценке $\theta$. Поэтому только состоятельные оценки имеют практический смысл.

*Замечание 1*. Несмещенная оценка будет состоятельной, если ее дисперсия стремится к нулю при $n \to \infty$.

*Замечание 2*. При ограниченном $n$ оценки, обладающие одновременно свойствами состоятельности и несмещенности, могут иметь разные дисперсии.

4. Оценка $\theta^*$ неизвестного параметра $\theta$ называется *эффективной* в некотором классе точечных оценок, если ее дисперсия наименьшая среди дисперсий других оценок из этого класса. Эффективная оценка имеет наименьший разброс по сравнению с другими несмещенными и состоятельными оценками относительно истинного значения оцениваемого параметра, поэтому эффективность является решающим свойством, определяющим качество оценки, и она, вообще говоря, не предполагает обязательного наличия свойства несмещенности.

*Замечание*. Для упрощения вычислений целесообразно использовать незначительно смещенные оценки или оценки, обла-

дающие большей дисперсией по сравнению с эффективными оценками, так как достичь выполнения перечисленных условий удается не всегда.

Статистические оценки параметров подразделяют по способу их представления на *точечные* и *интервальные*.

Оценка $\theta^*$, в приведенном выше примере, является несмещенной, состоятельной и эффективной оценкой неизвестной вероятности успеха $\theta$.

## 3.2. Числовые характеристики случайной величины

Если невозможно найти закон распределения, или этого не требуется, можно ограничиться нахождением значений, называемых числовыми характеристиками случайной величины. Эти величины определяют некоторое среднее значение (среднюю тенденцию) и степень их разбросанности вокруг этого среднего.

*Математическим ожиданием* дискретной случайной величины называется сумма произведений всех возможных значений случайной величины на их вероятности:

$$m_x = M(X) = x_1 p_1 + x_2 p_2 + ... + x_n p_n = \sum_{i=1}^{n} x_i p_i \ .$$

Несмещенной оценкой математического ожидания случайной величины (дискретной) называется среднее значение случайной величины, т. е.:

$$\bar{x} = \frac{x_1 + x_2 + ... + x_n}{n} \ .$$

Пример. Проведено четыре измерения (без систематических ошибок) некоторой случайной величины: 5, 6, 9, 12. Найти несмещенную оценку математического ожидания:

$$\bar{x} = \frac{5+6+9+12}{4} = 8.$$

Пример. Пусть $X$ – дискретная случайная величина, заданная законом распределения вероятностей:

| X | -1 | 3 |
|---|---|---|
| P | 0,4 | 0,6 |

Найти математическое ожидание этой случайной величины.

$$M(X) = (-1)\cdot 0{,}4 + 3\cdot 0{,}6 = 1{,}4.$$

*Математическим ожиданием* непрерывной случайной величины $X$, возможные значения которой принадлежат отрезку $[a;b]$, называется определенный интеграл

$$M(X) = \int_a^b xf(x)dx.$$

Если возможные значения случайной величины рассматриваются на всей числовой оси, то математическое ожидание находится по формуле:

$$M(X) = \int_{-\infty}^{\infty} xf(x)dx.$$

При этом, конечно, предполагается, что несобственный интеграл сходится.

*Модой* $M_0$ случайной величины называется ее наиболее вероятное значение. Для непрерывной случайной величины мода – такое значение случайной величины, при которой плотность распределения имеет максимум:

$$f(M_0) = \max.$$

Для дискретной случайной величины модой можно определить самое часто встречающееся значение измеренного признака, которым обладает максимальное число элементов выборки. Например, если исследовалось число правильно решенных учащимися задач, то модой будет такое значение, для которого количество учащихся, правильно решивших именно это число задач, максимально.

Числовые значения *моды* можно найти и по формулам:

$$M_0 = x_0 + \delta \frac{n_{M_0} - n^-}{2n_{M_0} - n^- - n^+},$$

где $x_0$ – нижняя граница модального интервала, $\delta$ – величина модального интервала, $n_{M_0}$ – частота модального интервала, $n^-$ – частота интервала, предшествующего модальному, $n^+$ – частота интервала, следующего за модальным.

Если многоугольник распределения для дискретной случайной величины или кривая распределения для непрерывной случайной величины имеет два или несколько максимумов, то такое распределение называется *бимодальным* или *полимодальным*.

*Медианой* $M_D$ случайной величины $X$ называется такое ее значение, относительно которого равновероятно получение большего или меньшего значения случайной величины $P(X < M_D) = P(X > M_D)$.

*Геометрически медиана* – абсцисса точки, в которой площадь, ограниченная кривой распределения делится пополам, иначе справа и слева от которой находится одинаковое число элементов выборки.

Расчетная формула для нахождения медианного значения имеет вид:

$$M_e = x_0 + \delta \frac{\frac{1}{2}n - n_H}{n_{M_e}},$$

где $x_0$ – нижняя граница медианного интервала, $\delta$ – величина медианного интервала, $n_{M_e}$ – частота медианного интервала, $n_H$ – частота, накопленная до медианного интервала.

Заметим, что если распределение унимодальное, то мода и медиана совпадают с математическим ожиданием. Средние тенденции помогают определить наиболее типичные значения (одно или несколько), которые наилучшим образом представляют весь комплекс признаков по данной переменной. Целесообразность использования того или иного типа средней величины зависит от цели усреднения, вида распределения, шкалы измерения признака, вычислительных соображений. Часто для характеристики рядов распределения оказывается недостаточным указание только средней величины данного признака, поскольку два ряда могут иметь одинаковые средние тенденции.

Кроме средних тенденций существует величина, которая характеризует отклонение значений случайной величины от среднего.

*Дисперсией* (*рассеиванием*) случайной величины называется математическое ожидание квадрата отклонения случайной величины от ее математического ожидания $D(X) = M\left[X - M(X)\right]^2$. Однако, на практике подобный способ вычисления дисперсии неудобен, т.к. приводит при большом количестве значений случайной величины к громоздким расче-

там. Поэтому применяется другой способ. *Дисперсия равна разности между математическим ожиданием квадрата случайной величины $X$ и квадратом ее математического ожидания* $D(X) = M(X^2) - [M(X)]^2$.

*Средним квадратичным отклонением* случайной величины $X$ называется квадратный корень из дисперсии:

$$\sigma(X) = \sqrt{D(X)}\,.$$

В практическом аспекте величины средних тенденций, вообще говоря, будут различными, поскольку разные объекты один респондент или разные респонденты один объект «в среднем» оценивают по-разному. Вероятно, естественным выглядит предложение считать «истинной» оценкой мнения респондента о рассматриваемом объекте соответствующее математическое ожидание. Однако и дисперсию рассматриваемых распределений можно проинтерпретировать естественным образом, например для нормального закона. Напомним, что нормальное распределение однозначно задается значениями математического ожидания и дисперсии. Рассмотрим рис. 3.1, на котором изображены распределения, отвечающие разным дисперсиям.

Нетрудно понять, что дисперсия говорит о степени уверенности (убежденности) респондента в своем мнении о рассматриваемом объекте. Если это мнение определяется распределением *I*, то респондент, будучи опрошенным в разное время, примерно с одинаковой вероятностью будет давать совершенно различные ответы, в том числе и весьма отличающиеся от среднего. Так, значения $x_1$ и $x_2$ в его ответах могут встретиться почти с той же вероятностью, что и среднее значение.

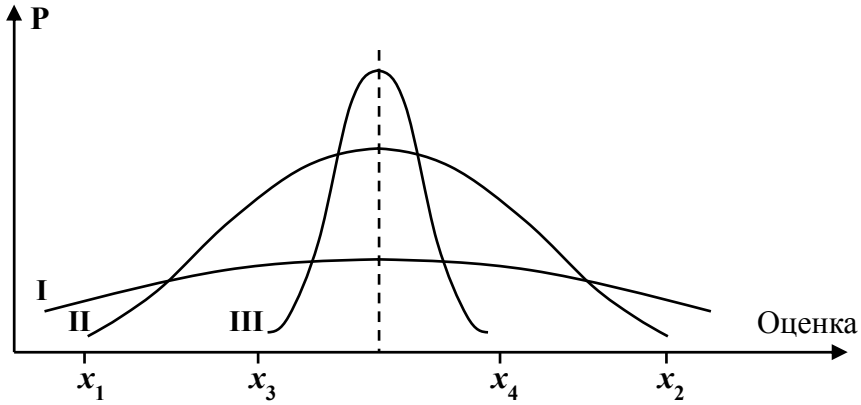



**Рис. 3.1. Нормальное распределение оценок *l*-м респондентом *i*-го объекта при разных дисперсиях**

Если мнение респондента определяется распределением III, то, напротив, значения, даже незначительно отличающиеся от среднего, такие, как $x_3$ и $x_4$ будут встречаться с меньшей вероятностью, чем само среднее. При реализации распределения II ситуация будет занимать промежуточное положение между двумя описанными выше. Ясно, что упомянутая степень уверенности может быть объяснена разными факторами: характером (принципиальностью) респондента, его знанием оцениваемых объектов, важностью этих объектов для респондента и т. д.

Рассмотрим вопрос: должны ли быть схожими, и, если должны, то в какой степени, распределения, отвечающие разным респондентам, при определенной однородности изучаемой совокупности респондентов. Покажем, что смысл задачи заставляет нас считать равными средние значения соответствующих распределений.

Предположим, что упомянутого равенства нет, то есть ситуация, отраженная на рис. 3.2 а. Наверное, социолог при нали-

чии в изучаемой совокупности таких респондентов, придет к выводу, что мнения относительно рассматриваемого объекта разделились: одним респондентам он нравится, другим – нет. В такой ситуации, вероятно, разумно разделить всю совокупность на две и для каждой из полученных совокупностей искать среднюю оценку отдельно.

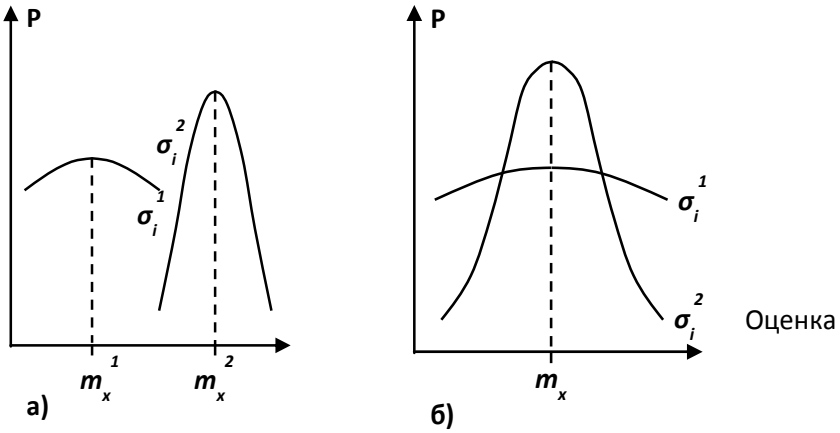



**Рис. 3.2. Распределение оценок *x*-го объекта, данных 1-м и 2-м респондентами: *а* – с разными средними и разными дисперсиями; *б* – с одинаковыми средними, но разными дисперсиями.**

Теперь, представим, что каким-то двум респондентам отвечают распределения, изображенные на рис. 3.2 б, где $\sigma^1$ и $\sigma^2$ – средние квадратичные отклонения. Один из респондентов хорошо знает объект и поэтому уверен в своих оценках. Его дисперсия мала, кривая «узка», вероятность дать ответ, сильно отличающийся от среднего, практически равна нулю. Напротив, другой респондент имеет об изучаемом объекте весьма смутное представление. Ему более или менее все равно, какие оценки

давать. Весьма сильно разнящиеся ответы могут встретиться примерно с одинаковой вероятностью. Его кривая «широка», а дисперсия велика.

Таким образом, степень концентрации (или, наоборот, разброса) значений признаков вокруг средней будет различной. Характеристикой такого разброса служат показатели колебания (вариативности) – разность между максимальным и минимальным значениями признака в некоторой совокупности (вариационный размах), а также другие показатели: квантильный ранг, среднее абсолютное (линейное) отклонение и т. п.

Кроме средних тенденций и разбросов исследователь может вычислить такие показатели, как асимметрия и эксцесс.

*Асимметрия* (коэффициент асимметрии, третий момент случайной величины) обозначает соотношение между длинами правого и левого «хвостов» распределения. Существует статистическая оценка отклонения данного распределения от симметричного, или, иначе говоря, его скошенность. Степень скошенности распределения и показывает величина его асимметрии, которая вычисляется по формуле:

$$A = \frac{\sum (x_i - \bar{x})^3}{n \cdot \sigma^3}.$$

Показатель асимметрии может быть использован для содержательной интерпретации полученных данных. Если наблюдаемый признак формируется под воздействием большого числа факторов, каждый из которых вносит свой небольшой вклад в величину этого признака, то мы вправе ожидать симметричного распределения. Однако если получена значительная величина асимметрии (большая по абсолютной величине, чем 0,4-0,5),

можно предположить, что присутствует значительное влияние одного или группы факторов.

*Эксцесс* характеризует островершинность распределения. Так, например, величина эксцесса для нормальной (гауссовой) кривой распределения $E_x = 3$. Исходя из целого ряда соображений, заостренность этой кривой принимают за стандарт, поэтому в качестве показателя эксцесса используют величину $y = E_x - 3$. Собственно сам эксцесс может быть вычислен по формуле:

$$E = \frac{\sum (x_i - \bar{x})^4}{n \cdot \sigma^4} - 3 .$$

Величина эксцесса может принимать очень большие значения, но он не может быть меньше единицы. Близость к единице свидетельствует о бимодальности (двувершинности) распределения. Бимодальность кривой распределения эмпирических данных может возникать за счет объединения в единую совокупность двух наборов разнородных измерений.

Кривые, полученные в результате графического представления эмпирических данных, могут иметь разнообразную форму. Среди них можно выделить относительно небольшое количество простых типов. Некоторые, наиболее распространенные, приведены на рис. 3.3.

Анализ формы кривых иногда поможет в выявлении внутренней, скрытой структуры исследуемой совокупности. Например, можно предположить, что форма кривой в) обусловлена наложением друг на друга двух кривых а) и б), иначе говоря, существует третья скрытая переменная (или группа переменных), определяющая деление совокупности на две группы.

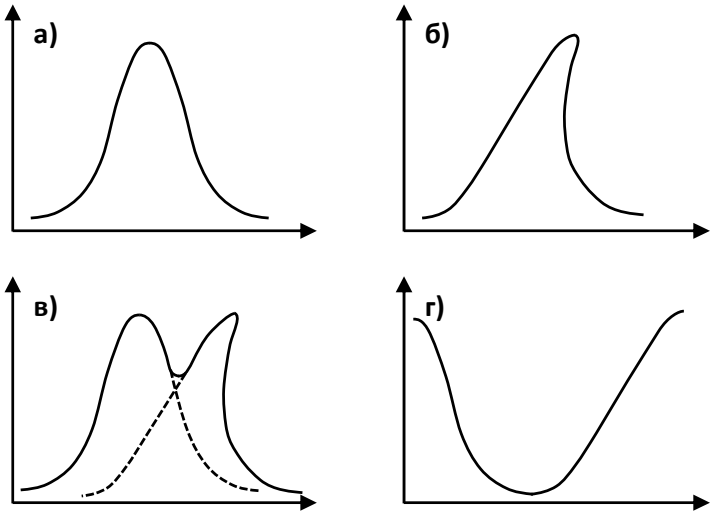



**Рис. 3.3. Различные формы кривых распределения**

## 3.3. Точечные и интервальные оценки

Точечной называют статистическую оценку $\theta^* = \theta^*(X_1, X_2, ..., X_n)$ параметра $\theta$, определяемую *одним* числом. Точечные оценки зависят от объема выборки и обычно используются в выборках большого объема. Оценки параметров совокупности, полученные по разным выборкам, как правило, отличаются друг от друга. *Ошибкой выборки* (оценивания) называют абсолютную разность $\left|\theta^* - \theta\right|$.

Выборочная средняя $\bar{x}_в = \frac{1}{n}\sum_{i=1}^{k} x_i \cdot n_i$ есть несмещенная и состоятельная точечная оценка математического ожидания $M(X)$. Выборочная дисперсия $D_в = \frac{1}{n}\sum_{i=1}^{k}(x_i - \bar{x}_в)^2 \cdot n_i$ – это точечная смещенная оценка генеральной дисперсии $D(X)$. Не-

смещенной оценкой генеральной дисперсии $D(X)$ является «исправленная» выборочная дисперсия $S_в^2 = \frac{n}{n-1} \cdot D_в$.

«Исправленная» выборочная дисперсия является состоятельной оценкой дисперсии $D(X)$. При достаточно больших значениях $n$ выборочная и «исправленная» дисперсии различаются мало.

Если известно математическое ожидание $M(X)$ случайной величины $X$, то выборочная дисперсия $D_в = \frac{1}{n}\sum_{i=1}^{k}(x_i - M(X))^2 \cdot n_i$ – несмещенная, состоятельная и эффективная оценка генеральной дисперсии $D(X)$.

Относительная частота $\frac{n_i}{n}$ – несмещенная и состоятельная оценка вероятности $P(X = x_i)$.

Пример. Необходимо найти несмещенную оценку дисперсии по данным, приведенным в таблице.

| **Федеральный округ** | **Средний размер заработной платы, $x_i$, тыс. руб.** | **Численность занятых, $n_i$, млн. чел.** |
|---|---|---|
| Центральный | 35,67 | 20,33 |
| Северо-Западный | 32,29 | 7,15 |
| Южный | 22,38 | 6,41 |
| Северо-Кавказский | 19,30 | 3,95 |
| Приволжский | 22,41 | 14,74 |
| Уральский | 34,03 | 6,08 |
| Сибирский | 26,48 | 9,05 |
| Дальневосточный | 37,26 | 3,19 |

*Решение*. Найдем средний размер заработной платы по всем округам:

$$\bar{x}_в = \frac{\sum_{i=1}^{8} x_i \cdot n_i}{\sum_{i=1}^{8} n_i} = (35{,}67 \cdot 20{,}33 + 32{,}29 \cdot 7{,}15 + 22{,}38 \cdot 6{,}41 +$$

$$+ 19{,}30 \cdot 3{,}95 + 22{,}41 \cdot 14{,}74 + 34{,}03 \cdot 6{,}08 + 26{,}48 \cdot 9{,}05 +$$

$$+ 37{,}26 \cdot 3{,}19 : (20{,}33 + 7{,}15 + 6{,}41 + 3{,}95 + 14{,}74 + 6{,}08 +$$

$$+ 9{,}05 + 3{,}19) = 2071{,}4646/70{,}9 \approx 29{,}21670804.$$

Для вычисления выборочной дисперсии воспользуемся формулой $D_в = \overline{x^2} - (\bar{x}_в)^2$, где $\overline{x^2}$ – выборочная средняя квадратов вариантов выборки.

$$\sum_{i=1}^{8} x_i^2 \cdot n_i = (35{,}67^2 \cdot 20{,}33 + 32{,}29^2 \cdot 7{,}15 + 22{,}38^2 \cdot 6{,}41 +$$

$$+ 19{,}30^2 \cdot 3{,}95 + 22{,}41^2 \cdot 14{,}74 + 34{,}03^2 \cdot 6{,}08 + 26{,}48^2 \cdot 9{,}05 +$$

$$+ 37{,}26^2 \cdot 3{,}19) = 63221{,}545186,$$

$$\overline{x^2} = \frac{\sum_{i=1}^{8} {x_i}^2 \cdot n_i}{\sum_{i=1}^{8} n_i} = \frac{63221{,}545186}{70{,}9} \approx 891{,}7002142,$$

$$D_в = 891{,}7002142 - (29{,}21670804)^2 \approx 38{,}08418554.$$

Найдем несмещенную оценку дисперсии:

$$S_в^2 = \frac{n}{n-1} \cdot D_в = \frac{70{,}9}{69{,}9} \cdot 38{,}08418554 \approx 38{,}62902367.$$

При большом объеме выборки выборочное среднее, выборочную и исправленную дисперсии удобно вычислять, используя вспомогательную таблицу:

| $x_i$ | $n_i$ | $x_i \cdot n_i$ | $x_i - \overline{x_в}$ | $(x_i - \overline{x_в})^2$ | $(x_i - \overline{x_в})^2 \cdot n_i$ |
|---|---|---|---|---|---|
| $x_1$ | $n_1$ | $x_1 \cdot n_1$ | $x_1 - \overline{x_в}$ | $(x_1 - \overline{x_в})^2$ | $(x_1 - \overline{x_в})^2 \cdot n_1$ |
| ... | ... | ... | ... | ... | ... |
| $x_k$ | $n_k$ | $x_k \cdot n_k$ | $x_k - \overline{x_в}$ | $(x_k - \overline{x_в})^2$ | $(x_k - \overline{x_в})^2 \cdot n_k$ |
| Сумма | $n = \sum_{i=1}^{k} n_i$ | $\sum_{i=1}^{k} x_i \cdot n_i$ | | | $\sum_{i=1}^{k} (x_i - \overline{x_в})^2 \cdot n_i$ |

*Замечание*. Для проведения статистической обработки информации используют также табличный процессор Microsoft Excel, включающий в себя программную надстройку «Пакет анализа» и библиотеку статистических функций. В повседневной деятельности такого набора инструментов бывает вполне достаточно для проведения довольно полного и качественного статистического анализа информации.

Статистическая оценка параметров закона распределения случайной величины $X$, характеризующаяся двумя числовыми значениями – концами интервала, называется *интервальной*.

Интервал, в который с заданной вероятностью (надежностью), попадает оцениваемый параметр, называется *доверительным*. На практике обычно используют два типа доверительных интервалов: *симметричные* и *односторонние*.

Вероятность $\gamma$, с которой выполняется неравенство $|\theta^* - \theta| < \varepsilon$, называется *доверительной вероятностью* (*надежностью*) оценки $\theta$ для заданного $\varepsilon$. Точностью оценки $\varepsilon$ называют число, равное половине длины доверительного интервала.

Доверительная вероятность связана не с величиной параметра $\theta$, а лишь с границами интервала, которые меняются при изменении выборки. Вероятность $\alpha = 1 - \gamma$ называется *уровнем значимости* (*вероятностью ошибок*). Общепринятые значения уровня значимости в социологических исследованиях – 0,95; 0,99.

*Доверительными границами* (*критическими значениями*) называют границы доверительного интервала. Доверительные границы зависят от закона распределения параметра $\theta^*$. Доверительный интервал, определяемый выборкой – случайной величиной, носит случайный характер и относительно длины, и по расположению относительно $\theta^*$, поэтому принято говорить, что доверительный интервал покрывает параметр $\theta$ с доверительной вероятностью $\gamma$.

Точность оценки $\varepsilon$, доверительная вероятность $\gamma$ и объем выборки $n$ связаны между собой. Методы построения доверительных интервалов различны. Впервые были разработаны американским статистиком Ю. Нейманом, использовавшим идеи англичанина Р. Фишера.

Алгоритм построения доверительного интервала состоит в следующем:

1. Производим выборку объема $n$ случайной величины $X$ из генеральной совокупности с известным распределением $f(x;\theta)$.

2. Находим точечную оценку $\theta^*$ неизвестного параметра $\theta$ по данным выборки.

3. Задаем доверительную вероятность $\gamma$.

4. Определяем границы доверительного интервала $(\theta^* - \varepsilon; \theta^* + \varepsilon)$. Для этого возьмем произвольное значение

$\theta$ и, используя плотность вероятности, найдем функцию распределения из условия

$$P\left\{\left|\theta-\theta^*\right|<\varepsilon\right\}=\int_{\theta^*-\varepsilon}^{\theta^*+\varepsilon} f(x,\theta)dx=\gamma\,.$$

Границы интервала определим из решения уравнений:

$$P(X(\theta)<\theta^*-\varepsilon)=\frac{1-\gamma}{2},$$

$$P(X(\theta)>\theta^*+\varepsilon)=\frac{1-\gamma}{2}.$$

Полученный интервал с доверительной вероятностью $\gamma$ покрывает неизвестный параметр $\theta$ и является его интервальной оценкой.

*Замечание*. При малом объеме выборки построение доверительных интервалов трудоемко, так как оно сводится к перебору значений неизвестного параметра.

Пример. Социологическое обследование образования 20 сотрудников организации дало следующие результаты.

| Номера анкеты | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Число лет, затраченных на получение образо-вания | 15 | 14 | 15 | 15 | 10 | 13 | 17 | 18 | 15 | 14 | 15 | 15 | 13 | 10 | 18 | 16 | 15 | 14 | 15 | 15 |

В данной организации считается, что в среднем 15 лет (10 лет в школе и 5 лет в вузе) – стандартный уровень образования для сотрудника, а если сотрудник обучался менее 15 лет, его уровень образования недостаточен. Оцените, можно ли на уров-

не значимости $\alpha = 0{,}05$ и $\alpha = 0{,}01$ говорить о том, что средний уровень образования сотрудников (оцениваемый с помощью среднего арифметического) ниже стандарта. Общее число сотрудников организации составляет 142 человека.

*Решение*. Найдем число степеней свободы $\upsilon = 142 - 1 = 141$. Значение *t*-статистики = 1,98 для $(\alpha = 0{,}05; \upsilon = 141)$ и значение *t*-статистики = 2,61 для $(\alpha = 0{,}01; \upsilon = 141)$.Рассчитаем среднее арифметическое и стандартное отклонение: $\bar{x} = 14{,}6$, $s = 2{,}06$.

При $\alpha = 0{,}05$ границы доверительного интервала находим по формуле:

$$\bar{x} \pm 1{,}98\frac{S}{\sqrt{n}} = 14{,}6 \pm 1{,}98\frac{2{,}00}{\sqrt{142}} = 14{,}6 \pm 0{,}34 .$$

Отсюда первый интервал:

$14{,}26 \le \mu \le 14{,}94$.

Находим границы доверительного интервала при $\alpha = 0{,}01$:

$$\bar{x} \pm 2{,}58\frac{S}{\sqrt{n}} = 14{,}6 \pm 0{,}45 .$$

Отсюда второй интервал:

$14{,}15 \le \mu \le 15{,}05$.

Таким образом, на уровне значимости $\alpha = 0{,}01$ уровень образования отвечает стандарту, а вот на уровне $\alpha = 0{,}05$ он находится несколько ниже стандарта. Обратим внимание, что во втором случае доверительный интервал меньше, но вероятность ошибки выше. Иными словами, если действительно уровень образования в организации соответствует стандарту, вероятность ошибочно заключить, что это не так, выше при $\alpha = 0{,}05$, чем при $\alpha = 0{,}01$.

## 3.4. Основные законы распределения и их статистики

Непрерывная случайная величина имеет равномерное распределение на отрезке [a, b], если на этом отрезке плотность распределения случайной величины постоянна, а вне этого отрезка равна нулю:

$$f(x)=\begin{cases}0, & x<a\\ C, & a\le x\le b\\ 0, & x>b\end{cases}.$$

Постоянная величина $C$ может быть определена из условия равенства единице площади, ограниченной кривой распределения. График приведен на рисунке 3.4.

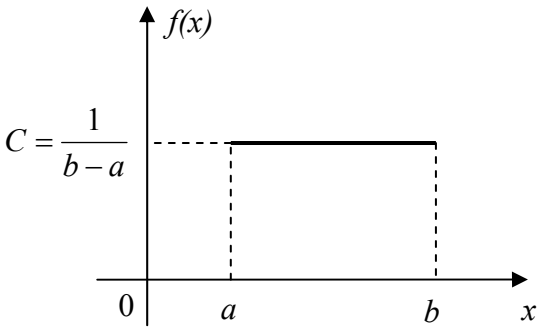
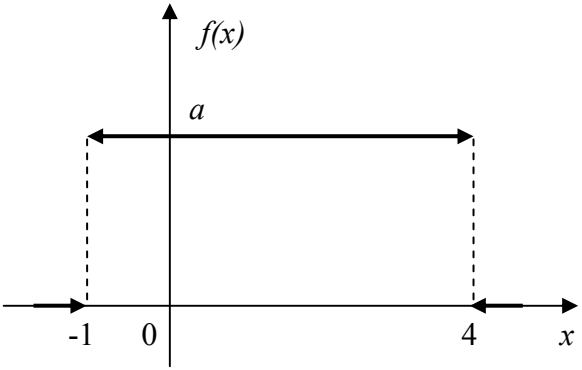



**Рис. 3.4. Плотность вероятности равномерного распределения**

Математическое ожидание и дисперсия равномерно распределенной непрерывной случайной величины вычисляются по формулам:

$$M(X)=\frac{a+b}{2};\ D(X)=\frac{(b-a)^2}{12}.$$

Пример. Задан график плотности распределения вероятностей непрерывной случайной величины $X$, распределенной равномерно в интервале (-1; 4). Найти значение $M(X)$.




$$M(X)=\frac{1}{4-(-1)}=\frac{1}{5}=0{,}2\,.$$

Пример. Случайная величина распределена равномерно на интервале (-10; 12). Найти ее математическое ожидание и дисперсию.

Параметры распределения $a=10$, $b=12$, тогда $M(X)=11$, $D(X)=\frac{1}{3}$.

*Нормальным* называется распределение вероятностей непрерывной случайной величины, которое описывается плотностью вероятности:

$$f(x)=\frac{1}{\sigma_x\sqrt{2\pi}}e^{-\frac{(x-m_x)^2}{2\sigma_x^2}};$$

Параметры $m_x$ и $\sigma_x$, входящие в плотность распределения являются соответственно математическим ожиданием и средним квадратичным отклонением случайной величины $X$.

В социологических исследованиях чаще всего ссылаются на нормальное распределение (распределение Гаусса), так как оно характеризуется тем, что крайние значения признака в нем встречаются достаточно редко, а значения, близкие к средней величине, – достаточно часто. Название это распределение получило, вследствие его распространенности в естественнонаучных исследованиях, что казалось нормой всякого массового случайного проявления признаков. График нормального распределения представляет собой колоколообразную кривую (рис. 3.5).

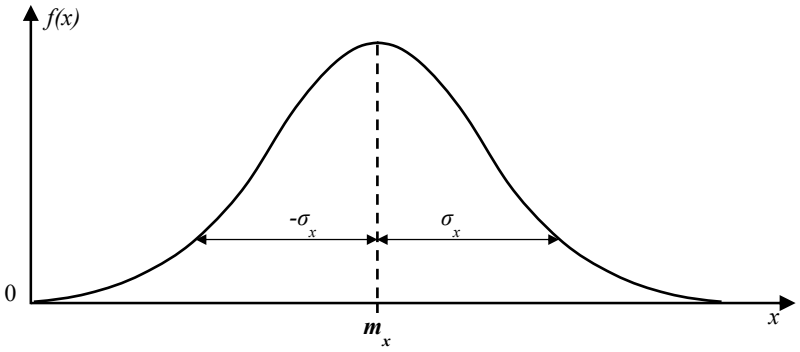



**Рис 3.5. Теоретическая кривая нормального распределения**

Нормальное распределение почти всегда имеет место, когда наблюдаемые случайные величины формируются под влиянием большого числа случайных факторов, ни один из которых существенно не превосходит остальные. На рисунке 3.6 приведены графики нормального распределения при $m_x = 0$ и трех возможных значениях среднего квадратичного отклонения $\sigma_x = 1$ (верхняя кривая), $\sigma_x = 2$ (средняя кривая) и $\sigma_x = 7$ (нижняя кривая).

Как видно, при увеличении значения среднего квадратичного отклонения график становится более пологим, а максимальное (модальное) значение уменьшается.

Нормальное распределение в математической статистике хорошо изучено. Следовательно, для него существуют статистические таблицы, которые позволяют по каждому значению случайной величины найти вероятность его встречаемости, а по каждой вероятности – значения, которые с этой вероятностью встречаются.

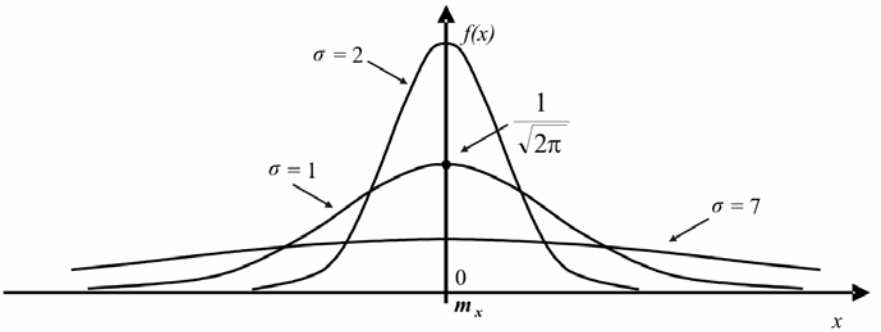



**Рис. 3.6. Варианты графиков нормального распределения**

Однако надо сделать некоторые дополнительные замечания. Дело в том, что известные статистические таблицы разработаны для, так называемого, стандартизованного нормального распределения. Иначе, для таких случайных величин, которым отвечает нулевое среднее и единичная дисперсия (поскольку, нельзя рассчитать таблицы для всех мыслимых нормальных распределений, у которых в качестве математического ожидания и дисперсии – любые положительные действительные числа). Соответствующая функция распределения носит название функции Лапласа и имеет вид:

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x_i} e^{-z^2/2} dz \text{, где } z = \frac{x - \mu}{\sigma}.$$

На практике случайные величины обычно не стандартизованы. Поскольку признаки могут изменяться в разных пределах (например, доход может варьировать от нуля до тысяч и миллионов единиц измерения, а число детей в семье редко превышает пять-шесть), при сравнении таких признаков иногда требуется привести их к одному виду, к единому стандарту. Наиболее распространенный вариант стандартизации – это приведение распределения к стандартизированному виду (z-стандартизация). Это достигается путем вычитания из каждого значения математического ожидания (децентрация) и деления полученной разности на стандартное отклонение (если эти параметры неизвестны, используются их выборочные оценки):

$$x' = \frac{x - \mu}{\sigma} \text{(или)} \, x' = \frac{x - x'}{S}.$$

Существуют и другие варианты стандартизации. Например, наглядной бывает стандартизация, преобразующая значения шкалы таким образом, чтобы они находились в интервале от 0 до 1:

$$x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}}.$$

Другие варианты:

максимальный размах значений равен единице $\left( x' = \frac{x}{x_{\max} - x_{\min}} \right)$;

математическое ожидание равное единице $\left( x' = \frac{x}{\mu} \right)$;

стандартное отклонение равно единице $\left( x' = \frac{x}{\sigma} \right)$.

В случаях, когда содержательный смысл признака подсказывает, что диапазон его изменения должен быть разбит на отрицательную и положительную части, оправдана стандартизация, приводящая значения случайной величины к интервалу от –1 до +1:

$$\left( x' = \frac{2x - 2x_{\min}}{x_{\max} - x_{\min}} - 1 \right).$$

Выбор корректного способа стандартизации зависит как от содержательных соображений, так и от статистических последствий преобразования.

Пример. При приеме на работу предприятие для повышения квалификации обязуется ежегодно организовывать сотрудникам не менее трех командировок за границу. Обследование показало, что распределение сотрудников по интенсивности таких командировок описывается следующим законом.

| Число командировок для повышения квалификации | 0 | 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|
| Число сотрудников | 26 | 43 | 27 | 12 | 5 |

Профсоюз утверждает, что командировки для повышения квалификации проводятся в компании слишком редко – вероятность получить обещанные предприятием три командировки в год не превышает 5%. Проведите z-стандартизацию и оцените, насколько корректны претензии профсоюза со статистической точки зрения (предполагая, что речь идет о нормальном распределении).

*Решение*. Поскольку речь идет об оценке вероятности того, что случайная величина примет значение больше $x_i$ следует ис-

пользовать одностороннюю критическую область. Чтобы оценить, справедливо ли утверждение $P(X \geq 3) \leq 0{,}05$ найдем соответствующее значение в таблицах для односторонней критической области. Оно составляет 1,64. Таким образом, если $X = 3$ отклоняется от математического ожидания на $1{,}64\sigma$ и более, выводы профсоюза верны. Чтобы оценить вероятность того, что случайная величина примет значение $X \geq 3$, проведем z-стандартизацию и воспользуемся таблицей стандартизованного нормального распределения:

$$\bar{x} = \frac{0 \cdot 26 + 1 \cdot 43 + 2 \cdot 27 + 3 \cdot 12 + 4 \cdot 5}{113} = 1{,}35;$$

$$\sigma = \sqrt{\frac{26(0-1{,}35)^2 + 43(1-1{,}35)^2 + 27(2-1{,}35)^2 + 12(3-1{,}35)^2 + 5(4-1{,}35)^2}{112}} = 1.17 \cdot$$

Отсюда стандартизированный ряд выглядит следующим образом:

| Исходные значения | 0 | 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|
| Стандартизированные значения | -1,16 | -0,30 | 0,56 | 1,41 | 2,27 |

Итак, значение $X = 3$ отклоняется от математического ожидания только на $1{,}41\sigma$ (что примерно соответствует вероятности 8%), и претензии профсоюза нельзя считать статистически корректными.

Существуют и другие распространенные в социологических исследованиях теоретические распределения. Все они, как правило, представляют семейства распределений; и каждое, в рамках одного семейства отличаются друг от друга некоторыми статистическими параметрами. Кроме рассмотренных, равномерного и нормального распределений, значительную роль в

теории статистической оценки играют распределения Пирсона, Стъюдента и Фишера.

*Семейство распределений Пирсона хи-квадрат* ($\chi^2$). Распределения $\chi^2$отличаются друг от друга числом степеней свободы. Соответственно для того чтобы определить вероятность появления некоторого значения случайной величины, требуется предварительно установить, сколько степеней свободы имеет данная величина, т. е. каким из распределений $\chi^2$ она характеризуется. По мере увеличения числа степеней свободы асимметрия и эксцесс распределения уменьшаются. Количество степеней свободы указывает, квадраты скольких величин использовались для получения величины $\chi^2$. Если взять из исходной совокупности по два значения, возвести их в квадрат и сложить, получим величину $\chi^2$ с двумя степенями свободы. Аналогично можно получить кривые распределения для сумм произвольного числа квадратов стандартизированных величин. Семейство распределений Пирсона характеризуется свойством: при больших значениях $n$ распределение $\chi^2$ стремится к нормальному. В приложении приведена таблица значений $\chi^2$ для разного числа степеней свободы и уровней доверительной вероятности. Как ею пользоваться? Например, что означает число 11,07 для $\alpha = 0{,}05$ и $\eta = 5$? Это число показывает, что сумма квадратов пяти значений, случайно выбранных из нормального распределения, только в пяти случаях из ста будет превышать величину 11,07. В остальных 95 случаях она будет меньше. Другими словами, вероятность получить значение критерия Пирсона в интервале от 0 до 11,07 равна 0,95. Критическая область для распределения $\chi^2$ всегда односторонняя (рис. 3.7).

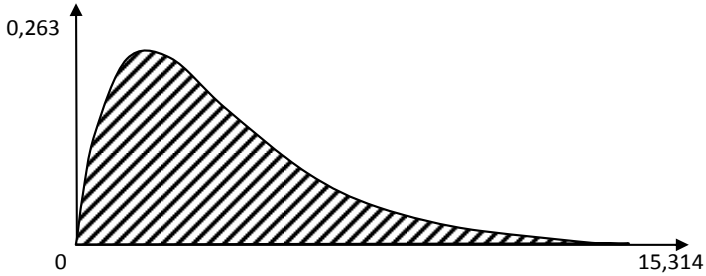



**Рис. 3.7. Функция плотности распределения $\chi^2$ ($\upsilon$= 4)**

*Семейство распределений Стьюдента*. t-распределения также меняются в зависимости от числа степеней свободы. При большом числе степеней свободы распределение Стьюдента приближается к нормальному распределению. Эксцесс распределения Стьюдента обратно пропорционален числу степеней свободы, а асимметрия при числе степеней свободы больших трех равна нулю. Критическая область распределения Стьюдента может быть как одно-, так и двусторонней, в зависимости от того, какую вероятность требуется определить (рис. 3.8).

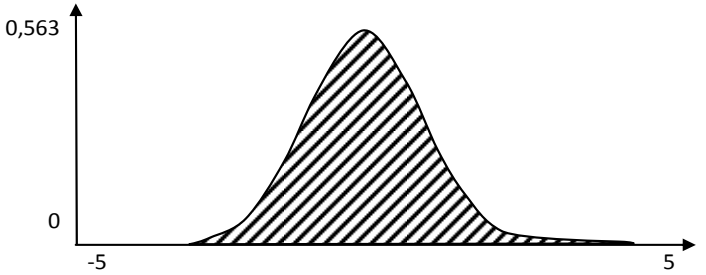



**Рис. 3.8. Функция плотности распределения Стьюдента ($\upsilon$ = 21)**

*Семейство распределений Фишера* (*F-распределения*). Распределение Фишера существует в виде семейства распределений, которые задаются двумя значениями степеней свободы. Оно представляет собой отношение двух случайных величин, распределен-

ных по закону $\chi^2$, стандартизованных на степени свободы. Поэтому *F*-распределение называют также распределением дисперсионного отношения. Критическая область *F*-распределения может быть односторонней – если проверяется гипотеза о наличии превосходства какой-либо из двух величин; или двусторонней – если проверяется гипотеза о превосходстве первой величины над второй и второй над первой (соответственно в данном случае вычисляются два критических значения $F_1$и $F_2$,причем $F_2 = 1 / F_1$).Асимметрия и эксцесс *F*-распределения зависят от соотношения между степенями свободы (рис. 3.9).

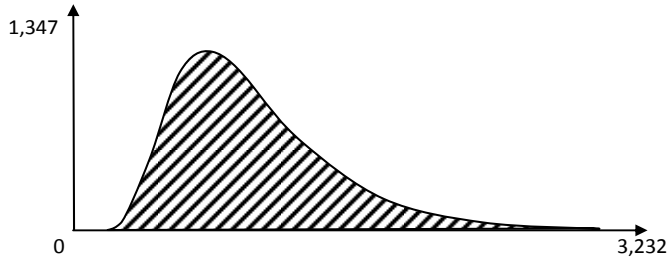



**Рис. 3.9. Функция плотности распределения Фишера ($\upsilon_1 = 22, \upsilon_2 = 16$)**

Практические расчеты с использованием приведенных распределений требуют тщательной группировки данных и достаточно сложных вычислений. Кроме того, возможности критериев согласия в полной мере проявляются на больших выборках, когда количество измерений превышает тридцать случаев. Одновременно они незаменимы для получения оценок в двух случаях: доказательство неслучайности предпочтений в выборе из нескольких альтернатив; обнаружение точки максимального расхождения между двумя распределениями, которая затем используется для перегруппировки данных.

# Раздел 4.
# Шкалы измерения

## 4.1. Основы теории измерений

Теория измерений является одной из составных частей прикладной статистики. Использование чисел в жизни и хозяйственной деятельности людей отнюдь не всегда предполагает, что эти числа можно складывать и умножать, производить иные арифметические действия. Что бы вы сказали о человеке, который занимается умножением автомобильных номеров? И не всегда 2 + 2 = 4. Если вечером поместить в клетку двух животных, а потом еще двух, то не всегда можно утром найти в этой клетке четырех животных. Их, например, может стать меньше – если к двум волкам поместили двух ягнят. Числа используются гораздо шире, чем арифметика.

*Шкала* – числовая система, в которой отношения между различными свойствами изучаемых явлений, процессов переведены в свойства того или иного множества, как правило – множества чисел. Можно выделить дискретные шкалы (в которых множество возможных значений оцениваемой величины конечно – например, школьная оценка в баллах – «1», «2», «3», «4», «5») и непрерывные шкалы (например, время, затрачиваемое учащимися на выполнение задания, в минутах).

Будем считать интуитивно ясным понятие *признака* – синонимы: переменная, характеристика, параметр, величина. Примеры: пол, возраст, удовлетворенность респондента работой.

И *его значения* – синонимы: градация, категория, альтернатива. Примеры: мужчина, 25 лет, совершенно не удовлетворен работой. Переменную, значения которой нельзя получить сразу, задав, скажем, определенный вопрос в анкете и получив соответствующий ответ респондента, будем называть *латентной* (скрытой). В противоположном случае будем говорить о *наблюдаемой* переменной. Процесс получения значений наблюдаемой переменной называется *прямым измерением*.

При проведении экспертиз проектных разработок или существующих процессов мнения экспертов часто выражены в *порядковой шкале* (подробнее о шкалах говорится ниже), т. е. эксперт может сказать, что один показатель более важен, чем другой, первый технологический объект более опасен, чем второй, и т. д. Но он не в состоянии сказать, *во сколько раз* или *на сколько* более важен, а значит более опасен. Экспертов просят провести ранжирование (упорядочение) объектов экспертизы, т. е. расположить их в порядке возрастания (или убывания) интенсивности интересующей организаторов экспертизы характеристики. *Ранг* – это номер объекта экспертизы в упорядоченном ряду значений характеристики у различных объектов. Формально ранги выражаются числами 1, 2, 3, ..., но с этими числами нельзя делать привычные арифметические операции. Например, хотя в арифметике 1 + 2 = 3, но нельзя утверждать, что для объекта, стоящем на третьем месте в упорядочении, интенсивность изучаемой характеристики равна сумме интенсивностей объектов с рангами 1 и 2.

Один из видов экспертного оценивания – школьные оценки учащихся. Вряд ли можно утверждать, что знания отличника

равны сумме знаний двоечника и троечника (хотя 5 = 2 + 3), хорошист соответствует двум двоечникам (2 + 2 = 4), а между отличником и троечником такая же разница, как между хорошистом и двоечником (5 –3 = 4 – 2). Поэтому очевидно, что для анализа подобного рода качественных данных необходима не всем известная арифметика, а другая теория, дающая базу для разработки, изучения и применения конкретных методов расчета. Уже в конце 1960-х годов было установлено, что баллы, присваиваемые экспертами при оценке объектов, измерены в порядковой шкале. Это положение характерно для задач педагогической квалиметрии, для проблем теории экспертных оценок, для агрегирования показателей качества продукции, в социологических исследованиях и др. Итогом теоретических и практических исследований явилась формулировка двух проблем: наряду с установлением типа шкалы измерения конкретных данных был выдвинут поиск *алгоритмов анализа* данных, результат работы которых не меняется при любом допустимом преобразовании шкалы (является инвариантным относительно этого преобразования). В соответствии с положениями теории измерений при моделировании реального явления или процесса следует, прежде всего, установить типы шкал, в которых измерены те или иные переменные. Тип шкалы задает группу допустимых преобразований шкалы. *Допустимые преобразования* не меняют соотношений между объектами измерения. Например, при измерении длины переход от футов к метрам не меняет соотношений между длинами рассматриваемых объектов – если первый объект длиннее второго, то это будет установлено и при измерении в футах, и при измерении в метрах. Обратим внимание, что

при этом численное значение длины в футах отличается от численного значения длины в метрах – не меняется лишь результат сравнения длин двух объектов. Установление типа шкалы, то есть задания группы допустимых преобразований шкалы измерения – дело специалистов соответствующей предметной области. Так, оценки привлекательности профессий можно считать измеренными в порядковой шкале. Однако отдельные социологи полагают, что выпускники школ пользуются шкалой с более узкой группой допустимых преобразований, например, интервальной шкалой. Очевидно, эта проблема относится не к математике, а к наукам о человеке.

Укажем основные виды шкал измерения и соответствующие им группы допустимых преобразований. Строго говоря, все шкалы измерения делят на две группы – шкалы качественных признаков и шкалы количественных признаков. Основные шкалы качественных признаков – порядковая шкала и шкала наименований. Поэтому во многих конкретных областях результаты качественного анализа можно рассматривать как измерения по этим шкалам.

В шкале *наименований* (другое название этой шкалы – *номинальная*) числа используются лишь как метки. Примерами такой шкалы являются, например, номера телефонов, паспортов, студенческих билетов, страховых свидетельств государственного пенсионного страхования, медицинского страхования, ИНН (индивидуальный номер налогоплательщика). Пол людей тоже измерен в шкале наименований, результат измерения принимает два значения – мужской, женский. Раса, национальность, цвет глаз, волос – номинальные признаки. Бессмысленно складывать

или умножать номера телефонов или сравнивать буквы и говорить, например, что буква П лучше буквы С, потому что она в алфавите стоит раньше (то есть, имеет меньший порядковый номер). Единственное, для чего годятся измерения в шкале наименований – это различать объекты. Во многих случаях только это от них и требуется. Например, шкафчики в раздевалках для взрослых различают по номерам, т. е. числам, а в детских садах используют рисунки, поскольку дети еще не знают чисел. Для такой шкалы допустимыми являются все взаимно-однозначные преобразования.

В порядковой шкале числа используются не только для различения объектов, но и для установления порядка между объектами. Простейшим примером являются оценки знаний учащихся. Символично, что в средней школе применяются оценки 2, 3, 4, 5, а в высшей школе ровно тот же смысл выражается словесно – неудовлетворительно, удовлетворительно, хорошо, отлично. Этим подчеркивается «нечисловой» характер оценок знаний учащихся. Почему мнения экспертов естественно выражать именно в порядковой шкале? Как показали многочисленные опыты, человек более правильно (и с меньшими затруднениями) отвечает на вопросы качественного, например, сравнительного, характера, чем количественного. Ему легче сказать, какая именно из двух гирь тяжелее, чем указать их примерный вес в граммах. В различных областях человеческой деятельности применяется много видов порядковых шкал. Так, например, в минералогии используется шкала Мооса, по которому минералы классифицируются согласно критерию твердости. А именно: тальк имеет балл 1, гипс – 2, кальций – 3, флюорит – 4, апатит –

5, ортоклаз – 6, кварц – 7, топаз – 8, корунд – 9, алмаз – 10. Минерал с большим номером является более твердым, чем минерал с меньшим номером и при нажатии царапает его. Порядковыми шкалами в географии являются – бофортова шкала ветров («штиль», «слабый ветер», «умеренный ветер» и т. д.), шкала силы землетрясений – шкала Рихтера. Очевидно, нельзя утверждать, что землетрясение в 2 балла (лампа качнулась под потолком – такое бывает и в Москве) ровно в 5 раз слабее, чем землетрясение в 10 баллов (полное разрушение всего на поверхности земли).

В медицине порядковыми шкалами являются – шкала стадий гипертонической болезни (по Мясникову), шкала степеней сердечной недостаточности (по Стражеско-Василенко-Лангу), шкала степени выраженности коронарной недостаточности (по Фогельсону) и т. д. Все эти шкалы построены по одной схеме: заболевание не обнаружено; первая стадия заболевания; вторая стадия; третья стадия и др. Каждая стадия имеет свойственную только ей медицинскую характеристику. При описании групп инвалидности числа используются в противоположном порядке: самая тяжелая – первая группа инвалидности, затем – вторая, самая легкая – третья. Номера домов также измерены в порядковой шкале – они показывают, в каком порядке стоят дома вдоль улицы. Номера томов в собрании сочинений писателя или номера дел в архиве предприятия обычно связаны с хронологическим порядком их создания. При оценке качества продукции и услуг единица продукции оценивается как годная или не годная. При более тщательном анализе используется шкала с тремя градациями: есть значительные дефекты – присутствуют только не-

значительные дефекты – нет дефектов. Иногда применяют четыре градации: имеются критические дефекты (делающие невозможным использование) – есть значительные дефекты – присутствуют только незначительные дефекты – нет дефектов. Аналогичный смысл имеет сортность продукции – высший сорт, первый сорт, второй сорт,… В порядковой шкале допустимыми являются строго возрастающие преобразования.

Шкалы количественных признаков – это шкалы интервалов, отношений, разностей, абсолютная. По шкале *интервалов* измеряют величину потенциальной энергии или координату точки на прямой. В этих случаях на шкале нельзя отметить ни естественное начало отсчета, ни естественную единицу измерения. Исследователь должен сам задать точку отсчета и сам выбрать единицу измерения. Допустимыми преобразованиями в шкале интервалов являются линейные возрастающие преобразования, т. е. линейные функции. Например, температурные шкалы Цельсия и Фаренгейта связаны именно такой зависимостью: при переводе из шкалы Фаренгейта в шкалу Цельсия из исходной цифры вычитают 32 и умножают на 5/9.

В шкале *отношений* есть естественное начало отсчета – нуль, но нет естественной единицы измерения. По шкале отношений измерены большинство физических единиц: масса тела, длина, заряд, а также цены в экономике. Допустимыми преобразованиями шкале отношений являются подобные (изменяющие только масштаб). Другими словами, линейные возрастающие преобразования без свободного члена. Примером является пересчет цен из одной валюты в другую по фиксированному курсу. Предположим, мы сравниваем экономическую эффективность

двух инвестиционных проектов, используя цены в рублях. Пусть первый проект оказался лучше второго. Теперь перейдем на валюту юани, используя фиксированный курс пересчета. Очевидно, первый проект должен опять оказаться более выгодным, чем второй. Это очевидно из общих соображений. Однако алгоритмы расчета не обеспечивают автоматически выполнения этого очевидного условия.

В *шкале разностей* есть естественная единица измерения, но нет естественного начала отсчета. Время измеряется по шкале разностей, если год (или сутки – от полудня до полудня) принимаем естественной единицей измерения, или по шкале интервалов в общем случае. На современном уровне знаний естественного начала отсчета указать нельзя. Дату сотворения мира различные авторы рассчитывают по-разному. Только для *абсолютной шкалы* – результаты измерений это числа в обычном смысле слова. Примером является число людей в комнате. Для абсолютной шкалы допустимым является только тождественное преобразование. В процессе развития соответствующей области знания тип шкалы может меняться. Так, сначала температура измерялась по порядковой шкале (холоднее – теплее). Затем – по*интервальной* (шкалы Цельсия, Фаренгейта, Реомюра). Наконец, после открытия абсолютного нуля температуру можно считать измеренной по шкале *отношений* (шкала Кельвина). Надо отметить, что среди специалистов иногда имеются разногласия по поводу того, по каким шкалам следует считать измеренными те или иные реальные величины. Другими словами, процесс измерения включает в себя и определение типа шкалы (вместе с обоснованием выбора определенного типа шкалы).

## 4.2. Свойства и статистики основных типов шкал

В соответствии с пониманием существа измерений совокупность шкальных значений – это определенная модель реальности. Результаты любых измерений относятся, как правило, к одному из перечисленных выше типов шкал. Однако получение результатов измерений не является самоцелью – эти результаты необходимо анализировать, а для этого нередко приходится строить на их основании *производные показатели*, которые могут измеряться в других шкалах, чем исходные. Например, для оценки знаний учащихся можно применять 100-балльную шкалу, как в оценивании ЕГЭ. Но она слишком детальна, и ее можно перестроить в десятибалльную («1» – от «1» до «10»; «2» – от «10» до «20» и т. д.), или дихотомическую (например, положительная оценка – все, что выше граничного балла (по физике, например, 37), отрицательная – ниже граничного балла). Следовательно, возникает проблема, какие преобразования можно применять к данным при переходе, от какой шкалы к какой он является корректным. Эта проблема в теории измерений называется *проблемой адекватности*.

Статистики переменных, измеренных в различных шкалах, те не менее имеют одинаковый набор: средняя тенденция, вариативность, симметрия, островершинность функции распределения. Рассмотрим основные типы шкал измерений, в аспекте адекватности, и их характерные статистики.

*Шкала наименований* (*номинальная шкала*), фактически, уже не связана с понятием «величина» и используется только с

целью отличить один объект от другого: фамилии учеников, номера автомобилей и т. п. Шкала выделяет различимые классы объектов. Например, при измерении значения признака «пол»: «девочки» и «мальчики» – объекты будут различимы независимо от того, какие термины или знаки для их обозначений будут использованы. «Лица женского пола» и «лица мужского пола», или «girls» и «boys», или «А» и «Б», или «1» и «2». Следовательно, для шкалы наименований применимы взаимнооднозначные преобразования, то есть сохраняющие четкую различимость объектов. Таким образом, самая слабая шкала – шкала наименований – допускает самый широкий диапазон преобразований.

Номинальные переменные представляют минимальную информацию об изучаемом явлении, такое измерение – это простое наименование объектов в соответствии с заранее заданной схемой классификации. Как уже отмечалось, арифметические действия с числами в номинальной шкале смысла не имеют. Измерение средней тенденции происходит путем определения моды. Напомним, что если в распределении признака наблюдается одно значение моды, то распределение называется унимодальным, если несколько, – полимодальным. Когда две категории обладают равным количеством модальных (наибольших) случаев проявлений изучаемой величины (переменной), то распределение будет бимодальное.

Для того чтобы определить, насколько типична мода для нашего распределения, вычисляем дисперсию. Дисперсией номинальных переменных называется *коэффициент вариации*, вычисляемый по формулам:

$$\nu = \frac{\sum f_{немод}}{n} \text{ или } \nu = 1 - \frac{\sum f_{мод}}{n},$$

где $\sum f_{немод}$ – сумма всех случаев, не входящих в модальную категорию;

$f_{мод}$ – количество случаев, входящих в модальную категорию;

$n$ – число измерений.

Значение коэффициента вариации колеблется между нулем, когда все случаи принимают одно и то же значение, и единицей, когда каждый случай имеет свое уникальное значение. Чем меньше коэффициент вариации, тем типичнее мода. В случае полимодального распределения для расчета величины коэффициента вариации *v* выбирается одно из модальных значений, относительно которого производятся вычисления. Для классификации номинальных переменных следует основываться на взаимоисключающих и исчерпывающихся категориях. Это означает, что невозможно отнести один объект более чем к одной категории, но при этом каждый объект обязательно должен быть отнесен к какой-либо категории.

Пример. Определить моду и ее значимость для выборки по величине дисперсии переменной «Тип занятия» среди опрошенных респондентов. Объем выборки равен 100 чел. Данные опроса приведены в табл. 4.1.

Наибольшее число случаев (равное 25) в выборке соответствует значению переменной «рабочие», следовательно, данное распределение унимодальное. Мода переменной – «рабочие». Оценим типичность полученного модального значения по величине коэффициента вариации (дисперсионная статистика для номинальных измерений).

**Таблица 4.1. Номинальная переменная «Тип занятий респондентов»**

| № п/п | Значения переменной | Число случаев |
|---|---|---|
| 1 | Рабочие | 25 |
| 2 | Бюджетные работники | 23 |
| 3 | Государственные служащие | 22 |
| 4 | Сельскохозяйственные рабочие | 20 |
| 5 | Безработные | 10 |

Вычисляем коэффициент вариации

$$\nu = \frac{75}{100} = 0{,}75\,,\ \nu = 1 - \frac{25}{100} = 0{,}75\,.$$

*Вывод*. Полученное значение коэффициента вариации достаточно высокое, что свидетельствует о не типичности модального значения «рабочий» для данной выборки.

*Порядковая шкала* (*шкала рангов*) – шкала, которая только упорядочивает объекты, приписывая им те или иные ранги (результатом измерений является нестрогое упорядочение объектов). В школах некоторых стран применяется такая итоговая оценка успеваемости учащихся как порядковое место, которое данный ученик занимает в данном классе (выпуске). Частным случаем порядковой шкалы является *дихотомическая шкала*, в которой имеются всего две упорядоченные градации – «справился с заданием», «не справился с заданием». В измеренных по порядковой шкале (шкале рангов) произвольным образом изменять значения признаков нельзя – должна сохраняться упорядоченность объектов (порядок следования одних объектов за другими). Следовательно, для порядковой шкалы допустимым является любое монотонное преобразование. Например, если

претендент Иванов набрал 5 голосов, а претендент Сидоров – 10, то их упорядочение не изменится, если мы число голосов умножим на одинаковое для всех претендентов положительное число, или сложим с некоторым одинаковым для всех числом, или возведем в квадрат и т. д. Например, вместо «1», «2», «3», «4», «5» используем соответственно «3», «5», «9», «17», «102». При этом изменятся разности и отношения «голосов», но упорядочение сохранится.

При порядковых измерениях нужно присваивать каждому объекту число, которое обозначает как именно данный объект связан с другими в терминах количества того конкретного свойства, которым он характеризуется. При этом возникает возможность расположить объекты по порядку, в зависимости от количества свойства, которое их характеризует. Например, понятие «социальный класс» имеет три порядковых уровня: низший, средний, высший. В порядковых измерениях должно быть не менее трех классов, при этом расстояние между классами не устанавливается, известно только, что они образуют некую последовательность.

Средняя тенденция порядковых переменных определяется медианой. Вычисление медианы требует отсчитать с обоих концов распределения признака равное количество до тех пор, пока мы не дойдем до серединного, который и определяет значение медианы, как категорию в ранжированном ряду. Если в ряду значений признаков четное число ($2k$), медиана определяется по среднему арифметическому из двух серединных значении признака. Например, для 100 это 50 и 51. Если они принадлежат одному значению признака, то оно и будет медианой, если они принадлежат разным значениям, то медиан будет две. При не-

четном числе членов $(2k + 1)$ медианным будет значение признака $(k + 1)$ объекта.

Пример. В выборке из 10 человек респонденты ранжированы по стажу работы на данном предприятии. Определить медиану для переменной «стаж работы».

| Ранг | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Стаж | 15 | 13 | 10 | 9 | 7 | 6 | 5 | 4 | 3 | 1 |

Серединные ранги 5 и 6, им соответствует стаж 7 и 6 лет, поэтому медиана равна $(7 + 6) / 2 = 6{,}5$ лет. Таким образом, установили, что 50% рабочих имеют стаж меньше 6,5 лет, и 50% рабочих – больше 6,5 лет.

Измерение дисперсии для порядковых переменных состоит в вычислении квантильного ранга. *Квантильный ранг* – это мера положения внутри распределения, означающая, какая часть значений из всей совокупности, считая от меньшего значения вверх, находится ниже рассмотренной точки. Чем меньше степень разброса величины между этими точками, тем плотнее сгруппированы случаи вокруг медианы и тем точнее медиана представляет всю совокупность.

Среди квантилей выделяют:

*персентиль*, деление совокупности на сто равных частей так, что первый персентиль – такое значение в этой совокупности (считая от меньшего значения вверх), ниже которого находится 1% всех случаев и так далее;

*дециль* – деление совокупности на десятые доли;

*квинтиль* – деление совокупности на пятые доли;

*квартиль* – деление совокупности на четверти.

Вычисление квартилей аналогично вычислению медианы (См. раздел 2). Например, различают нижний $Q_{1/4}$ и верхний $Q_{3/4}$ квартили. В этом случае величина $Q_{1/2}$ является медианой.

$$Q_{1/4} = x_0' + \delta' \frac{\frac{1}{4}\left(\sum n_i\right) - n_H'}{n_Q'}, \; Q_{3/4} = x_0'' + \delta'' \frac{\frac{3}{4}\left(\sum n_i\right) - n_H''}{n_Q''},$$

где $x_0'$ – минимальная граница интервала, содержащего нижний (верхний) квартиль;

$n_H'$ – частота (относительная частота), накопленная до первого квартального интервала;

$n_Q'$ – частота (относительная частота) первого квартального интервала;

$\delta'$ – величина первого квартального интервала;

$n_H''$ – частота (относительная частота), накопленная до четвертого квартального интервала;

$n_Q''$ – частота (относительная частота) четвертого квартального интервала;

$\delta''$ – величина четвертого квартального интервала.

На рисунке 4.1 показано схематичное определение квартилей.

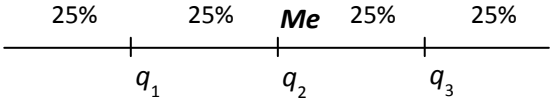



**Рис. 4.1. Схематичное определение квартилей**

Любой из указанных квантилей может быть использован для определения дисперсии порядковой величины вокруг медианы, хотя чаще используют децильные и квинтильные ранги.

Пример. Определение статистик порядковой переменной «Уровень образования». Проведены измерения в трех разных

группах респондентов, обозначенные как массивы. Ранжирование проведено по восхождению проявления измеряемого признака. Значения сведены в табл. 4.2.

**Таблица 4.2. Значения порядковой переменной «Уровень образования» для трех выделенных массивов**

| Ранг | Значение | Массив 1 | Массив 2 | Массив 3 |
|---|---|---|---|---|
| 1 | Основное общее | 25 | 25 | 10 |
| 2 | Среднее общее | 23 | 23 | 40 |
| 3 | Бакалавриат | 22 | 22 | 35 |
| 4 | Магистратура | 20 | 20 | 10 |
| 5 | Аспирантура | 9 | 10 | 5 |
| Сумма | | 99 | 100 | 100 |

Массив 1: серединный признак 50. Медиана – бакалавриат. Массив 2: серединные значения 50 и 51. Медиана – бакалавриат. Массив 3: серединные значения 50 и 51. Медианы – среднее общее и бакалавриат.

Квинтильный ранг $g$ определим по формуле: $g = g_4 - g_1$, где $g_4$ –четвертый квинтиль – значение, ниже которого находятся 4/5 или 80% всех признаков; $g1$ – первый квинтиль – значение, ниже которого находятся 1/5 или 20% всех признаков.

Рассчитаем квинтильный ранг для каждого массива нашего примера.

Массив 1: $g_4(81) = 4$, $g_1(21) = 1$, квинтильный ранг $g = 4 - 1 = 3$.

Массив 2: $g_4 = 4$, $g_1 = 1$, квинтильный ранг $g = 4 - 1 = 3$.

Массив 3: $g_4 = 3$, $g_1 = 2$, квинтильный ранг $g = 3 - 2 = 1$.

*Вывод*. Массив 3 лучше представлен своими медианами – среднее общее и бакалавриат, чем первый и второй массивы.

*Интервальные переменные* представляют наиболее полную количественную информацию об измеряемых данных. Исследователь получает возможность не только классифицировать и упорядочивать объекты, но и устанавливать, насколько большим или меньшим количеством измеряемого свойства по сравнению с другими объектами они характеризуются. Интервальные измерения основаны на представлении о существовании некоторой стандартной единицы измеряемого свойства. Для шкалы интервалов, как уже отмечалось, допустимо уже не любое монотонное преобразование, а только такое, которое сохраняет отношение разностей оценок, то есть линейное преобразование – умножение на положительное число и добавление постоянного числа.

Совокупность эмпирических отношений, отражаемых с помощью интервальной шкалы, богаче, она дает возможность отразить еще и порядок расстояний между шкалируемыми объектами. Предположим, например, что мы измерили отношение студентов к учебе и в результате получили, что четырем респондентам А, Б, В и Г оказались приписанными соответственно числа 1, 2, 3 и 8. Если мы знаем, что была использована порядковая шкала, то, интерпретируя результаты измерения, можно быть уверенными только в том, что респондент А хуже всех относится к учебе, респондент Б – получше и т. д. При использовании же интервальной шкалы мы можем получить дополнительную информацию: различие по отношению к учебе между респондентами А и Б меньше, чем различие между респондентами В и Г, а такого рода сведения весьма полезны.

Итак, если мы получаем числа, для которых «физически» осмыслены равенства типа $5 – 4 = 2 – 1$ или $8 – 3 > 3 – 2$, то счи-

таем, что они отвечают интервальной шкале. Эта шкала обычно считается «хорошей» в том смысле, что соответствующие шкальные значения в достаточной мере похожи на обычные числа. По интервальным шкалам обычно считают полученными значения таких признаков, как возраст или зарплата.

Для интервальных данных среднюю тенденцию определяет среднее арифметическое значение. Среднее арифметическое значение есть частное от деления суммы всех значений признака на их количество (См. раздел 2):

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \ldots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n},$$

где $x_1$, $x_2$, ..., $x_n$ – значения признака; $n$ – число наблюдений.

Пример. Вычислим среднее значение времени, ежедневно проведенного респондентами у телевизора, в выборке – 10 человек.

| Номер респондента | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Время на просмотр телепередач | 3 | 4 | 4 | 5 | 4 | 2 | 4 | 5 | 5 | 3 |

По формуле для *x* находим:

$$\bar{x} = \frac{39}{10} = 3{,}9 \text{ (часа)}.$$

Для сгруппированных данных формула вычисления среднего значения преобразуется в следующую:

$$\bar{x} = \frac{\sum_{i=1}^{h} x_i n_i}{n} = \frac{x_1 n_1 + x_2 n_2 + \ldots + x_h n_h}{n},$$

где $n_i$ – частота для *i*-го значения признака.

Процедуру вычисления среднего значения по сгруппированным данным удобно выполнять по схеме, приведенной в табл. 4.3.

**Таблица 4.3. Схема вычисления среднего по сгруппированным данным**

| **Интервал** | **Середина интервала** $x_i$ | **Относительная частота** $n_i$ | **Произведения** $x_i\, n_i$ |
|---|---|---|---|
| Последовательно выписываются все интервалы | $x_1$<br>$x_2$<br>…<br>$x_k$ | $n_1$<br>$n_2$<br>…<br>$n_k$ | $x_1\, n_1$<br>$x_2\, n_2$<br>…<br>$x_k\, n_k$ |
| | | $\sum_{i=1}^{k} x_i\, n_i = n$ | $\sum_{i=1}^{k} x_i\, n_i$ |

Пример. Используем данные предыдущего примера, сгруппируем их и вычислим среднее значение.

| Номер респондента *i* | 6 | 1 | 10 | 2 | 3 | 5 | 7 | 4 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|---|
| Время на просмотр телепередач *x* | 2 | 3 | 3 | 4 | 4 | 4 | 4 | 5 | 5 | 5 |
| Частота $n$ | 1 | | 2 | | | 4 | | | | 3 |

По результатам расчетов видно, что операция группировки на значение средней тенденции по данной выборке не повлияла.

Вариативность интервальных переменных определяется дисперсией и средним квадратичным отклонением. Геометрически среднее квадратичное отклонение является показателем того, насколько кривая распределения размыта относительно ее среднего арифметического. Измеряется в тех же единицах, что и изучаемый признак. При ручном счете для упрощения вычислений дисперсию рассчитывают по формуле методом отсчета от условного нуля (См. раздел 2).

Пример. Известно, что если заработная плата работников, выполняющих сходные функции в одном отделе, существенно различается, это негативно влияет на климат в коллективе и эффективность работы сотрудников. Будем считать, что условие приблизительного равенства доходов в рамках отдела соблюдается тогда, когда разброс в заработках (дисперсия) не превышает 4500 единиц. Сотрудники отдела, который состоит из 10 человек, зарабатывают соответственно 240, 256, 334, 176, 254, 219, 277, 414, 215, 366 усл. единиц.

Докажите, что условие равенства доходов не соблюдается. Оцените интервал разброса. Будет ли эта проблема решена, если перевести самого высокооплачиваемого работника в другой отдел?

*Решение*. Сначала следует вычислить среднюю тенденцию, как математическое ожидание.

$$M(X) = (240 + 256 + 334 + 176 + 254 + 219 + 277 + 414 + 215 + 366)/10 = 275{,}1.$$

Теперь рассчитаем несмещенную дисперсию заработков по формуле:

$$D(X) = \frac{\left[\sum (x_i - MX)^2\right]}{(n-1)},$$

которая равна 4965. Таким образом, условие не соблюдается (4965 > 4500). Интервал разброса определим по величине среднеквадратичного отклонения, равного 70,46, иначе, доходы сотрудников отдела преимущественно находятся в интервале от 204,64 до 282,56. Если перевести в другой отдел сотрудника с заработком 414, то изменится средняя тенденция $M(X) = 259{,}7$ и в этом случае, $D(X)$ составит 3114 и условие «комфортности» будет соблюдено.

Необходимо подчеркнуть, что проблема адекватности возникает не только при переходе от одной шкалы к другой, но и при выборе шкалы для получения первоначальных оценок – непосредственной информации об объекте. И здесь опять справедлив вывод о том, что шкала должна быть адекватна – если она слишком мощная, то возможно большое своеволие исследователя (например, при измерении качественных характеристик в шкале отношений), если слишком слабая, то происходят потери информации (например, при измерении количественных показателей в номинальной шкале). Например, нецелесообразно, с одной стороны, оценивать результаты решения одной задачи в 100-балльной шкале, а с другой стороны, результаты решения 100 задач в двухбалльной шкале. Таким образом, несмотря на то, что строго монотонное преобразование является допустимым для порядковой шкалы, соотношение между «средними тенденциями» изменилось. Это обусловлено тем, что операция вычисления среднего арифметического не является корректной в порядковой шкале. Однако, при порядковых шкалах, имеющих малое число «разрядов» – «баллов», медиана также мало информативна.

Для решения проблемы адекватности должно пользоваться свойствами взаимосвязи шкал и допустимых для них преобразований. Суммируем сказанное в таблице 4.4, которая отражает соответствие между шкалами и допустимыми преобразованиями.

**Таблица 4.4. Шкалы и допустимые преобразования**

| Название шкалы | Допустимое преобразование |
|---|---|
| Наименований | Взаимно-однозначное |
| Порядковая | Строго монотонное |
| Интервальная | Линейное |

Общий вывод таков – всегда возможен переход от более мощной шкалы к менее мощной, но не наоборот (например, на основании оценок, полученных в шкале отношений, можно строить балльные оценки в порядковой шкале, но не наоборот). Если же, например, каждому респонденту приписано число от 1 до 5 в соответствии с тем, как он ответил на вопрос типа: «Удовлетворены ли Вы своей работой?» (с вариантами ответов от «совершенно не удовлетворен» до «полностью удовлетворен», закодированными цифрами от 1 до 5 соответственно), то, кроме равенства и неравенства, можно судить и о некотором порядке между полученными числами: если одному респонденту приписано число 3, а другому – 5, то считаем, что первый меньше удовлетворен работой, чем второй. Но соотношения типа $5 – 4 = 2 – 1$ остаются бессмысленными с содержательной точки зрения.

Измерение в социологии зачастую переплетается с проблемой выбора возможных способов анализа собранных с его помощью данных. Это очень важно. Ведь измерение, в конце концов, нужно не само по себе, а именно для последующего изучения его результатов. Выбор способа анализа данных зависит от характера исходных шкал. Это обстоятельство на интуитивном уровне знакомо каждому социологу. И качество подходов к измерению должно оцениваться не в последнюю очередь с точки зрения возможности конструктивного определения того, что можно делать с этими результатами. Указанная проблема не встает для данных, полученных по шкалам низких типов, если использовать для их анализа специально предназначенные для этого методы. Но возникает вопрос другого рода. Далеко не для

всех методов, отвечающих естественной логике социолога, разработаны соответствующие математико-статистические концепции. Так, для них часто бывает совершенно неясно, каким образом переносить результаты с выборки на генеральную совокупность. Рассматриваемое положение говорит о том, что трактовка (интерпретация) данных обусловлена не только «доизмерительными» шагами (способом их физического получения), но и «послеизмерительными» представлениями о сути тех методов, которые предположительно будут использоваться для анализа результатов. Социолог, как правило, не задумывается о том, что в тех случаях, когда приписать объектам числа по интервальной шкале не удается, иногда все же бывает полезно получить хотя бы какие-нибудь соотношения для расстояний между объектами. Так, в дополнение к ранжированию телепередач по рейтингу неплохо было бы узнать, что, скажем, такие-то две передачи вызывают примерно одинаковый зрительский интерес, а вот две другие совершенно по-разному воспринимаются изучаемой аудиторией.

# Раздел 5.
# Способы графического представления выборки

## 5.1. Диаграммы

*Диаграммы и графики* позволяют наглядно и доступно отразить результаты обработки большого объема информации. В зависимости от целей использования выделяют структурные диаграммы, диаграммы сравнения и динамики; по форме графического образа – линейные диаграммы, столбиковые, радиальные, секторные, объемные и т. д.

*Линейная диаграмма* соединяет все значения одной переменной непрерывной ломаной линией и дает возможность сравнения значений разных переменных, как правило, различающихся цветом или формой изображения. При построении таких диаграмм в прямоугольной системе координат по оси абсцисс откладываются значения дат или периодов времени, по оси ординат – уровни или темпы роста. Линейные диаграммы с равномерной шкалой имеют недостаток, снижающий их познавательную ценность. Этот недостаток заключается в том, что равномерная шкала позволяет измерять и сравнивать только отраженные на диаграмме абсолютные приросты или уменьшения показателей на протяжении исследуемого периода. Однако при изучении динамики важно знать и относительные изменения исследуемых показателей по сравнению с достигнутым уровнем или темпы их изменения. Именно относительные изменения показателей в ди-

намике искажаются при изображении их на координатной диаграмме с равномерной вертикальной шкалой. Кроме того, в обычных координатах теряет всякую наглядность и даже становится невозможным изображение рядов динамики с резко изменяющимися уровнями, которые обычно имеют место в динамических рядах за длительный период времени.

На рис. 5.1 представлена линейная диаграмма, характеризующая динамику среднего балла ЕГЭ абитуриентов, зачисленных на бюджетные и платные места по техническим вузам 2011-2015 гг.

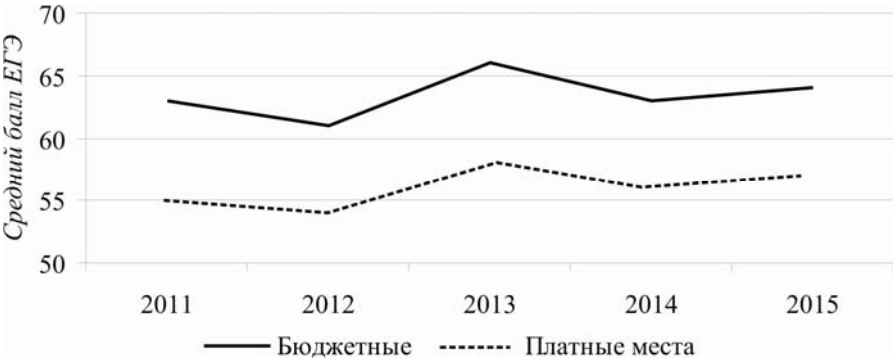



**Рис. 5.1. Линейная диаграмма сравнительной динамики среднего балла ЕГЭ абитуриентов технических вузов**

У выпускников школ, колледжей или техникумов, поступающих в вузы, на сегодняшний день большей популярностью пользуются экономические и гуманитарные направления высшего образования. Причиной падения престижности инженерных профессий послужил низкий уровень заработной платы в период кризисной ситуации в стране, сказавшейся на многих производственных предприятиях в последние годы. На сегодняшний день ситуация с производственной деятельностью по-

степенно нормализуется и, как следствие, выпускники технических специальностей и направлений становятся более востребованными. Значит, неизбежен рост числа абитуриентов, поступающих на технические направления обучения, а, следовательно, и рост баллов ЕГЭ. Другими словами, в сложившейся ситуации кадровой подготовки в стране помощью послужит только полное понимание происходящего и сознательный выбор направлений и специальностей технической направленности большинством абитуриентов, поступивших как в государственные, так и в коммерческие вузы за последние годы. Для достижения полного равновесия обучающегося контингента между гуманитарными и техническими специальностями и направлениями – необходимо время. Эти социальные тенденции и отражает линейная диаграмма.

*Столбиковые диаграммы* строятся в прямоугольной системе координат, где каждый столбик соответствует величине или уровню исследуемого статистического показателя, что позволяет их сравнивать. На горизонтали находится основание столбиков, ширина и расстояние между ними выбираются произвольно, но должны быть одинаковы. Высота меняется в зависимости от величины показателя. На одном графике возможно одновременное изображение нескольких показателей.

Согласно данным, приоритетными направлениями самоопределения выпускников сельских и городских школ большинство учащихся старших классов ориентированы на вузы. Порядка 10% выбирают техникумы, причем в сельских районах это число несколько выше. Около 6% – СПТУ и профессиональные курсы, меньшая часть молодежи планируют начать трудовую

жизнь (рис. 5.2). На диаграмме цифрами обозначены выборы выпускников (первый столбик выпускники сельских школ, второй – городских):

1 – пойдут работать;

2 – пойдут в училище или на курсы;

3 – пойдут в техникумы;

4 – пойдут в вузы;

5 – не определились.

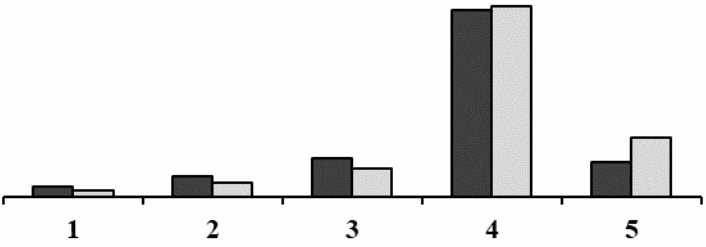



**Рис. 5.2. Виды деятельности, выбираемые выпускниками после окончания школы**

*Круговая диаграмма* строится путем деления круга на секторы, пропорционально удельному весу частей в целом. Сумма удельных весов всегда равна 100% и составляет 3600. Размер каждого сектора определяется величиной угла из расчета, что 1% соответствует 3,60. Для более четкого изображения рекомендуется использовать различные цвета или штриховку. Кроме того, есть возможность построения диаграмм с вырезанными долями, что значительно улучшает восприятие информации.

Пример круговых диаграмм показан на рис. 5.3. Динамичное развитие системы образования актуализировало проблему самоопределения личности. В целях выявления основных зако-

номерностей смыслообразующего самоопределения личности проведено исследование в одной из гимназий. В эксперименте принимали участие гимназисты 11-х классов, объем выборки составил 44 человека. Наиболее значимыми ценностями определены: духовные – 72%, затем – самостоятельность (13%), далее – безопасность (8%) и традиции (7%). Диаграмма результатов данного исследования представлена на рис. 5.3.

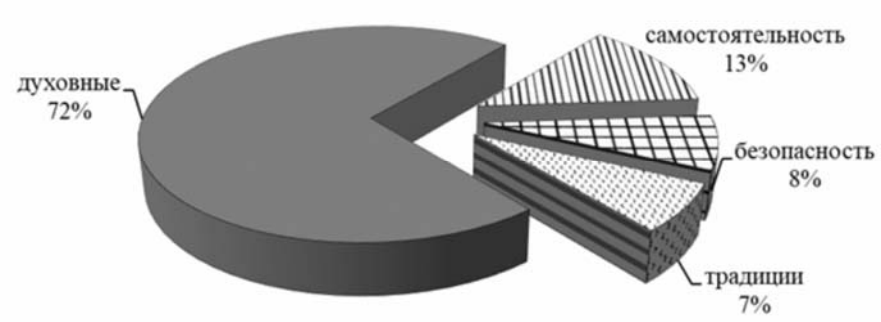



**Рис. 5.3. Ценности гимназистов в аспекте проблем самоопределения**

Результаты проведенного исследования показали, что старшеклассники не стремятся соблюдать единые ценности и правила окружающего их социума, в наименьшей степени проявляют уважения к традициям, смирение, благочестие.

Приведенные примеры визуальных представлений эмпирических данных не предусматривают учет типа измерений, то есть шкал, по которым получены данные. Диаграммы любого вида можно построить как для номинальных и порядковых переменных, так и для интервальных. Это обусловлено тем, что с их помощью есть возможность визуализировать большой объем однотипной информации, без отражения законов распределения и статистик измеренных данных. Для обеспечения нагляд-

ности в некоторых случаях приходится жертвовать точностью информации. Таким образом, выбор того или иного вида информационной модели зависит от цели, ради которой эта модель создается. Сделаем некоторые обобщения. Диаграмма – графическое изображение, дающее наглядное представление о соотношении каких-либо величин или нескольких значений одной величины, об изменении их значений. Используется множество разнообразных типов диаграмм. График – линия, дающая наглядное представление о характере зависимости какой-либо величины (например, пути) от другой (например, времени). График позволяет отслеживать динамику изменения данных. Круговая диаграмма служит для сравнения нескольких величин в одной точке. Она особенно полезна, если величины в сумме составляют нечто целое. Столбчатая диаграмма позволяет сравнивать несколько величин в нескольких точках. Диаграмма площадей позволяет одновременно проследить за изменением суммы нескольких величин в нескольких точках и при этом показать вклад каждой величины в общую сумму.

## 5.2. Прообразы законов распределения

Выборку можно представить полигоном частот распределения, кумулятивной кривой, гистограммой, которые являются прообразами законов распределения случайных величин. Гистограмма и полигон частот распределения, построенные на основе эмпирических данных выборки, позволяют выявить приближенную картину реального распределения в генеральной совокупности. При увеличении выборочной совокупности и все большем

дроблении величины интервалов эмпирическое распределение в виде гистограммы или полигона частот все более приближается к некоторой кривой, называемой *кривой распределения*.

*Полигон частот* – ломаная линия, прообраз кривой распределения. Для построения полигона частот величина признака откладывается на оси абсцисс, а частоты или относительные частоты – на оси ординат. Из точек, соответствующих значениям признака, восстанавливаются перпендикуляры, равные по высоте частотам. Вершины перпендикуляров соединяются прямыми линиями. В результате получается ломаная линия.

При построении *полигона частот распределения дискретной случайной величины* по оси абсцисс откладываются значения, которые принимает величина, а по оси ординат – частота появления этого значения в рассматриваемой совокупности. При построении *полигона частот распределения непрерывной случайной величины* или интервального ряда по оси абсцисс располагаются точки, обозначающие интервалы, в которые может попадать случайная величина (это могут быть как серединные значения интервалов, так и другие значения, в зависимости от принимаемой модели), а по оси ординат – частота появления значения признака (рис. 5.4). Площади полученных прямоугольников представляют собой выборочные оценки значений функции плотности распределения вероятностей.

Пример. Данные распределения рабочих в возрасте до 24 лет по тарифным разрядам дают возможность построить полигон частот.

| Разряд | I | II | III | IV | V | VI |
|---|---|---|---|---|---|---|
| Численность, % | 8,4 | 22,6 | 31,9 | 24,1 | 6,2 | 0,3 |

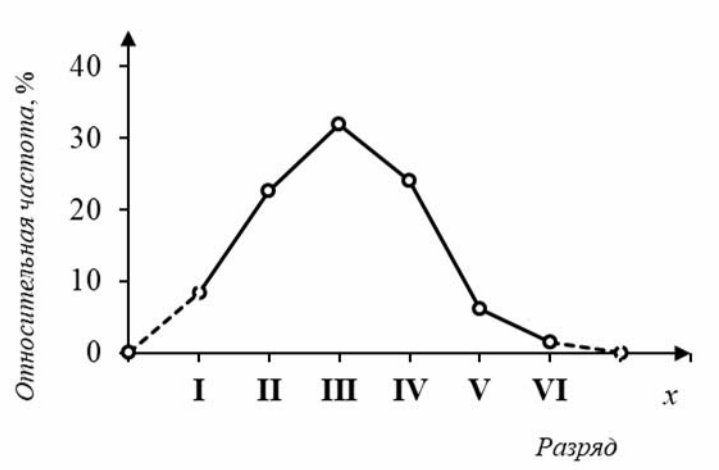



**Рис. 5.4. Полигон частот распределения рабочих по тарифным разрядам**

Условно принято крайние ординаты признака соединять с серединами примыкающих интервалов (на рисунке эти замыкающие линии нанесены пунктиром). Однако для распределения, где концентрация событий увеличивается на концах полигона, такое изображение может привести к ложным представлениям о сущности явления.

Пример. Из генеральной совокупности извлечена выборка объема $n = 50$, полигон частот которой имеет вид рис. 5.5.

Найти число вариант переменной $x_i = 4$ в данной выборке.

*Решение*. Так как объем выборки сумма всех вариант переменной, запишем ее в явном виде $n = n_1 + n_2 + n_3 + n_4 = 4 + 20 + 11 + n_4 = 50$, тогда $n_4 = 15$.

Для графического изображения вариационных рядов порядковой переменной используются также кумулятивные кривые (*кумуляты*). При построении кумуляты на оси абсцисс откладываются границы интервалов (либо значения дискрет-

ного признака), а на оси ординат – накопленные частоты (либо относительные частоты), соответствующие верхним границам интервалов. Таким образом, на графике кумуляты столбики, пропорциональные частотам, последовательно накладываются один на другой, так что высота последнего столбика является суммой всех высот, составляющей 100%. Кумулята округляет индивидуальные значения признака в пределах интервала и представляет собой возрастающую ломаную линию. Это графическое представление данных позволяет быстро определить процент случаев, находящихся ниже или выше заданной величины признака.

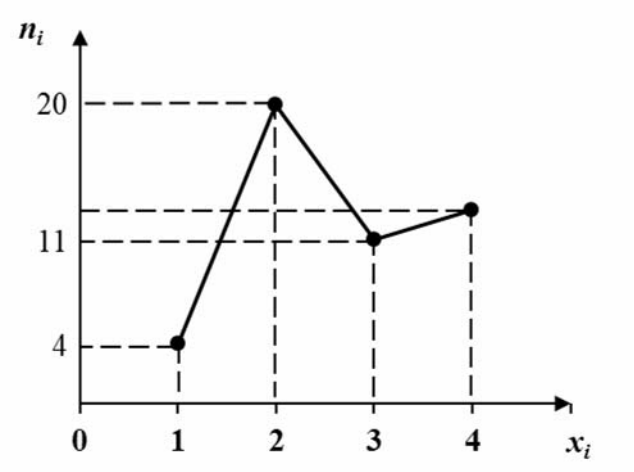



**Рис. 5.5. Полигон частот**

Составим вариационный ряд, для которого построим кумуляту при изучении результатов тестирования учащихся. Наибольшее количество баллов за тест равно 12, сложность заданий возрастает по мере увеличения порядкового номера, объем выборки – 35 учащихся.

| Количество баллов | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Частота | 1 | 1 | 2 | 3 | 4 | 4 | 6 | 5 | 3 | 3 | 2 | 1 |
| Накопленная частота $n$ | 1 | 2 | 4 | 7 | 11 | 15 | 21 | 26 | 29 | 32 | 34 | 35 |

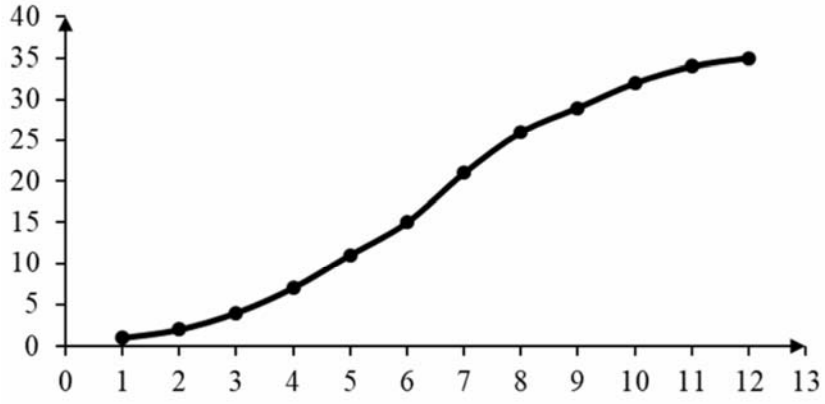



**Рис. 5.6. Кумулята распределения числа выполненных заданий в тесте**

*Гистограммой* называется графическое изображение зависимости частоты попадания элементов выборки от соответствующего интервала группировки (диапазона значений показателя). Гистограмма используется только для непрерывных признаков: предполагается, что внутри каждого интервала распределение случайной величины равномерно. По оси абсцисс откладываются крайние значения интервалов, по оси ординат – частота появления значения в рассматриваемой совокупности. Таким образом, концы интервала представляются на графике двумя точками, из которых опускаются проекции на ось абсцисс, после чего полученные четыре точки соединяются в прямоугольники. Гистограмма графическое средство для изображе-

ния ряда интервальных переменных. Рассмотрим пример построения гистограммы по данным табл. 5.1.

**Таблица 5.1. Распределение брачных возрастов разводящихся супругов**

| | **Менее года** | **1-2 года** | **2-3 года** | **3-4 года** | **4-6 года** | **6-8 лет** | **8-10 лет** | **10 и более** |
|---|---|---|---|---|---|---|---|---|
| Относительная частота, % | 7,2 | 14,5 | 13,2 | 22,9 | 16,9 | 8,4 | 1,2 | 15,7 |

На гистограмме общее число лиц в каждой категории выражается площадью соответствующего прямоугольника, а общая площадь равна численности совокупности (так как гистограмма строится по относительным частотам, то площадь равна единице или 100%). Поэтому для интервалов 4-6; 6-8; 8-10 (по табл. 5.1), длина которых в два раза больше предыдущих, нужно брать высоты прямоугольников в два раза меньше. При нанесении на график последнего открытого интервала «10 лет и более» условно будем считать верхней его границей 40 лет. Тогда ширина интервала равна 30 годам, а плотность распределения – около 0,5%.

Гистограмму следует отличать от столбчатой диаграммы – отличие состоит в том, что столбчатая диаграмма отображает частоту через высоту столбца, а гистограмма – через его площадь. Иными словами, столбцы столбчатой диаграммы всегда имеют одинаковую ширину, а столбцы гистограммы могут иметь разную ширину, в зависимости от интервалов.

Правильно построенные графические изображения эмпирических данных социологических измерений позволяют найти приближенные значения (с точностью до масштаба) средних

тенденций для измерений по различным типам шкал. Рассмотрим это на примере.

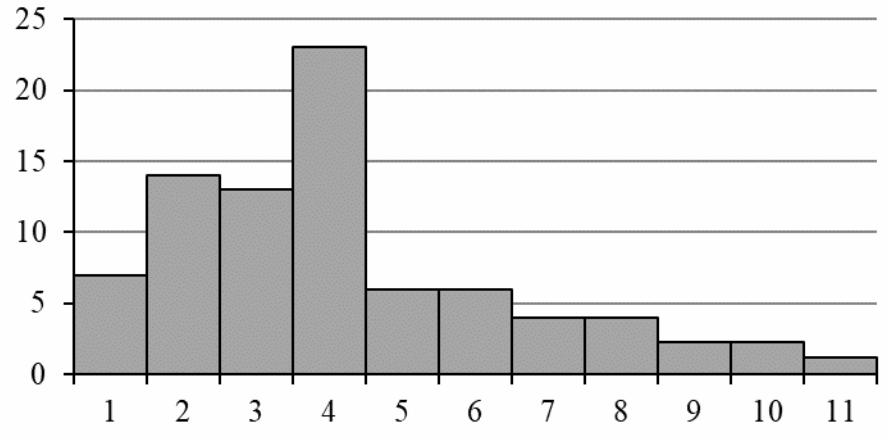



**Рис. 5.7. Гистограмма распределения соотношения брачных возрастов разводящихся супругов**

Пример. Имеются данные о возрастном составе рабочих (лет) некоторого предприятия:

18, 38, 28, 29, 26, 38, 34, 22, 28, 30, 22, 23, 35, 33, 27, 24, 30, 32, 28, 25, 29, 26, 31, 24, 29, 27, 32, 25, 29, 29.

1. Построить интервальный ряд распределения.

2. Построить графическое изображение ряда.

3. Графически определить моду и медиану.

*Решение*. Объем выборки равен 30 единиц. Определим число групп, на которые надо разделить всю совокупность по формуле Стерджесса:

$$1 + 3{,}322 \lg 30 = 6 \text{ групп (не более)}.$$

Размах выборки: максимальный возраст – 38, минимальный – 18. Найдем ширину интервала

$$\frac{38-18}{6}=\frac{20}{6}=3+1/3\,.$$

Так как концы интервалов должны быть целыми числами, разделим совокупность на 5 групп, при этом ширина интервала будет равна 4. Для облегчения подсчетов расположим данные в порядке возрастания

18, 22, 22, 23, 24, 24, 25, 25, 26, 26, 27, 27, 28, 28, 28,

29, 29, 29, 29, 29, 30, 30, 31, 32, 32, 33, 34, 35, 38, 38.

Составим таблицу распределение возрастного состава рабочих:

| **№ п/п** | **Возраст (границы групп), *x*** | **Частота, *f*** | **Накопленная частота, *S*** |
|---|---|---|---|
| 1 | 18-22 | 3 | 3 |
| 2 | 23-26 | 7 | 10 |
| 3 | 27-30 | 12 | 22 |
| 4 | 31-34 | 5 | 27 |
| 5 | 35-38 | 3 | 30 |
| | Всего | 30 | |

Изобразим ряд в виде гистограммы и полигона частот (рис. 5.8). Основание столбика гистограммы – ширина интервала. Высота столбика равна частоте.

Полигон (или многоугольник распределения) – график частот. Чтобы его построить по гистограмме, соединяем середины верхних сторон прямоугольников. Многоугольник замыкаем на оси OX на расстояниях, равных половине интервала от крайних значений *x*.

Для определения среднего значения по гистограмме, надо выбрать самый высокий прямоугольник, провести линию от

правой вершины этого прямоугольника к правому верхнему углу предыдущего прямоугольника, и от левой вершины модального прямоугольника провести линию к левой вершине последующего прямоугольника. От точки пересечения этих линий провести перпендикуляр к оси OX. Абсцисса и будет средним значением ≈ 27,5 лет – это означает средний возраст рабочих в выборке. Значение моды (Mo) определяем по полигону частот. С этой целью находим абсциссу вершины графика полигона с точностью до масштаба, она равна 29 годам, то есть самым часто встречающимся возрастом в выборке.

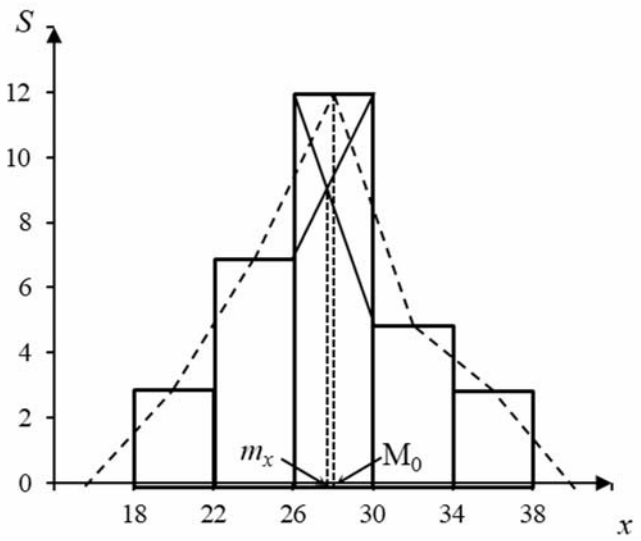



**Рис. 5.8. Гистограмма и полигон частот с указанием средних тенденций**

Медиану находим по кумуляте, графику накопленных частот. Абсциссы этого графика – варианты вариационного ряда. Ординаты – накопленные частоты.

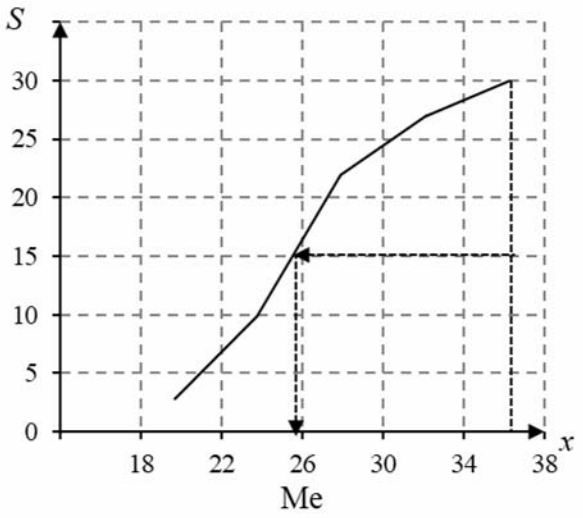



**Рис. 5.9. Кумулята с найденным значением медианы**

Для определения медианы по кумуляте находим по оси ординат точку, соответствующую 50% накопленных частот (в нашем случае 15), проводим через нее прямую, параллельно оси OX, и от точки ее пересечения с кумулятой проводим перпендикуляр к оси OX. Абсцисса этой точки является медианой: Me ≈ 25,9. Это означает, что половина рабочих в данной совокупности имеет возраст менее 26 лет.

В рассмотренном примере хорошо видно, что владение графическими методами представления эмпирических данных дает возможность провести предварительные оценки средних тенденций без трудоемких расчетов. Получившие широкое распространение в современных условиях пакеты прикладных программ компьютерной графики значительно облегчают задачи исследователя-социолога при практическом применении графического представления данных.

# Раздел 6.
# Регрессионный анализ данных

## 6.1. Регрессионные модели

*Регрессионный анализ* – статистический метод исследования влияния одной или нескольких независимых переменных на зависимую переменную. Независимые переменные иначе называют регрессорами или предикторами, а зависимые переменные критериальными. Терминология *зависимых* и *независимых* переменных отражает лишь математическую зависимость переменных, а не причинно-следственные отношения. Цели регрессионного анализа:

определение степени вариации зависимой переменной независимыми переменными;

предсказание значения зависимой переменной с помощью независимых;

определение вклада отдельных независимых переменных в вариацию зависимой.

Регрессионный анализ нельзя использовать для определения наличия связи между переменными, поскольку наличие такой связи и есть предпосылка для применения анализа. Проблема регрессии характерна тем, что о распределениях изучаемых величин нет достаточной информации. Пусть имеются основания предполагать, что случайная величина $Y$ имеет некоторое распределение вероятностей и нужно по результатам наблюдений определить значения параметров этого распределения.

В зависимости от природы задачи и целей анализа результаты эксперимента по-разному интерпретируются в отношении переменной x. Довольно часто в практике исследований имеет место ситуация, когда важнейшие переменные, описывающие некоторый процесс, известны заранее, но модель самого процесса неизвестна. В таких случаях возможны разные подходы. Одним из них является построение эмпирических моделей.

Для установления связи между величинами в эксперименте используется модель, основанная на допущениях:

величина $x$ является контролируемой величиной, значения которой заранее задаются при планировании эксперимента;

независимые переменные при различных измерениях одинаково распределены с нулевым средним и постоянной дисперсией.

В случае неконтролируемой переменной результаты наблюдений $(x_1, y_1,)$, ..., $(x_n, y_n)$ представляют собой выборку из некоторой двумерной совокупности. Методы регрессионного анализа одинаковы и в том, и в другом случае, однако интерпретация результатов различается (в последнем случае анализ существенно дополняется методами теории корреляции). Выбор вида функций иногда определяется по расположению экспериментальных значений $(x, y)$ на диаграмме рассеяния, чаще из теоретических соображений. Функциональная зависимость между переменными $x$ и $y$ это правило, которое каждому элементу $x$ из $\{X\}$ ставит в соответствие элемент $y$ из $\{Y\}$.

Если график функции регрессии $y_x = f(x)$ или $x_y = \varphi(y)$ изображается прямой линией, то регрессию называют *линейной*, а если кривой линией – *нелинейной*.

Например, функции регрессии, используемые при количественной оценке связей между переменными, могут иметь вид:

$y_x = ax + b$ – линейная;

$y_x = ax^2 + bx + c$ – квадратичная;

$y_x = bx^a$ – степенная;

$y_x = be^{ax}$ – экспоненциальная;

$y_x = ba^x$ – показательная;

$y_x = a \ln x + b$ – полулогарифмическая;

$y_x = b + \frac{a}{x}$; $y_x = \frac{1}{ax + b}$ – гиперболические.

Для определения вида функции регрессии строят точки $(x; y_x)$ – *диаграмму рассеяния* (рис. 6.1) и по их расположению делают заключение о примерном виде функции регрессии.

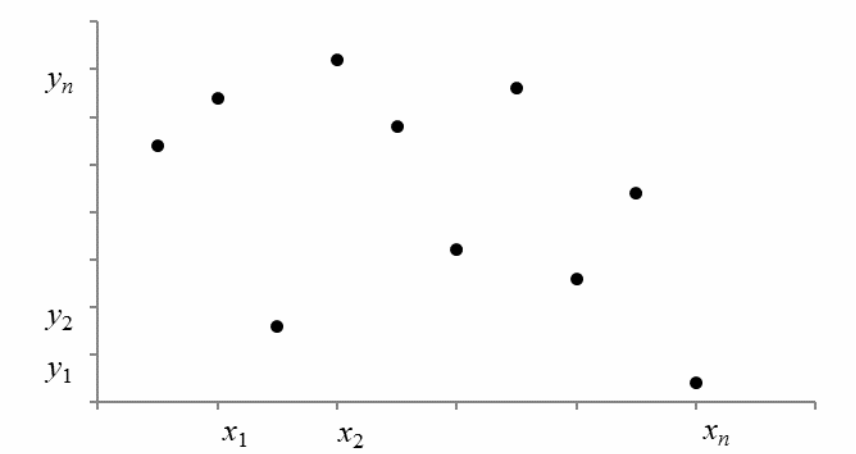



**Рис. 6.1. Диаграмма рассеяния**

Не существует общего правила для выбора подходящего вида функции. Можно лишь догадываться о виде уравнения регрессии. Однако существуют способы, с помощью которых можно проверить, является ли догадка удачной.

Процесс нахождения теоретической линии регрессии заключается в выборе и обосновании типа кривой и расчете параметров ее уравнения, в результате появляется возможность составить уравнение регрессии и получить количественную оценку влияния факторных признаков на результативный признак.

Регрессионный анализ естественным образом обобщается, когда зависимая переменная зависит не от одной, а от нескольких независимых переменных. Очевидно, что одновременный учет нескольких факторов, связанных с интересующей нас величиной, позволяет построить модель, точнее описывающую имеющиеся данные и лучше прогнозирующую зависимую переменную. Безусловно, процедура оценки наиболее эффективна при правильно спланированном эксперименте, но требуют рассмотрения и те случаи, когда в распоряжении имеются данные, полученные в заранее не спланированном эксперименте. В таком случае будем следовать некоторым правилам, позволяющим выбрать наиболее подходящую модель с наименьшими возможными затратами:

регрессионное уравнение должно содержать минимальное число коэффициентов, следовательно, и переменных;

желательно, чтобы уравнение имело под собой содержательное обоснование. Например, демографические изменения, если нет ограничений на пищевые и другие ресурсы, осуществляются по экспоненте, поэтому модель этого процесса должна иметь соответствующую функциональную зависимость.

Лучшая процедура отбора наиболее подходящих моделей – пошаговый регрессионный анализ. Суть его в том, что отдельные переменные последовательно включаются в первоначальную мо-

дель и на каждом этапе анализируются, приводит ли добавление переменной к существенному или статистически значимому приближению предсказанных значений к эмпирическим данным.

В значительной мере достоверность полученных оценок зависит от некоторых предположений относительно поведения случайной ошибки:

случайный характер – отдельные ошибки представляют собой случайные величины;

нулевое среднее – каждое отклонение ошибки характеризуется нулевым математическим ожиданием и не зависит от значений $x_i$;

дисперсия каждого отклонения равна для всех точек и независима от $x_i$ ;

отсутствие взаимосвязи (автокорреляции) ошибок;

ошибка должна иметь нормальное распределение.

При выполнении этих условий можно приближенно оценить точность предсказания влияния $x$ на $y$, так как именно предсказание является одной из главных целей регрессионного анализа.

Классический подход к оцениванию параметров линейной регрессии основан на методе наименьших квадратов (МНК), разработанный К. Гауссом и А. Лежандром. В его основу положена теория исследования на экстремум функции нескольких переменных.

## 6.2. Метод наименьших квадратов

Метод наименьших квадратов (МНК) позволяет получить такие оценки параметров $a$ и $b$, при которых сумма квадратов

отклонений $y$ фактических значений результативного признака от расчетных (теоретических) $\overline{y}_x$ минимальна:

$$S = \sum_i \left( y_i - \overline{y}_{x_i} \right)^2 \to \min .$$

Иными словами, из всего множества линий линия регрессии на графике выбирается так, чтобы сумма квадратов расстояний по вертикали между точками и этой линией была минимальной:

$$S = \sum_i \left( y_i - \overline{y}_{x_i} \right)^2 = \sum_i \left( y_i - ax_i - b \right)^2 = \sum_i \varepsilon_i^2 \to \min .$$

Процесс выражения опытных данных функциональной зависимостью с помощью МНК состоит из двух этапов: сначала выбирают вид искомой формулы, а затем для данной формулы подбирают параметры.

Рассмотрим в качестве эмпирической формулы линейную зависимость. Если выборочное уравнение регрессии имеет линейную зависимость $y_x = ax + b$, то параметры $a$ и $b$ находят из системы линейных уравнений, применяя метод Крамера:

$$\begin{cases} \left( \sum_{i=1}^{n} x_i^2 \right) \cdot a + \left( \sum_{i=1}^{n} x_i \right) \cdot b = \left( \sum_{i=1}^{n} x_i \, y_i \right), \\ \left( \sum_{i=1}^{n} x_i \right) \cdot a + n \cdot b = \left( \sum_{i=1}^{n} y_i \right), \end{cases}$$

где $n$ – количество точек измерений.

$$\Delta = \begin{vmatrix} \sum x_i^2 & \sum x_i \\ \sum x_i & n \end{vmatrix} = n \cdot \sum x_i^2 - (\sum x_i)^2$$

$$\Delta \, \text{a} = \begin{vmatrix} \sum y_i x_i & \sum x_i \\ \sum y_i & n \end{vmatrix} = n \cdot \sum y_i x_i - \sum x_i \cdot \sum y_i$$

$$\Delta \mathrm{b} = \begin{vmatrix} \sum x_i^2 & \sum y_i x_i \\ \sum x_i & \sum y_i \end{vmatrix} = \sum x_i^2 \sum y_i - \sum x_i y_i \sum x_i$$

Вычислим коэффициенты регрессии: $a = \frac{\Delta a}{\Delta}$; $b = \frac{\Delta b}{\Delta}$.

или $a = \frac{n\sum y_i x_i - \sum y_i \sum x_i}{n\sum x_i^2 - (\sum x_i^2)}$, $b = \frac{\sum x_i^2 \sum y_i - \sum x_i \sum y_i x_i}{n\sum x_i^2 - (\sum x_i)^2}$.

Критерием точности полученных коэффициентов уравнения регрессии является сумма квадратов отклонений значений расчетной и экспериментальной функций. В результате получим значение критерия выраженного одним числом

$$\sum \Delta_i^2 = (y_{p_i} - y_{э_i})^2.$$

Параметр $a$ называется коэффициентом регрессии. Его величина показывает среднее изменение признака *y* на *a* единиц при изменении фактора *x* на одну единицу. Коэффициент *b* равен значению *y* при $x = 0$, т. е. характеризует начальное состояние рассматриваемой зависимости – уровень отсчета. Например, если уравнение регрессии имеет вид $y_x = 2x + 5$, где*x* – уровень заработной платы персонала дошкольных учреждений, а $y_x$ – полученная средняя покупательная способность работников образовательных организаций, то смысл коэффициентов $a = 2$ и $b = 5$ таков: при увеличении заработной платы в дошкольных учреждениях на 1 условную единицу, покупательная способность (теоретически, в среднем) повышается на 2 две условных единицы. Если заработную плату «заморозить», покупательная способность составит 5 условных единиц. Уравнение регрессии может быть найдено также в виде $x_y = cy + d$.

Знак при коэффициенте $a$ показывает направление связи: при $a > 0$ связь прямая, при $a < 0$ – обратная. Процесс нахождения значений коэффициентов уравнения и их суть при этом сохраняются. Возможность интерпретации коэффициента регрессии сделала этот вариант регрессионной модели достаточно распространенным в прикладных исследованиях. Параметр $b$ может не иметь прикладного содержания. Интерпретировать можно лишь знак: если $b > 0$, то относительное изменение результата $y$ происходит медленнее, чем изменение фактора $x$. Иными словами, вариация результата $y$ меньше вариации фактора – коэффициент вариации по фактору $x$ выше коэффициента вариации для результата $y$: $V_x > V_y$.

Если уравнение регрессии имеет вид $y_x = ax^2 + bx + c$, то параметры $a$, $b$ и $c$ находят из системы линейных уравнений, тем же методом Крамера:

$$\begin{cases} \left(\sum_{i=1}^{n} x_i^4\right) \cdot a + \left(\sum_{i=1}^{n} x_i^3\right) \cdot b + \left(\sum_{i=1}^{n} x_i^2\right) \cdot c = \left(\sum_{i=1}^{n} x_i^2 \, y_i\right) \\ \left(\sum_{i=1}^{n} x_i^3\right) \cdot a + \left(\sum_{i=1}^{n} x_i^2\right) \cdot b + \left(\sum_{i=1}^{n} x_i\right) \cdot c = \left(\sum_{i=1}^{n} x_i \, y_i\right) \cdot \\ \left(\sum_{i=1}^{n} x_i^2\right) \cdot a + \left(\sum_{i=1}^{n} x_i\right) \cdot b + n \cdot c = \left(\sum_{i=1}^{n} y_i\right) \end{cases}$$

Зная уравнение регрессии, по точкам легко построить его наглядное представление – линию регрессии. С помощью полученного уравнения регрессии можно, подставляя задаваемые значения фактора $X$, в прогнозируемом периоде получить планируемую величину показателя.

Пример. Найти уравнение линейной регрессии по экспериментальным данным:

| Аргумент $x$ | $x_1 = 1$ | $x_2 = 2$ | $x_3 = 3$ | $x_4 = 4$ | $x_5 = 5$ |
|---|---|---|---|---|---|
| Значение $y_э$ | $y_1 = 2,3$ | $y_2=1,8$ | $y_3=3,8$ | $y_4=5,3$ | $y_5=4,3$ |

$$n = 5$$

$$\sum x_i = 15$$

$$\sum y_i = 17,5$$

$$\sum y_i x_i = 2,3+3,6+11,4+21,2+21,5 = 60$$

$$\sum x_i^2 = 55$$

$$\left(\sum x_i\right)^2 = 225$$

$$a = \frac{5 \cdot 60 - 15 \cdot 17,5}{5 \cdot 55 - 225} = \frac{300 - 262,5}{275 - 225} = \frac{37,5}{50} = 0,75;$$

$$b = \frac{55 \cdot 17,5 - 15 \cdot 60}{5 \cdot 55 - 225} = \frac{962,5 - 900}{275 - 225} = \frac{62,5}{50} = 1,25 \cdot$$

$y = 0,75x + 1,25$ – искомое уравнение линейной регрессии.

Вычислим «модельные» значения зависимой переменной построим график (рис. 6.2).

| $x$ | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| $y$ | 2 | 2,75 | 3,5 | 4,25 | 5 |

Рассчитаем критерий оценки построенного уравнения регрессии как сумму квадратов отклонений экспериментальных и расчетных значений $(y_э - y_р)^2$

$$\sum \Delta_i^2 = 0,09 + 0,9025 + 0,09 + 1,1025 + 1,69 = 3,875$$

Как отмечалось, линейная регрессия не является единственным вариантом описания экспериментальных данных. Мно-

го реальные процессы и явления предполагают нелинейный характер зависимости между переменными. В этих случаях применяется аналогичная методика расчета коэффициентов уравнения, только сложные случаи нелинейности следует привести к линейному виду, применив метод выравнивания исходных данных.

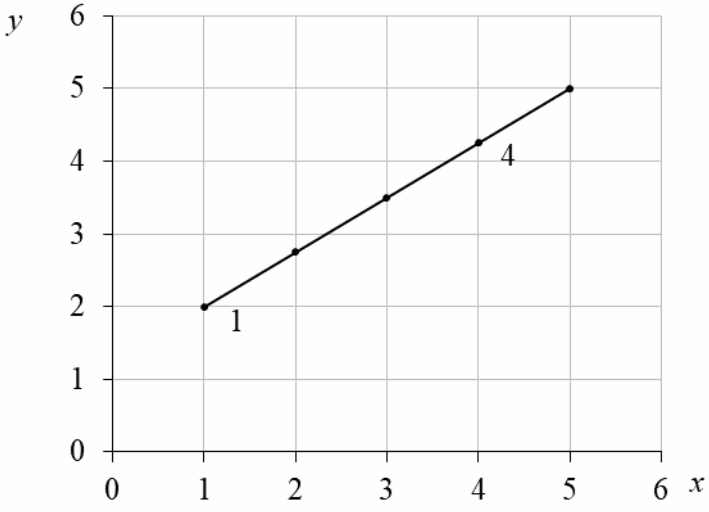



**Рис. 6.2. График линейной регрессии**

Приведем пример линеаризации нелинейной зависимости вида $y = ab^x$.

Прологарифмируем обе части уравнения $\lg y = \lg a + x\lg b$ и введем новые переменные $Y = \lg y$, $A = \lg a$, $B = \lg b$. В итоге получим уравнение линейного вида $Y = A + Bx$, коэффициенты которого находим стандартным методом наименьших квадратов. После чего, выполнив обратную замену, получим искомую нелинейную модель.

Другие, вышеприведенные нелинейные функции также легко приводятся к линейному виду путем введения новых переменных.

Пример. По исходным данным предыдущего примера построим уравнение нелинейной регрессии заданного вида, применив метод выравнивания к закону гиперболического типа

$$y = \frac{1}{ax + b}.$$

Введем новую переменную $Y = \frac{1}{y}$.

После введения новой переменной получим таблицу для расчета коэффициентов регрессионного уравнения по МНК:

| $x$ | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| $y$ | 2,3 | 1,8 | 3,8 | 5,3 | 4,3 |
| $Y=1/y$ | 0,4 | 0,5 | 0,3 | 0,2 | 0,2 |

$$n = 5$$

$$\sum x_i = 15$$

$$\sum \mathrm{Y}_i = 1{,}6$$

$$\sum \mathrm{Y}_i x_i = 0{,}4 + 1 + 0{,}9 + 0{,}8 + 1 = 4{,}1$$

$$\sum x_i^2 = 55$$

$$\left(\sum x_i\right)^2 = 225$$

Найдем коэффициент регрессии для выравненного закона $Y = ax + b$.

$$a = \frac{5 \cdot 4{,}1 - 15 \cdot 1{,}6}{5 \cdot 55 - 225} = \frac{20{,}5 - 24}{50} = \frac{-3{,}5}{50} = -0{,}07$$

$$b = \frac{55 \cdot 1{,}6 - 15 \cdot 4{,}1}{5 \cdot 55 - 225} = \frac{88 - 61{,}5}{50} = \frac{26{,}5}{50} = 0{,}53$$

Тогда $Y = -0{,}07x + 0{,}53$.

Вернемся к исходной переменной $\frac{1}{y} = 0{,}07x + 0{,}53$.

Окончательно уравнение нелинейной регрессии примет вид:

$$Y_p = \frac{1}{-0{,}07x + 0{,}53}.$$

Рассчитаем «модельные» значения, сумму квадратов отклонений и построим график нелинейной регрессии (рис. 6.3.):

| $x$ | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| $Y_p$ | 2,2 | 2,6 | 3,1 | 4 | 5,5 |

$$\sum \Delta_i^2 = 0{,}01 + 0{,}64 + 0{,}49 + 1{,}69 + 1{,}44 = 4{,}27$$

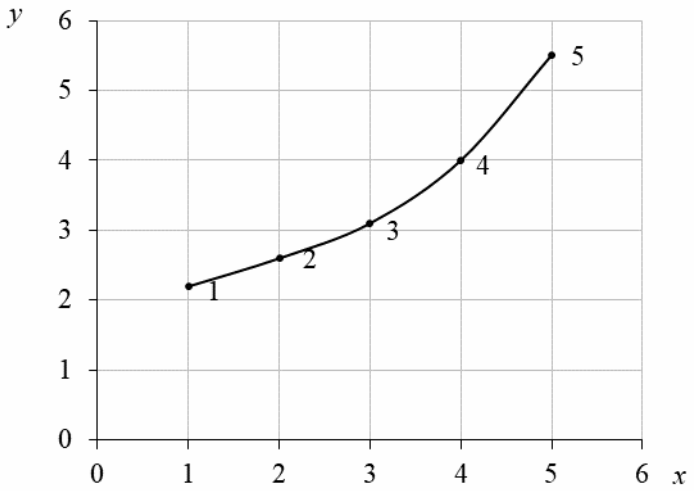



**Рис. 6.3. График нелинейной регрессии**

*Вывод*: в линейном законе сумма квадратов отклонений расчетных и экспериментальных точек меньше, чем в нелинейном законе (3,875 < 4,27), значит, линейный закон более адекватен исходным данным.

В практических исследованиях всегда имеет место некоторое рассеяние точек относительно линии регрессии. Оно обусловлено влиянием прочих, не учитываемых в уравнении регрессии факторов. Имеют место отклонения фактических данных от теоретических $y - \overline{y}_x$. Величина этих отклонений и лежит в основе расчета остаточной дисперсии (суммы квадратов отклонений):

$$D_{ост} = \frac{1}{n}\sum (y - \overline{y}_x)^2 .$$

Чем меньше величина остаточной дисперсии, тем лучше уравнение регрессии подходит к исходным данным. Если остаточная дисперсия оказывается примерно одинаковой для нескольких функций, то на практике предпочтение отдается более простым видам функций, так как они в большей степени поддаются интерпретации и требуют меньшего объема наблюдений. Результаты многих исследований подтверждают, что число наблюдений должно в 6-7 раз превышать число рассчитываемых параметров при переменной *x*. Это означает, что искать линейную регрессию, имея менее 6 наблюдений, бессмысленно. Для других видов функциональных зависимостей каждый параметр при *x* должен рассчитываться хотя бы по 6 наблюдениям. Следовательно, для параболы второй степени $y = ax^2 + bx + c$ необходимы не менее 12 данных, а для параболы третьей степени $y = ax^3 + bx^2 + cx + d$ – не менее 18 наблюдений.

## 6.3. Элементы теории корреляции

Интуитивно ясно, что о взаимозависимости между парой переменных можно говорить в тех случаях, когда уменьшению (увеличению) одной из них будет соответствовать уменьшение (увеличение) другой либо уменьшению (увеличению) первой будет соответствовать увеличение (уменьшение) второй переменной. В первом случае можно говорить о положительной корреляции между переменными (прямая зависимость), во втором – об отрицательной корреляции (обратная зависимость). Если рассмотреть разброс значений переменных относительно их средних, получим положительные и отрицательные разности, и знак их будет также различен. При однонаправленных изменениях обеих переменных произведение их отклонений положительно, если же изменения переменных разнонаправлены, то произведение отрицательно. Величина, полученная как отношение суммы произведений отклонений длине выборке без единицы, называется *ковариацией*. Ковариация вычисляется по формуле:

$$S_{xy} = \frac{\sum_{i=1}^{n} \left(x_i - \overline{x}\right)\left(y_i - \overline{y}\right)}{n-1}$$

Признаки $x$ и $y$, по которым рассчитывается взаимосвязь, могут измеряться в разных единицах, иметь произвольные средние и дисперсии. Поэтому, вычитание соответствующих средних по каждой переменной делает ковариацию независимой от средних. Если разделить ковариацию на произведение стандартных отклонений, получим безразмерный коэффициент связи, который называется коэффициентом корреляции. *Коэффициент*

*корреляции* (*Пирсона*) представляет собой численную меру степени взаимосвязи двух переменных, введенную в статистическую практику К. Пирсоном.

Коэффициент корреляции вычисляется по формуле:

$$r_{xy} = \frac{S_{xy}}{S_x S_y}$$

Очевидно, что $S_{xy} = S_{yx}$, *следовательно,* $r_{xy} = r_{yx}$, поэтому с помощью коэффициента корреляции можно численно оценить величину и направленность взаимосвязи. Имеются разные модификации формулы линейного коэффициента корреляции, например:

$$r_{xy} = a\frac{\sigma_x}{\sigma_y} = \frac{\mathrm{cov}(x,y)}{\sigma_x \sigma_y} = \frac{\overline{yx} - \bar{y}\cdot\bar{x}}{\sigma_x \sigma_y},$$

где $\overline{xy} = \dfrac{\sum_{i=1}^{n} x_i\ y_i}{n}$; $\sigma_x = \sqrt{\dfrac{\sum_{i=1}^{n} x_i^2}{n} - (\bar{x})^2}$;

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

По поводу интерпретации коэффициента корреляции сделаем два существенных замечания. При выводе расчетной формулы $r_{xy}$ не делалось предположений о характере совместного двумерного распределения величин $x$ и $y$. Однако, очевиден вывод о пределах его изменения в диапазоне от -1до +1. Выражения «сильная связь», «слабая связь», «умеренная связь» и т. д. справедливы только в рамках определенной статистической модели. Так, если частотные распределения величин $x$ и $y$ имеют разные значения асимметрии, то есть существенно скошены в

разных направлениях, то даже при максимально возможной линейной связи между $x$ и $y$ величина коэффициента корреляции не будет по абсолютной величине превышать значения 0,6-0,7. Эта зависимость максимальной величины коэффициента корреляции от характера распределения $x$ и $y$ приводит к трудностям интерпретации получаемых его конкретных значений. Что означает $r_{xy} = 0{,}6$? Максимально возможную линейную связь при положительной и отрицательной асимметрии распределений $x$ и $y$ или умеренную связь этих переменных при совместном распределении, подобном двумерному нормальному распределению? Ответы на эти вопросы можно получить из качественного анализа диаграмм рассеяния и гистограмм распределения.

Второе замечание связано со значением коэффициентов, близких к нулю. Равенство нулю коэффициентов корреляции между переменными не всегда свидетельствует об отсутствии статистической связи между $x$ и $y$. Так может проявляться, например, их нелинейная связь. Возможные варианты проявлений ложной корреляции могут быть еще связаны с появлением в совокупности исходных данных аномальных значений, или за счет неоднородности анализируемого материала, или за счет ошибок при регистрации данных.

Обобщим сказанное. Коэффициент корреляции находится в границах $-1 \le r_{xy} \le 1$. Если величина $r_{xy} > 0$, то зависимость между $X$ и $Y$ такова, что возрастание значений одной из переменных приводит к увеличению значений другой переменной. При значениях $r_{xy} < 0$ увеличение одной переменной приводит к уменьшению другой. Чем ближе значение $r_{xy}$ к $(\pm 1)$, тем сильнее (теснее) переменные $X$ и $Y$ связаны линейной

функциональной зависимостью. Если $r_{xy} = \pm 1$, то линии регрессии $\overline{y}_x$ и $\overline{x}_y$ сливаются в одну, что соответствует строгой линейной зависимости между $X$ и $Y$, в этом случае все наблюдаемые значения располагаются на общей прямой. При $r_{xy} = -1$ имеет место отрицательная линейная зависимость, при $r_{xy} = 1$ – положительная. Если $r_{xy} = 0$, то признаки $X$ и $Y$ некоррелированы. В этом случае линии регрессии $\overline{y}_x$ и $\overline{x}_y$ параллельны координатным осям.

Важно заметить, что величина линейного коэффициента корреляции оценивает тесноту связи рассматриваемых признаков в ее линейной форме. Поэтому близость абсолютной величины линейного коэффициента корреляции к нулю еще не означает, что $X$ и $Y$ независимы, если допускается отклонение этой зависимости от линейной. Следовательно, прямое утверждение $r_{xy} = 0$ не означает независимости исследуемой пары признаков. В то же время, если $X$ и $Y$ независимы, то верно утверждение $r_{xy} = 0$. Таким образом, при отклонении парной статистической зависимости от линейной коэффициент корреляции теряет свой смысл как характеристика степени тесноты связи. В этом случае следует воспользоваться другим измерителем связи – корреляционным отношением.

Допустим, что выборочный коэффициент корреляции (найденный по выборке) оказался отличным от нуля. Так как выборка случайна, то еще нельзя заключить, что коэффициент корреляции генеральной совокупности также отличен от нуля. В этом случае проверяют гипотезу о значимости (существенности) выборочного коэффициента корреляции. Другая формулировка нулевой гипотезы: коэффициент корреляции генеральной

совокупности $r_{xy}$ равен нулю. Для оценки коэффициента корреляции $r_{xy}$ нормально распределенной генеральной совокупности можно воспользоваться формулой:

$$r_{xy} - 3 \cdot \frac{1 - r_{xy}^2}{\sqrt{n}} \le r_{г.с.} \le r_{xy} + 3 \cdot \frac{1 + r_{xy}^2}{\sqrt{n}} \text{ (при } n \ge 50 \text{)}.$$

Для оценки качества подбора уравнения регрессии рассматривается квадрат линейного коэффициента корреляции $r_{xy}^2$, называемый *коэффициентом детерминации*:

$$r_{xy}^2 = \frac{\sigma_{y_{объясн.}}^2}{\sigma_{y_{общ.}}^2} \cdot$$

Коэффициент детерминации характеризует долю дисперсии результативного признака $Y$, объясняемую регрессией, в общей дисперсии результативного признака. Соответственно величина $1 - r^2$ характеризует долю дисперсии $Y$, вызванную влиянием остальных не учтенных в модели факторов.

Величина коэффициента детерминации является одним из критериев оценки качества линейной модели. Чем больше доля объясненной вариации, тем меньше роль прочих факторов, т. е. линейная модель хорошо аппроксимирует исходные данные, и ею можно воспользоваться для прогноза значений результативного признака. Например, если коэффициент корреляции $r_{xy} = 0{,}76$, то коэффициент детерминации $r_{xy}^2 = 0{,}76^2 \approx 0{,}58$. Как следует из полученного результата – 58% рассеяния зависимой переменной $Y$ объясняются линейной регрессией $y$ на $x$, а необъясненные 42% рассеяния могут быть вызваны либо случайными ошибками, либо тем, что линейная регрессионная модель плохо согласуется с опытными данными.

На практике при заданном уровне значимости $\alpha$ обычно проверяется гипотеза о том, что линейная регрессионная модель *не* согласуется с опытными данными, т. е. гипотеза, отвергающая наличие линейной связи между переменными $X$ и $Y$.

Коэффициент корреляции Пирсона адекватен только для интервальных шкал (он включает среднее арифметическое). Для оценки связи между признаками, измеренными на более низком уровне (номинальном и порядковом) существуют другие меры. Так, для оценки связи между порядковыми переменными существуют ранговые коэффициенты корреляции, которые рассчитываются с помощью ранжирования всех значений обоих признаков (ранг «1» приписывается самому большому значению признака, ранг «2» – второму по величине, ранг $n$ – наименьшему значению). Самые распространенные ранговые коэффициенты – это коэффициент Кендэлла и коэффициент Спирмена (см. Раздел 9).

Коэффициент Кендэлла определяется по формуле:

$$\tau = \frac{4K}{n\,(n-1)} - 1\,,$$

где $K$ – число перестановок, необходимое для того чтобы привести ранжировки обоих признаков к одному виду. Величину $K$ удобно рассчитывать по алгоритму: все наблюдения упорядочиваются таким образом, чтобы их ранги по первому признаку возрастали от ранга «1» до ранга «$n$». После этого для каждого наблюдения определяется число рангов второго признака, которые превосходят ранг этого наблюдения и следуют за ним. Например, пусть имеются ранжирование респондентов по доходу и возрасту:

| Признак | Ранги наблюдений | | | | |
|---|---|---|---|---|---|
| **Наблюдение** | **1** | **2** | **3** | **4** | **5** |
| 1. Возраст | 1 | 2 | 3 | 4 | 5 |
| 2. Доход | 3 | 5 | 4 | 1 | 2 |

Тогда для первого наблюдения (первый столбик по признаку «возраст») величина $K_1 = 2$, так как ранг этого наблюдения по второму признаку 3, а за ним следуют два наблюдения, которые превосходят его по рангу. Для второго наблюдения $K_2 = 0$, так как справа от него нет наблюдений, превосходящих его по рангу. Аналогично, $K_3 = 0$, $K_4 = 1$, $K_5 = 0$. Сумма этих значений составляет:

$$K = K_1 + K_2 + K_3 + K_4 + K_5 = 2 + 0 + 0 + 1 + 0 = 3,$$

тогда искомое значение $\tau = \frac{4 \cdot 3}{5\,(5-1)} - 1 = -0{,}4$ – умеренная отрицательная связь.

Коэффициент Кендэлла изменяется от -1 до +1: поэтому $\tau = -1$ свидетельствует о сильной отрицательной связи, $\tau = 1$ – о сильной положительной связи, $\tau = 0$ – об отсутствии связи.

Можно привести примеры, когда взаимосвязь переменных является очевидной, но, тем не менее, установить между ними функциональную зависимость (например, зависимость урожая от количества осадков) не представляется возможным. Такие зависимости называют статистическими (корреляционными).

*Корреляционная связь* – согласованное изменение двух признаков, показывающее, что изменчивость одного признака находится в зависимости от изменчивости другого. Корреляционные связи – вероятностные изменения, поэтому их можно изучать только методами математической статистики по выбор-

кам большого объема. Термин «корреляция» был введен английским естествоиспытателем Френсисом Гальтоном в 1886 г. При обработке экспериментальных данных наиболее часто требуется:

1. Установить направление, т. е. понять, с увеличением переменной $x$ переменная $y$ в среднем увеличивается (положительное направление) или имеет тенденцию к уменьшению (отрицательное направление).

2. По данным $(x_1; y_1), (x_2; y_2), \ldots, (x_n; y_n)$, полученным в результате $n$ независимых испытаний, определить форму корреляционной связи, т. е. вид функции.

3. Оценить силу корреляционной связи, т. е. дать оценку степени рассеяния значений $y$ около условного среднего $\overline{y}_x$ или $x$ около $\overline{x}_y$.

Пример. Пусть при исследовании 10 студентов при помощи тестов, проверяющих память (тест 1) и способность к логическому мышлению (тест 2), получены следующие данные (табл. 8.1).

**Таблица 8.1. Результаты тестирования студентов**

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Тест 1 | 5 | 8 | 7 | 10 | 4 | 7 | 9 | 6 | 8 | 6 |
| Тест 2 | 7 | 9 | 6 | 9 | 6 | 7 | 10 | 7 | 6 | 8 |

Приближенно определить значение коэффициента корреляции можно, анализируя диаграмму рассеяния (рис. 6.4). Чем теснее расположены точки относительно некоторой прямой, тем больше по абсолютной величине коэффициент корреляции, и наоборот, чем более расплывчато «облако» точек на диаграмме рассеяния, тем ближе к нулю коэффициент корреляции.

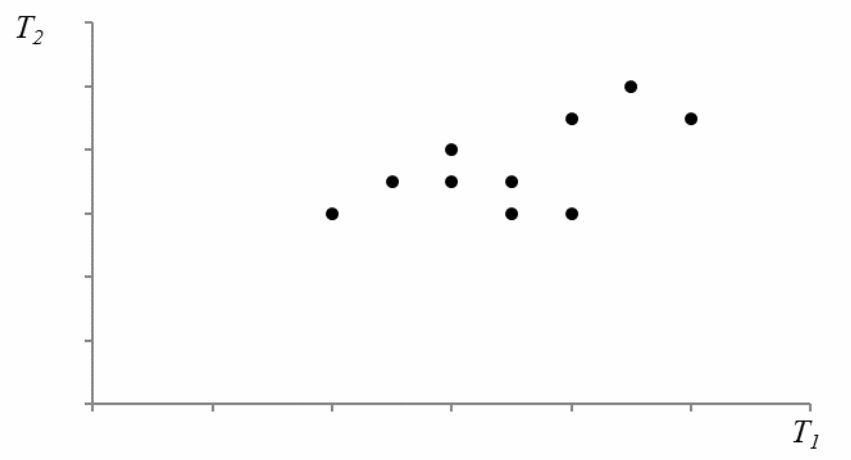



**Рис. 6.4. Диаграмма рассеяния для оценки вида корреляции**

Рассчитаем коэффициент корреляции по данным примера (табл. 6.1):

$$r_{xy} = \frac{10 \cdot 540 - 70 \cdot 75}{\sqrt{(10 \cdot 520 - 4900) \cdot (10 \cdot 581 - 5625)}} = \frac{150}{236{,}6} = 0{,}634 .$$

Полученное значение показывает, что существует умеренная по силе и положительная по направлению связь между памятью и логическим мышлением среди студентов, прошедших тестирование.

*Коэффициент ранговой корреляции* $r_S$-Спирмена. Не всегда, когда нас интересует изменение взаимосвязи между двумя признаками, эти признаки могут быть оценены количественно. Достаточно часто такую оценку получают качественно. Например, несколько городов с разной степенью урбанизации оцениваются по уровню загрязненности окружающей среды. Группа экспертов упорядочивает все города по обоим показателям, а затем интерес может представлять вопрос о согласованности уровня урбанизации и степени загрязненности окружающей

среды. Процесс упорядочивания носит название ранжирования, т. е. приписывания каждому городу ранга в общей иерархии (восходящей или нисходящей). В таком случае обычный коэффициент корреляции не вычисляется.

Метод ранговой корреляции Спирмена позволяет определить тесноту (силу) и направление корреляционной связи между двумя профилями (иерархиями) признаков. Сравниваемыми рядами значений могут быть:

два признака, измеренные в одной и той же группе испытуемых;

две индивидуальные иерархии признаков, выявленные у двух испытуемых по одному и тому же набору признаков (например, иерархии ценностей по методике Р. Рокича или последовательности предпочтений в выборе нескольких альтернатив);

две групповые иерархии признаков;

индивидуальная и групповая иерархия признаков.

Если абсолютная величина $r_S$ достигает критического значения или превышает его, корреляция достоверна.

Проверяемые гипотезы:

$H_0$ – корреляция между переменными (или иерархиями) не отличается от нуля;

$H_1$ – корреляция между переменными (или иерархиями) достоверно отличается от нуля.

Ограничения применения коэффициента Спирмена: по каждой переменной должно быть представлено не менее 5 и не более 40 наблюдений; при большом количестве одинаковых рангов коэффициент ранговой корреляции дает «огрубленные» значения.

Вариант классификации корреляционных связей по их силе:

сильная или тесная при значениях коэффициента корреляции $r > 0,7$;

средняя при $0,5 < r < 0,69$;

умеренная при $0,3 < r < 0,49$;

слабая при $0,2 < r < 0,29$;

очень слабая при $r < 0,19$.

Пример. Рассчитаем коэффициент ранговой корреляции в примере с обследованием городов. Группа городов ранжирована по восходящей схеме по степени урбанизации и загрязненности. Меньшему значению признака, как правило, присваивается меньший ранг. Данные сведены в табл. 8.2, в которой столбец А – уровень урбанизации, столбец В – уровень загрязненности окружающей среды.

**Таблица 8.2. Расчет коэффициента ранговой корреляции**

| $n_{i1}(A)$ | № | Города | $n_{i1}(B)$ | $d$ | $d^2$ |
|---|---|---|---|---|---|
| 3 | 1 | *А* | 2 | 1 | 1 |
| 7 | 2 | *Б* | 4 | 3 | 9 |
| 5 | 3 | *В* | 3 | 2 | 4 |
| 9 | 4 | *Г* | 5 | 4 | 16 |
| 1 | 5 | *Д* | 1 | 0 | 0 |
| 8 | 6 | *Е* | 9 | -1 | 1 |
| 6 | 7 | *Ж* | 8 | -2 | 4 |
| 10 | 8 | *З* | 10 | 0 | 0 |
| 4 | 9 | *И* | 7 | -3 | 9 |
| 2 | 10 | *К* | 6 | -4 | 16 |

Коэффициент ранговой корреляции Спирмена рассчитывается по формуле:

$$r_s = 1 - \frac{6\sum d^2}{n\left(n^2 - 1\right)}$$

В нашем примере $\sum d^2 = 60$, тогда:

$$r_s = 1 - \frac{6 \cdot 60}{10\left(100 - 1\right)} = 1 - \frac{360}{990} = 0{,}636 .$$

По таблице приложения для $n$ = 10 и $\alpha = 0{,}95$ определим $r_{s_{крит}} = 0{,}64$.

*Вывод*: $r_{s_{крит}} > r_{s_{выб}}$, поэтому принимаем гипотезу $H_0$: корреляция между переменными достоверно отличается от нуля. Присутствует прямая, средняя по силе связь $r_S < 0{,}7$ между урбанизацией и загрязненностью городов.

*Замечание*. При наличии одинаковых рангов необходимо в формулу расчета коэффициента Спирмена внести поправки.

Для изучения степени неравномерности распределения определенного суммарного показателя между единицами отдельных групп вариационного ряда в статистике используют *кривую Лоренца* (или кривая концентрации). Кривая Лоренца – это графическое изображение функции распределения. Она была предложена американским экономистом Максом Отто Лоренцем в 1905 году как показатель неравенства в доходах населения. В прямоугольной системе координат кривая Лоренца является выпуклой вниз и проходит под диагональю единичного квадрата, расположенного в I координатной четверти. Каждая точка на кривой Лоренца соответствует утверждению вида: «20 самых бедных процентов населения получают всего 7% дохода». В случае равномерного распределения каждая группа населения

имеет доход, пропорциональный своей численности. Такой случай описывается кривой равенства, являющейся прямой, соединяющей начало координат и точку (1; 1). В случае полного неравенства (когда лишь один член общества имеет доход) кривая сначала «прилипает» к оси абсцисс, а потом из точки (1; 0) «взмывает» к точке (1; 1).

Пример. Пусть имеется распределение городов по числу жителей и распределение населения в этих городах в одном из государств (графы 1, 2, 3). Построить кривую Лоренца.

| **Города с циклом жителей (тыс. чел.)** | **Число городов в % к итогу $W_i$** | **Численность населения % к итогу $Y_i$** | **Кумулятивные итого** | |
|---|---|---|---|---|
| | | | **% городов cum$W_i$** | **% населения cum$Y_i$** |
| До 3 | 4,2 | 0,2 | 4,2 | 0,2 |
| 3-5 | 4,6 | 0,2 | 8,8 | 0,5 |
| 5-10 | 13,1 | 1,7 | 21,9 | 2,2 |
| 10-20 | 28,3 | 6,8 | 50,2 | 9,0 |
| 20-50 | 28,7 | 14,8 | 78,9 | 23,8 |
| 50-100 | 9,7 | 10,3 | 88,6 | 34,1 |
| 100-500 | 9,7 | 33,8 | 98,3 | 67,9 |
| Свыше 500 | 1,7 | 32,1 | 100 | 100 |
| Итого | 100,0 | 100,0 | - | - |

На координатной плоскости наносим точки (cum$W_i$; cum$Y_i$) и по ним строим кривую.

Вывод: чем больше вогнутость (отличие кривой от линии равномерного распределения), тем выше концентрация (численности населения) в определенных группах единиц (крупных го-

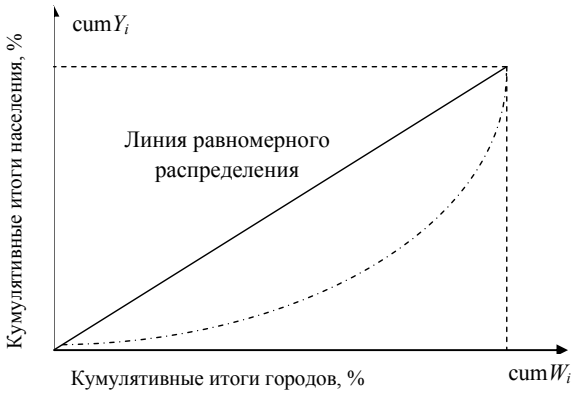



**Рис. 6.5. Кривая Лоренца**

родах). Кривая Лоренца заключена между кривыми равенства и неравенства. Кривые Лоренца применяют для распределений не только доходов, но и имущества домохозяйств, долей рынка для фирм в отрасли, природных ресурсов по государствам.

# Раздел 7.
# Статистические гипотезы

## 7.1. Виды статистических гипотез

Развитие научных знаний свидетельствует о том, что на определенной стадии формирования представлений о некотором объекте становится невозможным его описание только лишь путем установления непосредственных зависимостей между эмпирическими данными. И дело здесь не только в достоверности собранных фактов, а в способах их согласования и необходимости поиска гипотезы, позволяющей судить о природе закономерностей, доступных наблюдению.

*Гипотеза* – это научно обоснованное предположение о структуре социальных объектов, о характере элементов и связей, образующих эти объекты, о механизме их функционирования и развития.

Процесс установления истинности или ложности гипотезы есть процесс ее эмпирического обоснования. В результате такого исследования гипотезы либо опровергаются, либо подтверждаются и становятся положениями теории, истинность которых уже доказана. Общую гипотезу обычно получают в результате предварительного анализа изучаемого социального объекта. Однако в частном эмпирическом исследовании социологи сталкиваются с отдельными сторонами изучаемого объекта, с отдельными его элементами и связями. Поэтому в частных эмпирических исследованиях проверяются не сами гипотезы, а их

следствия, т.е. частные положения, логически вытекающие из гипотезы и описывающие отдельные элементы и связи изучаемого социального объекта.

Нередко одна и та же научная гипотеза подтверждается одними фактами и опровергается другими. Поэтому факты должны быть правильно истолкованы, для того чтобы служить средством проверки. Само использование эмпирических данных в качестве средства проверки гипотезы предполагает активную мыслительную деятельность исследователя и не сводится к пассивному созерцанию. Если установлено, что выведенные следствия ложны и не соответствуют данным, полученным в исследовании, то гипотеза опровергается. По содержанию предположений относительно изучаемого социального объекта различают описательные и объяснительные гипотезы.

*Описательные гипотезы* – это предположения о структурных и функциональных связях изучаемого объекта, которые могут относиться и к классификационным характеристикам социального объекта.

*Объяснительные гипотезы* – это предположения о причинно-следственных связях в изучаемом объекте.

В качестве примера рассмотрим гипотезы, сформулированные в одном исследовании социологов г. Санкт-Петербурга о влиянии содержания труда на отношение к нему в целом.

*Основная гипотеза* состоит в предположении, что содержание труда будет ведущим фактором, определяющим отношение человека к труду и фиксирующимся в объективных и субъективных показателях при данных общих социальных условиях трудовой деятельности. Из этой гипотезы были выведены следствия:

1. Чем выше творческие возможности содержания труда, тем выше объективные показатели отношения к труду.

2. Чем выше творческие возможности работы, тем выше субъективные показатели отношения к труду (удовлетворенность работой).

3. Величина корреляции между содержательностью труда по мере перехода от менее творческих к более творческим профессиям, с одной стороны, и отношением к труду по объективным и субъективным данным – с другой, будет выше, чем величина корреляции между повышением размера заработной платы и теми же показателями отношения к труду.

4. Структура мотивов труда в зависимости от его содержания будет колебаться больше, нежели в зависимости от различий в размере заработка.

*Альтернативная гипотеза* касается структуры мотивации труда. Если подтверждается, что содержание труда есть ведущий мотивационный фактор, определяющий отношение к труду в целом, но одновременно с этим имеются относительные различия в структуре мотивов. Это различие будет проявляться в том, что в группах с более творческим содержанием труда на первый план должны выдвигаться мотивы, связанные с содержанием труда, а в группах с менее творческим содержанием – мотивы, с ним не связанные. Как видно, вторая гипотеза есть развитие первой.

Эти две гипотезы являются не объяснительными, а описательными, поскольку причина здесь не анализируется. В их содержании высказывается лишь предположение о структуре мотивов и о возможной связи между отношением к труду от содержания труда и зависимостью его от заработной платы.

Проверка выводных гипотез возможна лишь в том случае, если все термины, в которых они формулируются, будут подвергнуты эмпирической интерпретации. Например, в первой гипотезе имеются термины: «творческие возможности работы (содержание труда)», «объективные показатели отношения к труду» и термин-связка «выше». При эмпирической интерпретации этих терминов определялись их показатели через наборы наблюдаемых признаков. Так, содержание труда фиксировалось в следующих трех показателях: уровень механизации работы, уровень требуемой квалификации и соотношение затрат физического и умственного труда по данным хронометража на операции физические и мыслительные. В зависимости от сочетания этих трех показателей все профессии были разделены на шесть, упорядоченных классов, в соответствии с содержанием труда - от неквалифицированного ручного труда с постоянной физической нагрузкой до высококвалифицированного труда. Объективных показателей отношения к труду было пять: выработка, качество продукции, уровень ответственности при выполнении срочных заданий, уровень инициативы в работе, повышение производственной квалификации. Эти показатели, выраженные количественно, сводились в единые числовые индексы. Связка «выше» означает, что все классы по содержанию труда упорядочены по указанным показателям от низшего к высшему. То же самое сделано с индексами объективных показателей отношения к труду.

В прикладной деятельности социологу часто приходится принимать решения по результатам выполненных измерений. Одним из компонентов этого процесса можно считать провер-

ку статистических гипотез. Любое предположение о виде и свойствах наблюдаемых в эксперименте случайных величин называется *статистической гипотезой*. Например, в схеме Бернулли одной из гипотез будет предположение «вероятность успеха равна 0,25», в случае нормального распределения – «теоретическая функция распределения нормальна с дисперсией, не превосходящей квадрата среднего значения». Нахождение точечных или интервальных оценок, как правило – предварительная стадия статистических исследований. На этом шаге обычно формулируется предположение относительно вида функции распределения (*непараметрическая гипотеза*), либо предположения относительно значений параметров функции распределения (дисперсия или математическое ожидание) известного вида (*параметрическая гипотеза*). Проблемы, возникающие при исследовании случайной величины, сводятся к оценке истинности одной или нескольких выдвигаемых гипотез на основе результатов анализа накопленной информации. В ходе решения практических задач выдвигаются следующие гипотезы:

об общем виде закона распределения исследуемой случайной величины (проверка таких гипотез дает возможность установить закон распределения с точностью до параметров, которые характеризуют неизвестный экспериментатору закон распределения);

об однородности двух или нескольких выборок (такие гипотезы позволяют сделать вывод о равенстве или различии законов распределения случайной величины, характеризующих изучаемое свойство);

о числовых значениях характеристик исследуемого явления или процесса (используя эти гипотезы, сравнивают значения числовых параметров с заданными значениями);

об общем виде зависимости, существующей между компонентами исследуемого многомерного признака (данные гипотезы позволяют определить характер зависимости между свойствами исследуемого признака);

о независимости и стационарности ряда наблюдений.

Главная задача статистических гипотез – выбор правильного решения из двух альтернативных. Основную (выдвинутую) гипотезу называют нулевой гипотезой $(H_0)$. Она называется нулевой, потому что выполняется равенство: $x_1 - x_2 = 0$, где $x_i$ – сопоставляемые значения признаков. Обычно нулевые гипотезы учитывают, что различие между сравниваемыми величинами (выборками) отсутствуют, а наблюдаемые отклонения объясняются лишь случайными колебаниями выборки.

*Альтернативной* $(H_1)$ называется гипотеза, конкурирующая с нулевой гипотезой в том смысле, что если нулевая гипотеза отвергается, то принимается альтернативная. Выбор альтернативной гипотезы зависит от поставленной задачи. Например, $H_0: \sigma = \sigma_0$, тогда $H_1: \sigma \neq \sigma_0$ или $\sigma < \sigma_0$, или $\sigma > \sigma_0$. Такая гипотеза доказывает предположение исследования, поэтому ее называют экспериментальной. Когда нужно убедиться, что выборки не различаются между собой по каким-либо характеристикам, то это означает подтверждение нулевой гипотезы. Чаще требуется доказать значимость различий, ибо они более информативны. Если в процессе исследований было замечено, что в одной из групп индивидуальные значения, например по соци-

альной смелости, выше, а в другой ниже, то для проверки значимости этих различий лучше сформулировать *направленные гипотезы*. Например, $H_0$: $x_1$ не превышает $x_2$; $H_1$: $x_1$ превышает $x_2$. Если нужно доказать, что различаются формы распределения признака по группам, то формулируются ненаправленные гипотезы: $H_0$: $x_1$ не отличается от $x_2$; $H_1$: $x_1$ отличается от $x_2$.

## 7.2. Правила принятия гипотез

Непосредственно определить истинность гипотезы нельзя, поэтому проверка статистических гипотез заключается в установлении согласования данных наблюдения с выдвинутой гипотезой. Процесс обоснованного сравнения сформулированной гипотезы с полученными в результате выборки данными называют *статистической проверкой гипотез*. Если данные наблюдения противоречат высказанной гипотезе, то говорят, что результат сопоставления высказанной гипотезы с выборочными данными отрицателен. В этом случае гипотезу следует *отклонить*. Проверка статистической гипотезы не дает логического доказательства ее верности или неверности и в чем-то похожа на обоснование математических утверждений. Чтобы опровергнуть утверждение, достаточно привести один контрпример, однако для доказательства справедливости недостаточно любого числа примеров. Так и в статистике говорят: «нулевая гипотеза не отвергается, так как полученные данные ей не противоречат». Кроме того, гипотеза может быть отвергнута на основании других выборочных данных или по другим причинам, поэтому принятие нулевой гипотезы нельзя расценивать как точно установленный,

содержащийся в ней факт, а только как достаточно правдоподобное, не противоречащее эксперименту утверждение.

Для проверки нулевой гипотезы используют специальным образом подобранную случайную величину, точное или приближенное распределение которой известно, и называют ее статистическим критерием (критерием). В зависимости от вида распределения случайную величину обозначают по-разному: в случае нормального распределения – $Z$, в случае распределения Фишера – $F$, в случае распределения Стьюдента – t, если рассматривается закон Пирсона (хи-квадрат) – $\chi^2$. В общем случае критерий обозначают $K$. Например, если проверяют гипотезу о равенстве дисперсий двух нормальных генеральных совокупностей, то в качестве критерия $K$ принимают отношение исправленных выборочных дисперсий:

$$K = \frac{S_1^2}{S_2^2}.$$

Под *статистическим критерием* также понимают правило, показывающее, когда статистическую гипотезу на основании опытных значений $x_1, x_2, \ldots, x_n$ величины следует принять, а когда – отвергнуть.

Критерии, применяемые в гипотезах, не требующих знаний о виде функции распределения, аналогично гипотезам называются *непараметрическими*. Непараметрические критерии основаны на оперировании частотами или рангами. Критерии, использующиеся для проверки гипотез о параметрах распределения генеральной совокупности, называются *параметрическими*. Параметрические критерии включают в формулу расчета параметры распределения,

то есть средние значения и дисперсии. Каждая группа критериев имеет свои преимущества и недостатки. Параметрические критерии, как правило, оказываются более мощными в случаях измерения признаков по интервальной шкале с нормальным распределением. Непараметрические критерии не требуют сложных расчетов, но они ограничены в том, что с их помощью невозможно оценить взаимодействие двух или более условий или факторов, влияющих на изменение признака.

При проверке гипотезы выборочные данные могут либо согласовываться с основной гипотезой, либо ей противоречить, а значит, всегда есть риск принять неверное решение. Причины расхождения результатов выборки и теоретических характеристик различны. Это и неудачный способ группировки данных, и недостаточный объем выборки, и неправильный выбор распределения, и др.

*Правило отклонения* $H_0$ и *принятия* $H_1$. Если эмпирическое значение критерия равняется критическому значению, соответствующему $p < 0{,}05$, или превышает его, то гипотеза $H_0$ отклоняется, но мы еще не можем определенно принять гипотезу $H_1$. Если эмпирическое значение критерия равняется критическому значению, соответствующему $p < 0{,}01$ или превышает его, то гипотеза $H_0$ отклоняется и принимается гипотеза $H_1$.

Ось значимости строится для наглядного представления процесса принятия решения (рис. 7.1).

*Исключения*: критерий знаков, критерий Вилкоксона, критерий Манна-Уитни. Для них устанавливаются обратные соотношения эмпирических и критических значений, а ось значимости отображается зеркально.


**Рис. 7.1. Ось значимости**

Критические значения критерия на рисунке обозначены, как $K_{0,05}$ и $K_{0,01}$, а эмпирическое значение критерия как $K_{эмп}$. Вправо от критического значения $K_{0,01}$ располагается зона значимости – в нее попадают безусловно значимые эмпирические значения. Влево от критического значения $K_{0,05}$располагается зона незначимости – в нее попадают безусловно незначимые эмпирические значения критерия. При попадании эмпирического значения в область между $K_{0,05}$ и $K_{0,01}$то есть зону неопределенности можно отклонить гипотезу о недостоверности различий $H_0$, но еще нельзя принять гипотезу об их достоверности $H_1$.

Сначала по данным выборки определяют частные значения входящих в критерий величин, а затем рассчитывают сам критерий. Значения критерия, вычисленные по выборке, называются наблюдаемым (эмпирическим) значением – $K_{набл}$. Для каждого критерия имеются соответствующие таблицы, по которым, используя данные выборки, находят $K_{крит}$. Для большинства критериев при $K_{набл} \geq K_{крит}$ нулевую гипотезу отвергают, при $K_{набл} < K_{крит}$ нет оснований, чтобы отвергнуть нулевую гипотезу (в этом случае считают, что данные наблюдений согласуются с нулевой гипотезой).

Множество значений критерия, при которых гипотеза $H_0$ отклоняется и принимается гипотеза $H_1$, называется *критиче-*

*ской областью*. Границы критической области называют *критическими точками* ($K_{крит}$). Критическая область, представляющая собой промежуток $(k_{кр}^{п}; \infty)$, называется *правосторонней* ($k_{кр}^{п}$ определяется из условия $P(K > k_{кр}^{п}) = \alpha$). Критическая область, представляющая собой промежуток $(-\infty; k_{кр}^{л})$, называется *левосторонней* ($k_{кр}^{л}$ определяется из условия $P(K < k_{кр}^{л}) = \alpha$). Критическая область, представляющая собой два промежутка $(-\infty; k_{кр}^{л})$ и $(k_{кр}^{п}; \infty)$, называется *двусторонней*^

$\left(k_{кр}^{л}\right.$ и $k_{кр}^{п}$ определяются из условий:

$$P(K < k_{кр}^{л}) = \frac{\alpha}{2} \text{ и } P(K > k_{кр}^{п}) = \left.\frac{\alpha}{2}\right).$$

Основной *принцип проверки статистических гипотез*: гипотеза, попадающая в критическую область, отвергается, а альтернативная гипотеза принимается как одна из возможных. В практической деятельности по обработке эмпирических данных существует возможность совершить массу различных ошибок. Не является исключением и проверка статистических гипотез, которые принимаются на основе выборочных данных. Ошибки могли возникнуть уже в процессе сбора и предварительной обработки данных. С точки зрения статистической проверки гипотез существует два вида ошибок, называемых *ошибкой I рода* и *ошибкой II рода*.

*Ошибкой I рода* называется ошибка отклонения верной нулевой гипотезы $(H_0)$, которая на самом деле верна. *Ошибкой II рода* называется ошибка принятия ложной гипотезы $(H_0)$. Вероятность $\alpha = P_{H_0}(H_1)$ совершения ошибки *I* рода называется *уровнем значимости* статистического критерия. Выбор величины уровня значимости $\alpha$ зависит от размера потерь в случае

ошибочного решения. Иными словами, уровень статистической значимости – это вероятность признать различия существенными (приняли альтернативную гипотезу и отклонили нулевую), а они в действительности случайные. Например, если указывается, что различия достоверны на 5%-ном уровне значимости, то подразумевается вероятность 0,05 того, что они все же недостоверные. Исторически сложилось так, что в гуманитарных исследованиях используют стандартные значения уровня значимости: $\alpha$ = 0,1; 0,05; 0,01. Наиболее распространенной является величина уровня значимости $\alpha$ = 0,05, которая означает, что высказанная гипотеза будет ошибочно отклонена в среднем в пяти случаях из ста. Если альтернативная гипотеза $H_1$ единственная, то можно вычислить вероятность ошибки *II* рода – $\beta = P_{H_1}(H_0)$.

Вероятность несовершения ошибки *II* рода, иначе вероятность отклонения неверной гипотезы, называется *мощностью статистического критерия*. Мощность критерия – это его способность выявлять различия, если они есть, то есть отклонить нулевую гипотезу об отсутствии различий, если она неверна. Одни и те же задачи могут быть решены с помощью различных критериев, при этом эмпирически устанавливается, что одни из них позволяют выявить различия там, где другие оказываются неспособными это сделать, или выявляют более высокий уровень значимости различий. Возникает вопрос: зачем использовать менее мощные критерии? Основанием для выбора критерия является не только его мощность, но и другие характеристики: простота расчетов; более широкий диапазон использования; применимость к выборкам разного объема; большая наглядность и информативность результатов.

Уровнем значимости, а значит, и вероятностью ошибки первого рода, можно управлять. Можно установить любую приемлемую степень риска для неправильного вывода по выборочным данным об ошибочности выдвинутой гипотезы. Заметим, что уровень значимости и мощность критерия связаны между собой, причем их связь нелинейная. Поэтому произвольно изменять уровень значимости не следует, так как неоправданное уменьшение ошибки первого рода может привести к существенной потере мощности критерия.

Пример. В ходе исследования рынка труда по группе компаний изучалась дискриминация по возрасту. Работодатели полагают, что ее доля составляет 3%, соискатели – 20%. Достигнута следующая договоренность в трудовой инспекции: если при проверке 10 случайно отобранных соискателей вакансий будет обнаружено не более одного случая дискриминации, то штрафные санкции на группу компаний не накладываются. Сформулируйте нулевую и альтернативную гипотезы с точки зрения соискателя вакансий. Определите критическую область и область принятия нулевой гипотезы. Сформулируйте, в чем состоят ошибки *I* и *II* рода. Найдите их вероятности.

*Решение*.

С точки зрения соискателя нулевой гипотезой $H_0$ будем считать гипотезу о 20% случаев дискриминации по возрасту, а альтернативной гипотезой $H_1$ версию работодателей о 3% таких.

Поскольку отбирается 10 человек, а затем фиксируется число случаев дискриминации, то множеством всевозможных результатов испытаний будет:

$$A = \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$$

и случаев дискриминации может оказаться 0, 1, …, 10. По условиям задачи гипотеза соискателя отвергается, если $K \le 1$, следовательно, критическая область – $A_{крит} = \{0, 1\}$, а область принятия гипотезы – $\{ 2, 3, 4, 5, 6, 7, 8, 9, 10 \}$.

Ошибка *I* рода: штрафных санкций не будет, даже, если будет выявлено 20% случаев дискриминации по возрасту.

Ошибка *II* рода: штрафных санкций не будет, даже, если будет выявлено 3% случаев дискриминации по возрасту.

Найдем вероятности этих ошибок. Если нулевая гипотеза верна, то вероятность того, что один случайно выбранный человек попал под случай дискриминации по возрасту, составляет 0,2. Ошибка *I* рода произойдет, если из 10 человек попал под случай дискриминации по возрасту не более чем один (0 или 1). В случае истинности нулевой гипотезы $H_0$ число тех, кто попал под дискриминацию по возрасту $K$, является биноминальной случайной величиной $B_i(10;\ 0{,}2)$, поэтому $P(K = 0) = 0{,}8^{10} \approx 0{,}107$, $P(K = 1) = 10 \cdot 0{,}8^9 \cdot 0{,}2 \approx 0{,}268$.

Аналогично вычислим вероятность $\alpha$ ошибки *I* рода:

$$\alpha = P(K \le 1 \,|\, H_0) = P(K = 0 \,|\, H_0) + P(K = 1 \,|\, H_0) = \\ = 0{,}107 + 0{,}268 = 0{,}375.$$

Если верна альтернативная гипотеза $H_1$, то вероятность того, что один случайно выбранный человек попал под дискриминацию по возрасту составляет 0,03. Ошибка *II* рода произойдет, если из 10 человек под случай дискриминации по возрасту попадут два или более.

В случае истинности альтернативной гипотезы $K$ также является биноминальной случайной величиной, но с другим па-

раметром – $B_i(10;0{,}03)$, поэтому $P(K=0)=0{,}97^{10}\approx 0{,}737$, $P(K=1)=10\cdot 0{,}97^9\cdot 0{,}03\approx 0{,}228$. Следовательно, для вероятности $\beta$ ошибки II рода имеем:

$$\beta = P(K>1\,|\,H_a)=1-P(K\le 1\,|\,H_a)=1-P(K=0\,|\,H_a)-P(K=1\,|\,H_a)=0{,}035.$$

*Замечание*. Из сравнения вероятностей $\alpha$ и $\beta$ следует, что оговоренная процедура проверки выгодна скорее работодателю.

Задачи, сводящиеся к оценке истинности нулевой гипотезы $H_0$ по отношению к конкурирующей гипотезе $H_1$, можно решить, используя различные статистические критерии.

Несмотря на разнообразие, как самих гипотез, так и применяемых статистических критериев, существует некий алгоритм:

1. На основе выборочных данных $x_1, x_2, \ldots, x_n$ с учетом конкретных условий задачи выдвигают нулевую гипотезу ($H_0$) и конкурирующую с ней альтернативную гипотезу $H_1$.

2. Выбирают уровень значимости α.

3. Выбирают (если он не задан) объем выборки $n$ и число степеней свободы $r$.

4. На основе выборочных данных выбирают критерий (статистику) – некоторую функцию $K$ от результатов наблюдений и условий рассматриваемой статистической задачи, подчиненную некоторому закону распределения.

5. По таблицам, в зависимости от объема выборки и уровня значимости α критических точек ($K_{крит}$), определяют критическую область *W*. В этом случае возможно совершение ошибки *I* рода.

6. Определяют на основе выборочных данных $x_1, x_2, \ldots, x_n$ численную величину критерия $K_{набл}$ по формулам, учитывающим характер проверяемой гипотезы.

7. Вырабатывают решение: если $K_{набл}$ попадает в критическую область, то нулевая гипотеза отклоняется и принимается альтернативная.

Проанализировать выбранное решение можно, используя таблицу:

| **Выбранное решение** | **Реальная ситуация** | |
|---|---|---|
| | $H_0$ **– истинная** $H_a$ **– ложная** | $H_0$ **– ложная** $H_a$ **– истинная** |
| Выбрали $H_0$ и отвергли $H_a$ | $P_{H_0}(H_0) = 1 - \alpha$ Правильное решение | $P_{H_a}(H_0) = \beta$ Ошибка *II* рода |
| Выбрали $H_a$ и отвергли $H_0$ | $P_{H_0}(H_a) = \alpha$ Ошибка *I* рода | $P_{H_a}(H_a) = 1 - \beta$ Правильное решение |

Для большей уверенности перед окончательным принятием гипотезы, желательно подвергнуть ее проверке с помощью других критериев или повторить эксперимент, увеличив объем выборки.

Пример. Компании необходимо принять решение о вхождении на рынок со своим товаром. В течение 8 месяцев при существующем маркетинге на одном из объектов остатки товара составили 10%, что принято за норму рентабельности. Для проверки была произведена случайная выборка 500 объектов компании, 60 из которых оказались малорентабельными. Согласуются ли данные выборки с утверждением «доля рентабельности

(генеральной совокупности) соответствует установленному нормативу»? Уровень значимости при проверке гипотезы принять $\alpha = 0{,}05$.

*Решение*.

В данной задаче нужно проверить гипотезу о том, что доля признака $p$ равна определенной величине $\delta$. Нулевая гипотеза $H_0$: $p = \delta = 0{,}1$, альтернативная гипотеза $H_1$: $p > 0{,}1$. С целью проверки нулевой гипотезы используем критерий

$K = \dfrac{m}{n}$ – выборочную долю признака в выборке.

Данный критерий имеет биноминальное распределение, но при большом $n$ и не очень маленьком $\delta$ ($n > 50$, $n\delta \geq 10$) это распределение приближается к нормальному с центром в точке $\delta$ и средним квадратичным отклонением

$$\sigma\left(\frac{m}{n}\right) = \sqrt{\frac{\delta(1-\delta)}{n}}.$$

Следовательно, величина

$$z = \frac{\dfrac{m}{n} - \delta}{\sqrt{\dfrac{\delta(1-\delta)}{n}}}$$

распределена по стандартному нормальному закону. Критическую точку определяем из условия:

$$P\left\{\frac{\dfrac{m}{n} - \delta}{\sqrt{\dfrac{\delta(1-\delta)}{n}}} \leq z\right\} = 0{,}5 + \Phi(z) = 1 - \alpha.$$

Отсюда $\Phi(z) = 1 - \alpha - 0{,}5 = 0{,}45$.

По таблице функций Лапласа (см. прил.) находим $z = 1{,}65$ и определяем критическую точку:

$$K_{крит} = \delta + z\sqrt{\frac{\delta(1-\delta)}{n}} = 0,1 + 1,65 \cdot \sqrt{\frac{0,1(1-0,1)}{500}} =$$

$$= 0,1 + 0,022 = 0,122 .$$

Выборочная доля $K_{набл} = \frac{m}{n} = \frac{60}{500} = 0,12$.

Имеем: $K_{набл} < K_{крит}$. Последнее означает, что нулевая гипотеза $H_0$ не отвергается. Доля нерентабельности объектов в выборочной совокупности равная12% превысила норматив, однако такое отклонение находится в допустимых пределах и может объясняться случайностью отбора.

Установление теоретического закона распределения случайной величины, характеризующей изучаемый признак по эмпирическому распределению, – основная задача математической статистики. Для ее решения необходимо установить вид и параметры закона распределения. Как бы хорошо ни подобрали теоретический закон распределения, расхождения между эмпирическим и теоретическим распределениями неизбежны. Случайно ли расхождение? Можно ли объяснить его малым числом наблюдений, способом группировки или другими причинами? Возможно, расхождение не случайно и объясняется ложной гипотезой о виде распределения. Ответить на эти вопросы и провести проверку согласия эмпирической функции распределения с предположением относительно теоретической функции распределения $F(x)$ позволяют критерии согласия, в которых гипотеза либо полностью, либо с точностью до небольшого числа параметров определяет закон распределения.

*Замечание*. Критерии согласия не доказывают справедливость гипотезы, а лишь устанавливают на принятом уровне ее согласие или несогласие с данными наблюдений.

Потребности социальной практики требуют применения методов количественного описания социальных процессов, позволяющих точно регистрировать не только количественные, но и качественные факторы. С этой целью разработаны несколько групп критериев различной мощности, о которых подробно будет рассказано в следующем разделе.

При проверке статистических гипотез принципиально существуют четыре возможности:

гипотеза верна, и она принимается;

гипотеза верна, но она отвергается (ошибка первого рода);

гипотеза неверна, и она отвергается;

гипотеза неверна, но она принимается (ошибка второго рода).

Ошибки первого и второго рода существенно различаются между собой по значимости, и это оказывает большое влияние на всю процедуру проверки статистических гипотез. Необходимо еще раз подчеркнуть, что никакая гипотеза не может быть окончательно принята или отвергнута. Поэтому используемые, довольно категорические, утверждения «принять» и «отвергнуть» являются просто условным сокращением выражений вида «опытные данные не противоречат выдвинутой гипотезе» и «опытные данные противоречат выдвинутой гипотезе».

# Раздел 8.
# Параметрические статистические критерии

## 8.1. Общие положения и задачи, критерий Пирсона

*Статистический критерий* – это решающее правило, которое обеспечивает математически обоснованное принятие истинной и отклонение ложной гипотезы Статистические критерии определяют в практической деятельности метод расчета определенного числа, которое обозначается как эмпирическое значение критерия, например, числа $\chi^2$ для критерия Пирсона.

Соотношение эмпирического и критического значений критерия является основанием для подтверждения или опровержения гипотезы. Согласно статистических гипотез и статистические критерии делятся на параметрические и непараметрические. Выбор критериев достаточно широк, в чем можно убедиться, ознакомившись с приведенными в списке литературы публикациями. Однако нашей целью является описание статистических критериев, адекватных типовым для социологических исследований задачам анализа данных.

Применение критериев позволяет устанавливать различия по уровню исследуемого признака между двумя, тремя и более выборками испытуемых, например, определение различий между работниками государственных предприятий и частных фирм, между людьми разной культуры, возрастные различия и т. д.

В результате таких исследований формируется статистически достоверный групповой профиль или усредненный портрет человека той или иной профессии, статуса, например, успешный менеджер, успешный политик.

Критерии различий предполагают, что сопоставляются независимые выборки, состоящие из разных испытуемых. Решение о выборе того или иного критерия принимается на основании количества и объема сопоставляемых выборок.

*Параметрические критерии* используются в задачах проверки параметрических гипотез и включают в свой расчет показателей распределения, например, средние, дисперсии и т. д. Это такие известные классические критерии, как $\chi^2$-критерий Пирсона, *t*-критерий Стьюдента, F-критерий Фишера.

*Параметрические критерии* позволяют прямо оценить уровень основных параметров генеральных совокупностей, разности средних и различия в дисперсиях Критерии способны выявить тенденции изменения признака при переходе от условия к условию, оценить взаимодействие двух и более факторов в воздействии на изменения признака. Параметрические критерии считаются несколько более мощными, чем непараметрические, при условии, что признак измерен в интервальной шкале и его распределение близко к нормальному. Однако в интервальной шкале могут возникнуть определенные проблемы, в частности, если данные, представлены не в стандартизированных оценках. К тому же проверка распределения «на нормальность» требует достаточно сложных расчетов.

Пример. В случайном порядке отобрано 60 студентов. Их оценки по итогам экзамена по курсу высшей математики:

4, 3, 5, 3, 4, 4, 3, 3, 4, 2, 3, 3, 3, 5, 3, 2, 3, 4, 5, 5, 4, 5, 4, 4, 3, 4, 2, 4, 3, 4, 4, 4, 3, 3, 4, 4, 3, 2, 4, 4, 2, 3, 3, 2, 5, 3, 2, 5, 3, 4, 3, 2, 3, 3, 5, 3, 4, 3, 3, 3.

Проверьте при уровне статистической значимости 0,05 гипотезу о нормальном законе распределения студентов по уровню знаний, продемонстрированных на экзамене по высшей математике.

*Решение*.

$H_0$: распределение студентов по уровню знаний продемонстрированных на экзамене по высшей математике (по оценкам) подчинено нормальному закону распределения.

$H_1$: распределение оценок значимо отличается от нормального распределения.

Чтобы получить предварительное представление о распределении студентов по уровню знаний по высшей математике (по оценкам), построим вариационный ряд. Изучаемый признак (оценка) принимает ограниченное число целых значений, поэтому удобно построить дискретный ряд и подсчитать число студентов, имеющих одинаковую оценку. Сгруппированные данные представим в виде вариационного ряда:

| Оценка ($x_i$) | 2 | 3 | 4 | 5 |
|---|---|---|---|---|
| Число студентов ($n_i$) | 8 | 25 | 19 | 8 |

Проверим длину выборки:

$$n = \sum_{i=1}^{4} n_i = 8 + 25 + 19 + 8 = 60 .$$

Это значение совпадает с количеством экспериментальных данных. Найдем параметры распределения: выборочную сред-

нюю (средний балл) $\overline{x}_в$, выборочную дисперсию $D_в$ и среднее квадратичное отклонение выборки $S$:

$$\overline{x}_в = \frac{2\cdot 8 + 3\cdot 25 + 4\cdot 19 + 5\cdot 8}{60} = \frac{207}{60} = 3{,}45;$$

$$\overline{x}_в^2 = \frac{2^2\cdot 8 + 3^2\cdot 25 + 4^2\cdot 19 + 5^2\cdot 8}{60} = \frac{761}{60} \approx 12{,}68,$$

$$D_в = \overline{x_в^2} - (\overline{x}_в)^2 = 12{,}68 - 3{,}45^2 \approx 0{,}78,$$

$$s = \sqrt{D_в} = \sqrt{0{,}78} \approx 0{,}88.$$

Для проверки нулевой гипотезы рассчитаем теоретические частоты нормального распределения с параметрами $a$ и $\sigma$. В качестве оценок $a$ и $\sigma$ возьмем оценки наибольшего правдоподобия, положив $a = \overline{x}_в = 3{,}45$ и $\sigma \approx s \approx 0{,}88$.

Расчеты сведем в таблицу:

| № | $x_i$ | $n_i$ | $x_i - a$ | $z_i = \frac{x_i - a}{\sigma}$ | $\varphi(z_i)$ | $P_i = \Delta z_i \cdot \varphi(z_i)$ | $P_i$ исправленное | $n \cdot P_i$ |
|---|---|---|---|---|---|---|---|---|
| 1 | 2 | 8 | -1,45 | -1,65 | 0,1023 | 0,117 | 0,123 | 7,4 |
| 2 | 3 | 25 | -0 | 0,51 | 0,3503 | 0,400 | 0,400 | 24,0 |
| 1 | 4 | 19 | 0,55 | 0,625 | 0,3281 | 0,374 | 0,374 | 22,4 |
| 4 | 5 | 8 | 1,55 | 1,76 | 0,0848 | 0,097 | 0,103 | 6,2 |
| $\Sigma$ | | 60 | | | | 0,988 | 1,000 | 60,0 |

*Замечание* $z_i$ – нормированное значение отклонений $x_i$ от математического ожидания $a$; $\varphi(z_i)$ – значения локальной функции Лапласа (См. Приложение)

$$\varphi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}};$$

$\Delta z_i = z_2 - z_1 = z_3 - z_2 = z_4 - z_3 = -0{,}51 + 1{,}65 = 1{,}14$;

$n \cdot P_i$ – теоретические частоты.

Расчетная сумма вероятностей $P_i$:

$\sum_{i=1}^{4} P_i = 0{,}988$ отлична от единицы.

Чтобы сумма вероятностей равнялась единице, к крайним значениям $z_i$, т. е. к $z_1$ и $z_4$, прибавим

$$\frac{1-0{,}988}{2} = 0{,}006.$$

В последнем столбце таблицы задействованы «исправленные» вероятности $P_i$.

Если нулевая гипотеза $H_0$ справедлива, то частоты выборки $n_i$ должны незначительно отличаться от теоретических частот $n \cdot P_i$. Для оценки расхождения эмпирических и теоретических частот используем критерий:

$$\sum_{i=1}^{4} \frac{(n_i - nP_i)^2}{nP_i} = \chi^2_{набл.}.$$

Расчеты сведем в таблицу:

| № | $n_i$ | $n \cdot P_i$ | $n_i - nP$ | $(n_i - nP)^2$ | $\frac{(m_i - nP_i)^2}{nP_i}$ |
|---|---|---|---|---|---|
| 1 | 8 | 7,4 | 0,6 | 0,36 | 0,05 |
| 2 | 25 | 24,0 | 1,0 | 1,00 | 0,04 |
| 1 | 19 | 22,4 | –3,4 | 11,56 | 0,52 |
| 4 | 8 | 6,2 | 1,8 | 3,24 | 0,52 |
| $\sum$ | 60 | 60,0 | | | $\chi^2_{набл.}$ =1,13 |

Наблюдаемое значение критерия $\chi^2_{набл.} = 1{,}13$. Определим число степеней свободы: случайная величина $X$ принимает всего четыре значения ($k = 4$), неизвестных параметров – два ($m = 2$), поэтому число степеней свободы $\nu = 4 - 3 = 1$. Из При-

ложения при уровне значимости $\alpha = 0{,}05$ и числе степеней свободы $\nu = 1$ находим: $\chi^2_{крит} = 3{,}84$. Таким образом $\chi^2_{набл.} = 1{,}13 < \chi^2_{крит.} = 3{,}84$.

*Вывод*: $\chi^2_{набл.}$ попадает в незначимую область критерия Пирсона. Нет оснований для отказа от гипотезы $H_0$, т. е. можно утверждать, что распределение студентов по уровню знаний продемонстрированных на экзамене по математике согласуется с нормальным распределением.

Некоторые замечания относительно типов шкал. Таблицы сопряженности обычно строятся для переменных, измеренных по номинальным шкалам, однако мы знаем, что шкалу любого типа можно перекодировать в номинальную. Например, есть переменная «доход», со значениями, лежащими в диапазоне от 10 000 до 25 000 руб. Эта переменная имеет 15 001 потенциальных значений. Реальных значений, естественно, будет меньше, но все равно было бы странно включать такую переменную в таблицу сопряженности, где для каждого значения переменной строится отдельная строка или столбец. Поэтому, если нам необходимо все-таки включить эту переменную в таблицу, то имеет смысл разбить диапазон ее значений на интервалы, например, 10 000-15 000, 15 000-20 000 и 20 000-25 000. В этом случае рассматриваемая переменная будет иметь только три возможных значения, и ее можно будет включить в таблицу сопряженности. Можно выбрать любой другой вариант разбиения на интервалы, причем здесь наблюдается интересный эффект – если одну и ту же переменную по-разному разбивать на интервалы и строить соответствующие таблицы сопряженности, то в зависимости от варианта разбиения результат проверки гипотезы о независимо-

сти может меняться, т. е. может получиться так, что в одном случае гипотезу придется принять, в другом – отвергнуть. Это объясняется следующим образом. Если вариант разбиения на интервалы выбирается из содержательных соображений, переменной придается некий дополнительный смысл. Например, у нас есть переменная «возраст» со значениями в диапазоне от 0 до 30. Можно просто разбить ее, скажем, на три равных интервала 0-10, 10-20, 20-30. А можно разбить на интервалы 0-7, 7-17, 17-22, 22-30. Здесь у каждого интервала будет смысл: первый интервал – раннее детство; второй – школа; третий – институт; четвертый – период, когда человек начинает трудовую деятельность. Можно разбивать переменную на интервалы исходя из каких-либо других содержательных соображений, и в каждом случае исходная переменная приобретает разный смысл, т. е. мы получаем *разные номинальные* переменные на основе исходной интервальной переменной. Поэтому неудивительно, что новые переменные ведут себя по-разному.

При работе с критерием $\chi^2$ есть важное ограничение. Если в таблице присутствуют малые теоретические частоты, статистика $\chi^2$ не будет иметь $\chi^2$-распределение, и соответственно нельзя будет проверить гипотезу. Однозначного ответа на вопрос, какие частоты считать малыми, нет. Обычно в качестве границы рассматривается число «5», т. е. если есть теоретические частоты, меньшие 5, критерий $\chi^2$ применять нельзя. Бывают ситуации, когда эту проблему удается решить за счет объединения категорий переменных.

Рассмотрим пример. Пусть у нас есть переменная «удовлетворенность работой» с возможными значениями «работа со-

всем не нравится», «работа скорее не нравится, чем нравится», «работа скорее нравится, чем нет», «работа очень нравится», и переменная категория сотрудника» с возможными значениями «руководитель» и «не руководитель». Построим таблицу сопряженности, чтобы посмотреть, зависит ли степень удовлетворенности работой от того, является ли человек начальником:

| Категория сотрудников | Работа совсем не нравится | Работа скорее не нравится, чем нравится | Работа скорее нравится, чем нет | Работа очень нравится | Сумма |
|---|---|---|---|---|---|
| Руководитель | 2 | 4 | 8 | 3 | 17 |
| Не руководитель | 3 | 4 | 9 | 2 | 18 |
| Сумма | 5 | 8 | 17 | 5 | 35 |

В таблице сопряженности приводятся наблюдаемые частоты, меньшие 5, но, как мы помним, с числом «5» надо сравнивать теоретические частоты. Найдем значение теоретической частоты для первой ячейки. Получим:

$$\mu_{11} = \frac{17 \cdot 5}{35} = 2{,}4 .$$

Аналогично найдем остальные теоретические частоты и запишем в таблицу:

| Категория сотрудников | Работа совсем не нравится | Работа скорее не нравится, чем нравится | Работа скорее нравится, чем нет | Работа очень нравится |
|---|---|---|---|---|
| Руководитель | 2,4 | 3,9 | 8,3 | 2,4 |
| Не руководитель | 2,6 | 4,1 | 8,7 | 2,6 |

Как видно из таблиц достаточно теоретических частот, меньших 5. Объединим категорию переменной «удовлетворен-

ность работой»: категории «работа совсем не нравится» и «работа скорее не нравится, чем нравится» в одну и назовем ее «негативное отношение к работе»; аналогично объединим категории «работа скорее нравится, чем нет» и «работа очень нравится» в одну и назовем ее «позитивное отношение к работе». При укрупнении разрядов часть информации теряется, но это, к сожалению, неизбежно.

В рассматриваемом примере наблюдаемые частоты для соответствующих категорий объединяются и записываются в таблицу:

| **Категория сотрудников** | **Негативное отношение к работе** | **Позитивное отношение к работе** |
|---|---|---|
| Руководитель | 2 + 4 = 6 | 8 + 3=11 |
| Не руководитель | 3 + 4 = 7 | 9 + 2=11 |

В итоге получаем таблицу сопряженности переменных после объединения категорий:

| **Категория сотрудников** | **Негативное отношение к работе** | **Позитивное отношение к работе** | **Сумма** |
|---|---|---|---|
| Руководитель | 6 | 11 | 17 |
| Не руководитель | 7 | 11 | 18 |
| Сумма | 13 | 22 | 35 |

Найдем теоретические частоты для данных:

| **Категория сотрудника** | **Негативное отношение к работе** | **Позитивное отношение к работе** |
|---|---|---|
| Руководитель | 6,3 | 10,7 |
| Не руководитель | 6,7 | 11,3 |

Таким образом, в полученной таблице нет теоретических частот меньше 5, значит, можно использовать критерий $\chi^2$.

Пример. Рассмотрим схему использования критерия $\chi^2$ для проверки гипотезы о независимости переменных. В качестве исходных данных возьмем результаты из таблицы сопряженности переменных после объединения категорий. Гипотеза формулируется следующим образом: переменные «категория сотрудника» и «удовлетворенность работой» независимы.

Вычисление теоретических частот уже проведено. Подставляем теоретические и эмпирические частоты в формулу и получим:

$$\chi^2 = \frac{(6{,}3-6)^2}{6{,}3} + \frac{(10{,}7-11)^2}{10{,}7} + \frac{(6{,}7-7)^2}{6{,}7} + \frac{(11{,}3-11)^2}{11{,}3} = 0{,}043.$$

Таким образом, $\chi^2_{выб} = 0{,}043$. Находим число степеней свободы. В нашем случае $\eta = (2-1)(2-1) = 1$. В качестве уровня значимости возьмем классическое значение 0,05. Соответствующее значение $\chi^2_{табл} = 3{,}84$. Сравним значения $\chi^2_{выб}$ и $\chi^2_{табл}$. Получается $\chi^2_{выб} < \chi^2_{табл}$, следовательно, гипотезу принимаем. Таким образом, можно сделать вывод, что рассматриваемые переменные независимы.

Для данных, измеренных в шкале отношений, для проверки гипотезы о совпадении характеристик двух групп целесообразно использование либо параметрического *критерия Крамера-Уэлча*, либо непараметрического *критерия Манна-Уитни*. Критерий Крамера-Уэлча предназначен для проверки гипотезы о равенстве средних (строго говоря – математических ожиданий) двух выборок. Критерий Крамера-Уэлча является более эффективным «заменителем» такого известного

критерия как *t*-критерий (критерий Стьюдента). Критерий Манна-Уитни является более «тонким» (но и более трудоемким) – он позволяет проверять гипотезу о том, что две выборки «одинаковы» (в том числе, что совпадают их средние, дисперсии и все другие показатели). Критерий Манна-Уитни плохо применим в условиях, когда число отличающихся друг от друга значений в выборках мало. Две выборки могут иметь одинаковые средние (то есть, критерий Крамера-Уэлча установит совпадение средних), но различаться, например, разбросом. Значит, те различия, которые не выявит критерий Крамера-Уэлча, могут быть выявлены критерием Манна-Уитни.

## 8.2. Критерий Крамера-Уэлча

Эмпирическое значение данного критерия рассчитывается на основании информации об объемах $n_1$ и $n_2$ двух выборок, выборочных средних $\overline{x_1}$ и $\overline{x_2}$ и выборочных дисперсиях $\sigma_1^2$ и $\sigma_2^2$ сравниваемых выборок по формуле:

$$T_{эмп} = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}.$$

Алгоритм определения достоверности совпадений или различий характеристик сравниваемых выборок с помощью критерия Крамера-Уэлча заключается в следующем:

1. Вычислить для сравниваемых выборок $T_{эмп}$ – эмпирическое значение критерия Крамера-Уэлча по приведенной выше формуле.

2. Сравнить это значение с критическим значением $T_{крит}$, если оно больше экспериментального, то сделать вывод: «характеристики сравниваемых выборок совпадают на уровне значимости 0,05»; если же $T_{эмп} > T_{крит}$, то сделать вывод «достоверность различий характеристик сравниваемых выборок составляет 95%».

Пример. В рамках исследования сравнивался уровень развития 4-летних и 5-летних детей. В качестве одного из показателей оценивалось время, затраченное ребенком на выполнение несложного задания. Были составлены две выборки, в одну были включены 4-летние, в другую – 5-летние дети, объемы выборок составили соответственно 20 и 40 человек. 4-летние дети тратили на выполнение задания в среднем 12 мин., 5-летние – 11,5 мин.

Из материалов предыдущих исследований известно, что генеральную дисперсию в обоих случаях можно считать равной единице. Следует проверить гипотезу о равенстве генеральных средних.

*Решение*. Запишем гипотезу: $H_0: \mu_1 = \mu_2$.

Нам известны генеральные дисперсии, поэтому используем формулу критерия Крамера-Уэлча.

Получим: $T_{эксп} = \dfrac{12 - 11{,}5}{\sqrt{\dfrac{1}{20} + \dfrac{1}{40}}} = 1{,}85$.

В расчетной формуле критерия Крамера-Уэлча по умолчанию принимается гипотеза о стандартном нормальном распределении, то есть используем таблицу критических значений для стандартизированной величины $z_{табл}$. Следовательно, нужно

посмотреть, насколько типично для этого распределения полученное значение $T_{эксп}$. Таким образом, мы должны сравнить $T_{эксп}$ с табличным значением для нормального распределения.

В качестве уровня значимости возьмем $\alpha = 0{,}05$. Здесь возникает некоторая особенность. Стандартное нормальное распределение симметрично относительно нуля, соответственно нетипичным значением может быть как слишком большое положительное, так и слишком большое по модулю отрицательное число. Иначе следует «отловить» нетипичные значения в двух полуплоскостях – положительной и отрицательной. Поэтому критическая область, соответствующая уровню значимости $\alpha$, делится на две части, и необходимо рассматривать табличное значение, соответствующее $\alpha/2$. Таким образом, типичными мы будем считать те значения, которые попадают в интервал $(-z_{табл}; z_{табл})$, т. е. не превосходят табличное значение по модулю.

*Вывод*. Для нашей задачи $T_{эксп} = 1{,}85$; $z_{табл} = 1{,}96$. Получаем, что $|z_{выб}| < |z_{табл}|$. Следовательно, нулевую гипотезу следует принять, генеральные средние равны.

При решении вопроса о наличии различий между выборками проводят проверку статистических гипотез о принадлежности обеих выборок одной генеральной совокупности или о равенстве средних.

В случае, когда вид распределения или функция распределения нам известны, задачу оценки различий двух групп независимых наблюдений можно решить с использованием параметрических критериев: критерия Стьюдента (если сравниваются средние значения выборок); критерия Фишера (если сравниваются дисперсии выборок).

## 8.3. Критерий Стьюдента

Критерий Стьюдента ($t$) был разработан Уильямом Госсетом для оценки качества пива в компании Гиннесс. В связи с обязательствами перед компанией по неразглашению коммерческой тайны (руководство Гиннесса считало таковой использование статистического аппарата в своей работе), статья Госсета вышла в 1908 году в журнале «Биометрика» под псевдонимом «Student» (Студент). Закон распределения случайной величины применяется для проверки гипотезы об отличии среднего значения $\bar{x}$ от некоторого известного значения $m$ по формуле:

$$t = \frac{\left|\bar{x} - m\right|}{S / \sqrt{n}}.$$

Количество степеней свободы рассчитывается как $\eta = n - 1$.

Закон Стьюдента положил начало созданию теории «малой выборки». При большом объеме выборки особенность распределения в генеральной совокупности не имеет значения, так как распределение отклонений выборочного показателя от генеральной характеристики при большой выборке всегда оказывается нормальным. В выборках небольшого объема ($n < 30$) на распределении ошибок выборки будет сказываться характер распределения генеральной совокупности. Распределение Стьюдента зависит от двух величин: значения $t$ и числа степеней свободы $\eta$. С увеличением $n$, т. е. числа наблюдений, это распределение быстро приближается к стандартизированному нормальному (с параметрами $\alpha = 0$ и $\delta = 1$). Уже при $n \geq 30$ распределение Стьюдента мало отли-

чается от стандартизированного нормального распределения. Для практического использования распределения Стьюдента существуют специальные таблицы, в которых содержатся критические значения *t* для разных уровней значимости $\alpha$ и чисел степеней свободы $\eta$.

*Замечание*. Если предварительная гипотеза о нормальности распределения попарных разностей окажется отвергнутой, то критерий Стьюдента применять не следует. В таких случаях нужно использовать непараметрические критерии.

*Замечание*. Критерий Стьюдента можно применять также и тогда, когда сравниваются не средние величины выборок, а их относительные частоты.

## 8.4. Критерий Фишера

Пусть имеются две выборки из нормальных генеральных совокупностей. Если сравниваются дисперсии выборок Р.Э. Фишер предложил рассматривать разность их натуральных логарифмов. Позже Д. Снедекор заменил разность логарифмов отношением выборочных дисперсий:

$$F_{эмп} = \frac{\sigma_1^2}{\sigma_2^2},$$

где $\sigma_1^2$ – большая дисперсия;

$\sigma_2^2$ – меньшая дисперсия рассматриваемых вариационных рядов.

Если вычисленное значение критерия $F_{эмп}$ больше критического для определенного уровня значимости и соответствующих чисел степеней свободы для числителя и знаменателя, то дисперсии считаются различными. Иными словами, проверяется

гипотеза, состоящая в том, что генеральные дисперсии рассматриваемых совокупностей равны между собой: $H_0 = \{D_x = D_y\}$.

Критическое значение критерия Фишера ($F$) следует определять по специальной таблице, исходя из уровня значимости $\alpha$ и степеней свободы числителя ($n_1 - 1$) и знаменателя ($n_2 - 1$).

Пример. Дисперсия такого показателя, как стрессоустойчивость для учителей составила 6,17 ($n_1 = 32$), а для менеджеров 4,41 ($n_2 = 33$). Можно ли считать уровень дисперсий примерно одинаковым для данных выборок на уровне значимости 0,05.

Для ответа на поставленный вопрос определим эмпирическое значение критерия:

$$F_{эмп} = \frac{6,17}{4,41} \approx 1,4 .$$

При этом критическое значение критерия $F_{кр}(0,05;31;32) = 2$.

Таким образом, $F_{эмп} = 1,4 < 2 = F_{кр}$, поэтому нулевая гипотеза о равенстве генеральных дисперсий на уровне значимости 0,05 принимается. Следует отметить, что критерий Фишера весьма чувствителен к отклонениям от нормальности изучаемого признака в рассматриваемых выборках.

Проверка гипотез о равенстве числовых характеристик генеральных совокупностей – это, по сути, методы проверки гипотез об однородности выборок. Параметрические критерии обладают большей мощностью. По этой причине, в случаях, когда выборки имеют нормальное распределение, нужно отдавать предпочтение именно параметрическим критериям.

# Раздел 9.
# Непараметрические критерии

## 9.1. Основные понятия

*Непараметрические критерии* проверки гипотез основаны на операциях не с параметрами, а с другими данными, в частности, частотами, рангами и т. п. Непараметрические критерии не позволяют осуществить прямую оценку уровня таких важных параметров, как среднее или дисперсия, с их помощью невозможно оценить взаимодействий действие двух и более условий или факторов, влияющих на изменение признака. Однако эти критерии позволяют решить многие важные задачи, которые сопровождают исследования в гуманитарных областях: выявление различий в уровне исследуемого признака, оценка сдвига значений исследуемого признака, выявление различий в распределениях.

При анализе статистических данных, во-первых, не всегда изучаемый признак подчиняется нормальному закону распределения, во-вторых, часто о законе распределения мало что известно. В этих случаях и применяют непараметрические критерии. Одна из возможных формулировок проверяемой ими гипотезы об однородности выборок – это предположение о совпадении законов распределения, описывающих разные выборки. При ограниченном объеме статистического материала возможным путем повышения достоверности является объединение имеющихся выборок в одну совокупность. Но для этого

следует убедиться в правомерности таких действий, т. е. доказать однородность выборок.

Наиболее популярные виды непараметрических статистических критериев при сравнении выборок:

*для зависимых распределений*: критерий Вилкоксона; критерий знаков;

*для независимых распределений*: критерий Манна-Уитни; критерий Колмогорова-Смирнова; угловое преобразование Фишера.

Рассмотрим алгоритмы применения некоторых непараметрических критериев.

Доказательство достоверности изменений (сдвигов) в значениях исследуемого признака в результате действия каких-либо факторов можно осуществить с помощью критериев изменений. Сдвиг – разность между вторым и первым замерами. Исторически *критерий Вилкоксона* был одним из первых критериев, основанный на рангах. Он применяется для сопоставления показателей, измеренных в двух разных условиях на одной и той же выборке испытуемых, позволяет установить не только направленность изменений, но и их интенсивность. Фактически проверяется гипотеза об однородности двух выборок. Сдвиги значений изучаемой величины происходят по разным причинам.

Сопоставление показателей, полученных у одних и тех же испытуемых по одним и тем же методикам, но в разное время дает *временной сдвиг*. Если показатели получены в разных условиях (покоя, стресса), то их сопоставление определяет *ситуационный сдвиг*. Изменение условий необязательно должно быть реальным. Можно попросить испытуемого «представить себе», что он оказался в позиции других людей, в будущем и т. д., в

таком случае получаем *умозрительный сдвиг*. Многократные исследования одних и тех же лиц на протяжении длительного отрезка их жизни, измеряемого иногда десятками лет, позволяют установить *лонгитюдинальные* (*продольные*) *сдвиги*. В общем случае можно создать специальные условия эксперимента, предположительно оказывающие влияние на те или иные показатели, и сравнить данные до и после экспериментального воздействия. Во всех отмеченных случаях можно говорить о *сдвигах под влиянием* контролируемых или неконтролируемых воздействий. Однако следует учитывать, что выводы будут ограничены, поскольку полученные результаты не проверены на контрольной группе, в которой проводились параллельные измерения. При отсутствии контрольной группы можно констатировать, что сдвиг произошел, но необязательно под воздействием изучаемого фактора.

## 9.2. Критерий Вилкоксона

Для использования критерия Т-Вилкоксона сдвиги между первым и вторым измерениями следует упорядочить. Критерий можно применять и в тех случаях, когда сдвиги принимают всего три значения: -1; 0; +1, хотя в этом случае результаты наверняка совпадут с выводами, полученными по критерию знаков. Если сдвиги существенны, то имеет смысл их ранжировать и оценивать суммы рангов. Суть метода состоит в сопоставлении выраженности сдвигов в каждом из направлений по абсолютной величине. Если сдвиги в ту или иную сторону случайны, то суммы рангов абсолютных значений будут примерно равны.

Гипотезы:

$H_0$ – интенсивность сдвигов в типичном направлении не превосходит интенсивности сдвигов в нетипичном направлении;

$H_1$ – интенсивность сдвигов в типичном направлении превышает интенсивности сдвигов в нетипичном направлении.

Ограничения применения критерия Вилкоксона: минимальное количество испытуемых, прошедших измерения в двух условиях, – 5 человек, максимальное – 50. Нулевые сдвиги из рассмотрения исключаются, и количество наблюдений уменьшается на это значение.

Данный критерий является одним из исключений общего правила. Зона значимости располагается слева, в стороне более низких значений, чем меньше нетипичных сдвигов, тем интенсивнее типичный сдвиг. Зона незначимости располагается справа, в стороне более высоких значений и характеризует однородность выборок. Если известно, что одна из выборок представляет характеристики объектов, подвергшихся какому-либо воздействию (обработке), то их однородность может свидетельствовать об отсутствии эффекта обработки. После проведения необходимых расчетов эмпирического значения критерия при $T_{эмп} < T_{крит}$ нулевая гипотеза отвергается, а принимается $H_1$.

Пример. Группе школьников младших классов был предложен стандартный тест на проверку скорости чтения. Затем со школьниками провели специальный курс занятий, после которого вновь предложили тест. Порядок проведения эксперимента позволяет предположить, что полученные данные на одном испытуемом независимы от аналогичных данных для остальных. В

ходе эксперимента измерялась скорость чтения в знаках каждого учащегося, до и после занятий. Ее значения приведены в табл. 9.1. Можно ли сказать, что проведенные занятия эффективны?

**Таблица 9.1. Расчет критерия Т-Вилкоксона**

| **Код ученика** | **Скорость чтения до занятий** | **Скорость чтения после занятий** $K_{после}$ | **Сдвиги после - до** | **Абсолютное значение сдвигов** | **Ранг сдвигов** |
|---|---|---|---|---|---|
| 1 | 181 | 181 | 0 | 0 | Исключаем |
| 2 | 194 | 104 | -90 | 90 | 12 |
| 3 | 173 | 209 | 36 | 36 | 10 |
| 4 | 153 | 183 | 30 | 30 | 8 |
| 5 | 168 | 180 | 12 | 12 | 3 |
| 6 | 176 | 168 | -8 | 8 | 1 |
| 7 | 163 | 215 | 52 | 52 | 11 |
| 8 | 152 | 172 | 20 | 20 | 6 |
| 9 | 155 | 155 | 0 | 0 | Исключаем |
| 10 | 191 | 156 | -35 | 35 | 9 |
| 11 | 178 | 197 | 19 | 19 | 4,5 |
| 12 | 160 | 183 | 23 | 23 | 7 |
| 13 | 164 | 174 | 10 | 10 | 2 |
| 14 | 195 | 176 | -19 | 19 | 4,5 |
| Сумма | | | | | 78 |

Из данных таблицы видно, что два испытуемых показали нулевой сдвиг, значит их значения следует исключить из расчетов и уменьшить количество испытуемых на это значение (в примере – 2). Типичным будем считать сдвиг в положитель-

ном направлении, а нетипичным – в отрицательном (в таблице выделены курсивом).

Гипотезы:

$H_0$ – интенсивность сдвигов в сторону увеличения скорости чтения после воздействия не превосходит интенсивности сдвигов до обучающего воздействия;

$H_1$ – интенсивность сдвигов в сторону увеличения скорости чтения после воздействия превосходит интенсивность сдвигов до обучающего воздействия.

При ранжировании значений абсолютных сдвигов выполняются правила:

меньшему значению присваивается меньший ранг;

одинаковые значения получают ранг, равный среднему арифметическому из тех, которые были бы присвоены в случае различных значений;

сумма рангов должна совпадать с расчетной суммой:

$$\sum R_i = \frac{n(n+1)}{2} = \frac{12 \cdot 13}{2} = 78.$$

Отметим сдвиги, которые являются нетипичными. В примере они отрицательные. Рассчитаем их сумму рангов, что и составит эмпирическое значение критерия Т-Вилкоксона:

$$T_{эмп} = \sum R_{(-)} = 12 + 1 + 9 + 4{,}5 = 26{,}5.$$

По таблице приложения для $n = 12$ и $\alpha = 0{,}05$ определяем критические значения $T_{крит} = 17$. Эмпирическое значение попало в зону незначимости различий, то есть $T_{эмп} > T_{крит}$, следовательно, принимается нулевая гипотеза и нельзя говорить о присутствии эффекта обучения. Достоверность полученного результата подтверждают и суммарные значения положительных и отрица-

тельных сдвигов, соответственно 202 и 152, что означает примерно одинаковую их интенсивность, а значит и слабый обучающий эффект.

## 9.3. Критерий знаков

Критерий знаков $G$ – один из простейших непараметрических критериев, с помощью которого проверяется сложная непараметрическая нулевая гипотеза о том, что две выборки извлечены из одной и той же генеральной совокупности. Данный критерий предназначен для установления общего направления сдвига исследуемого признака. Он позволяет установить, в какую сторону в выборке в целом изменяются значения признака при переходе от первого измерения к следующему. Критерий знаков применим как к сдвигам, которые определяются лишь качественно (изменение положительного восприятие чего-либо на отрицательное), так и к сдвигам, измеренным количественно (увеличение затрат на отдых или образование). Суть критерия знаков состоит в том, что он определяет, не слишком ли много наблюдается нетипичных сдвигов, чтобы сдвиг в типичном направлении считать преобладающим? Следовательно, $G_{эмп}$ – это количество нетипичных сдвигов. Чем меньше его значение, тем более вероятно, что сдвиг в типичном направлении статистически достоверен.

Гипотезы:

$H_0$ – преобладание типичного направления сдвига является случайным;

$H_1$ – преобладание типичного направления сдвига не является случайным.

Ограничения критерия знаков: количество наблюдений в обоих замерах не должно быть меньше 5 и больше 300. Данный критерий – обратный, он так же является одним из исключений из общего правила.

Преимущество этого критерия – в отсутствии ограничений относительно вида функции распределения, кроме ее непрерывности. Критерий знаков обычно применяется как критерий «спаренных» наблюдений. Обычно сравнивают результаты двух выборок одинакового объема. Если $r_{эмп.} > r_{крит.}$, то считается, что нет оснований для отклонения нулевой гипотезы о том, что две выборки извлечены из одной и той же генеральной совокупности. Если $r_{эмп.} \le r_{крит.}$, то нулевая гипотеза отклоняется, т. е. считают, что две выборки извлечены из генеральных совокупностей с различными функциями распределения.

Пример. Пусть известны результаты двух выборок: 3, 2, 2, 3, 3, 3, 5, 3, 3, 3, 3, 3, 3, 3, 4, 3, 3, 5, 3, 3, 4, 4 и 5, 3, 3, 3, 3, 3, 2, 3, 3, 3, 4, 4, 4, 3, 4, 3, 3, 3, 4, 4, 4, 3.

Проверить выполнение нулевой гипотезы о том, что две выборки извлечены из генеральных совокупностей одной функции распределения.

*Решение*.

Сведем данные в таблицу и определим знаки разностей спаренных результатов обеих выборок.

| | | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1-я выборка | 3 | 2 | 2 | 3 | 3 | 3 | 5 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 4 | 3 | 3 | 5 | 3 | 3 | 4 | 4 |
| 2-я выборка | 5 | 3 | 3 | 3 | 3 | 3 | 2 | 3 | 3 | 3 | 4 | 4 | 4 | 3 | 4 | 3 | 3 | 3 | 4 | 4 | 4 | 3 |
| Знаки разностей | – | – | – | | | | + | | | | – | – | – | | | | | + | – | – | | + |

Из таблицы видно меньше положительных знаков «+». Их число $G_{эмп.} = 3$. По таблице приложения определяем критическое значение для $n = 11$ и $\alpha = 0{,}05$, $G_{крит} = 2$. Получаем $r_{набл.} > r_{крит.}$, поэтому оснований для отклонения нулевой гипотезы нет.

## 9.4. Критерий Манна-Уитни

Критерий *U*-Манна-Уитни предназначен для оценки различий между двумя выборками по уровню какого-либо количественно измеренного признака. Он позволяет выявлять различия между небольшими выборками, начиная с трех измерений. Фактически происходит проверка гипотезы об однородности двух выборок. Иногда эту гипотезу называют гипотезой об отсутствии эффекта обработки. Другими словами, определяется эффект некоторого внешнего воздействия на испытуемых одной группы и сравнение их характеристик с контрольной группой.

Этот метод определяет, достаточно ли мала перекрещивающаяся зона значений между двумя группами. Чем меньше область перекрытия, тем более вероятно, что различия достоверны. Эмпирическое значение критерия *U*-Манна-Уитни фактически отражает величину зоны совпадения между группами. Поэтому, чем меньше эмпирическое значение критерия, тем более вероятно, что различия достоверны, то есть критерий *U*-Манна-Уитни также является исключением из общего правила, то есть обратный.

Гипотезы:

$H_0$ – уровень признака во второй выборке не ниже уровня признака в первой выборке;

$H_1$ – уровень признака во второй выборке ниже уровня признака в первой выборке.

Ограничения применения критерия незначительны: допускается, чтобы в одной из выборок было всего два наблюдения, но тогда в другой их должно быть не менее пяти; максимальный размер выборок не должен превосходить 60 наблюдений, хотя уже при длине выборки более 20 значений ранжирование становится достаточно трудоемким.

При подсчете эмпирического значения критерия *U*-Манна-Уитни проводится процедура ранжирования измеренных значений по правилам:

меньшему значению присваивается меньший ранг; наибольшему значению начисляется ранг, соответствующий общему количеству ранжируемых значений;

в случае равенства нескольких измеренных значений, им начисляется ранг, равный среднему арифметическому тех рангов, которые они получили бы, если бы не были равны;

ранжирование проводится для всех измеренных значений, как если бы это была одна большая выборка, принадлежность отдельных значений к каждой группе отмечается, например, разным цветом.

Общая сумма рангов должна совпадать с расчетной, которая определяется по формуле

$$\sum R_i = \frac{n(n+1)}{2},$$

аналогично критерию Вилкоксона.

По завершению ранжирования следует подсчитать суммы рангов по каждой группе и проверить их совпадение с расчетной суммой.

Определим наибольшую из ранговых сумм и вычислим эмпирическое значение критерия по формуле:

$$U = (n_1 \cdot n_2) + \frac{n_б \cdot (n_б + 1)}{2} - T_б .$$

где $n_1$, $n_2$ – количество испытуемых в выборках 1 и 2;

$n_б$ – количество испытуемых в группе с большей суммой рангов;

$T_б$ – значение большей из двух ранговых сумм.

Отметим что, чем меньше значение критерия, тем достоверность различий выше.

## 9.5. Критерий согласия Колмогорова

Критерий согласия $\lambda$ Колмогорова применяется для проверки простых гипотез о законах распределения только непрерывных случайных величин. Его отличие от критерия согласия $\chi^2$ Пирсона состоит в том, что при применении критерия $\chi^2$ сравнивались эмпирические и теоретические постоянные распределения, а при применении критерия $\lambda$ сравниваются эмпирическая и теоретическая функции распределения $\overline{F}(x)$ и $F(x)$. Кроме того, при применении критерия Колмогорова считаются известными теоретические значения параметров предполагаемой (теоретической) функции распределения, а в критерии согласия $\chi^2$ они определяются по данным выборки. Эти ограничения несколько сужают область практического применения критерия Колмогорова. Пусть выдвинута простая непараметри-

ческая нулевая гипотеза $H_0$ о том, что исследуемая случайная величина $X$ имеет непрерывную функцию распределения $F(x)$. Пусть $x_1, x_2, \ldots, x_n$ – выборка объема $n$ $(n \geq 50)$. Требуется проверить гипотезу $H_0$.

Алгоритм проверки критерия согласия Колмогорова.

1. Располагаем результаты наблюдения в возрастающем порядке или представляем их в виде интервального статистического ряда.

2. Находим эмпирическую функцию распределения

$$\overline{F}(x) = \frac{n_x}{x}.$$

3. Вычисляем значения теоретической функции распределения $F(x)$, соответствующие значениям случайной величины *X*.

4. Для каждого значения $X$ (интервала) находим $\left|\overline{F}(x) - F(x)\right|$.

5. Вычисляем эмпирическое значение выборочной статистики $\lambda$ Колмогорова $\lambda_{эмп.} = \sqrt{n} \cdot \max_x \left|\overline{F}(x) - F(x)\right|$.

6. Задаем уровень значимости и сравниваем $\lambda_{эмп.}$ с $\lambda_{крит.}$, используя таблицу значений $\lambda_{крит.}$:

| Уровень значимости $\alpha$ | 0,20 | 0,10 | 0,05 | 0,02 | 0,01 | 0,001 |
|---|---|---|---|---|---|---|
| $\lambda_{крит.}$ | 1,073 | 1,224 | 1,358 | 1,520 | 1,627 | 10 |

Заключение о правдоподобии гипотезы выносится так же, как и при использовании критерия Пирсона: если $\lambda_{эмп.} \geq \lambda_{крит.}$, гипотеза $H_0$ отвергается, в противном случае нет оснований для отклонения нулевой гипотезы и гипотеза признается правдоподобной.

## 9.6. Критерий Колмогорова–Смирнова

Если проведены две случайные выборки из одной и той же генеральной совокупности объема $n_1$ и $n_2$ $(n_{1,2} \geq 50)$, то, используется вариация $\lambda$-критерия – критерий Колмогорова–Смирнова, можно проверить нулевую гипотезу $H_0$. Заметим, что в методике Пирсона сопоставляются частоты нескольких распределений отдельно по каждому разряду, то в методе расчета критерия Колмогорова-Смирнова сопоставляем сначала по первому разряду, затем по сумме первого и второго, потом по сумме первого, второго и третьего и т. д. Таким образом, всякий раз сопоставляются частоты, накопленные к данному разряду.

Критерий Колмогорова-Смирнова предназначен для сопоставления либо двух эмпирических распределений, либо эмпирического с теоретическим: равномерным или нормальным. Если различия между распределениями существенны, то на некотором шаге вычислений разность накопленных частот достигнет наибольшего значения. В этом случае, сравнивая полученную разность накопленных частот с критическим значением, можно установить их статистическую достоверность. Чем больше эмпирическое значение критерия, тем более существенны различия.

Проверка основывается на вычислении выборочной статистики

$$\lambda : \lambda_{эмп.} = \sqrt{n} \cdot \max_x \left| \overline{F}_1(x) - \overline{F}_2(x) \right|,$$

где $n = \dfrac{n_1 \cdot n_2}{n_1 + n_2}$, $\overline{F}_1(x)$ и $\overline{F}_2(x)$ – эмпирические функции распределения, построенные по данным первой и второй выборки.

При $\lambda_{эмп.} \geq \lambda_{крит.}$ гипотеза $H_0$ отвергается, при $\lambda_{эмп.} < \lambda_{крит.}$ нет оснований для отклонения нулевой гипотезы $H_0$.

Гипотезы:

$H_0$ – различия между сопоставляемыми распределениями недостоверны;

$H_1$ – различия между сопоставляемыми распределениями достоверны.

Ограничения критерия Колмогорова-Смирнова:

выборка должна быть достаточно большой $> 50$ единиц, особенно это касается сопоставления эмпирических распределений. При сравнении с теоретическим распределением величина выборки допускается не менее 5;

разряды следует упорядочить по возрастанию или убыванию признака, поэтому данные, полученные в виде номинальных переменных, обработке этим критерием не подлежат. Например, при сопоставлении категорий национальность или специализация в профессии невозможно говорить о накопленных частотах по разрядам, поскольку нет однозначного однонаправленного изменения признака. В случаях таких переменных применяют критерий Пирсона.

Пример. Проведено социально-психологическое исследование стереотипов мужественности в двух выборках из женщин в возрасте от 22 до 49 лет с высшим образованием и без высшего образования. Испытуемым предлагались четыре карточки с описанием стереотипов мужественности: мифологический, национальный, современный, религиозный.

В таблице приведены эмпирические частоты попадания современного типа на каждую из четырех позиций.

| Разряды-позиции для стереотипа «современный» | 1 | 2 | 3 | 4 | Сумма |
|---|---|---|---|---|---|
| Группа 1 «высшее образование» | 25 | 15 | 13 | 8 | 61 |
| Группа 2 «без высшего» | 12 | 13 | 16 | 9 | 50 |

По процедуре исследования нужно было выбрать среди карточек одну, на которой представлен тип, в большей степени соответствующий представлению каждой женщины об идеальном мужчине, затем из оставшихся трех карточек опять выбрать одну и затем еще одну из двух оставшихся. Групповое предпочтение было явно за типом «современный». Различаются ли распределения предпочтений современного типа, выявленных по каждому из четырех типов между собой?

Гипотеза: $H_0$ – распределения выбора современного типа мужчин в двух группах женщин не различаются между собой;

Для расчета критического значения критерия Колмогорова-Смирнова сопоставление накопленных частот будем проводить по каждому разряду – позиции. Результаты вычислений сведем в таблицу:

| Позиции стереотипа «современный» | Эмпирические частоты | | Эмпирические частности | | Накопленные эмпирические частности | | Разность (абсолютная) |
|---|---|---|---|---|---|---|---|
| | $f_1$ | $f_2$ | $f_1^*$ | $f_2^*$ | $\sum f_1^*$ | $\sum f_2^*$ | $d = \left\lvert \sum f_1^* - \sum f_2^* \right\rvert$ |
| 1 | 25 | 12 | 0,409 | 0,24 | 0,409 | 0,24 | **0,169** |
| 2 | 15 | 13 | 0,246 | 0,26 | 0,655 | 0,5 | 0,155 |
| 3 | 13 | 16 | 0,213 | 0,32 | 0,868 | 0,82 | 0,042 |
| 4 | 8 | 9 | 0,132 | 0,18 | 1 | 1 | 0 |
| Сумма | 61 | 50 | 1,00 | 1,00 | | | |

Эмпирические частности рассчитываем по формуле

$$f_i^* = \frac{f_i}{n_i},$$

где $f_i$ – эмпирическая частота в данном разряде; $n_i$ – количество наблюдений.

Накопленные эмпирические частности находим по формуле

$$\sum f_j^* = \sum f_{j-1}^* + f_j^*,$$

где $j$ – порядковый номер разряда. В последнем столбце таблицы находим наибольшую абсолютную разность и обозначаем ее $d_{max}$. Эмпирическое значение критерия $\lambda_{эмп}$ рассчитываем по формуле

$$\lambda_{эмп} = d_{\max}\sqrt{\frac{n_1 \cdot n_2}{n_1 + n_2}}.$$

В рассматриваемом примере максимальная разность между накопленными эмпирическими частностями составляет 0,169 и попадает в первый разряд. Вычисляем эмпирическое значение по полученным данным:

$$\lambda_{эмп} = 0{,}169\sqrt{\frac{61 \cdot 50}{61 + 50}} = 0{,}88.$$

Сравним полученное значение с критическим из таблицы приложений, при этом длина выборки считается как объеденная, т. е. $n = 101$ и $\alpha = 0{,}05$, $\lambda_{крит} = 1{,}36$.

*Вывод*. Принимается гипотеза $H_0$ – распределения выбора современного типа мужчин в двух группах женщин не различаются между собой.

## 9.7. Критерий Фишера (углового преобразования)

Для достижения максимально точного результата критерий Колмогорова-Смирнова применяют в сочетании с *критерием* $\varphi^*$ *-Фишера* (*критерий углового преобразования*). Если сопоставляются выборки по количественно измеренным показателям, можно выявить такую точку распределения, которая может использоваться как критическая при разделе всех испытуемых на две группы есть эффект – нет эффекта. В таком случае имеет смысл применить угловое преобразование Фишера.

Этот метод относится к группе многофункциональных критериев, которые построены на сопоставлении частей, выраженных в долях единицы или процентах. Суть таких критериев состоит в определении того, какая доля наблюдений (реакций, выборов, испытуемых) в данной выборке характеризуется исследуемым эффектом, а какая часть этим эффектом не характеризуется.

Таким эффектом можно считать:

определенное *значение* качественно определяемого признака, например: согласие с каким-либо предложением; отношение к одному из полов и др.;

определенный *уровень* количественно измеренного признака, например: получение оценки, превосходящей проходной балл, выбор дистанции в разговоре более 1 метра и др.;

определенное *соотношение* значений или уровней исследуемого признака, например: более частый выбор альтернатив А и Б по сравнению с альтернативами С и Д, преимущественное проявление крайних значений признака высоких или низких.

Другими словами, путем сведения любых данных к альтернативной шкале *есть эффект – нет эффекта* критерий Фишера позволяет решать три перечисленные выше задачи, если обследованы две выборки испытуемых. Критерий оценивает достоверность различий между процентными долями двух выборок, в которых зарегистрирован изучаемый эффект.

Кратко суть углового преобразования Фишера состоит в переводе процентных долей в величины центрального угла, который измеряется в радианах. Большей процентной доле будет соответствовать больший угол $\varphi$ , но заметим, что это соотношение не линейное, а определяется формулой $\varphi = 2\arcsin\sqrt{P}$ , где $P$ – процентная доля, выраженная в долях единицы.

Чем больше величина $\varphi^*$ , тем более вероятно, что различия достоверны.

Гипотезы:

$H_0$ – доля лиц, у которых проявляется изучаемый эффект, в первой выборке не больше чем во второй;

$H_1$ – доля лиц, у которых проявляется изучаемый эффект, в первой выборке не больше чем во второй.

*Ограничения* критерия Фишера: ни одна из сопоставляемых долей не может быть равной нулю; верхний предел выборок отсутствует, они могут быть сколь угодно большими; нижний предел – 2 наблюдения в одной из выборок. Однако при этом между размерами обеих выборок должны выполняться следующие соотношения:

$$n_1 = 2 \longrightarrow n_2 > 30;\ n_1 = 3 \longrightarrow n_2 \geq 7;$$
$$n_1 = 3 \longrightarrow n_2 \geq 7;$$

$n_1, n_2 \geq 5$ – *любые размеры выборок*. Если эти соотношения не соблюдаются, достоверность различий установить не удается.

Пример. В совместном исследовании российских и английских ученых в Великобритании проводился опрос врачей общей практики двух категорий: а) имеющие собственный бюджет, б) полностью обеспечивающиеся государственным бюджетом. Каждый врач должен был спрогнозировать результаты медицинской реформы, то есть предположить, какова будет доля врачей с собственным бюджетом в предстоящем году? Различаются ли прогнозы врачей с собственным бюджетом и врачей на государственном обеспечении?

Для вычислений эмпирического значения критерия составим четырехпольную таблицу по исходным данным:

| **Группа** | **Есть эффект 41% - 100%** | **Нет эффекта 0% - 40%** | **Всего** |
|---|---|---|---|
| Врачи с собственным бюджетом | 26 (57,8%) | 19 (42,2%) | 45 |
| Врачи с государственным бюджетом | 9 (36,0%) | 16 (64,0%) | 25 |
| Суммы | 35 | 35 | 70 |

Гипотезы:

$H_0$ – доля лиц, прогнозирующих медицинскую реформу на 41% - 100% всех врачебных приемных, в группе врачей с собственным бюджетом не больше, чем во второй группе врачей;

По таблице приложения определяем величины $\varphi_1$ и $\varphi_2$, напомним, что угол $\varphi_1$ – это всегда угол, соответствующий большей процентной доле. Получены значения: $\varphi_1(57\%) = 1{,}727$ ; $\varphi_2(36\%) = 1{,}287$ . Рассчитаем эмпирическое значение критерия $\varphi^*$ Фишера по формуле

$$\varphi^*_{эмп} = (\varphi_1 - \varphi_2)\sqrt{\frac{n_1 \cdot n_2}{n_1 + n_2}}.$$

Получим

$$\varphi^*_{эмп} = (1{,}727 - 1.287)\sqrt{\frac{45 \cdot 25}{45 + 25}} = 1{,}764,\ \varphi^*_{крит} = 1{,}64.$$

*Вывод*. $H_0$ – отвергается ( $p = 0{,}039$ ). Доля лиц, прогнозирующих развитие медицинской реформы на 41% - 100% всех приемных, в группе врачей, имеющих собственный бюджет, превышает долю в группе врачей, не взявших фонда и оставшихся на государственном обеспечении.

Совместное применение двух критериев позволяет гарантировать, что выявленный уровень статистических различий – максимально возможный для этих реальных данных.

# Раздел 10.
# Методы многомерного анализа

## 10.1. Дисперсионный анализ

В предыдущих разделах обсуждались методы, позволяющие изучать проблемы, в которых изменчивость была представлена одной переменной. Измерение сразу нескольких признаков (свойств объектов) в одном эксперименте более естественно, чем измерение одного. Поэтому многомерный статистический анализ имеет более широкое поле применения. К тому же с формальной точки зрения, одномерный анализ представляет частный случай многомерного. К сожалению, построение теории для многомерных статистических данных – дело не простое. На сегодняшний день достаточно хорошо разработана лишь теория для гауссовских (имеющих нормальное распределение) данных. При анализе многомерных данных используются методы, не имеющие четкой статистической трактовки в смысле рассмотренной ранее концепции проверки гипотез, построения доверительных интервалов и т. д. Ограничимся в данном разделе пояснениями наиболее популярных методов, особенно тех, что нашли распространение в статистических пакетах.

*Дисперсионный анализ* представляет собой систему понятий и технических приемов, позволяющих обобщить процедуру сравнения двух средних для двух выборок, взятых из генеральных совокупностей с нормальным распределением, на случай большого числа выборок. Этот статистический метод предна-

значен для решения задачи об одновременном равенстве средних в $l$ выборках, каждая объемом $n$. Дисперсионный анализ, основанный на сравнении дисперсий, является одним из методов статистической обработки наблюдений и служит для оценки влияния на наблюдаемую величину различных факторных признаков, т. е. признаков, от которых зависит наблюдаемая величина. Первоначально дисперсионный анализ был предложен английским статистиком Р. Фишером (1925 г.) для обработки результатов агрономических опытов по выявлению условий, при которых испытываемый сорт сельскохозяйственной культуры дает максимальный урожай.

На практике дисперсионный анализ применяют, чтобы установить, оказывает ли существенное влияние некоторый качественный фактор $F$ (*однофакторный анализ*), имеющий $p$ уровней $F_1, F_2, \ldots, F_p$, на изучаемую величину $X$. Например, если требуется выяснить, какой вид удобрений наиболее эффективен для получения наибольшего урожая, то фактор $F$ – удобрение, а его уровни – виды удобрений. Суть дисперсионного анализа заключается в расчленении общей дисперсии признака на компоненты согласно влиянию конкретных факторов, в сравнении *факторной дисперсии*, порождаемой воздействием фактора, и *остаточной дисперсии*, обусловленной случайными причинами, проверке гипотез о значимости их влияния. Для этого совокупность разбивается на группы, различающиеся по уровню факторов. Если различие между факторной и остаточной дисперсиями значимо, то фактор оказывает существенное влияние на $X$. В этом случае средние наблюдаемых значений на каждом уровне значимо различают-

ся, что определяется в соответствии с законом распределения вероятностей случайных ошибок.

Объектом дисперсионного анализа являются средние квадраты (несмещенные оценки дисперсий), получающиеся делением сумм квадратов отклонений на соответствующее число степеней свободы. Число степеней свободы определяется как общее число наблюдений минус число связывающих их уравнений. Дисперсионный анализ, позволяющий лишь обнаружить наличие систематических расхождений, непригоден для их численной оценки с последующим исключением из результатов наблюдений. Эта цель может быть достигнута только при многократных измерениях.

Дисперсионный анализ применяется для установления однородности нескольких совокупностей. Исследуют воздействие нескольких факторов на нескольких постоянных или случайных уровнях и выясняют влияние отдельных уровней и их комбинаций (многофакторный анализ). Итак, на количественный нормально распределенный признак $X$ воздействует фактор $F$, который имеет $p$ уровней. Пусть число наблюдений на каждом уровне одинаково и равно $q$. Пусть наблюдалось $n = p \cdot q$ значений $x_{ij}$ признака $X$, где $i$ – номер испытания ($i = 1, 2, ..., q$), $j$ – номер уровня фактора ($j = 1, 2, ..., p$). Рассматривается нулевая гипотеза $H_0$ о равенстве групповых средних. Другими словами, гипотеза $H_0$ утверждает равенство факторной и остаточной дисперсий. Альтернативная гипотеза $H_1$ отрицает утверждение гипотезы $H_0$ (некоторые средние могут быть различными). По критерию $F$ Фишера для проверки нулевой гипотезы $H_0$ достаточно проверить нулевую гипотезу о равенстве факторной и остаточной дисперсий.

Схема метода дисперсионного анализа – это схема проверки нулевой гипотезы $H_0$.

Шаг 1. По данным испытаний вычисляют групповые средние $\bar{x}_{гр.j}$, $j = 1, 2, ..., p$, например

$$\bar{x}_{гр.1} = \frac{x_{11} + x_{21} + ... + x_{q1}}{q}.$$

Шаг 2. По данным испытаний вычисляют общую среднюю:

$$\bar{x}_0 = \frac{1}{pq} \cdot \sum_{i=1}^{q} \sum_{j=1}^{p} x_{ij};$$

$$\overline{x_0} = \frac{1}{pq} \cdot (x_{11} + x_{12} + ... + x_{1p} + x_{21} + x_{22} + ... + x_{2p} + ... + x_{q1} + x_{q2} + ... + x_{qp}).$$

Шаг 3. Находят общую сумму квадратов отклонений наблюдаемых значений $x_{ij}$ от общей средней $\bar{x}_0$:

$$S_{общ.} = \sum_{i=1}^{q} \sum_{j=1}^{p} (x_{ij} - \bar{x}_0)^2 .$$

Можно заметить, что $S_{общ.} = \overline{\sigma^2} \cdot n$.

Шаг 4. Находят факторную сумму квадратов отклонений групповых средних от общей средней. Факторная сумма характеризует рассеяние «между группами»:

$$S_{факт.} = q \cdot \sum_{j=1}^{p} (\bar{x}_{гр.j} - \bar{x}_0)^2 .$$

Справедливо соотношение: $S_{факт.} = n \cdot \delta^2$, где $\delta^2$ – межгрупповая дисперсия.

Шаг 5. Находят остаточную сумму квадратов отклонений наблюдаемых значений группы от своей групповой средней. Она характеризует рассеяние «внутри групп»: $S_{ост.} = S_{общ.} - S_{факт.}$. Легко заметить: $S_{ост.} = \tilde{\sigma}^2 \cdot n$, где $\tilde{\sigma}^2$ –

средняя арифметическая групповых дисперсий. $S_{общ.}$ характеризует влияние фактора и случайных причин, $S_{факт.}$ – воздействие фактора $F$; $S_{ост.}$ отражает влияние случайных причин.

Шаг 6. Находят общую, факторную и остаточную дисперсии:

$$S^2_{общ.} = \frac{S_{общ.}}{p \cdot q - 1},\ S^2_{факт.} = \frac{S_{факт.}}{p - 1},\ S^2_{ост.} = \frac{S_{ост.}}{p \cdot (q - 1)},$$

где $p$ – число уровней фактора; $q$ – число наблюдений на каждом уровне.

Шаг 7. Находят наблюдаемое значение фактора $F_{набл.}$ – отношение большей из найденных дисперсий к меньшей:

$$F_{набл.} = \frac{S^2_{факт.}}{S^2_{ост.}}, \text{ если } S^2_{факт.} > S^2_{ост.} \text{ и } F_{набл.} = \frac{S^2_{ост.}}{S^2_{факт.}},$$

$$\text{если } S^2_{ост.} > S^2_{факт.}.$$

Шаг 8. Число степеней свободы для $S^2_{факт.}$ равно $k_1 = p - 1$, число степеней свободы для $S^2_{ост.}$ $k_2 = p \cdot q - 1$. Учитывая число степеней свободы и уровень значимости $\alpha$, на котором рассматривается гипотеза $H_0$, по таблице критических точек распределения $F$ Фишера находят критическую точку: $F_{крит.}(\alpha; k_1; k_2)$.

Шаг 9. Вычисленное $F_{набл.}$ сравнивается с $F_{крит.}$.

Если $F_{набл.} > F_{крит.}$, то различие средних значимое и нулевая гипотеза $H_0$ о равенстве групповых средних отвергается.

Если $F_{набл.} < F_{крит.}$ – различие групповых средних незначимое и нет оснований для отказа от гипотезы $H_0$.

*Замечание 1*. Если $S^2_{ост.} > S^2_{факт.}$, то гипотеза $H_0$ справедлива.

*Замечание 2*. Критерий $F$ Фишера является равномерно наиболее мощным несмещенным критерием для проверки нулевой гипотезы $H_0$.

*Замечание 3*. Оценка доли влияния факторов и их сочетаний на изучаемый (результативный) признак производится с помощью корреляционных отношений

$$\eta_1^2 = \frac{S_{факт.}^2}{S_{общ.}^2} \text{ и } \eta_2^2 = \frac{S_{факт.}^2}{S_{общ.}^2}.$$

Чем ближе значение корреляционного отношения $\eta_1^2$ к единице, тем влияние соответствующего признака (фактора) больше.

*Замечание 4*. Если наблюдаемое значение $x_{ij}$ увеличивается (уменьшается) в одно и то же число раз $k$, то дисперсия увеличивается (уменьшается) в $k^2$ раз, но отношение дисперсий ($F_{набл.}/\eta_i$) не изменится.

*Замечание 5*. При увеличении (уменьшении) наблюдаемых значений $x_{ij}$ на одно и то же число $c$ дисперсия не меняется.

*Замечание 6*. Параметры $x_{ij}$ могут принимать как дискретные, так и непрерывные значения.

Пример. Указано число студентов в 4-х группах первого курса, не посещавших лекции по высшей математике в течение февраля. Проверьте выполнение нулевой гипотезы $H_0$ о равенстве групповых средних:

| № | Группы | | | | |
|---|---|---|---|---|---|
| | **1** | **2** | **3** | **4** | $\Sigma$ |
| 1 | 1 | 2 | 4 | 2 | 9 |
| 2 | 2 | 2 | 4 | 2 | 10 |
| 3 | 3 | 3 | 8 | 5 | 20 |
| 4 | 1 | 3 | 8 | 3 | 15 |
| 5 | 3 | 3 | 5 | 2 | 13 |
| 6 | 2 | 2 | 7 | 4 | 15 |
| $\Sigma$ | 12 | 15 | 36 | 18 | 81 |

*Решение*.

По условию $p = 4$, $q = 6$, поэтому $pq = 24$, $k_1 = p - 1 = 4 - 1 = 3$, $k_2 = p \cdot q - 1 = 24 - 1 = 23$. Используя выше приведенные соотношения, вычислим групповые и общую средние:

$$\overline{x}_{гр.1} = \frac{12}{6} = 2,\ \overline{x}_{гр.2} = \frac{15}{6} = 2{,}5,$$

$$\overline{x}_{гр.3} = \frac{36}{6} = 6,\ \overline{x}_{гр.4} = \frac{18}{6} = 3,\ \overline{x}_0 = \frac{81}{24} = 3{,}375.$$

Найдем теперь общую, факторную и остаточную суммы:

$$S_{общ.} \approx 89{,}63,\ S_{факт.} \approx 58{,}13,$$

$$S_{ост.} = S_{общ.} - S_{факт.} = 89{,}63 - 58{,}13 = 31{,}5.$$

Определим факторную и остаточную дисперсии:

$$S^2_{факт.} = \frac{58{,}13}{3} \approx 19{,}38,\ S^2_{ост.} = \frac{31{,}5}{20} \approx 1{,}58.$$

Для проверки нулевой гипотезы воспользуемся критерием $F$ Фишера, используя полученные величины, вычислим наблюдаемое значение критерия $F$ Фишера:

$$F_{набл.} = \frac{19{,}38}{1{,}57} \approx 12{,}34.$$

По таблице критических значений находим $F_{крит.}(0{,}05;3;23) = 3{,}028$.

*Вывод*. $F_{набл.} > F_{крит.}$, поэтому различие средних значимое и нулевая гипотеза $H_0$ о равенстве групповых средних *отвергается*.

Тождество $S_{ост.} = S_{общ.} - S_{факт.}$ является основным в дисперсионном анализе. Из него следует, что варьирование всех ре-

зультатов наблюдений около общего среднего может быть разложено на суммы квадратов, первая из которых характеризует варьирование, обусловленное изменчивостью эффектов различных уровней факторов, а вторая – варьирование под влиянием неучтенных факторов (ошибка эксперимента). Проблемы значимости результатов дисперсионного анализа основаны на предпосылке о нормальности исходных данных и равенстве (однородности) дисперсий. Нарушение этих предположений отражается на уровне статистической значимости, а именно:

если объемы выборок и их дисперсии не равны, а из совокупностей с большими дисперсиями выбирается меньшее число объектов, вероятность ошибки первого рода увеличивается;

при тех же условиях, если из совокупностей с большими дисперсиями берется большее число объектов, то вероятность ошибки первого рода уменьшается;

если объемы выборок равны, влиянием неоднородности дисперсий на уровень значимости F - критерия можно пренебречь;

влияние нарушения нормальности на номинальный уровень значимости F- критерия незначительно.

Таким образом, процедура дисперсионного анализа достаточно устойчива к нарушению предпосылок, лежащих в ее основе.

## 10.2. Кластерный анализ

При анализе многомерных данных часто возникает задача разбиения исходного множества на некоторые подмножества так, чтобы, с одной стороны, каждый объект наблюдения принадлежал только к одному подмножеству, с другой – объекты,

составляющие одно подмножество, были максимально сходными, а входящие в разные подмножества были существенно различными. Такие подмножества называют *кластерами*. Задача классификации решается методами *кластерного анализа* (от англ. cluster – гроздь). Суть этого метода кластерного анализа в следующем: вводится некая единая мера, охватывающая все измеряемые показатели; реализуется алгоритм чисто количественного решения вопроса о разбиении на подмножества.

Пусть в нашем распоряжении есть 11 объектов, у которых измеряется одна характеристика, то есть имеет место одномерный случай. Результаты измерений в таблице:

| Объекты | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Результаты измерений | 8 | 4 | 2 | 2 | 4 | 8 | 2 | 6 | 4 | 8 | 2 |

Рассчитаем сумму квадратов отклонений:

$$\sum_{i=1}^{11}(x_i - \bar{x})^2 = 64{,}27 .$$

Если теперь все эмпирическое множество данных разбить на 4 подмножества: $A_1 = \{8,8,8\}$, $A_2 = \{4,4,4\}$, $A_3 = \{2,2,2,2\}$, $A_4 = \{6\}$, то все внутриподмножественные суммы квадратов отклонений будут равны 0. В приведенном случае разбиение было очевидным и естественным. В общем случае все несколько сложнее.

Пусть каждый $k$-й объект характеризуется вектором измерений $x(k)$, имеющим длину $p$. Тогда его можно представить как точку в $p$-мерном пространстве. Пара объектов $O_k$ и $O_i$ будет попадать в один кластер, если расстояние между ними будет мало. Введем понятие расстояния между точками в $p$-мерном пространстве.

Евклидово расстояние между точками определяется формулой:

$$d(x_l, x_k) = \sqrt{\sum_{i=1}^{p} (x_{li} - x_{ki})^2}$$ , где $x_{li}$ и $x_{ki}$ – координаты векторов.

Расстояние Махаланобиса:

$$D^2(x_l, x_k) = (x_l - x_k)^T W^{-1} (x_l - x_k),$$

где $W^{-1}$ матрица, обратная матрице полной суммы квадратов и произведений (матрица рассеяния).

Примеры наиболее успешного применения кластерного анализа относятся к тем случаям, когда имеющаяся у экспериментатора предварительная информация позволяет заранее определить число кластеров.

Пример. Проведем кластерный анализ для эмпирических данных о политических предпочтениях респондентов. В качестве меры расстояния использовалось расстояние Махаланобиса.

Результаты представим в таблице.

| **Политические предпочтения** | **Кластер** | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** |
| Демократы | 49 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Либералы | 0 | 14 | 13 | 17 | 3 | 3 | 0 | 0 | 0 |
| Консерваторы | 0 | 7 | 5 | 6 | 18 | 0 | 3 | 5 | 6 |

Из таблицы видно, что демократические взгляды практически полностью попали в кластер 1, кластеры 2, 3, 4 и 6 можно отнести к либеральным и, наконец, кластеры 5, 7, 8, 9 – к консервативным. Каждый кластер характеризуется своим вектором средних и матрицей рассеяния. Заметим, что в одни и те же кластеры попали разные политические предпочтения, то есть, нет четкого разделения на три кластера, какое хотелось бы полу-

чить, поскольку заведомо известно, что данные принадлежат трем совокупностям.

Это может быть связано с невозможностью однозначно отнести к демократическим или консервативным 13 случаев из ста. С другой стороны, и без количественного анализа специалист может различить эти разновидности. При этом он, безусловно, использует качественные признаки, которые в количественном анализе не были учтены. Таким образом, возникает проблема выбора информативных признаков. Кроме того, результаты кластеризации зависят от выбранного метода, и эта зависимость тем сильнее, чем менее явно изучаемая совокупность разделяется на группы объектов.

Заметим также, что методы кластерного анализа не дают способа проверки статистической гипотезы об адекватности полученных классификаций, но часто служат подспорьем для содержательного анализа.

## 10.3. Факторный анализ

В *факторном анализе* речь идет о выделении из множества измеряемых характеристик объекта факторов, более адекватно отражающих свойства объектов. Множество показателей сложных объектов взаимосвязаны между собой и часто дублируют друг друга. Нахождение и оценка ненаблюдаемых переменных – факторов и является основной задачей факторного анализа.

Предположим, имеются две группы вопросов-задач (наблюдаемых переменных), требующих от отвечающего на них человека способностей соответственно к логическому мышле-

нию и к художественному воображению. Подсчитав корреляции между нашими вопросами, мы, вероятно, придем к выводу, что результаты ответов на вопросы каждой из этих групп коррелируют друг с другом. Человек, получивший высокую оценку по одному из «логических» вопросов, наверное, получит такую же оценку и по второму, и по третьему. Человек, проявивший высокие способности при решении одной из задач «на воображение», вероятно, не менее успешно решит и другие задачи подобного рода. То же будет иметь место и для низких оценок (напомним, что наличие корреляции между двумя признаками, грубо говоря, означает, что с ростом значений одного признака растут либо убывают значения другого).

Для объяснения описанных корреляций можно выдвинуть гипотезу, состоящую в том, что имеются два латентных фактора, которые условно можно назвать «логические способности» и «художественное воображение», принимающие разные значения у разных людей. И корреляции между нашими наблюдаемыми переменными объясняются действием именно этих латентных факторов: человек с высоким уровнем логических способностей будет, как правило, хорошо отвечать на вопросы первой группы, с низким уровнем таких способностей – плохо. Аналогичное утверждение будет справедливо и для второго фактора.

Таким образом, исходя из сформулированной гипотезы, полагаем, что все наблюдаемые нами изменения значений эмпирических признаков обусловлены изменением некоторых внутренних свойств этих объектов – значений латентных факторов. Предполагается, что совокупность этих факторов едина для всех наблюдаемых признаков. Такие факторы назовем общими. Из-

мерить их непосредственно мы не можем. Более того, мы не знаем заранее в точности, что из себя эти факторы представляют, сколько их. Однако предполагаем, что в принципе они существуют и что респонденты могут быть сопоставлены друг с другом по их значениям этих свойств (подчеркнем, что сказанным утверждается существование латентных переменных; напомним, что для социолога подобные утверждения далеко не всегда очевидны). Общие факторы имеют разное влияние на изменение того или иного наблюдаемого признака. Вес общего фактора, определяющий степень его влияния на изменение данного наблюдаемого признака, называют факторной нагрузкой фактора на признак.

Естественно предположить, что кроме тех изменений наблюдаемых признаков, которые вызваны действием общих латентных факторов, существуют индивидуальные изменения каждого наблюдаемого признака, вызываемые, например, случайными ошибками при их измерении. Причины, вызывающие невзаимосвязанные изменения исходных признаков, называются *специфическими* или *характерными факторами*. Таким образом, все причины изменений наблюдаемых признаков могут быть разделены на две составляющие: группу общих факторов и специфический фактор для каждого признака.

Итак, значения общих латентных факторов для какого-либо человека определяют ответы этого человека на рассматриваемые вопросы, или, как мы будем говорить, поведение этого человека. Именно действием указанных латентных факторов определяются все корреляции между нашими наблюдаемыми переменными. Это означает, что фиксация значений латентных

переменных должна привести к ликвидации связи между наблюдаемыми признаками. Другими словами, если мы зафиксируем, скажем, значение фактора «логические способности», то связи между отвечающими этому фактору наблюдаемыми переменными исчезнут. Возьмем только тех респондентов, которые имеют блестящие логические способности. Конечно, они в основном будут хорошо отвечать на наши логические тесты. Но могут встретиться и плохие ответы: скажем, кто-то из них слишком переволновался и забыл какую-то элементарную формулу, знание которой предполагалось тестом, и т. д. Однако связи между ответами на разные логические тесты уже не будет, поскольку сбои в ответах будут определяться не логическими способностями респондентов, а случайными по отношению к таким способностям обстоятельствами.

Факторный анализ родился в психологии как способ поиска латентных факторов, подобных описанным. Собственно, факторный анализ – это просто четкое выражение идей тестового подхода (рождение факторного анализа является хорошей иллюстрацией роли математического языка в развитии всякой науки). В этой связи целесообразно заметить, что, хотя факторный анализ – статистический метод и, в принципе, не может доказать наличие или отсутствие каких бы то ни было причинно-следственных отношений. Тем не менее при его использовании есть основания полагать, что латентная переменная олицетворяет собой причину, обусловливающую тот или иной уровень относящихся к ней наблюдаемых характеристик (хотя в практических задачах далеко не всегда бывает очевидным, что является причиной, что – следствием).

Соотношения между факторами и набором исходных измеряемых показателей могут быть найдены в виде матрицы факторных нагрузок $F$, имеющей размерность ( $p \times m$ ), где $p$ – число показателей; $m$ – число факторов. Основой для построения матрицы $F$ служит матрица парных коэффициентов корреляции $R$ размерностью ( $p \times p$ ). Факторная матрица характеризует степень связи между $m$ факторами и каждым из $p$ измеряемых показателей. При этом число факторов выбирается исходя из двух требований: оно должно быть много меньше числа показателей, а потери информации при этом должны быть минимально возможными. Таким образом, в результате факторного анализа выявляется группа показателей, наиболее тесно связанных с каждым из факторов. Следовательно, появляется возможность сравнивать между собой отдельные факторы, давать им содержательную интерпретацию и наименование.

Основная модель факторного анализа записывается в виде

$$x_i = \sum_{k=1}^{m} \omega_{ik} f_k + d_i u_i \text{ , где } i = 1, 2, ..., p;\ m << p.$$

Здесь $f_k$ – $k$-й общий фактор; $m$ – заданное число общих факторов; $\omega_{ik}$ – нагрузка $i$-го показателя на $k$-й общий фактор; $u_i$ – характерный фактор; $d_i$ – нагрузка на $i$-й характерный фактор.

Общие факторы учитывают корреляцию между измеряемыми показателями и являются стандартизированными величинами, характерный фактор учитывает оставшуюся дисперсию. Несложно показать, что квадраты факторных нагрузок показывают доли дисперсии измеряемого показателя, приходящегося на соответствующие факторы. Сформулируем фундаментальную факторную теорему.

**Теорема**. Если измеряемые показатели $x_1$, $x_2$ в основе своей имеют общий фактор, то для $m$ факторов имеет место равенство: $r_{x_1,x_2} = \omega_{11}\omega_{21} + \omega_{12}\omega_{22} + \omega_{1m}\omega_{2m}$.

Если измеряемых показателей $p$, то матрица факторных нагрузок $F$ будет иметь размерность $(p \times m)$, получим матричную форму теоремы $R = F \cdot F^T$.

Пример. Приведена матрица коэффициентов корреляции между восемью морфологическими признаками, вычисленными по выборке, состоящей из 305 девушек (данные Г. Хартмана). Заметим, что в силу симметричности матрицы, заполнена только одна ее половина. С помощью центроидного метода находим нагрузки факторов на отдельные измеряемые показатели. В результате можно выделить два фактора, обеспечивающие статистически значимое соответствие получаемого результата с исходным для анализа материалом: стройность и полнота.

**Корреляционная матрица**

| **Измеряемый показатель** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** |
|---|---|---|---|---|---|---|---|---|
| Рост | 1 | | | | | | | |
| Размах рук | 0,846 | 1 | | | | | | |
| Длина предплечья | 0,805 | 0,881 | 1 | | | | | |
| Длина ноги | 0,859 | 0,826 | 0,801 | 1 | | | | |
| Вес | 0,473 | 0,376 | 0,380 | 0,436 | 1 | | | |
| Окружность бедер | 0,389 | 0,326 | 0,319 | 0,329 | 0,762 | 1 | | |
| Окружность груди | 0,301 | 0,277 | 0,237 | 0,327 | 0,730 | 0,583 | 1 | |
| Ширина груди | 0,382 | 0,415 | 0,345 | 0,465 | 0,629 | 0,577 | 0,539 | 1 |

В связи с необходимостью придать полученному решению вид, удобный для содержательной интерпретации, было проведено «вращение» факторов и составлена результирующая матрица:

| Измеряемый показатель | Исходное решение Факторные нагрузки | | Финальное решение Факторные нагрузки | |
|---|---|---|---|---|
| | Фактор А | Фактор В | Фактор А | Фактор В |
| Рост | 0,830 | -0,396 | 0,879 | 0,272 |
| Размах рук | 0,818 | -0,469 | 0,919 | 0,210 |
| Длина предплечья | 0,777 | -0,470 | 0,890 | 0,182 |
| Длина ноги | 0,798 | -0,401 | 0,858 | 0,246 |
| Вес | 0,786 | 0,500 | 0,238 | 0,900 |
| Окружность бедер | 0,672 | 0,458 | 0,183 | 0,792 |
| Окружность груди | 0,594 | 0,444 | 0,135 | 0,729 |
| Ширина груди | 0,647 | 0,333 | 0,250 | 0,684 |

Анализ этой таблицы показывает, что уже исходное решение позволяет сделать предположение, что фактор *B*, имеющий в двух подгруппах измеряемых показателей нагрузки с разными знаками, может быть интерпретирован как полнота. Сделать какие-либо выводы по исходному решению относительно фактора *A* трудно. Поэтому используется поворот осей в плоскости полученных двух факторов такой, чтобы нагрузка на измеряемые переменные для каждого фактора приближалась к 0 или 1. В результате такой процедуры получаем финальное решение, которое убеждает нас в правильности предварительного вывода о том, что фактор *B* может быть интерпретирован как полнота. Фактор *A* подтверждает наименование «стройность», так как максимальные факторные нагрузки приходятся на первые четыре показателя, имеющие прямое отношение к комплексной оценке стройности. В частности, данные таблицы показывают,

что из восьми морфологических признаков важнейшим по вкладу в фактор «стройность» является размах рук, и только после него идут рост, длина предплечья и, наконец, длина ноги.

На этом простом примере видно, какие трудности возникают при попытке дать содержательную интерпретацию и наименование факторам, выявленным в процессе анализа.

История применения факторного анализа в социологии очень показательна. Математические методы начали широко использоваться советскими исследователями практически с самого начала возрождения отечественной социологии в 60-х годах. И факторный анализ сразу стал популярным. Было получено много результатов, как содержательных, так и методических, касающихся совершенствования аппарата факторного анализа применительно к специфике социологических задач, разработки приемов его использования в комплексе с другими методами. Считалось, что факторный анализ может способствовать успешному решению практически любой социологической задачи. Потом энтузиазм резко уменьшился. Начались разговоры о том, что этот метод не приспособлен для решения социологических задач. Из одной крайности преувеличения возможностей метода исследователи перешли в другую крайность – почти полное отрицание его полезности для социологии.

Упомянутые крайности были возможны по одной причине: из-за отсутствия внимания исследователя к анализу той модели, которая заложена в методе. Пока эта модель адекватна реальности, его использование полезно. Но как только метод начинает применять исследователь, не дающий себе отчета в том, что за формализмом стоит некоторая модель (и в силу этого не обеспе-

чивающий адекватности этой модели), применение метода перестает приносить пользу. Более того, оно зачастую становится вредным.

Назовем основные причины, снижающие эффективность применения факторного анализа в социологии.

Во-первых, факторный анализ рассчитан на количественные данные.

Во-вторых, социолог зачастую не имеет заранее, в своем сознании никаких гипотез, связанных с основной сутью модели факторного анализа. Поясним.

Основным элементом модели, заложенной в факторный анализ, является априорное предположение о наличии латентных факторов, стоящих за наблюдаемыми переменными, объясняющих связи между последними. Это предположение не означает, что количество и сущность факторов заранее точно определены. Предварительная гипотеза в процессе факторного анализа данных может быть скорректирована и даже вообще отвергнута. Анкета зачастую составляется из соображений, не имеющих никакого отношения к такому предположению. И только на этапе анализа данных приходит мысль использовать факторный анализ. Естественно, что в таком случае попытка разумно интерпретировать полученные с помощью факторного анализа результаты кончается крахом – в найденные факторы не удается вложить вменяемый смысл. В таких случаях обычно уровень объяснимой факторами дисперсии бывает малым, факторные нагрузки – низкими.

В-третьих, как уже было не раз отмечено, социолог чаще всего работает не с отдельными респондентами, а с большими

их совокупностями и поэтому не может позволить себе задать респонденту несколько сот вопросов (что, как правило, делает психолог). Из-за этого оказывается невозможным измерение такого количества наблюдаемых признаков, которого было бы достаточно для того, чтобы из них могли быть получены близкие к истине значения латентных факторов. Приведем цитату из работы П. Лазарсфельда относящуюся к латентно-структурному анализу, заметив предварительно, что этот метод по своей сути тождествен факторному анализу. В цитате речь идет о номинальном латентном факторе, и поэтому приписывание респонденту значения латентной переменной отождествляется с отнесением его к одному из латентных классов, с «положением в классификации»: «Показатели индивида по отдельному индикатору (т.е. значения нашей наблюдаемой переменной) могут случайно измениться, но его основное положение в классификации останется неизменным. Или же, наоборот, меняется основное положение, а показатели по каким-то индикаторам случайно остаются теми же. Но если для шкалы или индекса имеется много индикаторов, крайне мало вероятно, чтобы значительное их число случайно изменилось в одном направлении, в то время как изучаемый индивид фактически сохранял бы свое основное положение неизменным».

Описание объектов в терминах факторов по сути дела представляет собой математическую модель взаимосвязей, существующих между исходными параметрами. Эти взаимосвязи могут быть обусловлены самыми разными причинами. В моделях факторного анализа самих по себе, в математических построениях, на которых базируются вычислительные процедуры,

не содержатся представления о причинности. Это представление вносится исследователем при интерпретации.

В эмпирическом исследовании мы постоянно имеем дело с моделями реальности. Особенно остро этот вопрос стоит при использовании математических моделей. Это касается и изучения причин каких-либо явлений на базе анализа статистических связей. Даже разрабатывая анкету специально под факторный анализ, включая в нее довольно большое количество наблюдаемых индикаторов, социологи иногда некорректно ставят задачу. Ситуация переворачивается с ног на голову. Гипотетический латентный фактор (существование которого априори постулируется) в действительности может не являться причиной, обусловливающей изменения наблюдаемых индикаторов; может быть следствием таких изменений, а может и вообще к таким изменениям не иметь отношения. Фиксация его значений в таких случаях может не приводить к исчезновению связей между наблюдаемыми признаками. Исследователь же, не зная об этом и механически применив технику факторного анализа, либо получает очень плохую модель (вследствие того что его гипотеза об адекватности факторной модели не отвечает реальности), либо пытается искать интерпретацию найденного более или менее сносного латентного фактора на неправильном пути, полагая, что этот фактор тождествен той самой несостоятельной латентной переменной. В силу указанных причин интерпретацию результатов факторного анализа иногда имеет смысл расценивать не как финальный этап исследования, а как этап выдвижения гипотез.

В-пятых, интерпретация результатов факторного анализа часто бывает затруднена их принципиальной неоднозначностью.

Множество одинаково «хороших» факторных моделей может быть получено путем ротации некоторого первичного решения. Подчеркнем, что это не расценивается как недостаток метода. Напротив, в этом состоит его достоинство: постановка задачи была обусловлена жизненной ситуацией; и здесь мы снова сталкиваемся с той принципиальной невозможностью однозначно описать социальные явления формальными методами.

Несмотря на все сказанное, тестовая традиция в социологии работает. И в настоящее время успешно используется как сам факторный анализ, так и некоторые такие приемы, которые, будучи близки по своей логике к этому анализу, все же от него отличаются. В первую очередь имеются в виду известные шкалы Лайкерта и Гуттмана. Для того чтобы использование тестовой традиции было корректным, необходимо к тому же убедиться в том, что связи между наблюдаемыми признаками действительно определяются именно латентной переменной.

# Заключение

Изложенный в пособии материал можно успешно применять для изучения эмпирических зависимостей, полученных в результате проводимых экспериментов. Однако следует понимать, что это учебное пособие рассматривает не все вопросы получения и обработки эмпирических данных в социологии.

Методы формализации данных, построения моделей и математической обработки могут оказать исследователю неоценимую помощь, но они лишь средство, которое не должно заслонять собой цель. Достоверная статистическая тенденция – это все же не социальная закономерность, а выпадающие из общей картины индивидуальные значения есть отражение закономерностей более высокого порядка, чем те, что выявлены с помощью математических методов. В процессе подготовки и проведения социологических исследований неизбежно возникнут многие важные вопросы, ответы на которые недостаточно отражены в этой книге. Список литературы, приведенный в конце пособия, является возможностью получения дополнительной информации.

Автор надеется, что это учебное пособие будет полезно всем, кто хочет освоить методы статистической обработки данных, их анализа и применять эти знания в своей образовательной и исследовательской деятельности. Можно предположить, что после прочтения учебного пособия у читателя создалось правильное впечатление о том, насколько рассмотренные в ней вопросы, с одной стороны, сложны, а с другой, – неотделимы от

содержательных исследовательских представлений об изучаемом явлении, объекте. Измерение и обработка информации в социологии не сводится к выбору неких технических процедур. Проблема прикладных методов в социологии носит концептуальный характер. Ее решение определяется исследовательскими парадигмами, содержательными концепциями автора, его пониманием роли человека в изучаемых социальных процессах.

И выбор метода измерения, и интерпретация его результатов зависят от представлений исследователя о том, как респондент воспринимает рассматриваемые аспекты окружающего мира, как изучаемые процессы проявляются в ответах респондента на вопросы исследователя. Более того, успешное решение проблемы требует довольно глубокого освоения социологом смежных дисциплин.

Будущий и практикующий социолог должен внимательно анализировать, какую именно реальность он хочет отразить в математические конструкты (например, числа) с помощью измерения: какие именно соотношения между рассматриваемыми объектами его интересуют, каким образом формируются изучаемые им представления людей, как следует интерпретировать тот или иной ответ респондента, какую роль в интерпретации данных играет их рассмотрение в общем контексте исследования и т. д. Чтобы ответы на подобные вопросы действительно «вплетались» в общую стратегию исследования они должны быть достаточно четкими, должны быть сформулированы на логико-математическом языке.

В области теории социологического измерения много нерешенных вопросов и одним из основных пробелов является от-

сутствие необходимой стыковки между методами сбора данных (сюда входит, например, изучение влияния формулировки вопроса на результат исследования) и собственно методами измерения (например, теория, заложенная в шкале Лайкерта). Изложенный материал можно успешно применять для изучения эмпирических зависимостей, полученных в результате проводимых экспериментов.

Автору хотелось бы, чтобы студенты творчески восприняли главную идею: измерение в прикладной социологии — это моделирование реальности. Все приемы, методы, подходы о которых шла речь (да и не только они, по причине ограниченности объема), могут служить лишь некими «кирпичиками», из которых социологу предстоит построить «дом», для конкретной социологической ситуации.

# Список литературы


1. Айвазян С.А., Мхитарян В.С. Прикладная статистика и основы эконометрики. М.: ЮНИТИ, 1998.
2. Айвазян С.А., Енюков И.С., Мешалкин Л.Д. Прикладная статистика: Основы моделирования и первичная обработка данных. М.: Финансы и статистика, 1989.
3. Андерсен Т.Введение в многомерный статистический анализ. М.: Физматгиз, 1963.
4. Аренс Х., Лейтер Ю. Многомерный дисперсионный анализ. М.: Финансы и статистика, 1985.
5. Батыгин ГС. Обоснование научного вывода в прикладной социологии. М: Наука, 1986.
6. Благуш П. Факторный анализ с обобщениями. М.: Финансы и статистика, 1989.
7. Бородкин Л.И. Многомерный статистический анализ в исторических исследованиях. М.: Изд-во МГУ, 1986.
8. Вентцель Е.С., Овчаров Л.А. Теория вероятностей и ее инженерные приложения. М.: Наука, 1998.
9. Владимирский Б.М., Горстко А.Б., Ерусалимский Я.М. Математика. Общий курс. СПб.: Лань, 2002.
10. Воронов НЭП Методы сбора информации в социологическом исследовании. М.: Статистика, 1974.
11. Гласе Дж., Стэнли Дж. Статистические методы в педагогике и психологии. М.: Прогресс, 1976.
12. Голофаст В.Б. Методологический анализ в социологическом исследовании. Л.: Наука, 1981.

13. Джарол Б.Мангейм, Ричард К. Рич. Политология. Методы исследования. – М.: Весь мир, 1997.
14. Докторов Б.З. О надежности измерения в социологическом исследовании. Л.: Наука, 1979.
15. Дэвид Г. Метод парных сравнений. М.: Статистика, 1978.
16. Дэйвисон М. Многомерное шкалирование. М.: Финансы и статистика, 1988.
17. Жуковская В.М., Мучник И.Б. Факторный анализ в социально-экономических исследованиях. М.: Статистика, 1976.
18. Захаров В.П. Применение математических методов в социально-психологических исследованиях. Л.: ЛГУ, 1985.
19. Кендалл М.Дж., Стьюарт А. Статистические выводы и связи. М.: Наука, 1973.
20. Клигер С.А., Косолапое М.С., Телешова Ю.Н. Шкалирование при сборе и анализе социологической информации. М.: Наука, 1978.
21. Кокс Д., Снелл Э. Прикладная статистика. Принципы и примеры. – М.: Мир, 1984.
22. Колкот Э. Проверка значимости. М.: Статистика, 1978.
23. Лазарсфельд П. Измерение в социологии //Американская социология. М.: Прогресс, 1972.
24. Монсон П. Современная западная социология. Теории, традиции, перспективы. СПб.: Нотабене, 1992.
25. Морозов Е.И. Методология и методы анализа социальных систем. М.: Изд-во МГУ, 1995.
26. Осипов Г.В., Андреев Э.П. Методы измерения в социологии. М.: Наука, 1977.

27. Рунион Р. Справочник по непараметрической статистике. М.: Финансы и статистика, 1982.
28. Саганенко Г.И. Надежность результатов социологического исследования. Л.: Наука, 1983.
29. Сидоренко Е.В. Методы математической обработки в психологии. – СПб.: ООО «Речь», 2000.
30. Толстова Ю.Н. Логика математического анализа социологических данных. М.: Наука, 1991.
31. Тюрин Ю.Н., Литвак Б.Г., Орлов А.И., Сатаров Г.А., Шмерлинг Д.С. Анализ нечисловой информации. Москва, 1981.
32. Тюрин Ю.Н., Макаров А.А. Статистический анализ данных на компью-тере. – М.: ИНФРА, 1998.
33. Уемов А.И. Логические основы метода моделирования. М.: Мысль, 1971.
34. Ферстер Э., Ренц Б. Методы корреляционного и регрессионного анализа. М.: ФиС, 1983.
35. Хованов Н.В. Математические основы теории шкал измерения качества. Л., 1982.
36. Холлендер М., Вулф Д.А. Непараметрические методы статистики. М.: Финансы и статистика, 1983.
37. Чесноков СВ. Детермипационный анализ социально-экономических данных. М.: Наука, 1982.
38. Шерла К. Факторный анализ. М.: Статистика, 1980.
39. Ядов В.А. Социологическое исследование: методология, программа, методы. Саратов: Изд-во Саратовского ун-та, 1995.

# Приложения

## Таблицы критических значений

**Таблица 1. Критические значения критерия Пирсона $\chi^2$ при разном числе степеней свободы *v***

| *p* | | | *p* | | | *p* | | |
|---|---|---|---|---|---|---|---|---|
| ***v*** | **0,05** | **0,01** | ***v*** | **0,05** | **0,01** | ***v*** | **0,05** | **0,01** |
| **1** | 3,841 | 6,635 | **35** | 49,802 | 57,342 | **69** | 89,391 | 99,227 |
| **2** | 5,991 | 9.210 | **36** | 50,998 | 58,619 | **70** | 90,631 | 100,425 |
| **3** | 7,815 | 11,345 | **37** | 52,192 | 59,892 | **71** | 91,670 | 101,621 |
| **4** | 9,488 | 13,277 | **38** | 53,384 | 61,162 | **72** | 92,808 | 102,816 |
| **5** | 11,070 | 15,086 | **39** | 54,572 | 62,428 | **73** | 93,945 | 104,010 |
| **6** | 12,592 | 16,812 | **40** | 55,758 | 63,691 | **74** | 95,081 | 105,202 |
| **7** | 14,067 | 18,475 | **41** | 56,942 | 64,950 | **75** | 96,217 | 106,393 |
| **8** | 15,507 | 20,090 | **42** | 58,124 | 66,206 | **76** | 97,351 | 107,582 |
| **9** | 16,919 | 21,666 | **43** | 59,304 | 67,459 | **77** | 98,484 | 108,771 |
| **10** | 18,307 | 23,209 | **44** | 60,481 | 68,709 | **78** | 99,617 | 109,958 |
| **11** | 19,675 | 24,725 | **45** | 61,656 | 69,957 | **79** | 100,749 | 111,144 |
| **12** | 21,026 | 26,217 | **46** | 62,830 | 71,201 | **80** | 101,879 | 112,329 |
| **13** | 22,362 | 27,688 | **47** | 64,001 | 72,443 | **81** | 103,010 | 113,512 |
| **14** | 23,685 | 29,141 | **48** | 65,171 | 73,683 | **82** | 104,139 | 114,695 |
| **15** | 24,996 | 30,578 | **49** | 66,339 | 74,919 | **83** | 105,267 | 115,876 |
| **16** | 26,296 | 32,000 | **50** | 67,505 | 76,154 | **84** | 106,395 | 117,057 |
| **17** | 27,587 | 33,409 | **51** | 68,669 | 77,386 | **85** | 107,522 | 118,236 |
| **18** | 28,869 | 34,805 | **52** | 69,832 | 78,616 | **86** | 108,648 | 119,414 |
| **19** | 30,144 | 36,191 | **53** | 70,993 | 79.84'3 | **87** | 109,773 | 120,591 |
| **20** | 31,410 | 37,566 | **54** | 72,153 | 81,069 | **88** | 110,898 | 121,767 |
| **21** | 32,671 | 38,932 | **55** | 73,311 | 82,292 | **89** | 112,022 | 122,942 |
| **22** | 33,924 | 40,289 | **56** | 74,468 | 83,513 | **90** | 113,145 | 124,116 |

**Окончание табл. 1**

| *p* | | | *p* | | | *p* | | |
|---|---|---|---|---|---|---|---|---|
| ***v*** | **0,05** | **0,01** | ***v*** | **0,05** | **0,01** | ***v*** | **0,05** | **0,01** |
| **23** | 35,172 | 41,638 | **57** | 75,624 | 84,733 | **91** | 114,268 | 125,289 |
| **24** | 36,415 | 42,980 | **58** | 76,778 | 85,950 | **92** | 115,390 | 126,462 |
| **25** | 37,652 | 44,314 | **59** | 77,931 | 87,166 | **93** | 116,511 | 127,633 |
| **26** | 38,885 | 45.642 | **60** | 79,082 | 88,379 | **94** | 117,632 | 128,803 |
| **27** | 40,113 | 46,963 | **61** | 80,232 | 89,591 | **95** | 118,752 | 129,973 |
| **28** | 41,337 | 48,278 | **62** | 81,381 | 90,802 | **96** | 119,871 | 131,141 |
| **29** | 42,557 | 49,588 | **63** | 82,529 | 92,010 | **97** | 120,990 | 132,309 |
| **30** | 43,773 | 50,892 | **64** | 83,675 | 93,217 | **98** | 122,108 | 133,476 |
| **31** | 44,985 | 52,191 | **65** | 84,821 | 94,422 | **99** | 123,225 | 134,642 |
| **32** | 46,194 | 53,486 | **66** | 85,965 | 95,626 | **100** | 124,342 | 135,807 |
| **33** | 47,400 | 54,776 | **67** | 87,108 | 96,828 | | | |
| **34** | 48,602 | 56,061 | **68** | 88,250 | 98.028 | | | |

**Таблица 2. Критические значения критерия знаков G**

| *n* | *p* | | *n* | *p* | | *n* | *p* | | *n* | *p* | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | **0,05** | **0,01** | | **0,05** | **0,01** | | **0,05** | **0,01** | | **0,05** | **0,01** |
| **5** | 0 | - | **27** | 8 | 7 | **49** | 18 | 15 | **92** | 37 | 34 |
| **6** | 0 | - | **28** | 8 | 7 | **50** | 18 | 16 | **94** | 38 | 35 |
| **7** | 0 | 0 | **29** | 9 | 7 | **52** | 19 | 17 | **96** | 39 | 36 |
| **8** | 1 | 0 | **30** | 10 | 8 | **54** | 20 | 18 | **98** | 40 | 37 |
| **9** | 1 | 0 | **31** | 10 | 8 | **56** | 21 | 18 | **100** | 41 | 37 |
| **10** | 1 | 0 | **32** | 10 | 8 | **58** | 22 | 19 | **110** | 45 | 42 |
| **11** | 2 | 1 | **33** | 11 | 9 | **60** | 23 | 20 | **120** | 50 | 46 |
| **12** | 2 | 1 | **34** | 11 | 9 | **62** | 24 | 21 | **130** | 55 | 51 |
| **13** | 3 | 1 | **35** | 12 | 10 | **64** | 24 | 22 | **140** | 59 | 55 |
| **14** | 3 | 2 | **36** | 12 | 10 | **66** | 25 | 23 | **150** | 64 | 60 |
| **15** | 3 | 2 | **37** | 13 | 10 | **68** | 26 | 23 | **160** | 69 | 64 |
| **16** | 4 | 2 | **38** | 13 | 11 | **70** | 27 | 24 | **170** | 73 | 69 |
| **17** | 4 | 3 | **39** | 13 | 11 | **72** | 28 | 25 | **180** | 78 | 73 |
| **18** | 5 | 3 | **40** | 14 | 12 | **74** | 29 | 26 | **190** | 83 | 78 |
| **19** | 5 | 4 | **41** | 14 | 12 | **76** | 30 | 27 | **200** | 87 | 83 |
| **20** | 5 | 4 | **42** | 15 | 13 | **78** | 31 | 28 | **220** | 97 | 92 |
| **21** | 6 | 4 | **43** | 15 | 13 | **80** | 32 | 29 | **240** | 106 | 101 |
| **22** | 6 | 5 | **44** | 16 | 13 | **82** | 33 | 30 | **260** | 116 | 110 |
| **23** | 7 | 5 | **45** | 16 | 14 | **84** | 33 | 30 | **280** | 125 | 120 |
| **24** | 7 | 5 | **46** | 16 | 14 | **86** | 34 | 31 | **300** | 135 | 129 |
| **25** | 7 | 6 | **47** | 17 | 15 | **88** | 35 | 32 | | | |
| **26** | 8 | 6 | **48** | 17 | 15 | **90** | 36 | 33 | | | |

**Таблица 3. Критические значения критерия *U*-Манна-Уитни для уровня значимости p = 0,05**

| $N_1$ | $N^2$ | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **7** | **8** | **9** | **10** | **11** | **12** | **13** | **14** | **15** | **16** | **17** | **18** |
| **3** | 1 | 2 | 2 | 3 | 3 | 4 | 4 | 5 | 5 | 6 | 6 | 7 |
| **4** | 3 | 4 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 11 | 12 |
| **5** | 5 | 6 | 7 | 8 | 9 | 11 | 12 | 13 | 14 | 15 | 17 | 18 |
| **6** | 6 | 8 | 10 | 11 | 13 | 14 | 16 | 17 | 19 | 21 | 22 | 24 |
| **7** | 8 | 10 | 12 | 14 | 16 | 18 | 20 | 22 | 24 | 26 | 28 | 30 |
| **8** | 10 | 13 | 15 | 17 | 19 | 22 | 24 | 26 | 29 | 31 | 34 | 36 |
| **9** | 12 | 15 | 17 | 20 | 23 | 26 | 28 | 30 | 34 | 37 | 39 | 42 |
| **10** | 14 | 17 | 20 | 23 | 26 | 29 | 33 | 36 | 39 | 42 | 45 | 48 |
| **11** | 16 | 19 | 23 | 26 | 30 | 33 | 37 | 40 | 44 | 48 | 51 | 55 |
| **12** | 18 | 22 | 26 | 29 | 33 | 37 | 41 | 45 | 49 | 53 | 57 | 61 |
| **13** | 20 | 24 | 28 | 33 | 37 | 41 | 45 | 50 | 54 | 59 | 63 | 67 |
| **14** | 22 | 26 | 31 | 36 | 40 | 45 | 50 | 55 | 59 | 64 | 67 | 74 |
| **15** | 24 | 29 | 34 | 39 | 44 | 49 | 54 | 59 | 64 | 70 | 75 | 80 |
| **16** | 26 | 31 | 37 | 42 | 47 | 53 | 59 | 64 | 70 | 75 | 81 | 86 |
| **17** | 28 | 34 | 39 | 45 | 51 | 57 | 63 | 67 | 75 | 81 | 87 | 93 |
| **18** | 30 | 36 | 42 | 48 | 55 | 61 | 67 | 74 | 80 | 86 | 93 | 99 |
| **19** | 32 | 38 | 45 | 52 | 58 | 65 | 72 | 78 | 85 | 92 | 99 | 106 |
| **20** | 34 | 41 | 48 | 55 | 62 | 69 | 76 | 83 | 90 | 98 | 105 | 112 |

**Таблица 4. Критические значения критерия Т-Вилкоксона**

| *n* | *p* | | *n* | *p* | |
|---|---|---|---|---|---|
| | ***0.05*** | ***0.01*** | | ***0.05*** | ***0.01*** |
| **5** | 0 | - | **28** | 130 | 101 |
| **6** | 2 | - | **29** | 140 | 110 |
| **7** | 3 | 0 | **30** | 151 | 120 |
| **8** | 5 | 1 | **31** | 163 | 130 |
| **9** | 8 | 3 | **32** | 175 | 140 |
| **10** | 10 | 5 | **33** | 187 | 151 |
| **11** | 13 | 7 | **34** | 200 | 162 |
| **12** | 17 | 9 | **35** | 213 | 173 |
| **13** | 21 | 12 | **36** | 227 | 185 |
| **14** | 25 | 15 | **37** | 241 | 198 |
| **15** | 30 | 19 | **38** | 256 | 211 |
| **16** | 36 | 23 | **39** | 271 | 224 |
| **17** | 41 | 27 | **40** | 286 | 238 |
| **18** | 47 | 32 | **41** | 302 | 252 |
| **19** | 53 | 37 | **42** | 319 | 266 |
| **20** | 60 | 43 | **43** | 336 | 281 |
| **21** | 67 | 49 | **44** | 353 | 296 |
| **22** | 75 | 55 | **45** | 371 | 312 |
| **23** | 83 | 62 | **46** | 389 | 328 |
| **24** | 91 | 69 | **47** | 407 | 345 |
| **25** | 100 | 76 | **48** | 426 | 362 |
| **26** | 110 | 84 | **49** | 446 | 379 |
| **27** | 119 | 92 | **50** | 466 | 397 |

**Таблица 5. Уровни статистической значимости**

**разных значений критерия $\varphi^*$ Фишера.**

По полученному значению $\varphi^*_{эмп}$ определяется уровень значимости различий процентных долей

| **p равно или меньше** | **p равно или меньше (последний десятичный знак)** | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | **0** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** |
| **0,00** | 2,91 | 2,81 | 2,70 | 2,62 | 2,55 | 2,49 | 2,44 | 2,39 | 2,35 | |
| **0,01** | 2,31 | 2,28 | 2,25 | 2,22 | 2,19 | 2,16 | 2,14 | 2,11 | 2,09 | 2,07 |
| **0,02** | 2,05 | 2,03 | 2,01 | 1,99 | 1,97 | 1,96 | 1,94 | 1,92 | 1,91 | 1,89 |
| **0,03** | 1,88 | 1,86 | 1,85 | 1,84 | 1,82 | 1,81 | 1,80 | 1,79 | 1,77 | 1,76 |
| **0,04** | 1,75 | 1,74 | 1,73 | 1,72 | 1,71 | 1,70 | 1,68 | 1,67 | 1,66 | 1,65 |
| **0,05** | 1,64 | 1,64 | 1,63 | 1,62 | 1,61 | 1,60 | 1,59 | 1,58 | 1,57 | 1,56 |
| **0,06** | 1,56 | 1,55 | 1,54 | 1,53 | 1,52 | 1,52 | 1,51 | 1,50 | 1,49 | 1,48 |
| **0,07** | 1,48 | 1,47 | 1,46 | 1,46 | 1,45 | 1,44 | 1,43 | 1,43 | 1,42 | 1,41 |
| **0,08** | 1,41 | 1,40 | 1,39 | 1,39 | 1,38 | 1,37 | 1,37 | 1,36 | 1,36 | 1,35 |
| **0,09** | 1.34 | 1,34 | 1,33 | 1,32 | 1,32 | 1,31 | 1,31 | 1,30 | 1,30 | 1,29 |
| **0,10** | 1,29 | | | | | | | | | |

**Таблица 6. Критические значения модуля максимального расхождения $d_{max}$ при сопоставлении эмпирического распределения с теоретическим**

| ***n*** | Максимальный модуль разности накопленных частостей $d_{max}$ | | ***n*** | Максимальный модуль разности накопленных частостей $d_{max}$ | |
|---|---|---|---|---|---|
| | **p=0,05** | **p =0,01** | | **p=0,05** | **p=0,01** |
| **5** | 0,6074 | 0,7279 | **50** | 0,1921 | 0,2302 |
| **10** | 0,4295 | 0,5147 | **60** | 0,1753 | 0,2101 |
| **15** | 0,3507 | 0,4202 | **70** | 0,1623 | 0,1945 |
| **20** | 0,3037 | 0,3639 | **80** | 0,1518 | 0,1820 |
| **25** | 0,2716 | 0,3255 | **90** | 0,1432 | |
| **30** | 0,2480 | 0,2972 | **100** | 0,1358 | |
| **40** | 0,2147 | 0,2574 | **>100** | $1{,}36/\sqrt{n}$ | $1{,}63/\sqrt{n}$ |

**Таблица 7. Критерий λ Колмогорова-Смирнова**

**для сопоставления эмпирического распределения с теоретическим**

**(*n* > 50) или двух эмпирических распределений между собой (*n* > 50);**

**уровни статистической значимости разных значений $\lambda_{эмп}$**

| λ | **λ, последний десятичный знак** | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | **0** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** |
| | **p– десятичные знаки («0» опущен)** | | | | | | | | | |
| **0,3** | 99999 | 99998 | 99995 | 99991 | 99983 | 99970 | 99949 | 99917 | 99872 | 99807 |
| **0,4** | 99719 | 99603 | 99452 | 99262 | 99027 | 98741 | 98400 | 97998 | 97532 | 96998 |
| **0,5** | 96394 | 95719 | 94969 | 94147 | 93250 | 92282 | 91242 | 90134 | 88960 | 87724 |
| **0,6** | 86428 | 85077 | 83678 | 82225 | 80732 | 79201 | 77636 | 76042 | 74422 | 72781 |
| **0,7** | 71124 | 69453 | 67774 | 66089 | 64402 | 62717 | 61036 | 59363 | 57700 | 56050 |
| **0,8** | 54414 | 52796 | 51197 | 49619 | 48063 | 46532 | 45026 | 43545 | 42093 | 40668 |
| **0,9** | 39273 | 37907 | 36571 | 35266 | 33992 | 32748 | 31536 | 30356 | 29206 | 28087 |
| **1,0** | 27000 | 25943 | 24917 | 23922 | 22957 | 22021 | 21114 | 20236 | 19387 | 18566 |
| **1.1** | 17772 | 17005 | 16264 | 15550 | 14861 | 14196 | 13556 | 12939 | 12345 | 11774 |
| **1,2** | 11225 | 10697 | 10190 | 09703 | 09235 | 08787 | 08357 | 07944 | 07550 | 07171 |
| **1,3** | 06809 | 06463 | 06132 | 05815 | 05513 | 05224 | 04949 | 04686 | 04435 | 04196 |
| **1,4** | 03968 | 03751 | 03545 | 03348 | 03162 | 02984 | 02815 | 02655 | 02503 | 02359 |
| **1,5** | 02222 | 02092 | 01969 | 01852 | 01742 | 01638 | 01539 | 01446 | 01357 | 01274 |
| **1,6** | 01195 | 01121 | 01051 | 00985 | 00922 | 00864 | 00808 | 00756 | 00707 | 00661 |
| **1,7** | 00618 | 00577 | 00539 | 00503 | 00469 | 00438 | 00408 | 00380 | 00354 | 00330 |
| **1,8** | 00307 | 00285 | 00265 | 00247 | 00229 | 00213 | 00198 | 00186 | 00170 | 00158 |
| **1.9** | 00146 | 00136 | 00126 | 00116 | 00108 | 00100 | 00092 | 00085 | 00079 | 00073 |
| **2,0** | 00067 | 00062 | 00057 | 00053 | 00048 | 00045 | 00041 | 00038 | 00035 | 00032 |
| **2,1** | 00030 | 00027 | 00025 | 00023 | 00021 | 00019 | 00018 | 00016 | 00015 | 00014 |
| **2,2** | 00013 | 00011 | 00010 | 00010 | 00009 | 00008 | 00007 | 00007 | 00006 | 00006 |
| **2,3** | 00005 | 00005 | 00004 | 00004 | 00004 | 00003 | 00003 | 00003 | 00002 | 00002 |
| **2,4** | 00002 | 00002 | 00002 | 00001 | 00001 | 00001 | 00001 | 00001 | 00001 | 00001 |

**Таблица 8. Критические значения**

**выборочного коэффициента корреляции рангов**

| *n* | p | | *n* | p | | *n* | p | |
|---|---|---|---|---|---|---|---|---|
| | **0,05** | **0,01** | | **0,05** | **0,01** | | **0,05** | **0,01** |
| **5** | 0,94 | - | **17** | 0,48 | 0,62 | **29** | 0,37 | 0,48 |
| **6** | 0,85 | - | **18** | 0,47 | 0,60 | **30** | 0,36 | 0,47 |
| **7** | 0,78 | 0,94 | **19** | 0,46 | 0,58 | **31** | 0,36 | 0,46 |
| **8** | 0,72 | 0,88 | **20** | 0,45 | 0,57 | **32** | 0,36 | 0,45 |
| **9** | 0,68 | 0,83 | **21** | 0,44 | 0,56 | **33** | 0,34 | 0,45 |
| **10** | 0,64 | 0,79 | **22** | 0,43 | 0,54 | **34** | 0,34 | 0,44 |
| **11** | 0,61 | 0,76 | **23** | 0,42 | 0,53 | **35** | 0,33 | 0,43 |
| **12** | 0,58 | 0,73 | **24** | 0,41 | 0,52 | **36** | 0,33 | 0,43 |
| **13** | 0,56 | 0,70 | **25** | 0,40 | 0,51 | **37** | 0,33 | 0,43 |
| **14** | 0,54 | 0,68 | **26** | 0,39 | 0,50 | **38** | 0,32 | 0,41 |
| **15** | 0,52 | 0,66 | **27** | 0,38 | 0,49 | **39** | 0,32 | 0,41 |
| **16** | 0,50 | 0,64 | **28** | 0,38 | 0,48 | **40** | 0,31 | 0,40 |

**Таблица 9. Величины угла $\varphi$ (в радианах) для разных процентных долей**

| доля | **%, последний десятичный знак** | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | **0** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** |
| | **Значения** $\varphi = \arcsin\sqrt{P}$ | | | | | | | | | |
| **0,0** | 0,000 | 0,020 | 0,028 | 0,035 | 0,040 | 0,045 | 0,049 | 0,053 | 0,057 | 0,060 |
| **0,1** | 0,063 | 0,066 | 0,069 | 0,072 | 0,075 | 0,077 | 0,080 | 0,082 | 0,085 | 0,087 |
| **0,2** | 0,089 | 0,092 | 0,094 | 0,096 | 0,098 | 0,100 | 0,102 | 0,104 | 0,106 | 0,108 |
| **0,3** | 0,110 | 0,111 | 0,113 | 0,115 | 0,117 | 0,118 | 0,120 | 0,122 | 0,123 | 0,125 |
| **0,4** | 0,127 | 0,128 | 0,130 | 0,131 | 0,133 | 0,134 | 0,136 | 0,137 | 0,139 | 0,140 |
| **0,5** | 0,142 | 0,143 | 0,144 | 0,146 | 0,147 | 0,148 | 0,150 | 0,151 | 0,153 | 0,154 |
| **0,6** | 0,155 | 0,156 | 0,158 | 0,159 | 0,160 | 0,161 | 0,163 | 0,164 | 0,165 | 0,166 |

**Продолжение табл. 9**

| доля | **%, последний десятичный знак** | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | **0** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** |
| | **Значения** $\varphi = \arcsin\sqrt{P}$ | | | | | | | | | |
| **0,7** | 0,168 | 0,169 | 0,170 | 0,171 | 0,172 | 0,173 | 0,175 | 0,176 | 0,177 | 0,178 |
| **0,8** | 0,179 | 0,180 | 0,182 | 0,183 | 0,184 | 0,185 | 0,186 | 0,187 | 0,188 | 0,189 |
| **0,9** | 0,190 | 0,191 | 0,192 | 0,193 | 0,194 | 0,195 | 0,196 | 0,197 | 0,198 | 0,199 |
| **1** | 0,200 | 0,210 | 0,220 | 0,229 | 0,237 | 0,246 | 0,254 | 0,262 | 0,269 | 0,277 |
| **2** | 0,284 | 0,291 | 0,298 | 0,304 | 0,311 | 0,318 | 0,324 | 0,330 | 0,336 | 0,342 |
| **3** | 0,348 | 0,354 | 0,360 | 0,365 | 0,171 | 0,376 | 0,382 | 0,387 | 0,392 | 0,398 |
| **4** | 0,403 | 0,408 | 0,413 | 0,418 | 0,423 | 0,428 | 0,432 | 0,437 | 0,442 | 0,446 |
| **5** | 0,451 | 0,456 | 0,460 | 0,465 | 0,469 | 0,473 | 0,478 | 0,482 | 0,486 | 0,491 |
| **6** | 0,495 | 0,499 | 0,503 | 0,507 | 0,512 | 0,516 | 0,520 | 0,524 | 0,528 | 0,532 |
| **7** | 0,536 | 0,539 | 0,543 | 0,547 | 0,551 | 0,555 | 0,559 | 0,562 | 0.566 | 0,570 |
| **8** | 0,574 | 0,577 | 0,581 | 0,584 | 0,588 | 0,592 | 0,595 | 0,599 | 0,602 | 0,606 |
| **9** | 0,609 | 0,613 | 0,616 | 0,620 | 0,623 | 0,627 | 0,630 | 0,633 | 0,637 | 0,640 |
| **10** | 0,644 | 0,647 | 0,650 | 0,653 | 0,657 | 0,660 | 0,663 | 0,666 | 0,670 | 0,673 |
| **11** | 0,676 | 0,679 | 0,682 | 0,686 | 0,689 | 0,692 | 0,695 | 0,698 | 0,701 | 0,704 |
| **12** | 0,707 | 0,711 | 0,714 | 0,717 | 0,720 | 0,723 | 0,726 | 0,729 | 0,732 | 0,735 |
| **13** | 0,738 | 0,741 | 0,744 | 0,747 | 0,750 | 0,752 | 0,755 | 0,758 | 0,761 | 0,764 |
| **14** | 0,767 | 0,770 | 0,773 | 0,776 | 0,778 | 0,781 | 0,784 | 0,787 | 0,790 | 0,793 |
| **15** | 0,795 | 0,798 | 0,801 | 0,804 | 0,807 | 0,809 | 0,812 | 0,815 | 0,818 | 0,820 |
| **16** | 0,823 | 0,826 | 0,828 | 0,831 | 0,834 | 0,837 | 0,839 | 0,842 | 0,845 | 0,847 |
| **17** | 0,850 | 0,853 | 0,855 | 0,858 | 0,861 | 0,863 | 0,866 | 0,868 | 0,871 | 0,874 |
| **18** | 0,876 | 0,879 | 0,881 | 0,884 | 0,887 | 0,889 | 0,892 | 0,894 | 0,897 | 0,900 |
| **19** | 0,902 | 0,905 | 0,907 | 0,910 | 0,912 | 0,915 | 0,917 | 0,920 | 0,922 | 0,925 |
| **20** | 0,927 | 0,930 | 0,932 | 0,935 | 0,937 | 0,940 | 0,942 | 0,945 | 0,947 | 0,950 |
| **21** | 0,952 | 0,955 | 0,957 | 0,959 | 0,962 | 0,964 | 0,967 | 0,969 | 0,972 | 0,974 |
| **22** | 0,976 | 0,979 | 0,981 | 0,984 | 0,986 | 0,988 | 0,991 | 0,993 | 0,996 | 0,998 |

**Продолжение табл. 9**

| доля | %, последний десятичный знак | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | **0** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** |
| | **Значения** $\varphi = \arcsin\sqrt{P}$ | | | | | | | | | |
| **23** | 1,000 | 1,003 | 1,005 | 1,007 | 1,010 | 1,012 | 1,015 | 1,017 | 1,019 | 1,022 |
| **24** | 1,024 | 1,026 | 1,029 | 1,031 | 1,033 | 1,036 | 1,038 | 1,040 | 1,043 | 1,045 |
| **25** | 1,047 | 1,050 | 1.052 | 1,054 | 1,056 | 1.059 | 1,061 | 1,063 | 1,066 | 1,068 |
| **26** | 1,070 | 1,072 | 1,075 | 1,077 | 1,079 | 1,082 | 1,084 | 1,086 | 1,088 | 1,091 |
| **27** | 1,093 | 1,095 | 1,097 | 1,100 | 1,102 | 1,104 | 1,106 | 1,109 | 1,111 | 1,113 |
| **28** | 1,115 | 1,117 | 1,120 | 1,122 | 1,124 | 1,126 | 1,129 | 1,131 | 1,133 | 1,135 |
| **29** | 1,137 | 1,140 | 1,142 | 1,144 | 1,146 | 1,148 | 1,151 | 1,153 | 1,155 | 1,157 |
| **30** | 1,159 | 1,161 | 1,164 | 1,166 | 1,168 | 1,170 | 1,172 | 1,174 | 1,177 | 1,179 |
| **31** | 1,182 | 1,183 | 1,185 | 1,187 | 1,190 | 1,192 | 1,194 | 1,196 | 1,198 | 1,200 |
| **32** | 1,203 | 1,205 | 1,207 | 1,209 | 1,211 | 1,213 | 1,215 | 1,217 | 1,220 | 1,222 |
| **33** | 1,224 | 1,226 | 1,228 | 1,230 | 1,232 | 1,234 | 1,237 | 1,239 | 1,241 | 1,243 |
| **34** | 1,245 | 1,247 | 1,249 | 1,251 | 1,254 | 1,256 | 1,258 | 1,260 | 1,262 | 1,264 |
| **35** | 1,266 | 1,268 | 1,270 | 1,272 | 1,274 | 1,277 | 1,279 | 1,281 | 1,283 | 1,285 |
| **36** | 1,287 | 1,289 | 1,291 | 1,293 | 1,295 | 1,297 | 1,299 | 1,302 | 1,304 | 1,306 |
| **37** | 1,308 | 1,310 | 1,312 | 1,314 | 1,316 | 1,318 | 1,320 | 1,322 | 1,324 | 1,326 |
| **38** | 1,328 | 1,330 | 1,333 | 1,335 | 1,337 | 1,339 | 1,341 | 1,343 | 1,345 | 1,347 |
| **39** | 1,349 | 1,351 | 1,353 | 1,355 | 1,357 | 1,359 | 1,361 | 1,363 | 1,365 | 1,367 |
| **40** | 1,369 | 1,371 | 1,374 | 1,376 | 1,378 | 1,380 | 1,382 | 1,384 | 1,386 | 1,388 |
| **41** | 1,390 | 1,392 | 1,394 | 1,396 | 1,398 | 1,400 | 1,402 | 1,404 | 1,406 | 1,408 |
| **42** | 1,410 | 1,412 | 1,414 | 1,416 | 1,418 | 1,420 | 1,422 | 1,424 | 1,426 | 1,428 |
| **43** | 1,430 | 1,432 | 1,434 | 1,436 | 1,438 | 1,440 | 1,442 | 1,444 | 1,446 | 1,448 |
| **44** | 1,451 | 1,453 | 1,455 | 1,457 | 1,459 | 1,461 | 1,463 | 1,465 | 1,467 | 1,469 |
| **45** | 1,471 | 1,473 | 1,475 | 1,477 | 1,479 | 1,481 | 1,483 | 1,485 | 1,487 | 1,489 |
| **46** | 1,491 | 1,493 | 1,495 | 1,497 | 1,499 | 1,501 | 1,503 | 1,505 | 1,507 | 1,509 |

**Продолжение табл. 9**

| доля | %, последний десятичный знак | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | **0** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** |
| | **Значения** $\varphi = \arcsin\sqrt{P}$ | | | | | | | | | |
| **47** | 1,511 | 1,513 | 1,515 | 1,517 | 1,519 | 1,521 | 1,523 | 1,525 | 1,527 | 1,529 |
| **48** | 1,531 | 1,533 | 1,535 | 1,537 | 1,539 | 1,541 | 1,543 | 1,545 | 1,547 | 1,549 |
| **49** | 1,551 | 1,553 | 1,555 | 1,557 | 1,559 | 1,561 | 1,563 | 1,565 | 1,567 | 1,569 |
| **50** | 1,571 | 1,573 | 1,575 | 1,577 | 1,579 | 1,581 | 1,583 | 1,585 | 1,587 | 1,589 |
| **51** | 1,591 | 1,593 | 1,595 | 1,597 | 1,599 | 1,601 | 1,603 | 1,605 | 1,607 | 1,609 |
| **52** | 1,611 | 1,613 | 1,615 | 1,617 | 1,619 | 1,621 | 1,623 | 1,625 | 1,627 | 1,629 |
| **53** | 1,631 | 1,633 | 1,635 | 1,637 | 1,639 | 1,641 | 1,643 | 1,645 | 1,647 | 1,649 |
| **54** | 1,651 | 1,653 | 1,655 | 1.657 | 1,659 | 1,661 | 1,663 | 1,665 | 1,667 | 1,669 |
| **55** | 1,671 | 1,673 | 1,675 | 1,677 | 1,679 | 1,681 | 1,683 | 1,685 | 1,687 | 1,689 |
| **56** | 1,691 | 1,693 | 1,695 | 1,697 | 1,699 | 1,701 | 1,703 | 1,705 | 1,707 | 1,709 |
| **57** | 1,711 | 1,713 | 1,715 | 1,717 | 1,719 | 1,721 | 1,723 | 1,725 | 1,727 | 1,729 |
| **58** | 1,731 | 1,734 | 1,736 | 1,738 | 1,740 | 1,742 | 1,744 | 1,746 | 1,748 | 1,750 |
| **59** | 1,752 | 1,754 | 1,756 | 1,758 | 1,760 | 1,762 | 1,764 | 1,766 | 1,768 | 1,770 |
| **60** | 1,772 | 1,774 | 1,776 | 1,778 | 1,780 | 1,782 | 1,784 | 1,786 | 1,789 | 1,791 |
| **61** | 1,793 | 1,795 | 1,797 | 1,799 | 1,801 | 1,803 | 1,805 | 1,807 | 1,809 | 1,811 |
| **62** | 1,813 | 1,815 | 1,817 | 1,819 | 1,821 | 1,823 | 1,826 | 1,828 | 1,830 | 1,832 |
| **63** | 1,834 | 1,836 | 1,838 | 1,840 | 1,842 | 1,844 | 1,846 | 1,848 | 1,850 | 1,853 |
| **64** | 1,855 | 1,857 | 1,859 | 1,861 | 1,863 | 1,865 | 1.867 | 1,869 | 1,871 | 1,873 |
| **65** | 1,875 | 1,878 | 1,880 | 1,882 | 1,884 | 1,886 | 1,888 | 1,890 | 1,892 | 1,894 |
| **66** | 1,897 | 1,899 | 1,901 | 1,903 | 1,905 | 1,907 | 1,909 | 1,911 | 1,913 | 1,916 |
| **67** | 1,918 | 1,920 | 1,922 | 1,924 | 1,926 | 1,928 | 1,930 | 1,933 | 1,935 | 1,937 |
| **68** | 1,939 | 1,941 | 1,943 | 1,946 | 1,948 | 1,950 | 1,952 | 1,954 | 1,956 | 1,958 |
| **69** | 1,961 | 1,963 | 1,965 | 1,967 | 1,969 | 1,971 | 1,974 | 1,976 | 1,978 | 1,980 |
| **70** | 1,982 | 1,984 | 1,987 | 1,989 | 1,991 | 1,993 | 1,995 | 1,998 | 2,000 | 2,002 |

**Продолжение табл. 9**

| доля | %, последний десятичный знак | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | **0** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** |
| | **Значения** $\varphi = \arcsin\sqrt{P}$ | | | | | | | | | |
| **71** | 2,004 | 2,006 | 2,009 | 2,011 | 2,013 | 2,015 | 2,018 | 2,020 | 2,022 | 2,024 |
| **72** | 2,026 | 2,029 | 2,031 | 2,033 | 2,035 | 2,038 | 2,040 | 2,042 | 2,044 | 2,047 |
| **73** | 2,049 | 2,051 | 2,053 | 2,056 | 2,058 | 2,060 | 2,062 | 2,065 | 2,067 | 2,069 |
| **74** | 2.071 | 2,074 | 2,076 | 2,078 | 2,081 | 2,083 | 2,085 | 2,087 | 2,090 | 2,092 |
| **75** | 2,094 | 2,097 | 2,099 | 2,101 | 2,104 | 2,106 | 2,108 | 2,111 | 2,113 | 2,115 |
| **76** | 2,118 | 2,120 | 2,122 | 2,125 | 2,127 | 2,129 | 2,132 | 2,134 | 2,136 | 2,139 |
| **77** | 2,141 | 2,144 | 2,146 | 2,148 | 2,151 | 2,153 | 2,156 | 2,158 | 2,160 | 2,163 |
| **78** | 2,165 | 2,168 | 2,170 | 2,172 | 2,175 | 2,177 | 2,180 | 2,182 | 2,185 | 2,187 |
| **79** | 2,190 | 2,192 | 2,194 | 2,197 | 2,199 | 2,202 | 2,204 | 2,207 | 2,209 | 2,212 |
| **80** | 2,214 | 2,217 | 2,219 | 2.222 | 2,224 | 2,227 | 2,229 | 2,231 | 2,234 | 2,237 |
| **81** | 2,240 | 2,242 | 2,245 | 2,247 | 2,250 | 2,252 | 2,255 | 2,258 | 2,260 | 2,263 |
| **82** | 2,265 | 2,268 | 2,271 | 2,273 | 2,276 | 2,278 | 2,281 | 2,284 | 2,286 | 2,289 |
| **83** | 2,292 | 2,294 | 2,297 | 2,300 | 2,302 | 2,305 | 2,308 | 2,310 | 2,313 | 2,316 |
| **84** | 2,319 | 2,321 | 2,324 | 2,327 | 2,330 | 2,332 | 2,335 | 2,338 | 2,341 | 2,343 |
| **85** | 2,346 | 2,349 | 2,352 | 2,355 | 2,357 | 2,360 | 2,363 | 2,366 | 2,369 | 2,372 |
| **86** | 2,375 | 2,377 | 2,380 | 2,383 | 2,386 | 2,389 | 2,392 | 2,395 | 2,398 | 2,401 |
| **87** | 2,404 | 2,.407 | 2,410 | 2,413 | 2,416 | 2,419 | 2,422 | 2,425 | 2,428 | 2,431 |
| **88** | 2,434 | 2,437 | 2,440 | 2,443 | 2,447 | 2,450 | 2,453 | 2,456 | 2,459 | 2,462 |
| **89** | 2,465 | 2,469 | 2,472 | 2,475 | 2,478 | 2,482 | 2,485 | 2,488 | 2,491 | 2,495 |
| **90** | 2,498 | 2,501 | 2,505 | 2,508 | 2,512 | 2,515 | 2,518 | 2,522 | 2,525 | 2,529 |
| **91** | 2,532 | 2,536 | 2,539 | 2,543 | 2,546 | 2,550 | 2,554 | 2,557 | 2,561 | 2,564 |
| **92** | 2,568 | 2,572 | 2,575 | 2,579 | 2,583 | 2,587 | 2,591 | 2,594 | 2,598 | 2,602 |
| **93** | 2,606 | 2,610 | 2,614 | 2,618 | 2,622 | 2,626 | 2,630 | 2,634 | 2,638 | 2,642 |
| **94** | 2,647 | 2,651 | 2,655 | 2,659 | 2,664 | 2,668 | 2,673 | 2,677 | 2,681 | 2,686 |

**Окончание табл. 9**

| доля | %, последний десятичный знак | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | **0** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** |
| | **Значения** $\varphi = \arcsin\sqrt{P}$ | | | | | | | | | |
| **95** | 2,691 | 2,295 | 2,700 | 2,705 | 2,709 | 2,714 | 2,719 | 2,724 | 2,729 | 2,734 |
| **96** | 2,739 | 2,744 | 2,749 | 2,754 | 2,760 | 2,765 | 2,771 | 2,776 | 2,782 | 2,788 |
| **97** | 2,793 | 2,799 | 2,805 | 2,811 | 2,818 | 2,824 | 2,830 | 2,837 | 2,844 | 2,851 |
| **98** | 2,858 | 2,865 | 2,872 | 2,880 | 2,888 | 2,896 | 2,904 | 2,913 | 2,922 | 2,931 |
| **99.0** | 2,941 | 2,942 | 2,943 | 2,944 | 2,945 | 2,946 | 2,948 | 2,949 | 2,950 | 2,951 |
| **99.1** | 2,952 | 2,953 | 2,954 | 2,955 | 2,956 | 2,957 | 2,958 | 2,959 | 2,960 | 2,961 |
| **99.2** | 2,963 | 2,964 | 2,965 | 2,966 | 2,967 | 2,968 | 2,969 | 2,971 | 2,972 | 2,973 |
| **99.5** | 2,974 | 2,975 | 2,976 | 2,978 | 2,979 | 2,980 | 2,981 | 2,983 | 2,984 | 2,985 |
| **99.4** | 2,987 | 2,988 | 2,989 | 2,990 | 2,992 | 2,993 | 2,995 | 2,996 | 2,997 | 2,999 |
| **99.5** | 3,000 | 3,002 | 3,003 | 3,004 | 3,006 | 3,007 | 3,009 | 3,010 | 3,012 | 3,013 |
| **99.6** | 3,015 | 3,017 | 3,018 | 3,020 | 3,022 | 3,023 | 3,025 | 3,027 | 3,028 | 3,030 |
| **99.7** | 3,032 | 3,034 | 3,036 | 3,038 | 3,040 | 3,041 | 3,044 | 3,046 | 3,048 | 3,050 |
| **99.8** | 3,052 | 3,054 | 3,057 | 3,059 | 3,062 | 3,064 | 3,067 | 3,069 | 3,072 | 3,075 |
| **99.9** | 3,078 | 3,082 | 3.085 | 3,089 | 3,093 | 3,097 | 3,101 | 3,107 | 3,113 | 3,122 |
| **100** | 3,142 | | | | | | | | | |

**Таблица 10. Критические значения *F* критерия Фишера**

**($C_1$ число степеней свободы между уровнями, $C_2$ внутри уровней)**

Влияние фактора или взаимодействия факторов достоверно,

если $F_{эмп}$, равен или больше критического значения $F_{005}$

и тем более достоверно, если $F_{эмп} > F_{0,01}$

| $C_1$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $C_2$ | P < 0,05 | | | | | | | | | | | |
| 1 | 161 | 200 | 216 | 225 | 230 | 234 | 237 | 239 | 241 | 242 | 243 | 244 |
| 2 | 18,51 | 19,00 | 19,16 | 19,25 | 19,30 | 19,33 | 19,36 | 19,37 | 19,38 | 19,39 | 19,40 | 19,41 |
| 3 | 10,13 | 9,55 | 9,28 | 9,12 | 9,01 | 8,94 | 8,88 | 8,84 | 8,81 | 8,78 | 8,76 | 8,74 |
| 4 | 7,71 | 6,94 | 6,59 | 6,39 | 6,26 | 6,16 | 6,09 | 6,04 | 6,00 | 5,96 | 5,93 | 5,91 |
| 5 | 6,61 | 5,79 | 5,41 | 5,19 | 5,05 | 4,95 | 4,88 | 4,82 | 4,78 | 4,74 | 4,70 | 4,68 |
| 6 | 5,99 | 5,14 | 4,76 | 4,53 | 4,39 | 4,28 | 4,21 | 4,15 | 4,10 | 4,06 | 4,03 | 4,00 |
| 7 | 5,59 | 4,74 | 4,35 | 4,12 | 3,97 | 3,87 | 3,79 | 3,73 | 3,68 | 3,63 | 3,60 | 3,57 |
| 8 | 5,32 | 4,46 | 4,07 | 3,84 | 3,69 | 3,58 | 3,50 | 3,44 | 3,39 | 3,34 | 3,31 | 3,28 |
| 9 | 5,12 | 4,26 | 3,86 | 3,63 | 3,48 | 3,17 | 3,29 | 3,23 | 3,18 | 3,13 | 3,10 | 3,07 |
| 10 | 4,96 | 4,10 | 3,71 | 3,48 | 3,33 | 3,22 | 3,14 | 3,07 | 3,02 | 2,97 | 2,94 | 2,91 |
| 11 | 4,84 | 3,98 | 3,59 | 3,36 | 3,20 | 3,09 | 3,01 | 2,95 | 2,90 | 2,86 | 2,82 | 2,79 |
| 12 | 4,75 | 3,88 | 3,49 | 3,26 | 3,11 | 3,00 | 2,92 | 2,85 | 2,80 | 2,76 | 2,72 | 2,69 |
| 13 | 4,67 | 3,80 | 3,41 | 3,18 | 3,02 | 2,92 | 2,84 | 2,77 | 2,72 | 2,67 | 2,63 | 2,60 |
| 14 | 4,60 | 3,74 | 3,34 | 3,11 | 2,96 | 2,85 | 2,77 | 2,70 | 2,65 | 2,60 | 2,56 | 2,53 |
| 15 | 4,54 | 3,68 | 3,29 | 3,06 | 2,90 | 2,79 | 2,70 | 2,64 | 2,59 | 2,55 | 2,51 | 2,48 |
| 16 | 4,49 | 3,63 | 3,24 | 3,01 | 2,85 | 2,74 | 2,66 | 2,59 | 2,54 | 2,49 | 2,45 | 2,42 |

**Окончание табл. 10**

| $C_2$ | P < 0,01 | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **1** | 4052 | 4999 | 5403 | 5625 | 5764 | 5859 | 5928 | 5981 | 6022 | 6056 | 6082 | 6106 |
| **2** | 98,49 | 99,00 | 99,17 | 99,25 | 99,30 | 99,33 | 99,36 | 99,37 | 99,39 | 99,40 | 99,41 | 99,42 |
| **3** | 34,12 | 30,82 | 29,46 | 28,71 | 28,24 | 27,91 | 27,67 | 27,49 | 27,34 | 27,23 | 27,13 | 27,05 |
| **4** | 21,20 | 18,00 | 16,69 | 15,98 | 15,52 | 15,21 | 14,98 | 14,80 | 14,66 | 14,54 | 14,45 | 14,37 |
| **5** | 16,26 | 13,27 | 12,06 | 11,39 | 10,97 | 10,67 | 10,45 | 10,29 | 10,15 | 10,05 | 9,96 | 9,89 |
| **6** | 13,74 | 10,92 | 9,78 | 9,15 | 8,75 | 8,47 | 8,26 | 8,10 | 7,98 | 7,87 | 7,79 | 7,72 |
| **7** | 12,25 | 9,55 | 8,45 | 7,85 | 7,46 | 7,19 | 7,00 | 6,84 | 6,71 | 6,62 | 6,54 | 6,47 |
| **8** | 11,26 | 8,65 | 7,59 | 7,01 | 6,63 | 6,37 | 6,19 | 6,03 | 5,91 | 5,82 | 5,74 | 5,67 |
| **9** | 10,56 | 8,02 | 6,9 | 6,42 | 6,06 | 5,80 | 5,62 | 5,47 | 5,35 | 5,26 | 5,18 | 5,11 |
| **10** | 10,04 | 7,56 | 6,55 | 5,99 | 5,64 | 5,39 | 5,21 | 5,06 | 4,95 | 4,85 | 4,78 | 4,71 |
| **11** | 9,65 | 7,20 | 6,22 | 5,67 | 5,32 | 5,07 | 4,88 | 4,74 | 4,63 | 4,54 | 4,46 | 4,40 |
| **12** | 9,33 | 6,93 | 5,95 | 5,41 | 5,06 | 4,82 | 4,65 | 4,50 | 4,39 | 4,30 | 4,22 | 4,16 |
| **13** | 9,07 | 6,70 | 5,74 | 5,20 | 4,86 | 4,62 | 4,44 | 4,30 | 4,19 | 4,10 | 4,02 | 3,96 |
| **14** | 8,86 | 6,51 | 5,56 | 5,03 | 4,69 | 4,46 | 4,28 | 4,14 | 4,03 | 3,94 | 3,86 | 3,80 |
| **15** | 8,68 | 6,36 | 5,42 | 4,89 | 4,56 | 4,32 | 4,14 | 4,00 | 3,89 | 3,80 | 3,73 | 3,67 |
| **16** | 8,53 | 6,23 | 5,29 | 4,77 | 4,44 | 4,20 | 4,03 | 3,89 | 3,78 | 3,69 | 3,61 | 3,55 |

**Таблица 11. Таблица значений функции Лапласа**

Функция Лапласа $\Phi(t) = \frac{2}{\sqrt{2\pi}} \int_0^t e^{\frac{t^2}{2}} dt = \frac{1}{\sqrt{2\pi}} \int_{-t}^{+t} e^{-\frac{t^2}{2}} dt$.

При разных значениях $t$; $\Phi(-t) = -\Phi(t)$ (функция нормального распределения).

| $t$ | $\Phi(t)$ | $t$ | $\Phi(t)$ | $t$ | $\Phi(t)$ | $t$ | $\Phi(t)$ |
|---|---|---|---|---|---|---|---|
| **0,00** | 0,00000 | **1,00** | 0,68269 | **2,00** | 0,95450 | **3,00** | 0,99730 |
| **0,01** | 0,00798 | **1,01** | 0,68750 | **2,01** | 0,95557 | **3,01** | 0,99739 |
| **0,02** | 0,01596 | **1,02** | 0,69227 | **2,02** | 0,95662 | **3,02** | 0,99747 |
| **0,03** | 0,02393 | **1,03** | 0,69699 | **2,03** | 0,95764 | **3,03** | 0,99755 |
| **0,04** | 0,03191 | **1,04** | 0,70166 | **2,04** | 0,95865 | **3,04** | 0,99763 |
| **0,05** | 0,03988 | **1,05** | 0,70628 | **2,05** | 0,95964 | **3,05** | 0,99771 |
| **0,06** | 0,04784 | **1,06** | 0,71086 | **2,06** | 0,96060 | **3,06** | 0,99779 |
| **0,07** | 0,05581 | **1,07** | 0,71538 | **2,07** | 0,96155 | **3,07** | 0,99786 |
| **0,08** | 0,06376 | **1,08** | 0,71986 | **2,08** | 0,96247 | **3,08** | 0,99793 |
| **0,09** | 0,07171 | **1,09** | 0,72429 | **2,09** | 0,96338 | **3,09** | 0,99800 |
| **0,10** | 0,07966 | **1,10** | 0,72867 | **2,10** | 0,96427 | **3,10** | 0,99806 |
| **0,11** | 0,08759 | **1,11** | 0,73300 | **2,11** | 0,96514 | **3,11** | 0,99813 |
| **0,12** | 0,09552 | **1,12** | 0,73729 | **2,12** | 0,96599 | **3,12** | 0,99819 |
| **0,13** | 0,10348 | **1,13** | 0,74152 | **2,13** | 0,96683 | **3,13** | 0,99825 |
| **0,14** | 0,11134 | **1,14** | 0,74571 | **2,14** | 0,96765 | **3,14** | 0,99831 |
| **0,15** | 0,11924 | **1,15** | 0,74986 | **2,15** | 0,96844 | **3,15** | 0,99837 |
| **0,16** | 0,12712 | **1,16** | 0,75395 | **2,16** | 0,96923 | **3,16** | 0,99842 |
| **0,17** | 0,13499 | **1,17** | 0,75800 | **2,17** | 0,96999 | **3,17** | 0,99848 |
| **0,18** | 0,14285 | **1,18** | 0,76200 | **2,18** | 0,97074 | **3,18** | 0,99853 |
| **0,19** | 0,15069 | **1,19** | 0,76595 | **2,19** | 0,97148 | **3,19** | 0,99858 |
| **0,20** | 0,15852 | **1,20** | 0,76986 | **2,20** | 0,97219 | **3,20** | 0,99863 |
| **0,21** | 0,16633 | **1,21** | 0,77372 | **2,21** | 0,97289 | **3,21** | 0,99867 |

**Продолжение табл. 11**

| $t$ | $\Phi(t)$ | $t$ | $\Phi(t)$ | $t$ | $\Phi(t)$ | $t$ | $\Phi(t)$ |
|---|---|---|---|---|---|---|---|
| **0,22** | 0,17413 | **1,22** | 0,77754 | **2,22** | 0,97358 | **3,22** | 0,99872 |
| **0,23** | 0,18191 | **1,23** | 0,78130 | **2,23** | 0,97425 | **3,23** | 0,99876 |
| **0,24** | 0,18967 | **1,24** | 0,78502 | **2,24** | 0,97491 | **3,24** | 0,99880 |
| **0,25** | 0,19741 | **1,25** | 0,78870 | **2,25** | 0,97555 | **3,25** | 0,99855 |
| **0,26** | 0,20514 | **1,26** | 0,79233 | **2,26** | 0,97618 | **3,26** | 0,99889 |
| **0,27** | 0,21284 | **1,27** | 0,79592 | **2,27** | 0,97679 | **3,27** | 0,99892 |
| **0,28** | 0,22052 | **1,28** | 0,79945 | **2,28** | 0,97739 | **3,28** | 0,99896 |
| **0,29** | 0,22818 | **1,29** | 0,80295 | **2,29** | 0,97798 | **3,29** | 0,99900 |
| **0,30** | 0,23582 | **1,30** | 0,80640 | **2,30** | 0,97855 | **3,30** | 0,99903 |
| **0,31** | 0,24344 | **1,31** | 0,80980 | **2,31** | 0,97911 | **3,31** | 0,99907 |
| **0,32** | 0,25103 | **1,32** | 0,81316 | **2,32** | 0,97966 | **3,32** | 0,99910 |
| **0,33** | 0,25860 | **1,33** | 0,81648 | **2,33** | 0,98019 | **3,33** | 0,99913 |
| **0,34** | 0,26614 | **1,34** | 0,81975 | **2,34** | 0,98072 | **3,34** | 0,99916 |
| **0,35** | 0,27366 | **1,35** | 0,82298 | **2,35** | 0,98123 | **3,35** | 0,99919 |
| **0,36** | 0,28115 | **1,36** | 0,82617 | **2,36** | 0,98172 | **3,36** | 0,99922 |
| **0,37** | 0,28862 | **1,37** | 0,82931 | **2,37** | 0,98221 | **3,37** | 0,99925 |
| **0,38** | 0,29605 | **1,38** | 0,83241 | **2,38** | 0,98269 | **3,38** | 0,99928 |
| **0,39** | 0,30346 | **1,39** | 0,83547 | **2,39** | 0,98315 | **3,39** | 0,99930 |
| **0,40** | 0,31084 | **1,40** | 0,83849 | **2,40** | 0,98360 | **3,40** | 0,99933 |
| **0,41** | 0,31819 | **1,41** | 0,84146 | **2,41** | 0,98405 | **3,41** | 0,99935 |
| **0,42** | 0,32552 | **1,42** | 0,84439 | **2,42** | 0,98448 | **3,42** | 0,99937 |
| **0,43** | 0,33280 | **1,43** | 0,84728 | **2,43** | 0,98490 | **3,43** | 0,99940 |
| **0,44** | 0,34006 | **1,44** | 0,85013 | **2,44** | 0,98531 | **3,44** | 0,99942 |
| **0,45** | 0,34729 | **1,45** | 0,85294 | **2,45** | 0,98571 | **3,45** | 0,99944 |
| **0,46** | 0,35448 | **1,46** | 0,85571 | **2,46** | 0,98611 | **3,46** | 0,99946 |
| **0,47** | 0,36164 | **1,47** | 0,85844 | **2,47** | 0,98649 | **3,47** | 0,99948 |

**Продолжение табл. 11**

| $t$ | $\Phi(t)$ | $t$ | $\Phi(t)$ | $t$ | $\Phi(t)$ | $t$ | $\Phi(t)$ |
|---|---|---|---|---|---|---|---|
| **0,48** | 0,36877 | **1,48** | 0,86113 | **2,48** | 0,98686 | **3,48** | 0,99950 |
| **0,49** | 0,37587 | **1,49** | 0,86378 | **2,49** | 0,98723 | **3,49** | 0,99952 |
| **0,50** | 0,38292 | **1,50** | 0,86639 | **2,50** | 0,98758 | **3,50** | 0,99953 |
| **0,51** | 0,38995 | **1,51** | 0,86696 | **2,51** | 0,98793 | **3,51** | 0,99955 |
| **0,52** | 0,39694 | **1,52** | 0,87149 | **2,52** | 0,98826 | **3,52** | 0,99957 |
| **0,53** | 0,40389 | **1,53** | 0,87398 | **2,53** | 0,98859 | **3,53** | 0,99958 |
| **0,54** | 0,41080 | **1,54** | 0,87644 | **2,54** | 0,98891 | **3,54** | 0,99960 |
| **0,55** | 0,41768 | **1,55** | 0,87886 | **2,55** | 0,98923 | **3,55** | 0,99961 |
| **0,56** | 0,42452 | **1,56** | 0,88124 | **2,56** | 0,98953 | **3,56** | 0,99963 |
| **0,57** | 0,43132 | **1,57** | 0,88358 | **2,57** | 0,98983 | **3,57** | 0,99964 |
| **0,58** | 0,43809 | **1,58** | 0,88589 | **2,58** | 0,99012 | **3,58** | 0,99966 |
| **0,59** | 0,44481 | **1,59** | 0,88817 | **2,59** | 0,99040 | **3,59** | 0,99967 |
| **0,60** | 0,45149 | **1,60** | 0,89040 | **2,60** | 0,99068 | **3,60** | 0,99968 |
| **0,61** | 0,45814 | **1,61** | 0,89260 | **2,61** | 0,99095 | **3,61** | 0,99969 |
| **0,62** | 0,46474 | **1,62** | 0,89477 | **2,62** | 0,99121 | **3,62** | 0,99971 |
| **0,63** | 0,47131 | **1,63** | 0,89690 | **2,63** | 0,99146 | **3,63** | 0,99972 |
| **0,64** | 0,47783 | **1,64** | 0,89899 | **2,64** | 0,99171 | **3,64** | 0,99973 |
| **0,65** | 0,48431 | **1,65** | 0,90106 | **2,65** | 0,99195 | **3,65** | 0,99974 |
| **0,66** | 0,49075 | **1,66** | 0,90309 | **2,66** | 0,99219 | **3,66** | 0,99975 |
| **0,67** | 0,49714 | **1,67** | 0,90508 | **2,67** | 0,99241 | **3,67** | 0,99976 |
| **0,68** | 0,50350 | **1,68** | 0,90704 | **2,68** | 0,99263 | **3,68** | 0,99977 |
| **0,69** | 0,50981 | **1,69** | 0,90897 | **2,69** | 0,99285 | **3,69** | 0,99978 |
| **0,70** | 0,51607 | **1,70** | 0,91087 | **2,70** | 0,99307 | **3,70** | 0,99978 |
| **0,71** | 0,52230 | **1,71** | 0,91273 | **2,71** | 0,99327 | **3,71** | 0,99979 |
| **0,72** | 0,52848 | **1,72** | 0,91457 | **2,72** | 0,99347 | **3,72** | 0,99980 |
| **0,73** | 0,53461 | **1,73** | 0,91637 | **2,73** | 0,99367 | **3,73** | 0,99981 |

**Окончание табл. 11**

| $t$ | $\Phi(t)$ | $t$ | $\Phi(t)$ | $t$ | $\Phi(t)$ | $t$ | $\Phi(t)$ |
|---|---|---|---|---|---|---|---|
| **0,74** | 0,54070 | **1,74** | 0,91814 | **2,74** | 0,99386 | **3,74** | 0,99982 |
| **0,75** | 0,54675 | **1,75** | 0,91988 | **2,75** | 0,99404 | **3,75** | 0,99982 |
| **0,76** | 0,55275 | **1,76** | 0,92159 | **2,76** | 0,99422 | **3,76** | 0,99983 |
| **0,77** | 0,55870 | **1,77** | 0,92327 | **2,77** | 0,99439 | **3,77** | 0,99984 |
| **0,78** | 0,56461 | **1,78** | 0,92492 | **2,78** | 0,99456 | **3,78** | 0,99984 |
| **0,79** | 0,57047 | **1,79** | 0,92655 | **2,79** | 0,99473 | **3,79** | 0,99985 |
| **0,80** | 0,57629 | **1,80** | 0,92814 | **2,80** | 0,99489 | **3,80** | 0,99986 |
| **0,81** | 0,58206 | **1,81** | 0,92970 | **2,81** | 0,99505 | **3,81** | 0,99986 |
| **0,82** | 0,58778 | **1,82** | 0,93124 | **2,82** | 0,99520 | **3,82** | 0,99987 |
| **0,83** | 0,59346 | **1,83** | 0,93275 | **2,83** | 0,99535 | **3,83** | 0,99987 |
| **0,84** | 0,59909 | **1,84** | 0,93423 | **2,84** | 0,99549 | **3,84** | 0,99988 |
| **0,85** | 0,60468 | **1,85** | 0,93569 | **2,85** | 0,99563 | **3,85** | 0,99988 |
| **0,86** | 0,61021 | **1,86** | 0,93711 | **2,86** | 0,99576 | **3,86** | 0,99989 |
| **0,87** | 0,61570 | **1,87** | 0,93852 | **2,87** | 0,99590 | **3,87** | 0,99989 |
| **0,88** | 0,62114 | **1,88** | 0,93989 | **2,88** | 0,99602 | **3,88** | 0,99990 |
| **0,89** | 0,62653 | **1,89** | 0,94124 | **2,89** | 0,99615 | **3,89** | 0,99990 |
| **0,90** | 0,63188 | **1,90** | 0,94257 | **2,90** | 0,99627 | **3,90** | 0,99990 |
| **0,91** | 0,63718 | **1,91** | 0,94387 | **2,91** | 0,99639 | **3,91** | 0,99991 |
| **0,92** | 0,64243 | **1,92** | 0,94514 | **2,92** | 0,99650 | **3,92** | 0,99991 |
| **0,93** | 0,64763 | **1,93** | 0,94639 | **2,93** | 0,99661 | **3,93** | 0,99992 |
| **0,94** | 0,65278 | **1,94** | 0,94762 | **2,94** | 0,99672 | **3,94** | 0,99992 |
| **0,95** | 0,65789 | **1,95** | 0,94882 | **2,95** | 0,99682 | **3,95** | 0,99992 |
| **0,96** | 0,66294 | **1,96** | 0,95000 | **2,96** | 0,99692 | **3,96** | 0,99992 |
| **0,97** | 0,66795 | **1,97** | 0,95116 | **2,97** | 0,99702 | **3,97** | 0,99993 |
| **0,98** | 0,67291 | **1,98** | 0,95230 | **2,98** | 0,99712 | **3,98** | 0,99993 |
| **0,99** | 0,67783 | **1,99** | 0,95341 | **2,99** | 0,99721 | **3,99** | 0,99993 |

## Таблица 12. Критические значения распределения Стьюдента

| **Число степеней свободы** $f = n - 1$ | $n$ | **Доверительная вероятность** | | | |
|---|---|---|---|---|---|
| | | **0,90** | **0,95** | **0,99** | **0,999** |
| **1** | 2 | 6,3137515148 | 12,7062047364 | 63,6567411629 | 636,619249432 |
| **2** | 3 | 2,91998558036 | 4,30265272991 | 9,92484320092 | 31,599054577 |
| **3** | 4 | 2,3533634348 | 3,18244630528 | 5,84090929976 | 12,9239786366 |
| **4** | 5 | 2,13184678134 | 2,7764451052 | 4,60409487142 | 8,61030158138 |
| **5** | 6 | 2,01504837267 | 2,57058183661 | 4,03214298356 | 6,86882663987 |
| **6** | 7 | 1,94318028039 | 2,44691184879 | 3,70742802132 | 5,95881617993 |
| **7** | 8 | 1,89457860506 | 2,36462425101 | 3,49948329735 | 5,40788252098 |
| **8** | 9 | 1,85954803752 | 2,30600413503 | 3,35538733133 | 5,04130543339 |
| **9** | 10 | 1,83311293265 | 2,26215716274 | 3,24983554402 | 4,78091258593 |
| **10** | 11 | 1,81246112281 | 2,22813885196 | 3,16927266718 | 4,5868938587 |
| **11** | 12 | 1,7958848187 | 2,20098516008 | 3,10580651322 | 4,43697933823 |
| **12** | 13 | 1,78228755565 | 2,17881282966 | 3,05453958834 | 4,31779128361 |
| **13** | 14 | 1,77093339599 | 2,16036865646 | 3,01227583821 | 4,22083172771 |
| **14** | 15 | 1,76131013577 | 2,14478668792 | 2,97684273411 | 4,14045411274 |
| **15** | 16 | 1,75305035569 | 2,13144954556 | 2,94671288334 | 4,0727651959 |
| **16** | 17 | 1,74588367628 | 2,11990529922 | 2,92078162235 | 4,0149963326 |
| **17** | 18 | 1,73960672608 | 2,10981557783 | 2,89823051963 | 3,96512626361 |
| **18** | 19 | 1,73406360662 | 2,10092204024 | 2,87844047271 | 3,92164582001 |
| **19** | 20 | 1,72913281152 | 2,09302405441 | 2,86093460645 | 3,88340584948 |
| **20** | 21 | 1,72471824292 | 2,08596344727 | 2,84533970978 | 3,84951627298 |
| **21** | 22 | 1,72074290281 | 2,07961384473 | 2,83135955802 | 3,81927716303 |
| **22** | 23 | 1,71714437438 | 2,0738730679 | 2,8187560606 | 3,79213067089 |
| **23** | 24 | 1,71387152775 | 2,06865761042 | 2,80733568377 | 3,76762680377 |
| **24** | 25 | 1,71088207991 | 2,06389856163 | 2,79693950477 | 3,74539861893 |
| **25** | 26 | 1,70814076125 | 2,05953855275 | 2,78743581368 | 3,72514394948 |
| **26** | 27 | 1,70561791976 | 2,05552943864 | 2,77871453333 | 3,70661174331 |
| **27** | 28 | 1,70328844572 | 2,05183051648 | 2,77068295712 | 3,68959171334 |
| **28** | 29 | 1,70113093427 | 2,0484071418 | 2,76326245546 | 3,67390640062 |
| **29** | 30 | 1,69912702653 | 2,04522964213 | 2,75638590367 | 3,6594050194 |
| **30** | 31 | 1,69726089436 | 2,0422724563 | 2,74999565357 | 3,645958635 |
| **40** | 41 | 1,68385101139 | 2,021075383 | 2,70445926743 | 3,55096576086 |
| **60** | 61 | 1,67064886465 | 2,00029782106 | 2,66028303115 | 3,4602004692 |
| **120** | 121 | 1,65765089935 | 1,97993040505 | 2,61742114477 | 3,37345376507 |
| **999999,0** | 1000000,0 | 1,64485515072 | 1,95996635682 | 2,57583422011 | 3,29053646126 |

Учебное издание

Борисова Е. В.

Прикладные статистические модели и методы в социологии

(уровень бакалавриата)

Учебное пособие

Подписано в печать 16.12.2016. Формат 60x84/16.
Гарнитура «Times». Бумага офсетная. Печать офсетная.
Усл.-печатные листы: 14,76. Тираж 500 экз. Заказ № 7068
Мнения авторов могут не совпадать с мнением издательства.
Печатается в авторской редакции.

Издательство «АНАЛИТИКА РОДИС»
142400, Московская обл., Ногинск, ул. Рогожская д. 7.
http://www.publishing-vak.ru analitikarodis@yandex.ru
+7 (495) 210 0554, +7 915 056 9894
Отпечатано с готового оригинал-макета в типографии
«Книга по Требованию». 127918, г. Москва, Сущевский вал, д. 49.
ISBN 978-5-905277-73-3