О.В.Иванов

СТАТИСТИКА

учебный курс для социологов и менеджеров

Часть 1

Описательная статистика

Теоретико-вероятностные основания статистического вывода

Москва
2005

Иванов О.В. Статистика / Учебный курс для социологов и менеджеров. Часть 1. Описательная статистика. Теоретико-вероятностные основания статистического вывода. – М. 2005. – 187 с.

Учебный курс является новым и подготовлен для преподавания студентам-социологам и менеджерам в составе цикла математических дисциплин. Соответствует Государственному образовательному стандарту высшего профессионального образования по специальностям «Социология» и «Менеджмент». Содержит теоретическую часть, примеры, а также задачи для аудиторных и самостоятельных занятий.

Книга может быть полезна преподавателям, студентам, научным сотрудникам, аналитикам, всем, кто занимается прикладным статистическим анализом эмпирических социальных и экономических данных.

По вопросам, связанным с настоящим изданием, обращаться на кафедру социальной информатики социологического факультета МГУ им.М.В.Ломоносова по электронному адресу: info@socio.msu.ru

ОГЛАВЛЕНИЕ

Предисловие автора

Идея создания этого учебника появилась довольно давно. Студенты-социологи, как и студенты других гуманитарных факультетов, обычно испытывают трудности с пониманием и усвоением математических курсов. И дело скорее не в том, что они недостаточно подготовлены к изучению математических дисциплин, сколько в гуманитарном характере мышления. Общеизвестно, что студенты-социологи в процессе обучения с большим удовольствием работают с эмпирическим материалом, нежели с абстракциями и теоретическими обобщениями. Математические понятия лучше усваивают на примерах, нежели чем при помощи формальных определений с использованием непривычных символических знаков и формул. По этой причине в начало курса вынесены главы описательной статистики, в которых предполагается работа с эмпирическими данными: их представление, обработка, получение несложных характеристик. Теоретико-вероятностные понятия рассматриваются вслед за изучением их эмпирических аналогов, после работы с «цифрой из жизни».

Статистика предполагает наличие фундаментальных математических оснований. Поэтому в курсе присутствуют все необходимые понятия и теоремы, лежащие в основе статистических заключений. Особое место уделено логике статистического вывода.

Курс статистики предполагает знание основ математического анализа в объеме, предусмотренном программой по высшей математике первого семестра. Интегралы, в том числе несобственные, которые изучались студентами ранее, окажутся тесно связанными с вероятностью непрерывных случайных величин и найдут свое достойное место в настоящем курсе. Не останутся бесполезными и другие знания и навыки, полученные при изучении начального математического курса.

В учебник включены практические задачи - в качестве примеров, и в качестве задач для решения на семинарских занятиях и самостоятельно. Это позволяет использовать учебник в качестве основного по курсу теории вероятностей и математической статистики для студентов гуманитарных специальностей. Электронная версия учебника, материалы лекций и практические задания находятся в Интернете по адресу: informatics.socio.msu.ru.

Автор выражает большую признательность своим коллегам: Сосолихину А.А., Самыловскому А.И., Семенову К.Е., Астаховой Н.В., Коченкову А.И. за активную поддержку и помощь в подготовке настоящего курса, а также своим близким за прочтение и полезные советы и замечания.

Глава 1. Почему социологи применяют статистику

Первая глава является введением в курс статистики. В ней дается определение математической статистики как науки, включающей описание данных, статистическое оценивание и проверку гипотез. Обсуждена роль, которую статистические методы играют в социальных исследованиях. Второй параграф посвящен измерениям и шкалам. Рассматриваются типы данных, различные шкалы, а также критерии, которым должны отвечать измерения, чтобы быть полезными для получения практических выводов.

1-1 Что такое статистика

Во все времена люди стремились описать мир, в котором они живут. Для этого использовались самые разные средства. Кто-то выбирает прозу для описания чувств и эмоциональных переживаний, другие применяют поэзию и живопись.

Статистика всего лишь один из таких методов описания, хотя и уникальный, поскольку она описывает мир при помощи чисел. Как следствие, статистика может предоставить уникальные сведения относительно мира вокруг нас.

Чтобы получать пользу от числовых описаний, следует понимать правила и логику, которая применяется при их составлении. Статистические описания могут быть сделаны хорошо или безграмотно и в некоторых случаях бывает крайне сложно отличить одно от другого. Если вы собрались анализировать что-то при помощи чисел, вам необходимо делать это корректно. Этот курс сможет помочь вам в этом. Вы сможете воспринимать и критично оценивать предлагаемые вам числовые описания, а также сможете самостоятельно создавать их и корректно использовать. Некоторые описания окажутся относительно простыми для понимания, в то время как другие потребуют упорства и продолжительной работы для их изучения.

Рисунок 1-1. Я подготовил тезисы своего доклада, а вы подберите немного статистики, чтобы их обосновать.

Что такое статистика? В широком смысле, под статистикой понимают область деятельности людей, направленную на сбор информации и ее анализ с целью изучения массовых явлений в природе и обществе. Поскольку в статистике применяются научные методы, под статистикой принято понимать также науку о методах сбора данных, их обработки и анализа для выявления закономерностей, присущих изучаемому явлению.

Статистика - наука, изучающая количественные характеристики массовых явлений в неразрывной связи с учетом их качественного своеобразия.

Как наука, статистика представляет собой совокупность научных методов планирования эксперимента, сбора данных, их организации, представления, обобщения, анализа, интерпретации и получения выводов относительно изучаемого явления, основанных на этих данных. Значительная часть статистики основывается на теоретико-вероятностных моделях и методах, поэтому принято также говорить о статистике математической.

Математическая статистика - область науки, разрабатывающая математические методы для изучения количественных характеристик массовых явлений. Составными частями математической статистики являются: (1) описание данных, (2) статистическое оценивание и (3) проверка статистических гипотез.

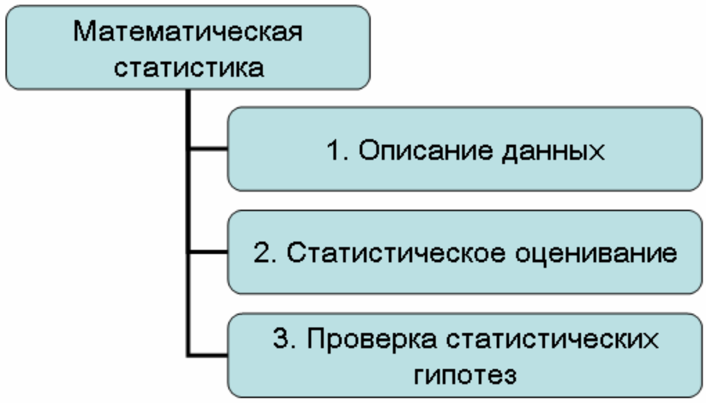


Рисунок 1-2. Разделы математической статистики

Поскольку статистика – наука древняя, крайне сложно установить точные границы для каждого из названных понятий статистики. В частности, возможно ли установить границы статистики математической и отделить ее от статистики «остальной»? Скорее нет, поскольку научные статистические исследования неотделимы от используемых статистических методов, которые всегда имеют математическую составляющую в качестве основы.

Статистика имеет дело с данными. Данные получают в результате наблюдений или специально организовываемых экспериментов, испытаний.

Данные (data) представляют собой результаты наблюдений, испытаний, накапливаемые с целью последующего изучения и анализа.

Данные могут представлять собой результаты анкетного опроса, сведения, полученные из официальных и неофициальных источников, Интернета, результаты наблюдений исследователя. Данные могут быть представлены на различных носителях и в различной форме.

Есть ли у вас проблемы с числами?

Не секрет, что многие люди менее комфортно чувствуют себя с числами, нежели чем со словами. Это не странно, поскольку в каждодневном общении люди именно словами выражают свои чувства и мысли. Мы не используем числа в качестве основного средства диалога. Люди с большим трудом воспринимают формулы, таблицы чисел и другие числовые выражения.

Кроме того, существует огромное *недоверие к числам*, а также к работе статистиков. Общеизвестна фраза: «есть ложь, есть наглая ложь, а есть статистика». Она означает, что статистики имеют склонность быть нечестными и при помощи чисел часто обосновывают ложные выводы. Эти слова, тем не менее, чаще употребляются в виде шутки, поскольку никто никогда не доказывал справедливость этого утверждения. На самом деле, нет оснований полагать, что статистики при помощи чисел обманывают людей чаще, чем кто-то другой из нас при помощи слов. Недоверие к числам, связанным со статистикой, есть скорее отражение расстройства людей при встрече с числовыми описаниями, которые иногда очень непросто понять и еще труднее использовать. Правда и то, что многие люди применяют числовые описания без осознания ограничений, отклоняются от своего первоначального замысла, а иногда напрямую незаконно подтасовывают их для достижения своих целей. Этот курс поможет читателю научиться понимать статистические описания, различать более точные числовые представления от менее точных, видеть, когда кто-то неправомерно использует числа или отклоняется от своих первоначальных намерений.

Иногда можно услышать, что использование чисел для описания людей и их деятельности есть «дегуманизация». Никто из нас не хочет «быть только числом». Тем не менее, хотя числа могли бы использоваться для достижения недостойных целей дегуманизации, у людей значительно больше возможностей, чтобы делать то же самое при помощи слов. Мы чаще имеем дело с плохими словами, чем с плохими числами. Для оскорблений люди чаще используют слова, нежели чем числа. Думается, тем самым, что использование чисел должно быть не более проблематичным, чем использование слов. В любом случае, следует разобраться в том, что означают числа и как использовать их должным образом и в правильных целях.

Еще одна проблема состоит в том, что *статистика во многом основывается на математике, теории вероятностей, логике*, довольно непростых математических дисциплинах. Это является серьезным препятствием для тех, кто чувствует недостаточную свою подготовку по общематематическим дисциплинам, чтобы адекватно воспринимать материал. Это заставляет их держаться подальше от предмета статистики в целом. Все это так. Тем не менее, не надо быть очень искушенным в математике, чтобы понимать и применять большую часть статистических методов. Наибольшее количество книг по статистике (включая эту) доступно всем, кто способен складывать, умножать, вычитать и делить.

И еще несколько важных замечаний. Студенты-социологи в целом стараются избегать *математических формул*. Чаще на экзамене по математике можно услышать от студентов математическое определение производной, заученное наизусть словами, чем записанное путем несложной формулы. Действительно, статистические методы используют формулы, составленные из абстрактных символов. Эти символы для закаленных пользователей всего лишь средство стенографии, в то время как для неопытных они выступают бессмысленными значками, которые могут только запутать. Мы будем своевременно разъяснять значение и смысл каждой формулы, чтобы устранить любые догадки.

Статистические методы *основаны на логике*. Будущим социологам следует опасаться применения статистических методов без их глубокого понимания и без контекста, который может оказаться крайне важным. Только после постижения внутренней логики каждого из методов можно с уверенностью говорить о способности исследователя без труда применять статистику для изучения социальных явлений.

Переменные (признаки)

Статистики описывают мир числами, подсчитывая, оценивая, и измеряя количество, размер, мнения и чувства. Например, они подсчитывают количество людей, работающих на различных должностях или имеющих различные профессии, исследуют, как жители города оценивают работу мэра в баллах от 1 до 10, или вычисляют размер годового дохода, приходящийся на одного члена семьи. В более сложных вопросах рассматривается связь между двумя или более измерениями.

Когда речь заходит об измерениях, следует понимать, что изучаемые исследователем понятия и категории в ходе проведения исследования должны утратить черты неопределенности и превратиться во что-то совершенно конкретное и *измеримое*. Любое исследование рано или поздно сосредотачивается вокруг тех свойств изучаемых объектов, которые могут быть измерены, упорядочены, подсчитаны. Такой процесс перехода от общих категорий к конкретным величинам называется *операционализацией понятий*. В результате операционализации мы получаем набор переменных, признаков изучаемых объектов.

Переменная, признак (variable) – это некоторая общая для всех изучаемых объектов характеристика или свойство, конкретные проявления которого могут меняться от объекта к объекту. Проявления признака называют значениями, альтернативами, градациями.

Например, переменная «пол» имеет два значения: «мужчина» и «женщина». Переменная «профессия» может принимать большое число различных значений. Субъективное понятие «счастье», например, может быть преобразовано в переменную со значениями в пределах от «очень несчастного» до «очень счастливого». Переменная «рост» может меняться от «очень низкий» до «очень высокий», или от ста пятидесяти сантиметров до двух метров десяти сантиметров, в зависимости от того, как мы захотим его использовать для целей нашего исследования.

В последнем примере с ростом хорошо видно, что понятие переменной или признака есть результат нашей собственной умственной деятельности. Хотя переменные и отражают определенные свойства объектов, тем не менее, выбор переменных, а также выбор градаций одной переменной напрямую зависит от исследователя, исследовательской цели, а также от множества субъективных особенностей. Умение «мыслить признаками», правильно определять переменные для достижения исследовательских целей является одним из важнейших качеств социолога-профессионала.

Иногда переменная может стать постоянной, например, когда респондентами являются только женщины. В таком случае мы перестаем рассматривать «пол» как изменяющуюся переменную, и строим наше исследование, считая пол постоянным.

Чем признак объекта отличается от переменной? Мы будем считать слова «признак» и «переменная» синонимами. Между ними есть отличие в части употребления в контексте. Признак больше связан с характеристиками объекта. Переменная, наоборот, отдалена от изучаемых характеристик объект, и на первом плане остаются способ измерения этой характеристики и ее возможные значения. Тогда мы в большей степени оказываемся вне качественной стороны исследуемой проблемы, абстрагируемся от нее в угоду рассмотрению математической, абстрактной модели.

Распределения переменных

Статистики не описывают один отдельно взятый случай. Их интересуют массовые явления. Они рассматривают большое число случаев, испытаний. Предмет их исследования – значения, которые переменная или признак принимает на различных объектах, для различных индивидуумов, составляющих группу или выборку. Различные значения переменной, которые она принимает для различных изучаемых объектов, приводят нас к необходимости рассматривать *распределение переменной (distribution of the variable)*.

Распределение переменной (distribution of the variable) – совокупность различных значений, которые переменная принимает для различных изучаемых объектов.

Предположим, например, что некоторое изучаемое нами сообщество состоит из 5 тысяч жителей района. Оно может быть описано нами в терминах пола, а именно процентного соотношения между мужчинами и женщинами. Если 55% среди сообщества - женщины, и 45% - мужчины, это есть распределение переменной пола в исследуемом сообществе. Далее, сообщество может быть описано в терминах возраста. Список возрастов 5 тысяч жителей - распределение переменной возраста. Обратите внимание, что каждая переменная, которая может нами рассматриваться, будь это пол, возраст, уровень образования, профессия, годовой доход и так далее – имеет определенное распределение. Кроме этого, распределение любой из этих переменных в изучаемом сообществе может отличаться от распределения этой же переменной, измеренной в другом сообществе.

Генеральная совокупность и выборка

Социолог, как правило, не имеет возможности (да и желания тоже) рассмотреть, подвергнуть эмпирическому изучению всю интересующую его совокупность объектов.

Сплошной опрос всех 5 тысяч человек изучаемого сообщества явился бы исключительно дорогим занятием и потребовал бы много времени. Кроме этого, изучение всех членов сообщества неминуемо привело бы ко вполне понятным человеческим ошибкам, связанным с обработкой большого объема информации. Оказывается, вполне достаточно изучить лишь некоторую выборку из всей совокупности и получить точную информацию о сообществе, или совокупности изучаемых объектов, если эта выборка была правильно сформирована и позволяет нам делать корректные выводы. При изучении сообщества из 5 тысяч человек, возможно, достаточно было бы использовать выборку из 200 или менее членов, чтобы получить сведения обо всем сообществе.

Интуитивно ясно, что размер выборки зависит от разнообразия изучаемой генеральной совокупности. Чем больше разнообразие, тем больше надо брать выборку. Это будет означать, что различные элементы генеральной совокупности и их сочетания представлены в выборке должным образом. Мы вернемся к этим соображениям, высказанным сейчас нестрого, позже, когда будем говорить о выборке в математических терминах. Дадим определение генеральной совокупности и выборки.

Генеральная совокупность (population) – вся интересующая исследователя совокупность изучаемых объектов.

Если вы собираетесь провести исследование о студентах своего факультета, генеральной совокупностью являются студенты факультета, возможно, только нынешние, а, возможно, и уже закончившие обучение. Если вы просто изучаете студентов, то генеральной совокупностью может стать совокупность студентов вашего города, а может, страны или всего мира. Исследователь сам определяет границы генеральной совокупности в зависимости от предмета и целей исследования.

Выборка, выборочная совокупность (sample) – некоторая, обычно небольшая, часть генеральной совокупности, отбираемая специальным образом и исследуемая с целью получения выводов о свойствах генеральной совокупности.

Когда исследователи используют слово «выборка», они, как правило, имеют в виду *репрезентативную*, то есть представительную выборку. Всякий раз, когда мы будем использовать слово выборка, мы будем подразумевать представительную выборку.

Мы будем использовать слово *параметры*, когда будет идти речь о характеристиках генеральной совокупности. Характеристики выборки мы будем называть *статистиками*.

Параметры (parameters) – числовые характеристики генеральной совокупности.

Статистики (statistics) – числовые характеристики выборки.

Мы будем использовать статистики для оценки тех параметров генеральной совокупности, которым они соответствуют. Мы увидим, что параметрами генеральной совокупности являются среднее значение и дисперсия, а среднее значение и дисперсия выборки есть статистики, которые позволяют оценивать параметры генеральной совокупности.

Описательная и аналитическая статистика

Статистические методы могут быть разделены на два типа – описательные и аналитические. Описательная статистика точно соответствует своему названию.

Описательная статистика (descriptive statistics) состоит из статистических методов, которые позволяют проводить сбор, упорядочение, обобщение и визуализацию данных.

Аналитическая статистика (inferential statistics) состоит из методов, которые на основе изучения статистик выборки позволяют получать выводы о параметрах генеральной совокупности.

Большинство исследований основывается на аналитической статистике, поскольку, как уже говорилось, слишком трудно и дорого изучать генеральные совокупности целиком. Статистика делает возможным избежать методов сплошного наблюдения. Если, например, мы интересуемся возрастом студентов в студенческой группе, мы находимся в области описательной статистики. А если нас интересует отношение пенсионеров к льготам на транспорте, то это относится уже к области аналитической статистики, поскольку для получения выводов нам потребуется выборка и возможность распространить результаты ее анализа на всю исследуемую генеральную совокупность.

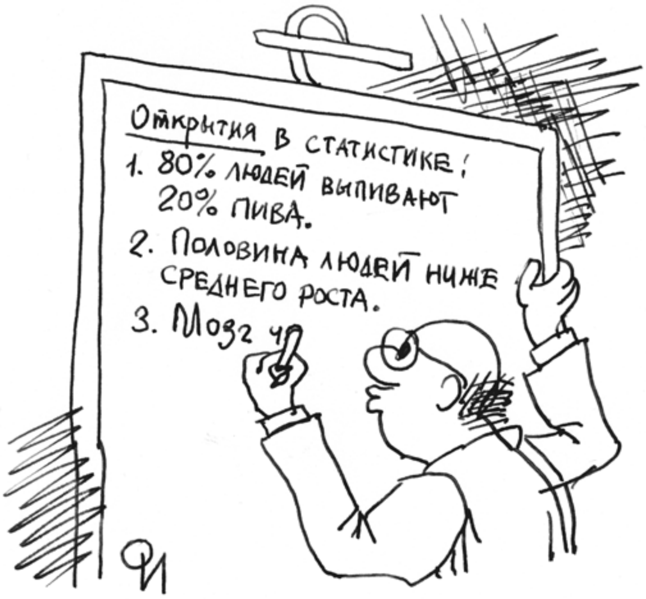


Рисунок 1-3. Статистики много знают о нас и нашей повседневной жизни.

Роль статистики в социальных исследованиях

Научное исследование предполагает определенную последовательность действий.

На первом этапе требуется выбрать и изложить *теоретическую модель*, которая будет использоваться для последующего изучения социальных явлений или событий. Теория является важным компонентом исследования и предполагает корректное определение изучаемых понятий и категорий, обзор публикаций и материалов предыдущих исследований, развитие предметной области. Письменное изложение теоретической модели позволяет увидеть, что и как исследователь думает об изучаемом явлении. Кроме того, детальное описание и развитие теории необходимо для последующих стадий получения и анализа эмпирического материала.

На втором этапе исследования проводится операционализация понятий. Под этим понимается переход от теоретических понятий и категорий к величинам, которые могут быть измерены количественно. Исследователь выделяет переменные, признаки исследуемых объектов, описывает возможные значения этих переменных. На этом же этапе исследователь формулирует предположения относительно характеристик переменных для рассматриваемых объектов. Именно эти предположения, называемые гипотезами, впоследствии подлежат эмпирической проверке.

Гипотеза (hypothesis) – предположение относительно параметров генеральной совокупности, которое подлежит проверке на основе анализа выборки.

Гипотезы являются очень важным для исследователя инструментом, поскольку могут быть проверены при помощи данных, полученных в результате наблюдений. Как мы потом увидим, большая часть статистической работы основана на проверке гипотез. Пример гипотезы: нет связи между продолжительностью рабочего дня и качеством работы.

Одна из целей исследования может состоять в том, чтобы объяснить, почему определенная переменная изменяется. Почему у людей разный рост или почему люди по-разному оценивают мэра? Вопрос, на который мы хотим получить ответ, содержит *зависимую переменную*. Мы можем выдвинуть гипотезу о том, что зависимая переменная изменяется в *зависимости* от некоторых обстоятельств или причин. Эти обстоятельства или причины после операционализации становятся *независимыми переменными*.

Переменная не является зависимой или независимой изначально, сама по себе. Она ставится той или иной после выбора теоретической модели и изложения предметной области, постановки вопросов исследования и формудилровки гипотез. Переменная, являющаяся зависимой при проверке одной гипотезы, может быть независимой переменной в другой гипотезе. Например, уровень образования может быть независимой переменной, если мы захотим проверить, зависит ли размер годового дохода от уровня образования. Эта же переменная является зависимой, если цель нашего исследования состоит в проверке предположения о существенных различиях в уровне образования взрослого населения в нескольких различных регионах.

Как только создана теоретическая модель, определены переменные и названы гипотезы, следует определить способы получения данных. Данные могут быть собраны различным путем, включая интервью, анкетные опросы, исследовательские наблюдения. Исследователь должен определить границы генеральной совокупности, описать ее. Кроме этого важно, чтобы стратегия получения выборки, которую изберет исследователь, позволяла получить репрезентативную выборку, то есть такую, которая будет хорошо представлять генеральную совокупность.

После уточнения методов, происходит сбор данных, проведение опросов. Если исследование правильно спланировано, сбор данных оказывается не слишком трудоемким и позволяет в короткий срок получить все необходимое для последующей обработки и применения статистических методов.

На следующем этапе применяются статистические методы: для обработки данных, их анализа, визуального представления. Параметры генеральной совокупности могут быть оценены на основе изучения статистик – характеристик полученной выборки. Кроме этого, могут быть проверены гипотезы относительно значений параметров генеральной совокупности, относительно связей между двумя или несколькими переменными.

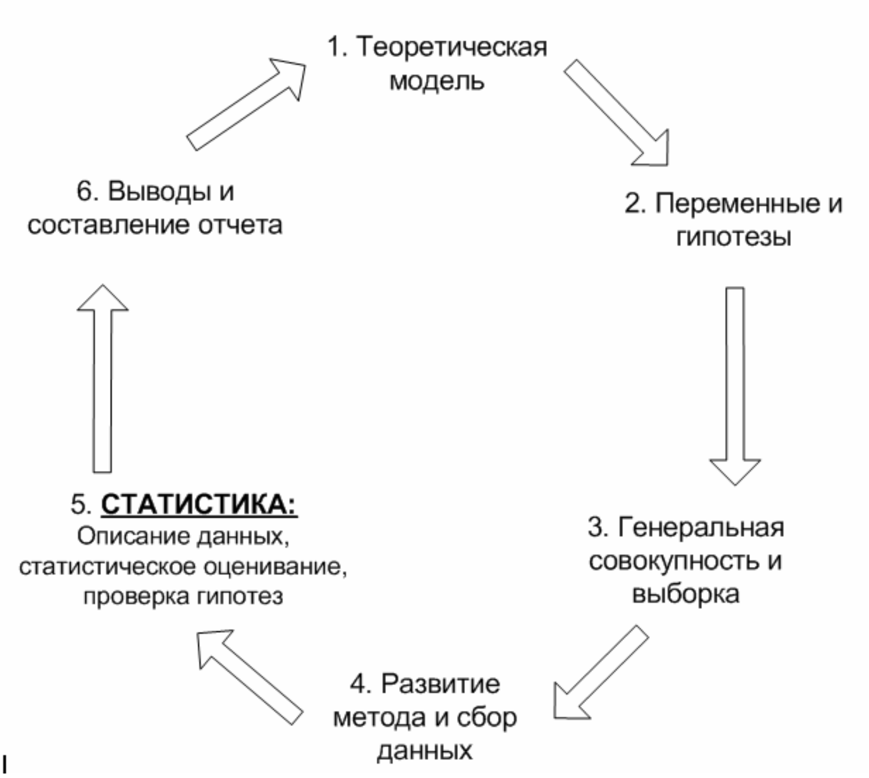


Рисунок 1-4. Роль статистики в проведении исследований

На рисунке 1-4 показана в некотором смысле «идеальная» последовательность проведения исследования. Ученые, занимающиеся исследованиями на протяжении многих лет, скажут, что на самом деле все значительно сложнее.

Тем не менее, рисунок дает представление о том, что статистика является частью общего процесса исследования. Ее роль может показаться незначительной, хотя это не так, поскольку именно статистика позволяет ответить на вопросы, должны ли мы принять или отклонить проверяемую гипотезу, является ли суждение, возникшее при рассмотрении теории, вероятным. Статистика служит важным инструментом развития новых теорий, позволяет определить направление для новых теоретических разработок.

Резюме

Цель этого раздела состояла в том, чтобы дать представление о предмете статистики. Статистика определена как совокупность методов для получения числовых описаний. В первой части курса речь пойдет об описательных методах статистики и теоретико-вероятностном подходе. В следующей части мы будем заниматься аналитической статистикой - методами, которые позволяют делать выводы обо всей генеральной совокупности на основе результатов изучения выборки.

Обсуждена роль и значение статистики в процессе проведения исследований. Ее роль может показаться малозначимой, но это не так, поскольку именно статистика позволяет ответить на вопросы, должны ли мы принять или отклонить проверяемую гипотезу, является ли суждение, возникшее при рассмотрении теории, вероятным. В следующем параграфе дается представление об измерениях и шкалах, применяемых в социальных исследованиях.

1-2 Измерения

Научный прогресс зависит от способности ученых проводить измерения, относящиеся к предмету исследования.

Измерение (measurement) означает присвоение чисел характеристикам изучаемых объектов, явлений согласно некоторому правилу.

Измеряется не сам объект, а его отдельные параметры, характеристики. Мы измеряем не респондентов самих по себе, а их возраст, пол, предпочтения и т.д. Обсуждение того, как наблюдаемым объектам или явлениям присваиваются числа, заслуживает серьезного внимания.

Шкала (scale) есть правило или алгоритм, в соответствии с которым изучаемым объектам, явлениям присваиваются числа.

Слово «шкала» обычно ассоциируется с измерительными приборами – часами, термометрами, спидометрами. Мы привыкли использовать эти приборы, когда хотим узнать время, температуру, время. Эти измерения являются настолько привычными, что мы не задумываемся над устройством этих приборов и просто используем их, когда нам необходимо. Со шкалами, о которых пойдет речь в нашем курсе, все обстоит иначе. Нам предстоит каждый раз создавать соответствующий измерительный прибор, задумываться о том, что мы собираемся измерить, какие деления будет иметь наша шкала, какие измерения нас удовлетворят и будут корректными.

Два типа данных

Данные, собираемые в ходе анкетных опросов, интервью или другими способами, есть результаты наблюдений, которые должны быть зафиксированы. Вернемся к примеру описания сообщества из 5 тысяч жителей. Предположим, в выборке из 250 человек мы выясняем социальные характеристики респондентов, включая возраст, пол, образование, семейное положение.

Таблица 1-1 Данные в табличной форме

РЕСПОНДЕНТ	ВОЗРАСТ	ПОЛ	ОБРАЗОВАНИЕ	СЕМЕЙНОЕ ПОЛОЖЕНИЕ
1	29	0	12	2
2	23	1	14	1
3	37	1	16	2
4	46	0	10	4
5	34	1	14	1

Занесем полученные данные в лист наблюдений. Для каждого респондента имеется одна строка, в которой каждая переменная получает определенное значение в соответствии с данными опроса.

Чтобы анализировать данные при помощи компьютера, нам следует перевести их в формат, совместимый с программным обеспечением, которое будет использовано для анализа. Такой формат обычно состоит из таблицы, в которой строки соответствуют респондентам, или случаям, а столбцы содержат значения переменных, как показано в Таблице 1-1. Первый столбец содержит нумерацию респондентов – от 1 и далее. Переменные начинаются со второго столбца: возраст, пол, образование и семейное положение. Например, респондент 1 - 29 лет, мужчина (0 = мужчина, 1 = женщина), имеет 12 лет образования, и так далее. Обратите внимание, что значения переменных представлены в виде чисел. Каждое число есть код ответа. Возраст представлен количеством лет. Пол представлен нулями и единицами, где 0 = мужчина и 1 = женщина.

Семейное положение закодировано так: 1 - холост, 2- женат, 3 - разведен и так далее. Компьютер тогда подсчитает число единиц, двоек, троек и сообщит о количестве случаев, в которых переменная принимает конкретное значение.

Данные, включаемые в таблицы, как правило, являются *числовыми* данными. Если значения переменных все-таки состоят из букв, их называют *текстовыми*. Мы могли использовать буквы: М = мужчина и Ж = женщина. Числовые данные могут быть дискретными и непрерывными.

Дискретные данные (discrete data) представляют собой отдельные значения признака, общее число которых конечно либо если бесконечно, то является счетным, т.е. может быть подсчитано натуральными числами от одного до бесконечности.

Примером дискретных данных является стаж работы на предприятии или продолжительность рабочего дня в минутах. В этом случае данные могут принимать различные значения, но количество вариантов конечно. Если речь идет о времени выполнения хирургической операции, то оно может быть измерено с высокой точностью, данные наблюдений перестают быть дискретными и становятся непрерывными.

Непрерывные данные (continuous data) могут принимать любое значение в некотором интервале.

Примером непрерывных данных является длина, вес, время и другие характеристики, измерение которых предполагает очень высокую точность. В социальных исследованиях мы будем редко встречать непрерывные данные. Мы чаще будем склонны считать любые данные дискретными и намеренно ограничивать наши возможности по измерениям с высокой точностью, поскольку это не будет иметь особого смысла.

Четыре критерия измерений

Как мы уже отметили, не всякие измерения бывают корректными. Поэтому уделим некоторое внимание обсуждению критериев, которые должны быть соблюдены, чтобы результатам измерения можно было доверять. Таких критериев мы назовем четыре: надежность, достоверность, завершенность и единственность.

1. Критерий надежности требует, чтобы при повторном проведении измерения в тех же самых условиях мы получили те же самые результаты. Например, вы используете напольные весы для определения собственного веса и считаете эти измерения надежными, поскольку если встать на весы повторно, вы получите тот же вес, что и в первый раз (мы пользуемся исправными весами и небольшие отклонения не учитываем). Это условие также называют воспроизводимостью.

Надежность измерения (reliability) означает возможность получить согласующиеся результаты при повторных измерениях характеристик объекта.

В научной деятельности проблема надежности измерений стоит очень остро. Более того, если измерение веса или роста легко воспроизвести, куда сложнее с измерениями в обществе. Социологи проводят измерения, связанные с такими сложными понятиями, как социальный класс, удовлетворенность социальным обеспечением и т.д. Понятие социального класса нетрудно использовать в устной беседе, но отнюдь не просто подобрать схему надежных измерений, связанных с этим понятием. Потребуется выработать точные критерии для установления границ социального класса. Результат измерения – принадлежность определенной категории людей к определенному классу - не должен зависеть от того, кем это измерение проводится и в каких условиях.

Критерий надежности требует, чтобы одни и те же люди были отнесены к одному и тому же социальному классу независимо от того, кто проводит измерение, и обстановки, в которой происходят измерения.

Социология часто использует измерения, основанные на ответах, которые люди дают в ходе анкетного опроса. Проблема надежности – даст ли респондент тот же самый ответ на вопрос, если повторить опрос. Возможно, ответы респондентов зависят от времени дня, событий, которые непосредственно предшествовали опросу, условий, в которых заполнялась анкета. Например, подростки могли бы иначе ответить на вопросы об отношении к спиртному, если бы рядом находились педагоги или родители.

Социологам хорошо известны эти проблемы, и они прилагают серьезные усилия, чтобы стандартизировать свои измерения, подробно исследовать проблемы надежности измерений.

2. Критерий достоверности. Если надежность имеет дело с воспроизводимостью, достоверность измерения означает, что мы действительно измеряем то, что нам нужно; любые измерения имеют силу только по отношению к определенным целям. Иначе обоснованность называют валидностью (англ. - validity). Показания уличного термометра не позволяют узнать температуру внутри помещения. С другой стороны, внутренний термометр сообщит нам лишь температуру, и мы не можем судить по нему о микроклимате или уровне комфорта внутри этого помещения, поскольку комфорт зависит от множества других факторов, таких как обстановка, непривычные запахи, присутствие других людей.

Достоверность измерения (validity) означает соответствие между результатами измерения и его целями, между выбранной шкалой и исследуемыми переменными.

При проведении анкетного опроса или интервью важно – действительно ли респонденты отвечают на тот вопрос, который подразумевает исследователь? Точность ответов зависит от формулировки вопроса

Достоверность измерений в социологии зависит от многих причин, включая навыки построения анкет, личный опыт исследователя в предметной области, его характер. Кроме всего, некоторые люди отказываются посвящать других в собственные мысли или обсуждать их поведение с посторонними. Иногда респондентов подводит память. В других случаях сложности связаны с получением информации о нетрадиционном или противозаконном поведении.

Точность проводимых измерений является тем фундаментом, на котором основывается любое научное исследование. Наши способности лучше понимать человеческое поведение зависят от наших возможностей проводить все более точные измерения.

3. Критерий завершенности означает, что для любого измеряемого объекта после проведения измерения мы сможем получить результат. Весы, имеющие верхнее ограничение в 90 килограммов не подойдут, поскольку не смогут указать вес более тяжелого человека. Для любого объекта, который может быть измерен, должно найтись некоторое, соответствующее ему значение.

Завершенность измерения (exhaustive) означает, что в результате измерения мы должны получить какой-либо результат.

На вопрос: «Кого бы вы избрали президентом?» могут быть предложены несколько фамилий. Это измерение не будет завершенным, если не добавить среди ответов «Другой кандидат» и «Никого». Следует понимать также, что если в результате опроса ответ «Никого» соберет большое количество голосов, такой результат окажется малоинформативным для исследователя.

4. Критерий единственности означает, что для любого исследуемого объекта в результате измерения мы получим единственное значение. Вряд ли нас устроит, если на вопрос о возрасте мы получим в качестве ответа два разных возраста респондента одновременно.

Единственность измерения (mutually exclusive) означает, что в результате измерения мы получим только одно значение переменной.

Следует отличать требования этого критерия от ситуаций, в которых на один вопрос респондент может выбрать одновременно несколько ответов из предложенных десяти вариантов. Соблюдается ли в этом случае требование единственности измерений? Да, соблюдается. Можно считать, что такой вопрос состоит из десяти вопросов, на каждый из которых респондент дает ответ. Такой переход от одного вопроса с десятью вариантами ответа к десяти переменным с двумя вариантами ответа приходится делать любому исследователю, и тем самым, соблюдать требование единственности измерения.

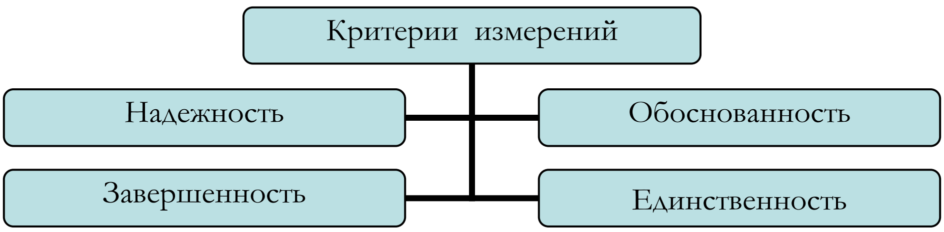


Рисунок 1-5. Четыре критерия измерений

Пять типов шкал

Данные, представляющие собой результаты измерений, бывают различных типов, в зависимости от типа рассматриваемых переменных. Разберем несколько принципиально разных типов данных и соответствующих им шкал.

Номинальная шкала (nominal scale) состоит из названий, имен или категорий для сортировки или классификации объектов, явлений по некоторому признаку. Результаты измерений, полученные при помощи номинальной шкалы, не могут быть упорядочены и с ними не могут производиться арифметические операции.

Примером номинальной шкалы служит часто используемая в анкетах шкала с тремя вариантами ответов: «Да» - «Нет» - «Не знаю». Другими примерами данных, измеряемых номинальными шкалами, являются пол, семейное положение, профессия, ученая степень. Предположим, вы строите шкалу профессий. Чтобы быть завершенной, она должна включать все возможные варианты профессий потенциальных респондентов, либо категорию «другая профессия» для тех респондентов, кто не найдет своего варианта в приведенном перечне. Если среди респондентов может оказаться человек, имеющий две профессии, шкала должна содержать категорию «две профессии», иначе шкала не будет удовлетворять критерию единственности.

Наше определение измерения указывает, что мы присваиваем объектам числовые значения, а не категории. Номинальная шкала может состоять из чисел. Например, социолог = 1, экономист = 2, юрист = 3 и далее. Тем не менее, любые действия с этими числами или их интерпретации оказываются бессмысленными, числа являются только ярлыками. Мы можем сказать только, что 1 отличается от 3 и не более того. Числа используются в номинальных шкалах, как правило, для кодировки данных и последующего компьютерного анализа.

Порядковая шкала (ordinal scale) означает, что числа присваиваются объектам, чтобы обозначить относительные позиции объектов, но не величину различий между ними.

Результаты измерений, полученные при помощи порядковой шкалы, могут быть определенным образом упорядочены, однако не могут указать на величину разницы между ними. Примером порядковой шкалы является общеизвестная классификация категории отелей. Когда вам сообщают, что гостиница имеет уровень «три звезды», вы понимаете, что гостиница, имеющая «четыре звезды» была бы лучше. Тем не менее, на сколько она была бы лучше, сказать не возможно. Это отличие не измеряется количественно. Порядковые номера не сообщают нам о величине различий.

Другой пример – выстраивание нескольких людей по росту и назначение им порядковых номеров. Как только номера присвоены, вы знаете, что первый – самый высокий, а четвертый по росту ниже третьего, но выше пятого. Однако в этом случае не известно, на сколько именно выше – сравниваются только порядки, а не сами значения роста, измеренные в сантиметрах.

В отличие от чисел, используемых в номинальной шкале, числа, используемые в порядковой шкале больше, чем просто ярлыки. Они позволяют сравнивать объекты и говорить, что у одного значение признака больше, чем у другого. Ограниченность порядковой шкалы состоит в том, что числа используются исключительно для упорядочения от меньшего к большему или наоборот, но не для сравнения количественных измерений.

Интервальная шкала (interval scale) позволяет находить разницу между двумя величинами. Обладает всеми свойствами номинальной и порядковой, но она позволяет указать количественное значение измеряемого признака. Недостатком служит отсутствие абсолютного нуля в качестве точки отсчета.

Пример - температура воздуха, измеренная в градусах. Когда на улице 20°C, это на 10°C ниже, чем если на улице 30°C. Нельзя сказать при этом, что это в полтора раза холоднее. Наличие 0°C еще не означает, что эту метку мы можем рассматривать в качестве привычного для нас нуля, относительно которого мы привыкли сравнивать числа и говорить, во сколько раз одно больше другого.

Второй пример – измерение времени в годах. Разница между 1962 и 2005 годами составляет 43 года. При этом, 2000 год не является в два раза большим по сравнению с 1000 годом.

Интервальная шкала состоит из интервалов одинаковой длины, называемых единицей измерения. Каждый единичный интервал может быть поделен на некоторое количество интервалов. Интервальные шкалы *делимы*. Шкала времени, например, может быть разделена на годы, каждый год разделен на дни, дни на часы и далее. Конечно, когда речь идет о возрасте, нас мало интересует измерение с точностью до секунд.

Тем не менее, для других целей мы можем рассматривать очень точные измерения. Результаты спортсменов измеряются до сотых долей секунды. Теоретически, интервальная шкала бесконечно делима, что позволяет увеличивать точность в зависимости от целей исследования.

Интервальная шкала является непрерывной, в то время, как номинальные и порядковые шкалы являются дискретными. Непрерывные шкалы позволяют производить точные измерения значений признака, с этими значениями можно проводить арифметические операции: складывать, вычитать, умножать и делить.

Таблица 1-2 Типы шкал

Шкала	Особенности	Пример
Номинальная	Содержит только категории, данные не могут упорядочиваться	Хобби студента. Только название.
Порядковая	Категории могут упорядочиваться, но разности не имеют смысла	Место на соревнованиях. Лучше результат – выше место.
Интервальная	Разности между значениями могут быть вычислены, но нет отношений	Температура студента. У больного выше на 1-2°С
Относительная	Имеется точка отсчета, возможны отношения между значениями	Рост студента. Один в 1,2 раза выше другого
Дихотомическая	Содержит две категории	Пол студента. Третьего не дано, если не рассматривать исключения.

Относительная шкала (ratio scale) обладает абсолютным нулем в качестве точки отсчета, что позволяет ей иметь все свойства интервальной шкалы. Для данных этой шкалы осмысленными являются все операции, включая вычитание и деление.

Цена товара измеряется относительной шкалой, в рублях или другой валюте. Можно говорить, что одна цена в два раза выше другой. Цена в ноль рублей означает, что товар бесплатный. Вес предмета также измеряется по шкале отношений. Ноль килограммов означает «нет веса». Бриллиант в 4 карата в два раза тяжелее бриллианта в 2 карата.

Интервальные и относительные шкалы называют *числовыми*. Числовые шкалы работают с непрерывными данными, в то время как номинальные и порядковые шкалы – с дискретными.

Еще один тип шкалы – дихотомическая. Дихотомической шкалой, как правило, измеряется пол человека. Вопрос с двумя вариантами ответа (Да = 1, Нет = 0) также представляет собой пример дихотомической шкалы.

Дихотомическая шкала (dichotomous scale) - номинальная шкала, которая состоит из двух категорий.

Дихотомическая шкала уникальна, поскольку к ней применимы некоторые арифметические операции. В первом примере, если мы имеем выборку из 10 людей, среди которых 2 - мужчины и 8 - женщины, мы можем сложить соответствующие нули и единицы и получить 8. Разделив на общее количество 10, получим 8/10, или 0,8. Значение 0,8 есть доля женщин в выборке. Получили простое и ясное описание соотношения пола в выборке.

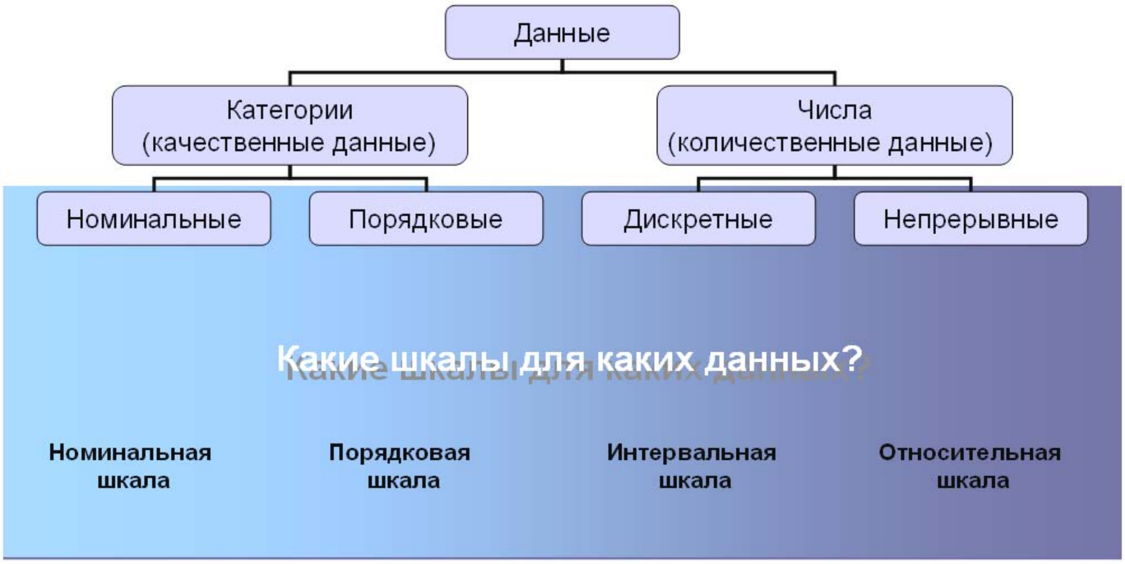


Рисунок 1-6. Какие шкалы применяются к разным типам данных?

Поскольку во многих случаях исследователям удавалось применять уникальные свойства дихотомической шкалы, ей уделяется особое внимание.

Резюме

Измерение необходимо для регистрации различий между объектами и явлениями окружающего мира. На элементарном уровне, шкалы измерений – это способ сортировки объектов с общими характеристиками. Измерения обязаны удовлетворять четырем критериям: (1) они должны быть надежны, т.е. позволять получить повторно один и тот же результат для одного и того же объекта при тех же условиях; (2) они должны быть достоверными, действительно измеряя то, что должно быть измерено; (3) шкала должна быть исчерпывающей, т.е. содержать все возможные значения признака; и (4) они должны удовлетворять критерию единственности, чтобы каждому объекту соответствовала только одна категория или значение.

Обсуждены пять типов шкал. Номинальные шкалы позволяют описывать наблюдения в терминах качественных признаков. Порядковые шкалы идут на один шаг дальше и позволяют упорядочивать наблюдаемые объекты от меньшего к большему по изучаемой характеристике. Номинальные и порядковые шкалы являются дискретными. Интервальные шкалы и шкалы отношений более сложны, состоят из интервалов равного размера, которые делают возможным определить количественное значение наблюдаемого признака. Дихотомические шкалы - уникальные номинальные шкалы, с которыми можно поступать так же, как с интервальными.

Что означают термины

Статистика	**Зависимая переменная**	**Завершенность измерения**
Генеральная совокупность	**Независимая переменная**	**Единственность измерения**
Выборка	**Описательная статистика**	**Шкала**
Репрезентативность выборки	**Аналитическая статистика**	**Номинальная шкала**
Данные	**Гипотеза**	**Порядковая шкала**
Дискретные данные	**Измерение**	**Интервальная шкала**
Непрерывные данные	**Надежность измерения**	**Относительная шкала**
Переменная, признак	**Достоверность измерения**	**Дихотомическая шкала**

Задачи и упражнения

1-1. Ответьте на вопросы по теме:

а) Что такое статистика?

б) Почему в статистике рассматривают выборку?

в) Опишите разницу между выборочной и генеральной совокупностью?

г) В чем различие между описательной статистикой и аналитической?

д) В чем различие между пятью различными видами шкал?

е) Приведите примеры «плохих» измерений, то есть не отвечающих хотя бы одному из четырех рассмотренных критериев измерений.

1-2. Описательная или аналитическая статистика? В следующих утверждениях укажите, где используется описательная, а где аналитическая статистика:

а) Средний возраст студентов в вашей группе равен 19,3 лет.

б) Средний возраст студентов факультета находится в пределах 19-20 лет.

в) На отделении «менеджмент» среди первокурсников 44% юношей и 56% девушек.

г) Обучение при помощи компьютера эффективнее, чем при прослушивании лекций.

д) Исследования перед выборами показывают, что действующий кандидат получит 63% голосов, а новый – 37%.

е) Существует связь между курением и риском заболевания рака легких.

ж) В соответствии с данными страховых компаний шанс любого человека дожить до 83 лет составляет 62,8%.

1-3. Дискретные или непрерывные переменные? В приведенных ниже примерах указать, какие переменные являются дискретными, а какие непрерывными:

а) Время, необходимое водителю, чтобы проехать определенную дистанцию.
б) Рост студента-первокурсника.
в) Рейтинг передач (плохо, средне, хорошо, отлично).
г) Зарплата кассиров крупных универмагов.
д) Семейное положение клиентов сберегательных банков.
е) Возраст студентов, записавшихся на военную кафедру.
ж) Температура внутри и вне самолета.
з) Вес новорожденного младенца.
и) Число книг на полке.
к) Километры, проезжаемые определенным автобусом в течение дня.

1-4. Качественные или количественные признаки? Укажите, какие из признаков, приведенных ниже, количественные:

а) Цвет автомобилей в автосалоне.
б) Число мест в кинотеатрах.
в) Длина кошек особых пород.
г) Число жалоб, полученных авиалинией за месяц.

1-5. Какой тип шкалы? В приведенных ниже примерах переменных, указать, шкалой какого типа измеряется значение этих переменных:

а) Температура воздуха в лекционной аудитории.
б) Возраст сотрудника.
в) Пол студента.
г) Семейное положение.
д) Место жительства.
е) Религиозные предпочтения.
ж) Время на подготовку домашнего задания.
з) Трудолюбие.

1-6. Генеральная совокупность и выборка. В следующих примерах указать исследуемую переменную (признак), границы генеральной совокупности и выборку:

а) Среди 200 случайно выбранных телезрителей 19% выключат телевизор в течение ближайших 15 минут.

б) 4 из 15 опрошенных читателей газеты поддержат кандидатуру нынешнего губернатора на очередных выборах.

в) Время подготовки к занятиям превышает 3 часа в день у половины студентов.

г) 48% выпускников университета работают по специальности.

1-7. Статистические заключения журналистов. По материалам газет и журналов приведите пример некорректного, с вашей точки зрения, статистического заключения. Объясните вашу точку зрения и укажите, как можно было бы избежать подобной ошибки.

1-8. Собственное исследование. Вам предстоит провести исследование по поводу отношения людей к смертной казни. Какие гипотезы вы можете предложить? Какова генеральная совокупность и как вы предполагаете делать выборку? Описательная или аналитическая статистика будет использоваться вами для получения выводов?

Глава 2. Представление данных

Цель этой главы состоит в том, чтобы познакомить читателя со способами обработки и представления результатов наблюдений. В реальных социологических исследованиях обычно выборка бывает большой и может включать несколько тысяч респондентов. Правильно распорядиться с таким объемом статистического материала возможно лишь при правильной организации работы и хорошем знании способов представления результатов наблюдений. Мы изучим понятие распределения частот, научимся вычислять относительные частоты, использовать таблицы и графики для представления данных.

2-1 Распределения частот

Статистики работают с эмпирическими данными. Если значение признака повторяется несколько раз, можно говорить о частоте появления определенного значения признака.

Частота (frequency) – количество наблюдений, в которых признак принимает определенное значение или находится в определенном интервале.

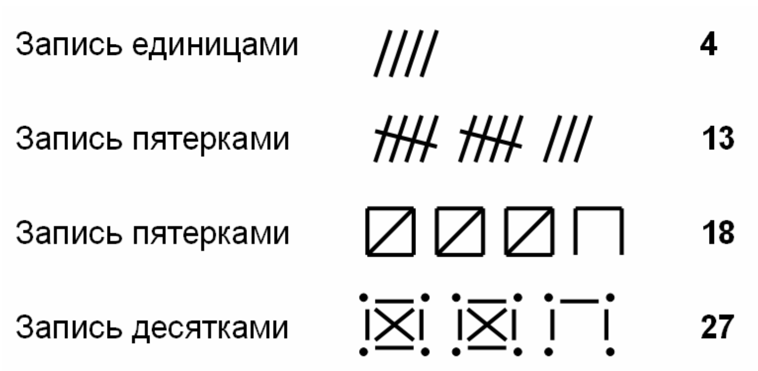
Распределение частот (frequency distribution) показывает частоты во взаимосвязи с результатами наблюдений.

Как мы обсудили в предыдущей главе, будем различать два случая: первый, когда признак измеряется номинальной или порядковой шкалой, принимает дискретные значения (будем называть их категории), и второй, когда значения признака измеряются с помощью числовой шкалы и непрерывны. В первом случае будем называть распределение частот *категориальным,* во втором будем строить *интервальное распределение.* Как позже будет объяснено, интервальное распределение можно строить также и для дискретного признака, если его значения являются редкими и могут быть объединены в несколько интервалов.

Подсчет частот может оказаться отнюдь не простым занятием. Представьте, что вам предстоит опросить более ста респондентов, а затем данные опроса представить в виде таблицы с распределением частот. Сначала ваши записи будут выглядеть следующим образом:

Номер респондента	Политические предпочтения
1	демократы
2	либералы
3	либералы
4	коммунисты
5	демократы
6	демократы
и так далее	

Предположим, 105 респондентов попадут в ваш длинный список, а ваши записи займут несколько страниц. Теперь предстоит обработка. Издавна известно несколько графических способов удобного ручного подсчета частот: пятерками, десятками и т.п. Выглядит это следующим образом:



При выборе одного из приведенных выше способов записи после получения ответа от очередного респондента нужно проставлять соответствующие метки, а затем сосчитать их общее число. Искусство ручного подсчета и обработки данных было распространено совсем недавно – до появления вычислительной техники и ее массового использования. Уметь пользоваться ручной записью оказывается полезным и до наших дней, особенно «в полевых условиях».

Категориальное распределение частот

Что мы хотим получить в результате обработки данных? Если в случайной выборке из 105 респондентов окажется, что 45 из них отдали политические предпочтения демократам - это и есть *частота* признака в выборке. Наблюдаемым признаком, или переменной, являются политические предпочтения избирателей.

Таблица 2-1 Политические предпочтения

КАТЕГОРИИ	f
Демократы	45
Коммунисты	43
Либералы	15
Всего	105

Измерения имеют номинальную шкалу с тремя категориями: демократы, коммунисты и либералы. Распределение частот для выборки 105 избирателей показано в таблице 2-1. Символ f обозначает частоты, n - общее количество наблюдений. Обратите внимание, что сумма частот должна равняться n.

Построенная нами таблица представляет собой категориальное распределение частот для политических предпочтений 105 опрошенных респондентов. Дадим также общее определение для категориального распределения частот признака.

Категориальное распределение частот состоит из категорий, являющихся значениями исследуемого признака, и соответствующих этим категориям частот. Категориальное распределение строится для данных, измеряемых номинальной или порядковой шкалой.

Для порядковых шкал распределение частот строится аналогично. В первом столбце располагаются возможные значения признака в порядке возрастания. Во втором – частоты, соответствующие каждому значению.

В таблице 2-2 приведен пример распределения частот по выборке 60 зрителей. Измерялось их отношение к просмотренному фильму по порядковой шкале из пяти значений. Результаты показывают, что 36 зрителям фильм понравился или очень понравился.

Таблица 2-2 Отношение к фильму

КАТЕГОРИИ	f
Очень понравился	24
Понравился	12
Фильм средний	10
Не понравился	6
Очень плохой	8
Всего	60

Таблица 2-3 Вес тела (упорядоченные данные)

(N = 77)						
46	59	65	69	71	74	79
49	60	65	69	71	75	80
50	60	65	69	72	75	81
50	60	66	70	72	75	81
52	61	67	70	73	76	83
53	62	67	70	73	76	84
54	62	67	70	73	77	84
55	63	68	70	73	77	85
55	64	68	71	74	78	87
56	64	68	71	74	79	89
58	64	69	71	74	79	90

Интервальное распределение частот

Для количественных шкал количество вариантов потенциально может быть очень большим, поскольку количественные шкалы могут включать непрерывное множество значений и признак может принимать уникальное значение для каждого наблюдения. Следовательно, представление распределения переменной, измеряемой количественной шкалой, требует иного подхода.

Представим себе, что результаты измерений записаны в виде последовательности чисел, являющихся значениями признака, измеренными для каждого отдельного объекта. Такие данные могут быть *неупорядоченными* или *упорядоченными*.

Неупорядоченные данные состоят из результатов, которые записаны в произвольном порядке, например в том, в котором были получены. **Упорядоченные данные** содержат результаты наблюдений, записанные в порядке возрастания (либо в порядке убывания).

В Таблице 2-3 представлены упорядоченные данные измерения веса произвольно выбранных 77 человек в килограммах. Недостатком такого представления является его слабая пригодность для последующего изучения выборки.

Построим другое, более удобное представление. Все наши наблюдения находятся в интервале от наименьшего 46 до наибольшего 90. Сгруппируем наши наблюдения и построим таблицу интервального распределения частот.

Интервальное распределение частот состоит из некоторого количества интервалов равной длины, на которые делится весь диапазон изменения признака, и соответствующих этим интервалам частот.

Таблица 2-4 Интервальное распределение частот

ИНТЕРВАЛЫ	*f*
45 - 49	2
50 - 54	5
55 - 59	5
60 - 64	10
65 - 69	14
70 - 74	20
75 - 79	11
80 - 84	6
85 - 89	3
90 - 94	1
Итого	77

Данные о весе 77 человек теперь преобразованы в таблицу по нескольким интервалам. Правый столбец, имеющий *f* в качестве заголовка, содержит количество людей, вес которых находится в соответствующем интервале. Например, вес 18 человек находится между 150 и 159 кг, а вес 15 человек - между 160 и 169 кг.

Существует разница между упорядоченными данными и интервальным распределением частот. Можно сказать, что в результате преобразования мы утратили часть информации, поскольку мы не имеем теперь 77 точных измерений, а получили лишь частоты попадания результатов наблюдений в определенный интервал. Нам пришлось пожертвовать деталями ради получения знаний о виде распределения. Очевидно, что даже неискушенным наблюдателям таблица с частотами больше говорит об особенностях веса людей в выборке, нежели чем просто упорядоченный набор наблюдений.

Построение интервального распределения

При построении интервального распределения требуется соблюдение несколько условий.

Условие 1. Интервалы не должны пересекаться, чтобы одно значение не могло попасть в два разных интервала.

Условие 2. Интервалы должны охватывать все возможные значения признака.

Условие 3. Интервалы должны иметь одинаковую длину, чтобы не искажать вид распределения данных. Если в крайние интервалы попадает небольшое количество значений признака, они могут быть объединены с соседними и только в этом случае интервалы будут иметь двойную длину по сравнению с обычными.

Таблица 2-5 Объявленные и точные границы интервалов веса тела

ОБЪЯВЛЕННЫЕ ГРАНИЦЫ	ТОЧНЫЕ ГРАНИЦЫ	f
45 – 49	менее 49,5	2
50 – 54	49,5–54,5	5
55 – 59	54,5–59,5	5
60 – 64	59,5–64,5	10
65 – 69	64,5–69,5	14
70 – 74	69,5–74,5	20
75 – 79	74,5–79,5	11
80 – 84	79,5–84,5	6
85 – 89	84,5–89,5	3
90 – 94	89,5 и более	1
	Итого	77

Условие 4. Интервалы не должны иметь пробелов. Если в какой-то интервал не попало ни одного значения, его не следует исключать из рассмотрения. Исключение составляют первый и последний интервалы. Они могут быть исключены без ущерба для вида распределения.

Условие 5 (не обязательное). Предпочтительно также, чтобы *середины интервалов были целым числом*. Это удобно для последующих вычислений и построения графиков. Кроме этого, обычно *выбирается от 5 до 20 интервалов*, другое количество не будет ошибкой, но существует такое негласное соглашение.

Если проверить вышеприведенные условия, таблица 2-4 не окажется безукоризненной. В частности, если человек имеет вес 79,5 кг, то возникает вопрос, в каком интервале находится это значение. Объявленные нами границы не являются *точными*.

В таблице 2-5 вычислены точные границы интервалов уменьшением нижней границы и увеличением верхней границы объявленных интервалов на 0,5 кг. Точные границы, скажем, седьмого интервала получены следующим образом:

(нижняя граница) = 75 – 0,5 = 74,5

(верхняя граница) = 79 + 0,5 = 79,5

Для окончательного устранения двусмысленности следует сказать, что значение 74,5 попадает в интервал 74,5-79,5 (следуя правилу округления – 74,5 округляется до 75, а 74,4 до 74), и не попадает в интервал 69,5–74,5. Интервал 69,5-74,5 содержит все числа большие или равные 69,5 и меньшие 74,5. Теперь мы полностью устранили пересечения интервалов, сделали их исчерпывающими, т.е. охватывающими все возможные значения. Для этого нам понадобилось расширить два крайних интервала.

Как уже говорилось, после перехода от упорядоченного массива к интервальному распределению частот, мы утратили информацию о каждом индивидуальном значении признака в выборке. Из таблицы 2.5 видно, что14 человек весят от 69,5 до 74,5 кг, но точный вес любого индивидуума при этом неизвестен. Лучшей оценкой веса каждого из этих людей служит середина интервала.

Середина интервала рассчитывается по следующей формуле: (середина интервала) = (начало интервала) + (длина интервала)/2. Для примера, применительно к интервалу 69,5–74,5, (середина интервала (69,5–74,5)) = 69,5 + 5/2 = 69,5 + 2,5 = 72. Удачно, что середина есть целое число.

Если бы мы выберем длину интервала 20 кг, то получим всего 3 интервала, что даст малоинформативное представление. При длине в 2 кг мы, наоборот, получим 30 интервалов и небольшие частоты, что не даст нам достаточно ясной картины. Выбор интервала длиной в 5 кг позволил построить 10 интервалов и получить вполне пригодное для анализа распределение частот.

Рассмотрим следующий пример. Имеются данные о продолжительности 20 разговоров по телефону в минутах. Получите распределение частот, используя 7 интервалов.

11	29	6	33	14
21	18	17	22	38
31	22	27	19	22
23	26	39	34	27

Решение. Построение интервального распределения частот представим по шагам.

ШАГ 1. Найдем самое большое и самое маленькое значения: 39 и 6.

ШАГ 2. Определим диапазон значений выборки: (диапазон) = (наибольшее значение) – (наименьшее значение) = 39 - 6 = 33.

ШАГ 3. Определим количество интервалов: (число интервалов) = 7.

ШАГ 4. Найдем длину интервала, разделив диапазон данных на количество интервалов: (длина интервала) = (диапазон) / (число интервалов) = 33 / 7 ≈ 4,7. Округлим до ближайшего целого: 4,7≈5.

Таблица 2-6 Вычисление интервального распределения частот

ИНТЕРВАЛ	ТОЧНЫЕ ГРАНИЦЫ	ЧАСТОТА
6-10	5,5-10,5	1
11-15	10,5-15,5	2
16-20	15,5-20,5	3
21-25	20,5-25,5	5
26-30	25,5-30,5	4
31-35	30,5-35,5	3
36-40	35,5-40,5	2

ШАГ 5. Нижней границей первого интервала является наименьшее значение 6. Прибавим 5 и получим нижнюю границу второго интервала. Продолжим прибавлять до тех пор, пока не получим 7 значений: 6, 11, 16, 21, 26, 31, 36 – нижние границы.

ШАГ 6. Вычтем единицу из нижних пределов, чтобы получить верхние пределы. Первый интервал 6-10, второй 11-15 и так далее.

ШАГ 7. Найдем точные границы интервалов, вычитая 0,5 от каждого нижнего предела и прибавляя 0,5 к верхнему пределу интервала.

ШАГ 8. Подсчитаем количество данных, попадающих в каждый из интервалов. Удобно вести подсчет графическими значками.

ШАГ 9. Запишем итоговые численные значения в столбце для частот.

Резюме

В этой части рассмотрены два типа распределения частот. Первый тип используется, когда данные – категориальные и называется категориальным распределением. Второй тип распределения используется, когда числовые данные необходимо группировать в интервалы. Он называется интервальным распределением частот. Оба типа распределений часто используются в статистических вычислениях и полезны для тех, кто описывает и представляет данные.

Как мы сможем убедиться, частотные распределения необходимы, чтобы представить данные правильным и понятным образом, визуально определить характер и форму распределения, построить таблицы (параграф 2-3), графики, диаграммы данных (параграф 2-4), упростить вычисление среднего значения и других параметров (параграф 3-1), сравнить различные данные между собой.

2-2 Относительные частоты, доли, проценты

Существуют общепринятые способы описания и использования распределения частот. Наиболее применимые – отношения частот, доли, относительные частоты, проценты, накопленные частоты и накопленные проценты. Любая форма представления частот связана с определенным сложившимся стандартом.

Отношения частот

Наиболее общий вид относительных частот есть обычное отношение двух чисел. Например, отношение числа студентов-социологов к числу студентов-экономистов:

$$\frac{\textit{Социологи}}{\textit{Экономисты}}$$

Если среди 60 студентов 40 социологов и 20 экономистов, то отношение социологов к экономистам есть 40/20 = 2,0, что означает, что есть два социолога на одного экономиста. Мы можем сосчитать также отношение экономистов к социологам: 20/40 = 0,5, и это означает, что имеется 0,5 экономистов на одного социолога.

Отношение частот (ratio) есть обычное деление одной частоты на другую и вычисление дроби.

Общая формула для отношения частот:

$$\textit{Отношение} = \frac{f_1}{f_2}$$

где f_1 = частота первого значения признака,
f_2 = частота второго значения признака,

Отношения очень полезны, поскольку они позволяют сравнивать различные измерения. Например, вы можете сосчитать количество студентов на одного преподавателя в вузе, размер платы за обучение к количеству аудиторных часов, количество автомобилей или мобильных телефонов в семье и так далее. Отношения можно вычислять для всех типов шкал - числовых, порядковых, номинальных, поскольку отношения основаны на количестве наблюдений с одним значением признака. Они являются мощным наглядным инструментом, потому что легко поняты и всегда под рукой. Далее мы часто будем использовать отношения в этой книге.

Доли и относительные частоты

Будем рассматривать долю как отношение некоторого подмножества частот к общей сумме частот. Мы можем записать это в виде следующей формулы:

$$Доля = \frac{Часть}{Целое}$$

Доля (proportion) есть отношение части к целому. Доля в распределении частот есть отношение одной из частот к общему количеству наблюдений, которое также принято называть **относительной частотой** значения признака.

Доля в распределении частот может быть выражена как

$$Доля = \frac{f_i}{n}$$

где f_i = одна из частот в распределении,
n = общее число наблюдений.

Доля социологов среди 60 студентов (социологов и экономистов) будет вычислена следующим образом:

$$\frac{Социологи}{Социологи + Экономисты}$$

В нашем примере, доля социологов в группе из 60 студентов есть 40/(40+20)=0,67. Доля экономистов есть 20/(40+20)=0,33. Заметим, что сумма 0,67 + 0,33 есть 1,00. Сумма долей распределения частот должна равняться единице. Вычисляя долю, мы получаем, что число социологов относительно общего числа студентов есть 0,67 и что число экономистов относительно общего количества – 0,33.

Мы можем рассматривать доли с еще одной точки зрения. Переход от первоначальных данных к долям может рассматриваться как *линейное преобразование* распределения частот. В нашем примере, первоначальное распределение социологов и экономистов было преобразовано от вида (40+20=60) к виду (0,67+0,33=1,00). Преобразование оказывается полезным, поскольку новый вид более понятен и нагляден, в особенности, если мы имеем дело с большими числами. Кроме того, преобразование позволяет нам приводить различные измерения к стандартному распределению значений, которое располагается от 0,00 до 1,00. Измерения, основанные на различных объемах выборки, можно сравнить, если преобразовать их в доли.

Таблица 2-7 Политические предпочтения

КАТЕГОРИИ	ЧАСТОТЫ f	ОТНОСИТЕЛЬНЫЕ ЧАСТОТЫ	ПРОЦЕНТЫ %
Демократы	45	0,429	43%
Коммунисты	43	0,410	41%
Либералы	15	0,143	14%
Всего	105	1,000	100%

Проценты

Проценты (percentages) есть доля, умноженная на 100%:

$$\textit{Проценты} = \frac{f_i}{n} \times 100$$

Проценты социологов и экономистов в группе 60 студентов: 40/60 × 100 = 66,7% и 20/60 × 100 = 33,3%. Сумма процентов всех частот распределения всегда равняется 100. Тем самым, сумма процентов социологов и экономистов есть 66,7 + 33,3 = 100. Преобразование частот распределения в проценты представляет собой другое линейное преобразование, которое переводит первоначальные частоты в новую числовую шкалу, располагающуюся от 0,00 до 100. Мы преобразовали первоначальное распределение частот 40 + 20 = 60 в 66,7 + 33,3 = 100. Преимущества использования процентов такие же, как использование отношений, их настолько часто используют повседневно, что они понимаются и применяются каждым.

Проценты в примере выше округлены до одного десятичного знака. Это общее правило позволяет сохранять данные в виде, подобном исходным измерениям. Первоначально мы говорили о социологах и экономистах в терминах индивидов, а затем в терминах процентов с одним десятичным знаком.

Возвратимся к таблице категориального распределения с политическими предпочтениями и продолжим ее двумя столбцами – третьим и четвертым (таблица 2.7). В третьем столбце мы вычисляем относительные частоты (доли) для каждого из значений признака. Для абсолютной частоты демократов 45, относительная частота вычисляется так: P = 45/105 = 0,429. Проценты вычисляются по формуле: 45/105 × 100% = 42,9%. До какого знака следует округлять? В данном случае это зависит от целей последующего использования результатов. Помните, что округление всегда приводит к ошибкам округления. В нашем примере, после округления сумма по четвертому столбцу не равна 100%, как бы этого нам не хотелось! Будьте внимательны!

Таблица 2-8 Частоты, доли и проценты результатов экзамена

КАТЕГОРИИ	f	Cf	P	CP	%	C%
Отл.	17	17	0,200	0,200	20,0%	20,0%
Хор.	41	58	0,482	0,682	48,2%	68,2%
Удовл.	20	78	0,236	0,918	23,6%	91,8%
Неуд.	7	85	0,082	1,000	8,2%	100,0%
ИТОГО	85		1,000		100,0%	

Обозначения столбцов: *f* = частота, C*f* = накопленная частота, P = относительная частота, CP = накопленная относительная частота, % = проценты, C% = накопленные проценты.

Накопленные частоты, относительные частоты и проценты

Отношения, доли и проценты могут применяться для всех типов шкал. Отношения и проценты, кроме этого, можно *накапливать*, чтобы показать число наблюдений выше или ниже любого выбранного значения распределения. Однако, такое накопление возможно только в интервальных и порядковых шкалах. Категории в номинальной шкале не могут быть упорядочены, и поэтому не имеет смысла думать в терминах «выше» или «ниже» для категорий номинальной шкалы.

Таблица 2-8 показывает результаты экзамена для 85 человек. Накопленная частота для оценки «Хорошо» получена сложением 17 + 41 = 58. Накопленная относительная частота для оценки «Хорошо» получена сложением 0,200 + 0,482 = 0,682. Накопленные проценты для оценки «Хорошо» получены сложением 20,0% + 48,2% = 68,2%.

Накопленные относительные частоты, CP, показывают, что 0,918 общего числа студентов получили оценки выше «Неудовлетворительно». Вычитая, мы получим: 1,000 – 0,682 = 0,318 - такая доля студентов получила менее 4 баллов.

Обратите внимание, что сумма столбца относительных частот P, равняется 1,00 и что накопленная частота CP, равна 1,00. Таким образом, первоначальное распределение частот *f*, включавшее 85 наблюдений, было преобразовано к новому распределению со значениями в пределах от 0,00 к 1,00. Тем самым, доли преобразовывают частоты распределения независимо от количества первоначальных наблюдений к относительному масштабу в пределах от 0,00 до 1,00, в то время как проценты преобразовывают их в масштаб от 0,00 до 100,00. Эти преобразования легко интерпретируются и позволяют сравнивать группы с различным количеством наблюдений.

Таблица 2-9 Результаты экзамена двух групп

	ЧАСТОТЫ	
СУММА БАЛЛОВ	ГРУППА 1	ГРУППА 2
100	10	47
95	20	94
90	30	141
85	30	141
80	20	94
75	10	47
ИТОГО	120	564

В Таблице 2-9 представлены данные результатов экзамена в двух различных группах студентов. С такими данными довольно сложно иметь дело, поскольку визуально трудно рассматривать и анализировать представленные частоты.

Составим другую таблицу. В Таблице 2-10 приведены также относительные частоты и проценты. Теперь можно сделать некоторые выводы. Очевидно, что распределения частот для обеих групп одинаковы. Преобразование первоначальных распределений частот к долям или процентам позволило нам увидеть, что нет никаких различий между группами.

Резюме

Относительные частоты, доли и проценты исключительно полезны, поскольку они представляют частоты в некотором стандартном, широко известном виде. Доли и проценты, к тому же, являются линейным преобразованием частот в стандартную форму. Далее мы опишем табличный способ представления данных.

Таблица 2-10 Частоты, доли и проценты результатов экзамена для двух групп

	ГРУППА 1			ГРУППА 2		
СУММА БАЛЛОВ	f	P	%	f	P	%
100	10	0,083	8,3	47	0,083	8,3
95	20	0,167	16,7	94	0,167	16,7
90	30	0,250	25	141	0,250	25
85	30	0,250	25	141	0,250	25
80	20	0,167	16,7	94	0,167	16,7
75	10	0,083	8,3	47	0,083	8,3
ИТОГО	120	1	100	564	1	100

2-3 Таблицы

Таблицы являются удобной и широко используемой формой представления данных. Следует обсудить основные правила и соглашения, которым следует следовать при построении и использовании таблиц.

Стандартный вид таблицы

На рисунке 2.1 представлены компоненты стандартной таблицы. Таблица всегда имеет название, которое ясно сообщает о ее содержании. Название должно быть кратким и должно указывать переменные, которые содержатся в таблице. Если в документе имеется более одной таблицы, они нумеруются. Если документ содержит более одной главы, таблицы могут нумероваться двумя числами, например, 1-1, 1-2, и 2-1, где первое число соответствует главе, и второе число номеру таблицы в главе. Статистические таблицы, кроме того, обычно содержат информацию о числе наблюдений, данные о которых представлены в таблице. Число наблюдений указывается в самой таблице или ее названии. Основное поле (или тело) таблицы содержит интересующие данные наблюдений.

Рисунок 2-11 Стандартный вид таблицы

<table>
<tr><td colspan="2">Номер, Заголовок таблицы</td></tr>
<tr><td rowspan="2">Название строк</td><td>Название столбцов</td></tr>
<tr><td>Заголовки столбцов</td></tr>
<tr><td>Заголовки строк</td><td>Поле (тело) таблицы</td></tr>
</table>

Таблицы сопряженности

Различают способы построения таблиц для одной, двух или нескольких переменных. Таблицы для двух переменных содержат одну переменную в строках, а другую в столбцах. Таблицы для двух переменных называются *таблицами сопряженности,* они показывают связи или отношения между ними. Пример таблицы сопряженности приведен в таблице 2-12.

Если изучается связь между независимой и зависимой переменной, то зависимая чаще размещается в строках, а независимая в столбцах. Если в таблице рассчитываются проценты, то они размещаются по направлению независимой переменной.

Таблица 2-12 Вид деятельности и удовлетворенность оплатой труда

УДОВЛЕТВОРЕННОСТЬ РАЗМЕРОМ ОПЛАТЫ ТРУДА	ВИД ДЕЯТЕЛЬНОСТИ		ВСЕГО
	РАБОЧИЙ	СЛУЖАЩИЙ	
Низкая	35	11	46
Высокая	12	49	61
ВСЕГО	47	60	107

Таблица 2-13 показывает связь между местом проживания и предпочитаемой формой досуга. Тело таблицы показывает измерения выборки 469 респондентов из четырех различных городов. Место проживания рассматривается как независимая переменная, и считается, что форма досуга зависит от места проживания. Каждая ячейка таблицы показывает количество наблюдений, для которых совпало место проживания и форма досуга. Таблица показывает частоту и процент для каждой ячейки. Проценты рассчитаны для независимой переменной (в столбце), чтобы показать процент людей, предпочитающих ту или иную форму досуга, в зависимости от места проживания. Если сложить проценты в любом столбце, получим 100%.

Можно было еще рассчитать проценты в строках и процент для каждой ячейки по отношению к общему числу наблюдений. Однако, существует правило, что в таблице должно быть столько вычислений, сколько требуется для исследования. Следует избегать перегрузки данными.

Заметим, что мы вычислили проценты и для строки итогов по столбцам. Это показывает нам, как выборка 469 человек распределена по отношению к формам досуга, или какова доля жителей конкретного города в выборке. Мы видим, например, что 44,8% предпочитают спорт, 41,4% - автомобили, и 13,8% - компьютеры. Мы также знаем, что 34,3% - проживают в Москве, 33,5% - Курске, 19% - Самаре, и 13,2 % - Волгограде.

Таблица 2-13 Форма досуга и место проживания

(N=169)					
ФОРМА ДОСУГА	МЕСТО ПРОЖИВАНИЯ				ВСЕГО ПО СТРОКЕ
	МОСКВА	КУРСК	САМАРА	ВОЛГОГРАД	
Спорт	94 58,4%	47 29,9%	43 48,3%	26 41,9%	210 44,8%
Автомобили	49 30,4%	93 59,2%	21 23,6%	31 50,0%	194 41,4%
Компьютер	18 11,2%	17 10,8%	25 28,1%	5 8,1%	65 13,9%
Всего по столбцу	161 34,3%	157 33,5%	89 19,0%	62 13,2%	469 100,0%

2-4 Визуализация данных

Важный инструмент представления числовых данных - использование графических изображений. Следующий параграф рассматривает некоторые графические методы. Мы рассмотрим гистограммы, полигоны частот, кумуляты.

Гистограммы частот

Гистограмма частот есть графическое представление, в котором по оси X откладываются значения переменной, а по оси Y соответствующие им частоты. Гистограмма строится в виде прямоугольников, высота которых соответствует частоте значения переменной.

Построим гистограмму частот для примера с телефонными звонками. Используем для этого результаты, полученные нами в таблице 2-6 распределения частот. По оси X откладываем соответствующие интервалы: 6-10, 11-15 и т.д., всего семь. В интервале 6-10 строим прямоугольник высотой 1, что означает, что лишь один звонок из 20 попал по продолжительности в первый интервал, во втором интервале – прямоугольник высотой 2 и так далее. Гистограмма частот полезна для визуальной оценки данных. Можно отметить, что наиболее популярное время: 21-25 минут, в этот интервал попало 5 звонков, это больше, чем в любой другой интервал.

Рисунок 2-3 Гистограмма частот для телефонных звонков

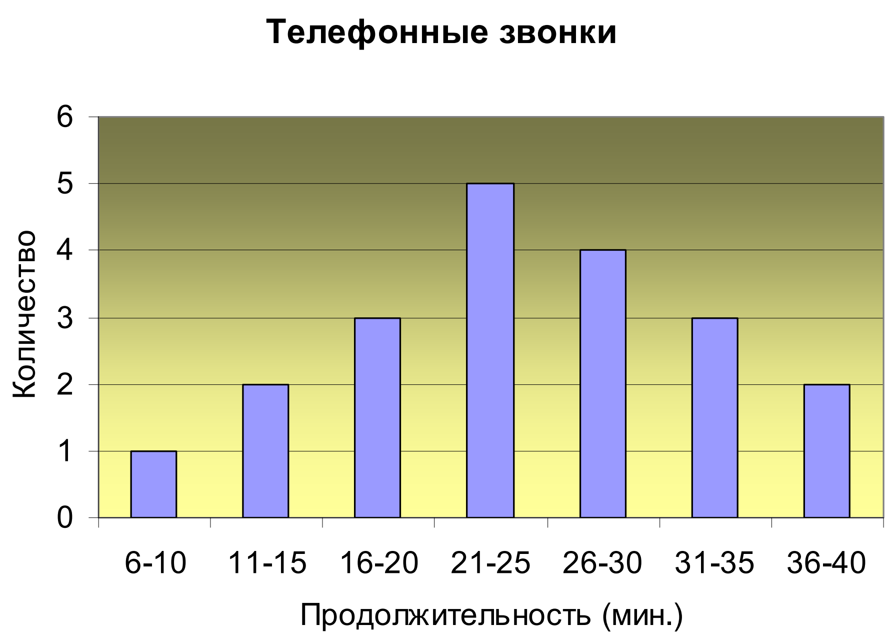


Рисунок 2-4 Гистограмма относительных частот

Гистограмма относительных частот есть графическое представление, в котором по оси X откладываются значения переменной, а по оси Y соответствующие им доли или проценты. Гистограмма строится в виде прямоугольников, высота которых соответствует доле или процентам для значения переменной.

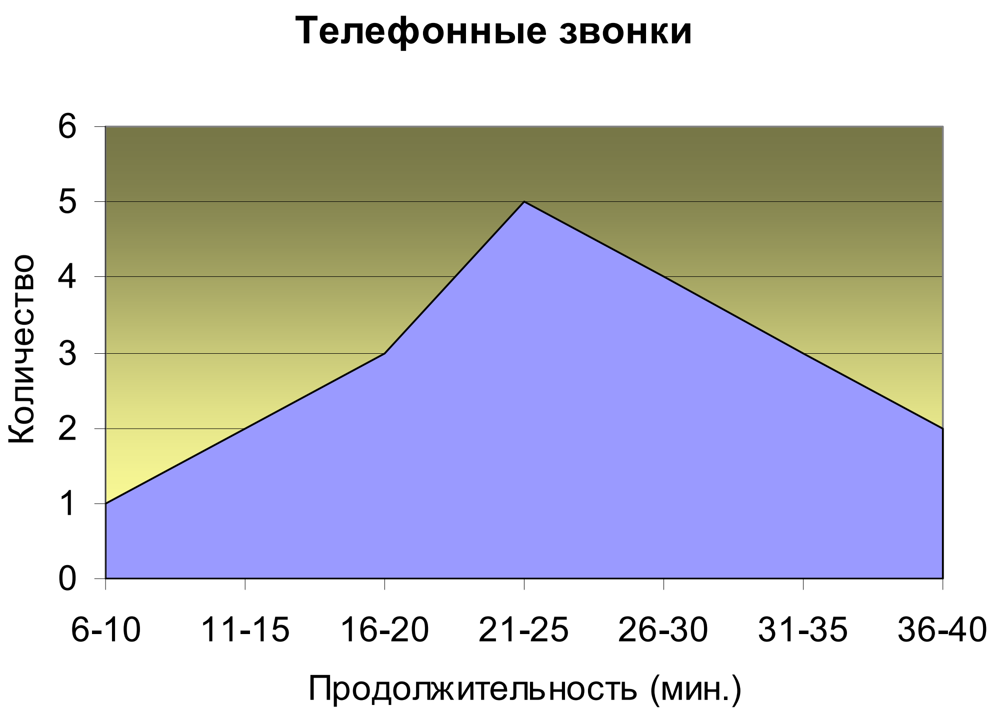
Гистограмма относительных частот строится таким образом, что вместо частот на оси Y отмечены доли или проценты для соответствующих значений переменной. Для рассмотренного нами примера гистограмма относительных частот построена на рисунке 2-4. Обе гистограммы идентичны, отличие состоит лишь в другой шкале по оси Y.

По гистограмме можно сказать, например, что 0,20 всех звонков попало по продолжительности в интервал 26-30 минут (или 20%).

Полигоны частот

Полигон частот строится подобно гистограмме. Вместо прямоугольников в полигоне строится линия по точкам, соответствующим серединам интервалов и частотам. Полигон дает зрительное представление о распределении частот, которое сильно отличается от гистограммы при одних и тех же данных. Не существует правил, которые говорили бы, какое из представлений лучше. Все зависит от конкретной ситуации и от вкусов исследователя, который выбирает вид представления из существующих альтернатив.

Рисунок 2-5 Полигон частот для телефонных звонков



Полигон есть графическое представление, в котором по оси X откладываются значения переменной, а по оси Y соответствующие им частоты. Полигон строится в виде области, ограниченной линией, проходящей по точкам (середина интервала, частота).

Полигоны, также как и гистограммы, можно строить для частот, относительных частот и процентов. Во всех трех случаях график останется прежним, за исключением оси Y. На рисунке 2.5 полигон построен для частот.

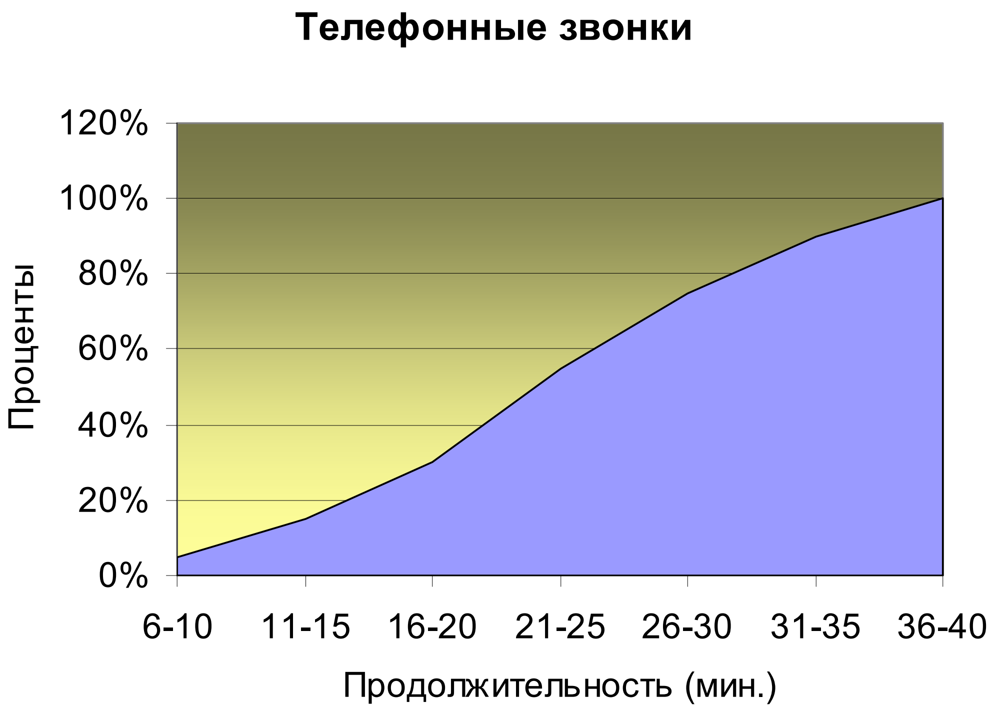
Кумуляты

Еще одним часто используемым графическим представлением данных является кумулята – от слова «кумулятивный», что означает накапливаемый. В этом случае по оси Y откладываются накопленные частоты или накопленные проценты.

На рисунке 2.6 график построен для накопленных процентов. По графику можно сказать, например, что около 70% звонков имеет продолжительность менее 30 мин, зато менее 20 минут разговаривают всего 30%.

Кумуляты могут быть построены как по возрастанию значений признака, так и по убыванию, в зависимости от того, какой анализ мы предполагаем проводить с использованием полученного графического изображения.

Рисунок 2-6 Кумулята частот для телефонных звонков



Анализ визуальных представлений

Мы не ставили целью назвать все возможные графические способы представления данных. Общеизвестны круговые диаграммы, линейные графики и много других. Со спектром возможностей можно познакомиться в любой компьютерной программе, занимающейся обработкой и графическим представлением данных. Мы остановились лишь на тех, которые часто используются для визуального анализа данных с точки зрения их распределения. В частности, они позволяют ответить на следующие вопросы:

- Какие значения являются минимальным и максимальным? Каков размах имеющихся данных?
- Какие значения встречаются в наборе данных чаще других? Какие значения являются наиболее типичными?
- Какова стандартная разница между значениями в наборе данных?
- Какой вид имеет распределение? Где сосредоточена основная часть данных? Симметрично ли они расположены вокруг типичного значения? В какую сторону смещены?
- Имеются ли характерные особенности? Выбросы? Есть ли значения признака, которые пропущены?

Используем компьютер

Эта глава предоставляет большой материал для компьютерных упражнений. Применение программного обеспечения к материалу главы поможет, одновременно, научиться использовать компьютер для представления данных и глубже понять содержание главы. Первое задание с использованием компьютера это организация и ввод данных. Продумайте список переменных, с которыми будете работать, определите их возможные значения. Затем создайте таблицы частот, включите в них групповые частоты. Сосчитайте абсолютные значения и проценты, накопленные частоты и накопленные проценты.

Статистический пакет, вероятно, содержит опции для создания различных типов диаграмм. Вы можете создать диаграммы, соответствующие интервальным, порядковым, и номинальным данным. Эти начальные шаги по применению статистического пакета очень важны для его последующего использования для решения более сложных задач.

Что означают термины

Частота	**Отношения и доли**	**Накопленные проценты**
Распределение частот	**Относительные частоты**	**Гистограмма**
Категориальное распределение частот	**Проценты**	**Полигон**
Интервальное распределение частот	**Накопленные частоты**	**Кумулята**
Заявленные и точные границы интервалов	**Накопленные относительные частоты**	**Круговая диаграмма**

Символы и формулы

x_i	Элементы выборки, варианты
f_i	Частоты
$P = f_i / n$	Относительные частоты
CP	Накопленные относительные частоты
%	Проценты
C%	Накопленные проценты

Задачи и упражнения

2-1. Вопросы по теме. Ответьте на следующие вопросы:

- Почему статистики строят распределение частот?
- Чем отличаются отношение частот, проценты и доли?
- Какие способы графических представлений данных вам известны?
- Для ответа на какие вопросы требуется строить кумуляту?

2-2. Чашки кофе. Исследование показало, что в течение дня несколько испытуемых выпили количество чашек кофе, приведенное ниже. Постройте распределение частот. Нарисуйте гистограмму. Сделайте выводы.

0	2	2	1	1	2
3	5	3	2	2	2
1	0	1	2	4	2
0	1	0	1	4	4
2	2	0	1	1	5

2-3. Возраст преподавателей. Ниже собраны данные о возрасте 40 преподавателей одной из школ. Постройте распределение частот, используя восемь интервалов. Нарисуйте гистограмму. Сделайте выводы.

37	41	41	47	62	27	44	43	40	58
62	43	50	61	53	65	58	45	50	27
36	65	43	41	30	42	29	32	48	31
63	38	37	47	26	50	35	31	49	34

2-4. Тестирование студентов. При тестировании уровня подготовки студентов, были получены данные о количестве выполненных заданий, приведенные ниже. Постройте распределение частот. Нарисуйте гистограмму. Сделайте выводы.

7	2	8	2	3	2	4	3	7
5	2	4	11	7	4	7	1	7
2	4	1	1	9	1	6	6	5
8	3	1	4	4	3	2	3	0

2-5. Посетители магазина. Ниже приведены данные, полученные в результате подсчета в течение 60 дней количества посетителей определенного магазина. Постройте распределение частот, используя шесть категорий. Нарисуйте гистограмму. Сделайте выводы.

55	115	118	114	59	109
63	97	90	59	105	81
84	81	82	61	103	77
82	76	68	86	97	80
77	85	69	62	101	83
58	83	101	86	84	78
59	92	88	97	87	92
70	86	72	84	82	84
101	80	93	56	65	91
75	78	100	74	74	90

2-6. Тест IQ. После проведения тестирования 108 случайно выбранных студентов были получены значения IQ, которые приводятся ниже. Нарисуйте гистограмму, полигон и кумуляту для этих величин.

Интервал	Частота
90-98	6
99-107	22
108-116	43
117-125	28
126-134	9

2-7. Почему женщины работают? В ходе опроса 100 работающих женщин изучались основные причины, по которым каждая женщина работает вне дома. Постройте круговую диаграмму.

Материальная независимость/содержание семьи	62
Дополнительные деньги	18
От нечего делать	12
Другое	8

2-8. Откуда берутся новости? В результате опроса 25 респондентов о получении новостей имеются следующие данные. Постройте частотное распределение для следующих данных (Г = газета, Т = телевидение, Р = радио, Ж = журналы):

Г	Г	Р	Т	Т
Р	Г	Т	Ж	Р
Ж	Ж	Г	Р	Ж
Т	Р	Ж	Г	Ж
Т	Р	Р	Г	Г

Глава 3. Описательная статистика

Распределения переменных являются основным материалом при проведении исследований. Поскольку социальные исследования обычно содержат большое число наблюдений, множество методов связано с представлением данных, чтобы показать их наиболее информативно и осмысленно. В главе 2 мы рассмотрели несколько способов представления данных, включая частотные распределения, таблицы и графические представления. Теперь мы обсудим способы численного описания переменных, способы получения простых числовых значений, которые описывают распределение данных. Числовые характеристики, рассмотренные здесь, включают измерение центральной тенденции и измерение вариации. Эти техники представляются простыми и легко понятными. Кроме этого, среднее значение и дисперсия играют роль рабочих лошадок в статистических моделях, поскольку в математических моделях, применяемых в статистическом анализе, используются для построения оценок математического ожидания и дисперсии изучаемых генеральных совокупностей.

3-1 Измерение центральной тенденции

Измерение центральной тенденции, обсуждаемое в настоящем разделе, включает три характеристики: моду, медиану и среднее значение. Эти характеристики могут быть легко вычислены и использованы для последующего анализа. Другие характеристики реже встречаются в социологических исследованиях, и поэтому нами не рассмотрены.

Каждая характеристика представляет уникальную информацию о распределении. Имеются ограничения при их использовании, поскольку не все характеристики могут быть вычислены для различных типов шкал. Например, мода может быть вычислена для номинальной, порядковой или интервальной шкалы, среднее может быть получено для данных, измеряемых по интервальной шкале, и с некоторыми оговорками, для порядковой шкалы.

Измерение центральной тенденции (measure of central tendency) состоит в выборе одного числа, которое наилучшим образом описывает все значения признака из набора данных. Такое число называют центром, типическим значением для набора данных, мерой центральной тенденции.

Почему это требуется? Получив такое число – одно-единственное, мы получаем информацию о распределении признака «в сжатой форме». При этом мы можем сравнивать при помощи этого числа два и более различных распределений. Главный недостаток состоит в том, что мы теряем, по сравнению с распределением частот много информации.

Мода

Мода – наиболее часто встречающееся значение в выборке, наборе данных. В случае, если данные сгруппированы и построено распределение частот, модой является значение, имеющее наибольшую частоту.

Моду будем обозначать *Mo*. Мода вполне пригодна для измерения центральной тенденции хотя бы потому, что это единственный способ описывать номинальное распределение не хуже порядкового или интервального. Ограничения в применении связаны с тем, что мода рассматривает лишь одну особенность распределения, а именно, расположение наиболее частого значения. Другие важные особенности, такие как число наблюдений выше или ниже моды, расстояние между модальными значениями и другие характеристики, остаются вне поля зрения.

В таблице 3-1 представлено категориальное распределение данных о выборе дисциплин специализации 641 студентом. Мода для этих данных - социология, поскольку ее выбрали в качестве дисциплины специализации наибольшее количество студентов - 149. Следующие по популярности специализации – политика и социальная психология.

Таблица 3-1 Дисциплины специализации

ДИСЦИПЛИНЫ	F
Антропология	97
Экономика	104
Политика	110
Психология	72
Социальная психология	109
Социология	149
Итого	641

Таблица 3-2 Вес тела двух групп людей

ВЕС (ФУНТЫ)	ГРУППА 1 F1	ГРУППА 2 F2
190-199	3	3
180-189	2	8
170-179	4	21
160-169	10	7
150-159	13	9
140-149	23	6
130-139	12	7
120-129	7	20
110-119	3	3
100-109	3	2
Всего	80	86

Мода для сгруппированного распределения частот есть середина интервала, в который попадает наибольшее количество наблюдений. В таблице 3-2. мода для группы 1 есть 144,5, поскольку это есть середина интервала, содержащего наибольшее количество наблюдений – 23 из 80 случаев.

Распределение может иметь более чем одну моду, как видно по группе 2 таблицы 3-2. Интервал 170-179 имеет наибольшее количество наблюдений, 21, но почти такое же значение 20 попадает в интервал 120-129. Поскольку разница является незначительной – всего одно наблюдение, это распределение может быть описано как имеющее две моды – 124,5 и 174,5. Его можно назвать *бимодальным*. Обе категории будут считаться одинаково популярными. Распределение может иметь более двух популярных значений, но если оно имеет более трех мод, описание такого распределения в терминах наиболее частых значений теряет всякий смысл.

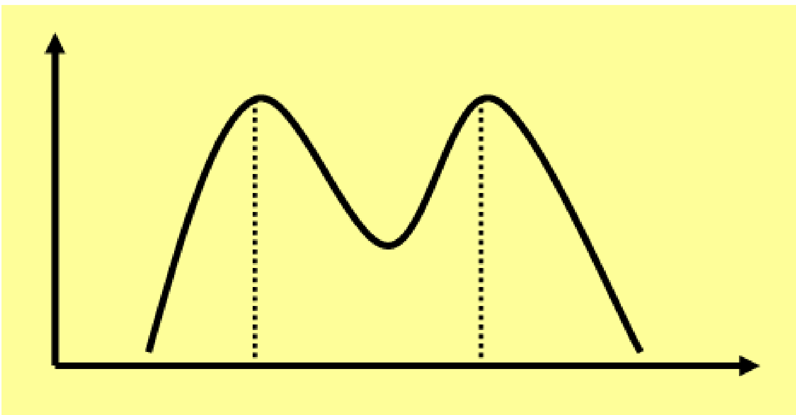


Рисунок 3-1. Вид бимодального распределения

Таблица 3-3 Выборки четного и нечетного размера

ВЫБОРКА 1 (N=5)	ВЫБОРКА 2 (N=6)
198	197
179	193
172	189
167	187
154	183
	179

Медиана

Медиана определяется как серединное значение выборки, или значение, выше и ниже которого располагается одинаковое число наблюдений. Для нахождения медианы обязательно упорядочить данные.

Медиана является точной серединой выборки. Обозначается Me и определяется по-разному для выборок с четным и нечетным числом элементов. Для нечетного количества наблюдений медиана есть наблюдение с номером (n+1)/2. Для четного количества наблюдений медиана вычисляется как среднее значение наблюдений с номерами n/2 и (n+2)/2.

В случае нечетного количества наблюдений медиана есть просто середина выборки, выше и ниже которой располагается одинаковое количество наблюдений.

В Таблице 3-3 первая выборка содержит пять наблюдений. Применяя формулу, получаем, что медиана есть (5+1)/2=3, то есть третье наблюдение, сосчитанное снизу или сверху, а именно, 172. Вторая выборка содержит четное количество наблюдений, 6, это означает, что медиана есть среднее значение наблюдений с номерами 6/2=3 и (6+2)/2=4, то есть значений 187 и 189. Тем самым, медиана есть (187+189)/2=188. В обоих случаях, для четного и нечетного числа наблюдений, медиана есть середина выборки.

«Выше» и «ниже» может иметь смысл лишь по отношению к данным, которые упорядочены по возрастанию или по убыванию. Медиана может быть определена также в терминах ранжирования вариационного ряда и нахождения ранга серединного элемента. Об этом мы скажем позже, а вначале мы определили понятие медианы более «наглядным» способом.

Медиана может быть определена для числовых данных и данных, измеряемых порядковой шкалой. Для номинальной шкалы медиану невозможно отыскать по причине невозможности упорядочить категории номинальной шкалы.

Отметим несколько очевидных свойств, которыми обладает медиана. Сильно отличающиеся от остальных данных крайние значения не влияют на величину медианы. Значение медианы является единственным для каждого набора данных, в отличие от моды. Медиана может быть определена не из полного набора данных. Достаточно знать их порядковое расположение, общее число и несколько значений, расположенных в середине.

Среднее

Среднее определяется как среднее арифметическое выборки, то есть как сумма всех значений выборки, деленная на ее объем.

Следуя определению, будем находить среднее значение по формуле:

$$\bar{x} = \frac{\sum x}{n},$$

где $\sum x$ = сумма всех значений выборки,

n = объем выборки.

Например, для выборки из семи значений: 1, 1, 3, 3, 6, 7, 7, среднее значение будет вычисляться следующим образом:

$$\bar{x} = \frac{1+1+3+3+6+7+7}{7} = \frac{28}{7} = 4.$$

Среднее значение может пониматься как «точка баланса». Если представить, что числовая прямая представляет собой весы, то попытаемся «взвесить» наши наблюдения. Выложим все имеющиеся наблюдения на числовую ось. Среднее значение будет той самой точкой, для которой правая часть и левая не перевешивают друг друга. Это показано для нашего примера на рисунке 3.1.

Понимаемое как точка баланса, среднее значение очень активно используется для описания выборочного распределения и применяется в статистических моделях и вычислениях высокого уровня.

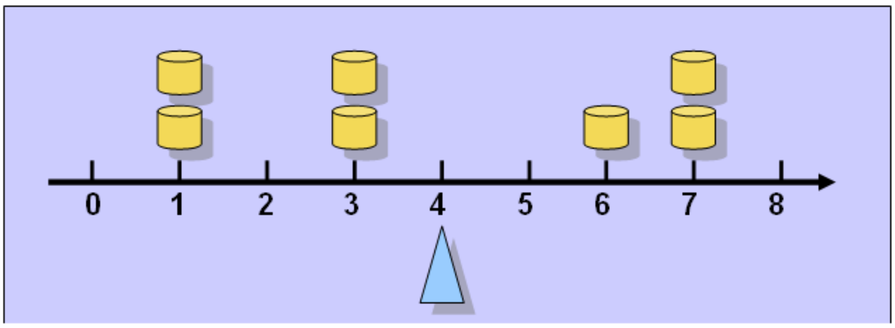


Рисунок 3-1 Среднее значение есть «точка баланса»

Таблица 3-6 Среднее для сгруппированных данных

ОЦЕНКА	ЧАСТОТА	ПРОИЗВЕДЕНИЕ
x	f	f · x
5	17	85
4	41	164
3	20	60
2	7	14
ИТОГО	85	323

Среднее для сгруппированных данных

Среднее значение для сгруппированных данных можно находить по более удобной формуле:

$$\bar{x} = \frac{\sum(f \cdot x)}{n}$$

В этой формуле $\sum(f \cdot x)$ обозначает сумму произведений каждого из значений признака на его частоту, *n* есть общее число наблюдений, которое равно также сумме всех частот.

Для примера, вычислим средний балл для данных, представленных в таблице распределения частот 3.6. В первом столбце перечислены все возможные значения признака: 5, 4, 3, 2. Во втором столбце указаны частоты, с которыми эти значения встретились в наборе данных или выборке. Всего по второму столбцу имеется 85 наблюдений. В третьем столбце мы вычисляем для начала в каждой строке произведение оценки на частоту, а затем находим сумму по столбцу. Сумма оказалась равна 323. Теперь воспользуемся формулой:

$$\bar{x} = \frac{\sum(f \cdot x)}{n} = \frac{323}{85} = 3{,}8$$

Получили, что средний балл составляет 3,8.

Такой способ расчета при помощи таблицы, существенно упрощает вычисления, проводимые вручную и даже на компьютере. Имеется еще несколько формул для вычисления среднего. Мы рассмотрим формулу вычисления среднего для распределения, составленного по интервалам, а также взвешенного среднего.

Таблица 3-7 Нахождение среднего по интервалам

ИНТЕРВАЛ	ЧАСТОТА	СЕРЕДИНА	ПРОИЗВЕДЕНИЕ
	f	m	f · m
0-99	11	49,5	544,5
100-199	12	149,5	1794,0
200-299	14	249,5	3493,0
300-399	1	349,5	349,5
400-499	2	449,5	899,0
Всего	Σf=40		$\Sigma(f \cdot m)$=7080,0

Среднее для интервального распределения

В случае, если мы имеем интервальное распределение, сами значения наблюдений, попадающие внутрь каждого из интервалов, неизвестны. Нахождение среднего в этом случае происходит следующим образом. В таблицу интервального распределения следует добавить столбец, в который проставить середины интервалов. Середины являются «представителями» всего интервала и умножаются на частоту для нахождения среднего.

Для примера рассмотрим таблицу 3-7. В первый интервал попадает 11 наблюдений, хотя конкретные их значения не известны. Выберем в качестве представителя интервала 0-99 его середину 49,5.

Формула для нахождения среднего значения в этом случае будет выглядеть таким образом:

$$\bar{x} = \frac{\sum (f \cdot m)}{n}$$

где f = частота попадания в интервал,
m = середина интервала,
n = объем выборки.

Формула означает, что мы перемножаем частоты на середины интервалов, складываем произведения и делим на общее число наблюдений.

Для нашего примера:

$$\bar{x} = \frac{\sum (f \cdot m)}{n} = \frac{7080,0}{40} = 177,0$$

Таблица 3-7 Нахождение взвешенного среднего

ГРУППА	СРЕДНЕЕ ПО ГРУППЕ	ОБЪЕМ ГРУППЫ
	$\bar{x}$	n
A	87	65
B	92	110
C	89	85
D	96	200
E	84	60
Всего		520

Взвешенное среднее

Часто мы имеем несколько групп наблюдений, средние значения внутри каждой из которых нам известны. Возникает вопрос – как вычислить групповое среднее, то есть среднее значение по всем наблюдениям, составленным из всех имеющихся групп наблюдений.

Взвешенное среднее – среднее значение, получаемое при объединении нескольких групп наблюдений.

Если группы имеют одинаковый объем, то групповое среднее можно вычислить как среднее арифметическое имеющихся средних значений по каждой группе. Если же группы имеют различный объем, то групповое среднее можно найти по следующей формуле:

$$\bar{X} = \frac{\sum(\bar{x} \cdot n)}{N},$$

где $\sum(\bar{x} \cdot n)$ = сумма произведений средних в группе на количество элементов в этой группе,

N = общее число наблюдений во всех группах.

Например, в таблице 3.7 класс А с 65 наблюдениями вносит меньший вклад в групповое среднее по сравнению с группой В. имеющей 110 наблюдений. Используем формулу для нашего примера:

$$\bar{X} = \frac{65 \cdot 87 + 110 \cdot 92 + 85 \cdot 89 + 200 \cdot 96 + 60 \cdot 84}{65 + 110 + 85 + 200 + 60} = \frac{47580}{520} = 91{,}5$$

Название «взвешенное среднее» используется потому, что для его нахождения мы учитываем веса, которые имеют средние значения по группам.

Среднее для дихотомической шкалы

Мы уже отмечали, что дихотомическая шкала обладает уникальным свойством. Для нее может вычисляться среднее значение, несмотря на то, что для номинальных шкал среднее не вычисляется, поскольку в них запрещены арифметические операции. Напомним, что в дихотомической шкале имеются всего два значения: да - нет, знаю – не знаю и т.п. Если два значения признака кодируются 0 и 1, то вычисленное среднее укажет долю единиц в выборке.

Например, для выборки {1, 0, 0, 0, 1, 1, 1, 1, 1, 0}среднее будет равно: число единиц / число элементов выборки = 6 / 10 = 0,6. Это означает, что 60% значений выборки принимают значение, равное единице.

Среднее не значит лучшее

Отметим некоторые очевидные свойства среднего.

- Среднее вычисляется только в числовых шкалах. К сожалению, среднего не существует для номинальных и порядковых шкал. Дихотомическая шкала составляет приятное исключение.
- При вычислении среднего необходимо использовать все данные. Пропущенные значения в данных не допускаются.
- Для каждого набора данных может быть вычислено только одно значение среднего. Этим свойством обладает медиана и не обладает мода.
- Среднее является единственной мерой центральной тенденции, для которой сумма отклонений каждого значения от него равна нулю: $\sum(x-\bar{x})=0$.

Среднее, несмотря на кажущиеся преимущества перед другими мерами центральной тенденции, имеет серьезные недостатки. Они становятся понятны после некоторой практики использования среднего в качестве меры.

Приведем только один пример. В некоей деревне проживает 50 жителей. Среди них 49 человек – крестьяне с месячным доходом в 1 тыс.рублей, а один житель – зажиточный владелец строительной фирмы, с месячным доходом 451 тыс.рублей. Вычислим среднее. Оно окажется равно 10 тыс. рублей. Однако, вряд ли в этом случае можно утверждать, что это число адекватно представляет доход жителей деревни. В этом случае, более разумно взять в качестве меры центральной тенденции моду или медиану (обе равны 1 тыс. рублей).

Таблица 3-8 Типы шкал и меры центральной тенденции

Типическое значение	Номинальные данные	Порядковые данные	Интервальные данные
Мода	●	●	●
Медиана		●	●
Среднее			●

Тогда возникает законный вопрос, а какая мера центральной тенденции является все-таки наилучшей? Ответ на этот вопрос существует и зависит от критериев. Каждая мера является наилучшей в своем, вполне определенном смысле.

- Если считать, что данные наилучшим образом представляет наиболее часто встречающийся элемент, тогда это **Мода**.
- Если считать, что наилучшим представителем данных является значение, для которого сумма абсолютных отклонений от него всех оставшихся значений будет наименьшей, тогда это **Медиана**.
- Если считать, что наилучшим представителем для данных является такое значение, для которого сумма квадратов отклонений от него всех значений будет наименьшей, тогда это **Среднее**.

Нам осталось обобщить соответствие различных мер центральной тенденции типам шкал, для которых они могут быть применены. Такое соответствие описано в таблице 3.8.

Резюме

Меры центральной тенденции представляют собой единственное число, которое может рассматриваться в качестве представителя набора данных. Такое единственное значение имеет преимущества и недостатки. Кроме того, три различные меры - мода, медиана и среднее - применимы не всегда. В некоторых случаях они просто не могут быть вычислены, а иногда не отражают главной цели – представлять набор данных для анализа и сопоставлений. Изученные в параграфе меры будут активно использоваться нами в последующих главах.

3-2 Измерение вариации

Измерение центральной тенденции не дает представления о различиях данных внутри выборки. Для этого существует *измерение вариации* выборки или набора данных.

Измерение вариации (measure of variation) состоит в нахождении чисел, которые характеризуют степень разброса данных относительно центра распределения.

Вариация может быть проиллюстрирована следующим примером. Предположим, у нас имеются измерения уровня достатка в нескольких провинциальных городах и пригородных зонах. Количество семей ниже уровня бедности на 100 семей отражено показателем и данные представлены в таблице 3.9.

В обоих случаях среднее значение бедных семей на каждые 100 составило 21,7. Это означает, что если для описания выборок использовать исключительно среднее, мы придем к выводу, что выборки идентичны. Однако, даже беглого взгляда достаточно, чтобы понять различия между ними. Для первой выборки характерно не слишком сильное отличие измерений от среднего значения 21,7. Наименьшее значение есть 19,6, наибольшее – 24,5. Выборка для центральной части городов имеет совершенно другую картину. Изменение значений происходит от 17,4 до 27,4. В этом случае среднее значение 21,7 не столь хорошо описывает количество бедных семей в случайно выбранном месте в центре города. Получилось, что семьи на окраинах городов более однородны в смысле числа бедных семей, чем в центральной части городов.

Таблица 3-9 Семьи ниже уровня бедности

ОКРАИНА ГОРОДА	ЦЕНТР ГОРОДА
24,5	27,4
23,8	24,6
23,1	23,0
22,4	22,5
21,7	21,8
21,0	21,6
21,0	20,9
20,3	19,7
19,6	18,1
19,6	17,4
217,0	217,0
$\bar{x}=21{,}7$	$\bar{x}=21{,}7$

Размах

Размах (range) – разница между наибольшим и наименьшим значениями.

Мы рассматриваем в качестве первой характеристики вариации размах. Для нахождения размаха прежде рекомендуется упорядочить данные в порядке возрастания. Можно записать размах с помощью формулы:

$$R = x_{\max} - x_{\min}$$

В нашем примере в таблице 3.9 разность составляет для первой выборки 24,5 – 19,6=4,9, а для второй 27,4 – 17,4 = 10. Мы обсуждали уже, что различие между двумя выборками существует, но не могли описать это различие количественно.

Разность дает нам вполне пригодную информацию о выборке путем описания расстояния между наибольшим и наименьшим значением. Очевидно, эта информация имеет прямое отношение к характеристике разброса изучаемого нами распределения.

Квартильный размах

Квартили (quartile) – значения, которые делят вариационный ряд на четыре равные по объему части. **Квартильный размах (Inter Quartile Range - IRQ)** – разница между третьим и первым квартилями.

Таких значений должно быть три: первая, вторая и третья квартиль соответственно. Для начала данные следует упорядочить. Затем отыскивается медиана, которая является вторым квартилем по определению. После этого находятся первый и третий квартили.

Существует несколько вариантов формального определения квартилей. Для учебных целей мы упростим задачу их поиска аналогично нахождению медианы. Если в качестве серединного элемента выступит два претендента – мы найдем их среднее арифметическое и его назовем соответствующим квартилем. Учебный пример нахождения квартилей будет рассмотрен в следующем параграфе.

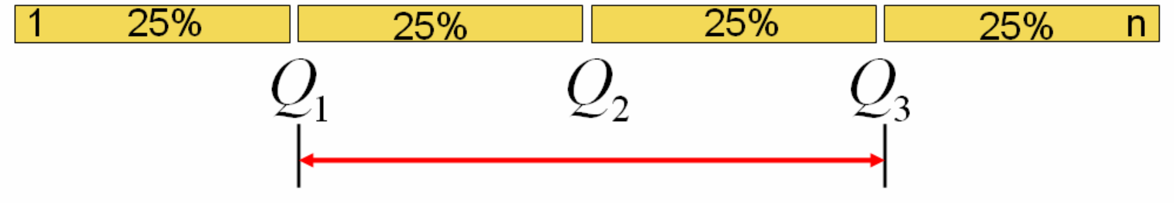


Рисунок 3-2 Квартили и квартильный размах

Квартильный размах находится по формуле:

$$IQR = Q_3 - Q_1$$

Если при вычислении размаха используются только наибольшее и наименьшее значения признака, а распределение данных между ними полностью игнорируется, то при вычислении квартильного размаха игнорируются «крайние» данные, расположенные за пределами первого и третьего квартилей. Между Q_1 и Q_3 расположено 50% всех данных.

Дисперсия

Дисперсия для набора данных или выборки – среднее арифметическое квадратов отклонений значений от их среднего.

Дисперсия обозначается s^2. Основная формула (по определению) для нахождения дисперсии:

$$s^2 = \frac{\sum (x - \bar{x})^2}{n - 1}$$

Формула означает, что нам следует вычитать среднее из каждого значения выборки, суммировать квадраты разности, а затем разделить полученную сумму на количество наблюдений минус 1. Почему в знаменателе при нахождении среднего арифметического квадратов отклонений используется (n-1) вместо n, будет разъяснено позднее, при обсуждении логики построения оценок параметров генеральной совокупности.

Вычислим дисперсию в приведенном примере. Расчет проведем путем достраивания таблицы 3.8 в таблицу 3.9. Получим:

Окраина города: $$s^2 = \frac{\sum (x - \bar{x})^2}{n - 1} = \frac{26{,}46}{10 - 1} = 2{,}94$$

Центр города: $$s^2 = \frac{\sum (x - \bar{x})^2}{n - 1} = \frac{79{,}34}{10 - 1} = 8{,}82$$

Результаты подтверждают наше интуитивное представление о том, что в центральной части городов дисперсия количества бедных семей больше, чем на окраинах городов. Это видно по результатам подсчета: 8,82 для центра города больше 2,94 для окраин. Стандартное отклонение, которое мы рассмотрим ниже, позволяет привести рассматриваемые дисперсии к стандартному виду, который более понятен для понимания и сравнений.

Таблица 3-10 Вычисление дисперсии для уровня бедности

ОКРАИНА ГОРОДА (N=10)			ЦЕНТР ГОРОДА (N=10)		
x	$x-\bar{x}$	$(x-\bar{x})^2$	x	$x-\bar{x}$	$(x-\bar{x})^2$
24,5	2,8	7,84	27,4	5,7	32,49
23,8	2,1	4,41	24,6	2,9	8,41
23,1	1,4	1,96	23,0	1,3	1,69
22,4	0,7	0,49	22,5	0,8	0,64
21,7	0	0	21,8	0,1	0,01
21,0	-0,7	0,49	21,6	-0,1	0,01
21,0	-0,7	0,49	20,9	-0,8	0,64
20,3	-1,4	1,96	19,7	-2,0	4,00
19,6	-2,1	4,41	18,1	-3,6	12,96
19,6	-2,1	4,41	17,4	-4,3	18,49
$\sum$=217,0		$\sum$=26,46	$\sum$=217,0		$\sum$=79,34

Имеется вторая формула для нахождения дисперсии выборки:

$$s^2 = \frac{n \cdot \sum x^2 - (\sum x)^2}{n \cdot (n-1)}$$

Считается, что эта формула более пригодна для ручного счета, поскольку требует меньшего количества арифметических операций. Проиллюстрируем эту формулу на следующем примере. Рассмотрим выборку из 4 значений: 2, 3, 6, 9. Вычислим дисперсию. Вспомогательная таблица 3-11 будет иметь всего два столбца: столбец значений выборки и квадраты значений. Сумма значений равна 20, а их квадратов 130. Это все, что необходимо нам для вычислений по второй формуле для дисперсии:

$$s^2 = \frac{n \cdot \sum x^2 - (\sum x)^2}{n \cdot (n-1)} = \frac{4 \cdot 130 - 20^2}{4 \cdot (4-1)} = 10$$

Таблица 3-11 Вычисление дисперсии

x	x^2
2	4
3	9
6	36
9	91
$\sum$=20	$\sum$=130

Таблица 3-12 Вспомогательная таблица для вычисления дисперсии

Стаж работы	f	x	$f \cdot x$	$f \cdot x^2$
2-4	2	3	6	18
5-7	5	6	30	180
8-10	10	9	90	810
11-13	4	12	48	576
14-16	2	15	30	450
	Σ=23		Σ=204	Σ=2034

Дисперсия для сгруппированных данных

Для сгруппированных данных дисперсия вычисляется по следующей формуле:

$$s^2 = \frac{n \cdot \sum (f \cdot x^2) - \left[\sum (f \cdot x)\right]^2}{n \cdot (n-1)}$$

Для вычисления по указанной формуле нам также потребуется вспомогательная таблица 3-12.

Рассмотрим в качестве примера данные, в которых 23 сотрудника указали свой стаж работы в компании и эти данные размещены в двух первых столбцах таблицы. В третьем столбце мы поставим средний стаж для каждого интервала (середину). Четвертый и пятый столбец являются вспомогательными для вычислений. Полученные значения подставляем в формулу:

$$s^2 = \frac{n \cdot \sum (f \cdot x^2) - \left[\sum (f \cdot x)\right]^2}{n \cdot (n-1)} =$$

$$= \frac{23 \cdot 2034 - 204^2}{23 \cdot (23-1)} = 10{,}2$$

Стандартное отклонение

Стандартное отклонение – квадратный корень из дисперсии выборки.

Обозначается s и вычисляется по формуле:

$$s = \sqrt{\frac{\sum (x - \bar{x})^2}{n-1}}$$

Коэффициент вариации

Коэффициент вариации вычисляется как отношение стандартного отклонения к среднему значению выборки.

Формула для коэффициента вариации:

$$CV = s / \bar{x}$$

Коэффициент вариации полезен, если сравниваются несколько совокупностей, измеряемых в разных величинах, либо сравниваются совокупности, измеряемые в одинаковых величинах, но имеющие сильно отличающиеся средние.

В качестве примера выясним, какие данные имеют большую вариацию: имеющие стандартное отклонение 20 при среднем 200 или имеющие стандартное отклонение 3 при среднем 30? Воспользуемся формулой для коэффициента вариации:

$$CV = s / \bar{x} = 20 / 200 = 0{,}1$$
$$CV = s / \bar{x} = 3 / 30 = 0{,}1$$

Получили, что в обоих случаях коэффициенты вариации равны. Это означает, что вариация одинакова.

Резюме

Вслед за изучением числовых характеристик центральной тенденции мы рассмотрели основные числовые характеристики вариации данных. Наиболее простыми с точки зрения нахождения являются размах и квартильный размах. Более сложны для вычислений, но чрезвычайно полезны и важны для последующих стадий статистического анализа такие характеристики как дисперсия и стандартное отклонение. Мы использовали в своих расчетах несколько различных формул для вычисления дисперсии, которые тем не менее дают одинаковые результаты на одних и тех же данных. Хотя эквивалентность разных формул нами не доказывалась, но подразумевалась.

Следующий параграф поможет нам использовать все рассмотренные числовые характеристики для так называемого исследовательского анализа данных. Этот анализ завершит краткое изучение основ описательной статистики.

3-3 Исследовательский анализ данных

Наше изучение описательной статистики завершает исследовательский анализ данных, который позволяет провести комплексный анализ важнейших характеристик распределения.

Исследовательский анализ данных (Exploratory Data Analysis - EDA) представляет собой применение статистических методов для представления, упорядочения данных и понимания их важнейших характеристик.

Основными разделами анализа являются:

1. **Измерение центральной тенденции**. Вычисление и анализ среднего, моды, медианы.
2. **Измерение вариации**. Нахождение минимума и максимума, размаха и квартильного размаха, вычисление дисперсии и стандартного отклонения.
3. **Нахождение и анализ выбросов**. Выделение границ для выбросов, анализ экстремальных и умеренных выбросов.
4. **Анализ формы распределения**. Вычисление и анализ коэффициентов асимметрии и куртозиса.

Исследовательский анализ данных позволяет анализировать числовые значения в качестве ключевых характеристик и делать выводы на основе этого анализа, относящиеся к имеющимся данным. К сожалению, такой анализ, при всей его комплексности и полноте не позволяет делать обоснованных выводов относительно генеральной совокупности, из которой получены данные. Это возможно будет сделать в рамках другой теории.

Коробковая диаграмма

При проведении анализа очень полезна так называемая коробковая диаграмма, которая имеет следующий вид (рис.3.3):

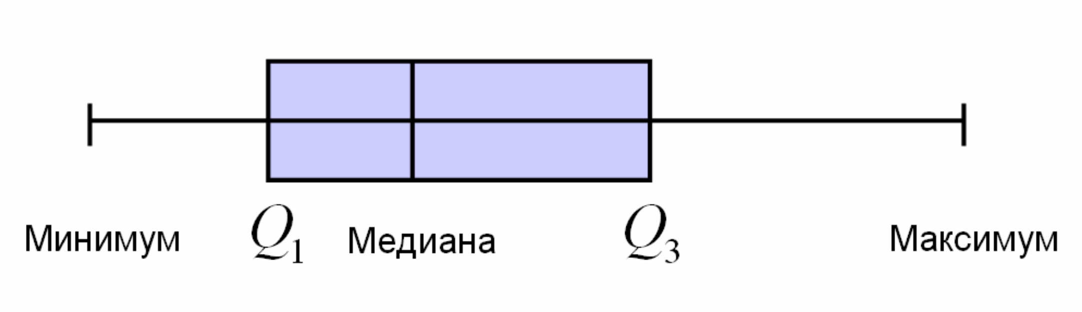


Рисунок 3-3 Коробковая диаграмма (Box Plot)

По виду коробковой диаграммы можно сделать вывод, где расположена медиана по отношению к минимуму и максимуму, по виду и размеру «коробки» можно судить, где расположены 50% данных.

Выбросы

Выбросами (outliers) называются данные, которые сильно отдалены от основного числа данных.

С коробковой диаграммой тесно связан анализ выбросов. Для нахождения и анализа выбросов кроме обычной коробковой диаграммы строится также расширенная коробковая диаграмма, которая содержит пометки на умеренных и экстремальных выбросах.

Чтобы отыскать выбросы нам должно быть известно значение IQR – квартильного размаха, который находится как разница между третьим и первым квартилем. Заметим, что IQR есть длина «коробки».

Умеренные выбросы (mild outliers) изображаются темными точками и удалены ниже первой квартили или выше третьей от 1,5 IQR, но не более 3 IQR.

Экстремальные выбросы (extreme outliers) изображаются светлыми точками и удалены ниже первой квартили или выше третьей более 3 IQR.

При анализе выбросов требуется принимать решение – либо отказаться от выбросов и вести дальнейшие исследования без них, либо оставить выбросы для целей последующего анализа. Если выбросы исключаются, это должно быть детально аргументировано и описано в отчете об исследовании. Если выбросы остаются, следует провести два параллельных исследования: с ними и без них, а затем сопоставить результаты и сделать дополнительные выводы.

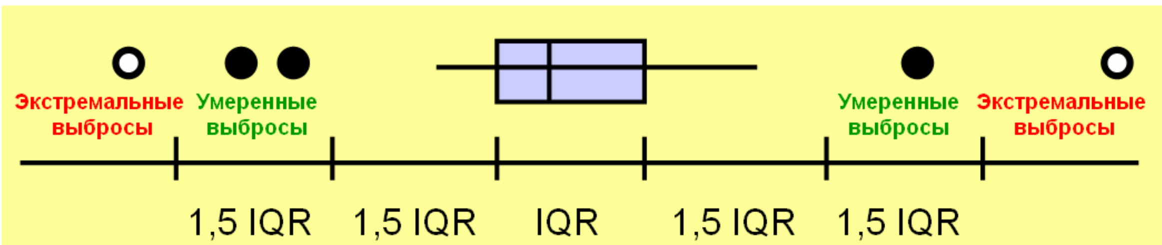


Рисунок 3-4 Расширенная коробковая диаграмма

Асимметрия

Асимметрия (skewness) – характеристика распределения, которая сообщает о наличии или отсутствии симметрии данных.

Асимметрия имеет тесную связь с расположением моды, среднего и медианы. Если распределение симметрично, **асимметрия равна нулю**. В этом случае совпадают значения моды, медианы и среднего значения (средний график).

Если одно или несколько значений существенно превышают остальные, имеется **положительная асимметрия**. Среднее больше моды и медианы (левый график).

Если одно или несколько значений существенно меньше остальных, имеется **отрицательная асимметрия**. Среднее меньше моды и медианы (правый грфик).

Асимметрия измеряется при помощи коэффициента, вычисляемого по формуле:

$$Sk = \frac{\bar{x} - Me}{s}$$

Иногда коэффициент асимметрии используется с коэффициентом 3 в числителе и тогда он изменяется в границах от -3 до +3. В приведенном случае коэффициент меняется от -1 до +1. Положительные значения характеризуют положительную асимметрию. В случае, когда медиана и среднее совпадают, коэффициент равен нулю.

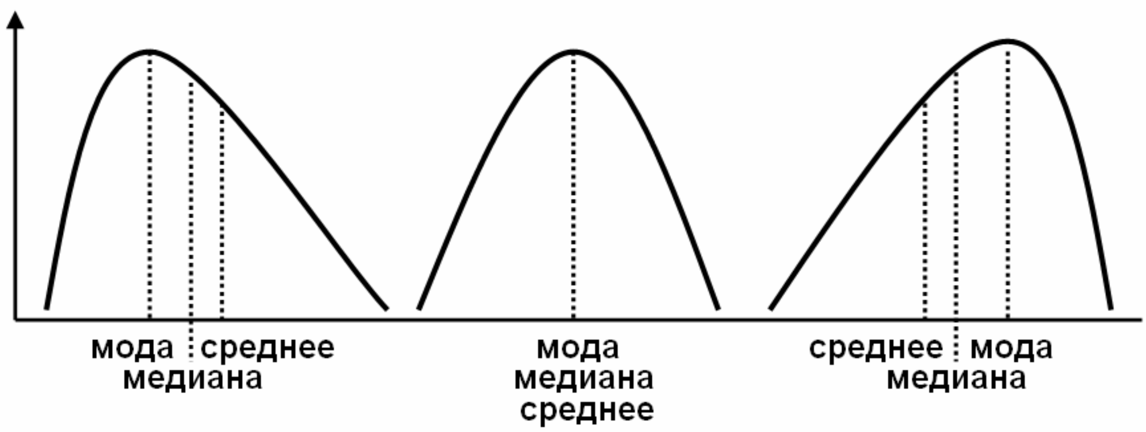


Рисунок 3-5 Виды асимметричных графиков

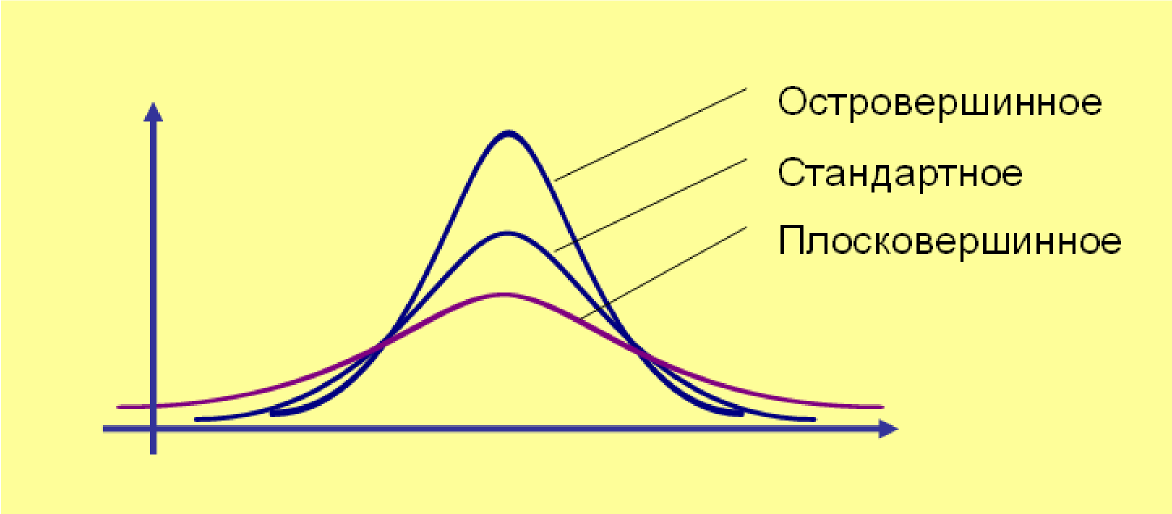


Рисунок 3-6 Виды распределения, отличающиеся крутостью

Эксцесс

Под **эксцессом** понимается крутость кривой распределения, которая определяется сопоставлением кривой с кривой стандартного нормального распределения (рисунок 3-6). Мы ограничимся этим и не будем изучать более подробно нахождение соответствующих числовых характеристик.

Резюме

Исследовательский анализ данных включает совокупность методов численного анализа данных. Выводы анализа не могут относиться к генеральной совокупности, а исключительно к самим данным, их распределению. Он называется исследовательским, поскольку используется для получения первичных выводов, формирования гипотез относительно генеральной совокупности.

Используем компьютер

Материал этой главы требует значительной работы с компьютером, поскольку изученные понятия и характеристики важно уметь вычислять в одном из статистических пакетов. Несколько заданий по исследовательскому анализу данных следует выполнить вручную, а затем получить результаты в компьютере и сравнить. Далеко не всегда результаты окажутся одинаковыми. Потребуется некоторая настойчивость, чтобы понять, в чем состоит причина расхождений. Кроме этого, исследовательский анализ данных на компьютере представит больше материала, поскольку мы изучили далеко не все.

Что означают термины

Измерение центральной тенденции	**Измерение вариации**	**Коэффициент вариации**
Мода	**Размах**	**Исследовательский анализ данных**
Медиана	**Квартильный размах**	**Выбросы**
Среднее	**Дисперсия**	**Асимметрия**
Взвешенное среднее	**Стандартное отклонение**	**Куртозис**

Символы и формулы

Mo	Мода
Me	Медиана
$\bar{x} = \frac{\sum x}{n}$	Среднее значение
$\bar{x} = \frac{\sum (f \cdot x)}{n}$	Среднее для сгруппированных данных и для интервального распределения
$\bar{X} = \frac{\sum (\bar{x} \cdot n)}{N}$	Взвешенное среднее
$R = x_{\max} - x_{\min}$	Размах
$IQR = Q_3 - Q_1$	Квартильный размах
$s^2 = \frac{\sum (x - \bar{x})^2}{n-1}$	Дисперсия выборки
$s^2 = \frac{n \cdot \sum x^2 - (\sum x)^2}{n \cdot (n-1)}$	Вторая формула для дисперсии
$s^2 = \frac{n \cdot \sum (f \cdot x^2) - [\sum (f \cdot x)]^2}{n \cdot (n-1)}$	Дисперсия для сгруппированных данных
$s = \sqrt{\frac{\sum (x - \bar{x})^2}{n-1}}$	Стандартное отклонение
$CV = s / \bar{x}$	Коэффициент вариации
$Sk = \frac{\bar{x} - Me}{s}$	Коэффициент асимметрии

Задачи и упражнения

В задачах 3-1 … 3-8 вычислить меры центральной тенденции и меры вариации. Сделать выводы.

3-1. Длина рассказа. Произвольно были выбраны десять рассказов, здесь приведен их объем в страницах: 415, 398, 402, 399, 400, 405, 395, 401, 412, 407.

3-2. Утренняя пробежка. Были протестированы двенадцать членов университетской туристической секции, сколько минут каждый из них совершает пробежку перед тренировкой: 32, 28, 35, 37, 43, 51, 61, 39, 48, 51, 53, 49.

3-3. Проданные кепки. Управляющий спортивного магазина вел записи продаж кепок за неделю, вот его данные о продаже: 132, 121, 119, 116, 130, 121, 131.

3-4. Подготовка к экзамену. Отобраны пятнадцать студентов третьего курса. Им задан вопрос: «Сколько времени вы потратили на подготовку к экзамену по статистике?» Их ответы записаны ниже (в часах): 8, 6, 3, 0, 0, 5, 9, 2, 1, 3, 7, 10, 0, 3, 6.

3-5. Сообразительность студентов. При тестировании 108 студентов коллежа были выявлены следующие показатели IQ:

IQ	Частота
90-98	6
99-107	22
108-116	43
117-125	28
126-134	9

3-6. Прочитанные книги. Количество книг, прочитанных студентами за ноябрь. Опрошено 28 человек:

Число книг	Частота
0	5
1	6
2	12
3	5
4	3

3-7. Лабиринт для мышей. Результаты исследования времени, требуемого нетренированной мыши для прохождения лабиринта:

Минуты	Частота
2.1-2.7	5
2.8-3.4	7
3.5-4.1	12
4.2-4.8	14
4.9-5.5	16
5.6-6.2	8

3-8. Секретари-машинистки. В ходе проверки скорости печати (слова в минуту) 25 машинисток были получены следующие результаты:

Слова	Частота
54-58	2
59-63	5
64-68	8
69-73	0
74-78	4
79-83	5
84-88	1

3-9. Учебники и возраст. На кафедрах факультета находится в среднем 16 учебников, а стандартное отклонение равно 5. Средний возраст профессоров составляет 43 года при стандартном отклонении 8 лет. Какие данные более изменчивы?

3-10. Журналы и продавцы. Осмотр книжных магазинов показал, что в среднее количество журналов в них равно 56, а стандартное отклонение – 12. Средняя продолжительность работы продавцов 6 лет со стандартным отклонением 2,5 года. Какие данные более изменчивы?

3-11. Актеры и актрисы. Имеются данные о возрасте актеров и актрис, в котором они были удостоены Оскара.

Актеры: 32, 37, 36, 32, 51, 53, 33, 61, 35, 45, 55, 39, 76, 37, 42, 40, 32, 60, 38, 56, 48, 48, 40, 43, 62, 43, 42, 44, 41, 56, 39, 46, 31, 47, 45, 60, 46, 40, 36.

Актрисы: 50, 44, 35, 80, 26, 28, 41, 21, 61, 38, 49, 33, 74, 30, 33, 41, 31, 35, 41, 42, 37, 26, 34, 34, 35, 26, 61, 60, 34, 24, 30, 37, 31, 27, 39, 34, 26, 25, 33.

Проведите исследовательский анализ данных и сделайте выводы.

Глава 4. Вероятность

Понятие вероятности для статистики является ключевым, поскольку на его основе строятся теоретические основания для получения статистических выводов при изучении явлений массового характера. В этой главе мы познакомимся с понятием случайного события, дадим несколько определений вероятности и обсудим их различия, а также изучим правило сложения и умножения вероятностей. Завершат главу две часто используемые при нахождении вероятности формулы – формула полной вероятности и формула Байеса.

4-1 Определение вероятности

Жизнью правит случай. Это известное философское изречение есть результат размышлений многих и многих поколений над случайными событиями и явлениями, которые окружают нас повсюду. Предсказывать случайные события, управлять ими до сих пор остается для человечества лишь мечтой. Пойдет ли завтра дождь, как выиграть в лотерею, какую оценку поставит преподаватель – таких вопросов множество и все они до поры до времени остаются без ответа.

Эти вопросы заставляли людей наблюдать происходящее и постепенно накапливать знания о случайных событиях. Так постепенно формировались теоретические представления о случайности и вероятности. Долгое время теория вероятностей развивалась учеными на основе наблюдений за азартными играми: игрой в кости, карточными играми, рулеткой. До сих пор при изучении элементарных основ вероятности эти игры дают классические учебные примеры.

Поскольку теория вероятностей является древней наукой, понятие вероятности не является простым и однозначным. Существуют одновременно субъективный и объективный подходы к определению вероятности. Субъективность может быть снижена, прежде всего, поиском закономерностей среди явлений и событий, имеющих массовый характер. Важнейшим открытием ученых в теории вероятностей является закон больших чисел, который показал, что массовые явления обладают удивительным свойством устойчивости, которое проявляется при большом числе испытаний. При этом конкретные особенности отдельного случайного явления почти не сказываются на результатах большого числа таких явлений.

Для начала рассмотрим математическую модель, используемую для классического определения вероятности. Она основана на понятии пространства элементарных исходов. Эта модель, как мы увидим, оказалась пригодна для описания многих случайных событий и явлений.

Случайные события

Эмпирические наблюдения, тестирование, проведение эксперимента можно назовем общим термином *испытание*. Для нас не важно, организовано ли испытание специально – в виде эксперимента с определенными целями, либо оно явилось результатом стечения внешних обстоятельств или следствием причин, которые нам не известны. Мы в любом случае будем говорить об испытаниях. Далее, в результате испытания мы можем получить некоторое количество *исходов*.

Если испытанием считать случайное падение стакана со стола, то исходов два – разобьется или не разобьется. Можно добавить еще один исход: стакан не упадет со стола. Тогда уже три исхода. А сколько их всего? А, может, стоит рассматривать только два первых? В самом начале изучения бывает трудно отыскать рамки испытания и выделить все возможные исходы. В конечном счете, это связано с моделью, которую выбирает исследователь в зависимости от целей исследования. Нам предстоит научиться выбирать модели для начала на простых примерах. В любом случае, мы будем считать, что различные исходы не могут произойти одновременно.

Теперь из всех возможных исходов научимся выделять элементарные. Исход «стакан упадет со стола» не является элементарным, поскольку в нашем рассмотрении в него входит целых два других: «стакан упадет со стола и разобьется» и «стакан упадет со стола и не разобьется».

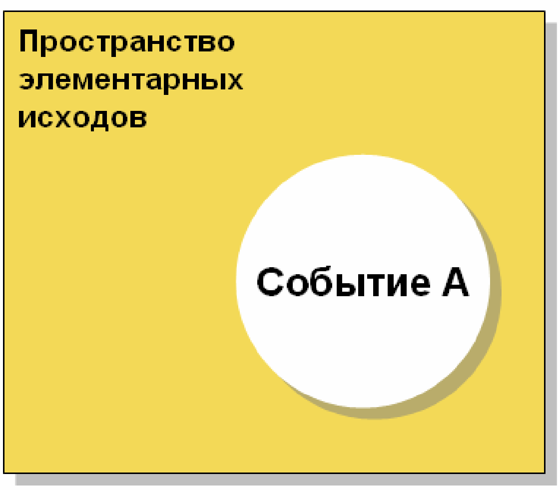


Рисунок 4-1 Случайное событие

Таблица 4-1 Примеры испытаний, элементарных исходов, событий

ИСПЫТАНИЕ	ПРИМЕРЫ СОБЫТИЙ	ПРОСТРАНСТВО ЭЛЕМЕНТАРНЫХ ИСХОДОВ
Подбрасывание одной кости	Выпадение числа 5 (элементарный исход) Выпадение четного числа (вероятное событие) Выпадение числа 7 (невероятное событие)	1, 2, 3, 4, 5, 6
Подбрасывание двух костей	Выпадение в сумме 7	1-1, 1-2, …, 6-6

Исход «стакан не упадет со стола» является элементарным, поскольку не может быть разделен на другие исходы.

Элементарный исход – исход испытания, который не может быть разделен на некоторое количество других исходов.

Подбрасываем кость. Примером элементарного исхода является выпадение единицы. Примером исхода, не являющегося элементарным, может служить выпадение четного числа. Это исход включает выпадение двойки, четверки или шестерки. А всего элементарных исходов шесть. Мы будем говорить, что они образуют пространство элементарных исходов.

Пространство элементарных исходов включает все элементарные исходы, которые могут произойти в результате испытания.

В пространстве элементарных исходов можно выделить подмножества, состоящие из одного или нескольких исходов. В этом случае говорят о случайном событии.

Случайное событие – некоторое подмножество пространства элементарных исходов испытания.

Некоторые примеры испытаний, случайных событий и элементарных исходов приведены в таблице 4-1. При подбрасывании кости выпадение числа семь является невероятным событием. Оно не включает ни одного элементарного исхода.

Достоверным назовем событие, наступающее при любом исходе испытания.

Пример достоверного события: при подбрасывании монеты выпадет Орел или Решка. Ситуации типа «Встанет на ребро» или «Повиснет в воздухе» считаются невозможными событиями.

Невозможным назовем событие, не наступающее ни при одном исходе испытания.

Обозначим ожидаемое нами событие А. Элементарные исходы, образующие событие А, будем называть *благоприятными*. Это вовсе не означает, что он чем-то лучше других. Просто если мы *ожидаем* событие А, то появление любого элементарного исхода, образующего событие А, для нас является благоприятным.

Равновозможными назовем события, для которых есть основания считать, что ни одно из них не является более возможным, чем другое.

Понятие равновозможных событий окажется принципиально важным для классического определения вероятности.

Алгебра событий

Рассмотрение двух и более событий приводит к необходимости отвечать на вопрос, могут ли они произойти одновременно, какова вероятность того, что произойдет одно из них или произойдут оба. Эти вопросы требуют обсуждения некоей общей модели, называемой алгеброй событий. В этой алгебре, как и алгебре чисел, существуют операции сложения, умножения событий. В результате эти операции окажутся важными для нахождения вероятностей, относящихся к двум и нескольким событиям.

Суммой А+В случайных событий А и В называется событие, состоящее в том, что произошло хотя бы одно из них.

Сумма А+В означает, что произошло событие А или событие В, не исключая того, что они могли произойти оба. Сумма событий есть их объединение. Любой элементарный исход, который входит в событие А или событие В, входит также и в их сумму А+В.

Произведением АВ событий А и В называется событие, состоящее в том, что произошли оба события.

Произведение АВ означает, что произошло и событие А, и событие В одновременно. Произведение событий есть их пересечение. В алгебре событий кроме суммы и произведения также определены операции вычитания и дополнения.

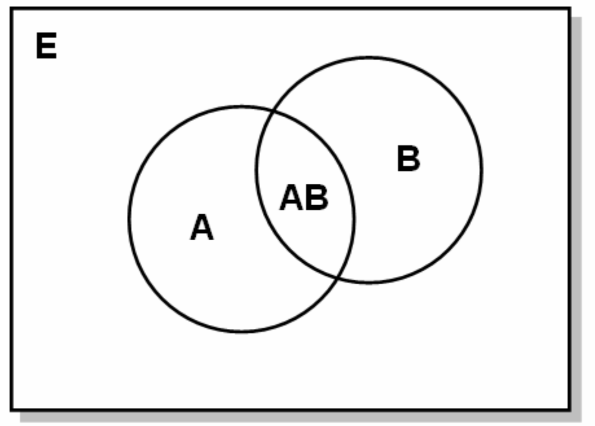


Рисунок 4-2 Совместные события

Для алгебры событий удобно использовать графическое представление, так называемую диаграмму Венна. Внешний прямоугольник обозначает все пространство элементарных исходов и обозначается E. Фигура A обозначает событие A как подмножество пространства элементарных исходов. Площадь прямоугольника теоретически считается равной единице, а площадь фигуры A – вероятности события A.

События *A* и *B* называются **несовместными**, если они не могут произойти одновременно. В противном случае, эти события совместны.

На рисунке 4-2 изображены совместные события A и B. Их пересечение составляет событие AB.

События H_1, H_2, … , H_n образуют **полную группу событий**, если они попарно несовместны, а их сумма является достоверным событием.

Одним из графических изображений полной группы событий, которое потребуется нам впоследствии, является изображение, представленное на рисунке 4-3.

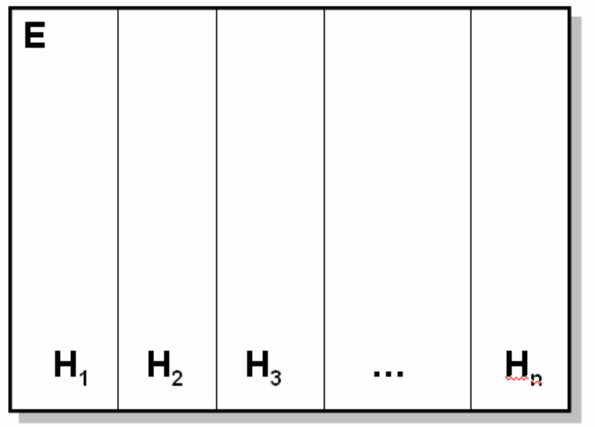


Рисунок 4-3 Полная группа событий

Полная группа событий, пространство элементарных исходов и достоверное событие являются в некотором смысле синонимами, поскольку означают и тот же объект. Тем не менее, мы не сможем отказаться ни от одного из этих терминов, поскольку каждый из них используется в своем контексте. В частности, пространство элементарных исходов потребуется для определения понятия вероятности.

Классическое определение вероятности

Итак, случайное событие представляет собой некоторое подмножество пространства элементарных исходов. Классическое определение вероятности основывается на модели, в которой *все элементарные исходы равновозможны*. Только в этом случае вероятность случайного события А можно определить как долю благоприятных исходов (образующих событие А) к общему числу элементарных исходов.

Вероятностью события А назовем отношение числа благоприятных исходов к общему числу элементарных исходов (классическое определение вероятности).

Запишем вероятность события А с помощью формулы:

$$P(A) = \frac{m}{n},$$

где m - число элементарных исходов, составляющих событие А,
n – общее число элементарных исходов.

Назовем несколько важных свойств вероятности случайного события.

Свойство 1. Вероятность достоверного события равна единице. Это означает, что событие, имеющее вероятность единица, наступит в любом случае. Вспомним событие «Выпадет Орел или Решка, но что-нибудь точно выпадет, если монета будет таки подброшена».

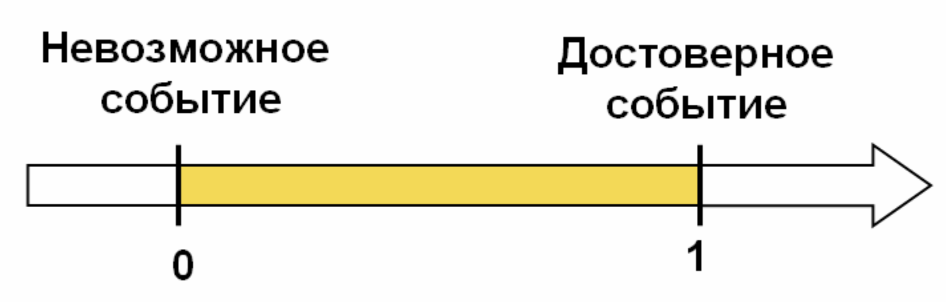


Рисунок 4-4 Вероятность измеряется по шкале от 0 до 1

Рисунок 4-5. - Профессор! Я подбросил монету 100 раз, но не записал результаты эксперимента… - Ничего, в следующий раз подбрасывайте вместо монеты куриные яйца!

Свойство 2. Вероятность невозможного события равна нулю. События «Монета после падения встанет на ребро» или «Повиснет в воздухе» имеют нулевую вероятность, поскольку являются невозможными.

Свойство 3. Вероятность любого события не может быть меньше нуля и больше единицы: $0 \le p(A) \le 1$.

Вероятность выступает своего рода мерой для случайных событий. Каждому случайному событию ставится в соответствие одно единственное число от 0 до 1 включительно, которое называется вероятностью этого события. У одних событий вероятность больше, чем у других. Одна из целей обучения - научиться правильно находить вероятности.

Решим задачу. Если подбросить две монеты, то с какой вероятностью выпадет хотя бы один орел? Начинаем размышлять. Возможны следующие исходы: на обеих монетах орел, на одной орел, а на другой решка, и, наконец, на обеих решка – всего три элементарных исхода. Благоприятными для нас являются первый и второй. Следовательно, вероятность выпадения хотя бы одного орла равна 2/3. Правильно? Нет. Допущена серьезная ошибка. Один из исходов не является равновозможным с другими. Выпадение орла на одной монете и решки на другой может произойти двумя способами. Это означает, что элементарных исходов всего четыре и вероятность равна 3/4, поскольку три исхода являются благоприятными.

Какова вероятность при бросании двух игральных костей получить в сумме пять очков? Всего имеется $6 \times 6 = 36$ элементарных исходов, поскольку каждая кость имеет 6 граней. Из этих 36 исходов благоприятными являются сочетания 1+4, 2+3, 3+2, 4+1, всего четыре. Вероятность равна 4/36 или 1/9.

Классическое определение вероятности обладает несомненными преимуществами – простотой и доступностью. Недостатком является невозможность его применения, если выделение равновозможных элементарных исходов не представляется возможным. Попробуйте применить классическое определение к ситуации со стаканом, падающим со стола. Какие исходы равновозможны? Таковых нет. В этих случаях приходится использовать другие, более пригодные определения понятия вероятности.

Статистическое определение вероятности

Статистический способ определения вероятности состоит в том, что мы изучаем частоту появления интересующего нас события А в серии испытаний и на этой основе получаем его вероятность. При увеличении числа опытов частота все более теряет свой случайный характер, приобретает свойство устойчивости, приближаясь с незначительными колебаниями к некоторой постоянной величине. Это означает, что для некоторых событий последовательность частот имеет предел, который называется *статистической вероятностью*.

Вероятность события А – предельная относительная частота появления события А при проведении серии испытаний, при неограниченном увеличении их числа (статистическое определение вероятности).

$$P(A) = \lim_{n \to \infty} \frac{s}{n},$$

где s – число испытаний, в которых произошло событие А,
n – общее число испытаний.

Статистическое определение хорошо подходит к проверке «правила бутерброда». Правило утверждает, по сути, что вероятность падения бутерброда маслом вниз равна единице. Мы должны провести серию из большого числа испытаний, чтобы понять, стремится ли к единице относительная частота падения бутерброда маслом вниз.

Отметим, что далеко не все события обладают статистической устойчивостью. Кроме этого, вряд ли нам удастся организовать серию экспериментов, чтобы понять вероятность события «Завтра пойдет дождь». Тем самым, еще более сужается круг событий и явлений, для которых удается вычислить их вероятность. Останется единственная соломинка – субъективное определение вероятности.

Субъективное определение вероятности

Если объективный подход к определению вероятности невозможен, единственный путь – субъективная оценка. Субъективная вероятность основана на индивидуальном или коллективном мнении людей, выступающих в роли экспертов. Они составляют свои оценки вероятности события на основе внешней, возможно, неточной, информации, а также своего опыта и интуиции.

Субъективная вероятность отражает степень уверенности индивида или группы в том, что данное событие произойдет.

Кроме перечисленных определений понятия вероятности имеются и другие. В математике используется так называемое аксиоматическое определение вероятности, которое было предложено академиком А.Н.Колмогоровым и является до настоящего времени единственным строгим и универсальным определением, пригодным для построения и развития теории вероятностей. Мы не приводим этого определения в нашем курсе.

Формулы комбинаторики

Для вычисления вероятности иногда приходится использовать несколько важных формул из комбинаторики. Комбинаторика является частью математики, которая занимается методами решения задач, связанных с перечислением и подсчетом. В комбинаторике имеются формулы для определения числа подмножеств заданного множества, подсчета числа перестановок, размещений и сочетаний.

Принцип произведения. Если одно множество состоит из n различных элементов, другое из m различных элементов, и эти множества не пересекаются, то сколько различных пар можно образовать из элементов этих множеств, если первый элемент берется из первого множества, а второй – из второго? Согласно принципу произведения количество пар будет равно $n \times m$.

Перестановки. Сколькими способами n разных объектов могут быть расположены на одной линии?

Например, сколькими способами 6 человек могут сесть на шесть стульев? Чтобы подсчитать, можно размышлять следующим образом. Для первого существует 6 возможностей, для второго, после того как первый уже выбрал, останется всего 5, для следующего – 4 и так далее. Последний, шестой, после пятерых будет иметь только одну возможность. Итак, $6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = 720$. Будем использовать обозначение $6!$ для записи таких произведений (произносится: шесть факториал). В общем виде количество перестановок из n элементов обозначается P_n и вычисляется по формуле:

$$P_n = n!$$

Размещения. Сколькими способами из n разных объектов можно выбрать упорядоченное подмножество из m объектов? Упорядоченным считается множество, в котором задан порядок элементов. Объекты после выбора не возвращаются и повторно не могут быть выбраны.

Например, сколькими способами из 6 человек можно выбрать четверых и рассадить на четыре стула? Способ подсчета аналогичен предыдущему. На первый стул сядет любой из шести, на следующий – уже из пяти. Всего четыре стула, поэтому: $6 \cdot 5 \cdot 4 \cdot 3 = 360$. В общем виде, количество возможных размещений из n элементов по m обозначается A_n^m и рассчитывается по формуле:

$$A_n^m = n(n-1)...(n-m+1) = \frac{n!}{(n-m)!}$$

Сочетания. Сколькими способами из n разных объектов можно выбрать m объектов? Выбор не упорядочен. Объекты после выбора не возвращаются.

Например, сколькими способами из шестерых человек можно выбрать четверых? Несложные размышления приведут к тому, что следует модифицировать формулу для размещений, а именно, отказаться от упорядоченности выбранных элементов, что будет стоить нам $m!$ в знаменателе. Количество сочетаний для множества из n элементов по m элементов определяется по формуле:

$$C_n^m = \frac{n!}{m!(n-m)!}$$

Выбор с повторением. Сколькими способами из n разных объектов можно выбрать m объектов с повторением? Объекты после выбора возвращаются.

Например, сколькими способами из шестерых человек можно выбрать четверых для дежурства, если можно выбирать с повторением и, потенциально, один из шестерых может быть выбран все четыре раза? В этом случае для каждого выбора у нас имеются все шесть кандидатов. Получим $6 \cdot 6 \cdot 6 \cdot 6 = 6^4$. В общем виде, количество способов для множества из n элементов по m элементов определяется по формуле: n^m.

4-2 Сложение и умножение вероятностей

Правила сложения

Чему равна вероятность суммы событий? Формула зависит от того, пересекаются они или нет. Для начала рассмотрим ситуацию, когда события не пересекаются.

Правило сложения для несовместных событий. Если события несовместны, то вероятность суммы этих событий равна сумме их вероятностей.

Вероятность суммы несовместных событий находится по формуле:

$$P(A + B) = P(A) + P(B)$$

Например, если вероятность выпадения единицы при подбрасывании игральной кости равна 1/6, а вероятность выпадения четного числа равна ½, то вероятность суммы этих событий равна 1/6 + ½ = 4/6. Мы воспользовались формулой сложения, поскольку события несовместны: выпадение четного числа и выпадение единицы не могут произойти одновременно.

В случае, если события пересекаются, сумма их вероятностей не равна вероятности суммы событий. Тогда следует воспользоваться правилом сложения для совместных событий.

Правило сложения для совместных событий. Если два события совместны, то вероятность их суммы находится как сумма вероятностей этих событий минус вероятность их пересечения.

Вероятность суммы совместных событий находится по формуле:

$$P(A+B) = P(A) + P(B) - P(AB)$$

Например, если из колоды в 32 карты вынимается одна, рассмотрим два события:

Событие A={карта красной масти}
Событие B={карта является королем или тузом}

Эти события совместны, поскольку эти события могут произойти одновременно. Тогда вероятность такого события определяется следующим образом.

P(A) = 16/32 = 1/2
P(B) = 8/32 = 1/4
P(AB) =P{взят туз или король красной масти}= 4/32 = 1/8

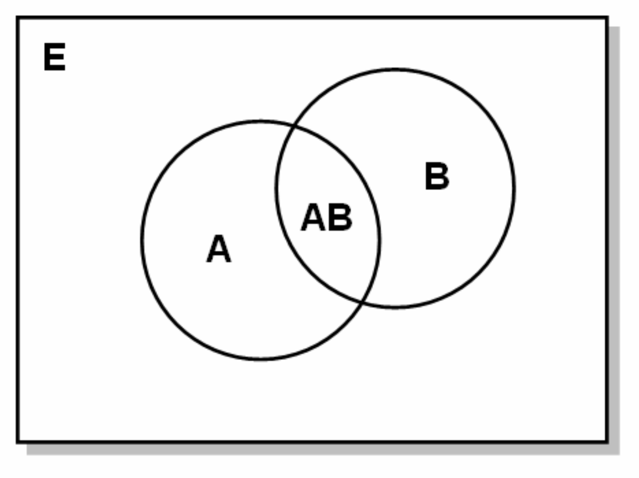


Рисунок 4-6. Совместные события

Далее по формуле сложения для совместных событий:

$$P(A+B) = P(A) + P(B) - P(AB) = \frac{1}{2} + \frac{1}{4} - 1/8 = 5/8$$

Получили, что вероятность того, что нам достанется туз или король красной масти равна 5/8.

Второй пример. В аудитории на научном семинаре присутствуют 5 экономистов и 8 философов. Среди них 7 философов и 3 экономиста женщины. Какова вероятность того, что случайно выбранный участник семинара окажется философом или мужчиной?

	Женщины	Мужчины	Всего
Философы	7	1	8
Экономисты	3	2	5
Всего	10	3	13

Находим вероятность по формуле:

P(философ или мужчина) =
= P(философ) + P(мужчина) – P(философ-мужчина) =
= 8/13 + 3/13 – 1/13 =
= 10/13

Получили, что выходящий из аудитории человек окажется философом или мужчиной с вероятностью 10/13.

Условная вероятность. Правила умножения

По отношению к событиям рассматривают кроме операции сложения произведение событий. Произведение событий есть их пересечение, то есть включает все исходы, которые включают оба события и обозначается AB. Вероятность произведения двух событий находится в зависимости от того, являются эти события зависимыми или нет.

События называются **независимыми**, если появление любого из них не влияет на вероятность появления другого. Если события не являются независимыми, то говорят, что они **зависимы**.

Для независимых событий выполняется равенство:

$$P(AB) = P(A) \cdot P(B)$$

Правило умножения для независимых событий: вероятность произведения двух независимых событий равна произведению их вероятностей.

В случае, если вероятность события B зависит от того, произошло или не произошло событие A, мы имеем дело с зависимыми событиями.

Условной вероятностью P(B/A) называется вероятность события B при условии, что событие A наступило.

Условную вероятность можно найти по формуле:

$$P(B / A) = \frac{P(AB)}{P(A)}$$

Из этой формулы, в частности, вытекает формула умножения для зависимых событий:

$$P(AB) = P(A) \cdot P(B / A)$$

Правило умножения для зависимых событий: вероятность произведения двух событий равна произведению вероятности одного из них на условную вероятность другого, при условии, что первое событие произошло.

Рассмотрим следующий пример. В урне находится десять шаров, из них три белых и семь черных. Выбираем один шар и, не возвращая его в урну, выбираем второй шар. Определим вероятность того, что оба выбранных шара окажутся белыми.

Мы рассматриваем вероятности следующих событий:

Событие A = {первый шар белый}
Событие B = {второй шар белый}

Вероятность первого события $P(A) = 3/10$. Вероятность второго события зависит от того, какой шар выбран первым. Если наступило событие A, то в урне осталось 2 белых шара и 7 черных. Вероятность вынуть второй белый шар в этом случае равна $P(B/A) = 2/9$. Событие AB состоит в том, что оба выбранных шара имеют белый цвет. Найдем искомую вероятность:

$$P(AB) = P(A) \cdot P(B/A) = \\ = 3/10 \cdot 2/9 = \\ = 1/15$$

Подведем итоги. Мы рассмотрели два различных варианта вероятности для произведения событий – для независимых событий и зависимых.

СОБЫТИЯ НЕЗАВИСИМЫ:	СОБЫТИЯ ЗАВИСИМЫ:
$P(B/A) = P(B)$	$P(B/A) \neq P(B)$
$P(AB) = P(A) \cdot P(B)$	$P(AB) = P(A) \cdot P(B/A)$

Формула $P(B/A) = P(B)$ показывает, что вероятность события B не зависит от того, произошло или не произошло событие A, именно это мы и подразумеваем интуитивно под независимостью события B от события A.

Противоположные события

Противоположное событие $\overline{A}$ включает все элементарные исходы, которые не включает A.

Вероятность противоположного события находится по формуле:

$$P(\overline{A}) = 1 - P(A)$$

Сумма двух событий: A и противоположного, является достоверным событием:

$$P(A) + P(\overline{A}) = 1$$

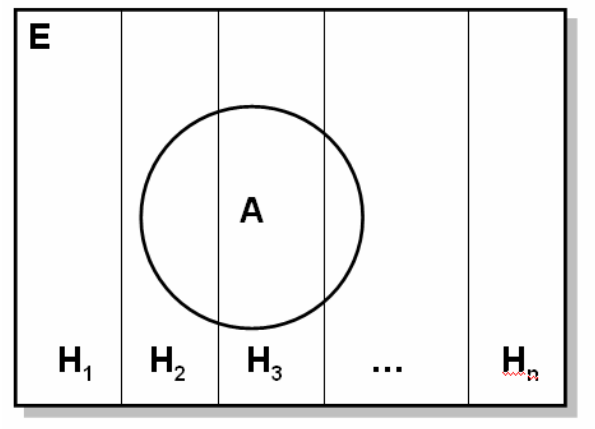


Рисунок 4-7. Формула полной вероятности

4-3 Формула полной вероятности. Формула Байеса

Для некоторых приложений важно уметь использовать формулу, называемую формулой полной вероятности. При наличии полной группы событий, вероятность случайного события А находится по *формуле полной вероятности:*

$$P(A) = P(H_1) \cdot P(A/H_1) + P(H_2) \cdot P(A/H_2) + \ldots + P(H_n) \cdot P(A/H_n)$$

Рассмотрим пример решения задачи при помощи этой формулы. Пусть имеются две урны с белыми и черными шарами.

	Черные шары	Белые шары	Всего
Первая урна	3	2	5
Вторая урна	5	3	8
Всего	8	5	13

Из первой урны случайным образом переложили во вторую 2 шара, а затем из второй вынули один шар. Требуется определить вероятность того, что будет взят белый шар. Эта вероятность зависит от того, какие шары были переложены. Поскольку это нам неизвестно, выделим полную группу событий и найдем их вероятности:

H_1={переложили два белых шара}
H_2={переложили один белый и один черный шар}
H_3={переложили два черных шара}

$P(H_1) = 2/5 \cdot 1/4 = 1/10 = 0{,}1$
$P(H_2) = 2/5 \cdot 3/4 + 3/5 \cdot 2/4 = 6/10 = 0{,}6$
$P(H_3) = 3/5 \cdot 2/4 = 3/10 = 0{,}3$

Теперь находим условные вероятности:

$$P(A/H_1) = 5/10 = 0{,}5$$
$$P(A/H_2) = 4/10 = 0{,}4$$
$$P(A/H_3) = 3/10 = 0{,}3$$

Подставляем полученные результаты в формулу полной вероятности и получим ответ:

$$P(A) = P(H_1)\cdot P(A/H_1) + P(H_2)\cdot P(A/H_2) + P(H_3)\cdot P(A/H_3) =$$
$$= 0{,}1 \cdot 0{,}5 + 0{,}6 \cdot 0{,}4 + 0{,}3 \cdot 0{,}3 =$$
$$= 0{,}38$$

Получили, что шар окажется белым с вероятностью 0,34.

Мы можем решить обратную задачу. Если нам достоверно известно, что вынутый шар оказался белым, то какова вероятность, что переложили два белых шара? Нам требуется в этом случае, зная, что P(A)=0,34, найти вероятность события $P(H_1/A)$. Для этого используется *формула Байеса*:

$$P(H_1 / A) = \frac{P(H_1) \cdot P(A / H_1)}{P(A)}$$

Формула была открыта британским министром Томасом Байесом (1702-1761) впервые опубликована в 1747 году и носит его имя. Зная окончательный исход, можно определить вероятность того, что этот исход явился результатом предыдущего события.

Решаем нашу задачу по этой формуле.

$$P(H_1 / A) = \frac{P(H_1) \cdot P(A / H_1)}{P(A)} =$$
$$= \frac{0{,}1 \cdot 0{,}5}{0{,}38} = 0{,}132$$

Тем самым, если известно, что в результате испытания вынут белый шар, то вероятность того, что перед этим переложили два белых шара, равна 13,2%.

Используем компьютер

Эта глава - теоретическая и не предполагает активного использования компьютера. Следует научиться вычислять комбинаторные формулы при помощи функций, имеющихся в электронных таблицах, поскольку ручные вычисления занимают неоправданно много времени.

Что означают термины

Элементарный исход	**Сумма случайных событий**	**Перестановки**
Пространство элементарных исходов	**Произведение случайных событий**	**Размещения**
Случайное событие	**Несовместные события**	**Сочетания**
Достоверное событие	**Полная группа событий**	**Независимые события**
Невозможное событие	**Классическое определение вероятности**	**Условная вероятность**
Равновозможные события	**Статистическое определение вероятности**	**Принцип сложения вероятностей**
Противоположное событие	**Субъективное определение вероятности**	**Принцип умножения вероятностей**

Символы и формулы

$P(A)=\dfrac{m}{n}$	Классическое определение вероятности
$P(A)=\lim\limits_{n\to\infty}\dfrac{s}{n}$	Статистическое определение вероятности
$P_n=n!$	Перестановки
$A_n^m=\dfrac{n!}{(n-m)!}$	Размещения
$C_n^m=\dfrac{n!}{m!(n-m)!}$	Сочетания
$P(A+B)=P(A)+P(B)-P(AB)$	Правило сложения для совместных событий
$P(AB)=P(A)\cdot P(B)$	Правило умножения для независимых событий

$P(AB) = P(A) \cdot P(B/A)$	Правило умножения для зависимых событий
$P(\overline{A}) = 1 - P(A)$	Вероятность противоположного события
$P(A) = \sum_{i=1}^{n} P(H_i) \cdot P(A/H_i)$	Формула полной вероятности
$P(H_1/A) = \dfrac{P(H_1) \cdot P(A/H_1)}{P(A)}$	Формула Байеса

Задачи и упражнения

4-1. Костюмы для девушки. В гардеробе девушки висят три юбки, пять блузок и четыре шарфика. Сколько различных костюмов может составить девушка, если считать, что цвета одежды хорошо сочетаются друг с другом?

4-2. Игральные кости. Брошено семь игральных костей. Сколько выпадет различных вариантов? (Варианты 1-1-2 и 1-2-1 различны).

4-3. Три предмета. На полке лежат различные книги: 8 по статистике, 6 по геометрии и 3 по алгебре. Студент должен выбрать одну книгу по каждому предмету. Сколько у него возможностей?

4-4. Фильмы кинофестиваля. На кинофестивале будет представлено восемь фильмов. Сколько существует вариантов последовательности их показа?

4-5. Подбрасывание монеты. Монета брошена шесть раз. Сколько возможно различных исходов?

4-6. Выбор двух сотрудников. Президент компании должен выбрать менеджера и ассистента менеджера в каждом из двух отделений фирмы, при этом перемещать сотрудников нельзя. На первом работает 9 человек, на втором – 7. Сколько вариантов у президента для выбора?

4-7. Выбор трех сотрудников. В списке директора находится девять человек. Сколькими способами можно назначить руководителя, помощника руководителя и бухгалтера?

4-8. Статистика. Сколько различных перестановок можно составить из букв слова *СТАТИСТИКА*?

4-9. Контрольные работы. Преподаватель имеет пять вариантов контрольных для пяти разных групп. Сколькими способами он может распределить контрольные работы по группам?

4-10. Группа из четырех человек. Сколькими способами можно составить группу из 4 человек, если всего имеется десять человек?

4-11. Две монеты сто раз. Взяв две монеты, подбросьте их 100 раз и запишите, сколько раз выпадет «орел» при каждом подбрасывании (0, 1 или 2). Вычислите теоретическую вероятность каждого исхода и сравните с эмпирическими результатами.

4-12. Сотрудники супермаркета. В супермаркете работают 2 менеджера, 3 начальника отдела, 5 продавцов-консультантов и 8 кассиров. Найти вероятность того, что выбранный наугад работник магазина окажется кассиром или менеджером.

4-13. Заседание кафедры. На заседании кафедры присутствовали 7 преподавателей математики, 5 преподавателей информатики, 3 преподавателя статистики и 4 научных сотрудника. Найти вероятность того, что наугад выбранный участник заседания окажется преподавателем математики или научным сотрудником.

4-14. Большой театр или зоопарк. Вероятность того, что турист посетит Зоопарк, составляет 0,80, а вероятность посещения Большого театра равна 0,55. В Зоопарк и Большой театр турист попадет в один и тот же день с вероятностью 0,42. Найти вероятность того, что турист посетит или Зоопарк, или Большой театр.

4-15. Новая библиотека. Студентам университета было предложено ответить на вопрос: «Будете ли вы посещать новую библиотеку?». Результаты опроса сведены в таблицу:

Курс	Да	Нет	Не знаю
I	15	8	6
II	25	3	2

Найти вероятность того, что наугад выбранный студент:
a. Ответил «не знаю».
b. Ответил «нет» или является студентом первого курса.
c. Ответил «да» или является студентом второго курса.

4-16. Три телевизионных канала. По трем телевизионным каналам идут телевикторины, комедии и сериалы. Считаем, что все передачи имеет одинаковую продолжительность. Найти вероятность того, что случайно включенный телевизор будет показывать:

a. Телевикторину или канал A.
b. Сериал или комедию.
c. Сериал или канал C.

	A	B	C
Телевикторины	5	2	1
Комедии	3	2	8
Сериалы	4	4	2

4-17. Неподготовленный студент. Если на 10 вопросов экзамена студент отвечает наугад «да» или «нет», какова вероятность того, что ответы будут верными?

4-18. Тяжелые пациенты. Изучение пациентов с избыточным весом показало, что 56% из них имеют также повышенное кровяное давление. Определить вероятность того, что у двух произвольно выбранных пациентов с избыточным весом будет повышенное давление.

4-19. Три дня рождения. Какова вероятность того, что трое случайно выбранных человека окажутся рожденными в одном и том же месяце?

4-20. Дни рождения в марте. Какова вероятность того, что у троих случайно выбранных человек день рождения будет в марте?

4-21. Математика и социология. Вероятность того, что студент увлекается математикой и социологией одновременно равна 0,092. Вероятность того, что студент увлечен социологией, равна 0,73. Найти вероятность того, что студент увлечен математикой, если известно, что он увлекается социологией.

4-22. Пиво и бильярд. В местном баре 72% посетителей играют в бильярд и пьют пиво. В бильярд играют 80% посетителей. В баре случайным образом выбирают посетителя. Найти вероятность того, что он пьет пиво, если известно, что он играет в бильярд.

4-23. Успеваемость студентов. Имеются результаты успеваемости двух отделений социологического факультета:

ОТДЕЛЕНИЕ	УДОВЛ.	ХОР.	ОТЛ.
Социология	5	8	15
Менеджмент	7	12	8

a. Найдите вероятность того, что студент учится на отделении социологии, если он – отличник.

b. Если известно, что студент учится на отделении менеджмент, то какова вероятность, что он – троечник?

4-24. Некачественные пакетики. Два производителя обеспечивают ресторан продуктами питания. Производитель А поставляет 2400 пакетиков с супом, 3% из которых не пригодны для продажи. От производителя В ресторан получает 3600 таких же пакетиков, но с 1% брака. а) Найдите вероятность того, что случайно выбранный пакетик окажется дефектным. б) Если пакетик оказался дефектным, то какова вероятность, что он произведен компанией В?

4-25. Бейсбольные кепки. Магазин закупает бейсбольные кепки у трех различных производителей. В коробке производителя А было 12 синих кепок, 6 красных и 6 зеленых. В коробке производителя В было 10 синих кепок, 10 красных и 4 зеленых. В коробке производителя С - 8 синих, 8 красных и 8 зеленых кепок. Наугад выбрали коробку и кепку из нее. Какова вероятность, что она красная?

4-26. Пробки на дорогах. Водитель может добраться из одного города в другой 3-мя путями. Вероятность попасть в автомобильную пробку на 1-м пути составляет 80%, на 2-м – 60% и 30% на 3-м. Водитель решил проехать 50% времени по первому пути и по 25% по 2-му и 3-му. Водитель позвонил и проинформировал диспетчера о том, что попал в пробку. Какова вероятность того, что это произошло на 1-м пути?

4-27. Сотрудники в лаборатории. В лаборатории работают восемь сотрудников: трое мужчин и пять женщин. Какова вероятность встретить в лаборатории женщину, если двое сотрудников вышли?

4-28. Дети в бассейне. Еженедельно бассейн посещают трое детей из детского сада и пятеро первоклассников. Двое детей заболели и не придут в бассейн. Какова вероятность, что среди заболевших окажется, по крайней мере, один ребенок из детского сада?

4-29. Неисправные тормоза. Статистика показывает, что 6% автомобилей имеют неисправности с тормозами. Если остановить 5 автомобилей и проверить их техническое состояние, то какова вероятность того, что среди них будет хотя бы один автомобиль с неисправными тормозами?

4-30. Встреча сотрудников факультета. На встрече присутствовало 7 профессоров, 5 доцентов, 6 ассистентов и 12 преподавателей. Для выступления случайным образом выбрали 4-х человек. Найти вероятность того, что среди них окажется хотя бы один профессор.

Глава 5. Вероятностные распределения

В этой главе мы объединим методы описательной статистики (глава 2) и вероятностные методы (глава 4). В случае с подбрасыванием игральной кости в главе 2 мы научились собирать данные, оформлять их в таблицу, строить гистограммы, находить среднее значение и стандартное отклонение. Применяя методы главы 4, мы можем найти теоретическую вероятность для каждого возможного исхода. В этой главе мы объединим оба эти подхода и рассмотрим распределения вероятностей, которые будут описывать возможные исходы и вероятности, с которыми эти исходы могут произойти. Мы начнем с изучения случайной величины, а затем рассмотрим наиболее распространенные вероятностные распределения дискретных случайных величин.

5-1 Случайные величины

В этом параграфе мы обсудим понятие случайной величины, вероятностного распределения, научимся вычислять математическое ожидание и дисперсию случайной величины.

Случайной величиной называют переменную, которая в результате испытания принимает единственное значение, которое зависит от случая и не может быть известно заранее.

Рассмотрим случайную величину, равную количеству мальчиков среди 14 новорожденных. Эта случайная величина может принять одно из следующих возможных значений: 0, 1, 2, ..., 14. Вполне вероятно, что количество мальчиков и девочек окажется равным. В то же время, существует теоретическая вероятность, что среди новорожденных окажутся только мальчики или только девочки. Однако вероятность таких исходов представляется существенно меньшей, чем для других исходов. Определим понятие распределения случайной величины.

Таблица 5.1 Вероятности мальчиков

X (МАЛЬЧИКИ)	P(X)
0	0,000
1	0,001
2	0,006
3	0,022
4	0,061
5	0,122
6	0,183
7	0,209
8	0,183
9	0,122
10	0,061
11	0,022
12	0,006
13	0,001
14	0,000

Вероятностное распределение случайной величины это график, таблица или формула, которые указывают на соответствие между принимаемыми значениями и вероятностью. Соответствие между значениями случайной величины и их вероятностями называется **законом распределения случайной величины**.

Случайная величина в приведенном примере является дискретной. Дадим общее определение дискретной случайной величины.

Дискретная случайная величина принимает конечное или счетное количество значений. Счетное количество может быть бесконечным, но, тем не менее, может быть подсчитано при помощи определенной процедуры. Счетными являются, например, целые числа.

Примером дискретной случайной величины является число очков, выпадающих при бросании кости. Известно, что оно может быть равно от 1 до 6 – всего шесть значений. Оно не может принять значение 5,35 или 2,14. В то же время, если рассмотреть в качестве случайной величины вес ребенка при рождении, он может оказаться равным 3,62 кг или 4,123 кг. Здесь мы имеем дело с непрерывной случайной величиной.

Непрерывная случайная величина, в противоположность дискретной, принимает бесконечное количество значений из определенного непрерывного множества на числовой прямой. Множество значений непрерывной случайной величины несчетно.

В этой главе мы будем изучать только дискретные случайные величины. Непрерывные мы будем рассматривать позже.

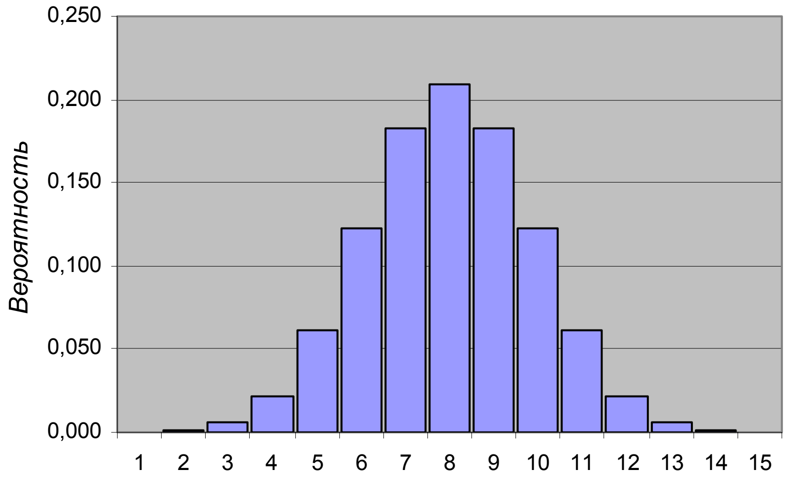


Рисунок 5-1 Мальчики среди 14 новорожденных

Вероятностные распределения можно различными способами изображать графически. Мы приведем пример построения гистограммы для случайной величины новорожденных мальчиков. По горизонтальной оси обозначены возможные значения случайной величины, по вертикальной оси измеряется вероятность для случайной величины принять то или иное значение. Гистограмма дает нам графическое представление о распределении случайной величины.

Для всякой дискретной случайной величины должно выполняться условие:

$$\sum P(x) = 1,\ 0 \le P(x) \le 1,$$

где сумма берется по всем возможным значениям случайной величины. Это условие означает, что с вероятностью 1 случайная величина примет одно из возможных значений.

Для примера проверим, существует ли случайная величина, которая принимает значения 0, 1, 2, 3 с вероятностями, представленными в таблице 5-2?

Таблица 5-2 Это случайная величина?

X	P(X)
0	0,2
1	0,5
2	0,4
3	0,3

Для ответа на вопрос проверим выполнение необходимого условия:

$$\sum P(x) = P(0) + P(1) + P(2) + P(3) =$$
$$= 0{,}2 + 0{,}5 + 0{,}4 + 0{,}3 = 1{,}4$$

Условие равенства единице не выполнено. Это означает, что таблица не описывает никакой случайной величины.

Другой пример. В праздничной лотерее выпущено 100 билетов. Имеется один выигрыш в 1000 рублей и десять выигрышей по 10 рублей, оставшиеся билеты без выигрыша. Определим случайную величину x, равную размеру возможного выигрыша при получении одного билета. Такая случайная величина может принять три значения: 1000, 100 и 0. При этом P(1000) = 0,01; P(100) = 0,1; P(0) = 0,89. Соблюдено условие P(1000)+P(100)+P(0)=1.

5-2 Числовые характеристики случайных величин

В этом параграфе мы определим основные числовые характеристики дискретной случайной величины.

Математическое ожидание

Математическое ожидание случайной величины есть ее среднее значение.

Математическое ожидание дискретной случайной величины вычисляется по формуле:

$$M(X) = \sum \left(x \cdot P(x)\right)$$

Математическое ожидание обладает следующими свойствами.

Свойство 1. Математическое ожидание постоянной величины равно этой величине: MC=C.

Свойство 2. Постоянную можно выносить: M(CX)=CM(X).

Свойство 3. Математическое ожидание суммы случайных величин равно сумме их математических ожиданий: M(X+Y)= M(X)+M(Y).

Свойство 4. Математическое ожидание произведения независимых случайных величин равно произведению математических ожиданий этих величин: M(X $\cdot$ Y)= M(X) $\cdot$ M(Y).

Таблица 5-3 Страхование жизни

X	P(X)
10	0,992
-990	0,008

Для примера вычислим математическое ожидание для следующей случайной величины. Согласно статистике в одной из стран вероятность того, что двадцатипятилетний человек проживет еще год, равна 0,992. Страховая компания предлагает таким молодым людям застраховать жизнь на сумму $1000 сроком на один год с уплатой $10 страхового взноса. Какую прибыль ожидает компания от страховки одного человека?

Рассмотрим в качестве случайной величины прибыль от страховки одного человека. Она принимает два значения: 10, если человек жив к моменту, когда срок страхования истек, и значение –990=10-1000 в противоположном случае.

Распределение случайной величины записано в таблице 5.3. Подсчитаем математическое ожидание этой случайной величины:

$$M(X) = \sum (x \cdot P(x)) = 10 \cdot 0{,}992 + (-990) \cdot 0{,}008 =$$
$$= 9{,}92 - 7{,}92 = 2{,}0$$

Это означает, что от страховки одного человека кампания ожидает получить в среднем $2 прибыли.

Дисперсия

Дисперсия случайной величины характеризует отклонение случайной величины от ее среднего значения.

Дисперсия дискретной случайной величины вычисляется по формуле:

$$D(X) = \sum \left((x - M(X))^2 \cdot P(x) \right)$$

Из формулы видно, что дисперсия есть среднее значение квадрата отклонения случайной величины от ее среднего значения. Назовем основные свойства дисперсии.

Свойство 1. Дисперсия постоянной величины равна нулю: $D(C)=0$.

Свойство 2. Постоянный множитель можно выносить за знак дисперсии, возведя в квадрат: $D(Cx)=C^2D(x)$.

Свойство 3. Дисперсия суммы *независимых* случайных величин равна сумме дисперсий: $D(x+y)= D(x)+D(y)$.

Если использовать приведенные свойства дисперсии, можно получить другую формулу для дисперсии дискретной случайной величины:

$$D(X) = M(X^2) - (M(X))^2$$

Вторая формула более пригодна для практических вычислений, поскольку позволяет вести более простые и короткие расчеты.

Стандартное отклонение

Стандартное отклонение случайной величины есть квадратный корень из дисперсии.

Стандартное отклонение случайной величины вычисляется по формуле:

$$\sigma = \sqrt{D(x)}$$

Стандартное отклонение имеет ту же размерность, что и случайная величина, в отличие от дисперсии, которая имеет размерность, равную квадрату размерности случайной величины. Например, если случайная величина измеряется в метрах, то дисперсия – в квадратных метрах, стандартное отклонение – в метрах. Это делает стандартное отклонение более пригодным для интерпретации результатов, чем саму дисперсию.

Подсчитаем для примера числовые характеристики для числа новорожденных мальчиков на основании распределения, которое было представлено в таблице 5-1. Для подсчета составим вспомогательную таблицу 5-4. Дополняем таблицу распределения тремя новыми столбцами, сумма по которым понадобится нам для включения результатов в формулу для дисперсии. Если для вычисления числовых характеристик использовать компьютер, то значения внутри электронных таблиц будут рассчитываться автоматически.

Применяя результаты вычислений в таблице, получаем:

$$M(x) = \sum \left(x \cdot P(x) \right) = 6{,}993 \approx 7{,}0$$

$$D(x) = \sum \left((x - \mu)^2 \cdot P(x) \right) = 52{,}467 - 6{,}993^2 = 3{,}565 \approx 3{,}6$$

$$\sigma = \sqrt{D(x)} = \sqrt{3{,}6} \approx 1{,}9$$

Таблица 5-4 Вычисление μ, σ для вероятностного распределения

X	P(X)	X · P(X)	X^2	$X^2 \cdot P(X)$
0	0,000	0,000	0	0,000
1	0,001	0,001	1	0,001
2	0,006	0,012	4	0,024
3	0,022	0,066	9	0,198
4	0,061	0,244	16	0,976
5	0,122	0,610	25	3,050
6	0,183	1,098	36	6,588
7	0,209	1,463	49	10,241
8	0,183	1,464	64	11,712
9	0,122	1,098	81	9,882
10	0,061	0,610	100	6,100
11	0,022	0,242	121	2,662
12	0,006	0,072	144	0,864
13	0,001	0,013	169	0,169
14	0,000	0,000	196	0,000
		$\Sigma[x \cdot P(x)]=6{,}993$		$\Sigma[x^2 \cdot P(x)]=52{,}467$

Мы получили, что для 14 новорожденных математическое ожидание количества мальчиков равно 7; дисперсия 3,6; стандартное отклонение 1,9.

Если считать, что значение случайной величины *редко* отклоняется от среднего значения более чем на ±2σ, то можно определить границы ее изменения. Получим, что минимальное и максимальное значение интервала для значения случайной величины.

максимальное значение $\mu + 2\sigma = 7{,}0 + 2 \cdot 1{,}9 = 10{,}8$
минимальное значение $\mu - 2\sigma = 7{,}0 - 2 \cdot 1{,}9 = 3{,}2$

Этот результат можно интерпретировать следующим образом: среди 14 случайно выбранных новорожденных число мальчиков будет *обычно* находиться в интервале от 3,2 до 10,8. Эти нестрогие соображения мы сможем высказывать в строгой форме несколько позже.

Правило округления

Правило округления результатов вычислений состоит в том, что обычно результат должен иметь на один знак после запятой больше, чем точность случайной величины. Если случайная величина принимает целые значения, среднее значение, дисперсия, стандартное отклонение следует округлять до одного знака после запятой.

Мы следовали этому правилу при вычислении числовых характеристик в приведенном примере.

5-3 Биномиальное распределение

В предыдущем параграфе мы для примера рассмотрели несколько различных вероятностных распределений. В этом параграфе мы сосредоточимся на одном специальном распределении – биномиальном. Оно является очень важным, поскольку мы часто имеем дело с ситуациями, в которых результат случайного исхода описывается в терминах двух категорий: ответ на вопрос анкеты да/нет, пол интервьюируемого мужской/женский, продукция имеет/не имеет отклонений в качестве и т.п.

Схема Бернулли – схема проведения испытаний, которая предполагает соблюдение следующих условий : (1) Проводится определенное фиксированное количество испытаний. (2) Испытания являются независимыми, то есть результат одного не зависит от исходов других. (3) В каждом испытании могут произойти все возможные исходы, которые классифицируются в терминах двух категорий. (4) Вероятности исходов должны быть постоянны для каждого испытания.

Будем рассматривать две возможные категории: «успех» и «неуспех». Пусть исходы, которые являются для нас благоприятными, составляют в совокупности событие А. Мы будем называть эти исходы успехом. Обратное событие $\overline{A}$ составляют такие исходы, появление которых в результате испытания является для нас неблагоприятным.

Обозначения:

$P(A) = p$	вероятность успеха в одном испытании,
$P(\overline{A}) = 1 - p = q$	вероятность неуспеха в одном испытании,
n	общее количество испытаний,
k	количество успехов в n испытаниях, целое число между 0 и n, включительно,
$P(k)$	вероятность получить точно k успехов в n испытаниях.

Слово «успех» применяется в данном случае условно, оно вовсе не обязано означать, что в действительности для нас происходит нечто хорошее. Любая из двух категорий могла бы именоваться успехом, при условии, что вероятность такого события составит p.

Случайная величина, равная количеству успехов при проведении испытаний по схеме Бернулли, имеет **биномиальное распределение**.

В реальных исследованиях, например, при проведении опроса, мы делаем выборку без возвращения. Это означает, что один и тот же респондент не может быть опрошен дважды. В этом случае мы не можем говорить о независимых испытаниях. Подходит ли наша схема в этом случае? Можем ли мы применять эту теорию в случае выборки без возвращения? Примем следующее правило. Будем считать исходы «практически» независимыми (хотя они остаются зависимыми), если объем выборки не превышает 5% от общего числа изучаемых объектов (т.е. $n \le 0{,}05N$). В этом случае разница между результатами для выборки с возвращением и выборки без возвращения оказывается незначительной.

Предположим, в тест включены 4 вопроса, каждый из которых предусматривает 5 различных ответов (а, б, в, г, д), один из которых правильный. Предположим, что неподготовленные студенты будут пытаться «оттадывать» правильные ответы и нас интересует вероятность правильно ответить на три вопроса теста из четырех. Можно ли считать указанное распределение биномиальным? Конечно, да, поскольку выполнены необходимые условия:

(1) проводится фиксированное количество испытаний – четыре,

(2) все испытания независимы, поскольку правильный ответ на один вопрос не зависит от правильного ответа на любой другой вопрос,

(3) в каждом испытании ответ может быть двух категорий: правильным или неправильным,

(4) для каждого испытания имеется пять возможных исходов, один из которых правильный. Вероятность правильного ответа в каждом испытании составляет 1/5 (или 0,2). Тем самым, вероятности постоянны для каждого испытания.

Какие значения принимают в этом примере n, k, p, q? Поскольку тест имеет 4 вопроса, $n = 4$. Мы хотим иметь три правильных ответа, значит, $k = 3$. Вероятность правильно ответить на любой вопрос теста 0,2, следовательно, $p = 0{,}2$. Вероятность «неуспеха»: $q = 0{,}8$.

Теперь мы представим три метода для нахождения вероятностей, соответствующих разным значениям случайной величины x, имеющей биномиальное распределение. Первый метод применяет аналитическую формулу для биномиального распределения и является базисом для двух других методов. Второй метод включает использование Таблицы А-1, а третий метод основан на применении компьютера. Если вы будете использовать компьютер и получать значение вероятности автоматически, мы рекомендуем перед этим несколько раз воспользоваться первым методом, провести вычисления вручную, чтобы понять «логику» вычислений.

Метод 1. Применение формулы биномиального распределения. В биномиальном распределении вероятности могут быть вычислены по формуле:

$$P(k) = C_n^k p^k q^{n-k},$$

где

n = общее количество испытаний,
k = количество успехов в n испытаниях (k =0, 1, 2, …, n),
p = вероятность успеха в одном испытании,
q = вероятность неуспеха в одном испытании ($q = 1 - p$).

Напомним, что сочетания вычисляются по формуле:

$$C_n^k = \frac{n!}{k!(n-k)!}$$

Применяя формулу для биномиального распределения, найдем вероятность дать три правильных ответа на 4 вопроса теста в предыдущем примере, т.е. найдем *P(3)*, если известно, что *n=4, x=3, p=0,2* и *q=0,8*. Подставив в формулу, получим:

$$P(k) = C_n^k \cdot p^k \cdot q^{n-k} = C_4^3 \cdot 0{,}2^3 \cdot 0{,}8^{4-3} =$$
$$= \frac{4!}{3!(4-3)!} \cdot 0{,}008 \cdot 0{,}8 = 4 \cdot 0{,}008 \cdot 0{,}8 = 0{,}0256$$

Вероятность получить три правильных ответа составит 0,0256, что означает, что только два-три студента из ста смогут случайным образом достичь указанного результата.

Метод 2. Применение таблицы. Мы можем легко отыскивать значения вероятностей в таблице А-1. Для начала следует найти n, а затем значения k (в строке) и p (в столбце). Искомое значение вероятности находится на пересечении соответствующей строки и столбца.

Таблица 5-4 Правильные ответы на тест

x	P(X)
0	0,410
1	0,410
2	0,154
3	0,026
4	0,002

Применяя таблицу А-1, находим распределение случайной величины x из обсуждавшегося выше примера.

Из полученного результата видно, что наиболее вероятно не ответить ни на один вопрос теста правильно, либо ответить правильно всего на один вопрос. Только 15 студентам из 100 удается, случайным образом отвечая на вопросы, получить два правильных ответа.

Метод 3. **Применение компьютера**. Мы можем воспользоваться встроенными функциями во многих компьютерных приложениях (Excel, Minitab, SPSS, STATDISK и др.). В данном параграфе мы не будем пояснять детали.

Числовые характеристики биномиального распределения

Числовые характеристики случайной величины, имеющей биномиальное распределение, могут быть получены из общих формул для математического ожидания, дисперсии и стандартного отклонения для дискретной случайной величины. Мы не приводим соответствующие выкладки, оставив это занятие для пытливого читателя.

ЧИСЛОВАЯ ХАРАКТЕРИСТИКА	ЛЮБОЕ ДИСКРЕТНОЕ РАСПРЕДЕЛЕНИЕ	БИНОМИАЛЬНОЕ РАСПРЕДЕЛЕНИЕ
Математическое ожидание	$M(x) = \sum\left(x \cdot P(x)\right)$	$\mu = np$
Дисперсия	$D(x) = \sum\left((x-\mu)^2 \cdot P(x)\right)$	$\sigma^2 = npq$
Стандартное отклонение	$\sigma = \sqrt{Dx}$	$\sigma = \sqrt{npq}$

Вычислим числовые характеристики для количества мальчиков среди 14 новорожденных.

Подставляя *n=14, p=0,5* и *q=0,5*, получим:

$$\mu = np = 14 \cdot 0{,}5 = 7{,}0$$
$$\sigma^2 = npq = 14 \cdot 0{,}5 \cdot 0{,}5 = 3{,}5$$
$$\sigma = \sqrt{npq} = 1{,}87$$

Эти вычисления значительно проще по сравнению с вычислениями, которые мы провели для этого же примера при помощи таблицы 5-4.

5-4 Распределение Пуассона

Распределение Пуассона полезно, когда требуется оценить число появлений некоторого события за определенный промежуток времени. Эти ситуации возникают при появлении очереди из людей, прибытии самолетов в аэропорт, автомобилей на заправку. Мы ограничимся рассмотрением пуассоновской случайной величины, равной числу успехов в определенном интервале.

Распределение Пуассона есть распределение дискретной случайной величины, которая равна числу успехов в определенном интервале. Интервал может измеряться временем, расстоянием, площадью, объемом или другими единицами измерения.

Случайная величина x равна числу успехов в некотором интервале, x = 0, 1, 2 … и может принимать неограниченно большие значения. Успехи случайны, не зависят друг от друга и распределены равномерно внутри рассматриваемого интервала. Вероятности распределения Пуассона находятся по формуле:

$$P(x) = \frac{\mu^x \cdot e^{-\mu}}{x!}$$

Распределение Пуассона имеет следующие параметры:

- Среднее значение μ.
- Стандартное отклонение $\sigma = \sqrt{\mu}$.

Как видно, распределение Пуассона задается одним единственным параметром - значением μ, которое равно среднему количеству появления успехов в рассматриваемом интервале.

Распределение Пуассона принципиально отличается от биномиального распределения следующими особенностями:

- Биномиальное распределение определяется количеством испытаний n и вероятностью p, в то время как распределение Пуассона определяется исключительно средним значением μ.
- Возможные значения случайной величины, имеющей биномиальное распределение, есть 0, 1, 2, … , n. Случайная величина, имеющая распределение Пуассона, может принимать значения 0, 1, 2, … без ограничения верхней границы.

Приведем пример, известный под названием бомбы второй мировой войны. При анализе попаданий самолетов снарядов V-1 Южный Лондон был поделен на 576 зон, каждая площадью 0,25 км^2. Всего произошло 535 попаданий бомб. В предположении, что зона выбрана случайно, какова вероятность двойного попадания? Как много зон будет поражено дважды?

В этом случае распределение Пуассона применимо, поскольку мы имеем дело со случайными исходами (попаданиями снарядов) в некотором интервале (зона площадью 0,25 км^2). Среднее значение попаданий в зону есть:

$$\mu = \frac{\textit{попадания}}{\textit{зоны}} = \frac{535}{576} = 0{,}929$$

Поскольку нас интересует точно двукратное попадание в одну зону, положим x=2 и воспользуемся формулой для распределения Пуассона:

$$P(x) = \frac{\mu^x \cdot e^{-\mu}}{x!} = \frac{0{,}929^2 \cdot 2{,}71828^{-0{,}929}}{2!} = \frac{0{,}863 \cdot 0{,}395}{2} = 0{,}170$$

Вероятность двойного попадания снаряда в определенную зону составляет P(2)=0,170. Основываясь на полученной вероятности, мы ожидаем, что среди 576 зон число пораженных снарядом точно дважды, составит 576·0,170=97,9.

В предыдущем примере мы также можем вычислить вероятности и ожидаемые значения для 0, 1, 2, 3, 4 и 5 попаданий. Мы остановимся на 5, поскольку не ожидается зон, в которые произойдет больше пяти попаданий (вероятность для x>5 равна 0,000 с округлением до трех знаков после запятой). Вероятности и ожидаемые значения количества зон показаны в таблице 5-6. Четвертая колонка содержит реальные данные, собранные во время войны. Как видно из таблицы, имеется 229 зон, в которые не случилось попаданий, в 211 зон случилось однократное попадание, в 7 зон снаряды попали 4 раза.

Таблица 5-6 Попадание снарядов в 576 зон Южного Лондона

ЧИСЛО СНАРЯДОВ	ВЕРОЯТНОСТЬ	ОЖИДАЕМОЕ ЧИСЛО ЗОН	ДЕЙСТВИТЕЛЬНОЕ ЧИСЛО ЗОН
0	0,395	227,5	229
1	0,367	211,4	211
2	0,170	97,7	93
3	0,053	30,5	35
4	0,012	6,9	7
5	0,002	1,2	1

Мы имеем возможность сравнить значения, предсказанные при помощи распределения Пуассона с реальными значениями, наблюдаемыми в ходе войны. Как видно, теоретические значения очень близки фактическим значениям, распределение Пуассона выполняет неплохую работу по прогнозированию результатов.

Приближение биномиального распределения

Распределение Пуассона иногда используется для приближения биномиального распределения, когда число испытаний *n* велико, а вероятность *p* мала. Будем полагать выполнение двух условий: $n \geq 100$ и $np \leq 10$. Если эти условия выполнены, нам необходимо вычислить параметр распределения Пуассона при помощи формулы: $\mu = np$.

Предположим, вы ежедневно играете на одном том же автомате и ожидаете выпадения набора из трех цифр, например, 111. Если вы играете в эту игру один раз в день, найдем вероятность однократного выигрыша в течение 365 дней.

Поскольку интервал времени составляет 365 дней, n=365. Поскольку имеется только один выигрыш из 1000 (от 000 до 999), p=1/1000. Оба условия, $n \geq 100$ и $np \leq 10$, выполнены, поэтому мы можем применить распределение Пуассона в качестве аппроксимации биномиального распределения. Прежде всего, нам необходимо вычислить:

$$\mu = np = 365 \cdot \frac{1}{1000} = 0{,}365$$

Найдя среднее значение, мы найдем теперь значение вероятности P(1):

$$P(x) = \frac{\mu^x \cdot e^{-\mu}}{x!} = \frac{0{,}365^1 \cdot 2{,}71828^{-0{,}365}}{1!} = 0{,}253$$

Применяя распределение Пуассона в качестве приближения биномиального распределения, мы нашли, что с вероятностью 0,253 мы выиграем точно один раз в течение 365 дней. Применяя биномиальное распределение, мы получим более точный ответ, вероятность будет равна 0.254. Как мы видим, приближение при помощи распределения Пуассона является хорошим.

Используем компьютер

Применение компьютера для материалов этой главы полезно для нахождения математического ожидания, дисперсии и стандартного отклонения случайной величины, заданной таблицей распределения. Кроме этого, следует научиться пользоваться компьютером для получения значений вероятности биномиального распределения и распределения Пуассона. Например, в таблицах Excel функция биномиального распределения в русскоязычной версии выглядит так: БИНОМРАСП (k, n, p, INT). Разобраться с использованием этой функции в первый раз помогут подсказки, которые появляются при заполнении полей формы для этой функции. Распределение Пуассона вычисляется при помощи функции ПУАССОН (x, μ, INT). В обоих случаях последнему параметру INT следует присвоить значение INT=ЛОЖЬ.

Что означают термины

Случайная величина	**Непрерывная случайная величина**	**Схема Бернулли**
Вероятностное распределение	**Математическое ожидание**	**Биномиальное распределение**
Закон распределения	**Дисперсия**	**Распределение Пуассона**
Дискретная случайная величина	**Стандартное отклонение**	

Символы и формулы

$M(X)=\sum\left(x\cdot P(x)\right)$	Математическое ожидание дискретной случайной величины
$D(X)=\sum\left((x-M(X))^2\cdot P(x)\right)$	Дисперсия дискретной случайной величины
$D(X)=M(X^2)-(M(X))^2$	Дисперсия дискретной случайной величины – вторая формула
$\sigma=\sqrt{D(x)}$	Стандартное отклонение
$P(k)=C_n^k p^k q^{n-k}$	Вероятность k успехов в n испытаниях в схеме Бернулли
$\mu=np$	Математическое ожидание биномиального распределения
$\sigma^2=npq$	Дисперсия биномиального распределения
$\sigma=\sqrt{npq}$	Стандартное отклонение биномиального распределения
$P(x)=\dfrac{\mu^x\cdot e^{-\mu}}{x!}$	Вероятности распределения Пуассона

Задачи и упражнения

5-1. Количество сыновей. В семье трое детей. Случайная величина X равна количеству сыновей. Постройте ее распределение.

5-2. Выздоровление после операции. В результате некоторой операции от болезни излечиваются 80% пациентов. Если предполагается сделать четыре операции, построить закон распределения случайной величины X, равной количеству выздоровевших.

5-3. Вычисляем числовые характеристики. Найти математическое ожидание и дисперсию случайной величины X, закон распределения которой задан таблицей:

а)

X	2	4	8
P	0,1	0,6	0,3

б)

X	-1	0	2
P	0,2	0,5	0,3

в)

X	-2	1	2	3
P	0,3	0,3	0,3	0,1

г)

X	-1	0	1	3
P	0,1	0,3	0,2	0,4

5-4. Два шара из десяти. В урне 10 шаров, из них два белых. Выбираем наудачу два шара без возвращения. Случайная величина X - число белых шаров в выборке. Составить ее закон распределения и найти математическое ожидание и стандартное отклонение.

5-5. Пиджаки и вероятности. Количество пиджаков, продаваемых в день в салоне одежды, занесено в таблицу. Найдите среднее, дисперсию и среднеквадратичное отклонение распределения.

Кол-во проданных пиджаков, X	8	9	10	11	12
Вероятность, P(X)	0,1	0,2	0,2	0,3	0,2

5-6. Размер ботинок. Администратор обувного магазина вычислил вероятности продажи ботинок каждого размера, как показано ниже. Найдите среднее, дисперсию и среднеквадратичное отклонение для распределения.

Размер, X	40	41	42	43	44	45
Вероятность, P(X)	0,1	0,2	0,3	0,2	0,1	0,1

5-7. Автомобили клиентов. Страховая компания провела исследование, в котором анализировалось число автомобилей у клиентов компании. Найдите среднее, дисперсию и среднеквадратичное отклонение для распределения.

Число автомобилей, X	1	2	3	4
Вероятность, P(X)	0,4	0,3	0,2	0,1

5-8. Реклама в детских программах. Обеспокоенные родители посчитали количество рекламы, показанной в течение пяти детских программ. Найдите среднее, дисперсию и среднеквадратичное отклонение для распределения.

Кол-во рекламы, X	5	6	7	8	9
Вероятность, P(X)	0,2	0,25	0,38	0,10	0,07

5-9. Сколько телевизоров в семье. Исследование, проведенное телекомпанией, показало количество телевизоров в семьях и соответствующие вероятности. Найдите среднее, дисперсию и среднеквадратичное отклонение.

Число телевизоров, X	1	2	3	4
Вероятность, P(X)	0,32	0,51	0,12	0,05

5-10. Коробка с банкнотами. В коробке находятся десять $1 банкнот, пять $5 банкнот, три $20 банкноты, одна $50 банкнота и одна $100 банкнота. Человек платит $20, чтобы вытащить одну банкноту. Найдите математическое ожидание выигрыша. Является ли игра справедливой?

5-11. Игра в кости. Если игрок кидает две игральные кости и получает сумму очков, равную 2 или 12, он выигрывает $20. Если человек набирает 7 очков, он выигрывает $5. Стоимость одной игры – $3. Найдите математическое ожидание выигрыша.

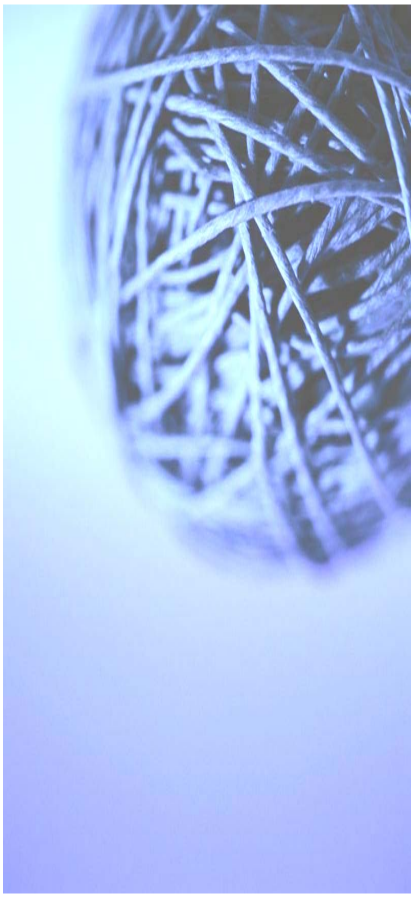
5-12. Лотерея из 100 билетов. В лотерее продано 100 билетов, по 10 рублей каждый. Из них один билет выигрывает 500 рублей, 3 билета по 50 рублей и 10 билетов по 20 рублей. Построить распределение выигрыша. Найти математическое ожидание чистого выигрыша для человека, купившего один билет.

5-13. Лотерея из 1000 билетов. Лотерея предлагает один приз 1000 рублей, один приз 500 рублей, и пять призов 100 рублей. Продали одну тысячу билетов по 3 рубля за билет. Найдите математическое ожидание выигрыша, если приобретен один билет.

5-14. Игра в рулетку. Игровая рулетка состоит из 37 чисел: числа от 1 до 36 и ноль (zero). Половина чисел от 1 до 36 красная, а другая – черная. Шарик останавливается на одном из 37 чисел, определяя номер и цвет. Ноль зеленого цвета. При единичной ставке чистый выигрыш будет следующим:

Красное или черное	1
Нечетное или четное	1
Первая половина (1-18)	1
Вторая половина (19-36)	1
Один любой номер	35
Ноль	35

Найдите математическое ожидание выигрыша для каждого из перечисленных случаев, если ставка равна 5.



Глава 6. Распределения непрерывных случайных величин

Для описания непрерывных случайных величин требуются специальные функции, важнейшими из которых являются функция распределения и плотность. Поскольку непрерывные случайные величины принимают бесконечно большое количество значений и даже несчетное, вероятность того, что непрерывная случайная величина примет определенное значение равна нулю. Важно при помощи функции распределения и плотности находить вероятность того, что случайная величина примет значение в определенном интервале, а также вычислять числовые характеристики непрерывных случайных величин.

6-1 Функция распределения, плотность, числовые характеристики

В этом параграфе мы выясним, каким образом можно описать распределение непрерывной случайной величины, обсудим специальные функции, называемые функцией распределения и плотностью распределения. Напомним, что непрерывные случайные величины принимают бесконечное количество значений из определенного интервала числовой прямой.

Функция распределения

Мы предполагаем, что точность наших измерений непрерывной случайной величины не ограничена в принципе, иначе наши измерения становятся дискретными. Например, рост и вес человека являются непрерывными случайными величинами. Тем не менее, если проводить измерение роста шкалой с делениями в 1 см, мы не сможем отличить двух людей с ростом 178,7 и 179,2 см. Нам придется считать рост дискретной случайной величиной, и признать рост этих людей равным 179 см.

Важно понимать, что на самом деле не существует двух людей, имеющих одинаковый рост. Будучи измеренным с очень высокой точностью, рост двух людей обязательно будет отличаться в каком-то знаке и, значит, будет различным. Ключевое свойство для непрерывной случайной величины – вероятность принять любое определенное значение равно нулю. Поэтому для описания распределения вероятностей непрерывной случайной величины не пользуются вероятностью события $X = x$, а рассматривают вероятность события $X < x$, где – x некоторая текущая переменная. Очевидно, вероятность события $X < x$ зависит от значения переменной x и является функцией от этой переменной. Эта функция называется функцией распределения случайной величины и обозначается $F(x)$.

Функция распределения есть функция F(x), равная для каждого значения x вероятности того, что случайная величина X примет значение меньше x. Функция распределения называется также интегральной функцией или законом распределения.

Символически, определение функции распределения записывается следующим образом:

$$F(x) = P(X < x)$$

Геометрически это означает вероятность того, что значение случайной величины окажется левее заданного значения переменной x. Функция распределения, определяемая таким образом, может быть построена для любой случайной величины, в том числе, и для дискретной. Для дискретной случайной величины функция распределения имеет вид ступенек, а для непрерывной – вид линии, поднимающейся от 0 до 1 непрерывно. Вид ступенек или непрерывной линии зависит, очевидно, от самого распределения случайной величины.

Назовем основные свойства функции распределения.

Свойство 1. Функция распределения есть неотрицательная функция, заключенная между нулем и единицей: $0 \le F(x) \le 1$. Это следует из того, что функция распределения есть вероятность.

Свойство 2. Функция распределения есть неубывающая функция. Это следует из того, что для двух различных значений аргумента, вероятность для правого может быть только больше, поскольку она включает в себя вероятность для левого значения аргумента.

Свойство 3. Функция распределения от минус бесконечности равна нулю. Это следует из того, что вероятность такого события, что наша случайная величина окажется меньше минус бесконечности, равна нулю.

Свойство 4. Функция распределения от плюс бесконечности равна единице. Это следует из того, что с вероятностью единица наша случайная величина примет значение, меньшее плюс бесконечности.

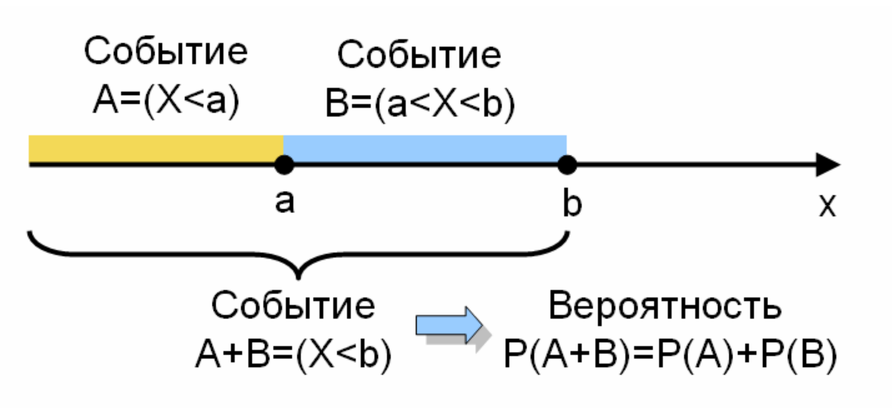


Рисунок 6-1 Вероятности событий А и В

Свойство 5. Вероятность попадания случайной величины в интервал (a, b) равна приращению функции распределения на этом интервале:

$$P(a < X < b) = F(b) - F(a)$$

Последнее свойство является важнейшим и позволяет при помощи функции распределения находить вероятности попадания значения случайной величины в определенный заданный интервал.

Плотность вероятности

С функцией распределения непрерывной случайной величины тесно связана другая очень полезная функция – плотность распределения.

Плотность распределения непрерывной случайной величины есть функция f(x), равная производной от функции распределения. Функция плотности называется также плотностью вероятности, дифференциальной функцией или весовой функцией.

Символически, функция плотности записывается следующим образом:

$$f(x) = F'(x)$$

Назовем основные свойства плотности.

Свойство 1. Плотность распределения есть неотрицательная функция. Это следует из того, что плотность является производной от неубывающей функции.

Свойство 2. Площадь под графиком плотности распределения равна единице. Это следует из того, что площадь под графиком есть несобственный интеграл от функции плотности в границах от минус бесконечности до плюс бесконечности, а он равен значению функции распределения в плюс бесконечности, то есть единице.

Свойство 3. Вероятность попадания случайной величины в интервал (a, b) равна определенному интегралу от плотности в пределах от a до b:

$$P(a < X < b) = \int_a^b f(x)dx$$

Это следует из того, что F(x) есть первообразная и выполняется формула Ньютона-Лейбница (смотри свойство 5 функции распределения).

Математическое ожидание, дисперсия

Назовем теперь числовые характеристики для непрерывных случайных величин, дадим их определение и приведем формулы.

Математическое ожидание непрерывной случайной величины вычисляется как несобственный интеграл от произведения переменной x на плотность f(x) в пределах от минус бесконечности до плюс бесконечности. Обозначается M(X) или символом μ.

Символически, формула вычисления математического ожидания записывается следующим образом:

$$M(X) = \int_{-\infty}^{+\infty} x \cdot f(x)dx$$

Эта формула является прямым аналогом формулы для математического ожидания дискретной случайной величины, только для непрерывной величины обычная сумма по всем возможным значениям превратилась в интегральную сумму.

Дисперсия непрерывной случайной величины вычисляется как несобственный интеграл от произведения квадрата разности переменной x и математического ожидания на плотность f(x) в пределах от минус бесконечности до плюс бесконечности. Обозначается D(X) или σ^2.

Символически, формула вычисления дисперсии записывается следующим образом:

$$D(X) = \int_{-\infty}^{+\infty} (x - M(X))^2 \cdot f(x)dx$$

Кроме дисперсии, определим также стандартное отклонение.

Стандартное отклонение непрерывной случайной величины есть квадратный корень из ее дисперсии.

Таблица 6.1 Соответствие числовых характеристик

ЧИСЛОВАЯ ХАРАКТЕРИСТИКА	ДИСКРЕТНАЯ СЛУЧАЙНАЯ ВЕЛИЧИНА	НЕПРЫВНАЯ СЛУЧАЙНАЯ ВЕЛИЧИНА
Математическое ожидание	$\mu = \sum\left(x \cdot P(x)\right)$	$\mu = \int\limits_{-\infty}^{+\infty} x \cdot f(x)dx$
Дисперсия	$\sigma^2 = \sum\left((x - M(x))^2 \cdot P(x)\right)$	$\sigma^2 = \int\limits_{-\infty}^{+\infty} \left(x - M(X)\right)^2 \cdot f(x)dx$
Стандартное отклонение	$\sigma = \sqrt{D(X)}$	$\sigma = \sqrt{D(X)}$

Стандартное отклонение случайной величины вычисляется по формуле:

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

Очевидно, приведенные определения и формулы есть полные аналоги уже данных в предыдущей главе определений и формул для дискретных случайных величин. Все свойства также сохранены и могут быть использованы. В частности, может быть использована вторая формула для дисперсии, которая приводилась нами как следствие свойств дисперсии.

Типы задач для решения

Назовем основные типы задач, которые следует научится решить на основе изложенной теории.

Задача типа 1. Найти плотность непрерывной случайной величины, если известна ее функция распределения. Плотность находится как производная от функции распределения.

Задача типа 2. По известной плотности непрерывной случайной величины найти функцию распределения этой случайной величины. Функция распределения является первообразной для функции плотности и находится как несобственный интеграл:

$$F(x) = \int\limits_{-\infty}^{x} f(t)dt$$

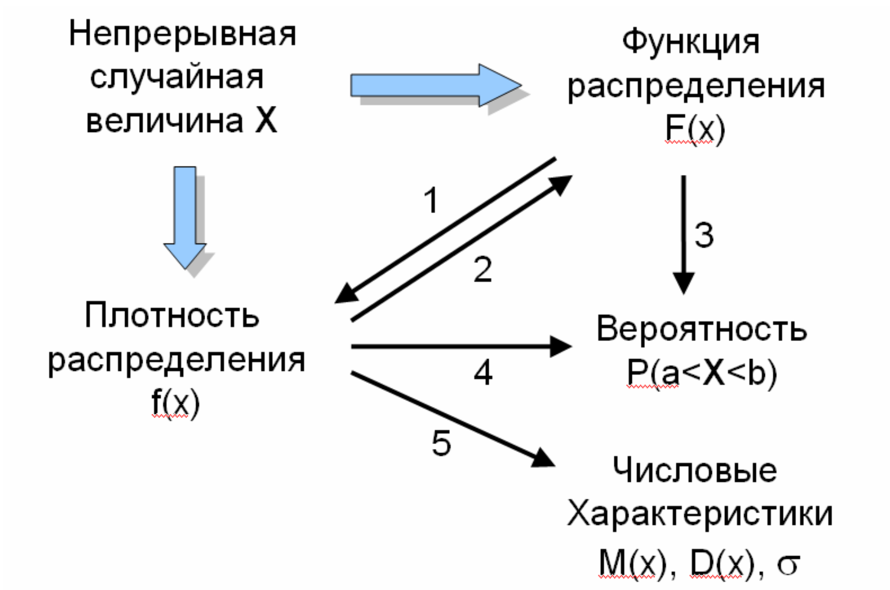


Рисунок 6-2 Основные типы задач

Задача типа 3. По заданной функции распределения непрерывной случайной величины требуется вычислить вероятность того, что случайная величина примет значение в интервале (a,b). Искомая вероятность находится по формуле $P(a < x < b) = F(b) - F(a)$.

Задача типа 4. Известна плотность распределения непрерывной случайной величины. Необходимо вычислить вероятность того, что случайная величина примет значение в интервале (a,b). Искомая вероятность находится по формуле:

$$P(a < X < b) = \int_a^b f(x)dx.$$

Задача типа 5. По заданной плотности распределения непрерывной случайной величины найти ее числовые характеристики: математическое ожидание, дисперсию и стандартное отклонение.

Решим для примера следующую задачу. Дана функция распределения непрерывной случайной величины:

$$F(x) = \begin{cases} 0 & при \quad x \le 0 \\ x^2/4 & при \quad 0 < x \le 2 \\ 1 & при \quad x > 2 \end{cases}$$

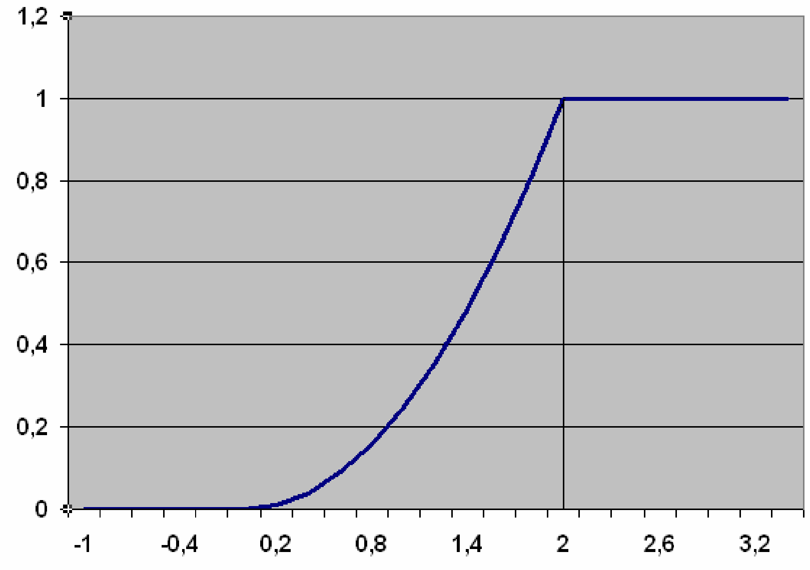


Рисунок 6-3 График функции распределения

Найдем плотность как производную от функции распределения:

$$f(x) = F'(x) = \begin{cases} 0 & при \quad x \le 0 \\ x/2 & при \quad 0 < x \le 2 \\ 0 & при \quad x > 2 \end{cases}$$

Математическое ожидание:

$$M(X) = \int_{-\infty}^{+\infty} x \cdot f(x)dx = \int_{-\infty}^{0} (x \cdot 0)dx + \int_{0}^{2} x \cdot (x/2)dx + \int_{2}^{+\infty} (x \cdot 0)dx =$$

$$= 0 + \left.\frac{x^3}{6}\right|_0^2 + 0 = \frac{2^3}{6} - \frac{0^3}{6} = \frac{4}{3}.$$

Вычислим дисперсию и стандартное отклонение:

$$M(X^2) = \int_{-\infty}^{+\infty} x^2 \cdot f(x)dx = 0 + \int_{0}^{2} x^2 \cdot (x/2)dx + 0 = 2$$

$$D(X) = M(X^2) - (M(X))^2 = 2 - \left(\frac{4}{3}\right)^2 = \frac{2}{9}$$

$$\sigma = \sqrt{D(X)} = \sqrt{\frac{2}{9}} = 0,47$$

Теперь построим графики функции распределения (рисунок 6-3) и график плотности (6-4).

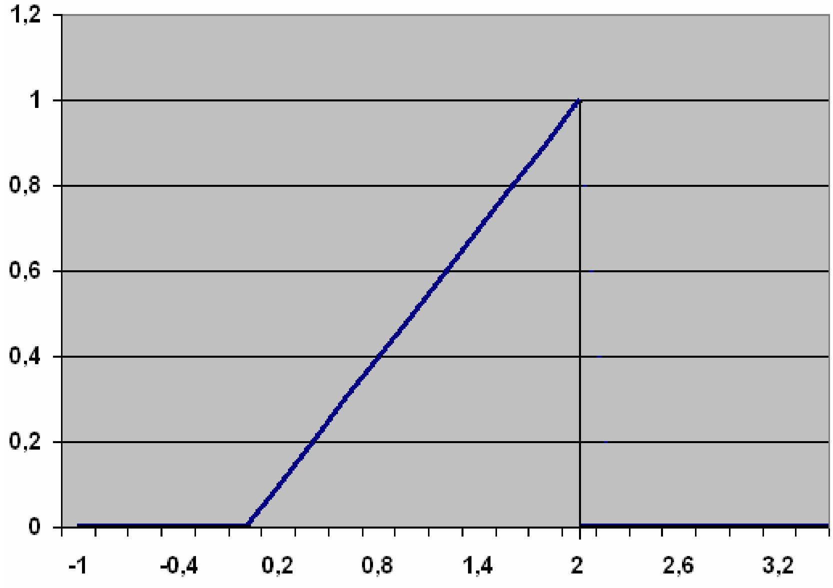


Рисунок 6-4 График плотности распределения

На основе полученной функции плотности или с использованием функции распределения мы можем отыскать вероятность того, что наша непрерывная случайная величина примет значение в интервале. Пусть нас интересует, например, вероятность того, что случайная величина примет значение от 0 до 1.

Воспользуемся формулой и вычислим соответствующий определенный интеграл:

$$P(a < X < b) = \int_a^b f(x)dx = \int_0^1 (x/2)dx = \left.\frac{x^2}{4}\right|_0^1 = 1/4$$

Если бы мы использовали функцию распределения, то получили бы:

$$P(0 < X < 1) = F(1) - F(0) = 1/4 - 0 = 1/4$$

Резюме

В этом параграфе мы рассмотрели основные теоретические понятия, связанные с непрерывными случайными величинами. Функция распределения, которую мы рассмотрели для примера, носит скорее учебный характер. Чаще мы будем иметь дело с другими законами распределения непрерывных случайных величин, которые используются для описания реальных процессов и моделей, и поэтому давно стали классическими. Мы рассмотрим равномерное распределение, показательное распределение, а всю следующую главу посвятим нормальному закону.

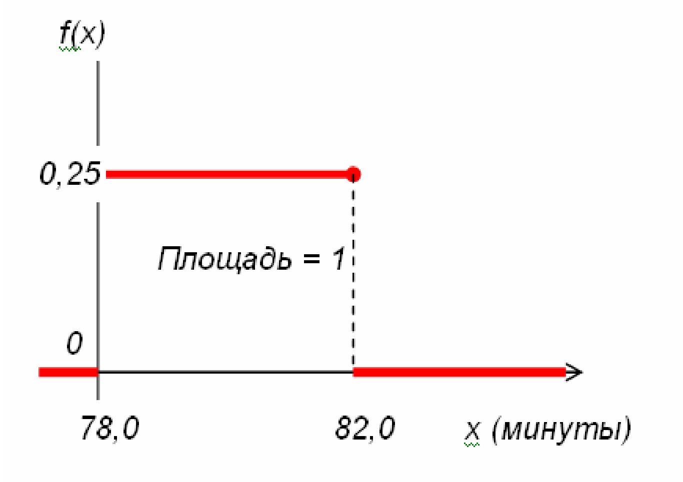


Рисунок 6-5 Продолжительность лекции

6-2 Равномерное распределение

Рассмотрим непрерывное распределение, называемое равномерным.

Непрерывная случайная величина имеет **равномерное распределение**, если ее значения равномерно распределены на отрезке. График равномерного распределения имеет прямоугольную форму.

Как известно, обычная продолжительность лекции для студентов составляет один час двадцать минут, что составляет ровно 80 минут. Предположим, что профессор статистики не носит часов и при этом очень хорошо планирует свои лекции. Это приводит к тому, что продолжительность лекции равномерно распределена на отрезке времени от 78 до 82 минут. Будем считать, что лекция может случайным образом закончиться в любой момент времени в указанном интервале. Какова в этих условиях вероятность того, что лекция задержится более чем на одну минуту?

Мы имеем дело с равномерным законом распределения случайной величины, равной продолжительности лекции. Плотность для такой случайной величины выглядит следующим образом:

$$f(x)=\begin{cases} 0 & при \quad x\le 78 \\ 0{,}25 & при \quad 78<x\le 82 \\ 0 & при \quad x>82 \end{cases}$$

График плотности представлен на рисунке 6-4. Плотность выглядит как прямоугольник высотой 0,25 и имеет сторону 82 – 78 = 4. Площадь такого прямоугольника равна единице.

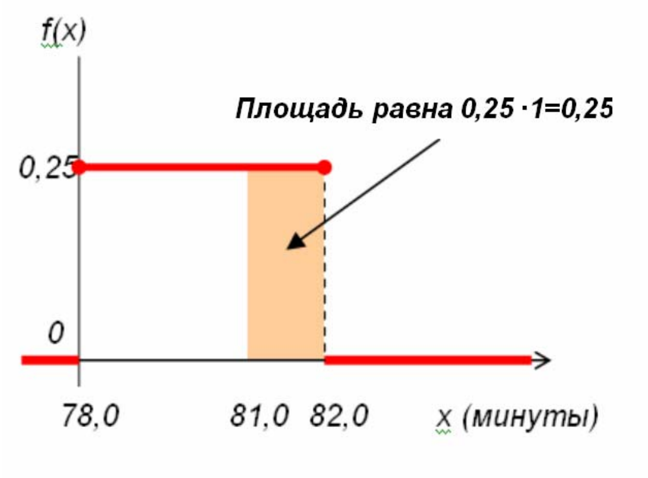


Рисунок 6-6 Нахождение вероятности

Для нахождения вероятности мы можем найти площадь под графиком на интересующем нас интервале времени. Интервал от 81 до 82 минут составляет одну четверть от всего интервала возможных значений случайной величины. Искомая вероятность равна площади прямоугольника высотой 0,25 и длиной 1, т.е. равна 0,25.

В общем виде плотность равномерного распределения на отрезке [a, b] записывается в виде следующей формулы:

$$f(x) = \begin{cases} \dfrac{1}{b-a} & \textit{при} \quad a < x \le b \\ 0 & \textit{вне отрезка} \end{cases}$$

Функция распределения для равномерного закона будет тогда иметь общий вид:

$$F(x) = \begin{cases} 0 \quad \textit{при} \quad x \le a \\ (x-a)(b-a) \quad \textit{при} \quad a < x \le b \\ 1 \quad \textit{при} \quad x > b \end{cases}$$

Числовые характеристики для равномерного распределения:

$$M(X) = \int_a^b \frac{x}{b-a} dx = \frac{a+b}{2}$$

$$D(X) = \frac{1}{b-a} \int_a^b \left(x - \frac{a+b}{2} \right)^2 dx = \frac{(b-a)^2}{12}$$

$$\sigma = \sqrt{\frac{(b-a)^2}{12}} = \frac{b-a}{2\sqrt{3}}$$

Используем компьютер

Материал этой главы не требует длительных практических занятий на компьютере. Для равномерного распределения и вычисления вероятностей не требуются специальные функции, поскольку эти функции могут быть созданы пользователем самостоятельно при помощи несложных формул.

Для непрерывных случайных величин полезно научиться строить графики функции распределения и плотности при помощи электронных таблиц или статистических пакетов. Для этого с определенным шагом задаются значения переменной, а для каждого из этих значений вычисляется соответствующее значение функции. Если шаг выбран небольшим, график получится гладким и правильным.

Что означают термины

Функция распределения **Плотность распределения** **Равномерное распределение**

Символы и формулы

$F(x)=P(X<x)$	Функция распределения случайной величины
$P(a<X<b)=F(b)-F(a)$	Вероятность попадания значения непрерывной случайной величины в интервал
$f(x)=F'(x)$	Плотность распределения непрерывной случайной величины
$P(a<X<b)=\int_a^b f(x)dx$	Вероятность попадания значения непрерывной случайной величины в интервал
$M(X)=\int_{-\infty}^{+\infty} x\cdot f(x)dx$	Математическое ожидание непрерывной случайной величины
$D(X)=\int_{-\infty}^{+\infty}(x-M(X))^2\cdot f(x)dx$	Дисперсия непрерывной случайной величины
$\sigma=\sqrt{D(X)}$	Стандартное отклонение случайной величины
$F(x)=\int_{-\infty}^{x} f(t)dt$	Нахождение функции распределения непрерывной случайной величины по плотности
$f(x)=\begin{cases}\dfrac{1}{b-a} & при \quad a<x\le b\\ 0 & вне\ отрезка\end{cases}$	Плотность равномерного закона распределения
$F(x)=\begin{cases}0 & при \quad x\le a\\ (x-a)(b-a) & при \quad a<x\le b\\ 1 & при \quad x>b\end{cases}$	Функция распределения равномерного закона

Задачи и упражнения

6-1. Вероятность попадания в интервал. Найти вероятность того, что в результате испытания случайная величина X примет значение в интервале (0; 1), если она задана следующей функцией распределения:

а) Найти $P(0 < X < 1)$, если $F(x) = \begin{cases} 0, & -\infty < x \le -1 \\ \frac{1}{2}x + \frac{1}{2}, & -1 < x \le 2 \\ 1, & 2 < x \end{cases}$

б) Найти $P\left(0 < X < \frac{1}{3}\right)$, если $F(x) = \begin{cases} 0, & -\infty < x \le -1 \\ \frac{3}{4}x + \frac{3}{4}, & -1 < x \le \frac{1}{3} \\ 1, & \frac{1}{3} < x \end{cases}$

в) Найти $P\left(\frac{1}{4} < X < \frac{3}{4}\right)$, если $F(x) = \begin{cases} 0, & -\infty < x \le 0 \\ x^2, & 0 < x \le 1 \\ 1, & 1 < x \end{cases}$

г) Найти $P(0 < X < 1)$, если $F(x) = \frac{1}{2} + \frac{1}{\pi} arctg \quad x$

6-2. Найти плотность по функции распределения. Найти плотность случайной величины, если она задана следующей функцией распределения:

а) $F(x) = \begin{cases} 0, & -\infty < x \le 0 \\ \sin x, & 0 < x \le \frac{\pi}{2} \\ 1, & \frac{\pi}{2} < x \end{cases}$

б) $F(x) = \begin{cases} 0, & -\infty < x \le 0 \\ \sin 2x, & 0 < x \le \frac{\pi}{4} \\ 1, & \frac{\pi}{4} < x \end{cases}$

6-3. Найти функцию распределения по плотности. Найти функцию распределения случайной величины, если известна ее плотность:

а) $$f(x)=\begin{cases} 0, & -\infty < x \le 0 \\ \sin x, & 0 < x \le \dfrac{\pi}{2} \\ 0, & \dfrac{\pi}{2} < x \end{cases}$$

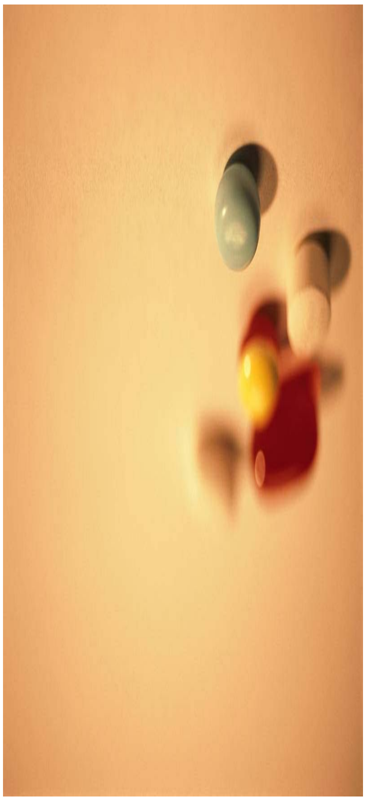
б) $$f(x)=\begin{cases} 0, & -\infty < x \le 1 \\ \dfrac{2}{3}x, & 1 < x \le 2 \\ 0, & 2 < x \end{cases}$$

6-4. Равномерное распределение. Найти математическое ожидание, дисперсию и стандартное отклонение случайной величины, равномерно распределенной на интервале:

а) (1; 5)
б) (-1; 3)

6-5. График равномерного распределения. Построить график плотности и график функции распределения для равномерно распределенной на интервале (1; 2) случайной величины.

6-6. Автобусная остановка. Автобус некоторого маршрута идет строго по расписанию с интервалом в 8 мин. Найти вероятность того, что пассажир, подошедший к остановке, будет ожидать очередной автобус менее трех минут.



Глава 7. Нормальное распределение

Значение нормального закона чрезвычайно важно, поскольку это распределение встречается в очень большом числе приложений и играет важную роль для построения оснований статистических выводов. Известно стандартное нормальное распределение, а также другие нормальные законы, которые отличаются от стандартного значениями двух параметров. Тем не менее, для распределений, отличных от стандартного существует специальный прием, называемый стандартизацией, который позволяет преобразовывать эти распределения к стандартному виду.

7-1 Стандартное нормальное распределение

Кроме рассмотренного нами в предыдущей главе равномерного распределения нам предстоит изучить еще некоторое число популярных непрерывных распределений, важнейшим из которых является нормальное распределение.

Непрерывная случайная величина имеет **нормальное распределение,** если ее плотность задается уравнением $f(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$.

График плотности нормального распределения имеет форму колокола и симметричен относительно вертикальной прямой, проходящей через μ. Этот график называют нормальной кривой или кривой Гаусса.

В формуле плотности присутствуют два параметра: μ и σ. Формула для плотности нормального закона имеет весьма сложный и специальный вид. Это оправдано тем, что при вычислении математического ожидания мы сможем получить μ, а стандартное отклонение после вычислений окажется равно σ. Тем самым, имеется много случайных величин, имеющих нормальный закон распределения. Они отличаются друг от друга двумя параметрами: μ и σ, которые представляют собой математическое ожидание и стандартное отклонение случайной величины.

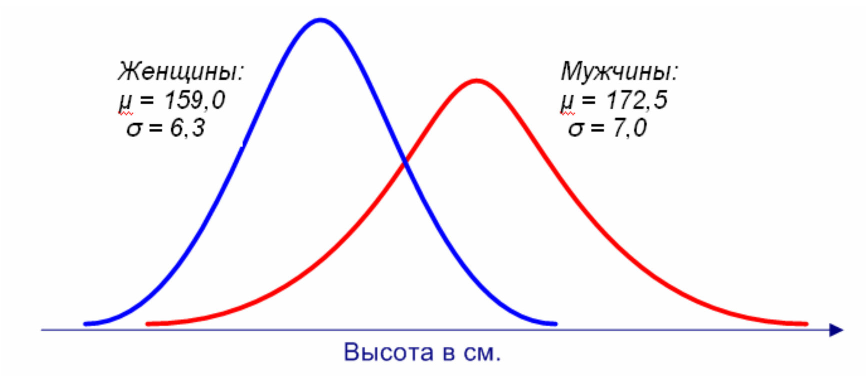


Рисунок 7-1 График плотности имеет форму колокола

В качестве примера рассмотрим случайную величину, равную росту мужчин, и другую случайную величину, равную росту женщин. Если речь идет о взрослом населении, то мы имеем дело с двумя распределениями, которые можно считать нормальными. Тем не менее, эти распределения отличаются средним значением и дисперсией. Как видно по рисунку 7-1, средний рост мужчин составляет 172,5 см, в то время как женщин 159 см. Стандартное отклонение для роста женщин немного меньше и составляет 6,3 см. Заметим, что площадь под обоими графиками равна единице, что обязательно для графика плотности распределения. В общем виде, от параметра μ зависит смещение графика вдоль горизонтальной оси, а от значения параметра σ зависит форма колокола. Чем больше стандартное отклонение, тем шире расходятся стороны графика.

Среди всех случайных величин, имеющих нормальное распределение, принято выделять одну, имеющую стандартное нормальное распределение.

Случайная величина имеет **стандартное нормальное распределение**, если ее среднее значение $\mu = 0$, а стандартное отклонение $\sigma = 1$.

Плотность стандартного нормального закона распределения определяется формулой:

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

Эта формула может быть получена после подстановки значений двух параметров в основную формулу плотности. Общепринято в качестве переменной для стандартного закона использовать z вместо x. Соответствующая кривая называется стандартной или нормированной.

Кроме функции плотности при использовании нормального закона часто бывает необходима функция распределения, обозначаемая Ф(x).

Формула для функции распределения получается при помощи формул, обсуждавшихся в предыдущей главе.

$$\Phi(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{(t-\mu)^2}{2\sigma^2}} dt$$

Для стандартного нормального закона функция распределения примет вид:

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-\frac{t^2}{2}} dt$$

Нам потребуется воспользоваться некоторыми очевидными свойствами плотности и функции распределения нормального закона, о которых мы сейчас кратко расскажем.

Свойством плотности является то, что эта функция четная. Это означает, что ее график симметричен относительно вертикальной оси и ее значения для отрицательных значений аргумента совпадают со значениями от соответствующих положительных значений:

$$f(-x) = f(x)$$

Далее, эта функция стремится к нулю, если переменная х стремится к плюс или минус бесконечности.

Свойством функции распределения для нахождения ее значений при отрицательных аргументах является следующее равенство:

$$\Phi(-x) = 1 - \Phi(x)$$

Это свойство окажется важным, поскольку нам предстоит использовать таблицы, которые составлены, как правило, только для положительных значений. Свойство позволяет узнавать значения этой функции и для отрицательных значений, если менять их знак

Если значение аргумента стремится к плюс бесконечности, функция распределения стремится к единице, монотонно возрастает на всей числовой прямой. При стремлении аргумента к минус бесконечности убывает и стремится к нулю. Эти несложные, но полезные свойства могут быть получены, если провести исследование плотности и функции распределения методами исследования функций, изучавшимися студентами в курсе математического анализа.

Таблица А-2. Нормальное распределение

z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753
0,2	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517
0,4	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879
0,5	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224
0,6	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549
0,7	0,7580	0,7611	0,7642	0,7673	0,7704	0,7734	0,7764	0,7794	0,7823	0,7852
0,8	0,7881	0,7910	0,7939	0,7967	0,7995	0,8023	0,8051	0,8078	0,8106	0,8133
0,9	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,8340	0,8365	0,8389
1,0	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621
1,1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	0,8790	0,8810	0,8830
1,2	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,8980	0,8997	0,9015
1,3	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147	0,9162	0,9177
1,4	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9279	0,9292	0,9306	0,9319
1,5	0,9332	0,9345	0,9357	0,9370	0,9382	0,9394	0,9406	0,9418	0,9429	0,9441
1,6	0,9452	0,9463	0,9474	0,9484	0,9495	0,9505	0,9515	0,9525	0,9535	0,9545
1,7	0,9554	0,9564	0,9573	0,9582	0,9591	0,9599	0,9608	0,9616	0,9625	0,9633
1,8	0,9641	0,9649	0,9656	0,9664	0,9671	0,9678	0,9686	0,9693	0,9699	0,9706

Рисунок 7-2 Пример применения таблицы

Нахождение вероятностей по таблице

Нам необходимо научиться вычислять значения вероятностей для различных значений x, применяя таблицы для нормального закона. Вообще имеется два вида таблиц – для плотности и для функции распределения. Обычно любая таблица сопровождается формулой или рисунком. Рисунок над таблицей А-2, приведенной в приложении, означает, что таблица содержит значения функции распределения, поскольку они соответствуют площади, заштрихованной под графиком нормальной кривой. Это площадь, накопленная от минус бесконечности до значения z.

Все таблицы нормального закона составлены для стандартного распределения, то есть имеющего параметры $\mu = 0$ и $\sigma = 1$. Тем не менее, позже мы узнаем, как применять эту таблицу для других нормальных распределений.

Таблица позволяет делать следующее.

(1) По заданному z-значению находить площадь под графиком, то есть соответствующее ему значение функции распределения.

(2) По известной площади под графиком, то есть для известного значения функции распределения находить соответствующее ему z-значение.

Обе возможности будут нами использованы в соответствующих разделах изучаемого курса статистики. Курс предполагает хорошее владение таблицами любым студентом. Это не означает, правда, что таблицы – единственный путь к нахождению необходимых нам значений. Некоторые значения мы будем использовать так часто, что будем знать их наизусть. Некоторые впоследствии поможет отыскать компьютер.

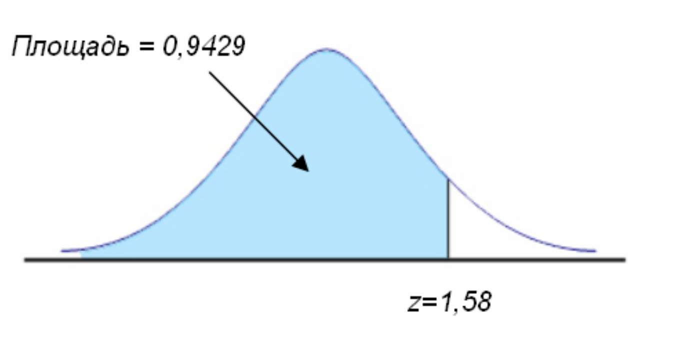


Рисунок 7-3 Площадь для z=1,58

Для примера, найдем площадь для значения *z = 1,58*. Как показано на рисунке 7-2, находим 1,5 в левой колонке и 0,08 в шапке таблицы. На пересечении получаем значение площади 0,9429. Если, наоборот, нам известна площадь 0,9429, то мы находим это значение внутри таблицы, а на полях получаем соответствующее им *z*-значение, равное 1,58.

Решим теперь следующую задачу. Компания, производящая термометры для научных исследований, проводит проверку качества своей продукции и получает, что некоторые термометры показывают точку замерзания воды ниже 0ºC, а другие – выше. Полагая, что среднее значение показаний есть 0ºC, а стандартное отклонение 1ºC, найти вероятность того, что показание случайно выбранного термометра: (1) Окажется ниже 1,58ºC. (2) Окажется выше -1,23ºC. (3) Будет находиться в интервале от -2ºC до 1,5ºC.

(1) Мы хотим найти вероятность того, что показание окажется ниже 1,58ºC, то есть значение функции распределения от 1,58. Мы уже нашли в таблице это значение, вероятность равна 0,9429. Ее можно проинтерпретировать следующим образом: 94,29% термометров будут иметь показание точки замерзания воды ниже 1,58ºC.

(2) Площадь для *z>-1,23*, т.е. для отрицательных значений будем находить, пользуясь свойством симметрии графика. Симметрия означает, что интересующая нас площадь равна площади для *z<1,23*. Из таблицы А-2 находим, что площадь, а значит, и соответствующая вероятность равна 0,8907.

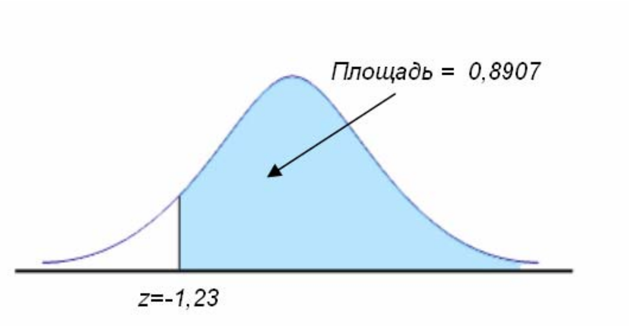


Рисунок 7-4 Нахождение площади для x>-1,23

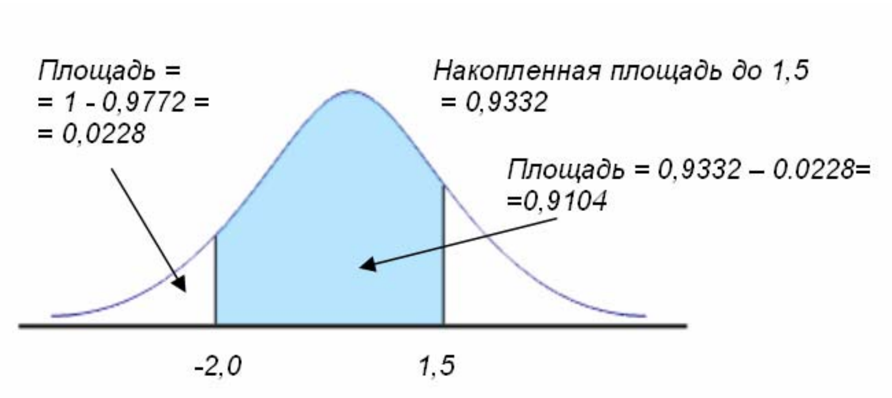


Рисунок 7-5 Нахождение площади для -2ºC<z<1,5ºC

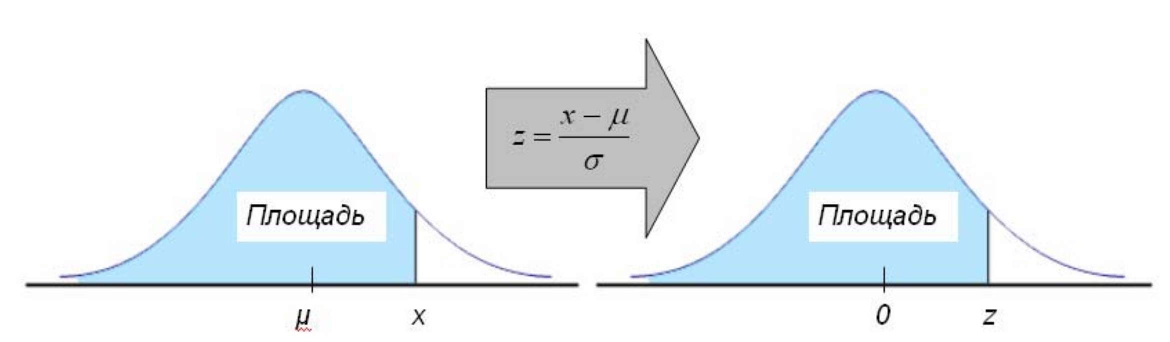
(3) Площадь для -2ºC<*z*<1,5ºC находится путем вычитания из накопленной площади слева до 1,5ºC площади, накопленной до -2ºC. Площадь до -2ºC составила 0,0228, до 1,5ºC составила 0,9332. Вычитаем из одной площади другую и получим, что вероятность, что термометр окажется в заданном интервале составляет 0,9104. Можно проинтерпретировать, что в этом интервале окажется 91% всех термометров.

Нахождение вероятностей при помощи таблиц требует некоторых усилий. С таблицами надо «повозиться». Следует решить несколько задач, чтобы почувствовать уверенность. Не следует думать, что компьютер легко заменит таблицы, и навыки их использования впоследствии не понадобятся. Даже весьма опытные пользователи статистических пакетов весьма часто предпочитают использовать таблицы, поскольку это быстрее и надежнее. Не случайно таблицы продолжают публиковать не только в учебных книгах, а в любой литературе по статистике.

Резюме

Стандартное нормальное распределение определяется формулой плотности и имеет параметры 0, 1, в отличие от других нормальных распределений, которые мы рассмотрим в следующем параграфе. Для стандартного нормального закона составлены таблицы значений, которые отличаются по форме, но позволяют решать задачи двух типов: по заданному z-значению находить вероятность, и, наоборот, по заданной вероятности находить соответствующее z-значение.

Рисунок 7-6 Операция стандартизации



7-2 Другие нормальные распределения

Для изучения произвольных нормальных распределений, то есть тех, у которых среднее значение и стандартное отклонение не равны 0 и 1 соответственно, применяется операция стандартизации.

Операция стандартизации

Операция стандартизации означает преобразование произвольного нормального распределения с параметрами μ, σ, в стандартное с параметрами 0 и 1.

Формула для операции стандартизации:

$$z = \frac{x-\mu}{\sigma}$$

Стандартизация означает, что мы смещаем распределение и изменяем его форму так, чтобы оно стало стандартным. При этом площадь под кривой до значения x равна площади под стандартной кривой до значения z.

Применяя формулу, следует округлять получаемое значение z до двух знаков после запятой. Операция стандартизации позволяет находить значения площади, применяя таблицу для стандартного нормального распределения.

В качестве примера разберем ситуацию с расчетом характеристик салона нового автомобиля. При разработке новой модели автомобиля инженеры-конструкторы учитывают, что рост человека в сидячем состоянии в среднем составляет 90 см со стандартным отклонением 3,5 см. Исходя из этого, они предполагают, что рост человека в сидячем состоянии не превысит 97 см. Проверим, насколько разумны эти предположения. Какое количество потенциальных покупателей останутся «за бортом»?

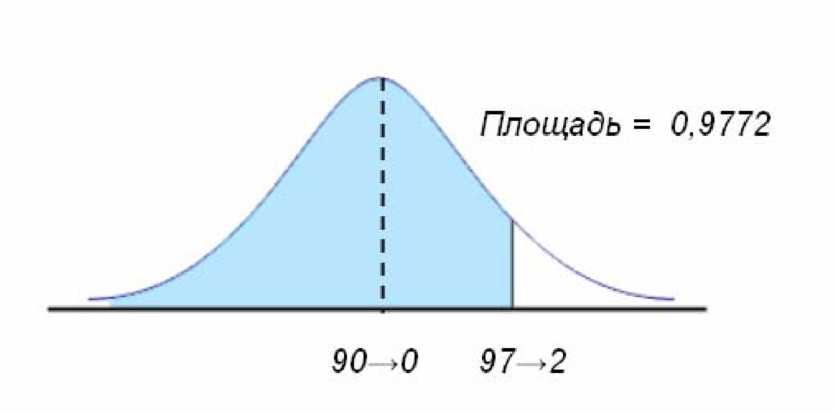


Рисунок 7-7 Операция стандартизации

Если максимально возможный рост составляет 97 см, найдем в условиях задачи вероятность того, что произвольный человек окажется ниже этого роста. На рисунке 7-7 заштрихована площадь, которая нас интересует. Если мы найдем ее, задача будет решена.

Чтобы воспользоваться таблицей, проведем преобразование и перейдем к стандартному распределению.

$$z = \frac{x - \mu}{\sigma} = \frac{97 - 90}{3,5} = 2,00$$

Этот расчет означает, что 97 отклоняется от среднего значения 90 на 2,0 стандартных отклонения. По таблице А-2 находим, что при z=2 площадь под графиком равна 0,9772. Это есть вероятность того, что случайно выбранный человек будет иметь рост в сидячем положении не более 97 см:

$$P(x < 97см) = P(z < 2,00) = 0,9772$$

Инженерам-конструкторам следует решить, считают ли они возможным разрабатывать автомобиль, в котором из соображений роста поместятся 97,72% мужчин и 2,28% окажутся «за бортом».

Может случиться, что инженеры, после советов с продавцами, выберут границу роста в 5%. Тогда требуется решить обратную задачу: для какого роста 95% мужчин окажутся в его пределах? Теперь необходимо отыскать в основном поле таблицы число 0,9500. Такого значения, к сожалению, нет. Зато есть 0,9495 и 0,9505. Соответствующие им z-значения равны 1,64 и 1,65 соответственно. Берем среднее и получим, что z-значение, соответствующее 95% вероятности есть 1,645.

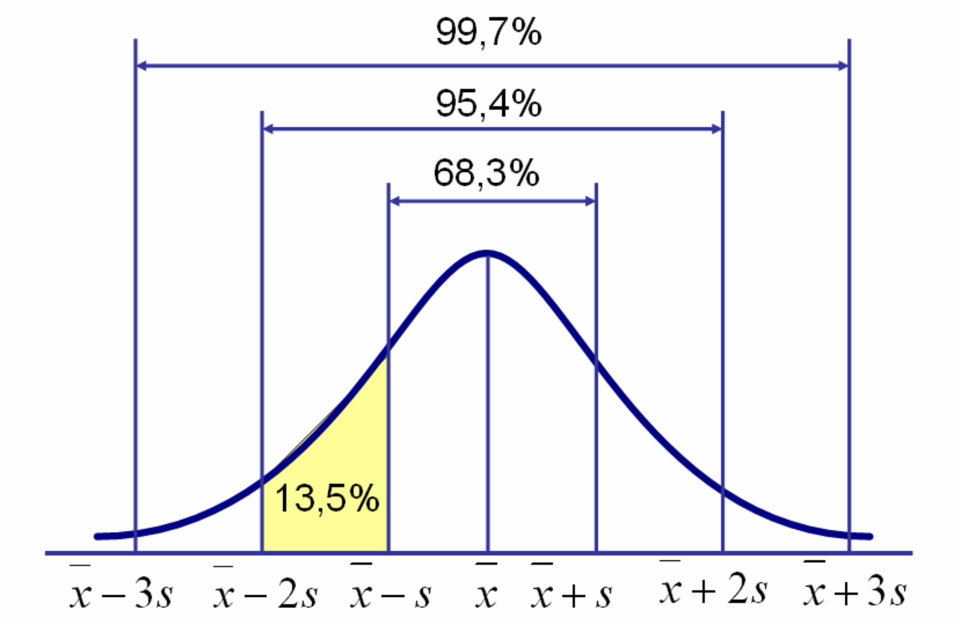


Рисунок 7-8 Правило «трех сигм» для нормального закона

Проводим обратное стандартизации преобразование. Для этого решим уравнение:

$$1{,}645 = \frac{x-90}{3{,}5}$$

Находим x = 95,7575 ≈ 95,8. Это означает, что 5% мужчин имеют в сидячем состоянии рост выше 95,8. Эту характеристику можно учитывать при проектировке салона нового автомобиля.

Правило «трех сигм»

Для нормального закона выполняется так называемое правило «трех сигм». Несложные преобразования и вычисления для нормального закона позволяют получить, что на расстоянии одного σ - стандартного отклонения от среднего значения находится 68,3% всех наблюдений, на расстоянии двух стандартных отклонений от среднего значения находится 95,4% всех наблюдений, и, наконец, на расстоянии трех сигм от среднего значения находится 99,7% всех наблюдений. Это означает, что на расстоянии трех сигм находятся практически все возможные значения случайной величины и лишь в трех случаях из тысячи это будет не выполнено.

Правило трех сигм позволяет с высокой точностью оценивать диапазон возможного изменения значений переменной относительно среднего, если она распределена нормально.

7-3 Нормальное приближение биномиального распределения

Изучив некоторые важные свойства нормального распределения и использование таблиц, мы можем вернуться к биномиальному распределению и сделать важные дополнения. Эти дополнения основаны на том, что биномиальное распределение хорошо описывается нормальным законом. Если бы мы могли визуально сравнить два распределения – биномиальное и нормальное, то существенным отличием является исключительно то, что одно является дискретным, а другое непрерывным. Во всем остальном они практически одинаковы. Если считать, что можно было бы «сгладить» дискретный характер биномиального распределения, то мы получили бы нормальный закон.

На этом основано следующее утверждение, называемое локальной теоремой Муавра-Лапласа. В схеме Бернулли, когда вероятность наступления события А постоянна и находится между 0 и 1, то вероятность того, что в n независимых испытаниях событие А произойдет ровно k раз, приближенно равна:

$$P_n(k) \approx \frac{f(z)}{\sqrt{npq}}$$

где

$$z = \frac{k - np}{\sqrt{npq}}$$

Локальная теорема Муавра-Лапласа при помощи нормального закона позволяет определить вероятность получить ровно k успехов, если проводится n испытаний по схеме Бернулли. Формула вероятности приведена выше.

Чем больше n, тем точнее приближение, которое дает формула. Считается, что формулу можно использовать, если $np \geq 5$ и $nq \geq 5$. В противном случае она дает большие погрешности.

Решим для примера следующую задачу. В отдельном регионе на 100 семей приходится 24 многодетных. Найдем вероятность того, что из 400 случайно выбранных семей многодетными окажутся ровно 90. Для этого вычислим для начала необходимые величины:

$$k = 90$$

$$n = 400, p = 24/100 = 0,24, q = 0,86$$

$$npq = 400 \cdot 0,24 \cdot 0,86 = 82,56 > 20$$

$$z = \frac{k - np}{\sqrt{npq}} = \frac{90 - 400 \cdot 0,24}{\sqrt{82,56}} = \frac{-6}{9,086} = -0,66$$

Теперь при помощи таблиц находим значение функции плотности:

$$f(z) = f(-0,66) = f(0,66) = 0,3209$$

Мы воспользовались свойством симметрии для плотности стандартного нормального распределения. При этом значения такой функции от отрицательных значений аргумента равны значениям от соответствующих положительных значений. Подставляем полученное в формулу:

$$P_n(k) \approx \frac{f(z)}{\sqrt{npq}} = \frac{0,3209}{82,56} = 0,0039$$

Это означает, что в 3,9% случаев среди 400 семей ровно 90 окажутся многодетными. Очевидно, нас могла бы заинтересовать вероятность того, что из 400 случайно выбранных семей многодетными окажутся от 80 до 100 семей. Эта вероятность, очевидно, существенно больше полученной нами, поскольку включает вероятности того, что многодетных семей будет ровно 80, ровно 81, …, ровно 100. Мы могли бы применить локальную теорему двадцать раз и получили бы искомую вероятность. Этого делать нам не придется, поскольку кроме локальной теоремы мы рассмотрим также интегральную теорему, которая позволяет вычислять вероятность того, что в схеме Бернулли из n испытаний количество наступлений интересующего нас события будет находиться в интервале от a до b.

Интегральная теорема Муавра-Лапласа при помощи нормального закона позволяет определить вероятность получить от a до b успехов, если проводится n испытаний по схеме Бернулли. Формула вероятности приведена ниже.

В схеме Бернулли, когда вероятность наступления события A постоянна и находится между 0 и 1, то вероятность того, что в n независимых испытаниях число k наступления события A будет находиться в пределах от a до b, приближенно равна:

$$P_n(a \le k \le b) \approx \Phi(z_2) - \Phi(z_1)$$

где

$$z_1 = \frac{a - np}{\sqrt{npq}}, \ z_2 = \frac{b - np}{\sqrt{npq}}$$

Вернемся к решению. Нас интересует, чтобы количество многодетных семей находилось в интервале от 80 до 100.

$$npq = 82{,}56$$

$$a = 80, b = 90$$

$$z_1 = \frac{a - np}{\sqrt{npq}} = \frac{80 - 400 \cdot 0{,}24}{9{,}086} = -1{,}76$$

$$z_2 = \frac{b - np}{\sqrt{npq}} = \frac{100 - 96}{9{,}086} = 0{,}44$$

По таблице находим значения функции распределения, а затем вероятность:

$$\Phi(z_1) = \Phi(-1{,}76) = 1 - \Phi(1{,}76) = 1 - 0{,}9608 = 0{,}0392$$

$$\Phi(z_2) = \Phi(0{,}44) = 0{,}6700$$

$$P_n(a \le k \le b) \approx \Phi(z_2) - \Phi(z_1) = 0{,}6700 - 0{,}0392 = 0{,}6308$$

Получили, что в 63% случаев количество многодетных семей будет находиться в интервале от 80 до 100 включительно, если мы случайно выберем всего 400 семей.

Резюме

Мы вновь рассмотрели схему Бернулли и теперь смогли получить для нее новые формулы для нахождения вероятностей. Эти формулы основаны на нормальном приближении для биномиального закона. Имеется два вида формул. Локальная формула позволяет вычислять вероятность того, что количество успехов примет из общего числа испытаний определенное точечное значение. Интегральная формула позволяет вычислить вероятность того, что число успехов будет находиться в заданном интервале. Формулы дают хорошее приближение, если количество испытаний достаточно велико.

7-4 Распределения, связанные с нормальным

При изучении статистических методов мы встретим несколько распределений, связанных с нормальным. Кратко дадим описание каждого из них.

Распределение χ^2

Пусть имеется n независимых случайных величин, каждая из которых имеет стандартное нормальное распределение. Тогда сумма квадратов этих случайных величин распределена по закону χ^2 («хи квадрат») с количеством степеней свободы, равным количеству этих случайных величин. Количество степеней свободы есть параметр распределения.

Распределением χ^2 с n степенями свободы называется распределение суммы квадратов n независимых случайных величин, распределенных по стандартному нормальному закону.

Символически случайную величину, имеющую χ^2-распределение, можно записать в виде суммы:

$$\chi^2 = z_1^2 + z_2^2 + \ldots + z_n^2$$

Плотность χ^2-распределения:

$$f(x) = \begin{cases} 0 \quad \textit{при} \quad x \le 0 \\ \dfrac{1}{2^{k/2}\Gamma(k/2)} e^{-x/2} x^{(k/2)-1} \quad \textit{при} \quad x > 0 \end{cases}$$

где $\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} dt$ - гамма-функция Эйлера.

Для χ^2-распределения составлены таблицы. Мы будем использовать их при изучении аналитической статистики.

Назовем некоторые полезные свойства χ^2-распределения. Оно всегда неотрицательно. Зависит от n – числа степеней свободы. Математическое ожидание равно n - количеству переменных, а стандартное отклонение равно 2n. С увеличением числа степеней свободы χ^2-распределение стремится к нормальному.

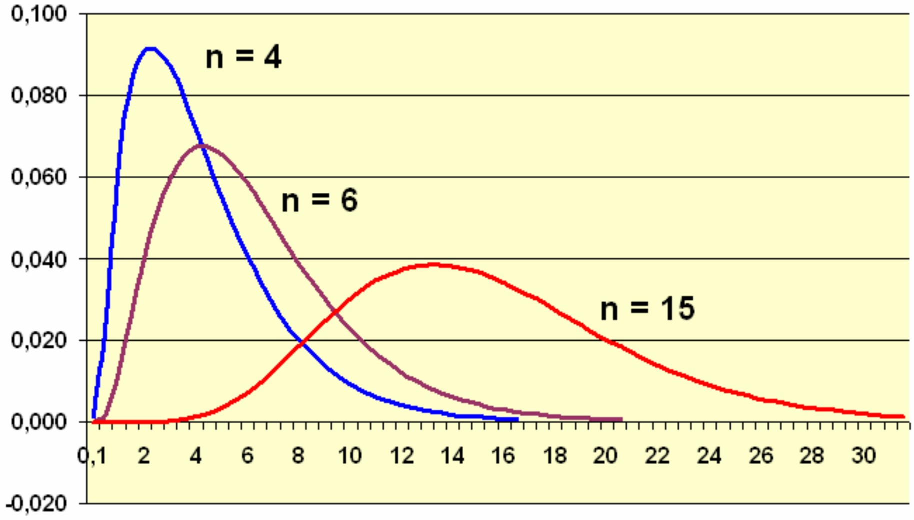


Рисунок 7-9 Вид χ^2 распределения для n=4, 6 и 15

На рисунке 7-9 показаны кривые распределения. Все кривые обладают правосторонней (положительной) асимметрией. При этом с увеличением количества степеней свободы распределение постепенно приближается к нормальному.

Распределение Стьюдента

Распределением Стьюдента (или t-распределением) называется распределение случайной величины: $t = \dfrac{z}{\sqrt{\chi^2 / k}}$, где z имеет стандартное нормальное распределение, а χ^2 – независимая от z случайная величина, имеющая χ^2 распределение с n степенями свободы.

Мы не приводим выражение для плотности распределения Стьюдента. Назовем основные свойства этого распределения. Распределение симметрично относительно вертикальной оси, кривая более пологая по сравнению с нормальной. При увеличении числа степеней свободы распределение приближается к нормальному. Можно считать это практически выполненным для n>30. Математическое ожидание в силу симметрии равно нулю, а дисперсия равна n/(n-2).

Распределение Фишера

Распределением Фишера (или F-распределением) называется распределение случайной величины: $F = \dfrac{\chi^2(k_1)/k_1}{\chi^2(k_2)/k_2}$, где $\chi^2(k_1)$ и $\chi^2(k_2)$ – случайные величины, имеющие χ^2 распределение с k_1 и k_2 степенями свободы соответственно.

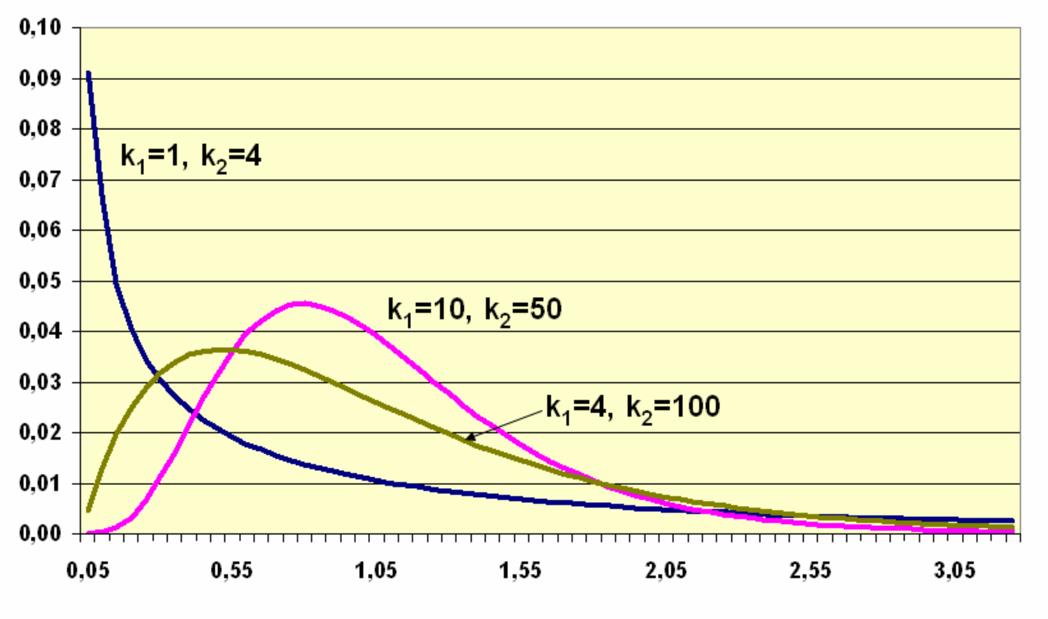


Рисунок 7-10 F-распределение для трех наборов степеней свободы

Мы не приводим громоздкого выражения для плотности распределения Фишера. Распределение характеризуется двумя параметрами степеней свободы. На рисунке показаны несколько кривых для разных наборов степеней свободы.

Резюме

Кратко обсуждены три распределения, связанных с нормальным законом – χ^2-распределение (хи-квадрат), Стьюдента и Фишера. Кроме формального определения названы некоторые полезные свойства и для двух представлены графики. Распределения характеризуются числом степеней свободы – новым понятием, которое обозначает параметр распределения, зависящий от количества случайных величин, образующих рассматриваемую случайную величину. Например, для χ^2-распределения если случайную величину образуют n независимых случайных величин, то в общем случае сумма квадратов имеет n степеней свободы.

Используем компьютер

Для материала настоящей главы полезно научиться использовать компьютер для нахождения табличных значений рассмотренных распределений. Довольно непростой задачей, которую нужно решить, является построение графиков плотности распределений в электронных таблицах. Требуется для начала подобрать соответствующие функции, рассчитать плотность для различных значений аргумента, а затем построить графики распределения подобные тем, которые имеются в книге.

Что означают термины

Нормальное распределение	**Операция стандартизации**	**Распределение χ^2**
Стандартное нормальное распределение	**Локальная теорема Муавра-Лапласа**	**Распределение Стьюдента**
Плотность и функция нормального распределения	**Интегральная теорема Муавра-Лапласа**	**Распределение Фишера**

Символы и формулы

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

Плотность нормального распределения

$$\Phi(x)=\frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{x}e^{-\frac{(t-\mu)^2}{2\sigma^2}}dt$$

Функция распределения нормального закона

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

Плотность стандартного нормального распределения

$$\Phi(z)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{z}e^{-\frac{t^2}{2}}dt$$

Функция распределения стандартного нормального закона

$$\Phi(-x)=1-\Phi(x)$$

Свойство функции распределения для отрицательных значений

$$z=\frac{x-\mu}{\sigma}$$

Операция стандартизации

$$P_n(k)\approx\frac{f(z)}{\sqrt{npq}}$$

$$z=\frac{k-np}{\sqrt{npq}}$$

Вероятность в схеме Бернулли (локальная теорема Муавра-Лапласа)

$$P_n(a\le k\le b)\approx\Phi(z_2)-\Phi(z_1)$$

$$z_1=\frac{a-np}{\sqrt{npq}}$$

$$z_2=\frac{b-np}{\sqrt{npq}}$$

Вероятность в схеме Бернулли (интегральная теорема Муавра-Лапласа)

$$\chi^2=z_1^2+z_2^2+\ldots+z_n^2$$

Случайная величина, имеющая распределение χ^2

$$t=\frac{z}{\sqrt{\chi^2/k}}$$

Случайная величина, имеющая распределение Стьюдента

$$F=\frac{\chi^2(k_1)/k_1}{\chi^2(k_2)/k_2}$$

Случайная величина, имеющая распределение Фишера

Задачи и упражнения

7-1. Ответьте на вопросы.

а) Какая процентная доля площади находится до среднего значения?
б) После него?
в) Какая процентная доля площади под кривой нормального распределения находится в пределах одного стандартного отклонения до и после среднего значения?
г) В пределах двух стандартных отклонений?
д) В пределах трех стандартных отклонений?

7-2. Найдите площадь. Найдите площадь под графиком плотности нормального распределения:

а) Между $z = 0$ и $z = 1{,}97$
б) Между $z = 0$ и $z = 0{,}56$
в) Между $z = 0$ и $z = -0{,}48$
г) Между $z = 0$ и $z = -2{,}07$
д) Справа от $z = 1{,}02$
е) Справа от $z = 0{,}23$
ж) Слева от $z = -0{,}42$
з) Слева от $z = -1{,}43$
и) Между $z = 1{,}23$ и $z = 1{,}90$
к) Между $z = 0{,}79$ и $z = 1{,}28$
л) Между $z = -0{,}87$ и $z = -0{,}21$
м) Между $z = -1{,}56$ и $z = -1{,}83$
н) Между $z = 0{,}24$ и $z = -1{,}12$
о) Между $z = 2{,}47$ и $z = -1{,}03$
п) Слева от $z = +1{,}22$
р) Слева от $z = +2{,}16$
с) Справа от $z = -1{,}92$
т) Справа от $z = -0{,}18$
у) Слева от $z = -2{,}15$ и справа от $z = 1{,}62$
ф) Справа от $z = 1{,}98$ и слева от $z = -0{,}59$

7-3. Найдите вероятности. Найдите вероятности, используя стандартное нормальное распределение:

а) $P(0 < z < 1{,}69)$
б) $P(0 < z < 0{,}67)$
в) $P(-1{,}23 < z < 0)$
г) $P(-1{,}57 < z < 0)$
д) $P(z > 2{,}59)$
е) $P(z > 2{,}83)$

ж) $P(z < -1{,}77)$
з) $P(z < -1{,}51)$
и) $P(-0{,}05 < z < 1{,}10)$
к) $P(-2{,}46 < z < 1{,}74)$
л) $P(1{,}32 < z < 1{,}51)$
м) $P(1{,}46 < z < 2{,}97)$
н) $P(z > -1{,}39)$
о) $P(z < -1{,}42)$

7-4. Найдите z-значение.
а) Найдите значение z справа от среднего, такое, что 53,98% распределения лежит слева от него.
б) Найдите значение z слева от среднего, такое, что 96,86% распределения лежит справа от него.
в) Найдите два значения z, таких, что 40% площади посередине заключено между ними.

7-5. Вес шоколадной конфеты. Средний вес шоколадной конфеты составляет 16 граммов со стандартным отклонением 0,3 грамма. Определите вероятность того, что конфета весит от 16 до 16,2 граммов.

7-6. Более 26 миль. Автомобиль некоторой марки при расходовании одного галлона горючего в состоянии преодолеть расстояние 22 мили при городском цикле езды. Стандартное отклонение составляет 3 мили на один галлон. Определите вероятность того, что в произвольно выбранный день, автомобиль преодолеет расстояние более 26 миль, израсходовав один галлон топлива при городском цикле езды.

7-7. Сколько горожане смотрят телевизор. Исследование телевизионного журнала показало, что горожане смотрят телевизионные передачи в среднем 18,6 часов в неделю со стандартным отклонением 2,3 часа. Определите следующие вероятности:
а) Горожанин смотрит телевизионные передачи менее 15 часов в неделю.
б) Горожанин смотрит телевизионные передачи более 22 часов в неделю.

7-8. Возраст директоров. Во время проведенного исследования выяснилось, что возраст директоров заводов в городе N в среднем составляет 56 лет. Стандартное отклонение равно 4 годам. При произвольном выборе завода определите вероятность того, что возраст ее директора находится:

а) Между 53 и 59 годами
б) Между 58 и 63 годами
в) Между 50 и 55 годами

7-9. Время в зоопарке. В среднем каждый человек находится в зоопарке города 62 минуты со стандартным отклонением 12 минут. При произвольном выборе посетителя определите вероятность того, что он или она проведут в зоопарке:

а) Как минимум 82 минуты
б) Максимум 50 минут

7-10. Задача по статистике. Среднее время для решения определенной задачи по статистике студентами первого курса составляет 24,6 минуты со стандартным отклонением 5,8 минут. Определите следующие вероятности:

а) Студенту потребуется от 15 до 30 минут для решения задачи
б) Студенту потребуется менее 18 минут или более 28 минут

7-11. Претенденты на работу. Показатели теста на уровень умственного развития IQ находятся в средних пределах 100, а стандартное отклонение от нормы составляет 15. Если руководитель кадровой службы должен отобрать лучших 75% претендентов, которые проходят через тестирование, то определите уровень возможного отсева.

7-12. Велосипедные гонки. Ассоциация спортсменов хочет осуществить руководство проведением велосипедных гонок. Среднее время проезда по всей дистанции составляет 62,5 минуты при стандартном отклонении, равном 5,8 минуты. Если ассоциация принимает решение использовать только 25% лучших гонщиков, то какой отсев произойдет во время квалификационного заезда?

7-13. Найти стандартное отклонение. Дано нормальное распределение, у которого $\mu = 100$. Определите значение σ, если 2,68% площади лежит справа от 105.

7-14. Найти математическое ожидание. Для нормального распределения определите значение μ при $\sigma = 6$, если 3,75% площади лежит слева от 85.

7-15. Приближение биномиального закона. Используя приближение биномиального распределения нормальным, найдите вероятности для следующих условий:

а) $n = 30, p = 0{,}5, k = 18$
б) $n = 50, p = 0{,}8, k = 44$
в) $n = 100, p = 0{,}1, k = 12$
г) $n = 10, p = 0{,}5, k \geq 7$
д) $n = 20, p = 0{,}7, k \leq 12$
е) $n = 50, p = 0{,}6, k \leq 40$

7-16. Покупки в Интернет. 12% посетителей интернет-магазина делают покупки. Если в течение дня на сайт заходили 500 посетителей, определите вероятность того, что как минимум 55 из них сделали покупки.

7-17. Номера в гостинице. В гостинице имеется 80 номеров, а заполнение гостиницы в марте составляет 60%. Определите следующие вероятности:

а) Как минимум 60 номеров будут заняты.
б) Будет занято меньше 50 номеров.

7-18. Сколько девочек. Если вероятность того, что родившийся ребенок будут девочкой, составляет 50%, то определите вероятность того, что на 100 родов 55 или более будут девочками.

7-19. Сколько автомобилей в семье. Если 20% семей города N имеют как минимум 2 машины, то определите вероятность того, что при произвольном выборе 100 семей ровно в 20 семьях будет две или более машин.

7-20. **Двойки на экзаменах**. Если количество двоек по курсу математики обычно составляет 20%, то определите вероятность того, что в студенческой группе из 25 человек ровно 5 студентов провалятся на экзаменах.

Глава 8. Основания для статистических выводов

В этой главе мы рассмотрим логику статистических выводов и основания, которые позволяют на основе изучения выборки делать выводы о параметрах генеральной совокупности. Для начала будут рассмотрены способы получения случайной выборки, поскольку нерепрезентативные и другие виды детерминированных выборок не позволяют делать статистических заключений. Затем мы установим связь между тремя распределениями – распределением генеральной совокупности, распределением выборки и, наконец, распределением выборочных средних. Завершит изучение центральная предельная теорема и приложения, которые имеют существенное значение для статистики.

8-1 Выборочное наблюдение

В самом начале мы уже рассматривали выборку из генеральной совокупности и обсуждали цели, которые заставляют нас получать ее и анализировать.

Теперь, после изучения теоретико-вероятностных основ, мы возвратимся к началу и вновь обсудим некоторые понятия. Генеральной совокупностью мы назвали всю интересующую исследователя совокупность изучаемых объектов. Выборкой назвали некоторую, обычно небольшую, часть генеральной совокупности, отбираемую специальным образом и исследуемую с целью получения выводов о свойствах генеральной совокупности. Мы отметили, что репрезентативной называется такая выборка, которая *хорошо* представляет генеральную совокупность. Это означает, что каждое свойство (или комбинация свойств) наблюдается в выборке с той же частотой, что и в генеральной совокупности. Выборка, которая не является репрезентативной, говорят, что имеет смещение. Как должна быть получена выборка, чтобы быть репрезентативной? В каком случае мы на основе изучения выборки сможем получить корректные выводы, относящиеся ко всей генеральной совокупности?

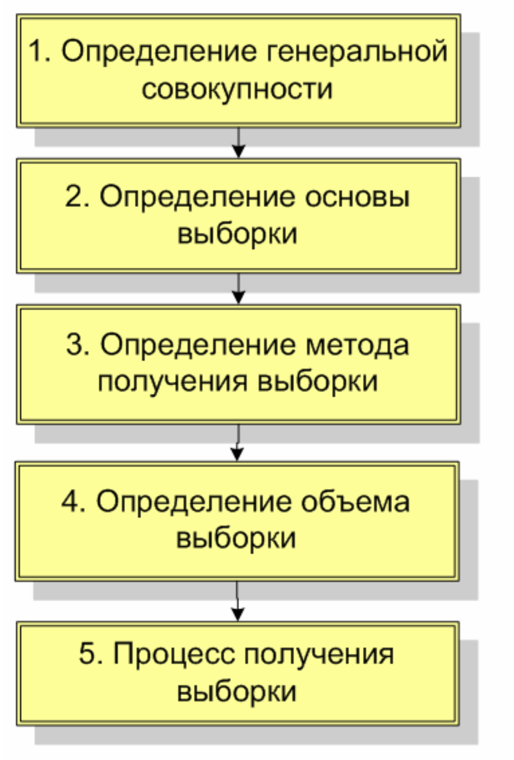


Рисунок 8-1 Выборочное наблюдение

Обсудим действия исследователя при проведении выборочного наблюдения.

1. Определение генеральной совокупности. На первом этапе определяют целевую генеральную совокупность. Важно понять, какие объекты попадут в эту совокупность, где ее границы.

Целевая генеральная совокупность (target population) – совокупность объектов, обладающих информацией, которую желает получить исследователь и о которой требуется сделать заключение.

Как соотносятся между собой понятия генеральной совокупности и целевой генеральной совокупности? Когда мы говорим о генеральной совокупности, мы подразумеваем готовую для исследования, возможно, сформированную без нашего участия, совокупность. Это просто теоретическая модель. Когда мы говорим о целевой генеральной совокупности, мы подразумеваем совокупность, которую нам требуется выделить и представить для последующей работы. Указание на целевой характер совокупности означает, что эта совокупность формируется самим исследователем исходя из стоящих перед ним целей.

Если требуется, например, провести опрос жителей и выяснить их отношение к транспортному обслуживанию, на первой стадии исследования важно понять, кого мы включим в целевую генеральную совокупность. Возможно, это будут граждане, проживающие в определенном районе и имеющие постоянную прописку. А если в районе есть проживающие с временной пропиской или без регистрации, то их мнение о качестве обслуживания тоже может оказаться важным для улучшения транспортного обслуживания, поскольку эти люди, возможно, очень часто пользуются транспортом и оплачивают услуги транспортных компаний. А будут ли учитываться мнения детей, скажем, школьного возраста?

Как только появляется тема исследования, определение границ целевой генеральной совокупности становится важным занятием исследователя и первой содержательной стадией его работы при проведении выборочного наблюдения. Начиная с первой стадии, все действия вплоть до отдельных деталей должны документироваться, чтобы зафиксировать логику, которой придерживался исследователь, принимая решения о границах генеральной совокупности.

2. Определение основы выборки. После выделения целевой генеральной совокупности на второй стадии определяется основа выборочного наблюдения.

Основа выборочного наблюдения (sampling frame) есть представление элементов изучаемой генеральной совокупности. Обычно это список всех объектов или перечень инструкций для определения границ и объектов изучаемой совокупности.

Основа, или представление целевой генеральной совокупности необходима для формирования выборки. Примерами основы выборочного наблюдения могут служить телефонные справочники, отраслевые справочники предприятий, список адресатов и т.п.

Если исследователь не может составить подробный перечень элементов, следует, по крайней мере, установить правила для отбора объектов изучаемой генеральной совокупности, например процедуру случайного набора номеров при проведении опроса по телефону.

Ошибки при формировании основы выборочного наблюдения состоят в пропуске некоторых объектов или включении в основу объектов, не относящихся к изучаемой генеральной совокупности. Имеются методы, позволяющие устранить такие ошибки или свести их действие к минимуму.

3. Определение метода получения выборки. Когда основа сформирована, необходимо определиться с методом получения выборки. Существуют детерминированные методы, включающие нерепрезентативную выборку, поверхностную выборку, квотную выборку и выборку по принципу «снежного кома». Каждая из них пригодна для определенных целей исследования, но обладает главным недостатком: не позволяет применить вероятностные, статистические методы для получения выводов о генеральной совокупности.

К другой группе методов получения выборки относятся вероятностные методы: простая случайная выборка, систематическая выборка, стратифицированная выборка и кластерная выборка. Мы перечислили распространенные методы получения выборок, хотя этот перечень содержит существенно большее количество методов и имеет много особенностей. В нашем курсе мы будем иметь дело лишь с простой случайной выборкой и стратифицированной выборкой.

4. Определение объема выборки. Следующим важным шагом является определение объема выборки. К этому времени известен объем генеральной совокупности. Предположим, она включает N объектов. Задача состоит в том, чтобы определить n - объем выборки. Он должен быть оптимален с точки зрения затрат на исследование и достаточен для обоснованного получения выводов об исследуемой генеральной совокупности. Если изучается вся генеральная совокупность (N = n), то выборка называется переписью.

Переписью (census) называется ситуация, при которой выборка совпадает со всей исследуемой генеральной совокупностью.

Как правило, изучаемая исследователем генеральная совокупность конечна. Если объем совокупности достаточно велик, а выборка небольшая, тогда считают, что:

$$\frac{n}{N} \approx \frac{n}{\infty}$$

В случае, если объем выборки сопоставим с объемом генеральной совокупности, исследователи, как мы увидим, делают поправку на конечность генеральной совокупности, для чего используют специальные поправочные коэффициенты.

5. Процесс получения выборки. Последняя стадия, будучи хорошо подготовленной, проводится исследователем по заранее разработанной программе и поэтому не требует дополнительных интеллектуальных усилий. Тем не менее, в зависимости от метода получения выборки и ее объема, эта стадия может оказаться продолжительной, в отличие от четырех предыдущих, которые могут, несмотря на высокую содержательность, уложиться в один день или даже несколько часов.

Вероятностные методы получения выборки

Как уже отмечалось, к вероятностным методам получения выборки относятся простая случайная выборка, систематическая выборка, стратифицированная выборка и кластерная выборка.

Простая случайная выборка формируется при помощи методов случайного отбора или случайных чисел.

Один из методов заключается в нумерации каждого объекта генеральной совокупности и случайном выборе номеров. Случайный выбор может проводиться при помощи таблицы случайных чисел, калькулятора или компьютера. Таблицу случайных чисел мы включили в приложение (таблица А-0). Она была составлена нами при помощи EXCEL.

Если, скажем, среди 68 объектов случайным образом нужно выбрать шесть, мы, начиная с любого места таблицы, последовательно определяем шесть номеров, которые определяют объекты, включаемые в выборку. Если нам встретится, например, номер 84, мы пропустим его, поскольку в генеральной совокупности отсутствует объект с таким номером.

Практика показывает, что применение случайных чисел является едва ли не самым надежным методом случайного отбора, поскольку гарантирует случайный отбор объектов. Нарушение принципа случайности приводит к невозможности делать статистически обоснованные заключения о генеральной совокупности на основе исследования и анализа выборки.

Стратифицированная выборка получается путем разбиения генеральной совокупности на группы (или страты) в зависимости от характеристик, важных для изучения.

Для примера, рассмотрим получение стратифицированной выборки из числа студентов двух факультетов: филологического и биологического. Выделены два признака: пол студента и факультет. По этим двум признакам образованы четыре страты: филологи-мужчины, филологи-женщины, биологи-мужчины, биологи-женщины. Всего генеральная совокупность включает 2000 студентов, среди которых 60% филологов и 40% биологов. Соотношение мужчин и женщин 30÷70. Тогда при формировании выборки объема 100 нам следует учесть имеющиеся пропорции и из каждой страты выбрать соответствующее количество респондентов. Для того, чтобы обеспечить случайность выборки, внутри страты должен быть обеспечен случайный отбор.

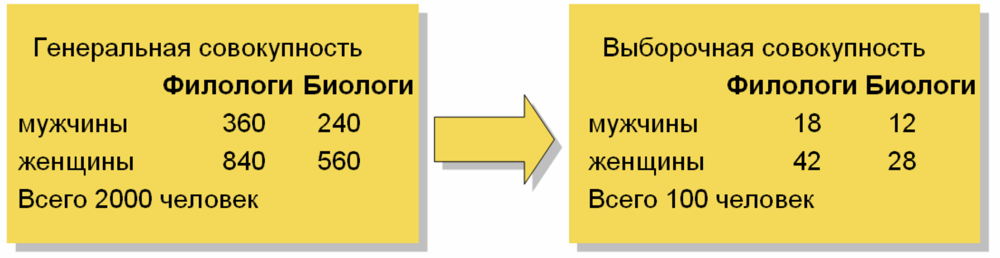


Рисунок 8-2 Получение стратифицированной выборки

Преимуществом стратифицированной выборки является наличие представителей каждой страты в выборке в соотношении, сходном с генеральной совокупностью. Недостатком является сложность организации процесса получения выборки при наличии нескольких признаков. Предположим, если в качестве изучаемых признаков выбраны возраст, доход, уровень образования и социальный статус, то такой набор приведет к выделению большого числа страт, что затруднит получение выборки.

Систематическая выборка получается путем нумерации каждого члена генеральной совокупности и выбором каждого k-ого номера, начиная с некоторого, выбранного случайно.

Пусть, например, генеральная совокупность включает 2000 единиц, требуется отобрать 50. Поскольку 2000/50=40, то будем выбирать каждый 40-й элемент. Для начала случайным образом выберем первый элемент выборки среди первых сорока элементов генеральной совокупности. Если первым оказался номер 12, тогда выборка будет включать объекты с номерами 12, 52, 92 и так далее, всего 50 объектов.

Большим недостатком систематической выборки является отсутствие случайности при выборе объектов, кроме первого. В простом случайном отборе случайным является любой следующий объект, что существенно повышает случайный характер выборки.

Кластерная выборка образуется при выделении отдельных групп, которые называются кластерами.

Предположим, исследователю необходимо опросить жителей некоторого района. Район включает 100 пятиэтажных домов, в каждом из которых проживает, в среднем, 160 жителей. Население района составляет 16 000 жителей. Мы можем случайным образом отобрать 5 домов и опросить всех их жителей. В выборке окажется тогда 800 человек, что составит 5% населения района. Мы не будем рассматривать далее кластерную и систематическую выборку.

После того, как объект извлечен из генеральной совокупности для включения в выборку, его либо возвращают в генеральную совокупность, либо нет. Если его возвратили, он может попасть в выборку повторно.

Выборка без возвращения – любой объект не может попасть в выборку больше одного раза. **Выборка с возвращением** – любой объект может оказаться в выборке более одного раза.

Например, интервьюер не обратится дважды к одному и тому же участнику протестного митинга, а выберет несколько различных человек. Тем не менее, иногда делаются выборки с возвращением, и это может быть оправдано целями исследования.

8-2 Распределение выборочного среднего

Параметры и статистики

Мы уже говорили, что *параметрами* будем называть характеристики генеральной совокупности. Параметр генеральной совокупности есть фиксированное число, которое нам не известно, при его вычислении случайность отсутствует. Тем самым, параметр есть неизвестная и фиксированная величина.

С другой стороны, *статистикой* мы назвали числовую характеристику выборки. Статистика является случайной величиной, так как в ее основе лежат данные, полученные в результате случайного отбора. Тем самым, статистика является известной и случайной величиной.

Статистики являются *оценочными функциями* параметров генеральной совокупности. Фактическое значение статистики, рассчитанное по данным выборки, назовем *оценкой* параметра генеральной совокупности.

Чтобы проследить связь между характеристиками генеральной совокупности и характеристиками выборки (между параметрами и статистиками), разберем следующий учебный пример. Предположим, мы имеем генеральную совокупность, состоящую из чисел 1, 2 и 5. Параметры такой генеральной совокупности:

Среднее значение	= 2,7
Медиана	= 2
Размах	= 4
Дисперсия	= 2,9
Стандартное отклонение	= 1,7
Доля нечетных чисел	= 0,67

Таблица 8-1 Параметры генеральной совокупности и выборки

Все возможные выборки объема n=2	Среднее	Медиана	Размах	Дисперсия	Стандартное отклонение	Доля нечетных чисел
1 1	1,0	1,0	0	0,0	0,0	1,0
1 2	1,5	1,5	1	0,5	0,7	0,5
1 5	3,0	3,0	4	8,0	2,8	1,0
2 1	1,5	1,5	1	0,5	0,7	0,5
2 2	2,0	2,0	0	0,0	0,0	0,0
2 5	3,5	3,5	3	4,5	2,1	0,5
5 1	3,0	3,0	4	8,0	2,8	1,0
5 2	3,5	3,5	3	4,5	2,1	0,5
5 5	5,0	5,0	0	0,0	0,0	1,0
Среднее значение статистики	2,7	2,67	1,78	2,9	1,3	0,7
Параметр генеральной совокупности	2,7	2	4	2,9	1,7	0,7

Это редкий случай, когда генеральная совокупность нам так хорошо известна. Обычно про генеральную совокупность известно очень мало, она сложнее устроена. В данном случае пример позволяет сравнить, какие значение принимают параметры генеральной совокупности и какие значения – статистики выборки.

Составим выборку из этой генеральной совокупности объема 2. Всего возможно составить девять различных выборок с возвращением. Мы перечислили их в таблице 8-1. Каждая выборка имеет равную вероятность 1/9. Стандартное отклонение генеральной совокупности и стандартное отклонение выборки вычисляются по следующим формулам:

$$\sigma = \sqrt{\frac{\sum (x-\mu)^2}{N}} \text{ и } s = \sqrt{\frac{\sum (x-\bar{x})^2}{n-1}}$$

Из таблицы видно, что совпадение среднего значения статистики со значением параметра генеральной совокупности происходит только для трех характеристик: среднего, дисперсии и доли. Эти три статистики **могут служить оценками параметров** генеральной совокупности. Три другие: медиана, размах и стандартное отклонение, как показал учебный расчет, **не могут служить** оценками соответствующих параметров генеральной совокупности.

Теперь нас будут интересовать три распределения: распределение генеральной совокупности, распределение выборки и новое для нас распределение – распределение статистики выборки, например, распределение выборочных средних или распределение выборочной доли.

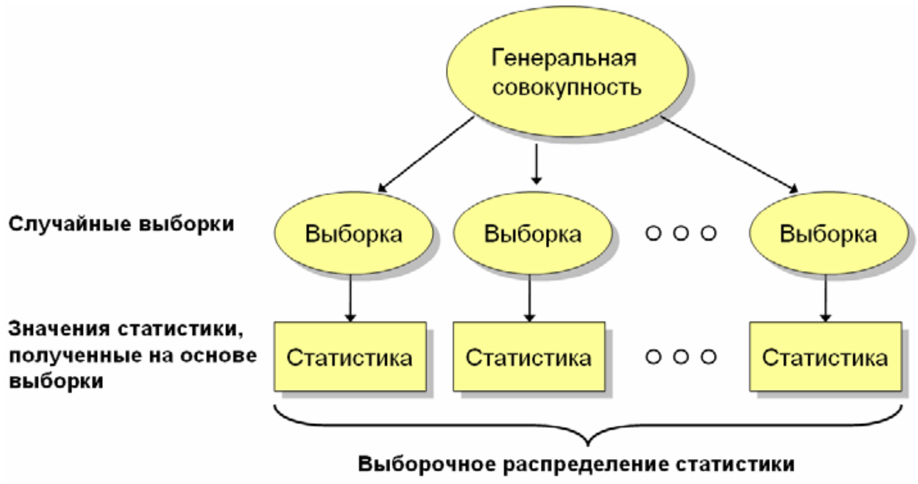


Рисунок 8-3 Выборочное распределение статистики

Три ключевых распределения

Поскольку из генеральной совокупности мы можем получить различные выборки, и каждая из них является случайной, мы имеем дело с выборочным распределением статистики. В учебном примере выборочное среднее, например, принимает значения 1, 1,5, 3 и другие, в зависимости от выборки.

Распределение выборочного среднего есть вероятностное распределение среднего значения выборки при условии, что рассматриваемые выборки имеют одинаковый объем *n*.

Поскольку сама выборка является случайной, тогда и статистики выборки являются случайными величинами и их значения зависят от состава элементов выборки. В этом смысле среднее значение выборки, дисперсия, доля и другие числовые характеристики могут рассматриваться как случайные величины, они должны иметь те или иные распределения и изучение распределений статистик, как функций выборки есть предмет изучения в настоящей главе.

Таблица 8-2 Распределение выборочного среднего

X	*P(X)*
1	1/9
1,5	2/9
2	1/9
3	2/9
3,5	2/9
5	1/9

Распределения	**Среднее значение**	**Стандартное отклонение**
Распределение генеральной совокупности	μ	σ
Распределение выборочных средних	$\mu_{\bar{x}}$	$\sigma_{\bar{x}}$
Распределение выборки	$\bar{x}$	s

Логика статистических заключений (или статистический вывод, statistical inference) основывается на трех ключевых распределениях: распределении генеральной совокупности, распределении выборочных средних и распределении выборки.

Центральная предельная теорема

В учебном примере мы получили, что среднее значение выборки равно среднему значению генеральной совокупности. Центральная предельная теорема объясняет нам, что если объем выборки достаточно большой, то распределение среднего значения выборки может быть приближено нормальным распределением, даже в том случае, если генеральная совокупность не распределена нормально. Хотя мы будем употреблять слово «теорема», мы не будем ее доказывать. Мы сосредоточимся на ее содержании и на том, как ее применять.

Центральная предельная теорема

Если: (1) Случайная величина имеет распределение, которое не обязано быть нормальным, со средним значением μ и стандартным отклонением σ, (2) Выборки объема n берутся из генеральной совокупности случайным образом.

То: (1) С ростом объема выборки распределение выборочного среднего стремится к нормальному, (2) Среднее значение всех выборочных средних совпадает со средним значением генеральной совокупности μ, (3) Стандартное отклонение выборочных средних равно $\sigma / \sqrt{n}$.

При применении теоремы следует иметь в виду следующее.

Во-первых, распределение выборочных средних *стремится к нормальному вне зависимости от вида распределения* генеральной совокупности. Это означает, что оно будет нормальным и в том случае, когда генеральная совокупность имеет ассиметричное или равномерное распределение.

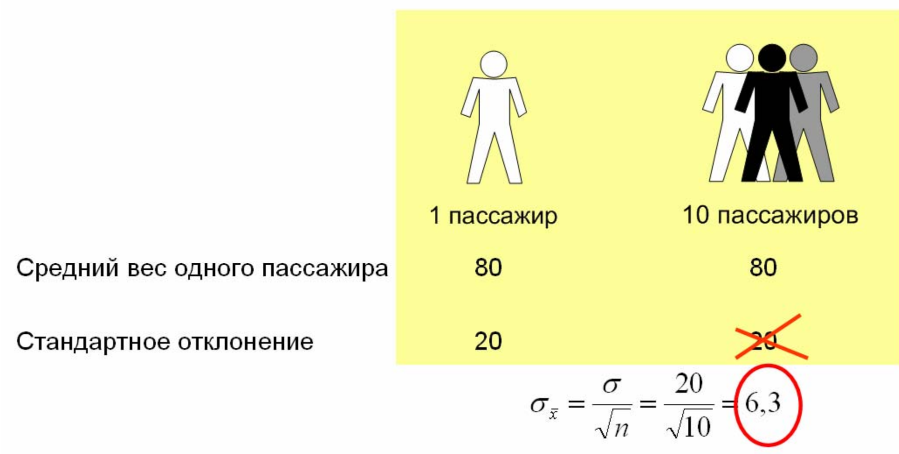


Рисунок 8-4 Стандартное отклонение выборочного среднего

Во-вторых, чем сильнее распределение генеральной совокупности отличается от нормального, тем большее влияние оказывает увеличение объема выборки на точность результата. Считается, что центральная предельная теорема дает для статистических заключений приемлемые результаты, если *объем выборки больше 30*.

В-третьих, если генеральная совокупность имеет нормальное распределение, тогда выборочная средняя будет распределена нормально *для выборок любого объема*.

Для всех возможных случайных выборок объема *n* из генеральной совокупности со средним μ и стандартным отклонением σ, выборочное среднее обозначим $\mu_{\bar{x}}$, стандартное отклонение выборочного среднего обозначим $\sigma_{\bar{x}}$. Тогда по теореме:

$$\mu_{\bar{x}} = \mu$$

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

Разберем значение этих формул на следующем примере. Средний вес пассажира лифта равен 80 кг со стандартным отклонением 20 кг. При проектировании лифта важно знать, на какую максимальную грузоподъемность следует рассчитывать, если в лифте планируется перевозить не более десяти пассажиров. Если вес одного человека имеет среднее μ=80 и стандартное отклонение σ=20, то вес одного человека в выборке из десяти человек имеет среднее 80 кг со стандартным отклонением 6,3 кг.

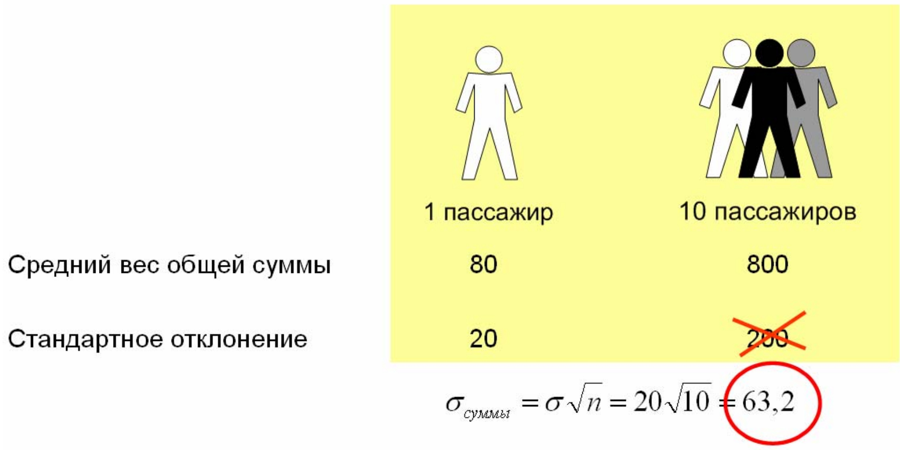


Рисунок 8-5 Стандартное отклонение общей суммы

При этом средний вес всех десяти человек составит 800 кг со стандартным отклонением 63,2 кг. Если воспользоваться правилом трех сигм, то мы получим максимально возможный вес:

$$\mu + 3\sigma = 800 + 3 \cdot 63{,}2 = 989{,}6$$

Мы получили, что вес 10 человек составит не более 990 кг, если выборка подбирается случайным образом.

В общем виде, среднее и стандартное отклонение для суммы запишется в виде формул следующим образом:

$$\mu_{суммы} = n \cdot \mu$$
$$\sigma_{суммы} = \sigma \cdot \sqrt{n}$$

При решении этой задачи мы существенно опирались на случайность выборки. Это следует иметь в виду, поскольку если выборка не будет случайной, вполне возможна ситуация, когда мы пригласим «прокатиться» на лифте десять толстяков, вес каждого из которых составит 80+3 · 20=140 кг. Тогда их сумма весов будет равна 1400 кг, что существенно превышает полученную нами границу.

Резюме

Мы рассмотрели распределение выборочного среднего и выяснили, что оно стремится к нормальному и имеет стандартное отклонение, зависящее от объема выборки. Чем больше выборка, тем меньше стандартное отклонение выборочного среднего. Центральная предельная теорема гарантирует нам соблюдение этих правил, что дает основания для получения обоснованных статистических заключений о генеральной совокупности на основе изучения выборки.

8-3 Стандартная ошибка

Любое распределение характеризуется стандартным отклонением. Точное значение стандартного отклонения генеральной совокупности не известно. Для получения оценок рассматривают стандартную ошибку среднего.

Стандартная ошибка среднего

Стандартная ошибка среднего оценивает изменчивость выборочного среднего, приближенно показывая, насколько выборочное среднее отличается от среднего генеральной совокупности.

Стандартное отклонение среднего, как мы установили в предыдущем параграфе, рассчитывается по формуле:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

Поскольку значение σ не известно, для оценивания стандартного отклонения среднего вычисляют стандартную ошибку среднего:

$$s_{\bar{x}} = \frac{s}{\sqrt{n}}$$

Это есть формула для стандартной ошибки среднего. Стандартное отклонение выборки s показывает, на сколько значения выборки отклоняются от выборочного среднего в то время, как стандартная ошибка среднего $s_{\bar{x}}$ показывает, на сколько выборочные средние отличаются от среднего генеральной совокупности.

Поправка для малой генеральной совокупности

Если объем генеральной совокупности небольшой и выборка составляет значительную часть совокупности, стандартную ошибку можно уменьшить, введя поправочный коэффициент для конечной генеральной совокупности.

Скорректированная стандартная ошибка среднего есть стандартная ошибка среднего, умноженная на поправочный коэффициент в случае, если генеральная совокупность конечна и сопоставима по размерам с выборкой.

Скорректированная стандартная ошибка запишется в виде:

$$\sqrt{\frac{N-n}{N-1}} \cdot s_{\bar{x}}$$

Если размер выборки приближается к размеру генеральной совокупности, значение (N–n) уменьшается, значение скорректированной ошибки также уменьшается, что отражает высокое качество оценки, полученной по большой части генеральной совокупности. Если N велико, то поправочный коэффициент близок к 1 и не оказывает влияния на величину ошибки.

Как определить, когда корректировка необходима? Статистики обычно соглашаются с тем, что корректирующий коэффициент необходимо применять, когда в выборке содержится более 5% значений генеральной совокупности, то есть, когда объем выборки n превышает 5% размера генеральной совокупности N.

Стандартная ошибка доли

В случае биномиального распределения имеют место две ошибки: для частоты m и для доли m/n.

Стандартная ошибка доли есть оценка стандартного отклонения доли признака в генеральной совокупности, полученная по выборке.

Стандартное отклонение для частоты и для доли признака запишутся при помощи формул следующим образом:

	Частота событий, m	**Доля,** $\hat{p} = \frac{m}{n}$
Стандартное отклонение (для генеральной совокупности)	$\sigma_m = \sqrt{np(1-p)}$	$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$
Стандартная ошибка (оценка по выборке)	$s_m = \sqrt{n\hat{p}(1-\hat{p})}$	$s_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

Неизвестная доля признака в генеральной совокупности обозначена *p*.

Рассмотрим несложный пример. Предположим, что при обследовании 50 индивидуумов у 8 обнаружены отклонения по здоровью. Это означает, что 16% обследованных имеют отклонения. Мы имеем возможность вычислить стандартные отклонения для частоты и для доли признака.

Используем формулы:

$$s_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{0{,}16\cdot(1-0{,}16)}{50}} = 0{,}0518$$

$$s_m = \sqrt{n\hat{p}(1-\hat{p})} = \sqrt{50\cdot 0{,}16\cdot(1-0{,}16)} = 2{,}59$$

Мы получили, что доля имеющих отклонения по здоровью в генеральной совокупности равна 16% с неопределенностью 5,18%. При этом наблюдаемая частота равна 8 с неопределенностью 2,59.

Используем компьютер

По материалам этой главы при помощи компьютера следует научиться создавать случайные выборки. В электронных таблицах и статистических пакетах имеются встроенные генераторы случайных чисел. В таблицах EXCEL, например, функция СЛЧИС( ) позволяет получить случайным образом число в интервале от нуля до единицы. Если использовать сдвиг и растяжение этой функции, можно получить случайное число в любом интервале. Компьютер позволяет также проводить эксперимент, в котором можно смоделировать генеральную совокупность и проверить, насколько случайные выборки дают хорошие приближения параметров генеральной совокупности. Если выборка достаточно велика, будет хорошо видно, что статистики хорошо приближают соответствующие параметры.

Что означают термины

Целевая генеральная совокупность	**Систематическая выборка**	**Стандартная ошибка среднего**
Основа выборочного наблюдения	**Кластерная выборка**	**Скорректированная ошибка среднего**
Перепись	**Выборка без возвращения**	**Стандартная ошибка доли**
Простая случайная выборка	**Распределение выборочного среднего**	**Параметры**
Стратифицированная выборка	**Центральная предельная теорема**	**Статистики**

Символы и формулы

n	Объем выборки
N	Объем генеральной совокупности
$\mu_{\bar{x}}$	Среднее для распределения выборочных средних
$\sigma_{\bar{x}}$	Стандартное отклонение для распределения выборочных средних
$\mu_{\bar{x}} = \mu$	Среднее для распределения выборочных средних
$\sigma_{\bar{x}} = \dfrac{\sigma}{\sqrt{n}}$	Стандартное отклонение для распределения выборочных средних
$\mu_{суммы} = n \cdot \mu$	Среднее для суммы выборки
$\sigma_{суммы} = \sigma \cdot \sqrt{n}$	Стандартное отклонение для суммы выборки
$s_{\bar{x}} = \dfrac{s}{\sqrt{n}}$	Стандартная ошибка среднего
$\sqrt{\dfrac{N-n}{N}} \cdot s_{\bar{x}}$	Поправка для малой генеральной совокупности
$s_{\hat{p}} = \sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}}$	Стандартная ошибка для доли
$s_m = \sqrt{n\hat{p}(1-\hat{p})}$	Стандартная ошибка для частоты

Задачи и упражнения

8-1. Нужен ли корректирующий коэффициент? Определите необходимость использования корректирующего коэффициента в перечисленных ниже ситуациях и, если это необходимо, вычислите его.

а) N = 500, n = 50
б) N = 100, n = 20
в) N = 2000, n = 50
г) N = 5000, n = 200
д) N =981, n = 50
е) N = 2032, n = 100
ж) N =562, n = 30
з) N = 3215, n = 150
и) N = 8000, n = 500
к) N = 10000, n = 400

В задачах 8-2…8-5 мы предполагаем, что выборка сделана из большой генеральной совокупности и корректирующий

коэффициент не требуется. Если размер выборки меньше 30, то предполагаем, что переменная распределена нормально.

8-2. Средний балл. Предположим, что средний балл студентов большого университета составляет 3,23 при стандартном отклонении 0,72. На отделении обучается 48 студентов. Определите вероятность того, что средний балл на отделении будет ниже 3,15.

8-3. Тестирование детей. Средняя оценка тестов на сообразительность для 12-летних детей составляют 30 со стандартным отклонением 5. Когда психолог проводит тест в классе из 22 учащихся, определите вероятность того, что средний показатель в отобранной группе будет находиться в пределах от 27 до 31.

8-4. Возраст бухгалтеров. Средний возраст бухгалтеров составляет 43 года со стандартным отклонением 5 лет. Если в фирме бухгалтерского сервиса работает 30 бухгалтеров, то определите вероятность того, что средний возраст в отобранной группе будет превышать 44,2 года.

8-5. Время выполнения теста. Среднее время, необходимое для группы взрослых людей, чтобы выполнить тест составляет 46,2 минуты. Стандартное отклонение 8 минут.

а) Определите вероятность того, что произвольно выбранный взрослый человек закончит выполнение этого теста быстрее, чем за 43 минуты.

б) Определите вероятность того, что в случае выполнения теста группой взрослых людей из 50 человек, среднее время выполнения теста в группе составит менее 43 минут.

8-6. Галстуки в бутике. Средняя цена галстуков, продаваемых в одном из бутиков, составляет 40 долларов со стандартным отклонением 5 долларов. Если из 20 галстуков продано 8, то определите вероятность того, что их средняя цена составляет 43 доллара. Не забудьте проверить необходимость применения корректирующего коэффициента.

8-7. Вычислите стандартную ошибку. Стандартное отклонение при измерении IQ составляет 15. Для объема выборки из 100 человек вычислите стандартную ошибку среднего. Какого размера следует взять выборку, чтобы увеличить стандартную ошибку вдвое? Чтобы уменьшить вдвое?

Приложение А. Таблицы

Приведенные таблицы часто используются для нахождения значений некоторых функций, важнейшей из которых является плотность распределения стандартного нормального закона. Часть таблиц не будет использована при решении задач первой части курса, а понадобится при решении задач во второй части курса.

ПЕРЕЧЕНЬ ТАБЛИЦ

Таблица А-0. Таблица случайных чисел

79	41	71	93	60	35	04	67	96	04	79	10	86
26	52	53	13	43	50	92	09	87	21	83	75	17
18	13	41	30	56	20	37	74	49	56	45	46	83
19	82	02	69	34	27	77	34	24	93	16	77	00
14	57	44	30	93	76	32	13	55	29	49	30	77
29	12	18	50	65	33	15	79	50	28	50	45	45
01	27	92	67	62	31	97	55	29	21	64	27	29
55	75	65	68	82	73	07	95	66	43	43	92	16
84	95	95	96	13	30	91	64	74	83	47	89	71
62	62	21	37	29	62	19	44	08	64	34	50	11
66	57	28	69	75	99	74	31	58	19	47	66	89
48	13	69	97	01	01	75	58	05	40	40	18	29
94	31	73	19	80	76	33	18	05	53	04	51	41
00	06	53	98	62	55	08	38	49	42	10	44	38
46	16	44	27	39	15	28	01	64	27	89	03	27
77	49	85	95	23	93	25	39	63	74	54	82	85
81	96	43	27	06	53	85	61	12	90	67	96	02
40	46	15	73	93	75	96	68	13	99	49	64	11

Таблица А-1. Биномиальное распределение

n = общее количество испытаний
k = количество успехов
p = вероятность успеха в одном испытании

		P													
n	**k**	**0,01**	**0,05**	**0,10**	**0,20**	**0,30**	**0,40**	**0,50**	**0,60**	**0,70**	**0,80**	**0,90**	**0,95**	**0,99**	**k**
2	**0**	980	903	810	640	490	360	250	160	090	040	010	003	000	**0**
	1	020	095	180	320	420	480	500	480	420	320	180	095	020	**1**
	2	000	003	010	040	090	160	250	360	490	640	810	903	980	**2**
3	**0**	970	857	729	512	343	216	125	064	027	008	001	000	000	**0**
	1	029	135	243	384	441	432	375	288	189	096	027	007	000	**1**
	2	000	007	027	096	189	288	375	432	441	384	243	135	029	**2**
	3	000	000	001	008	027	064	125	216	343	512	729	857	970	**3**
4	**0**	961	815	656	410	240	130	063	026	008	002	000	000	000	**0**
	1	039	171	292	410	412	346	250	154	076	026	004	000	000	**1**
	2	001	014	049	154	265	346	375	346	265	154	049	014	001	**2**
	3	000	000	004	026	076	154	250	346	412	410	292	171	039	**3**
	4	000	000	000	002	008	026	063	130	240	410	656	815	961	**4**
5	**0**	951	774	590	328	168	078	031	010	002	000	000	000	000	**0**
	1	048	204	328	410	360	259	156	077	028	006	000	000	000	**1**
	2	001	021	073	205	309	346	313	230	132	051	008	001	000	**2**
	3	000	001	008	051	132	230	313	346	309	205	073	021	001	**3**
	4	000	000	000	006	028	077	156	259	360	410	328	204	048	**4**
	5	000	000	000	000	002	010	031	078	168	328	590	774	951	**5**
6	**0**	941	735	531	262	118	047	016	004	001	000	000	000	000	**0**
	1	057	232	354	393	303	187	094	037	010	002	000	000	000	**1**
	2	001	031	098	246	324	311	234	138	060	015	001	000	000	**2**
	3	000	002	015	082	185	276	313	276	185	082	015	002	000	**3**
	4	000	000	001	015	060	138	234	311	324	246	098	031	001	**4**
	5	000	000	000	002	010	037	094	187	303	393	354	232	057	**5**
	6	000	000	000	000	001	004	016	047	118	262	531	735	941	**6**
7	**0**	932	698	478	210	082	028	008	002	000	000	000	000	000	**0**
	1	066	257	372	367	247	131	055	017	004	000	000	000	000	**1**
	2	002	041	124	275	318	261	164	077	025	004	000	000	000	**2**
	3	000	004	023	115	227	290	273	194	097	029	003	000	000	**3**
	4	000	000	003	029	097	194	273	290	227	115	023	004	000	**4**
	5	000	000	000	004	025	077	164	261	318	275	124	041	002	**5**
	6	000	000	000	000	004	017	055	131	247	367	372	257	066	**6**
	7	000	000	000	000	000	002	008	028	082	210	478	698	932	**7**
8	**0**	923	663	430	168	058	017	004	001	000	000	000	000	000	**0**
	1	075	279	383	336	198	090	031	008	001	000	000	000	000	**1**
	2	003	051	149	294	296	209	109	041	010	001	000	000	000	**2**
	3	000	005	033	147	254	279	219	124	047	009	000	000	000	**3**
	4	000	000	005	046	136	232	273	232	136	046	005	000	000	**4**
	5	000	000	000	009	047	124	219	279	254	147	033	005	000	**5**
	6	000	000	000	001	010	041	109	209	296	294	149	051	003	**6**
	7	000	000	000	000	001	008	031	090	198	336	383	279	075	**7**
	8	000	000	000	000	000	001	004	017	058	168	430	663	923	**8**

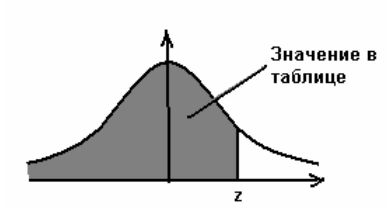
		P													
n	**k**	**0,01**	**0,05**	**0,10**	**0,20**	**0,30**	**0,40**	**0,50**	**0,60**	**0,70**	**0,80**	**0,90**	**0,95**	**0,99**	**k**
9	**0**	914	630	387	134	040	010	002	000	000	000	000	000	000	**0**
	1	083	299	387	302	156	060	018	004	000	000	000	000	000	**1**
	2	003	063	172	302	267	161	070	021	004	000	000	000	000	**2**
	3	000	008	045	176	267	251	164	074	021	003	000	000	000	**3**
	4	000	001	007	066	172	251	246	167	074	017	001	000	000	**4**
	5	000	000	001	017	074	167	246	251	172	066	007	001	000	**5**
	6	000	000	000	003	021	074	164	251	267	176	045	008	000	**6**
	7	000	000	000	000	004	021	070	161	267	302	172	063	003	**7**
	8	000	000	000	000	000	004	018	060	156	302	387	299	083	**8**
	9	000	000	000	000	000	000	002	010	040	134	387	630	914	**9**
10	**0**	904	599	349	107	028	006	001	000	000	000	000	000	000	**0**
	1	091	315	387	268	121	040	010	002	000	000	000	000	000	**1**
	2	004	075	194	302	233	121	044	011	001	000	000	000	000	**2**
	3	000	010	057	201	267	215	117	042	009	001	000	000	000	**3**
	4	000	001	011	088	200	251	205	111	037	006	000	000	000	**4**
	5	000	000	001	026	103	201	246	201	103	026	001	000	000	**5**
	6	000	000	000	006	037	111	205	251	200	088	011	001	000	**6**
	7	000	000	000	001	009	042	117	215	267	201	057	010	000	**7**
	8	000	000	000	000	001	011	044	121	233	302	194	075	004	**8**
	9	000	000	000	000	000	002	010	040	121	268	387	315	091	**9**
	10	000	000	000	000	000	000	001	006	028	107	349	599	904	**10**
11	**0**	895	569	314	086	020	004	000	000	000	000	000	000	000	**0**
	1	099	329	384	236	093	027	005	001	000	000	000	000	000	**1**
	2	005	087	213	295	200	089	027	005	001	000	000	000	000	**2**
	3	000	014	071	221	257	177	081	023	004	000	000	000	000	**3**
	4	000	001	016	111	220	236	161	070	017	002	000	000	000	**4**
	5	000	000	002	039	132	221	226	147	057	010	000	000	000	**5**
	6	000	000	000	010	057	147	226	221	132	039	002	000	000	**6**
	7	000	000	000	002	017	070	161	236	220	111	016	001	000	**7**
	8	000	000	000	000	004	023	081	177	257	221	071	014	000	**8**
	9	000	000	000	000	001	005	027	089	200	295	213	087	005	**9**
	10	000	000	000	000	000	001	005	027	093	236	384	329	099	**10**
	11	000	000	000	000	000	000	000	004	020	086	314	569	895	**11**
12	**0**	886	540	282	069	014	002	000	000	000	000	000	000	000	**0**
	1	107	341	377	206	071	017	003	000	000	000	000	000	000	**1**
	2	006	099	230	283	168	064	016	002	000	000	000	000	000	**2**
	3	000	017	085	236	240	142	054	012	001	000	000	000	000	**3**
	4	000	002	021	133	231	213	121	042	008	001	000	000	000	**4**
	5	000	000	004	053	158	227	193	101	029	003	000	000	000	**5**
	6	000	000	000	016	079	177	226	177	079	016	000	000	000	**6**
	7	000	000	000	003	029	101	193	227	158	053	004	000	000	**7**
	8	000	000	000	001	008	042	121	213	231	133	021	002	000	**8**
	9	000	000	000	000	001	012	054	142	240	236	085	017	000	**9**
	10	000	000	000	000	000	002	016	064	168	283	230	099	006	**10**
	11	000	000	000	000	000	000	003	017	071	206	377	341	107	**11**
	12	000	000	000	000	000	000	000	002	014	069	282	540	886	**12**

Таблица А-1. Биномиальное распределение (продолжение)

n	k	0,01	0,05	0,10	0,20	0,30	0,40	0,50	0,60	0,70	0,80	0,90	0,95	0,99	k
13	0	878	513	254	055	010	001	000	000	000	000	000	000	000	0
	1	115	351	367	179	054	011	002	000	000	000	000	000	000	1
	2	007	111	245	268	139	045	010	001	000	000	000	000	000	2
	3	000	021	100	246	218	111	035	006	001	000	000	000	000	3
	4	000	003	028	154	234	184	087	024	003	000	000	000	000	4
	5	000	000	006	069	180	221	157	066	014	001	000	000	000	5
	6	000	000	001	023	103	197	209	131	044	006	000	000	000	6
	7	000	000	000	006	044	131	209	197	103	023	001	000	000	7
	8	000	000	000	001	014	066	157	221	180	069	006	000	000	8
	9	000	000	000	000	003	024	087	184	234	154	028	003	000	9
	10	000	000	000	000	001	006	035	111	218	246	100	021	000	10
	11	000	000	000	000	000	001	010	045	139	268	245	111	007	11
	12	000	000	000	000	000	000	002	011	054	179	367	351	115	12
	13	000	000	000	000	000	000	000	001	010	055	254	513	878	13
14	0	869	488	229	044	007	001	000	000	000	000	000	000	000	0
	1	123	359	356	154	041	007	001	000	000	000	000	000	000	1
	2	008	123	257	250	113	032	006	001	000	000	000	000	000	2
	3	000	026	114	250	194	085	022	003	000	000	000	000	000	3
	4	000	004	035	172	229	155	061	014	001	000	000	000	000	4
	5	000	000	008	086	196	207	122	041	007	000	000	000	000	5
	6	000	000	001	032	126	207	183	092	023	002	000	000	000	6
	7	000	000	000	009	062	157	209	157	062	009	000	000	000	7
	8	000	000	000	002	023	092	183	207	126	032	001	000	000	8
	9	000	000	000	000	007	041	122	207	196	086	008	000	000	9
	10	000	000	000	000	001	014	061	155	229	172	035	004	000	10
	11	000	000	000	000	000	003	022	085	194	250	114	026	000	11
	12	000	000	000	000	000	001	006	032	113	250	257	123	008	12
	13	000	000	000	000	000	000	001	007	041	154	356	359	123	13
	14	000	000	000	000	000	000	000	001	007	044	229	488	869	14
15	0	860	463	206	035	005	000	000	000	000	000	000	000	000	0
	1	130	366	343	132	031	005	000	000	000	000	000	000	000	1
	2	009	135	267	231	092	022	003	000	000	000	000	000	000	2
	3	000	031	129	250	170	063	014	002	000	000	000	000	000	3
	4	000	005	043	188	219	127	042	007	001	000	000	000	000	4
	5	000	001	010	103	206	186	092	024	003	000	000	000	000	5
	6	000	000	002	043	147	207	153	061	012	001	000	000	000	6
	7	000	000	000	014	081	177	196	118	035	003	000	000	000	7
	8	000	000	000	003	035	118	196	177	081	014	000	000	000	8
	9	000	000	000	001	012	061	153	207	147	043	002	000	000	9
	10	000	000	000	000	003	024	092	186	206	103	010	001	000	10
	11	000	000	000	000	001	007	042	127	219	188	043	005	000	11
	12	000	000	000	000	000	002	014	063	170	250	129	031	000	12
	13	000	000	000	000	000	000	003	022	092	231	267	135	009	13
	14	000	000	000	000	000	000	000	005	031	132	343	366	130	14
	15	000	000	000	000	000	000	000	000	005	035	206	463	860	15

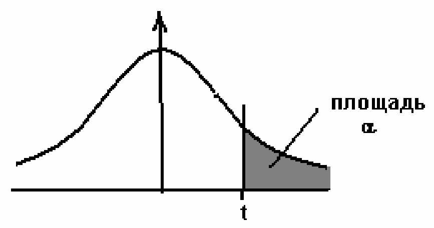
n	k	0,01	0,05	0,10	0,20	0,30	0,40	0,50	0,60	0,70	0,80	0,90	0,95	0,99	k
16	**0**	851	440	185	028	003	000	000	000	000	000	000	000	000	**0**
	1	138	371	329	113	023	003	000	000	000	000	000	000	000	**1**
	2	010	146	275	211	073	015	002	000	000	000	000	000	000	**2**
	3	000	036	142	246	146	047	009	001	000	000	000	000	000	**3**
	4	000	006	051	200	204	101	028	004	000	000	000	000	000	**4**
	5	000	001	014	120	210	162	067	014	001	000	000	000	000	**5**
	6	000	000	003	055	165	198	122	039	006	000	000	000	000	**6**
	7	000	000	000	020	101	189	175	084	019	001	000	000	000	**7**
	8	000	000	000	006	049	142	196	142	049	006	000	000	000	**8**
	9	000	000	000	001	019	084	175	189	101	020	000	000	000	**9**
	10	000	000	000	000	006	039	122	198	165	055	003	000	000	**10**
	11	000	000	000	000	001	014	067	162	210	120	014	001	000	**11**
	12	000	000	000	000	000	004	028	101	204	200	051	006	000	**12**
	13	000	000	000	000	000	001	009	047	146	246	142	036	000	**13**
	14	000	000	000	000	000	000	002	015	073	211	275	146	010	**14**
	15	000	000	000	000	000	000	000	003	023	113	329	371	138	**15**
	16	000	000	000	000	000	000	000	000	003	028	185	440	851	**16**
17	**0**	843	418	167	023	002	000	000	000	000	000	000	000	000	**0**
	1	145	374	315	096	017	002	000	000	000	000	000	000	000	**1**
	2	012	158	280	191	058	010	001	000	000	000	000	000	000	**2**
	3	001	041	156	239	125	034	005	000	000	000	000	000	000	**3**
	4	000	008	060	209	187	080	018	002	000	000	000	000	000	**4**
	5	000	001	017	136	208	138	047	008	001	000	000	000	000	**5**
	6	000	000	004	068	178	184	094	024	003	000	000	000	000	**6**
	7	000	000	001	027	120	193	148	057	009	000	000	000	000	**7**
	8	000	000	000	008	064	161	185	107	028	002	000	000	000	**8**
	9	000	000	000	002	028	107	185	161	064	008	000	000	000	**9**
	10	000	000	000	000	009	057	148	193	120	027	001	000	000	**10**
	11	000	000	000	000	003	024	094	184	178	068	004	000	000	**11**
	12	000	000	000	000	001	008	047	138	208	136	017	001	000	**12**
	13	000	000	000	000	000	002	018	080	187	209	060	008	000	**13**
	14	000	000	000	000	000	000	005	034	125	239	156	041	001	**14**
	15	000	000	000	000	000	000	001	010	058	191	280	158	012	**15**
	16	000	000	000	000	000	000	000	002	017	096	315	374	145	**16**
	17	000	000	000	000	000	000	000	000	002	023	167	418	843	**17**
18	**0**	835	397	150	018	002	000	000	000	000	000	000	000	000	**0**
	1	152	376	300	081	013	001	000	000	000	000	000	000	000	**1**
	2	013	168	284	172	046	007	001	000	000	000	000	000	000	**2**
	3	001	047	168	230	105	025	003	000	000	000	000	000	000	**3**
	4	000	009	070	215	168	061	012	001	000	000	000	000	000	**4**
	5	000	001	022	151	202	115	033	004	000	000	000	000	000	**5**
	6	000	000	005	082	187	166	071	015	001	000	000	000	000	**6**
	7	000	000	001	035	138	189	121	037	005	000	000	000	000	**7**
	8	000	000	000	012	081	173	167	077	015	001	000	000	000	**8**
	9	000	000	000	003	039	128	185	128	039	003	000	000	000	**9**
	10	000	000	000	001	015	077	167	173	081	012	000	000	000	**10**
	11	000	000	000	000	005	037	121	189	138	035	001	000	000	**11**
	12	000	000	000	000	001	015	071	166	187	082	005	000	000	**12**
	13	000	000	000	000	000	004	033	115	202	151	022	001	000	**13**
	14	000	000	000	000	000	001	012	061	168	215	070	009	000	**14**
	15	000	000	000	000	000	000	003	025	105	230	168	047	001	**15**
	16	000	000	000	000	000	000	001	007	046	172	284	168	013	**16**
	17	000	000	000	000	000	000	000	001	013	081	300	376	152	**17**
	18	000	000	000	000	000	000	000	000	002	018	150	397	835	**18**

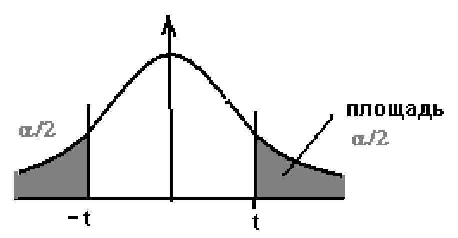
Таблица А-2. Нормальное распределение



z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753
0,2	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517
0,4	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879
0,5	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224
0,6	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549
0,7	0,7580	0,7611	0,7642	0,7673	0,7704	0,7734	0,7764	0,7794	0,7823	0,7852
0,8	0,7881	0,7910	0,7939	0,7967	0,7995	0,8023	0,8051	0,8078	0,8106	0,8133
0,9	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,8340	0,8365	0,8389
1,0	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621
1,1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	0,8790	0,8810	0,8830
1,2	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,8980	0,8997	0,9015
1,3	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147	0,9162	0,9177
1,4	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9279	0,9292	0,9306	0,9319
1,5	0,9332	0,9345	0,9357	0,9370	0,9382	0,9394	0,9406	0,9418	0,9429	0,9441
1,6	0,9452	0,9463	0,9474	0,9484	0,9495	0,9505	0,9515	0,9525	0,9535	0,9545
1,7	0,9554	0,9564	0,9573	0,9582	0,9591	0,9599	0,9608	0,9616	0,9625	0,9633
1,8	0,9641	0,9649	0,9656	0,9664	0,9671	0,9678	0,9686	0,9693	0,9699	0,9706
1,9	0,9713	0,9719	0,9726	0,9732	0,9738	0,9744	0,9750	0,9756	0,9761	0,9767
2,0	0,9772	0,9778	0,9783	0,9788	0,9793	0,9798	0,9803	0,9808	0,9812	0,9817
2,1	0,9821	0,9826	0,9830	0,9834	0,9838	0,9842	0,9846	0,9850	0,9854	0,9857
2,2	0,9861	0,9864	0,9868	0,9871	0,9875	0,9878	0,9881	0,9884	0,9887	0,9890
2,3	0,9893	0,9896	0,9898	0,9901	0,9904	0,9906	0,9909	0,9911	0,9913	0,9916
2,4	0,9918	0,9920	0,9922	0,9925	0,9927	0,9929	0,9931	0,9932	0,9934	0,9936
2,5	0,9938	0,9940	0,9941	0,9943	0,9945	0,9946	0,9948	0,9949	0,9951	0,9952
2,6	0,9953	0,9955	0,9956	0,9957	0,9959	0,9960	0,9961	0,9962	0,9963	0,9964
2,7	0,9965	0,9966	0,9967	0,9968	0,9969	0,9970	0,9971	0,9972	0,9973	0,9974
2,8	0,9974	0,9975	0,9976	0,9977	0,9977	0,9978	0,9979	0,9979	0,9980	0,9981
2,9	0,9981	0,9982	0,9982	0,9983	0,9984	0,9984	0,9985	0,9985	0,9986	0,9986
3,0	0,9987	0,9987	0,9987	0,9988	0,9988	0,9989	0,9989	0,9989	0,9990	0,9990
3,1	0,9990	0,9991	0,9991	0,9991	0,9992	0,9992	0,9992	0,9992	0,9993	0,9993
3,2	0,9993	0,9993	0,9994	0,9994	0,9994	0,9994	0,9994	0,9995	0,9995	0,9995
3,3	0,9995	0,9995	0,9995	0,9996	0,9996	0,9996	0,9996	0,9996	0,9996	0,9997
3,4	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9998
3,5	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998

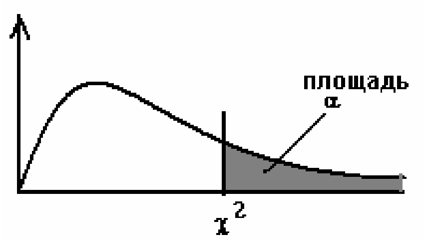
Таблица А-3. Критические точки распределения Стьюдента





Df	Односторонняя область				
	0,005	0,01	0,025	0,05	0,10
	Двусторонняя область				
	0,01	0,02	0,05	0,10	0,20
1	63,656	31,821	12,706	6,314	3,078
2	9,925	6,965	4,303	2,920	1,886
3	5,841	4,541	3,182	2,353	1,638
4	4,604	3,747	2,776	2,132	1,533
5	4,032	3,365	2,571	2,015	1,476
6	3,707	3,143	2,447	1,943	1,440
7	3,499	2,998	2,365	1,895	1,415
8	3,355	2,896	2,306	1,860	1,397
9	3,250	2,821	2,262	1,833	1,383
10	3,169	2,764	2,228	1,812	1,372
11	3,106	2,718	2,201	1,796	1,363
12	3,055	2,681	2,179	1,782	1,356
13	3,012	2,650	2,160	1,771	1,350
14	2,977	2,624	2,145	1,761	1,345
15	2,947	2,602	2,131	1,753	1,341
16	2,921	2,583	2,120	1,746	1,337
17	2,898	2,567	2,110	1,740	1,333
18	2,878	2,552	2,101	1,734	1,330
19	2,861	2,539	2,093	1,729	1,328
20	2,845	2,528	2,086	1,725	1,325
21	2,831	2,518	2,080	1,721	1,323
22	2,819	2,508	2,074	1,717	1,321
23	2,807	2,500	2,069	1,714	1,319
24	2,797	2,492	2,064	1,711	1,318
25	2,787	2,485	2,060	1,708	1,316
26	2,779	2,479	2,056	1,706	1,315
27	2,771	2,473	2,052	1,703	1,314
28	2,763	2,467	2,048	1,701	1,313
29	2,756	2,462	2,045	1,699	1,311
30	2,750	2,457	2,042	1,697	1,310
40	2,704	2,423	2,021	1,684	1,303
60	2,660	2,390	2,000	1,671	1,296
100	2,626	2,364	1,984	1,660	1,290
200	2,601	2,345	1,972	1,653	1,286
300	2,592	2,339	1,968	1,650	1,284
400	2,588	2,336	1,966	1,649	1,284
500	2,586	2,334	1,965	1,648	1,283

Таблица А-4. Критические точки распределения хи-квадрат (χ^2)



df	0,995	0,99	0,95	0,90	0,10	0,05	0,01	0,005
1	0,000	0,000	0,004	0,016	2,706	3,841	6,635	7,879
2	0,010	0,020	0,103	0,211	4,605	5,991	9,210	10,597
3	0,072	0,115	0,352	0,584	6,251	7,815	11,345	12,838
4	0,207	0,297	0,711	1,064	7,779	9,488	13,277	14,860
5	0,412	0,554	1,145	1,610	9,236	11,070	15,086	16,750
6	0,676	0,872	1,635	2,204	10,645	12,592	16,812	18,548
7	0,989	1,239	2,167	2,833	12,017	14,067	18,475	20,278
8	1,344	1,647	2,733	3,490	13,362	15,507	20,090	21,955
9	1,735	2,088	3,325	4,168	14,684	16,919	21,666	23,589
10	2,156	2,558	3,940	4,865	15,987	18,307	23,209	25,188
11	2,603	3,053	4,575	5,578	17,275	19,675	24,725	26,757
12	3,074	3,571	5,226	6,304	18,549	21,026	26,217	28,300
13	3,565	4,107	5,892	7,041	19,812	22,362	27,688	29,819
14	4,075	4,660	6,571	7,790	21,064	23,685	29,141	31,319
15	4,601	5,229	7,261	8,547	22,307	24,996	30,578	32,801
16	5,142	5,812	7,962	9,312	23,542	26,296	32,000	34,267
17	5,697	6,408	8,672	10,085	24,769	27,587	33,409	35,718
18	6,265	7,015	9,390	10,865	25,989	28,869	34,805	37,156
19	6,844	7,633	10,117	11,651	27,204	30,144	36,191	38,582
20	7,434	8,260	10,851	12,443	28,412	31,410	37,566	39,997
21	8,034	8,897	11,591	13,240	29,615	32,671	38,932	41,401
22	8,643	9,542	12,338	14,041	30,813	33,924	40,289	42,796
23	9,260	10,196	13,091	14,848	32,007	35,172	41,638	44,181
24	9,886	10,856	13,848	15,659	33,196	36,415	42,980	45,558
25	10,520	11,524	14,611	16,473	34,382	37,652	44,314	46,928
26	11,160	12,198	15,379	17,292	35,563	38,885	45,642	48,290
27	11,808	12,878	16,151	18,114	36,741	40,113	46,963	49,645
28	12,461	13,565	16,928	18,939	37,916	41,337	48,278	50,994
29	13,121	14,256	17,708	19,768	39,087	42,557	49,588	52,335
30	13,787	14,953	18,493	20,599	40,256	43,773	50,892	53,672
40	20,707	22,164	26,509	29,051	51,805	55,758	63,691	66,766
50	27,991	29,707	34,764	37,689	63,167	67,505	76,154	79,490
60	35,534	37,485	43,188	46,459	74,397	79,082	88,379	91,952
70	43,275	45,442	51,739	55,329	85,527	90,531	100,425	104,215
80	51,172	53,540	60,391	64,278	96,578	101,879	112,329	116,321
90	59,196	61,754	69,126	73,291	107,565	113,145	124,116	128,299
100	67,328	70,065	77,929	82,358	118,498	124,342	135,807	140,170

Таблица А-5. Критические точки распределения Фишера

$\alpha=0,05$

Степени свободы df_2	Степени свободы df_1: 1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	161	199	216	225	230	234	237	239	241	242	244	246	248	249	250	251	252	253	254
2	18,5	19,0	19,2	19,2	19,3	19,3	19,4	19,4	19,4	19,4	19,4	19,4	19,4	19,5	19,5	19,5	19,5	19,5	19,5
3	10,13	9,55	9,28	9,12	9,01	8,94	8,89	8,85	8,81	8,79	8,74	8,70	8,66	8,64	8,62	8,59	8,57	8,55	8,53
4	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6,00	5,96	5,91	5,86	5,80	5,77	5,75	5,72	5,69	5,66	5,63
5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,77	4,74	4,68	4,62	4,56	4,53	4,50	4,46	4,43	4,40	4,37
6	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10	4,06	4,00	3,94	3,87	3,84	3,81	3,77	3,74	3,70	3,67
7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,64	3,57	3,51	3,44	3,41	3,38	3,34	3,30	3,27	3,23
8	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39	3,35	3,28	3,22	3,15	3,12	3,08	3,04	3,01	2,97	2,93
9	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,14	3,07	3,01	2,94	2,90	2,86	2,83	2,79	2,75	2,71
10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,98	2,91	2,85	2,77	2,74	2,70	2,66	2,62	2,58	2,54
11	4,84	3,98	3,59	3,36	3,20	3,09	3,01	2,95	2,90	2,85	2,79	2,72	2,65	2,61	2,57	2,53	2,49	2,45	2,40
12	4,75	3,89	3,49	3,26	3,11	3,00	2,91	2,85	2,80	2,75	2,69	2,62	2,54	2,51	2,47	2,43	2,38	2,34	2,30
13	4,67	3,81	3,41	3,18	3,03	2,92	2,83	2,77	2,71	2,67	2,60	2,53	2,46	2,42	2,38	2,34	2,30	2,25	2,21
14	4,60	3,74	3,34	3,11	2,96	2,85	2,76	2,70	2,65	2,60	2,53	2,46	2,39	2,35	2,31	2,27	2,22	2,18	2,13
15	4,54	3,68	3,29	3,06	2,90	2,79	2,71	2,64	2,59	2,54	2,48	2,40	2,33	2,29	2,25	2,20	2,16	2,11	2,07
16	4,49	3,63	3,24	3,01	2,85	2,74	2,66	2,59	2,54	2,49	2,42	2,35	2,28	2,24	2,19	2,15	2,11	2,06	2,01
17	4,45	3,59	3,20	2,96	2,81	2,70	2,61	2,55	2,49	2,45	2,38	2,31	2,23	2,19	2,15	2,10	2,06	2,01	1,96
18	4,41	3,55	3,16	2,93	2,77	2,66	2,58	2,51	2,46	2,41	2,34	2,27	2,19	2,15	2,11	2,06	2,02	1,97	1,92
19	4,38	3,52	3,13	2,90	2,74	2,63	2,54	2,48	2,42	2,38	2,31	2,23	2,16	2,11	2,07	2,03	1,98	1,93	1,88
20	4,35	3,49	3,10	2,87	2,71	2,60	2,51	2,45	2,39	2,35	2,28	2,20	2,12	2,08	2,04	1,99	1,95	1,90	1,84
21	4,32	3,47	3,07	2,84	2,68	2,57	2,49	2,42	2,37	2,32	2,25	2,18	2,10	2,05	2,01	1,96	1,92	1,87	1,81
22	4,30	3,44	3,05	2,82	2,66	2,55	2,46	2,40	2,34	2,30	2,23	2,15	2,07	2,03	1,98	1,94	1,89	1,84	1,78
23	4,28	3,42	3,03	2,80	2,64	2,53	2,44	2,37	2,32	2,27	2,20	2,13	2,05	2,01	1,96	1,91	1,86	1,81	1,76
24	4,26	3,40	3,01	2,78	2,62	2,51	2,42	2,36	2,30	2,25	2,18	2,11	2,03	1,98	1,94	1,89	1,84	1,79	1,73
25	4,24	3,39	2,99	2,76	2,60	2,49	2,40	2,34	2,28	2,24	2,16	2,09	2,01	1,96	1,92	1,87	1,82	1,77	1,71
26	4,23	3,37	2,98	2,74	2,59	2,47	2,39	2,32	2,27	2,22	2,15	2,07	1,99	1,95	1,90	1,85	1,80	1,75	1,69
27	4,21	3,35	2,96	2,73	2,57	2,46	2,37	2,31	2,25	2,20	2,13	2,06	1,97	1,93	1,88	1,84	1,79	1,73	1,67
28	4,20	3,34	2,95	2,71	2,56	2,45	2,36	2,29	2,24	2,19	2,12	2,04	1,96	1,91	1,87	1,82	1,77	1,71	1,65
29	4,18	3,33	2,93	2,70	2,55	2,43	2,35	2,28	2,22	2,18	2,10	2,03	1,94	1,90	1,85	1,81	1,75	1,70	1,64
30	4,17	3,32	2,92	2,69	2,53	2,42	2,33	2,27	2,21	2,16	2,09	2,01	1,93	1,89	1,84	1,79	1,74	1,68	1,62
30	4,17	3,32	2,92	2,69	2,53	2,42	2,33	2,27	2,21	2,16	2,09	2,01	1,93	1,89	1,84	1,79	1,74	1,68	1,62
40	4,08	3,23	2,84	2,61	2,45	2,34	2,25	2,18	2,12	2,08	2,00	1,92	1,84	1,79	1,74	1,69	1,64	1,58	1,51
60	4,00	3,15	2,76	2,53	2,37	2,25	2,17	2,10	2,04	1,99	1,92	1,84	1,75	1,70	1,65	1,59	1,53	1,47	1,39
120	3,92	3,07	2,68	2,45	2,29	2,18	2,09	2,02	1,96	1,91	1,83	1,75	1,66	1,61	1,55	1,50	1,43	1,35	1,25

$\alpha=0{,}01$

Степени свободы df_2	Степени свободы df_1: 1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	4052	4999	5404	5624	5764	5859	5928	5981	6022	6056	6107	6157	6209	6234	6260	6286	6313	6340	6366
2	98,50	99,00	99,16	99,25	99,30	99,33	99,36	99,38	99,39	99,40	99,42	99,43	99,45	99,46	99,47	99,48	99,48	99,49	99,50
3	34,12	30,82	29,46	28,71	28,24	27,91	27,67	27,49	27,34	27,23	27,05	26,87	26,69	26,60	26,50	26,41	26,32	26,22	26,13
4	21,20	18,00	16,69	15,98	15,52	15,21	14,98	14,80	14,66	14,55	14,37	14,20	14,02	13,93	13,84	13,75	13,65	13,56	13,46
5	16,26	13,27	12,06	11,39	10,97	10,67	10,46	10,29	10,16	10,05	9,89	9,72	9,55	9,47	9,38	9,29	9,20	9,11	9,02
6	13,75	10,92	9,78	9,15	8,75	8,47	8,26	8,10	7,98	7,87	7,72	7,56	7,40	7,31	7,23	7,14	7,06	6,97	6,88
7	12,25	9,55	8,45	7,85	7,46	7,19	6,99	6,84	6,72	6,62	6,47	6,31	6,16	6,07	5,99	5,91	5,82	5,74	5,65
8	11,26	8,65	7,59	7,01	6,63	6,37	6,18	6,03	5,91	5,81	5,67	5,52	5,36	5,28	5,20	5,12	5,03	4,95	4,86
9	10,56	8,02	6,99	6,42	6,06	5,80	5,61	5,47	5,35	5,26	5,11	4,96	4,81	4,73	4,65	4,57	4,48	4,40	4,31
10	10,04	7,56	6,55	5,99	5,64	5,39	5,20	5,06	4,94	4,85	4,71	4,56	4,41	4,33	4,25	4,17	4,08	4,00	3,91
11	9,65	7,21	6,22	5,67	5,32	5,07	4,89	4,74	4,63	4,54	4,40	4,25	4,10	4,02	3,94	3,86	3,78	3,69	3,60
12	9,33	6,93	5,95	5,41	5,06	4,82	4,64	4,50	4,39	4,30	4,16	4,01	3,86	3,78	3,70	3,62	3,54	3,45	3,36
13	9,07	6,70	5,74	5,21	4,86	4,62	4,44	4,30	4,19	4,10	3,96	3,82	3,66	3,59	3,51	3,43	3,34	3,25	3,17
14	8,86	6,51	5,56	5,04	4,69	4,46	4,28	4,14	4,03	3,94	3,80	3,66	3,51	3,43	3,35	3,27	3,18	3,09	3,00
15	8,68	6,36	5,42	4,89	4,56	4,32	4,14	4,00	3,89	3,80	3,67	3,52	3,37	3,29	3,21	3,13	3,05	2,96	2,87
16	8,53	6,23	5,29	4,77	4,44	4,20	4,03	3,89	3,78	3,69	3,55	3,41	3,26	3,18	3,10	3,02	2,93	2,84	2,75
17	8,40	6,11	5,19	4,67	4,34	4,10	3,93	3,79	3,68	3,59	3,46	3,31	3,16	3,08	3,00	2,92	2,83	2,75	2,65
18	8,29	6,01	5,09	4,58	4,25	4,01	3,84	3,71	3,60	3,51	3,37	3,23	3,08	3,00	2,92	2,84	2,75	2,66	2,57
19	8,18	5,93	5,01	4,50	4,17	3,94	3,77	3,63	3,52	3,43	3,30	3,15	3,00	2,92	2,84	2,76	2,67	2,58	2,49
20	8,10	5,85	4,94	4,43	4,10	3,87	3,70	3,56	3,46	3,37	3,23	3,09	2,94	2,86	2,78	2,69	2,61	2,52	2,42
21	8,02	5,78	4,87	4,37	4,04	3,81	3,64	3,51	3,40	3,31	3,17	3,03	2,88	2,80	2,72	2,64	2,55	2,46	2,36
22	7,95	5,72	4,82	4,31	3,99	3,76	3,59	3,45	3,35	3,26	3,12	2,98	2,83	2,75	2,67	2,58	2,50	2,40	2,31
23	7,88	5,66	4,76	4,26	3,94	3,71	3,54	3,41	3,30	3,21	3,07	2,93	2,78	2,70	2,62	2,54	2,45	2,35	2,26
24	7,82	5,61	4,72	4,22	3,90	3,67	3,50	3,36	3,26	3,17	3,03	2,89	2,74	2,66	2,58	2,49	2,40	2,31	2,21
25	7,77	5,57	4,68	4,18	3,85	3,63	3,46	3,32	3,22	3,13	2,99	2,85	2,70	2,62	2,54	2,45	2,36	2,27	2,17
26	7,72	5,53	4,64	4,14	3,82	3,59	3,42	3,29	3,18	3,09	2,96	2,81	2,66	2,58	2,50	2,42	2,33	2,23	2,13
27	7,68	5,49	4,60	4,11	3,78	3,56	3,39	3,26	3,15	3,06	2,93	2,78	2,63	2,55	2,47	2,38	2,29	2,20	2,10
28	7,64	5,45	4,57	4,07	3,75	3,53	3,36	3,23	3,12	3,03	2,90	2,75	2,60	2,52	2,44	2,35	2,26	2,17	2,06
29	7,60	5,42	4,54	4,04	3,73	3,50	3,33	3,20	3,09	3,00	2,87	2,73	2,57	2,49	2,41	2,33	2,23	2,14	2,03
30	7,56	5,39	4,51	4,02	3,70	3,47	3,30	3,17	3,07	2,98	2,84	2,70	2,55	2,47	2,39	2,30	2,21	2,11	2,01
30	7,56	5,39	4,51	4,02	3,70	3,47	3,30	3,17	3,07	2,98	2,84	2,70	2,55	2,47	2,39	2,30	2,21	2,11	2,01
40	7,31	5,18	4,31	3,83	3,51	3,29	3,12	2,99	2,89	2,80	2,66	2,52	2,37	2,29	2,20	2,11	2,02	1,92	1,80
60	7,08	4,98	4,13	3,65	3,34	3,12	2,95	2,82	2,72	2,63	2,50	2,35	2,20	2,12	2,03	1,94	1,84	1,73	1,60
120	6,85	4,79	3,95	3,48	3,17	2,96	2,79	2,66	2,56	2,47	2,34	2,19	2,03	1,95	1,86	1,76	1,66	1,53	1,38
∞	6,64	4,61	3,78	3,32	3,02	2,80	2,64	2,51	2,41	2,32	2,18	2,04	1,88	1,79	1,70	1,59	1,47	1,32	1,00

Приложение В. Ответы к задачам

Глава 1.

1-2.
а, в, ж – описательная
б, г, д, е – аналитическая

1-3.
а, б, ж, з, к – непрерывные
в, г, д, е, и – дискретные

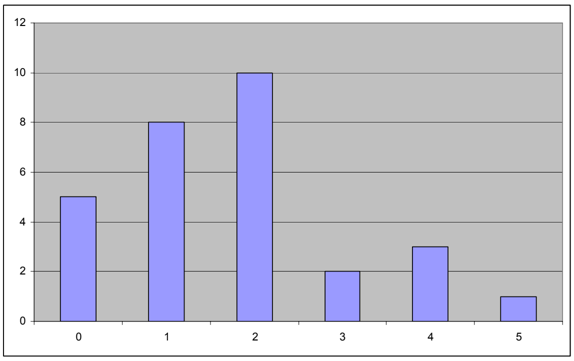
1-4.
а - качественные
б, в, г - количественные

1-5.
а – интервальная
б, ж – относительная
в – дихотомическая
г, д, е – номинальная
з – порядковая

Глава 2.

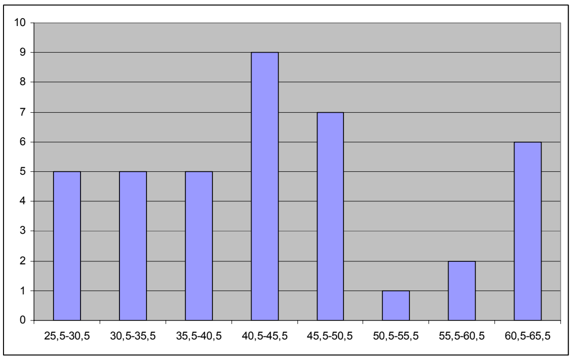
2-2.
Распределение частот и гистограмма:

x	f
0	5
1	8
2	10
3	2
4	3
5	1



2-3.
Интервальное распределение частот и гистограмма:

интервал	точные границы	частота
26-30	25,5-30,5	5
31-35	30,5-35,5	5
36-40	35,5-40,5	5
41-45	40,5-45,5	9
46-50	45,5-50,5	7
51-55	50,5-55,5	1
56-60	55,5-60,5	2
61-65	60,5-65,5	6



2-8.
Распределение частот:

x	f
Газета	7
Телевидение	5
Радио	6
Журнал	6

Глава 3.

3-1. Меры центральной тенденции и вариации:

Объем выборки	10
Среднее	403,4
Медиана	401,5
Мода	395,0
Стандартное отклонение	6,3
Дисперсия	40,27
Размах	20,0
Минимум	395,0
Максимум	415,0
Квартили 1	398,75
2	401,50
3	408,25

Распределение мультимодальное. В качестве моды взято наименьшее значение.

3-5. Таблица данных, дополненная вычислительными столбцами:

интервал	m	f	fm	fm
90-98	94	6	564	53 016
99-107	103	22	2 266	233 398
108-116	112	43	4 816	539 392
117-125	121	28	3 388	409 948
126-134	130	9	1 170	152 100
		108	12 204	1 387 854

Меры центральной тенденции и вариации:

Объем выборки	108
Среднее	113,0
Медиана	112
Мода	112
Стандартное отклонение	9,1
Дисперсия	82,26
Размах	44
Минимум	90
Максимум	134
Квартили 1	103
2	112
3	121

Глава 4.

4-1. 60.

4-2. 279 936.

4-3. 144.

4-4. 40 320.

4-5. 64.

4-6. 756.

4-7. 84.

4-8. 75 600.

4-9. 120.

4-10. 84.

4-12. 0,56.

4-13. 0,58.

4-14. 0,51.

4-15. а) 0,14; б) 0,54; в) 0,76.

4-16. а) 0,48; б) 0,74; в) 0,61.

4-**17**. 1/1024.

4-18. 0,31.

4-19. 0,00694.

4-20. 0,00058.

4-21. 0,126.

4-22. 0,90.

4-23. а) 0,65; б) 0,26.

4-24. а) 0,018; б) 0,33.

4-25. 0,33.

4-26. 0,64.

4-27. 0,625.

4-28. 0,643.

4-29. 0,266.

4-30. 0,766.

Глава 5.

5-1.
Закон распределения:

X	0	1	2	3
P	0,125	0,375	0,375	0,125

5-2.
Закон распределения:

X	0	1	2	3	4
P	0,0016	0,0256	0,1536	0,4096	0,4096

5-3.
а) MX=5; DX=4,2.
б) MX=0,4; DX=1,24.

5-4. MX=0,4; DX=0,28.

X	0	1	2
P	0,622	0,356	0,022

5-7. Среднее 2, дисперсия 1, стандартное отклонение 1.

5-9. Среднее 1,9; дисперсия 0,63; стандартное отклонение 0,8.

5-10. Построим распределение выигрыша.

X	-19	-15	0	30	80
P	0,5	0,25	0,15	0,05	0,05

Тогда MX=-7,75. Игра несправедлива, поскольку игрок в среднем проигрывает почти восемь долларов.

Глава 6.

6-1.

а) 1/2
б) ¼
в) ½
г) 0,25

6-2.
а) после дифференцирования получим ответ:

$$f(x)=\begin{cases} 0, & -\infty < x \le 0 \\ \cos x, & 0 < x \le \dfrac{\pi}{2} \\ 0, & \dfrac{\pi}{2} < x \end{cases}$$

6-3.
а) После интегрирования получим ответ:

$$F(x)=\begin{cases} 0, & -\infty < x \le 0 \\ 1-\cos x, & 0 < x \le \dfrac{\pi}{2} \\ 1, & \dfrac{\pi}{2} < x \end{cases}$$

6-4.
а) MX=3, DX=4/3, σ=1,15.

6-6. Вероятность равна 3/8.

Глава 7.

7-2.
а) 0,476
б) 0,212
в) 0,184
г) 0,481
д) 0,154
е) 0,409
ж) 0,337
з) 0,076
и) 0,081
к) 0,114
л) 0,225
м) 0,026
н) 0,463
о) 0,842
п) 0,889
р) 0,985
с) 0,973
т) 0,571
у) 0,068
ф) 0,301

7-3.
а) 0,454
б) 0,249
в) 0,391
г) 0,442
д) 0,005
е) 0,002
ж) 0,038
з) 0,066
и) 0,384
к) 0,952
л) 0,028
м) 0,071
н) 0,918
о) 0,078

7-4.
а) 0,0999
б) -1,8606
в) -0,5244; 0,5244.

7-5. 0,248

7-6. 0,091

7-7.
а) 0,059
б) 0,070

7-8.
а) 0,547
б) 0,268
в) 0,334

7-9.
а) 0,048
б) 0,159

7-13. 2,591.

7-15.
а) 0,080
б) 0,052
в) 0,106
г) 0,103
д) 0,165
е) 0,998

7-16. 0,754

7-18. 0,159

7-20. 0,197

Глава 8.

8-1.
Поправочный коэффициент не требуется, кроме указанных ситуаций:

а) 0,95
б) 0,90
д) 0,97
ж) 0,97
и) 0,97

8-2. 0,221

8-3. 0,823

8-4. 0,094

8-5.
а) 0,401
б) 0,039

8-7.
Стандартная ошибка 1,5.
Чтобы увеличить - 25.
Чтобы уменьшить – 400.

Примечание. Если вы обнаружили ошибку в ответе, не забудьте сообщить об этом по электронной почте автору. Любые ошибки и опечатки должны быть найдены и исправлены!

Приложение C. Библиография

Айвазян С.А., Мхитарян В.С. Теория вероятностей и прикладная статистика. – М. ЮНИТИ-ДАНА, 2001.

Вентцель Е.С. Теория вероятностей. – М. Физматгиз (Любое издание).

Гмурман В.Е. Теория вероятностей и математическая статистика. – М. Высшая школа (Любое издание).

Гмурман В.Е. Руководство к решению задач по теории вероятностей и математической статистике. – М. Высшая школа (Любое издание).

Гнеденко Б.В., Хинчин А.Я. Элементарное введение в теорию вероятностей. – М. Наука (Любое издание).

Ефимова М.Р., Петрова Е.В., Румянцев В.Н. Общая теория статистики. Учебник. – М. Инфра-М, 1999.

Калинина В.Н., Панкин В.Ф. Математическая статистика. – М. Высшая школа, 1998.

Малхотра Нэреш. Маркетинговые исследования. – М. Вильямс, 2003.

Рабочая книга социолога. Под ред. Осипова Г.В. – М. Едиториал УРСС, 2003.

Сигел Эндрю. Практическая бизнес-статистика. – М. Вильямс, 2004.

Теория статистики. Учебник. Под ред. Шмойловой Р.А. – М. Финансы и статистика, 2002.

Тюрин Ю.Н., Макаров А.А. Анализ данных на компьютере. Под ред Фигурнова В.Э. - М. Инфра-М, 2003.

Сулицкий В.Н. Методы статистического анализа в управлении. – М. Дело, 2002.

ПРЕДМЕТНЫЙ УКАЗАТЕЛЬ

СОДЕРЖАНИЕ

СВЕДЕНИЯ ОБ АВТОРЕ

Иванов Олег Валентинович, заведующий кафедрой социальной информатики социологического факультета МГУ им.М.В.Ломоносова, кандидат физико-математических наук.

На социологическом факультете МГУ в течение ряда лет руководит преподаванием цикла математических дисциплин, включающего основы математического анализа, теорию вероятностей, математическую статистику, информатику, теорию измерений и анализ данных.