Министерство образования Республики Беларусь

Учреждение образования
«Гомельский государственный университет
имени Франциска Скорины»

В. И. Харламова

ТЕОРИЯ ВЕРОЯТНОСТЕЙ И МАТЕМАТИЧЕСКАЯ СТАТИСТИКА

ПЕРВИЧНАЯ ОБРАБОТКА СТАТИСТИЧЕСКИХ ДАННЫХ

ПРАКТИЧЕСКОЕ ПОСОБИЕ
для студентов университета

Гомель
УО «ГГУ им. Ф. Скорины»
2009

УДК 519.22(075.8)
ББК 22.172я73
Х 211

Рецензенты: Ю. В. Малинковский, профессор, доктор физико-математических наук, заведующий кафедрой экономической кибернетики и теории вероятностей учреждения образования «Гомельский государственный университет имени Франциска Скорины»;
кафедра высшей математики учреждения образования «Гомельский государственный университет имени Франциска Скорины»

Рекомендовано к изданию научно-методическим советом учреждения образования «Гомельский государственный университет имени Франциска Скорины»

Харламова, В. И.

Х 211 Теория вероятностей и математическая статистика. Первичная обработка статистических данных: практическое пособие для студентов университета / В. И. Харламова; М-во образования РБ, Гомельский гос. ун-т им. Ф. Скорины. – Гомель : УО «ГГУ им. Ф. Скорины», 2009. – 112 с.
ISBN 978–985–439–451–0

В соответствии с учебной программой курса «Теория вероятностей и математическая статистика» для студентов экономического факультета в пособии содержится изложение важного раздела математической статистики, посвященного задачам и методам первичной математической обработки статистических данных. Большое внимание уделяется доступному и мотивированному описанию основных теоретических понятий математической статистики, четкому разъяснению вычислительных формул и построению алгоритмов статистических процедур, необходимых для обработки экспериментальных данных. Все вводимые понятия, рассуждения, вычислительные схемы иллюстрируются простыми конкретными примерами. К каждому разделу даны соответствующие упражнения для самостоятельной работы.

Пособие предназначено для активного изучения курса математической статистики студентами экономических специальностей, оно поможет студентам приобрести необходимые знания и выработать практические навыки правильной обработки экономической информации.

УДК 519.22(075.8)
ББК 22.172я73

ISBN 978–985–439–451–0

Содержание

Введение

Большинство видов коллективной и индивидуальной деятельности человека характеризуется количественными показателями. Повседневно мы погружаемся в поток разнообразной цифровой информации, например, о погоде и природных катаклизмах, о политике и политических рейтингах, об экономике и экономических прогнозах. Любые количественные сведения и оценки, полученные в результате систематических наблюдений или испытаний, являются статистическими данными, имеющими огромное информационное значение.

Математическая статистика является наукой о методах систематизации, анализа и интерпретации статистических данных.

Математическая статистика условно делится на 2 части: описательную и аналитическую. Описательная статистика помогает организовать правильную систему сбора статистических данных и их группировку в удобную информационную форму. Аналитическая статистика разрабатывает методы принятия решений и прогнозов на основе выявленных общих закономерностей.

Математическая статистика всегда имела широкое практическое применение. Статистические методы исследования являются основной частью научного метода познания. Тесное взаимодействие математической статистики с другими науками связано с наличием таких общих базовых понятий как массовость и измеримость большинства реальных явлений. Подчеркнем особую важность математической статистики для научных исследований в области физики, химии, биологии, психологии, медицины, социологии и других наук. Возможности для более широкого применения статистических методов обусловлены, прежде всего, бурным развитием самой математической статистики и ее приложений. К настоящему времени разработаны математические модели большого круга социальных и гуманитарных процессов, многие процедуры статистического анализа унифицированы и алгоритмизированы. Их применение стало общедоступным. Эффективность статистических методов существенно возрастает при использовании вычислительной техники. Современные компьютеры позволяют решать сложные задачи с большим объемом данных.

Интенсивное внедрение статистических методов в практику научных исследований требует улучшения математического образования будущих специалистов и руководителей самых разных профессиональных направлений. Традиционный университетский курс теории вероятностей и математической статистики для студентов нематематических специальностей особенно нуждается в модернизации. Современный курс должен соответствовать последним достижениям науки и иметь большую прикладную направленность с учетом специализации.

Данное практическое пособие представляет собой доступное введение в математическую статистику. Оно содержит изложение основных задач и методов первичной статистической обработки экспериментальных данных. Большое внимание уделяется мотивированному введению математических понятий, вычислительных формул и рабочих алгоритмов статистических процедур. Все теоретические рассуждения сопровождаются конкретными простыми примерами и задачами. Большое количество упражнений для самостоятельной работы поможет приобрести практические исследовательские навыки.

Пособие предназначается студентам, аспирантам, специалистам, впервые изучающим математическую статистику после предварительного курса теории вероятностей.

1 Первичная обработка статистических данных

1.1 Генеральная совокупность и случайная выборка

Возможность использования методов математической статистики для решения практических задач технического, экономического, социологического, медицинского, политического и любого другого содержания обусловлена тем, что математические результаты доказываются и формулируются в самом общем виде. Для понимания и применения результатов любой математической теории необходимо изучение языка, или терминологии. В теории математической статистики исследуются отношения не между конкретными предметами действительности, а между абстрактными математическими объектами, или элементами. Рассмотрим описание базисных понятий и допущений математической статистики, которые необходимы для дальнейшего изложения материала.

Определение 1.1 Генеральной совокупностью называется множество, состоящее из всех однородных элементов, которые подлежат исследованию относительно определенного свойства.

Для каждой конкретной задачи определяется соответствующая генеральная совокупность и перечисляются все ее возможные элементы. Объединение элементов в генеральную совокупность происходит в соответствии с целью исследования. Например, при исследовании продолжительности горения электрических ламп генеральную совокупность составляют все электрические лампы; при исследовании влияния повышенных доз радиации на здоровье человека все граждане, подвергнувшиеся радиоактивному облучению, составляют генеральную совокупность. В каждом конкретном случае элементами генеральной совокупности могут быть предметы, люди, животные, города, числа и другие, реально существующие объекты. Но могут быть и воображаемые, или мысленно возможные генеральные совокупности, например, такую совокупность образуют предполагаемые результаты измерений какой-либо физической величины или гипотетические промышленные изделия, которые будут выпущены в будущем.

Генеральные совокупности могут быть конечными или бесконечными. Бесконечную генеральную совокупность можно рассматривать тогда, когда число элементов, подлежащих определенному исследованию, неограниченно возрастает.

Определение 1.2 Статистическим экспериментом называется планируемая осуществляемая процедура для исследования определенного свойства генеральной совокупности.

Статистический эксперимент часто называют другими словами: опыт, наблюдение или просто эксперимент. Предполагается, что эксперимент может многократно повторяться при относительно одинаковых условиях. Самым лучшим способом исследования генеральной совокупности является тот, при котором тщательно рассматривается каждый элемент генеральной совокупности.

Определение 1.3 Исследование каждого элемента генеральной совокупности называется переписью.

Наиболее известной формой переписи является перепись населения. Обследование каждого элемента генеральной совокупности даёт возможность получить самую полную информацию для её анализа. Однако условия для проведения переписи генеральной совокупности осуществляются достаточно редко по ряду причин. В общем случае это невозможно тогда, когда ее элементы существуют только гипотетически либо претерпевают бесконечные изменения в процессе исследования. Кроме того, для проведения переписи может потребоваться слишком много времени или цена переписи может оказаться очень высокой. Чаще всего единственным способом исследования генеральной совокупности является так называемый выборочный метод, когда обследованию подвергается лишь часть элементов совокупности.

Определение 1.4 Выборкой называется любое подмножество генеральной совокупности.

Идея выборочного метода состоит в том, чтобы выявлять определенные закономерности генеральной совокупности на основе исследования свойств её выборки. Преимущество выборочного метода состоит в том, что при сокращении времени и уменьшении

затрат он позволяет получить значимую информацию тогда, когда полное обследование генеральной совокупности невозможно.

Любая выборка представляет собой как бы уменьшенную, хотя и не совсем точную копию генеральной совокупности. Очевидно, что выборка должна наиболее полно отражать существенные свойства всей генеральной совокупности, то есть должна быть **репрезентативной**. В репрезентативной выборке исследуемые свойства генеральной совокупности проявляются в такой же пропорции и с такой же частотой, с которой они представлены во всей совокупности. Именно репрезентативность выборки даёт возможность обобщить её свойства на всю генеральную совокупность.

Главным признаком непрезентативности выборки является нарушение условия объективной случайности выборочного процесса. Например, экзамен является конкретной формой применения выборочного метода. Целью семестрового экзамена является оценивание знаний студентов по определенному учебному предмету. Как правило, на экзамене студент отвечает лишь на несколько вопросов, которые касаются небольшой части изучаемого курса. При соблюдении случайного выбора билета преподаватель достаточно точно может оценить знания студента по этому предмету. Если бы студент сам выбирал только те вопросы, которые он знает, то адекватность оценки его знаний была бы сомнительной.

Определение 1.5 Случайной выборкой из генеральной совокупности называется выборка, образованная таким способом, при котором каждый элемент генеральной совокупности имеет равную вероятность быть выбранным.

Отсюда следует, что все случайные выборки объема n из данной генеральной совокупности имеют одинаковую вероятность быть выбранными. В дальнейшем мы будем рассматривать только случайные выборки, называя их кратко выборками.

Определение 1.6 Число элементов генеральной совокупности или выборки называется объемом генеральной совокупности или выборки, соответственно.

Очевидно, что чем больше объем выборки, тем точнее она отражает свойства генеральной совокупности. Известно, что предельная

ошибка выборки относительно всей совокупности обратно пропорциональна квадрату объёма выборки n. Это означает, что при необходимости удвоения точности выборки её объём должен быть увеличен в $2^2 = 4$ раза. В математической статистике принято считать большим объём выборки, содержащей более 30 элементов ($n > 30$). Методы исследования малых и больших выборок могут существенно отличаться.

1.2 Сбор статистических данных

Сбор статистических данных является одной из первых задач статистики. Объективная статистическая информация может дать точные характеристики исследуемых явлений. При использовании недостоверной статистической информации трудно получить верные выводы и прогнозы.

Получение правильных данных для статистического анализа является трудоемким процессом, состоящим из нескольких последовательных этапов:

– осознание и формулировка конкретной цели наблюдения эксперимента;

– выделение соответствующей генеральной совокупности и определение результирующей случайной величины;

– описание способа образования выборки, подлежащей обследованию, определение необходимого объема выборки и единицы измерения;

– подбор необходимой формы представления выборочных данных.

Разнообразные статистические данные имеют огромное информационное значение, поэтому существует регламентированная система государственных и частных структур, занимающихся сбором статистической информации. Библиотеки таких данных используются для анализа и прогнозирования природных, экономических, общественных и многих других явлений. Отметим, что данные, собранные другими людьми, называются ***вторичными***. К таким данным всегда надо относится с определенной долей недоверия, так как условия, методы и цели собирания таких данных обычно остаются неизвестными. Для получения достоверных

научных результатов преимущественно используются ***первичные*** статистические данные, которые получают для конкретной цели при контролируемых условиях сами исследователи.

Выделяют два основных метода сбора данных: наблюдение и эксперимент. Однако разница между ними довольно условная. Считается, что при проведении эксперимента контролируются определенные условия, а при простом наблюдении такой контроль отсутствует. Существуют определенные требования, предъявляемые к любому методу получения статистических данных. Одним из них является условие многократной повторяемости случайного эксперимента в относительно одинаковых условиях. Единичный эксперимент не может считаться достаточным доказательством правильности статистических выводов. Наиболее существенным является требование соблюдения условия случайности выборочного процесса, так как именно тенденциозно организованная выборка чаще всего оказывается источником ошибочных выводов.

В математической статистике разработан ряд процедур и методов, которые в определенных условиях обеспечивают случайность отбора объектов из генеральной совокупности при формировании случайной выборки.

Самым распространенным способом получения выборочных данных является простой случайный **выбор без возвращения**, когда каждый случайно отобранный объект в исследуемую генеральную совокупность обратно не возвращается. Полученные таким способом выборки называются **бесповторными**. Такой выбор, например, используется для контроля качества однородных промышленных изделий.

При **выборе с возвращением** происходит формирование выборки с повторяющимися элементами. Например, при регистрации числа пациентов, посещающих поликлинику в течение определенного промежутка времени, может получиться **повторная** выборка.

Осуществить простой случайный выбор возможно с помощью обычной жеребьевки. Однако не всегда такой способ является наилучшим.

Рассмотрим один из известных методов случайного выбора элементов генеральной совокупности с использованием ***таблиц случайных чисел***. Таблица случайных чисел состоит из случайных

наборов цифр, размещенных по строкам и столбцам. Каждая из возможных цифр 0, 1, 2, …, 9 выбирается случайным образом. Для удобства чтения цифры в таблице группируются парами (иногда по три, четыре, пять) и записываются по строкам и столбцам. Существуют таблицы с миллионами случайных чисел. Такая таблица есть и в приложении А. При отсутствии готовой таблицы можно воспользоваться генератором случайных чисел, который содержится в каждом компьютере.

Для создания действительно случайной выборки с помощью таблицы случайных чисел необходимо выполнить следующие действия.

1. Каждому элементу генеральной совокупности приписывается определенный номер. Если наибольший номер является, например, трехзначным (n-значным), то и остальные номера записываются трехзначным (n-значным) числом с помощью дополнительных нулей. Таким образом, номер 1 записывается в виде 001, номер 15 – в виде 015.

2. Случайным образом выбирается начальная точка в таблице случайных чисел.

3. От начальной точки по порядку слева направо, как обычно, читаем по одной цифре, если все номера однозначные, по две цифры, если номера двузначные, и так далее до тех пор, пока не получим необходимое число различных номеров, совпадающее с объемом выборки n. Повторные номера вычеркиваются.

Преимущество такого выборочного процесса состоит в том, что исключается влияние определенных личностных пристрастий.

Пример 1.1 В художественной школе занимается 98 учащихся. Для экскурсии в музей надо выбрать 12 учеников. Сделаем справедливый выбор.

Составим список всех детей школы и получим номер каждого от 1 до 98. Выберем с закрытыми глазами начальную точку в таблице случайных чисел и прочитаем последовательно пары 29, 03, 06, 11, 80, 72, 96, 20, 74, 41, 56, 23. Значит для экскурсии отобраны учащиеся с номерами:

3, 6, 11, 20, 23, 29, 41, 56, 72, 74, 80, 96.

■

В определенных ситуациях используется так называемый

стратифицированный выбор элементов из генеральной совокупности. Для этого исходная совокупность делится на непересекающиеся группы, которые называются слоями, или **стратами**, после чего из каждого слоя осуществляется необходимый простой случайный выбор элементов. Такой метод, например, подходит для отбора образцов промышленных изделий разного ассортимента.

Существуют также типовые и комбинированные **многоступеньчатые** модификации случайного выбора. Например, для опроса группы студентов университета сначала выбираются факультеты, затем последовательно выбираются курсы, группы и, наконец, студенты.

В любом случае при формировании выборки необходимо учитывать следующие важные условия:

– все элементы исследуемой генеральной совокупности должны иметь одинаковую вероятность для включения в выборку;

– отбор элементов должен осуществляться из одной и той же генеральной совокупности.

Отметим еще одно требование к выборочным данным, без которого даже абсолютно случайная выборка может дать неправильное решение, – это точность и объективность измерений. Иногда либо из-за большого желания получить нужное заключение, либо из-за профессиональной некомпетентности, либо из-за каких-то других причин исследователи просто подгоняют результаты опытов в нужную им сторону. Естественно, такие «статистические данные» не дают объективных выводов.

Любые отклонения характеристик выборки от соответствующих свойств генеральной совокупности считаются ошибками репрезентативности, которые могут быть как случайными, так и систематическими. Полного исключения ошибок достичь невозможно. Однако в математической статистике разработаны методы, позволяющие устанавливать необходимую точность статистических выводов на основе выборочных данных. Для профессиональных статистических исследований существуют специальные методы фильтрации выборок для выявления ошибок репрезентативности и ошибок регистрации выборочных данных. Выборки, которые обеспечивают отклонение значений основных характеристик не более чем на 5 % от соответствующих значений генеральной совокупности, принято считать удовлетворительными.

В любом случае прежде, чем проводить анализ статистических данных, необходимо убедиться в том, что они соответствуют цели исследования, не тенденциозно подобраны, точны и полны.

1.3 Закон распределения случайной дискретной величины

Для получения исходных статистических данных проводится целенаправленное исследование либо всех элементов генеральной совокупности, либо ее случайной выборки. В большинстве случаев результат любого отдельного эксперимента может выражаться одним или несколькими числовыми значениями. Фактически это значит, что каждому элементу генеральной совокупности по исходу эксперимента ставится в соответствие определенное число, зависящее от степени проявления наблюдаемого свойства при исследовании этого элемента. Следовательно, статистические данные, полученные в результате обследования всех элементов случайной выборки объема *n*, представляются множеством, состоящим из n числовых значений. Такую совокупность чисел можно считать значениями, некоторой переменной, которая количественно характеризует наличие определенного свойства у элементов генеральной совокупности.

Определение 1.7 Случайной величиной называется такая переменная X, которая в результате эксперимента принимает единственное значение для каждого элемента генеральной совокупности.

Таким образом, случайная величина X является числовой функцией, определенной на множестве элементов генеральной совокупности. Значения случайной величины зависят от многих случайных факторов и выясняются только по завершении опыта.

Подчеркнем, что множество значений случайной величины X само является генеральной совокупностью, а любая случайная выборка из генеральной совокупности отождествляется с определенным подмножеством значений случайной величины X, полученных в результате конечной серии наблюдений. Такая замена реальных объектов числовыми значениями дает возможность для

развития и широкого использования аналитических методов математической статистики.

В теории вероятностей и математической статистике принято выделять **одномерные** и **многомерные** случайные величины. При исследовании только одного свойства генеральной совокупности результат единичного эксперимента характеризуется одним числовым значением, и соответствующая случайная величина считается одномерной. Если же результат эксперимента характеризуется несколькими числовыми значениями, то соответствующая случайная величина называется многомерной. Далее мы будем рассматривать только одномерные случайные величины.

Определение 1.8 Случайная величина, имеющая конечное или счетное число значений, называется дискретной.

Полное описание любой случайной дискретной величины включает в себя перечисление не только множества ее возможных значений, но и соответствующих вероятностей этих значений.

Пример 1.2 Два стрелка делают по одному выстрелу в мишень. Вероятность попадания для первого стрелка равна 0,7, а для второго – 0,9. Обозначим через X число попаданий в мишень.

Очевидно, что случайная величина X может принимать три значения: 0, 1, 2. Вычислим вероятность каждого возможного значения:

$P(X = 0) = 0{,}3 \cdot 0{,}1 = 0{,}03$

$P(X = 1) = 0{,}7 \cdot 0{,}1 + 0{,}3 \cdot 0{,}9 = 0{,}34$

$P(X = 2) = 0{,}7 \cdot 0{,}9 = 0{,}63$

Занесем все значения случайной величины X и соответствующие вероятности в следующую таблицу.

Таблица 1.1 – Распределение вероятностей числа попаданий в мишень

X	0	1	2
P	0,03	0,34	0,63

■

Определение 1.9 Совокупность всех возможных значений дискретной случайной величины X и соответствующих им вероятностей называется законом распределения вероятностей случайной величины.

Более кратко **закон распределения вероятностей** называют просто **распределением случайной величины X.** Часто распределение вероятностей задается таблицей, содержащей все возможные значения случайной величины X и соответствующие вероятности этих значений.

Таблица 1.2 – Закон распределения вероятностей случайной дискретной величины X

X	x_1	x_2	…	x_n	…
P	p_1	p_2	…	p_n	…

Очевидно, что сумма вероятностей всех возможных значений дискретной случайной величины X равна 1:

$$p_1 + p_2 + \ldots + p_n + \ldots = 1.$$

Это свойство называется **условием нормированности** распределения.

Однако более удобно закон распределения вероятностей дискретной случайной величины выражать некоторой функцией, позволяющей найти вероятность реализации каждого конкретного значения случайной величины X.

Пример 1.3 Монета бросается семь раз. Обозначим через X число появлений герба. Найдем закон распределения случайной величины X.

Очевидно, что случайная величина X может принимать следующие значения: 0, 1, 2, 3, 4, 5, 6, 7. При этом вероятность того, что значение переменной X равно k вычисляется по формуле Бернулли:

$$P(X = k) = P_7(k) = C_7^k (0{,}5)^k (0{,}5)^{7-k},$$

где $k = 0, 1, 2, 3, 4, 5, 6, 7$. Данная функция полностью определяет распределение вероятностей случайной величины X.

■

В некоторых случаях используется графическое изображение закона распределения вероятностей случайной величины. Для этого на оси абсцисс отмечаются все значения случайной величины x_1, x_2, …, x_n, а на оси ординат откладываются соответствующие вероятности p_1, p_2, …, p_n. Затем точки с координатами (x_1; p_1), (x_2; p_2), …, (x_n; p_n) последовательно соединяются отрезками. Полученная ломаная линия называется ***многоугольником распределения вероятностей***.

Графическое изображение закона распределения позволяет получить визуальное представление об исследуемой зависимости. С помощью графика можно заметить основные тенденции варьирования значений случайной величины. Особенно полезны графики в тех случаях, когда нужно показать постоянство некоторых характеристик закона распределения.

Пример 1.4 Построим многоугольник распределения вероятностей для случайной величины X по данным предыдущего примера 1.3.

Прежде всего, вычислим вероятность каждого возможного значения:

$$p_0 = P(X = 0) = C_7^0 (0,5)^0 (0,5)^7 = 0,0078125$$

$$p_1 = P(X = 1) = C_7^1 (0,5)^1 (0,5)^6 = 0,0546875$$

$$p_2 = P(X = 2) = C_7^2 (0,5)^2 (0,5)^5 = 0,1640625$$

$$p_3 = P(X = 3) = C_7^3 (0,5)^3 (0,5)^4 = 0,2734375$$

$$p_4 = P(X = 4) = C_7^4 (0,5)^4 (0,5)^3 = 0,2734375$$

$$p_5 = P(X = 5) = C_7^5 (0,5)^5 (0,5)^2 = 0,1640625$$

$$p_6 = P(X = 6) = C_7^6 (0,5)^6 (0,5)^1 = 0,0546875$$

$$p_7 = P(X = 7) = C_7^7 (0,5)^7 (0,5)^0 = 0,0078125$$

Проверка показывает, что сумма вероятностей всех значений равна 1.

Округлим значения вероятностей:

$p_0 = 0,008$	$p_1 = 0,055$	$p_2 = 0,164$	$p_3 = 0,273$
$p_4 = 0,273$	$p_5 = 0,164$	$p_6 = 0,055$	$p_7 = 0,008$

Выполним построение многоугольника распределения вероятностей.

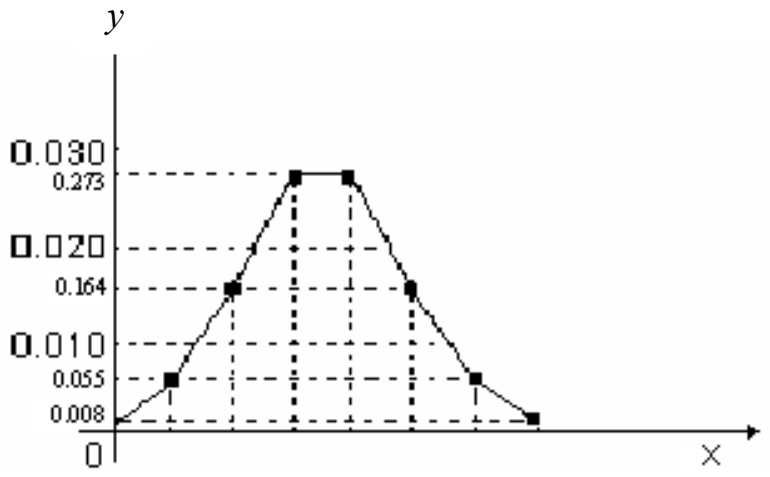


Рисунок 1.1 – Многоугольник распределения вероятностей для числа появления герба при семикратном бросании монеты

■

1.4 Функция распределения случайной величины

Все рассмотренные способы задания закона распределения случайной величины являются неприемлемыми тогда, когда случайная величина имеет слишком много значений, которые невозможно перечислить. Для исследования закона распределения вероятностей произвольной случайной величины X можно использовать не вероятность события $X = x$, а вероятность события $X < x$, где x – некоторое действительное число. Вероятность того, что случайная величина X в результате опыта примет значение, которое будет меньше числа x, является функцией аргумента x.

Определение 1.10 Функцией распределения случайной величины X называется функция $F_x(x)$ действительной переменной $x \in (-\infty; \infty)$, равная вероятности того, что X принимает значения, меньшие числа x.

Таким образом, функция распределения случайной величины X определяется следующим соотношением:

$$F_x(x) = P(X < x), \quad x \in (-\infty; \infty).$$

В дальнейшем вероятностную функцию распределения $F_x(x)$ мы

будем называть теоретической. Иногда вместо термина «функция распределения» используется равнозначный термин «интегральная функция распределения».

Пример 1.5 Обозначим через X число нечетных цифр в произвольном четырехзначном номере. Найдем функцию распределения случайной величины X и построим ее график.

Очевидно, что случайная величина X может принимать следующие значения: 0, 1, 2, 3, 4.

Вероятность выбора одной нечетной цифры равна 0,5, вероятность выбора четной цифры также равна 0,5. Вычислим вероятности соответствующих значений.

$p_0 = P(X = 0) = C_4^0 (0,5)^0 (0,5)^4 = 0,0625$

$p_1 = P(X = 1) = C_4^1 (0,5)^1 (0,5)^3 = 0,2500$

$p_2 = P(X = 2) = C_4^2 (0,5)^2 (0,5)^2 = 0,3750$

$p_3 = P(X = 3) = C_4^3 (0,5)^3 (0,5)^1 = 0,2500$

$p_4 = P(X= 4) = C_4^4 (0,5)^4 (0,5)^0 = 0,0625$

Суммирование вероятностей подтверждает условие нормированности распределения.

Пусть $x \in (-\infty; 0]$, тогда

$F_x(x) = P(X < x) = P(X < 0) = 0.$

При $x \in (0; 1]$ имеем

$F_x(x) = P(X < x) = P(X < 1) = P(X = 0) = 0,0625.$

При $x \in (1; 2]$ имеем

$F_x(x) = P(X < x) = P(X < 2) = P(X = 0) + P(X = 1) =$

$= 0,0625 + 0,2500 = 0,3125.$

При $x \in (2; 3]$ имеем

$F_x(x) = P(X < x) = P(X < 3) = P(X = 0) + P(X = 1) + P(X = 2) = 0,0687.$

При $x \in (3; 4]$ имеем

$F_x(x) = P(X < x) = P(X < 4) =$

$= P(X = 0) + P(X = 1) + P(X = 2) + P(X =3) = 0{,}9375.$

Наконец, при $x \in (4; \infty)$ имеем

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

$= P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3) + P(X = 4) = 1.$

В итоге получается следующее выражение функции распределения:

$$F_x(x) = \begin{cases} 0, & x \in (-\infty;0]; \\ 0{,}0625, & x \in (0;1]; \\ 0{,}3125, & x \in (1;2]; \\ 0{,}6875, & x \in (2;3]; \\ 0{,}9375, & x \in (3;04]; \\ 1, & x \in (4;\infty). \end{cases}$$

Построим график этой функции.

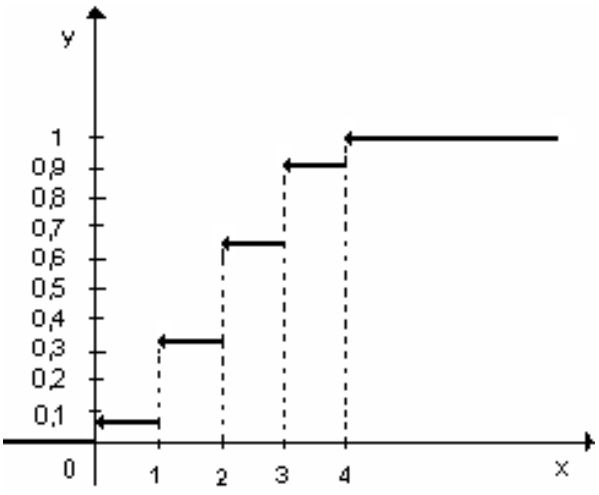


Рисунок 1.2 – График функции распределения числа четных цифр в произвольном четырехзначном номере

■

В общем случае функция распределения любой дискретной случайной величины X находится по формуле:

$$F_x(x) = \sum_{x_i < x} P(X = x_i),$$

то есть суммируются вероятности всех значений случайной

величины, которые являются меньшими числа x.

Напомним основные свойства функции распределения.

1. Функция распределения является неубывающей функцией, то есть

$$F_x(x_1) \le F_x(x_2) \text{ при } x_1 < x_2.$$

2. Справедливы следующие равенства:

$$\lim_{x \to -\infty} F_x(x) = 0 \quad \text{и} \quad \lim_{x \to +\infty} F_x(x) = 1.$$

3. Все значения функции распределения изменяются от 0 до 1, то есть

$$\mathbf{0} \le F_x(x) \le 1.$$

4. Функция распределения является непрерывной слева, то есть

$$\lim_{x \to x_0 - 0} F_x(x) = F_x(x_0).$$

5. Вероятность неравенства $a \le X < b$ для любых a и b вычисляется по формуле

$$P(a \le X < b) = F_x(b) - F_x(a).$$

Подчеркнем, что функция распределения является наиболее общей формой закона распределения и самой полной характеристикой случайной величины любого типа.

1.5 Плотность распределения вероятностей

Результаты не всех реальных экспериментов можно исчерпывающе описать с помощью дискретных случайных величин. Например, основные метеорологические параметры погоды и физиологические показатели состояния здоровья человека такие, как температура, давление, вес, размеры, промежутки времени, концентрация химических веществ в жидкости изменяются

непрерывно от наименьшего возможного значения до наибольшего. Если множество возможных значений случайной величины заполняют некоторый конечный или бесконечный числовой интервал, то такая случайная величина называется **непрерывной**.

Подчеркнем, что непрерывная случайная величина в отличие от дискретной имеет бесконечное множество значений, которые невозможно перечислить. Из определения непрерывности вытекает и другая ее особенность.

Вероятность любого конкретного значения *x* непрерывной случайной величины X равна нулю, то есть

$$P(X = x) = 0.$$

Заметим, что из равенства нулю вероятности конкретного значения x не следует невозможность принятия случайной величины X этого значения. Так как непрерывная случайная величина имеет бесконечное множество значений, то каждое из них реализуется крайне редко.

Поэтому для непрерывных величин определяется вероятность попадания значений в бесконечно малый интервал Δx. Эта вероятность вычисляется по формуле:

$$P(x \le X < x + \Delta x) = F_x(x + \Delta x) - F_x(x).$$

Долю вероятности, соответствующую единице длины интервала Δx, показывает соотношение:

$$\frac{P(x \le X < x + \Delta x)}{\Delta x} = \frac{F_x(x + \Delta x) - F_x(x)}{\Delta x}.$$

Очевидно, что предел этого соотношения при $\Delta x \to 0$ является производной функции распределения:

$$\lim_{\Delta x \to 0} \frac{P(x \le X < x + \Delta x)}{\Delta x} = \lim_{\Delta x \to 0} \frac{F_x(x + \Delta x) - F_x(x)}{\Delta x} = F'_x(x).$$

Определение 1.11 Плотностью распределения вероятностей $p_x(x)$ непрерывной случайной величины X называется первая производная (если она существует) ее функции распределения:

$$\mathbf{p_x(x) = F'_x(x).}$$

Плотность распределения вероятностей $p_x(x)$ сокращенно называют **плотностью вероятности** случайной величины X. Иногда вместо термина «плотность вероятности» используется эквивалентный термин «дифференциальная функция распределения». Случайная величина имеет функцию плотности вероятности только тогда, когда ее функция распределения $F_x(x)$ непрерывна и дифференцируема всюду, за исключением отдельных точек на конечном промежутке. Функция плотности вероятности $p_x(x)$ дает возможность судить о характере распределения непрерывной случайной величины в малых окрестностях точек числовой оси.

Из предшествующих рассуждений следует, что плотность вероятности $p_x(x)$ пропорциональна вероятности того, что случайная величина примет значение, находящееся на бесконечно малом расстоянии $\Delta x \to 0$ от точки x. Отсюда также вытекает следующее приближенное равенство:

$$P(x \le X < x + \Delta x) \approx p_x(x)\,\Delta x.$$

Произведение $p_x(x)\Delta x$ называют ***элементом вероятности***, оно равно площади прямоугольника с основанием Δx и высотой $p_x(x)$. Значит, вероятность попадания значений случайной величины X в полузамкнутый интервал $[x; x + \Delta x)$ приближенно равна площади прямоугольника с основанием Δx и высотой, равной $p_x(x)$.

Рассмотрим геометрическую интерпретацию данного утверждения.

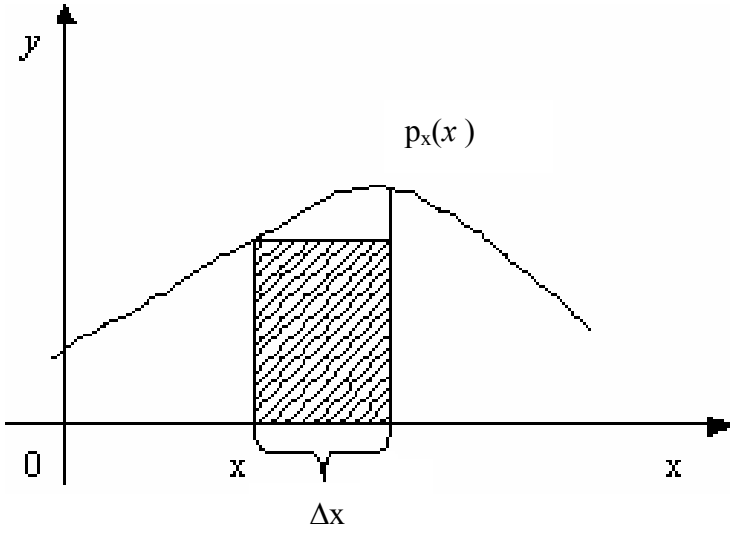


Рисунок 1.3 – Вероятность попадания значений случайной величины X в интервал [x; x + Δx)

Площадь заштрихованного прямоугольника приближенно равна вероятности того, что случайная величина X примет значение, принадлежащее интервалу $[x; x + \Delta x)$.

Рассмотрим основные свойства плотности вероятности.

1. Плотность вероятности $p_x(x)$ всегда неотрицательна:

$$p_x(x) \geq 0.$$

2. Вероятность того, что непрерывная случайная величина X примет значения из интервала $[x_1; x_2)$ вычисляется по формуле:

$$P(\tilde{o}_1 \leq \tilde{O} < \tilde{o}_2) = \int_{\tilde{o}_1}^{\tilde{o}_2} \eth_{\eth}(\tilde{o}) d\tilde{o}.$$

3. Для плотности вероятности всегда справедливо равенство:

$$\int_{-\infty}^{\infty} p_x(x)\, dx = 1.$$

4. Функция распределения $F_x(x)$ и плотность вероятности $p_x(x)$ непрерывной случайной величины X связаны формулой:

$$F_x(x) = \int_{-\infty}^{x} p_x(t) dt.$$

Рассмотрим геометрическую интерпретацию основных свойств плотности вероятности.

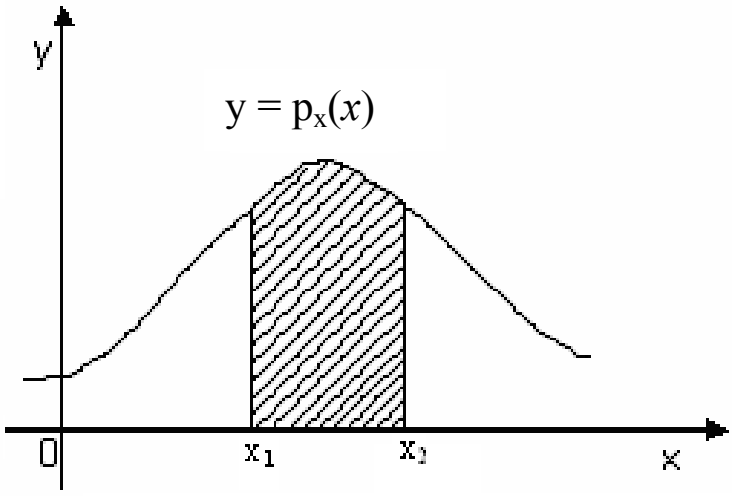


Рисунок 1.4 – Вероятность попадания значений случайной величины X в интервал $[x_1; x_2)$

Данный рисунок показывает, что в соответствии со вторым свойством площадь заштрихованной криволинейной трапеции, заключенной между кривой плотности вероятности $y = p_x(x)$ и осью O_x, равна вероятности неравенства $P(x_1 \le X < x_2)$.

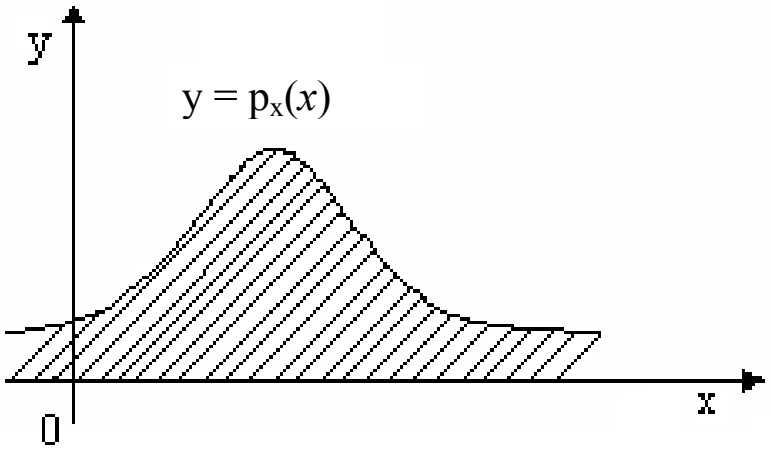


Рисунок 1.5 – Полная площадь
под кривой плотности вероятности

Так как по третьему свойству плотности

$$P(-\infty < \tilde{O} < \infty) = \int_{-\infty}^{\infty} p_x(x)dx = 1,$$

то площадь заштрихованной фигуры, ограниченной графиком плотности вероятности $y = p_x(x)$ и осью O_x, всегда равна 1, что показывает рисунок 1.5.

Так как функция распределения $F_x(x)$ равна вероятности того, что случайная величина X примет значения, меньшие числа x, то по свойству 4 имеем:

$$F_x(x) = P(X < x) = P(-\infty < X < x) = \int_{-\infty}^{x} p_x(t)dt.$$

Следовательно, геометрически функция распределения $F_x(x)$ равна площади заштрихованной фигуры на рисунке 1.6, расположенной левее точки x, ограниченной графиком плотности вероятности $y = p_x(x)$ и осью O_x.

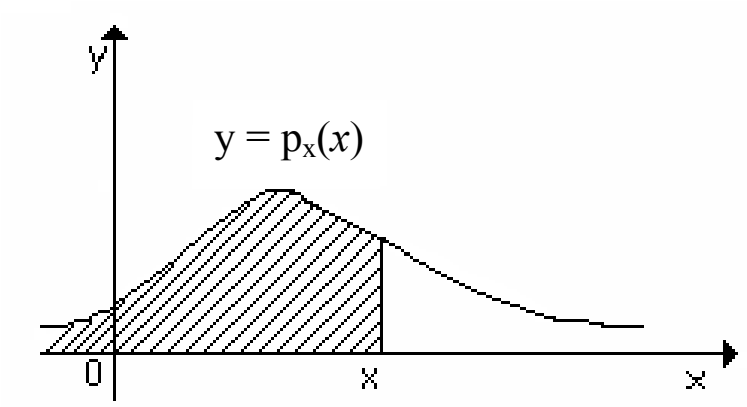


Рисунок 1.6 – Геометрическая интерпретация
функции распределения $F_x(x)$

Подчеркнем, что функция распределения и плотность вероятности случайной величины являются фундаментальными понятиями как теории вероятности, так и математической статистики. Заданные в аналитической форме, они дают существенную информацию об исследуемой случайной величине и позволяют теоретически обосновать статистические выводы, сделанные на основе эмпирических данных.

1.6 Группировка статистических данных

Полученные в результате экспериментов или наблюдений первичные статистические данные, как правило, записываются в рабочую таблицу наблюдений. К сожалению, на основе неорганизованного скопления числовых значений сложно сделать какие-либо статистические выводы. Прежде всего необходимо представить результаты экспериментов в рабочем виде. Существуют определенные способы группировки статистических данных в специальные таблицы.

Допустим, что в результате проведения n экспериментов получена некоторая выборка значений случайной величины X. Расположим данные выборочные значения в порядке их возрастания, при этом некоторые из них могут повторяться несколько раз.

Определение 1.12 Все различные значения случайной величины, содержащиеся в выборке, называются вариантами.

Определение 1.13 Число m_i, показывающее, сколько раз варианта x_i встречается в выборке, называется частотой варианты x_i.

Очевидно, что если $x_1, \; x_2 \; , \cdots , \; x_k$ – все варианты выборки, то сумма их соответствующих частот равна объему всей выборки:

$$m_1 + m_2 + \cdots + m_k = n .$$

Определение 1.14 Если объем всей выборки равен n, то относительной частотой варианты x_i называется число p_i^*, равное отношению частоты m_i к объему n:

$$p_i^* = \frac{m_i}{n}.$$

Каждая варианта выборки $x_1, \; x_2 , \; \ldots , \; x_k$ имеет соответствующую относительную частоту:

$$p_1^* = \frac{m_1}{n}, \quad p_2^* = \frac{m_2}{n}, \; \cdots , \; p_k^* = \frac{m_k}{n} .$$

Не трудно проверить, что сумма относительных частот всех вариант выборки равна 1:

$$p_1^* + p_2^* + \cdots + p_k^* = 1 .$$

Определение 1.15 Множество всех вариант выборки, расположенных в порядке возрастания их значений, вместе с их соответствующими частотами или относительными частотами называется вариационным рядом.

Вариационный ряд удобно представлять в виде следующей таблицы.

Таблица 1.3 – Вариационный ряд

x_i	x_1	x_2	…	x_k
m_i	m_1	m_2	…	m_k
p_i^*	p_i^*	p_2^*	…	p_k^*

Пример 1.6 В течение недели было проведено исследование посещаемости университетской библиотеки студентами группы, состоящей из 25 человек. Зафиксированное число посещений каждого студента представляет следующая выборка:

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

Объем выборки n = 25. Выпишем все варианты в порядке их возрастания:

0, 1, 2, 3, 4, 5.

Найдем частоту каждой варианты:

$$m_0 = 5; \quad m_1 = 6; \quad m_2 = 7; \quad m_3 = 2; \quad m_4 = 3; \quad m_5 = 2.$$

Для проверки найдем сумму всех частот:

$$5 + 6 + 7 + 2 + 3 + 2 = 25.$$

Теперь вычислим относительные частоты соответствующих вариант:

$$p_0^* = \frac{5}{25} = 0{,}20; \quad p_1^* = \frac{6}{25} = 0{,}24; \quad p_2^* = \frac{7}{25} = 0{,}28;$$

$$p_3^* = \frac{2}{25} = 0{,}08; \quad p_4^* = \frac{3}{25} = 0{,}12; \quad p_5^* = \frac{2}{25} = 0{,}08.$$

Проверим сумму:

$$0{,}20 + 0{,}24 + 0{,}28 + 0{,}08 + 0{,}12 + 0{,}08 = 1.$$

Составим вариационный ряд данной выборки.

Таблица 1.4 – Вариационный ряд данных посещаемости библиотеки

Число посещений	0	1	2	3	4	5
m_i	5	6	7	2	3	2
p_i^*	0.20	0.24	0.28	0.08	0.12	0.08

■

Вариационный ряд часто помогает сгруппировать и более организованно записать результаты статистических экспериментов. Однако преимущества вариационного ряда теряются в тех случаях, когда выборка имеет большой объем и не содержит повторяющихся значений. Для выборочных данных с большим объемом существует более общая форма представления.

Рассмотрим произвольную выборку значений случайной величины X объема n. Обозначим через a наименьшее выборочное значение, а через b – наибольшее. Тогда вся выборка принадлежит отрезку [a; b]. Разделим этот отрезок точками на k меньших интервалов равной длины:

$$a = x_0 < x_1 < x_2 < \ldots < x_k = b.$$

Для каждого *i*-го интервала $[x_{i-1};\ x_i)$, $i = 1,2,\ldots, k$, найдём частоту m_i, равную числу выборочных значений, принадлежащих этому интервалу. Частное $\dfrac{m_i}{n}$ является относительной частотой выборочных значений, принадлежащих *i*-му интервалу. Используем стандартное обозначение: $p_i^* = \dfrac{m_i}{n}$.

Очевидно, что сумма частот всех интервалов равна объему выборки:

$$m_1 + m_2 + \cdots + m_k = n,$$

а сумма всех относительных частот равна 1:

$$p_1^* + p_2^* + ... + p_k^* = 1.$$

Определение 1.16 Таблица интервалов, содержащая данную выборку значений случайной величины X и соответствующие частоты или относительные частоты, называется статистическим рядом.

В общем случае статистический ряд представляется следующим образом.

Таблица 1.5 – Интервальный статистический ряд

Интервалы значений X	$[x_0; x_1)$	$[x_1; x_2)$	…	$[x_{k-1}; x_k]$
m_i	m_1	m_2	…	m_k
p_i^*	$\frac{m_1}{n}$	$\frac{m_2}{n}$	…	$\frac{m_k}{n}$

Заметим, что вариационный ряд является частным случаем статистического ряда.

Пример 1.7 Администрацией поликлиники были собраны данные о возрасте 250 наиболее нуждающихся в лечении пациентов. Результаты этого исследования представлены следующим статистическим рядом.

Таблица 1.6 – Данные исследования возраста пациентов поликлиники

Возраст в годах	10–20	20–30	30–40	40–50	50–60	60–70	70–80	80–90
m_i	17	24	35	48	57	42	21	6
p_i^*	0,068	0,096	0,140	0,192	0,228	0,168	0,084	0,024

Следующие суммы находятся для проверки вычислений:

$$n = \sum_{i=1}^{8} m_i = 17 + 24 + 35 + 48 + 57 + 42 + 21 + 6 = 250,$$

$$\sum_{i=1}^{8} p_i^* = 0{,}068 + 0{,}096 + 0{,}140 + 0{,}192 + 0{,}228 + 0{,}168 + 0{,}084 + 0{,}024 = 1.$$

■

Рассмотрим общую последовательность действий при построении статистического ряда.

Алгоритм построения статистического ряда случайной выборки

1. Заранее определяется число интервалов k.

Число интервалов можно выбирать произвольно, оно не должно быть слишком большим и слишком малым. Можно воспользоваться известной оценочной формулой:

$$k \approx 1 + 3{,}2 \lg n,$$

где правая часть равенства округляется до ближайшего целого.

2. Определяется длина интервала.

Для этого вычисляется число

$$\ell \approx \frac{\text{íàèáîëüøåå çíà÷åíèå âûáîðêè} - \text{íàèìåíüøåå çíà÷åíèå âûáîðêè}}{k}$$

при этом значение ℓ округляется так, чтобы получилось простое и удобное число.

3. Определяется начало и конец всего интервала наблюдений, содержащего всю данную выборку.

Число a, обозначающее нижнюю границу, должно быть чуть меньше, чем наименьшее выборочное значение. Число b, обозначающее верхнюю границу, находится по формуле:

$$b = \ell \cdot k,$$

где ℓ – длина интервала, k – число всех интервалов.

4. Находится частота каждого интервала, равная числу выборочных значений, принадлежащих этому интервалу.

При этом пограничные значения приписываются только одному интервалу. В результате получится k значений:

$$m_1, m_2, \ldots, m_k, \text{ причем } \sum_{i=1}^{k} m_i^* = n.$$

5. Вычисляются соответствующие относительные частоты:

$$p_i^* = \frac{m_1}{n}, \; p_2^* = \frac{m_2}{n}, \ldots, \; p_k^* = \frac{m_k}{n}.$$

Для проверки находится их сумма: $\sum_{i=1}^{k} p_i^* = 1$.

6. Полученные результаты заносятся в таблицу представляющую статистический ряд.

Пример 1.8 Составим статистический ряд по данным измерений высот 40 зданий города.

7,2	15,2	20,8	24,6	27,5	30,6	34,7	38,5
9,6	16,3	22,4	25,1	28,3	31,8	35,6	42,3
10,5	17,2	23,5	25,7	28,8	32,3	35,8	43,4
12,8	18,4	23,7	26,4	29,4	33,2	36,2	44,5
14,2	18,8	24,5	27,2	30,4	34,5	37,4	48,5

Будем строить статистический ряд для девяти интервалов: k = 9, n = 40.

Наименьшее выборочное значение равно 7,2 , наибольшее – 48,5. Определим длину каждого интервала. Найдем

$$\ell = \frac{48{,}5 - 7{,}2}{9} = 4{,}89.$$

Округлив полученное число до ближайшего целого, будем считать, что длина интервала $\ell = 5$.

В качестве нижней границы всех наблюдаемых значений выберем число 5. Тогда имеем следующие интервалы:

[5;10), [10;15), [15;20), [20;25), [25;30), [30;35), [35;40), [40;45), [45;50].

Теперь считаем частоту выборочных значений для каждого интервала:

$$m_1 = 2;\ m_2 = 3;\ m_3 = 5;\ m_4 = 6;\ m_5 = 8;\ m_6 = 7;\ m_7 = 5;$$
$$m_8 = 3;\ m_9 = 1.$$

Проверка подтверждает правильность равенства

$$\sum_{i=1}^{9} m_i = 40.$$

Далее вычисляем относительные частоты:

$$p_1^* = \frac{2}{40} = 0{,}05;\ \ p_2^* = 0{,}075;\ p_3^* = 0{,}125;\ p_4^* = 0{,}150;$$

$$p_5^* = 0{,}200;\ p_6^* = 0{,}175;\ p_7^* = 0{,}125;\ p_8^* = 0{,}075;\ p_9^* = 0{,}025.$$

Суммирование относительных частот показывает, что равенство $\sum_{i=1}^{9} p_i = 1$ выполняется. Запишем статистический ряд.

Таблица 1.7 – Статистический ряд измерений высоты зданий

Высота зданий	5–10	10–15	15–20	20–25	25–30	30–35	35–40	40–45	45–50
m_i	2	3	5	6	8	7	5	3	1
p_i^*	0,050	0,075	0,125	0,150	0,200	0,175	0,125	0,075	0,025

■

Вместо короткого названия **статистический ряд** часто используются более точные по смыслу такие названия, как **статистическое распределение выборки**, или **статистический закон распределения**.

Напомним, что в теории вероятностей **законом распределения дискретной случайной величины** называется соответствие между всеми её возможными значениями и их вероятностями. Статистическое распределение отличается от теоретического вероятностного распределения тем, что вероятность отдельных значений в математической статистике заменяется их относительной частотой.

1.7 Графическое представление статистических данных

Существуют разные графические способы представления статистических распределений выборок или статистических рядов. Любое наглядное изображение статистических данных способствует лучшему пониманию их общих закономерностей.

Вначале рассмотрим традиционный способ графического изображения вариационного ряда. Для этого возьмем прямоугольную систему координат. По определению вариационный ряд состоит из пар чисел вида:

$$(x_1; m_1),\ (x_2; m_2),\ \ldots,\ (x_k; m_k),$$

где значения x_i являются вариантами выборки, а m_i являются соответствующими частотами, i = 1, 2, ... , k. В данной системе координат отмечаем k точек с соответствующими координатами:

$$(x_1; m_1),\ (x_2; m_2),\ \ldots,\ (x_k; m_k),$$

и соединяем их последовательно отрезками прямых.

Определение 1.17 Полигоном частот вариационного ряда называется ломаная линия, состоящая из отрезков прямых, последовательно соединяющих точки с координатами (x_i; m_i) , где x_i пробегает все варианты выборки, а m_i – их соответствующие частоты, i = 1,2, …,k **.**

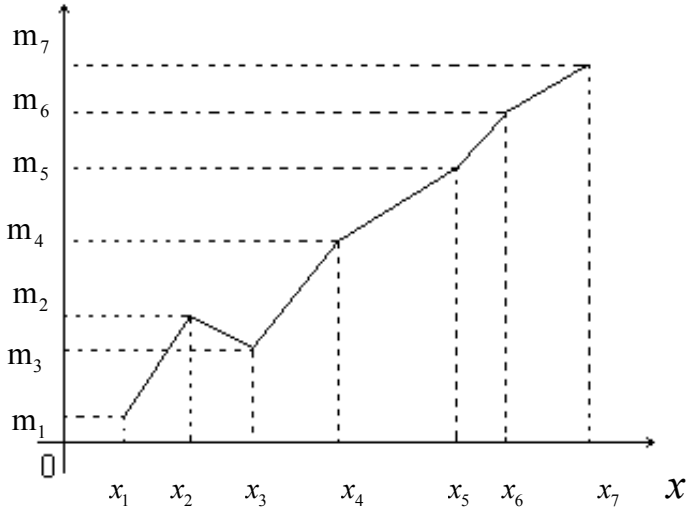


Рисунок 1.7 – Полигон частот вариационного ряда

Пример 1.9 Построим полигон частот для вариационного ряда из примера 1.6, представляющего результаты исследования посещаемости библиотеки студентами.

Для этого мы отмечаем точки (0;5), (1;6), (2;7), (3;2), (4;3), (5;2) и последовательно соединяем их.

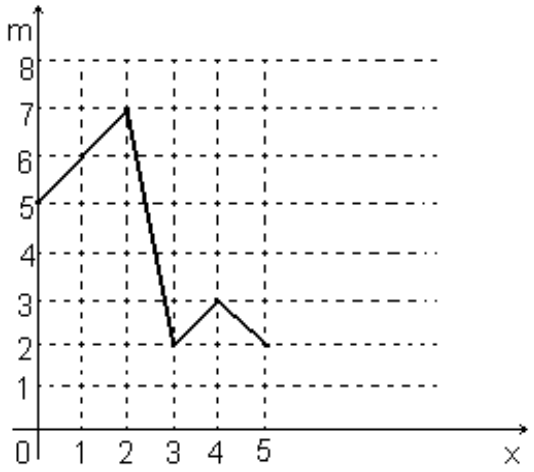


Рисунок 1.8 – Полигон частот данных посещаемости библиотеки студентами

■

Если частоты вариант заменить соответствующими относительными частотами, то получится полигон относительных частот.

Определение 1.18 Полигоном относительных частот вариационного ряда называется ломаная линия, состоящая из отрезков прямых, последовательно соединяющих точки с координатами $(x_i; p_i^*)$, **где** x_i **пробегает все варианты выборки, а** p_i^* **– их соответствующие относительные частоты**, **i = 1, 2, … , k .**

Пример 1.10 Построим полигон относительных частот для вариационного ряда примера 1.6.

Для этого в системе координат отмечаем точки (0; 0,20), (1; 0,24), (2; 0,28), (3; 0,08), (4; 0,12), (5; 0,08) и соединяем их последовательно.

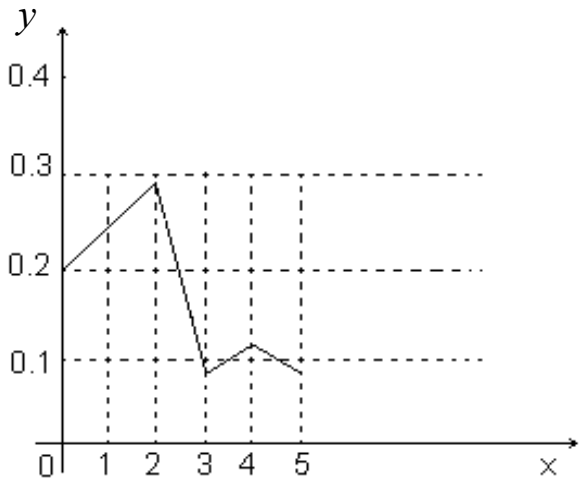


Рисунок 1.9 – Полигон относительных частот данных посещаемости библиотеки студентами

■

Интервальный статистический ряд удобно представлять графически в виде плоской ступенчатой диаграммы, которая называется **гистограммой**.

Гистограмма частот статистического ряда строится следующим образом. В прямоугольной системе координат на оси абсцисс отмечаются интервалы данного статистического ряда. Затем на каждом интервале, как на основании, рисуется прямоугольник, высота которого равна частоте этого интервала.

Пример 1.11 Построим гистограмму частот статистического ряда из примера 1.7 о возрасте пациентов поликлиники.

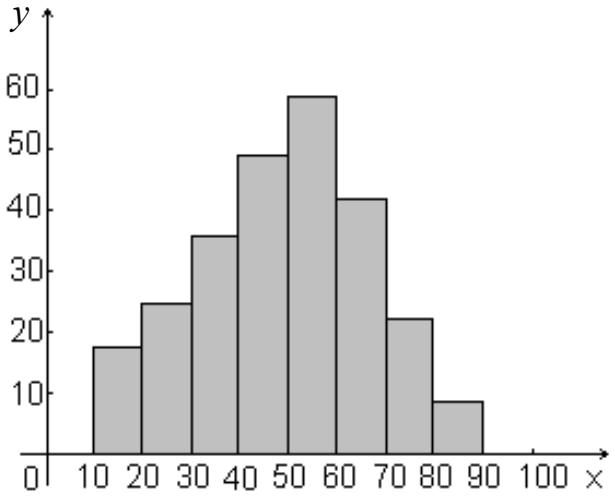


Рисунок 1.10 – Гистограмма частот данных о возрасте пациентов поликлиники

■

Иногда более полезным может оказаться другой вид гистограммы, которая называется **гистограммой относительных частот** статистического ряда. Чтобы её получить, на каждом интервале статистического ряда строится прямоугольник, **площадь** которого равна относительной частоте этого интервала. Из формулы площади прямоугольника легко получается выражение для его высоты:

$$h_i = \frac{p_i^*}{\ell_i}\ , \ i = 1, 2, \ldots, k\ ,$$

где p_i^* – относительная частота i-го интервала, а $\ell_i = x_i - x_{i-1}$ – это длина i-го интервала.

При изображении гистограммы удобно выбирать такой масштаб, в котором длина интервалов ℓ равна 1. Тогда высоты прямоугольников равны соответствующим относительным частотам: $h_i = p_i^*$, $i = 1,2, \ldots, k$. Это упрощает построение гистограммы. В дальнейшем будем считать, что длина каждого интервала равна 1.

Пример 1.12 Построим гистограмму относительных частот статистического ряда из примера 1.7.

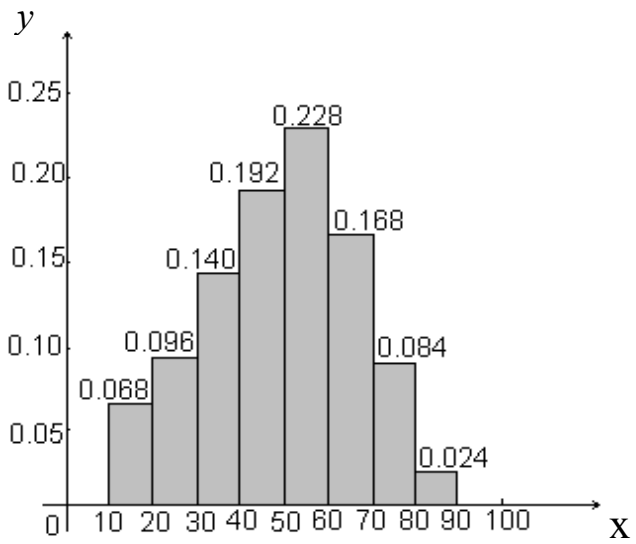


Рисунок 1.11 – Гистограмма относительных частот данных о возрасте пациентов поликлиники

■

Подчеркнем, что форма гистограммы существенно определяется

статистическим рядом. Но любой статистический ряд является обобщением определенной выборки значений некоторой случайной величины. Следовательно, форма гистограммы определенным образом характеризует не только статистическое распределение выборки, но и распределение исследуемой случайной величины. Напомним, что в теории вероятностей закон распределения непрерывной случайной величины X определяется формулой плотности распределения вероятностей, обозначаемой через $\mathrm{p}_x(\boldsymbol{x})$.

Отметим важное свойство гистограммы.

Высота прямоугольника каждого интервала гистограммы относительных частот является приближенным значением плотности распределения вероятностей исследуемой случайной величины на этом интервале:

$$\mathrm{h}_i \approx \mathrm{p}_\mathrm{x}(\boldsymbol{x}_i),\ \ \mathrm{i} = 1, 2, ..., \mathrm{k}.$$

Таким образом, верхнюю ступенчатую границу гистограммы можно считать статистическим аналогом кривой плотности распределения непрерывной случайной величины. Гистограмма на каждом интервале показывает средние плотности распределения вероятностей. Если середины верхних оснований прямоугольников гистограммы соединить прямолинейными отрезками, то получится полигон распределения частот. Очевидно, что при увеличении числа испытаний и при уменьшении длины интервалов ступенчатая ломаная, ограничивающая гистограмму сверху, будет более соответствовать кривой плотности распределения. Итак, если мы сгладим верхнюю ступенчатую ломаную гистограммы плавной кривой, то получим наглядное представление о кривой плотности распределения вероятностей исследуемой случайной величины. Таким образом гистограмма является удобным способом представления сгруппированных данных.

Пример 1.13 Построим гистограмму относительных частот статистического ряда по данным примера 1.8 о высотах городских зданий и покажем приближенный график плотности распределения вероятностей.

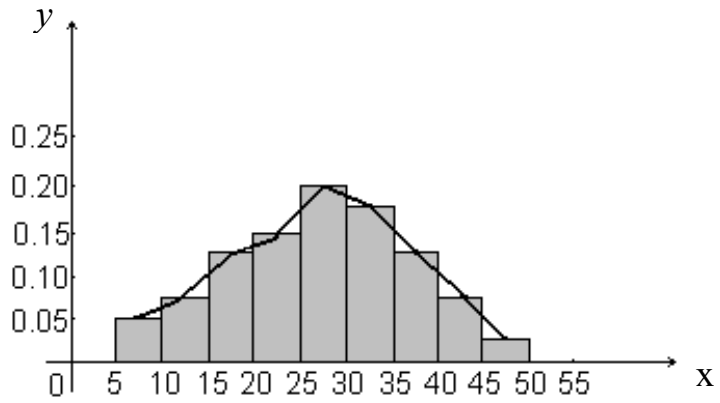


Рисунок 1.12 – Гистограмма относительных частот данных высот зданий города

■

Сформулируем ещё одно свойство гистограммы, тесно связанное с предыдущим.

Площадь гистограммы относительных частот, построенной на интервалах единичной длины, равна 1.

Итак, гистограмма и полигон частот дают возможность реально увидеть в общих чертах закономерности распределения исследуемой генеральной совокупности. Заметим, что полигон частот имеет большее преимущество при сравнении двух разных распределений. В любом случае на первоначальном этапе графические формы представления статистических данных являются важным рабочим методом статистики.

1.8 Эмпирическая функция распределения

В теории вероятностей рассматриваются и изучаются вероятностные законы распределения случайных величин. Как правило, любое вероятностное распределение случайной величины задается в аналитическом виде каким-либо математическим выражением. В математической статистике рассматриваются статистические распределения. Ранее мы говорили о близости понятия статистического ряда с понятием статистического распределения выборки. На практике часто возникает задача нахождения аналитического

выражения, которое, хотя бы приближенно, представляло бы неизвестную теоретическую функцию распределения исследуемой случайной величины. В математической статистике рассматривается аналог этой функции.

Определение 1.19 Если x_1, x_2, …x_n – выборка значений случайной величины X, то эмпирической функцией распределения называется функция действительного аргумента $x \in (-\infty; \infty)$, обозначаемая через $F_x^*(x)$, равная относительной частоте выборочных значений, меньших числа x.

Таким образом,

$$F_x^*(x) = \frac{n_x}{n},$$

где n – это объем выборки значений случайной величины X, а n_x – это количество выборочных значений, удовлетворяющих неравенству $X < x$, $x \in (-\infty; \infty)$.

Так как относительная частота значений случайной величины X, удовлетворяющих неравенству $X < x$, в выборке объема n стремится к вероятности выполнения этого неравенства, то при $n \to \infty$ имеем, что

$$F_x^*(x) = \frac{n_x}{n} \to P(X < x) = F_x(x).$$

Таким образом эмпирическая функция распределения стремится к теоретической функции распределения. Чем больше объём выборки, тем точнее оценивается теоретическое распределение выборочными данными. Данный вывод обосновывается следующей теоремой В. И. Гливенко, которая считается фундаментальной теоремой математической статистики.

Теорема 1.1 Если $F_x(x)$ и $F_x^*(x)$ – теоретическая и эмпирическая функции распределения для выборки объема n, то для любого $\varepsilon > 0$

$$\lim_{n\to\infty} P(\,|\,F_x(x) - F_x^*(x)\,|\,< \varepsilon) = 1.$$

Аналитическое выражение эмпирической функции распределения хорошо находится по данным вариационного ряда. Для этого определяются значения функции для всех вариант выборки.

Пример 1.14 Найдем эмпирическую функцию распределения по данным вариационного ряда из примера 1.6 о посещаемости университетской библиотеки.

Таблица 1.8 – Вариационный ряд данных посещаемости библиотеки

Число посещений	0	1	2	3	4	5
m_i	5	6	7	2	3	2
p_i^*	0,20	0,24	0,28	0,08	0,12	0,08

Объём данной выборки n = 25. По определению 1.19 находим значения эмпирической функции распределения для всех вариант

$$F_x^*(0) = \frac{n_0}{n} = \frac{0}{25} = 0;$$

$$F_x^*(1) = \frac{n_1}{n} = \frac{5}{25} = 0{,}20;$$

$$F_x^*(2) = \frac{n_2}{n} = \frac{5+6}{25} = 0{,}20 + 0{,}24 = 0{,}44;$$

$$F_x^*(3) = 0{,}20 + 0{,}24 + 0{,}28 = 0{,}72;$$

$$F_x^*(4) = 0{,}20 + 0{,}24 + 0{,}28 + 0{,}08 = 0{,}80;$$

$$F_x^*(5) = 0{,}80 + 0{,}12 = 0{,}92;$$

для $x > 5$ $F_x^*(x) = 0{,}92 + 0{,}08 = 1$.

Объединим полученные результаты и найдем выражение для $F_x^*(\boldsymbol{x})$.

$$F_x^*(x) = \begin{cases} 0, \text{ ïðè } & x \le 0; \\ 0{,}20 \text{ ïðè } & 0 < x \le 1; \\ 0{,}44 \text{ ïðè } & 1 < x \le 2; \\ 0{,}72 \text{ ïðè } & 2 < x \le 3; \\ 0{,}80 \text{ ïðè } & 3 < x \le 4; \\ 0{,}92 \text{ ïðè } & 4 < x \le 5; \\ 1, \text{ ïðè } & x > 5. \end{cases}$$

Построим графическое изображение данной функции.

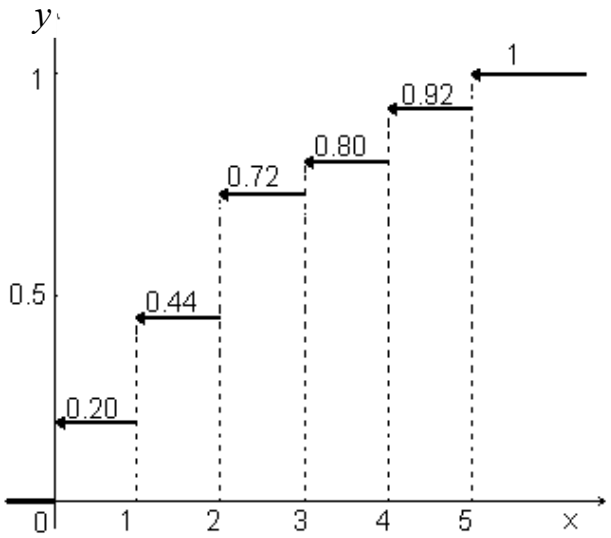


Рисунок 1.13 – График эмпирической функции распределения посещаемости библиотеки студентами

■

Интервальный статистический ряд скрывает конкретные выборочные значения случайной величины. Поэтому точные значения эмпирической функции распределения можно определить только на границах интервалов, внутри интервалов при отсутствии полной информации о выборочных данных значения эмпирической функции распределения можно определить только приближенно. В том случае, когда исследуется непрерывная случайная величина, то в системе координат отмечаются найденные значения эмпирической функции распределения на границах интервалов и полученные точки последовательно соединяются плавной линией. В результате получается

приближенный график эмпирической функции распределения.

Пример 1.15 Построим эмпирическую функцию распределения по данным статистического ряда из примера 1.8 обследования высот городских зданий.

Ранее мы получили следующий статистический ряд:

Высота зданий	5–10	10–15	15–20	20–25	25–30	30–35	35–40	40–45	45–50
m_i	2	3	5	6	8	7	5	3	1
p_i^*	0,050	0,075	0,125	0,150	0,200	0,175	0,125	0,075	0,025

Объём исследуемой выборки n = 40. Вычислим значения эмпирической функции распределения на границах интервалов.

$$F_x^*(5) = 0;\ \ F_x^*(10) = 0{,}050;\ \ F_x^*(15) = 0{,}050 + 0{,}075 = 0{,}125;$$

$$F_x^*(20) = 0{,}050 + 0{,}075 + 0{,}125 = 0{,}250;\ \ F_x^*(25) = 0{,}400;$$

$$F_x^*(30) = 0{,}600;\ \ F_x^*(35) = 0{,}775;\ \ F_x^*(40) = 0{,}900;$$

$$F_x^*(45) = 0{,}975;\ \ F_x^*(50) = 1.$$

По соответствующим точкам (5;0), (10; 0,050), (15; 0,125), (20; 0,250), (25, 0,400), (30; 0,600), (35; 0,775), (40; 0,920), (45; 0,975), (50;1) построим приближенный график эмпирической функции распределения.

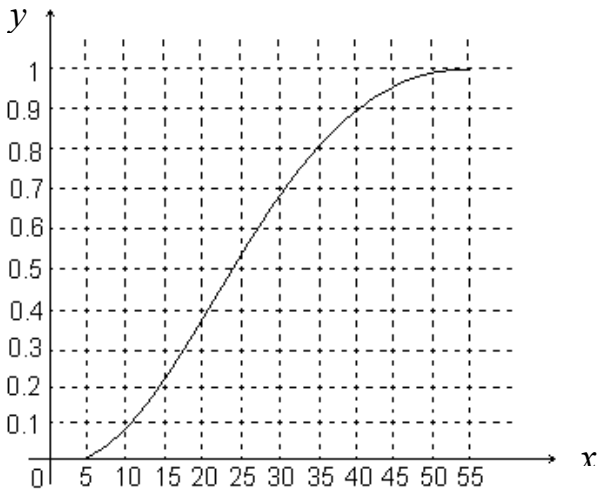


Рисунок 1.14 – График эмпирической функции распределения высот зданий города

■

Эмпирическая функция распределения обладает всеми свойствами теоретической функции распределения.

1. Эмпирическая функция распределения является неубывающей функцией, то есть

$$F_x^*(x_1) \le F_x^*(x_2) \text{ при } x_1 < x_2 .$$

2. Справедливы следующие равенства:

$$\lim_{x \to -\infty} F_x^*(x) = 0 \quad \text{и} \quad \lim_{x \to +\infty} F_x^*(x) = 1 .$$

3. Все значения эмпирической функции распределения находятся между 0 и 1, то есть

$$0 \le F_x^*(x) \le 1,$$

4. Эмпирическая функция распределения является непрерывной слева, то есть

$$\lim_{x \to x_0 - 0} F_x^*(x) = F_x^*(x_0) .$$

Подчеркнем, что эмпирическая функция распределения является статистическим аналогом теоретической функции распределения. Она помогает приближенно оценить общий характер закона распределения исследуемой случайной величины. Однако необходимо понимать, что на любом эмпирическом распределении отражается влияние многих случайных факторов.

Упражнения

1.1 Группа студентов, состоящая из 25 человек, должна выбрать 5 делегатов для участия в студенческой профсоюзной конференции. Было проведено голосование и выбрано 5 студентов, получивших большинство голосов. Можно ли считать такой выбор случайным? Обоснуйте свой ответ.

1.2 В библиотеке имеется только 15 задачников и 20 учебников по высшей математике на группу, состоящую из 28 студентов. Опишите процедуру справедливого на ваш взгляд распределения книг.

1.3 На 500 первокурсников, нуждающихся в общежитии, выделено только 70 мест. Распределите места в общежитие с помощью таблицы случайных чисел.

1.4 На участие в телешоу поступило 2700 заявок от зрителей. Для первого тура необходимо выбрать 50 участников. Сделайте справедливый выбор с помощью таблицы случайных чисел.

1.5 Коллектив предприятия состоит из 240 мужчин и 80 женщин. Для анкетирования нужно выбрать 10 % работников предприятия так, чтобы отношение 3:1 мужчин и женщин сохранилось. Составьте список анкетируемых.

1.6 При регистрации кандидат в депутаты представил 10000 подписных листов. Избирательная комиссия должна проверить не менее 1000 из представленных листов. Каким образом провести случайный отбор проверяемых листов с помощью таблицы случайных чисел?

1.7 Опишите возможную процедуру выбора 12 присяжных заседателей для суда.

1.8 По данным о числе отсутствующих на занятиях студентов, собранных в течение 25 дней, составьте вариационный ряд и постройте полигон частот и полигон относительных частот.

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

1.9 Составьте вариационный ряд и постройте полигон относительных частот по данным экзаменационным оценкам студентов 1 курса. Найдите эмпирическую функцию распределения оценок.

5	4	2	6	9	7	9	6	5	3
4	5	7	8	7	8	5	7	6	4
2	6	3	7	6	9	2	4	3	7
5	3	5	10	5	1	6	7	1	8
3	4	6	4	3	3	7	8	4	3
6	6	6	2	5	4	8	5	7	2
7	7	7	5	4	6	5	6	8	5
8	9	8	4	2	5	4	4	5	6

1.10 Для проверки качества семян было сделано 40 опытных посевов по 10 семян. По данным результатам всхожести семян составьте вариационный ряд, постройте полигон частот и найдите эмпирическую функцию распределения.

8	10	6	5	9	5	7	8	9	3
7	6	4	8	7	9	6	6	5	7
7	3	4	6	9	8	7	5	8	6
5	9	9	10	7	8	8	6	7	7

1.11 По анкетам 50 безработных собраны сведения о том, сколько раз они меняли место работы до регистрации в бюро занятости. По этим данным составьте вариационный ряд, постройте полигоны частот, найдите эмпирическую функцию распределения.

1	1	2	3	4	6	5	3	4	2
3	3	1	2	5	1	2	1	2	1
2	1	2	1	1	2	4	2	3	1
1	2	1	4	2	1	1	5	1	2
4	5	3	1	3	7	3	3	6	4

1.12 По результатам медицинского обследования получены данные о весе студентов группы. Составьте статистический ряд данной выборки и постройте гистограмму частот. Найдите эмпирическую функцию распределения и её график.

58,2	48,5	51,5	65,2	56,2	93,5	56,4	70,4	90,2	84,2
67,3	50,4	64,1	72,3	46,4	54,3	45,3	50,5	44,8	59,1
72,4	62,7	76,3	70,2	62,7	68,2	69,5	60,3	51,8	61,4
54,3	58,8	47,8	53,7	74,6	52,2	76,5	52,6	62,4	54,5
70,8	78,3	70,5	80,4	81,3	74,2	84,2	73,1	78,5	73,2

1.13 В процессе биологического исследования было проведено измерение длины 100 листьев одного дерева. По данным этого эксперимента составьте статистический ряд, постройте гистограмму относительных частот. Найдите эмпирическую функцию распределения и постройте её график.

1,5	12,1	7,8	12,5	7,6	6,8	4,4	11,5	7,5	9,4
10,1	6,2	3,2	2,4	12,7	8,4	10,2	3,6	4,3	10,8
4,2	11,2	8,4	9,5	1,8	3,5	6,5	17,1	13,2	2,7
8,3	2,8	14,8	5,7	8,4	10,7	12,4	7,4	1,8	7,3
5,3	13,2	6,3	13,8	4,5	5,8	2,3	12,6	13,4	4,8
11,3	8,1	11,7	8,3	11,4	12,8	8,1	5,3	6,5	11,2
6,7	10,2	10,2	6,7	10,4	9,6	11,4	10,6	16,4	7,8
12,2	9,1	9,5	9,5	9,5	14,4	15,8	8,7	10,6	13,6
9,4	14,6	12,3	10,6	12,7	11,5	9,2	11,3	8,6	8,2
14,6	13,4	17,8	11,4	15,2	15,1	17,8	15,4	11,4	11,8

1.14 При изучении кадровой политики крупного объединения были собраны сведения о том, сколько лет проработали в данной отрасли руководители объединения до первого назначения на руководящую должность. Исследуйте распределение данной выборки, составьте статистический ряд, постройте гистограмму. Найдите эмпирическую функцию распределения и постройте её график.

21	4	8	12	17	21	8	22
16	11	16	10	8	13	14	15
7	9	10	7	12	2	7	9
12	6	11	14	9	6	12	3
17	13	18	19	21	10	16	12

2 Числовые характеристики выборочного распределения

2.1 Мода и медиана

Оптимально сгруппированные и визуально представленные статистические данные, тем не менее, не позволяют увидеть и обосновать глубокие закономерности исследуемых явлений. Для проведения статистического анализа необходимы определенные числовые характеристики, которые наилучшим образом отражают полученные результаты экспериментов. Поэтому вычисление простых и понятных показателей, обобщающих наиболее существенные свойства статистических данных, является одним из основных методов математической статистики. Традиционно для обобщения большого количества экспериментальных данных используются определенные так называемые **средние** выборочные показатели. Рассмотрим наиболее известные типы средних. В любой совокупности выборочных данных естественно выделяется значение, которое появляется чаще других.

Определение 2.1 Если в выборке $x_1, x_2, \ldots, x_n$ есть одно выборочное значение, имеющее наибольшую частоту, то оно называется модой данной выборки.

Для обозначения моды используется символ $X_{мод}$.

Пример 2.1 Рассмотрим выборочные данные о размерах обуви двадцати женщин.

Таблица 2.1 – Вариационный ряд данных размеров обуви

Размер	35	36	37	38	39	40
m_i	1	3	7	5	3	1

В данной выборке преобладает 37-й размер обуви, поэтому мода этой выборки равна 37, то есть $X_{мод} = 37$.

■

Выборка не имеет моды, когда в ней нет единственного элемента

с наибольшей частотой. Если в выборке несколько значений повторяются больше других одинаково часто, то выборка называется мультимодальной (бимодальной, тримодальной и т. д.). Для определения моды надо знать частоту каждого выборочного значения. В случае, когда мы имеем только интервальный статистический ряд, то из всех интервалов можно выделить **модальный интервал**, который содержит самые повторяемые значения выборки. Если длина каждого интервала статистического ряда равна ℓ, то для вычисления моды выполняется следующая процедура.

Алгоритм вычисления моды статистического ряда

Условие: длина каждого интервала статистического ряда одинакова.

1. Определяется модальный интервал статистического ряда.

Будем считать, что именно i-й интервал $[x_{i-1}; x_i)$ имеет наибольшую частоту, то есть является модальным. Вместе с ним также рассматриваются предыдущий $(i-1)$-й интервал $[x_{i-2}; x_{i-1})$ и последующий $(i+1)$-й интервал $[x_i; x_{i+1})$.

2. Для каждого из этих интервалов находятся соответствующие частоты:

$$m_{i-1},\ m_i \quad \text{è} \quad m_{i+1}.$$

3. Значение моды вычисляется по формуле:

$$X_{\text{ìîä}} = x_{i-1} + \frac{m_i - m_{i-1}}{(m_i - m_{i-1}) + (m_i - m_{i+1})}\ell.$$

Подчеркнем, что x_{i-1} – это нижняя граница модального интервала, m_{i-1} – частота предшествующего ему интервала, m_{i+1} – частота последующего интервала, ℓ – длина каждого интервала.

Пример 2.2 Найдем моду статистического ряда по данным примера 1.7 о возрасте пациентов клиники.

Таблица 2.2 – Данные исследования возраста пациентов поликлиники

Возраст	10–20	20–30	30–40	40–50	50–60	60–70	70–80	80–90
m_i	17	24	35	48	57	42	21	6

Очевидно, что наибольшую частоту имеет пятый интервал [50; 60), для которого $m_5 = 57$. Частота предыдущего интервала $m_4 = 48$, частота последующего интервала $m_6 = 42$, длина каждого интервала $\ell = 10$. Тогда мода вычисляется по формуле:

$$X_{\text{мод}} = 50 - \frac{57-48}{(57-48)+(57-42)} \cdot 10 = 50 - \frac{9}{24} \cdot 10 = 53{,}75.$$

Учитывая, что выборочные данные характеризуют возраст пациентов в годах, округляем найденное значение:

$$X_{\text{мод}} = 53 \text{ ãîäà} \quad è \quad 9 \text{ ìåñÿöåâ}.$$

Таким образом, чаще других обращаются в поликлинику пациенты в возрасте от 50 до 60 лет, причем в этой группе наиболее проблемный возраст составляет 53,75 года.

■

Мода – одна из немногих характеристик, которая используется при анализе не только количественных, но и качественных данных.

Пример 2.3 Рассмотрим данные анкетирования 40 посетителей автосалона о предпочитаемом ими цвете автомобиля:

Таблица 2.3 – Результаты анкетирования о любимом цвете автомобиля

Цвет	белый	черный	красный	синий	зеленый	серый	другие
m_i	10	8	6	4	3	5	4

В этой выборке модой является белый цвет, имеющий наибольшую частоту.

■

Понятие моды используется главным образом в прикладных исследованиях тогда, когда возникает необходимость выявления в выборке большого объема наиболее преобладающих вариант. Такие ситуации часто встречаются при изучении потребительского спроса, качественного состава продукции массового производства, результатов опроса населения и в других случаях. Но так как мода не всегда существует, то в аналитической статистике это понятие используется крайне редко.

Результаты многочисленных исследований показывают, что значительная часть выборочных данных имеет тенденцию собираться вокруг некоторого центра. Это свойство обобщается введением следующего понятия медианы.

Определение 2.2 Пусть все выборочные данные $x_1, x_2, \ldots, x_n$ расположены в порядке возрастания с сохранением повторяющихся значений. Если n – нечетное число, то медианой этой выборки называется число $X_{мед}$, равное выборочному значению $x_{\frac{n+1}{2}}$, стоящему на $\frac{n+1}{2}$-м месте. Если n – четное число, то медиана $X_{мед}$ равна полусумме выборочных значении $x_{\frac{n}{2}}$ и $x_{\frac{n}{2}+1}$, стоящих соответственно на $\frac{n}{2}$-м и $\left(\frac{n}{2}+1\right)$-м местах:

$$X_{\text{ìåä}} = x_{\frac{n+1}{2}} \text{ при нечётном n}; \quad X_{мед} = \frac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2} \text{ при чётном n}.$$

Другими словами, если объем выборки является нечетным числом, то медиана $X_{мед}$ равна единственному выборочному значению, расположенному в самой середине упорядоченной выборки. Если же объем выборки является четным числом, то посередине выборки находятся два соседних значения. В этом случае медиана равна сумме этих значений, деленной на 2.

Пример 2.4 Ниже приводятся две выборки данных о количестве новых слов, выученных каждым из девяти учеников одной группы и каждым из десяти учеников другой группы в течение одного урока английского языка.

I группа: 1, 3, 3, 4, 4, 5, 6, 7, 8.
II группа: 1, 2, 2, 3, 3, 5, 5, 6, 6, 7.

Обе выборки записаны в порядке возрастания их значений. Объем первой выборки n = 9 является нечетным числом, поэтому посередине находится одно пятое значение $x_5 = 4$, которое и является медианой этой выборки:

$$X_{мед(I)} = 4.$$

Посередине второй выборки находятся два значения 3 и 5, поэтому

$$X_{\text{мед}\,(II)} = \frac{3+5}{2} = 4.$$

Значит, эти выборки имеют равные медианы $X_{мед} = 4$.

■

Медиана делит выборку на две части, каждая из которых содержит одинаковое количество элементов. Первая часть состоит из выборочных значений, расположенных до медианы. Их величина не может быть больше величины медианы. Вторую часть составляют выборочные значения, расположенные после медианы. Их величина не может быть меньше величины медианы. Например, медиана $X_{\text{мед}} = 4$ из предыдущего примера показывает, что половина учеников каждой группы за урок выучила не более 4-х новых английских слов, а вторая половина запомнила не менее 4-х слов. Таким образом, медиана является определенным граничным значением исследуемой случайной величины, показывающим, что в половине всех испытаний получаются выборочные значения, не превосходящие медиану, а в половине испытаний получаются значения, практически превосходящие медиану по величине. Фактически медиана выборки характеризует структуру и конфигурацию составляющих элементов исследуемой совокупности.

Теперь рассмотрим способ нахождения медианы по сгруппированным данным.

Таблица 2.4 – Произвольный статистический ряд

Интервалы значений X	$[x_0; x_1)$	$[x_1; x_2)$	…	$[x_{i-1}; x_i)$	…	$[x_{k-1}; x_k]$
m_i	m_1	m_2	…	m_i	…	m_k

Алгоритм вычисления медианы статистического ряда

Условие: длина каждого интервала статистического ряда равна ℓ.

1. Прежде всего, определяется так называемый медианный интервал. Для этого вычисляется число $\frac{n}{2}$, равное половине

всего количества выборочных значений. Затем последовательно складываются частоты первого, второго и так далее интервалов до тех пор, пока не получится сумма, которая либо равна $\frac{n}{2}$, либо чуть больше $\frac{n}{2}$. Интервал, соответствующий последней прибавленной частоте, и будет являться медианным. Допустим, что сумма частот первых s интервалов не меньше числа $\frac{n}{2}$, то есть

$$\sum_{i=1}^{s} m_i \geq \frac{n}{2},$$

но сумма частот нижних $s-1$ интервалов меньше $\frac{n}{2}$, то есть

$$\sum_{i=1}^{s-1} m_i < \frac{n}{2}.$$

Тогда именно s-й интервал $[x_{s-1}; x_s)$ является медианным.

2. Далее медиана статистического ряда вычисляется по формуле:

$$X_{\text{мед}} = x_{s-1} + \frac{\frac{n}{2} - \sum_{i=1}^{s-1} m_i}{m_s} \cdot \ell.$$

Подчеркнем, что x_{s-1} – это нижняя граница медианного интервала, n – объем всей выборки, $\sum_{i=1}^{s-1} m_i$ – сумма частот всех интервалов, расположенных ниже медианного, m_s – частота медианного интервала, ℓ – длина каждого интервала.

Пример 2.5 Найдем медиану статистического ряда по данным о возрасте пациентов поликлиники.

Таблица 2.5 – Данные исследования о возрасте пациентов поликлиники

Возраст	10–20	20–30	30–40	40–50	50–60	60–70	70–80	80–90
m_i	17	24	35	48	57	42	21	6

Объем всей выборки n = 250, поэтому $\frac{n}{2}$ = 125. Последовательно

складываем частоты пока не получим сумму, равную или большую 125:

$$17 + 24 + 35 + 48 = 124, \quad \text{но}$$

$$17 + 24 + 35 + 48 + 57 = 181.$$

Сумма частот первых четырех интервалов меньше 125, а сумма частот пяти интервалов больше 125, поэтому именно пятый интервал [50; 60) является медианным. Вычислим медиану по данной формуле:

$$X_{ìåä} = 50 + \frac{125 - 124}{57} \cdot 10 \quad = \quad 50{,}175$$

или

$$X_{ìåä} \quad = \quad 50 \text{ ëåò } è \quad 2 \text{ ìåñÿöà} \ .$$

Это значение медианы показывает, что возраст половины пациентов в данной выборке не больше 50 лет и 2 месяцев.

■

Пример 2.6 Найдем медиану статистического ряда из примера 1.8, представляющего данные о высоте зданий.

Таблица 2.6 – Статистический ряд измерений высоты зданий

Высота зданий	5–10	10–15	15–20	20–25	25–30	30–35	35–40	40–45	45–50
m_i	2	3	5	6	8	7	5	3	1

Объем всей выборки $n = 40$, поэтому $\frac{n}{2} = 20$. Находим последовательно суммы частот:

$$2 + 3 + 5 + 6 = 16 < 20,$$

$$\text{но} \quad 2 + 3 + 5 + 6 + 8 = 24 > 20.$$

Поэтому пятый интервал [25; 30) является медианным.
Вычислим медиану по указанной выше формуле:

$$X_{ìåä} \quad = \quad 25 + \frac{20 - 16}{8} \cdot 5 \quad = \quad 25 + 2{,}5 \quad = \quad 27{,}5 \, .$$

Следовательно, $X_{мед} = 27{,}5$. Это значение показывает, что в

исследуемой выборке половина зданий имеет высоту не более 27,5 метров.

■

Заметим, что медиана существует в любой статистической выборке. Следует подчеркнуть и такое полезное свойство медианы как её нечувствительность к месторасположению экстремальных значений в больших выборках. Наличие в выборке сильно отклоняющихся значений создает определенные проблемы при анализе, поэтому использование медианы позволяет в определенных случаях обойти некоторые трудности.

Известно, что существуют симметричные и асимметричные распределения случайных величин. В том случае, когда выборочные данные реализуют симметричное распределение, то значение медианы и моды практически совпадают. Для асимметричных распределений равенство не выполняется.

2.2 Выборочное среднее

Теперь рассмотрим наиболее часто используемое понятие **среднего**, которое в отличие от моды и медианы объединяет все выборочные значения.

Определение 2.3 Средним арифметическим выборки x_1, x_2, …, x_n значений случайной величины X называется число $\overline{X}$, равное сумме всех выборочных значений, деленной на число всех наблюдений n:

$$\overline{X} = \frac{x_1 + x_2 + \ldots + x_n}{n} \quad \text{èëè} \quad \overline{X} = \frac{1}{n} \sum_{i=1}^{n} \tilde{o}_i \ .$$

Обычно **среднее арифметическое** $\overline{X}$ называется **выборочным средним** или просто **средним.**

В случае, когда выборка совпадает со всей исследуемой генеральной совокупностью, то таким же образом вычисляется её генеральное среднее. В дальнейшем выборочное среднее всей генеральной совокупности значений случайной величины X будем обозначать символами μ или μ_x и называть **генеральным средним.**

Пример 2.7 Студент Денисов по пяти предметам получил следующие экзаменационные оценки по десятибальной системе: 8, 9, 8, 7, 6. Вычислим среднее данных оценок:

$$\overline{\mathrm{X}} = \frac{8+9+8+7+6}{5} = 7{,}6 .$$

■

Подчеркнем, что выборочное среднее не обязательно должно быть элементом самой выборки. В том случае, когда выборка представляется вариационным рядом, содержащим k вариант x_1, x_2, ..., x_k с соответствующими частотами $\mathrm{m}_1, \mathrm{m}_2, \ldots, \mathrm{m}_\mathrm{k}$, то среднее $\overline{\mathrm{X}}$ вычисляется по следующей формуле:

$$\overline{\mathrm{X}} = \frac{x_1 \cdot \mathrm{m}_1 + x_2 \cdot \mathrm{m}_2 + \ldots + x_\mathrm{k} \cdot \mathrm{m}_\mathrm{k}}{\mathrm{m}_1 + \mathrm{m}_2 + \ldots + \mathrm{m}_\mathrm{k}},$$

или

$$\overline{\mathrm{X}} = \frac{1}{\mathrm{n}} \sum_{i=1}^{k} \tilde{o}_i \cdot \mathrm{m}_i \text{, где } \mathrm{n} = \mathrm{m}_1 + \mathrm{m}_2 + \ldots + \mathrm{m}_\mathrm{k} .$$

Пример 2.8 Результаты экзамена по математике для группы студентов, состоящей из 25 человек, представлены следующим вариационным рядом:

Таблица 2.7 – Экзаменационные оценки группы студентов

Оценки	1	2	3	4	5	6	7	8	9	10
m_i	0	0	2	5	4	4	5	2	3	0

Найдем среднее:

$$\overline{\mathrm{X}} = \frac{1\cdot 0 + 2\cdot 0 + 3\cdot 2 + 4\cdot 5 + 5\cdot 4 + 6\cdot 4 + 7\cdot 5 + 8\cdot 2 + 9\cdot 3 + 10\cdot 0}{25} = 5{,}92$$

■

Пример 2.9 Для вариационного ряда данных о посещаемости университетской библиотеки из примера 1.6., найдем среднее, моду и медиану:

Таблица 2.8 – Вариационный ряд данных посещаемости библиотеки

Число посещений	0	1	2	3	4	5
m_i	5	6	7	2	3	2

Вычислим среднее:

$$\overline{\mathrm{X}} = \frac{0 \cdot 5 + 1 \cdot 6 + 2 \cdot 7 + 3 \cdot 2 + 4 \cdot 3 + 5 \cdot 2}{5 + 6 + 7 + 2 + 3 + 2} = 1{,}92 .$$

Очевидно, что мода данного ряда $\mathrm{X}_{\text{моды}} = 2$. Легко находится и медиана $\mathrm{X}_{\text{меды}} = 2$. Среднее $\overline{\mathrm{O}} = 1{,}92$ меньше моды и медианы, равным 2. Равенство $\mathrm{X}_{\text{моды}} = \mathrm{X}_{\text{меды}} = 2$ приводит к предположению о симметричности этого распределения.

■

В случае, когда выборка сгруппирована в виде статистического ряда, состоящего из k интервалов

$$[x_0; x_1), \ [x_1; x_2), \ \ldots, \ [x_{i-1}; x_i), \ \ldots, \ [\tilde{o}_{\mathrm{k}-1}; \tilde{o}_{\mathrm{k}}],$$

то конкретные выборочные значения могут быть неизвестными, и рассмотренные выше формулы не пригодны для вычисления среднего. Поэтому для каждого интервала вычисляется его середина, или **интервальное среднее** по формуле:

$$\widetilde{x}_i = \frac{x_{i-1} + x_i}{2}, \quad i = 1, 2, \ldots, \mathrm{k} .$$

При выполнении вычислений каждое выборочное значение, принадлежащее интервалу, заменяется интервальным средним.

Определение 2.4 Если $\widetilde{x}_1, \ \widetilde{x}_2, \ \ldots, \ \widetilde{x}_{\mathrm{k}}$ – интервальные средние, а $\mathrm{m}_1, \mathrm{m}_2, ..., \mathrm{m}_{\mathrm{k}}$ – соответствующие частоты всех интервалов статистического ряда, то выборочное среднее определяется формулой:

$$\overline{\mathrm{X}} = \frac{\widetilde{x}_1 \cdot \mathrm{m}_1 + \widetilde{x}_2 \cdot \mathrm{m}_2 + \ldots + \widetilde{x}_{\mathrm{k}} \cdot \mathrm{m}_{\mathrm{k}}}{\mathrm{m}_1 + \mathrm{m}_2 + \ldots + \mathrm{m}_{\mathrm{k}}},$$

или

$$\overline{\mathrm{X}} = \frac{\sum_{i=1}^{\mathrm{k}} \widetilde{x}_i \cdot \mathrm{m}_1}{\mathrm{n}}, \quad \text{ãäå} \quad \mathrm{m}_1 + \mathrm{m}_2 + \ldots + \mathrm{m}_{\mathrm{k}} = \mathrm{n} .$$

Пример 2.10 Найдем выборочное среднее возраста пациентов поликлиники по данным статистического ряда из примера 1.3.

Таблица 2.9 – Данные исследования возраста пациентов поликлиники

Возраст	10–20 15	20–30 25	30–40 35	40–50 45	50–60 55	60–70 65	70–80 75	80–90 85
m_i	17	24	35	48	57	42	21	6

Вычислим интервальные средние всех интервалов:

$$\tilde{x}_1 = \frac{10+20}{2} = 15; \quad \tilde{x}_2 = \frac{20+30}{2} = 25; \quad \tilde{x}_3 = \frac{30+40}{2} = 35;$$

$$\tilde{x}_4 = 45; \quad \tilde{x}_5 = 55; \quad \tilde{x}_6 = 65; \quad \tilde{x}_7 = 75; \quad \tilde{x}_8 = 85.$$

Найдем выборочное среднее:

$$\overline{X} = \frac{15 \cdot 17 + 25 \cdot 24 + 35 \cdot 35 + 45 \cdot 48 + 55 \cdot 57 + 65 \cdot 42 + 75 \cdot 21 + 85 \cdot 6}{17 + 24 + 35 + 48 + 57 + 42 + 21 + 6} =$$

$$= \frac{12190}{250} = 48{,}76.$$

Среднее $\overline{X} = 48{,}76$ характеризует средний возраст наиболее нуждающихся в лечении пациентов.

■

Пример 2.11 Найдем выборочное среднее для статистического ряда из примера 1.8. по данным о высоте зданий:

Таблица 2.10 – Данные исследования высоты зданий

Высота зданий	5–10 7,5	10–15 12,5	15–20 17,5	20–25 22,5	25–30 27,5	30–35 32,5	35–40 35,4	40–45 42,5	45–50 47,5
m_i	2	3	5	6	8	7	5	3	1

Объем исследуемой выборки $n = 40$. Вычислим интервальные средние:

$$\widetilde{x}_1 = \frac{5+10}{2} = 7{,}5;\ \widetilde{x}_2 = \frac{10+15}{2} = 12{,}5;\ \widetilde{x}_3 = \frac{15+20}{2} = 17{,}5;\ \widetilde{x}_4 = 22{,}5;$$

$$\widetilde{x}_5 = 27{,}5; \quad \widetilde{x}_6 = 32{,}5; \quad \widetilde{x}_7 = 37{,}5; \quad \widetilde{x}_8 = 42{,}5; \quad \widetilde{x}_9 = 47{,}5.$$

Найдем выборочное среднее:

$$\overline{\mathrm{X}} = \frac{7{,}5\cdot 2 + 12{,}5\cdot 3 + 17{,}5\cdot 5 + 22{,}5\cdot 6 + 27{,}5\cdot 8 + 32{,}5\cdot 7 + 37{,}5\cdot 5 + 42{,}5\cdot 3 + 47{,}5\cdot 1}{40} =$$

$$= 27{,}125.$$

Итак, $\overline{\mathrm{X}} = 27{,}125$ метров – это среднее высот зданий данной выборки.

■

Существуют различные приемы, облегчающие вычислительную работу при нахождении выборочного среднего.

Если выборка содержит большие числа или многократно повторяющиеся близкие значения, то статистические данные можно преобразовать с помощью следующего равенства:

$$u_i = \frac{x_i - \alpha}{\beta}, \quad i = 1, 2, \ldots, \mathrm{n},$$

где α и β – любые действительные числа, причем $\beta \neq 0$.

Число α подбирается так, чтобы разности $x_i - \alpha$ были бы наименьшими, число β изменяет масштаб данных.

В результате преобразования данная выборка x_1, x_2, ..., x_n заменяется выборкой $u_1, u_2, \ldots, u_n$ с такими же соответствующими частотами.

Выборочная средняя $\overline{\mathrm{X}}$ выражается через выборочную среднюю $\overline{\mathrm{U}}$ по формуле:

$$\overline{\mathrm{X}} = \beta \cdot \overline{\mathrm{U}} + \alpha\,.$$

Правильный подбор значений α и β обычно облегчает нахождение среднего.

Отметим, что при вычислении выборочных числовых показателей удобно пользоваться определенными расчетными таблицами, в которые обычно записывают необходимые промежуточные результаты.

Пример 2.12 Найдем среднее возраста пациентов поликлиники по статистическому ряду из примера 1.3., используя преобразования выборочных данных.

Таблица 2.11 – Данные исследования возраста пациентов поликлиники

Возраст	10–20 15	20–30 25	30–40 35	40–50 45	50–60 55	60–70 65	70–80 75	80–90 85
m_i	17	24	35	48	57	42	21	6

Положим, что $\alpha = 55$. Такой выбор обусловлен тем, что именно это значение имеет наибольшую частоту. В качестве β выберем длину интервалов, то есть $\beta = 10$.

Запишем формулу преобразования выборки:

$$u_i = \frac{\widetilde{x}_i - 55}{10}, \quad \mathrm{i} = 1, 2, \ldots, 8.$$

Все необходимые расчеты будем записывать в следующей таблице.

Таблица 2.12 – Вычисления среднего возраста пациентов поликлиники

Возраст X	Интервальное среднее $\widetilde{x}_i$	Частота m_i	$\widetilde{x}_i - 55$	$u_i = \frac{\widetilde{x}_i - 55}{10}$	$u_i \mathrm{m}_i$
10 – 20	15	17	-40	-4	-68
20 – 30	25	24	-30	-3	-72
30 – 40	35	35	-20	-2	-70
40 – 50	45	48	-10	-1	-48
50 – 60	55	57	0	0	0
60 – 70	65	42	10	1	42
70 – 80	75	21	20	2	42
80 – 90	85	6	30	3	18
		$\mathrm{n} = 250$			$\sum_{i=1}^{\beta} u_i \mathrm{m}_i = -156$

Найдем $\overline{U} = \frac{-156}{250} = -0{,}624$.

Выборочное среднее $\overline{X}$ вычислим по следующей формуле:

$$\overline{X} = 10 \cdot (-0{,}624) + 55 = 48{,}76.$$

Это значение $\overline{X} = 48{,}76$ совпадает с ранее найденным значением из примера 2.10.

■

Пример 2.13 Следующий статистический ряд представляет результаты проведенного измерения роста пятидесяти семнадцатилетних девушек.

Таблица 2.13 – Данные измерения роста

Рост X	150–155 152,5	155–160 157,5	160–165 162,5	165–170 167,5	170–175 172,5	175–180 177,5
Частота m_i	3	12	14	10	7	4

Для вычисления среднего, прежде всего найдем интервальные средние:

$\tilde{x}_1 = 152{,}5;\ \tilde{x}_2 = 157{,}5;\ \tilde{x}_3 = 162{,}5;$

$\tilde{x}_4 = 167{,}5;\ \tilde{x}_5 = 172{,}5;\ \tilde{x}_6 = 177{,}5.$

В качестве α выберем интервальное среднее с наибольшей частотой: $\alpha = 162{,}5$. Положим $\beta = 5$, что совпадает с длиной интервалов.

Преобразуем данные по формуле:

$$u_i = \frac{\tilde{x}_i - 162{,}5}{5}.$$

Дальнейшие вычисления записываем в следующей таблице:

Таблица 2.14 – Вычисление среднего роста девушек

Рост X	Интервальное среднее $\widetilde{x}_i$	Частота m_i	$\widetilde{x}_i - 162{,}5$	$u_i = \frac{\widetilde{x}_i - 162{,}5}{5}$	$u_i m_i$
150–155	152,5	3	-10	-2	-6
155–160	157,5	12	-5	-1	-12
160–165	162,5	14	0	0	0
165–170	167,5	10	5	1	10
170–175	172,5	7	10	2	14
175–180	177,5	4	15	3	12
		$n = 50$			$\sum_{i=1}^{6} u_i m_i = 18$

Получаем $\overline{U} = \frac{18}{50} = 0{,}36.$

Найдем выборочное среднее по формуле:

$$\overline{X} = 5 \cdot 0{,}36 + 162{,}50 = 164{,}30$$

Таким образом, выборочное среднее данной выборки равно 164,3.

■

Заметим, что выборочное среднее несгруппированной выборки в общем виде отличается от выборочного среднего этой выборки, вычисленного после группировки. Однако, для большинства исследований это различие является несущественным.

Формулу для вычисления среднего вариационного ряда можно преобразовать следующим образом:

$$\frac{x_1 m_1 + x_2 m_2 + \ldots + x_k m_k}{n} = x_1 \frac{x_1}{n} + x_2 \frac{m_2}{n} + \ldots + x_k \frac{m_k}{n} = x_1 p_1^* + x_2 p_2^* + \ldots + x_k p_k^*,$$

где p_1^*, p_2^*, …, p_k^* – относительные частоты соответствующих значений. Известно что, при неограниченном увеличении числа испытаний $n \to \infty$ относительная частота стремится к вероятности события:

$$p_1^* \to P(X = x_1) = p_1, \quad p_2^* \to P(X = x_2) = p_2, \quad \ldots,$$

$$p_k^* \to P(X = x_k) = p_k.$$

Отсюда следует, что

$$\overline{\mathrm{X}} = x_1 \mathrm{p}_1^* + x_2 \mathrm{p}_2^* + \ldots + x_\mathrm{k} \mathrm{p}_\mathrm{k}^* \to x_1 \mathrm{p}_1 + x_2 \mathrm{p}_2 + \ldots + x_\mathrm{k} \mathrm{p}_\mathrm{k} = \mathrm{MX}\,.$$

Таким образом, при увеличении числа испытаний среднее $\overline{\mathrm{X}}$ стремится к математическому ожиданию MX случайной величины X, поэтому выборочное среднее $\overline{\mathrm{X}}$ является статистическим аналогом математического ожидания и обладает основными свойствами математического ожидания. Главное же отличие состоит в том, что математическое ожидание исследуемой случайной величины является постоянной величиной, а среднее $\overline{\mathrm{X}}$ является случайной величиной, так как его значение определяется случайной выборкой. Разные выборки из одной и той же генеральной совокупности могут иметь разные средние.

Определение 2.5 Число $d_i = x_i - \overline{\mathrm{X}}$ называется отклонением выборочного значения x_i от среднего $\overline{\mathrm{X}}$.

Пример 2.14 Администратор учреждения зафиксировал реальное время, затраченное на обеденный перерыв шестью сотрудниками: 55, 58, 62, 64, 65, 68.

Вычислим среднее:

$$\overline{\mathrm{X}} = \frac{55 + 58 + 62 + 64 + 65 + 68}{6} = 62\,.$$

Найдем отклонения от среднего:

$$d_1 = 55 - 62 = -7; \qquad d_4 = 64 - 62 = 2;$$
$$d_2 = 58 - 62 = -4; \qquad d_5 = 65 - 62 = 3;$$
$$d_3 = 62 - 62 = 0; \qquad d_6 = 68 - 62 = 6.$$

Теперь найдем сумму всех отклонений:

$$\sum_{i=1}^{6} d_i = -7 - 4 + 0 + 2 + 3 + 6 = 0\,.$$

■

В данном примере сумма всех отклонений от среднего равна 0. Такой результат справедлив и в самом общем случае.

Теорема 2.1 Сумма всех отклонений выборочных значений $x_1, x_2, \ldots, x_n$ от их среднего $\overline{X}$ равна 0.

Это свойство среднего подтверждает его центральную роль в совокупности выборочных данных. Равенство

$$(x_1 - \overline{X}) + (x_2 - \overline{X}) + \cdots + (x_n - \overline{X}) = 0$$

показывает, что выборочные значения окружают среднее $\overline{X}$ как справа, так и слева.

Интересное свойство среднего связано с суммой квадратов отклонений выборочных значений:

$$(x_1 - \overline{X})^2 + (x_2 - \overline{X})^2 + \cdots + (x_n - \overline{X})^2.$$

Теорема 2.2 Если x_1, x_2, …, x_n – случайная выборка со средним $\overline{X}$, то сумма квадратов разностей

$$(x_1 - x)^2 + (x_2 - x)^2 + \cdots + (x_n - x)^2$$

принимает свое минимальное значение при $x = \overline{X}$.

Например, запишем сумму квадратов разностей для выборки 55, 58, 62, 64, 65, 68 из предыдущего примера:

$$(55 - x)^2 + (58 - x)^2 + (62 - x) + (64 - x)^2 + (65 - x)^2 + (68 - x)^2.$$

Ранее мы нашли среднее данной выборки $\overline{X} = 62$. Из теоремы 2.2 следует, что эта сумма имеет минимальное значение при $x = 62$.

Выборочная средняя находит более широкое применение, чем другие средние, в практических и теоретических исследованиях. Среднее арифметическое обобщает все значения исследуемой выборки и часто используется в качестве единого представителя всей совокупности выборочных данных. Например, при многократных экспериментальных измерениях некоторой величины за истинное значение часто принимается выборочное среднее. Тем не менее, между средним и каждым индивидуальным выборочным значением существует определенное различие.

2.3 Геометрическое среднее и гармоническое среднее

Существуют определенные генеральные совокупности, для которых больше подходят другие виды средних. Рассмотрим понятия **среднего геометрического** и **среднего гармонического.**

Определение 2.6 Средним геометрическим выборки x_1, x_2, …, x_n с положительными значениями называется число $\overline{X}_{ãåîì}$, равное корню n-ой степени из произведения всех элементов выборки:

$$\overline{\mathbf{X}}_{ãåîì} = \sqrt[n]{x_1 x_2 \ldots x_n}.$$

Пример 2.15 Рассмотрим данные о посещаемости студентами группы консультаций по высшей математике. На первой консультации присутствовало 3 студента, на второй – 6, на третьей – 12 студентов.

Итак, выборка состоит из трех значений: 3, 16, 12, n = 3. Вычислим среднее геометрическое по формуле:

$$\overline{X}_{ãåîì} = \sqrt[3]{3 \cdot 6 \cdot 12} = 6.$$

Для сравнения вычислим и среднее арифметическое:

$$\overline{X} = \frac{3+6+12}{3} = 7.$$

Оба эти значения $\overline{X}_{ãåîì} = 6$ и $\overline{X} = 7$, являясь достаточно близкими, характеризуют среднее число студентов, присутствующих на одной консультации. Заметим, что

$$\overline{X}_{ãåîì} < \overline{\tilde{O}}.$$

■

В общем случае выборочное среднее и среднее геометрическое связаны неравенством:

$$\overline{X}_{ãåîì} \le \overline{\tilde{O}}.$$

Отметим, что формула, определяющая среднее геометрическое, не слишком удобна для расчетов. Поэтому она предпочтительно

используется в другом виде:

$$\lg\left(\overline{X}_{\text{ãåîì}}\right) = \frac{\lg x_1 + \lg x_2 + \cdots + \lg x_n}{n}.$$

Пример 2.16 Рассмотрим показатели доходности инвестиций финансового фонда в течение пяти последовательных лет:

$$3{,}2;\quad 4{,}5;\quad 7{,}4;\quad 8{,}1;\quad 10{,}5.$$

Найдем среднее геометрическое этой выборки:

$$\overline{X}_{\text{ãåîì}} = \sqrt[5]{3{,}2 \cdot 4{,}5 \cdot 7{,}4 \cdot 8{,}1 \cdot 10{,}5}.$$

Для вычислений используем логарифмическую формулу:

$$\lg\left(\overline{X}_{\text{ãåîì}}\right) = \frac{\lg 3{,}2 + \lg 4{,}5 + \lg 7{,}4 + \lg 8{,}1 + \lg 10{,}5}{5}.$$

$$\begin{aligned}
&\lg 3{,}1 = 0{,}5051\\
&\lg 4{,}5 = 0{,}6532\\
&\lg 7{,}4 = 0{,}8692\\
&\lg 8{,}1 = 0{,}9085\\
&\lg 10{,}5 = 1{,}0212\\
&\text{————————}\\
&\text{ñóììà} = 3{,}9572
\end{aligned}$$

Значит,

$$\lg\left(\overline{X}_{\text{ãåîì}}\right) = \frac{3{,}9572}{5} = 0{,}79144.$$

Отсюда получаем, что $\overline{X}_{\text{ãåîì}} = 6{,}14$.

Мы получили среднегодовой показатель доходности инвестиций в течение данного пятилетнего периода.

■

Необходимость использования среднего геометрического возникает при исследованиях темпов изменения величин, когда результаты последующих измерений пропорционально зависят от

ранее достигнутых значений. Среднее геометрическое дает более точную характеристику центра выборки и в том случае, когда она состоит из ряда достаточно отдаленных друг от друга значений.

Решим следующую простую задачу.

Автомобиль половину пути проехал со скоростью ν_1 км/час, а вторую половину – со скоростью ν_2 км/час. Какой была средняя скорость автомобиля на этом пути?

Обозначим через x расстояние, равное половине пути. Тогда время, затраченное на первую половину пути равно $t_1 = \frac{x}{\nu_1}$, а на вторую – $t_2 = \frac{x}{\nu_2}$. Время всего движения составляет $t_1 + t_2 = \frac{x}{\nu_1} + \frac{x}{\nu_2}$. Значит, средняя скорость на всем пути равна:

$$\nu = \frac{2x}{t_1 + t_2} = \frac{2x}{\frac{x}{\nu_1} + \frac{x}{\nu_2}} = \frac{2}{\frac{1}{\nu_1} + \frac{1}{\nu_2}}.$$

Итак, средняя скорость движения определяется выражением:

$$\nu = \frac{2}{\frac{1}{\nu_1} + \frac{1}{\nu_2}}$$

В общем случае для подобных задач существует понятие **среднего гармонического.**

Определение 2.7 Средним гармоническим выборки $x_1, x_2, \ldots, x_n$ называется число $\overline{X}_{ãàðì}$, которое вычисляется по формуле:

$$\overline{X}_{ãàðì} = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \cdots + \frac{1}{x_n}}.$$

Другими словами, среднее гармоническое получается делением объема выборки n на сумму обратных чисел для выборочных значений.

Пример 2.17 За один час работы на компьютере первый студент набрал 2 страницы текста, второй – 3 станицы, третий – 6, четвертый – 12. Найдем среднее гармоническое числа страниц, которые может набрать на компьютере за один час студент этой группы.

Мы имеем выборку из четырех наблюдений:

$$2,\ \ 3,\ \ 16,\ \ 12.$$

Вычислим среднее гармоническое:

$$\overline{X}_{\text{ãàðì}} = \frac{4}{\frac{1}{2}+\frac{1}{3}+\frac{1}{6}+\frac{1}{12}} = 3{,}69.$$

Следовательно, можно предполагать, что в среднем студент этой группы может набрать 3,69 станицы за один час.

■

Определение 2.8 Средним гармоническим вариационного ряда x_1, x_2, …, x_n с соответствующими частотами $m_1, m_2, \ldots, m_k$ называется число $\overline{X}_{\text{ãàðì}}$, которое вычисляется по формуле:

$$\overline{X}_{\text{ãàðì}} = \frac{m_1 + m_2 + \cdots m_k}{\frac{m_1}{x_1} + \frac{m_2}{x_2} + \cdots + \frac{m_k}{x_k}}.$$

Пример 2.18 На решение одной задачи два студента тратят по 10 минут, 3 студента – по 20 минут, 5 студентов – по 30 минут, 4 студента – по 40 минут. Найдем среднее гармоническое времени, необходимого на решение одной задачи.

По данным задачи составим вариационный ряд.

Таблица 2.15 – Затраты времени студентов на решение задачи

Время	10	20	30	40
m_i	2	3	5	4

Вычислим среднее гармоническое:

$$\overline{X}_{\text{ãàðì}} = \frac{2+3+5+4}{\frac{2}{10}+\frac{3}{20}+\frac{5}{30}+\frac{4}{40}} = 22{,}7\,.$$

Итак среднее гармоническое времени, необходимого для решения одной задачи для этой группы студентов составляет 22,7 минут.

Мы видим, что $\overline{X}_{\text{ãàðì}} = 22{,}7$ меньше $\overline{X} = 27{,}8$.

■

Подчеркнем, что во всех случаях среднее гармоническое, среднее геометрическое и среднее арифметическое удовлетворяют неравенству:

$$\overline{X}_{\text{ãàðì}} \le \overline{X}_{\text{ãåîì}} \le \overline{\tilde{O}}\,.$$

Понятие среднего гармонического используется при исследовании некоторых физических, химических, биологических явлений, а также в экономике.

Мы рассмотрели несколько разных понятий выборочных характеристик среднего. Поэтому применение любого из них необходимо сопровождать дополнительным пояснением.

2.4 Выборочная дисперсия и стандартное отклонение

Выборочное среднее является важной, но не достаточной числовой характеристикой распределения исследуемой случайной величины. Любая случайная выборка состоит из индивидуальных значений, которые могут существенно отличаться и друг от друга, и от среднего. Некоторые значения могут располагаться близко к центру, а другие могут быть значительно отдалены от него. Очевидно, что экспериментальные выборочные значения характеризуются определенной степенью рассеяния вокруг среднего.

Степень различия между отдельными значениями генеральной совокупности или между выборочными значениями называется изменчивостью, или **вариацией.** Аналогичный смысл вкладывается в такие понятия, как рассеяние и разброс.

Рассмотрим три основных характеристики степени изменчивости статистических данных.

Самым простым показателем изменчивости является **вариационный размах.**

Определение 2.9 Вариационным размахом выборки x_1, x_2, …, x_n называется число R, равное разности между наибольшим и наименьшим значениями данной выборки:

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

Вариационный размах может называться одним словом – размах.

Пример 2.19 Рассмотрим метеорологические данные о дневной температуре воздуха за одну неделю наблюдений:

$$4^\circ, 1^\circ, -2^\circ, -7^\circ, -10^\circ, -16^\circ, -20^\circ.$$

Размах этой выборки легко находится:

$$\mathrm{R} = 4^\circ - \left(-20^\circ\right) = 24^\circ.$$

■

Фактически, размах дает максимальную величину отклонения между выборочными значениями.

Определение 2.10 Диапазоном наблюдений выборки x_1, x_2, …, x_n называется отрезок $[x_{\min}; x_{\max}]$, заключенный между минимальным выборочным значением $x_{\min}$ и максимальным $x_{\max}$.

Диапазон наблюдений содержит все выборочные значения. Например, отрезок $[-20^\circ; 4^\circ]$ является диапазоном наблюдений для выборки температуры воздуха за неделю наблюдений.

Заметим, что размах равен длине диапазона наблюдений.

Так как размах находится лишь по двум экстремальным выборочным значениям, то он не дает информации об изменчивости остальных наблюдений. Размах, в основном, используется для выборок небольшого объема, он дает слишком поверхностное представление об изменчивости исследуемого явления.

В отличие от размаха следующая числовая характеристика является показателем изменчивости внутри диапазона наблюдений. Рассмотрим отклонения всех выборочных значений $x_1, x_2, \ldots, x_n$ от среднего $\overline{X}$ этой выборки:

$$d_1 = x_1 - \overline{X}, \qquad d_2 = x_2 - \overline{X}, \qquad \ldots, \; d_n = x_n - \overline{X}.$$

Некоторые из этих отклонений являются положительными числами, а другие отрицательными, при этом сумма всех отклонений равна 0 для любой выборки. Заметим, что модуль отклонения $|d_i| = |x_i - \overline{X}|$ равен расстоянию между выборочным значением x_i и средним $\overline{X}$. Тогда сумма модулей отклонений учитывает все случайные выборочные значения и является положительным числом. Чем теснее выборочные значения группируются вокруг среднего, тем меньше эта сумма, и, наоборот, при широком разбросе выборочных значений сумма модулей отклонений увеличивается. Среднее значение суммы модулей отклонений характеризует усредненное расстояние выборочных значений от центра.

Определение 2.11 Средним абсолютным отклонением выборки x_1, x_2, …, x_n со средним $\overline{X}$ называется число $\overline{d}$, которое вычисляется по формуле:

$$\overline{d} = \frac{|x_1 - \overline{X}| + |x_2 - \overline{X}| + \cdots + |x_n - \overline{X}|}{n}.$$

В сокращенном виде данное выражение записывается так:

$$\overline{d} = \frac{\sum_{i=1}^{n} |x_i - \overline{X}|}{n}.$$

Среднее абсолютное отклонение, которое называется также **средним линейным отклонением**, является простой и полезной характеристикой степени рассеяния выборочных данных. К сожалению, из-за определенных неудобств при работе с модулями величин это понятие не используется в теоретической статистике.

Пример 2.20 Рассмотрим выборочные данные о годовой стоимости обучения в восьми вузах города:

900, 1200, 1500, 1700, 1800 2100, 2400, 2800.

Найдем среднее абсолютное отклонение этой выборки. Прежде всего, вычислим среднее:

$$\overline{\mathrm{X}} = \frac{900+1200+1500+1700+1800+2100+2400+2800}{8} = 1800\,.$$

Определим отклонения всех выборочных значений:

$$d_1 = 900-1800 = -900; \qquad d_5 = 1800-1800 = 0;$$
$$d_2 = 1200-1800 = -600; \qquad d_6 = 2100-1800 = 300;$$
$$d_3 = 1500-1800 = -300; \qquad d_7 = 2400-1800 = 600;$$
$$d_4 = 1700-1800 = -100; \qquad d_8 = 2800-1800 = 1000.$$

Для проверки правильности расчетов можно использовать равенство:

$$d_1 + d_2 + \cdots d_n = 0.$$

Теперь вычислим:

$$\overline{d} = \frac{|-900|+|-600|+|-300|+|-100|+|0|+|300|+|600|+|1000|}{8} = 475.$$

Итак, среднее абсолютное отклонение стоимости обучения в данных вузах равно 475. Заметим, что реальные отклонения могут быть меньше или больше среднего отклонения.

■

Основными характеристиками степени рассеяния выборочных данных являются дисперсия и стандартное отклонения.

Определение 2.12 Дисперсией выборки $x_1, x_2, \ldots, x_n$ называется число $\mathbf{S}^2$, которое вычисляется по формуле:

$$\mathbf{S}^2 = \frac{\left(x_1 - \overline{\mathbf{X}}\right)^2 + \left(x_2 - \overline{\mathbf{X}}\right)^2 + \cdots + \left(x_n - \overline{\mathbf{X}}\right)^2}{\mathbf{n}-1}$$

при малом объеме выборки ($\mathbf{n} \le 30$) и

$$\mathbf{S}^2 = \frac{\left(x_1 - \overline{\mathbf{X}}\right)^2 + \left(x_2 - \overline{\mathbf{X}}\right)^2 + \cdots + \left(x_n - \overline{\mathbf{X}}\right)^2}{\mathrm{n}}$$

при большом объеме выборки (n > 30).

Сокращенно формулы записываются в таком виде:

$$\mathrm{S}^2 = \frac{\sum\limits_{i=1}^{n}\left(x_i - \overline{\mathrm{X}}\right)^2}{\mathrm{n}-1} \quad \text{или} \quad \mathrm{S}^2 = \frac{\sum\limits_{i=1}^{n}\left(x_i - \overline{\mathrm{X}}\right)^2}{\mathrm{n}}$$

Выборочная дисперсия при малых значениях объема $\mathrm{n} \leq 30$ и при больших значениях $\mathrm{n} > 30$ вычисляется по разным формулам. Замена делителя n на n – 1 для выборок малых объемов устраняет систематическую ошибку, или «смещение» относительно дисперсии всей генеральной совокупности. Исключение систематической ошибки – это одно из необходимых условий получения правильной оценки любой числовой характеристики генеральной совокупности.

Определение 2.13 Стандартным отклонением выборки x_1, x_2, …, x_n называется число S, которое вычисляется по формуле:

$$\mathbf{S} = \sqrt{\mathbf{S}^2}.$$

Таким образом, выборочное стандартное отклонение равно квадратному корню из выборочной дисперсии, следовательно, справедливы формулы:

$$\mathrm{S} = \sqrt{\frac{\sum\limits_{i=1}^{n}\left(x_i - \overline{\mathrm{X}}\right)^2}{\mathrm{n}-1}} \quad \text{либо} \quad \mathrm{S} = \sqrt{\frac{\sum\limits_{i=1}^{n}\left(x_i - \overline{\mathrm{X}}\right)^2}{\mathrm{n}}}$$

Пример 2.21 В течение пяти дней студент Ковалев записывал стоимость обедов в студенческой столовой: 3,2; 4,8; 5,6; 4,5; 5,4. Найдем выборочную дисперсию и стандартное отклонение.

Сначала определим среднее:

$$\overline{\mathrm{X}} = \frac{3{,}2 + 4{,}8 + 5{,}6 + 4{,}5 + 5{,}4}{5} = 4{,}7.$$

Вычислим дисперсию:

$$S^2 = \frac{(3{,}2-4{,}7)^2 + (4{,}8-4{,}7)^2 + (5{,}6-4{,}7) + (4{,}5-4{,}7)^2 + (5{,}4-4{,}7)^2}{5-1} =$$

$$= \frac{2{,}25 + 0{,}01 + 0{,}81 + 0{,}04 + 0{,}49}{4} = \frac{3{,}6}{4} = 0{,}9.$$

Найдем стандартное отклонение:

$$S = \sqrt{0{,}9} = 0{,}94868\,.$$

Округлим полученное значение: S = 0,95 условных рублей.

■

Определение 2.14 Выборочной дисперсией вариационного ряда $x_1, x_2, \ldots, x_n$ с соответствующими частотами $m_1, m_2, ..., m_k$ называется число S^2, определяемое формулой:

$$\mathbf{S^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{X})^2 m_i}{n-1}} \quad \textbf{или} \quad \mathbf{S^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{X})^2 m_i}{n}}$$

соответственно, при малом и большом значении n, где $n = m_1 + m_2 + \cdots + m_k$**.**

Пример 2.22 Для социологического исследования были собраны данные о количественном составе 20 семей, приведенные в следующей таблице.

Таблица 2.16 – Количественный состав семей

Количество членов семьи	1	2	3	4	5	6
m_i	2	3	8	5	1	1

Найдем среднее, дисперсию и стандартное отклонение:

$$n = 2 + 3 + 8 + 5 + 1 + 1 = 20;$$

$$\overline{X} = \frac{1\cdot 2 + 2\cdot 3 + 3\cdot 8 + 4\cdot 5 + 5\cdot 1 + 6\cdot 1}{20} = 3{,}15;$$

$$S^2 = \frac{(1-3{,}15)^2 \cdot 2 + (2-3{,}15)^2 \cdot 3 + (3-3{,}15)^2 \cdot 8 + (4-3{,}15)^2 \cdot 5 + (5-3{,}15)^2 \cdot 1 + (6-3{,}15)^2 \cdot 1}{20-1} =$$

$$= \frac{9{,}2450 + 3{,}9675 + 0{,}1800 + 3{,}6125 + 3{,}4225 + 8{,}1225}{19} = \frac{28{,}5500}{19} = 1{,}5026$$

$$S = \sqrt{S^2} = \sqrt{1{,}5026} = 1{,}2258.$$

Округлим $S^2 = 1{,}50$ и $S = 1{,}23$. Итак, $\overline{X} = 3{,}15$ – это среднее число членов семьи, $S = 1{,}23$ – это стандартное отклонение от среднего.

■

Определение 2.15 Выборочной дисперсией статистического ряда, состоящего из k интервалов с соответствующими интервальными средними $\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_k$ и интервальными частотами $m_1, m_2, ..., m_k$, называется число S^2, равное:

$$S^2 = \frac{\sum_{i=1}^{k} (\tilde{x}_i - \overline{X})^2 m_i}{n-1} \quad \text{или} \quad S^2 = \frac{\sum_{i=1}^{k} (\tilde{x}_i - \overline{X})^2 m_i}{n},$$

соответственно, при малом и большом значении n, где $n = m_1 + m_2 + \cdots + m_k$.

Пример 2.23 Результаты экзамена по высшей математике пятидесяти студентов представлены следующим статистическим рядом. Используется десятибалльная система оценок. Найдем среднее и стандартное отклонение.

Таблица 2.17 – Итоги экзамена по высшей математике

Оценка	0–2	2–4	4–6	6–8	8–1
m_i	3	9	16	14	8

Итак, $n = 3 + 9 + 16 + 14 + 8 = 50$.

Найдем интервальные средние:

$\tilde{x}_1 = 1, \quad \tilde{x}_2 = 3, \quad \tilde{x}_3 = 5, \quad \tilde{x}_4 = 7, \quad \tilde{x}_5 = 9.$

Вычислим среднее:

$$\overline{X} = \frac{1 \cdot 3 + 3 \cdot 9 + 5 \cdot 16 + 7 \cdot 14 + 9 \cdot 8}{50} = 5{,}6.$$

Найдем дисперсию данной выборки:

$$S^2 = \frac{(1-5{,}6)^2 \cdot 3 + (3-5{,}6)^2 \cdot 9 + (5-5{,}6)^2 \cdot 16 + (7-5{,}6)^2 \cdot 14 + (9-5{,}6)^2 \cdot 8}{49} = \frac{250}{49} = 5{,}102$$

Определим значение стандартного отклонения:

$$S = \sqrt{5{,}102} = 2{,}259.$$

Итак, средняя оценка студентов I курса составляет 5,6 баллов. Стандартное отклонение $S = 2{,}26$ баллов показывает, что оценки большинства студентов отличаются от среднего не более, чем на 2,26 баллов.

■

Таким образом, для вычисления выборочной дисперсии необходимо найти значение среднего $\overline{X}$, вычислить сумму квадратов отклонений выборочный значений от среднего и разделить ее на $n - 1$, где n – число всех наблюдений. Извлечение квадратного корня при нахождении стандартного отклонения возвращает к первоначальному масштабу единицы измерения.

Обработка и анализ статических данных требует кропотливой и нелегкой вычислительной работы. Для организации вычислений в математической статистике часто используются специальные таблицы.

Пример 2.24 Найдем среднее и стандартное отклонение для статистического ряда из примера 1.4 о высоте городских зданий. Все необходимые вычисления будем записывать в таблицу 2.18.

Из таблицы 2.18 берем необходимые промежуточные результаты:

$$n = 40; \qquad \overline{X} = \frac{1085}{40} = 27{,}125;$$

$$S^2 = \frac{3869{,}3747}{39} = 99{,}214735; \qquad S = \sqrt{99{,}214735} = 9{,}9606593.$$

Итак, среднее высоты зданий равно 27,12 метров, а стандартное отклонение S равно 9,96 метров.

Таблица 2.18 – Вычисление среднего и стандартного отклонения высоты зданий

Высота	Интервальное среднее $\widetilde{x}_i$	Частота m_i	$\widetilde{x}_i m_i$	$\widetilde{x}_i - \overline{X}$ $(\widetilde{x}_i - 27{,}125)$	$(\widetilde{x}_i - \overline{X})^2$	$(\widetilde{x}_i - \overline{X})^2 m_i$
5–10	7,5	2	15	-19,625	385,14062	770,28124
10–15	12,5	3	37,5	-14,625	213,89062	641,67186
15–20	17,5	5	87,5	-9,625	92,64065	463,20312
20–25	22,5	6	135	-4,625	21,390625	128,34375
25–30	27,5	8	220	0,375	0,140625	1,12500
30–35	32,5	7	227,5	5,375	28,890625	202,23437
35–40	37,5	5	187,5	10,375	107,64062	538,20310
40–45	42,5	3	127,5	15,375	236,39062	709,17186
35–50	47,5	1	47,5	20,375	415,14062	415,14062
Сумма		40	1085			3869,3447

■

Мы округлили полученные значения $\overline{X}$ и S так, чтобы они были соизмеримы с наблюдаемыми значениями.

В математической статистике принято соблюдать два **правила округления результатов**:

– округлению подвергаются только значения результирующих показателей. Промежуточные значения не округляются;

– конечные значения округляются так, чтобы оставалось на одну (две) значащие цифры больше, чем в первоначальных данных.

Если в выражении, определяющем дисперсию, выполнить следующее преобразование

$$(x_i - \overline{X})^2 = x_i^2 - 2x_i\overline{X} + (\overline{X})^2$$

то получится другая эквивалентная формула, которая помогает облегчать вычисление дисперсии.

Теорема 2.3 **Дисперсия выборки $x_1, x_2, \ldots, x_n$ вычисляется по формуле:**

$$S^2 = \frac{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}{n-1}.$$

Дисперсия вариационного ряда $x_1, x_2, \ldots, x_n$ с соответствующими частотами $m_1, m_2, \ldots, m_k$ вычисляется по формуле:

$$S^2 = \frac{\sum_{i=1}^{k} x_i^2 m_i - \frac{\left(\sum_{i=1}^{k} x_i m_i\right)^2}{n}}{n-1}, \quad \text{ãääå} \quad n = \sum_{i=1}^{k} m_i .$$

Дисперсия статистического ряда с соответствующими интервальными средними $\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_n$ и частотами $m_1, m_2, \ldots, m_k$ вычисляется по формуле:

$$S^2 = \frac{\sum_{i=1}^{k} \tilde{x}_i^2 m_i - \frac{\left(\sum_{i=1}^{k} \tilde{x}_i m_i\right)^2}{n}}{n-1}, \quad \text{ãääå} \quad n = \sum_{i=1}^{k} m_i .$$

Заметим, что ни одна из этих формул не требует предварительного вычисления среднего $\overline{X}$. Все данные формулы используются при малых объемах выборок, при больших объемах делитель n – 1 заменяется на n. Они не только облегчают вычислительную работу, но и дают более точный результат в тех случаях, когда при нахождении среднего делаются округления.

Пример 2.25 Вычислим среднее и стандартное отклонение для статистического ряда из примера 1.7 о возрасте пациентов поликлиники. Необходимые расчеты будем записывать в следующей таблице:

Таблица 2.19 – Вычисление среднего и стандартного отклонения возраста пациентов поликлиники.

Возраст	Интервальное среднее $\widetilde{x}_i$	Частота m_i	$\widetilde{x}_i m_i$	$\widetilde{x}_i^2 m_i$
10–20	15	17	255	3825
20–30	25	24	600	15000
30–40	35	35	1225	42875
40–50	45	48	2160	97200
50–60	55	57	3135	172425
60–70	65	42	2730	177450
70–80	75	21	1575	118125
80–90	85	6	510	43350
		250	12190	670250

Используя суммы столбцов, получим:

$$\overline{X} = \frac{12190}{250} = 48{,}76;$$

$$S^2 = \frac{670250 - \frac{(12190)^2}{250}}{249} = 304{,}68112;$$

$$S = \sqrt{304{,}68112} = 17{,}455117.$$

Округлим полученные значения:

$$\overline{X} = 48{,}8; \quad S = 17{,}5.$$

Таким образом, средний возраст пациентов поликлиники равен 48,8 лет, стандартное отклонение равно 17,4 лет.

■

В том случае, когда обследованию подвергается вся генеральная совокупность значений исследуемой случайной величины, то выборочная дисперсия генеральной совокупности совпадает с

теоретической дисперсией исследуемой случайной величины X, которая определяется формулой:

$$DX = M(X - MX)^2.$$

Далее для обозначения дисперсии генеральной совокупности мы будем использовать обозначение $\sigma^2 = DX$, а стандартное отклонение генеральной совокупности будем обозначать через $\sigma = \sqrt{\sigma^2}$. Среднее μ и стандартное отклонение σ генеральной совокупности в основном используются в теоретической части математической статистики. Подчеркнем, что выборочное стандартное отклонение S всегда больше теоретического стандартного отклонения σ. Однако, при увеличении объема выборки различие между ними уменьшается.

Следует отметить, что вместо термина стандартное отклонение часто используются такие названия этого же понятия, как **среднее квадратическое отклонение** или **среднее квадратичное отклонение**.

Еще раз подчеркнем, что стандартное отклонение характеризует степень случайного рассеяния выборочных значений вокруг среднего. Чем меньше значение S, тем ближе разбросаны выборочные данные вокруг среднего $\overline{X}$. В предельном случае, когда $S = 0$, случайное рассеяние отсутствует, так как из равенства

$$(x_1 - \overline{X})^2 + (x_2 - \overline{X})^2 + \cdots + (x_n - \overline{X})^2 = 0$$

следует, что $x_1 = x_2 = \cdots = x_n = \overline{X}$, то есть случайная величина является константой.

Правомерность использования стандартного отклонения σ в качестве меры рассеяния конкретных значений случайной величины X вокруг среднего μ теоретически подтверждается известным неравенством Чебышева:

для любой случайной величины X, имеющей конечную дисперсию, при каждом $\varepsilon > 0$ справедливо неравенство

$$P(|X - \mu| \le \varepsilon) \ge 1 - \frac{\sigma^2}{\varepsilon^2}$$

В частном случае, когда $\varepsilon = k\sigma$, где k – целое число большее 1, имеет место следующее неравенство

$$P\left(\left|X-\mu\right|\le k\sigma\right)\ge 1-\frac{1}{\sigma^2}.$$

Отсюда при $k=2$ и $k=3$ получаются следствия:

$$P\left(\left|X-\mu\right|\le 2\sigma\right)\ge 1-\frac{1}{2^2}=\frac{3}{4}=0{,}75;.$$

$$P\left(\left|X-\mu\right|\le 3\sigma\right)\ge 1-\frac{1}{3^2}=\frac{8}{9}=0{,}89.$$

Таким образом, не менее 75 % значений случайной величины имеют отклонение от среднего μ, не превышающее двух стандартных отклонений 2σ, и не менее 89 % значений отличаются от среднего не более чем на три стандартных отклонения 3σ.

Конкретные значения случайных величин с нормальным распределением имеют еще более выраженную центральную тенденцию. Многочисленные статистические исследования стали основанием для следующего утверждения, подходящего для многих реальных выборок.

Если $\overline{X}$ – среднее, а S – стандартное отклонение выборки $x_1, x_2, \ldots, x_n$, то в интервале $\left(\overline{X}-S;\ \overline{X}+S\right)$ содержится около 68 % выборочных значений, в интервале $\left(\overline{X}-2S;\ \overline{X}+2S\right)$ содержится около 95 % выборочных значений, в интервале $\left(\overline{X}-3S;\ \overline{X}+3S\right)$ содержится около 99,7 % выборочных значений.

Другими словами, около 68 % выборочных значений имеют отклонение от среднего, не превышающее одного стандартного отклонения S, около 95 % выборочных значений имеют отклонения не более 2S, а между значениями $\overline{X}-3S$ и $\overline{X}+3S$ находится около 99,7 % выборки.

Пример 2.26 В примере 2.24 мы нашли $\overline{X}=27{,}125$ метров и $S=9{,}96$ метров для высоты 40 зданий.

Рассмотрим интервал $\left(\overline{X}-S;\ \overline{X}+S\right)=\left(17{,}165;\ 37{,}085\right)$. Вернемся к первоначальным не сгруппированным данным примера 1.4 и подсчитаем, сколько выборочных значений содержится в этом интервале. Их оказалось 27, что составляет 67,5 % от объема всей выборки.

Рассмотрим интервал $(\overline{X} - 2S; \overline{X} + 2S) = (7{,}205; 47{,}045)$, он содержит 39 выборочных значений, что составляет 97,5 % от всей выборки. И только одно значение не попало в этот интервал. В интервал $(\overline{X} - 3S; \overline{X} + 3S)$ попадает вся выборка.

■

В любом случае полученные по конкретным выборкам значения среднего $\overline{X}$ и стандартного отклонения S не совпадают с соответствующими значениями среднего μ и стандартного отклонения σ всей генеральной совокупности. С одной стороны возникает вопрос о степени согласованности между выборочными и теоретическими характеристиками исследуемой случайной величины, с другой – об оценке достоверности характеристик, найденных по конкретной единичной случайной выборке. Для каждого параметра распределения существуют так называемые **стандартные ошибки,** которые дают возможность по результатам одной выборки оценивать параметры других выборок исследуемой совокупности. Так, например, **стандартной ошибкой среднего $\overline{X}$** называется отношение стандартного отклонения S к $\sqrt{n}$, где n – объем выборки, то есть величину $S_{\overline{X}}$, равную

$$S_{\overline{X}} = \frac{S}{\sqrt{n}}.$$

Очевидно, что значение стандартной ошибки существенно зависит от объема выборки n. Величина стандартной ошибки приблизительно показывает, насколько среднее $\overline{X}$ одной выборки отличается от средних других выборок объема n из исследуемой генеральной совокупности.

Стандартная ошибка $S_{\overline{X}}$ всегда меньше стандартного отклонения S, которое характеризует изменчивость отдельных значений относительно среднего $\overline{X}$ внутри одной выборки.

Степень рассеяния выборочных значений случайной величины также показывает следующая относительная характеристики.

Определение 2.16 Коэффициентом вариации выборки x_1, x_2, …, x_n называется отношение её стандартного отклонения S к среднему $\overline{X}$:

$$V = \frac{S}{\overline{X}}, \quad \text{где} \quad \overline{X} \neq 0.$$

Определение 2.17 Коэффициентом вариации случайной величины X называется отношение её стандартного отклонения σ к математическому ожиданию μ:

$$V_{\text{òåîð}} = \frac{\sigma}{\mu}, \quad \text{где} \quad \mu \neq 0.$$

Заметим, что при $\overline{X} = 1$ и $\mu = 1$ получаем, соответственно, что $V = S$ и $V_{\text{теор}} = \sigma$. Часто выборочный и теоретический коэффициенты вариации задаются в процентах:

$$V = 100 \cdot \frac{S}{\overline{X}} \ \% \quad \text{и} \quad V_{\text{òåîð}} = 100 \cdot \frac{\sigma}{MX}.$$

Во всех случаях коэффициент вариации является безразмерной относительной характеристикой рассеяния значений случайной величины, которая используется для сравнения нескольких выборок из генеральной совокупности одного типа.

Пример 2.27 Найдем коэффициент вариации высот городских зданий по данным примера 2.24.

Ранее мы нашли, что $\overline{X} = 27{,}125$ и $S = 9{,}961$. Тогда выборочный коэффициент вариации $V = \frac{S}{\overline{X}} = \frac{9{,}961}{27{,}125} = 0{,}367$, что составляет около 37 %.

■

2.5 Выборочные и теоретические моменты распределения

Рассмотрим важные выборочные числовые характеристики распределений, обобщающие понятия среднего и дисперсии. Пусть k – целое неотрицательное число.

Определение 2.18 Начальным моментом k-го порядка выборки $x_1, x_2, \ldots, x_n$ называется среднее k-тых степеней данных выборочных значений, то есть

$$\widehat{\nu}_k = \frac{1}{n}\sum_{i=1}^{n} x_i^k.$$

Очевидно, что начальный выборочный момент $\widehat{\nu}_0$ нулевого порядка всегда равен 1, а начальный выборочный момент первого порядка $\widehat{\nu}_1 = \overline{X}.$

Определение 2.19 Центральным моментом k-го порядка выборки $x_1, x_2, \ldots, x_n$ называется среднее k-тых степеней отклонений данных выборочных значений от среднего $\overline{X}$, то есть

$$\widehat{\mu}_k = \frac{1}{n}\sum_{i=1}^{n} (x_i - \overline{X})^k.$$

Из данного определения следует, что центральный выборочный момент $\widehat{\mu}_0$ нулевого порядка равен 1. При k = 1 получается, что

$$\widehat{\mu}_1 = \frac{1}{n}\sum_{i=1}^{n} (x_i - \overline{X}) = 0,$$

а при k = 2 имеем

$$\widehat{\mu}_2 = \frac{1}{n}\sum_{i=1}^{n} (x_i - \overline{X})^2 = S^2.$$

Следовательно, выборочная дисперсия S^2 является центральным выборочным моментом второго порядка. Для вычисления центрального выборочного момента третьего порядка используем стандартные алгебраические преобразования:

$$\widehat{\mu}_3 = \frac{1}{n}\sum_{i=1}^{n} (x_i - \overline{X})^3 = \frac{1}{n}\sum_{i=1}^{n} (x_i^3 - 3x_i^2\overline{X} + 3x_i(\overline{X})^2 - (\overline{X})^3) =$$

$$= \frac{1}{n}\sum_{i=1}^{n} x_i^3 - 3\overline{X}\cdot\frac{1}{n}\sum_{i=1}^{n} x_i^2 + 3(\overline{X})^2\cdot\frac{1}{n}\sum_{i=1}^{n} x_i - (\overline{X})^3 =$$

$$= \widehat{\nu}_3 - 3\widehat{\nu}_1\cdot\widehat{\nu}_2 + 3\widehat{\nu}_1^3 - \widehat{\nu}_1^3 = \widehat{\nu}_3 - 3\widehat{\nu}_1\widehat{\nu}_2 + 2\widehat{\nu}_1^3.$$

В результате получилось выражение центрального момента третьего порядка через начальные моменты. Таким же способом находятся выражения для центральных моментов более высоких порядков. Приведем ряд формул, которые на практике используются чаще других:

$$\hat{\mu}_2 = \hat{\nu}_2 - \hat{\nu}_1^2;$$

$$\hat{\mu}_3 = \hat{\nu}_3 - 3\hat{\nu}_1\hat{\nu}_2 + 2\hat{\nu}_1^3;$$

$$\hat{\mu}_4 = \hat{\nu}_4 - 4\hat{\nu}_1\hat{\nu}_3 + 6\hat{\nu}_1^2\hat{\nu}_2 - 3\hat{\nu}_4.$$

При вычислении начальных и центральных выборочных моментов используются приемы и таблицы, аналогичные тем, которые применялись ранее для вычисления среднего $\overline{X}$ и дисперсии S^2.

Пример 2.28 В ходе социологического исследования собраны ответы 25 рядовых сотрудников учреждения о количестве стрессовых ситуаций, возникавших на работе в течение недели. Данные опроса приведены в следующей таблице. Найдем начальные и центральные выборочные моменты первого, второго, третьего и четвертого порядков.

Таблица 2.20 – Данные исследования стрессовых ситуаций

Количество стрессов	0	1	2	3	4	5
m_i	1	2	8	10	2	2

Необходимые промежуточные расчеты будем фиксировать в следующей таблице.

Таблица 2.21 – Вычисления начальных и центральных моментов

x_i	m_i	$x_i m_i$	$x_i^2 m_i$	$x_i^3 m_i$	$x_i^4 m_i$
0	1	0	0	0	0
1	2	2	2	2	2
2	8	16	32	64	128
3	10	30	90	270	810
4	2	8	32	128	512
5	2	10	50	250	1250
	25	66	206	714	2702

Объем выборки n = 25. Вычислим начальные выборочные моменты:

$$\hat{\nu}_1 = \frac{66}{25} = 2{,}64; \qquad \hat{\nu}_2 = \frac{206}{25} = 8{,}24;$$

$$\hat{\nu}_3 = \frac{714}{25} = 28{,}56; \qquad \hat{\nu}_4 = \frac{2702}{25} = 108{,}08.$$

Используя соответствующие формулы, вычислим центральные выборочные моменты:

$$\hat{\mu}_1 = 0; \quad \hat{\mu}_2 = \hat{\nu}_2 - \hat{\nu}_1^2 = 8{,}24 - (2{,}64)^2 = 1{,}2704;$$

$$\hat{\mu}_3 = \hat{\nu}_3 - 3\hat{\nu}_1\hat{\nu}_2 + 2\hat{\nu}_1^3 = 28{,}56 - 3 \cdot 2{,}64 \cdot 8{,}24 + 2(2{,}64)^3 = 0{,}098688;$$

$$\hat{\mu}_4 = \hat{\nu}_4 - 4\hat{\nu}_1\hat{\nu}_3 + 6\hat{\nu}_1^2\hat{\nu}_2 - 3\hat{\nu}_1^4 =$$

$$= 108{,}08 - 4 \cdot 2{,}64 \cdot 28{,}56 + 6 \cdot (2{,}64)^2 \cdot 8{,}24 - 3(2{,}64)^4 = 5{,}33745.$$

Округлим полученные значения центральных моментов:

$$\hat{\mu}_1 = 0; \qquad \hat{\mu}_2 = 1{,}27; \qquad \hat{\mu}_3 = 0{,}10; \qquad \hat{\mu}_4 = 5{,}34.$$

■

Начальные и центральные выборочные моменты являются аналогами соответствующих понятий теоретических моментов всей генеральной совокупности значений исследуемой случайной величины.

Определение 2.20 Начальным моментом k-го порядка случайной величины X называется число ν_k, равное математическому ожиданию k-й степени величины X:

$$\nu_k = \mathrm{M}(\mathbf{X}^k).$$

Для вычисления начального момента k-го порядка используются следующие формулы:

$$\nu_k = \begin{cases} \sum_{i=1}^{\infty} (x_i)^k p_i, & \text{åñëè } X \text{ äèñêðåòíàÿ ;} \\ \int_{-\infty}^{\infty} (x - MX)^k p_x(x)dx, & \text{åñëè } X \text{ íåïðåðûâíà ÿ.} \end{cases}$$

Говорят, что момент ν_k существует, если он конечен, в противном случае считается, что момент не существует.

Определение 2.21 Центральным моментом k-го порядка случайной величины X называется число μ_k, равное математическому ожиданию величины

$$\mu_k = \mathbf{M(X - MX)^k}.$$

Для вычисления центрального момента k–го порядка используются формулы:

$$\mu_k = \begin{cases} \sum_{i=1}^{\infty} (x_i - MX)^k p_i, & \text{åñëè } X \text{ äèñêðåòíàÿ ;} \\ \int_{-\infty}^{\infty} (x - MX)^k p_x(x)dx, & \text{åñëè } X \text{ íåïðåðûâíà ÿ.} \end{cases}$$

Заметим, что формулы, выражающие центральные моменты через начальные, аналогичны соответствующим формулам для выборочных моментов. В частности, имеют место соотношения:

$$\mu_1 = 0;$$

$$\mu_2 = \nu_2 - \nu_1^2;$$

$$\mu_3 = \nu_3 - 3\nu_1\nu_2 + 2\nu_1^3;$$

$$\mu_4 = \nu_4 - 4\nu_1\nu_3 + 6\nu_1^2\nu_2 - 3\nu_1^4.$$

Очевидно, что математическое ожидание случайной величины является начальным моментом первого порядка, а дисперсия – центральным моментом второго порядка. Как теоретические, так и выборочные моменты используются при исследовании закона распределения случайной величины. Все центральные моменты четных порядков, как и дисперсия, характеризуют рассеяние значений случайной величины вокруг математического ожидания.

Центральные моменты нечетных порядков выявляют асимметрию распределения относительно центра. В частности, если значения случайной величины распределены симметрично относительно математического ожидания, то все ее существующие моменты нечетных порядков равны нулю. С другой стороны, существование отличного от нуля центрального момента нечетного порядка показывает наличие асимметрии распределения.

2.6 Асимметрия и эксцесс

В математической статистике для выяснения геометрической формы плотности вероятности случайной величины используются две числовые характеристики, связанные с центральными моментами третьего и четвертого порядков.

Определение 2.22 Коэффициентом асимметрии выборки x_1, x_2, …, x_n называется число $\hat{\gamma}$, равное отношению центрального выборочного момента третьего порядка $\hat{\mu}_3$ к кубу стандартного отклонения S:

$$\hat{\gamma} = \frac{\hat{\mu}_3}{\mathbf{S}^3}.$$

Так как $\mathrm{S} = \sqrt{\hat{\mu}_2}$ и $\mathrm{S}^3 = \sqrt{\hat{\mu}_2^3}$, то коэффициент асимметрии выражается через центральные моменты следующей формулой:

$$\hat{\gamma} = \frac{\hat{\mu}_3}{\sqrt{\hat{\mu}_2^3}}.$$

Отсюда получается формула, выражающая коэффициент асимметрии через начальные моменты:

$$\hat{\gamma} = \frac{\hat{\nu}_3 - 3\hat{\nu}_1\hat{\nu}_2 + 2\hat{\nu}_1^3}{\sqrt{(\hat{\nu}_2 - \hat{\nu}_1^2)^3}},$$

которая облегчает практические вычисления.

Соответствующая теоретическая характеристика вводится

с помощью теоретических моментов.

Определение 2.23 Коэффициентом асимметрии случайной величины X называется число γ_1 равное отношению центрального момента третьего порядка μ_3 к кубу стандартного отклонения σ:

$$\gamma = \frac{\mu_3}{\sigma^3}.$$

Если случайная величина X имеет симметричное распределение относительно математического ожидания μ, то её теоретический коэффициент асимметрии равен 0, если же распределение вероятностей несимметрично, то коэффициент асимметрии отличен от нуля. Положительное значение коэффициента асимметрии говорит о том, что большая часть значений случайной величины расположена правее математического ожидания, то есть правая ветвь кривой плотности вероятности более удлинена, чем левая. Отрицательное значение коэффициента асимметрии говорит о том, что более длинная часть кривой расположена слева. Данное утверждение иллюстрирует следующий рисунок.

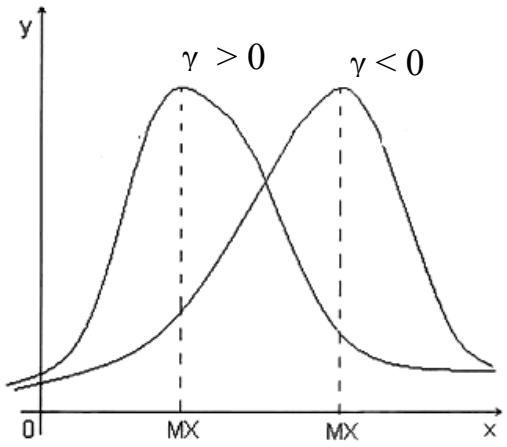


Рисунок 2.1 – Положительная и отрицательная асимметрия распределений

Пример 2.29 Найдем выборочный коэффициент асимметрии по данным исследования стрессовых ситуаций из примера 2.28.

Пользуясь ранее вычисленными значениями центральных

выборочных моментов, получим

$$\hat{\gamma} = \frac{0{,}098688}{\sqrt{(1{,}2704)^3}} = 0{,}0689213.$$

Округлим $\hat{\gamma}$ = 0,07. Найденное отличное от нуля значение коэффициента асимметрии показывает скошенность распределения относительно среднего. Положительное значение $\hat{\gamma}$ говорит о том, что более длинная ветвь кривой плотности вероятности расположена справа.

■

Особенности распределения значений случайной величины вокруг её модального значения $X_{мод}$ характеризует следующая постоянная.

Определение 2.24 Эксцессом выборки x_1, x_2, …, x_n называется число $\hat{e}$, равное

$$\hat{e} = \frac{\hat{\mu}_4}{\mathbf{S}^4} - 3,$$

где $\hat{\mu}_4$ – выборочный центральный момент четвёртого порядка, $\mathbf{S}^4$ – четвёртая степень стандартного отклонения S.

Теоретическое понятие эксцесса является аналогом выборочного.

Определение 2.25 Эксцессом случайной величины X называется число *e*, равное

$$e = \frac{\mu_4}{\sigma^4} - 3,$$

где μ_4 – теоретический центральный момент четвёртого порядка, σ^4 – четвёртая степень стандартного отклонения σ.

Значение эксцесса *e* характеризует относительную крутость вершины кривой плотности распределения вокруг точки максимума. Если эксцесс является положительным числом, то соответствующая кривая распределения имеет более острую вершину.

Распределение с отрицательным эксцессом имеет сглаженную и более плоскую вершину. Следующий рисунок иллюстрирует возможные случаи.

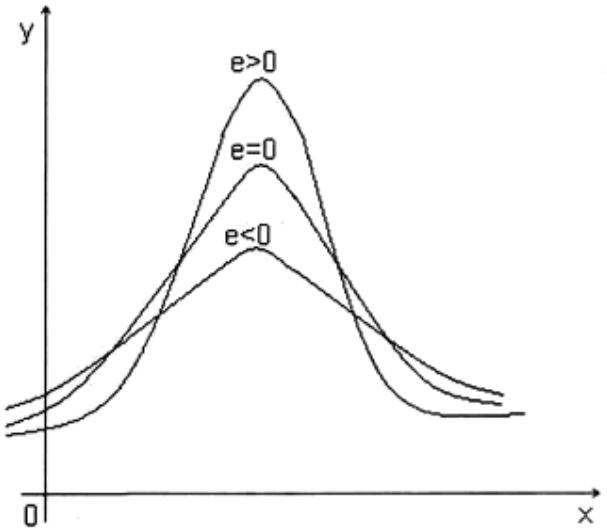


Рисунок 2.2 – Распределения с положительным, нулевым и отрицательным значениями эксцессов

Пример 2.30 Вычислим значение выборочного эксцесса по данным исследования стрессовых ситуаций примера 2.28.

Возьмем найденные ранее значения центральных выборочных моментов

$\hat{\mu}_4 = 5{,}33745$ и $\hat{\mu}_2 = 1{,}2704$.

Так как $\hat{\mu}_2 = S^2$, то $\hat{\mu}_2^2 = S^4$. Следовательно

$$\hat{e} = \frac{\hat{\mu}_4}{S^4} - 3 = \frac{\hat{\mu}_4}{\hat{\mu}_2^2} - 3 = \frac{5{,}33745}{(1{,}2704)^2} - 3 = 0{,}3071431.$$

После округления $\hat{e} = 0{,}31$. Положительное значение эксцесса указывает на более острую вершину кривой плотности вероятности.

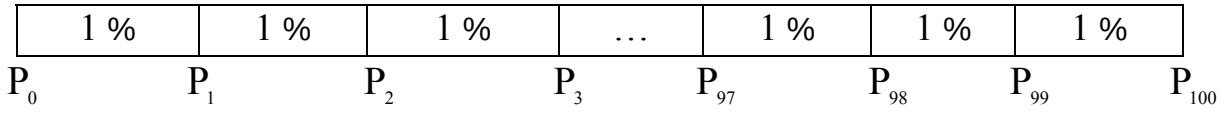
■

Отметим, что коэффициент асимметрии $\hat{\gamma}$ и эксцесс $\hat{e}$ вместе с $\overline{X}$ и стандартным отклонением S являются важными числовыми характеристиками закона распределения исследуемой величины.

2.7 Процентные точки и квантили распределения

Существуют определенные числовые характеристики, которые описывают месторасположение одной части выборки относительно остальной части упорядоченной выборки. Рассмотрим понятие **процентиля**, с помощью которого локализуются позиции наблюдений относительно всей выборки.

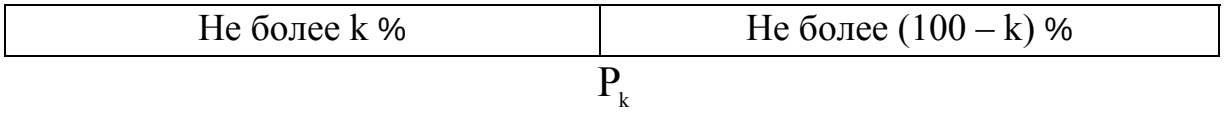
Процентили – это числа, обозначаемые через $P_0, P_1, \ldots, P_{100}$, которые делят исследуемую выборку на 100 равных частей.



Определение 2.26 k-ым процентилем упорядоченной выборки ($0 \le k \le 100$) называется число P_k, удовлетворяющее двум условиям:

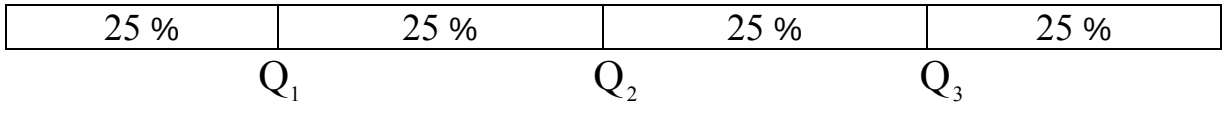
1) в выборке имеется не более k % значений, меньших числа P_k;

2) в выборке имеется не более $(100 - k)$ % значений больших числа P_k.



Например, если двадцатый процентиль $P_{20} = 14$, то в выборке содержится не более 20 % значений, меньших 14, и не более (100 – 20) % = 80 % значений, больших 14. Очевидно, что 50-й процентиль P_{50} совпадает с медианой выборки: $P_{50} = X_{\text{мед}}$. Нулевой процентиль P_0 является наименьшим выборочным значением. Сотый процентиль P_{100} равен наибольшему выборочному значению.

Более часто используемыми характеристиками локализации отдельных частей выборки являются так называемые **квартили**. Квартили – это числа, обозначаемые через Q_1, Q_2, Q_3, которые делят упорядоченную выборку на четыре части:



Первый квартиль выборки Q_1 совпадает с 25-м процентилем выборки, то есть Q_1 является числом, большим не более 25 % выборочных значений, и меньшим не более 75 % значений.

Второй квартиль Q_2 является медианой. Третий квартиль Q_3 совпадает с 75-ым процентилем. Число Q_3 делит выборку на две части: первая часть, содержащая значения меньшие числа Q_3, составляет не более 75 % выборки, вторая часть, содержащая выборочные значения, большие числа Q_3, составляет не более 25 % выборки. Итак, $Q_1 = P_{25}$, $Q_2 = P_{50}$, $Q_3 = P_{75}$.

Алгоритм вычисления k-го процентиля

1. Выборка записывается в порядке возрастания выборочных значений от меньшего к большему. Если объем выборки равен n, то после упорядочения каждое выборочное значение занимает определенную позицию или определенный номер от 1 до n.

2. Если $k = 50$, то процентиль P_{50} совпадает с медианой и находится по алгоритму определения медианы.

3. При $k < 50$, вначале вычисляется вспомогательное значение $d_k = \frac{n \cdot k}{100}$, где n – объем выборки. Если d_k получается дробным числом, то оно округляется до следующего за ним целого числа d. Например, если $d_k = 18{,}3$, то $d = 19$. Если d_k выражается целым числом, то берутся два целых числа $d = d_k$ и следующее за ним $d+1$. Например, если $d_k = \frac{n \cdot k}{100} = 11$, то берутся $d = 11$ è $d+1 = 12$.

4. При $k > 50$ вместо k используется значение $k' = 100 - k$, которое меньше 50. Затем для k' выполняются необходимые вычисления из пункта 3 для нахождения либо одного целого числа d, либо двух чисел d и $d+1$. Находится значение процентиля P_k.

5. Если $k < 50$, то от начала выборки отсчитывается позиция с номером d. Если d – единственное значение, найденное по третьему пункту, то выборочное значение x_d, стоящее на месте с номером d и будет являться k-ым процентилем: $P_k = x_d$. Если были

взяты два значения d и $d+1$, то находятся два выборочных значения x_d è x_{d+1}, стоящих на местах с номерами d и $d+1$. Процентиль P_k равен полусумме этих значений:

$$P_k = \frac{x_d + x_{d+1}}{2}.$$

Если $k > 50$, то позиция с номером d или позиции с номерами d и $d+1$ отсчитываются от конца выборки. Затем значение P_k находится так же, как и для $k < 50$.

По этой схеме вычисляются и квартили.

Пример 2.31 Рассмотрим данные о весе багажа, зарегистрированного пассажирами самолета одного авиарейса.

5,7	10,6	14,8	23,6	29,7	35,5	46,4	56,5
7,4	10,8	15,6	24,4	32,2	36,7	48,1	58,2
8,2	11,7	16,7	25,7	32,6	38,4	49,5	64,8
9,4	12,5	20,4	27,2	33,5	44,3	52,8	68,7
9,8	13,4	22,5	28,5	34,6	45,2	54,7	70,2

По условию $n = 40$ – это объем данной выборки. Найдем первый квартиль Q_1, совпадающий с 25-м процентилем P_{25}. Итак, $k = 25 < 50$. Вычислим $d_{25} = \frac{n \cdot k}{100} = \frac{40 \cdot 25}{100} = 10$. Получилось целое число, поэтому берем два значения $d = 10$ и $d+1 = 11$. Находим два выборочных значения, стоящие на 10-м и 11-м местах: $x_{10} = 13{,}4$ и $x_{11} = 14{,}8$, 25-й процентиль равен их полусумме: $P_{25} = \frac{13{,}4 + 14{,}8}{2} = 14{,}1$. Итак $Q_1 = P_{25} = 14{,}1$. Найдем второй квартиль $Q_2 = P_{50} = X_{\text{мед}}$. Так как n – четное число, то находим два выборочных значения, стоящих на $\frac{n}{2}$-м месте и на $\left(\frac{n}{2}+1\right)$– месте: $x_{20} = 28{,}5$ и $x_{21} = 29{,}7$. Медиана $X_{\text{мед}}$ равна их полусумме: $X_{\text{мед}} = \frac{28{,}5 + 29{,}7}{2} = 29{,}1$. Итак, $Q_2 = P_{50} = X_{\text{мед}} = 29{,}1$.

Найдем третий квартиль $Q_3 = P_{75}$. Так как 75 > 50, то берем значение $k' = 100 - 75 = 25$. Для $k' = 25$ найдено два целых числа $d = 10$ и $d + 1 = 11$. На 10-м и 11-м местах от конца выборки стоят соответствующие значения $x'_{10} = 46{,}4$ и $x'_{11} = 45{,}2$. Вычислим их полусумму $\frac{46{,}4 + 45{,}2}{2} = 45{,}8$. Полученное число и является 75-м процентилем. Следовательно, $Q_3 = P_{75} = 45{,}8$.

Теперь вычислим 37-ой процентиль. Найдем $d_k = \frac{nk}{100} = \frac{40 \cdot 37}{100} = 14{,}8$. Округлим до следующего целого: $d = 15$. На 15-м месте от начала выборки находится значение 22,5, которое и является 37-м процентилем: $P_{37} = 22{,}5$.

Квартили $Q_1 = 14{,}1$, $Q_2 = 29{,}1$, $Q_3 = 45{,}8$ делят выборку на четыре равные части:

5,7 10,6	14,823,6	29,735,5	46,456,5	
7,4 10,8	15,624,4	32,236,7	48,158,2	
8,2 11,7	16,725,7	32,638,4	49,564,8	
9,4 12,5	20,427,2	33,544,3	52,868,7	
9,8 13,4	22,528,5	34,645,2	54,770,2	

$P_0 = 5{,}7$ $Q_1 = 14{,}1$ $Q_2 = 29{,}1$ $Q_3 = 45{,}8$ $P_{100} = 70{,}2$

■

Сгруппированные по интервалам наблюдения скрывают конкретные выборочные значения, поэтому точные значения процентилей и квартилей не определяются. По статистическому ряду можно найти только их приближенные оценки. Метод нахождения оценок процентилей объясняет следующий конкретный пример.

Пример 2.32 Рассмотрим сведения о сроках эксплуатации 50 легковых автомобилей, зарегистрированных страховой фирмой.

Таблица 2.22 – Данные о сроках эксплуатации автомобилей

Срок эксплуатации	0–5	5–10	10–15	15–20	20–25	25–30	30–35	35–40
m_i	6	12	14	6	5	4	2	1
p_i^*	0,12	0,24	0,28	0,12	0,10	0,08	0,04	0,02
%	12 %	24 %	28 %	12 %	10 %	8 %	4 %	2 %

В третьей строке даны процентные количества выборочных значений в каждом интервале.

Простроим гистограмму данного статистического ряда.

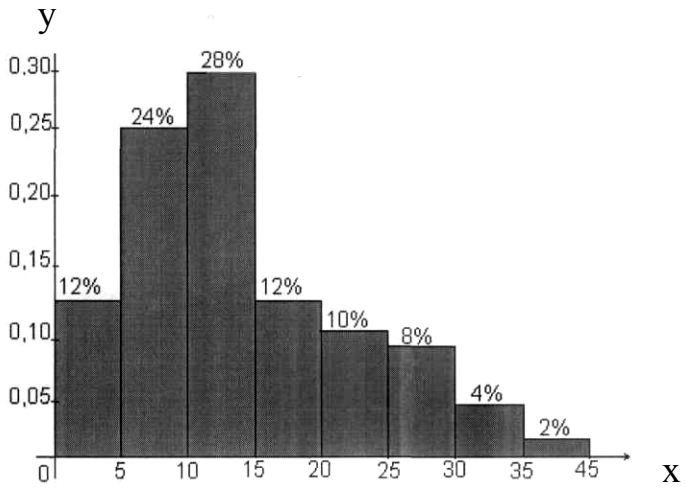


Рисунок 2.3 – Гистограмма данных о сроках эксплуатации автомобилей

Будем считать, что внутри каждого интервала выборочные значения распределены одинаково равномерно. Это значит, что всем элементам интервала соответствуют отрезки одинаковой длины. Например, первый интервал содержит 12 % выборочных значений, тогда одному проценту соответствует отрезок длиной $\frac{\ell}{12}$, где ℓ – длина интервала.

Чтобы найти k-ый процентиль статистического ряда, необходимо, прежде всего, выяснить, в каком интервале он находится. Допустим, мы хотим найти 70-й процентиль P_{70}. Последовательно складываем проценты 1-го, 2-го, … интервалов до тех пор, пока не получим максимальную сумму, не превосходящую числа 70:

$$12\ \% + 24\ \% + 28\ \% = 64\ \%.$$

Следовательно, 70-й процентиль попадает в следующий четвертый интервал (15–20]. Чтобы получить 70 % надо к 64 % прибавить 6 % значений из четвертого интервала. В четвертом интервале содержится 12 % значений, длина его равна 5, поэтому каждому проценту соответствует длина $\frac{5}{12}$, но тогда 6 % значений

лежит на отрезке, длина которого равна $\frac{5}{12}\cdot 6 = 2{,}5$. Прибавляем к нижней границе интервала это значение 15 + 2,5 = 17,5. Полученное число и является 70-м процентилем: $P_{70} = 17{,}5$. Это означает, что все выборочные значения из первого, второго и третьего интервалов и не более 6 % значений из четвертого интервала являются меньшими числа $P_{70} = 17{,}5$.

■

В том случае, когда выборка не сгруппирована и упорядочена по возрастанию от меньшего к большему, каждому значению выборки соответствует определенный **процентный ранг**. Рассмотрим это понятие на конкретном примере.

Пример 2.33 Найдем процентный ранг каждого элемента следующей выборки, характеризующей количественный состав 10 семей:

1, 2, 2, 3, 3, 3, 3, 4, 4, 5.

Рассмотрим значение, равное 2. В выборке только одно значение меньше 2-х, что составляет 10 % всей выборки. После значения 2 следующим выборочным значением является 3. В выборке есть три значения, которые меньше 3-х, что составляет 30 % всей выборки. Складываем 10 % + 30 % и делим на 2, полученное число $\frac{10+30}{2} = 20$ и будет процентным рангом выборочного значения 2, обозначаемого символом $\mathrm{PR}(2) = 20$.

Рассмотрим значение 3. В выборке имеется три значения меньших 3-х, что составляет 30 % выборки, и семь значений меньших 4-х, что составляет 70 %. Получаем процентный ранг

$$\mathrm{PR}(3) = \frac{30+70}{2} = 50.$$

Рассмотрим значение 4. В выборке 7 значений меньших 4-х (70 %) и 9 значений меньших 5-ти (90 %), поэтому $\mathrm{PR}(4) = \frac{70+90}{2} = 80$.

Рассмотрим значение 1. В выборке нет значений меньших 1 (0 %) и одно значение меньше 2-х (10%), поэтому $\mathrm{PR}(1) = \frac{0+10}{2} = 5$.

Наконец возьмем значение 5. В выборке 90 % значений меньших 5 и 100 % значений меньших следующего условного значения, поэтому $\mathrm{PR}(5)=\dfrac{90+100}{2}=95$.

Таким образом, мы нашли процентные ранги всех выборочных значений:

$$\mathrm{PR}(1)=5,\quad \mathrm{PR}(2)=20,\quad \mathrm{PR}(3)=50,\quad \mathrm{PR}(4)=80,\quad \mathrm{PR}(5)=95.$$

Простые вычисления дают следующие значения соответствующих процентилей:

$$P_5=1,\quad P_{20}=2,\quad P_{50}=3,\quad P_{80}=4,\quad P_{95}=5.$$

■

Рассмотрим понятие процентного ранга в общем виде.

Определение 2.27 Пусть *x* – произвольное значение, имеющее частоту m в упорядоченной выборке объёма n. Процентным рангом значения *x* называется число $\mathrm{PR}(x)$, равное

$$\mathrm{PR}(x)=\frac{\mathrm{k}+\frac{1}{2}\cdot \mathrm{m}}{\mathrm{n}}\cdot 100,$$

где k – число выборочных значений, меньших *x*.

Например, для предыдущей выборки найдем процентный ранг значения 4:

$$\mathrm{PR}(4)=\frac{7+\frac{1}{2}\cdot 2}{10}\cdot 100=80.$$

Заметим, что понятие процентиля (персентиля) совпадает с понятием процентной точки. Процентили, или процентные точки используются для обозначения границ изменчивости исследуемой случайной величины. Несколько значений процентилей могут довольно хорошо показать основные черты распределения. Наиболее часто используются **пять основных процентелей**:

1. Наименьшее выборочное значение, или нулевой процентиль: P_0.

2. Первый нижний квартиль Q_1, совпадает с 25-м процентилем: $Q_1 = P_{25}$.

3. Медиана, совпадающая со вторым квартилем и с 50-м процентилем: $X_{\text{мед}} = Q_2 = P_{50}$.

4. Третий верхний квартиль, или 75-й процентиль: $Q_3 = P_{75}$.

5. Наибольшее выборочное значение, или 100-й процентиль P_{100}.

Пример 2.34 Рассмотрим данные измерений частоты пульса у двенадцати пациентов поликлиники:

$$58,\ 62,\ 64,\ 65,\ 72,\ 74,\ 78,\ 80,\ 82,\ 84,\ 88,\ 93.$$

Наименьшим значением является 58, а наибольшим – 93.

Найдем медиану. Так как $n = 12$ – четное число, то берем два значения $\frac{n}{2} = 6$ и $\frac{n}{2} + 1 = 7$. На 6-м и 7-м местах находятся значения $x_6 = 74$ и $x_7 = 78$. Медиана $X_{\text{мед}} = \frac{74 + 78}{2} = 76$.

Найдем первый квартиль $Q_1 = P_{25}$. Итак, $k = 25 < 50$ и $d_k = \frac{nk}{100} = \frac{12 \cdot 25}{100} = 3$ является целым числом, поэтому берем два числа $d = 3$ и $d + 1 = 4$. На 3-м и 4-м местах от начала выборки находятся $x_3 = 64$ и $x_4 = 65$. Тогда $Q_1 = \frac{64 + 65}{2} = 64{,}5$.

Для определения верхнего квартиля Q_3 берем значения 84 и 82, стоящие на 3-м и 4-м местах от конца выборки. Тогда $Q_3 = \frac{84 + 82}{2} = 83$.

Следовательно, мы получили пять основных показателей локализации выборочных значений:

$$x_{\min} = 58;\quad Q_1 = 64{,}5;\quad x_{\text{мед}} = 76;\quad Q_3 = 83 \text{ и } x_{\max} = 93.$$

■

Подчеркнем, что пять основных процентилей делят выборку на четыре части, содержащих по 25 % выборочных значений. Квартили Q_1 и Q_3 выделяют центральную часть выборки, которая, как

считается, дает более устойчивые оценки исследуемого распределения.

Статистическое понятие процентиля тесно связано со следующим теоретическим аналогом.

Определение 2.28 Квантилью порядка p, или p-квантилью, $0 \le p \le 1$, случайной величины X с функцией распределения $F_x(x)$ называется число x_p, для которого

$$F_x(x_p) = p.$$

Это значит, что p-квантиль x_p является корнем данного уравнения. Из определения функции распределения следует, что значение p является вероятностью события $X < x_p$. Отметим, что 0,5-квантиль совпадает с медианой X_{med}. Если функция распределения $F_x(x)$ строго монотонна, то уравнение имеет только одно решение, причем большим значениям вероятностей p соответствуют большие значения квантилей. Если случайная величина X является дискретной, то ее функция распределения изменяется скачками, поэтому для некоторых значений p решения указанного уравнения образуют целый отрезок между двумя соседними α и β, такими, что $F_x(\alpha) < p$, но $F_x(\beta) \ge p$. При этом любая точка отрезка $[\alpha; \beta]$ является p-квантилью.

Подчеркнем, что квантиль порядка p совпадает с k -ым процентилем при k = 100 p. Квантили и процентили связаны следующим соотношением:

$$x_p = P_{100p}.$$

Таким образом, 0,95-квантиль является 95-ым процентилем; соответственно, квантили

$$x_{0,25} = P_{25}, \quad x_{0,50} = P_{50}, \quad x_{0,75} = P_{75}$$

также называются **квартилями.** Квантили $x_{0,1}, x_{0,2}, x_{0,3}, ..., x_{0,9}$ и соответствующие процентили называются **децилями**. Для ряда наиболее известных законов распределений составлены специальные таблицы, по которым находятся значения квантилей и, соответственно, процентилей.

Определенная совокупность квантилей для заранее подобранных значений p дает возможность составить представление о виде функции распределения исследуемой случайной величины. Например, по следующим децилям

$$x_{0,1} = -1{,}28;\ \ x_{0,2} = -0{,}84;\ \ x_{0,3} = -0{,}52;\ \ x_{0,4} = -0{,}25;$$

$$x_{0,5} = 0;\ \ x_{0,6} = 0{,}25;\ \ x_{0,7} = 0{,}52;\ \ x_{0,8} = 0{,}84;\ x_{0,9} = 1{,}28$$

можно построить график функции распределения случайной величины X, имеющей стандартное нормальное распределение, то есть функции

$$\mathrm{F_x}(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}} dt,$$

который изображен на следующем рисунке.

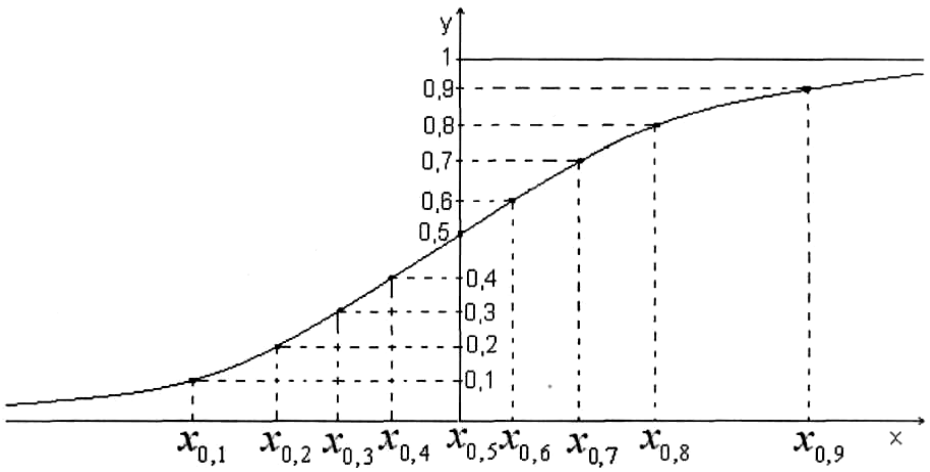


Рисунок 2.4 – Построение графика функции распределения по децилям

Упражнения

2.1 В автошколе было проведено тестирование по правилам дорожного движения. Каждому из 32 учащихся надо было решить 10 задач. В результате получены следующие данные о числе правильных ответов. Найдите моду и медиану данной выборки:

4	7	5	9	7	3	8	5
10	6	7	6	8	6	4	7
7	8	8	7	5	8	8	8
8	5	8	8	5	5	10	6

2.2 Следующая выборка состоит из данных о конкурсе абитуриентов нескольких вузов по разным специальностям. Найдите медиану и моду данной выборки. Затем составьте статистический ряд и вычислите его моду и медиану. Сравните полученные значения.

2,7	1,5	3	2,1	1,4	2,3	3,1	4
1,5	2	1,4	3,2	3,1	5,2	2,2	1,5
2,4	1,1	5,7	1,4	2,7	2,5	1,7	1,8
4,1	3,6	2,6	2	4,3	3,5	6,2	2,5

2.3 Следующий вариационный ряд содержит данные о том, сколько льготных путевок получил каждый член профсоюза за последние 10 лет. Найдите моду и медиану данной выборки.

Кол-во путевок X	0	1	2	3	4	5	6	7	8
Частота m_i	336	180	115	68	42	28	17	10	4

2.4 Вычислите $\overline{X}$ по данным тестирования учащихся по правилам дорожного движения в задаче 2.1. Сравните значения $\overline{X}$, $X_{ìîä}$ и $X_{ìåä}$. Объясните полученные результаты.

2.5 Вычислите среднее $\overline{X}$ по данным задачи 2.2. о конкурсных баллах абитуриентов в вузах. Сравните $\overline{X}$, $X_{ìîä}$ и $X_{ìåä}$. Объясните

полученные результаты.

2.6 Вычислите среднее $\overline{X}$ по данным вариационного ряда из задачи 2.3. Сравните его с модой и медианой. Объясните полученные результаты.

2.7 Верить или не верить гороскопам? Группа из 80 человек внимательно изучила опубликованные недельные гороскопы, и в конце недели каждый участник эксперимента указал, какой процент предсказаний оказался правильным. По данным наблюдений вычислите среднее, характеризующее процент сбывшихся предсказаний. Какой ответ подсказывает найденное значение?

Проценты X	0–20	20–40	40–60	60–80	80–100
Частота m_i	45	18	9	6	2

2.8 При исследовании транспортных проблем города были собраны данные об интервалах между рейсами автобусов для каждого часа движения в течение суток. Составьте вариационный ряд. Вычислите среднее $\overline{X}$, геометрическое среднее, гармоническое среднее. Сравните их значения и объясните результаты.

30	10	10	10	20
20	15	15	10	20
10	20	10	15	30
5	15	5	20	30

2.9 Декан факультета составил список студентов I курса, имеющих спортивные разряды. В списке оказалось 5 студентов 1-й группы, 4 студента – 2-й группы, 8 студентов – 3-й и 20 студентов – 4-й группы. Сколько спортсменов-разрядников в среднем приходится на одну группу? Вычислите арифметическое, геометрическое и гармоническое среднее. Какое из них более адекватно отражает исходные данные?

2.10 По данным Госкомитета ежемесячная пенсия в России в 1999, 2000, 2001, 2002, 2003 годах составляла соответственно 449,

694, 3, 1024, 1379, 1627 рублей. Вычислите среднее арифметическое и среднее гармоническое пенсий за эти годы. Найдите среднее ежегодного роста пенсий. Объясните полученные результаты.

2.11 По данным ООН о плотности населения 50 стран вычислите вариационный размах, среднее плотности, среднее абсолютное и стандартное отклонение. Найдите квартили распределения плотности населения.

32	39	5	28	262	26	22	59	11	14
9	224	283	54	114	83	68	140	74	30
834	57	20	101	21	883	150	60	385	11
49	98	27	30	68	11	7	13	84	29
13	5	15	30	20	2	205	15	35	56

2.12 Для импортеров текстиля составлен список цен за одну условную единицу товара для 32 стран мира, являющихся самыми главными производителями текстиля. Вычислите среднее цены текстиля и стандартное отклонение. Найдите квартили распределения цен. Найдите 10-й и 85-й процентили. Найдите процентный ранг следующих цен: 0,31; 0,49; 0,56.

0,31	0,39	0,49	0,44	0,49	0,46	0,51	0,46
0,25	0,56	0,49	0,46	0,38	0,46	0,55	0,49
0,55	0,51	0,46	0,42	0,42	0,56	0,49	0,27
0,46	0,56	0,56	0,41	0,49	0,49	0,52	0,52

2.13 Вычислите стандартное отклонение для выборки из задачи 2.1. Найдите квартили выборки.

2.14 Вычислите стандартное отклонение для выборки из задачи 2.2. Найдите квартили выборки.

2.15 Вычислите стандартное отклонение для выборки из задачи 2.3. Найдите квартили выборки.

2.16 Вычислите стандартное отклонение для выборки из задачи 2.7. Найдите квартили и децили выборки.

2.17 Вычислите стандартное отклонение для выборки из задачи 2.8. Найдите квартили и децили выборки.

2.18 По отчетам крупных компаний России опубликованы сведения о процентной доле налогов в выручке соответствующих компаний. Вычислите среднее и стандартное отклонение. Найдите пять основных процентилей выборки.

41,8	46,3	38,1	37	36,8	33,7	32,7	30,8	26,6	33,5	19,6
42	38	34,4	27,5	41,2	26,2	31,9	32,3	25,6	25,4	19,4
45,4	38,7	31,5	23,8	43,1	26,1	30,5	21,9	22,1	18,3	24,5

2.19 Брокерская фирма собрала данные о котировке акций крупной нефтяной компании на биржевых торгах. Вычислите среднее стоимости одной акции и стандартное отклонение. Найдите пять процентилей данной выборки.

96	93	96	97	100	101	102	99	94	91
97	94	95	98	102	104	100	97	92	90

2.20 Проведите обработку статистических данных о количестве телевизоров, приходящихся на 100 жителей разных стран. Представьте графическое распределение исходных данных. Вычислите все числовые характеристики.

10	224	227	239	201	48	80	18
103	495	454	378	5	5	409	9
89	497	178	6	1	88	290	48
222	33	6	2	215	433	446	334
368	229	115	9	205	353	162	370
7	467	19	24	286	51	684	34
424	6	220	714	56	66	80	9
219	284	2111	4	9	63	256	477

Заключение

Первичная обработка статистических данных является необходимой подготовкой статистической информации к статистическому анализу. Основными задачами этого этапа являются:

– группировка данных в компактную информативную форму;

– представление данных в наглядном изображении, отражающем их существенные закономерности;

– вычисление основных числовых характеристик.

Эффективное использование статистических методов в прикладных исследованиях обеспечивается не только глубокими теоретическими знаниями математической статистики, но и уверенными навыками пользования соответствующим программным обеспечением ЭВМ. К настоящему времени методы статистического анализа унифицированы, для их практической реализации создан мощный аппарат прикладной математической статистики в форме пакетов статистических программ. Существуют профессиональные статические пакеты с максимальным набором методов исследования любых баз данных. Несколько меньшие возможности имеют универсальные статические пакеты, в которых могут отсутствовать некоторые узкоспециальные методы. Для широкого круга практических исследований статистических данных подходят специализированные пакеты, содержащие основные процедуры прикладной математической статистики. Методы первичной обработки данных и вычислительные алгоритмы описательной статистики достаточно полно представлены в электронных таблицах Excel 97\2000\XP из широко распространенного программного приложения MS Office. Технология работы с данным табличным процессором, регламентируемая встроенной справочной системой и соответствующими примерами, является вполне доступной для студентов, имеющих элементарные навыки работы в Excel. С помощью компьютерного процессора решаются основные задачи первичной обработки данных, а именно:

– ввод и создание баз данных;

– сортировка данных по возрастанию либо по убыванию;

– вычисление частот;

– графическое изображение полигона частот;
– формирование интервального ряда;
– построение гистограммы интервального ряда;
– построение эмпирической функции распределения;
– вычисление числовых характеристик выборки.

В приложении Б приведен список статистических функций для вычисления числовых характеристик в MS Excel. Практическое использование программного обеспечения при изучении курса математической статистики с одной стороны решает проблему трудоемкости вычислительных статистических процедур и с другой стороны способствует более прочному усвоению теоретических основ прикладной статистики.

Приложение А

(обязательное)

Таблица случайных чисел

51867	21562	07079	08899	14894	30068	49608	95350	98901
12116	97983	19322	91039	05030	82800	54738	16371	84550
65558	45040	56325	67251	19561	69692	56324	03426	25769
51904	19200	84819	28701	56517	05851	31093	13895	35983
93123	16383	62615	03846	39284	58653	77924	19138	03742
27887	14031	52342	94589	33737	99949	28622	31200	76822
53138	00936	82968	78471	42512	63505	83543	30616	12073
21488	81518	75540	57741	86411	40409	28912	14639	59463
09095	48440	80045	13599	23753	85551	15059	44406	84420
78777	02218	53069	84390	29690	90729	80192	44236	15868
71240	04756	20665	32146	26096	64938	83964	57360	99505
99187	19506	21282	28701	81361	52403	78192	81644	11426
19258	60695	07768	03846	93099	42396	21626	94761	12651
86421	88494	93945	94589	33922	40112	91399	28155	61646
16401	60677	06293	78471	37329	11469	07235	03521	11769
59981	15076	22879	57741	89911	03476	07104	36415	75109
68155	92554	08161	13599	55876	03328	73652	78452	86996
45673	26042	01442	84390	28379	84238	64425	92359	97669
76210	23472	75071	32146	81031	26570	85149	81091	25757
58219	69869	21427	00871	22058	51790	75409	56513	32535
45738	62877	94842	09354	21487	42122	64666	88321	07122
29550	19584	26210	22745	54613	13318	34767	97910	81769
24736	39576	75689	65806	78355	14192	97298	87971	74436
09574	66314	76131	89242	54013	98167	92708	29031	02630
46251	05212	96837	79337	50774	75631	01994	51780	72310
25437	67859	67450	59293	30666	74141	53188	27376	79337
69654	89356	44511	47481	61205	22369	78476	81056	59253
99716	20056	50424	07740	42574	36757	78048	86155	45049
11563	30648	82848	43345	47773	89117	91170	55488	18029
08803	87349	41975	25716	36027	54998	54998	50590	07469
86027	20389	71663	70020	27174	60571	60571	74514	42341
51867	53805	02921	54005	08845	54786	62404	58147	98173
12116	20416	16919	14955	99145	26281	82201	68841	36737

65558 87410 35424 59592 94316 01855 75694 53625 98863
51904 75646 93209 97035 88974 30706 02808 02059 77240
93123 64176 52133 80430 29828 66578 65983 75223 76251
27887 82752 87327 87220 97069 32019 74373 16783 00654
53138 63606 95897 06392 90327 65884 66693 19272 64688
21488 37011 65171 79028 61848 58485 13094 61994 09343
09095 57346 20376 57123 29604 09531 74183 71090 70278
78777 69512 14295 52872 01769 81853 73020 18875 67331
71240 28701 34969 42446 71825 59334 15360 52809 82861
99187 56992 14216 41880 55957 70929 73776 70594 54371
19258 40423 03191 37415 98271 03544 40914 41649 76610
86421 62415 61647 47472 02784 18510 85190 32935 94934
32942 35641 74579 05303 18534 17639 72865 68142 75109
95416 00301 33844 91109 22346 38284 16829 67957 56474
42339 16096 33426 82403 54556 59478 86542 70896 74111
59045 34775 07570 40312 17558 90409 00396 37983 31966
26693 21562 00728 62191 73689 21997 20363 20487 29969
49057 97983 07079 67023 14894 56199 13010 95350 70093
8796 45040 19322 90073 05030 30068 69645 16371 98901
20624 19200 56325 83205 19561 82800 49608 03426 84550
14819 16383 84819 71344 56517 69692 54738 13895 25769
07410 14031 62615 57071 39284 05851 56324 19138 35983
99859 00936 52342 90357 33737 58653 31093 31200 03742
83828 81518 82968 12901 42512 99949 77924 30616 76822
21409 48440 75540 08899 86411 63505 28622 14639 12073
29094 02218 80045 91039 23753 40409 83543 44406 59463
65114 04756 53069 67251 29690 85551 28912 44236 84420
36701 19506 20665 28701 26096 90729 15059 57360 15868
25762 60695 21282 03846 81361 64938 80192 81644 99505
12827 88494 07768 94589 93099 52403 83964 94761 11426
59981 60677 93945 78471 33922 42396 78192 28155 12651
68155 15076 06293 57741 37329 40112 21626 03521 61646
45673 92554 22879 13599 89911 11469 91399 36415 11769
76210 26042 08161 84390 55876 03476 07235 78452 75109
58219 23472 01442 32146 28379 03328 07104 92359 86996
45738 69869 75071 00871 81031 84238 73652 81091 97669
29550 62877 21427 09354 22058 26570 64425 56513 25757
24736 19584 94842 22745 21487 51790 85149 88321 32535

Приложение Б

(справочное)

Таблица статистических функций для вычисления числовых характеристик в MS Excel

Числовые характеристики	Стандартные функции в MS Excel
Среднее	=СРЗНАЧ (число 1; число 2; …) =СРЗНАЧ (значение1; значение 2; …)
Дисперсия выборки	=ДИСП (число 1; число 2; …) =ДИСПА (значение1; значение 2; …)
Дисперсия генеральной совокупности	=ДИСПР (число 1; число 2; …) =ДИСПРА (значение1; значение 2; …)
Среднее абсолютное отклонение	=СРОТК (число 1; число 2; …)
Сумма квадратов отклонений	=КВАДРОТКЛ (число 1; число 2; …)
Стандартное отклонение выборки	=СТАНДОТКЛОН (число 1; число 2; …) =СТАНДОТКЛОНА (значение1; значение 2; …)
Стандартное отклонение генеральной совокупности	=СТАНДОТКЛОНП (число 1; число 2; …) =СТАНДОТКЛОНПА (значение1; значение 2; …)
Мода	=МОДА (число 1; число 2; …)
Медиана	=МЕДИАНА (число 1; число 2; …)
Гармоническое среднее	=СРГАРМ (число 1; число 2; …)
Геометрическое среднее	=СРГЕОМ (число 1; число 2; …)
Асимметрия	=СКОС (число 1; число 2; …)
Эксцесс	=ЭКСЦЕСС (число 1; число 2; …)
k-й процентиль	=ПЕРСЕНТИЛЬ (массив; k)
Процентный ранг	=ПРОЦЕНТРАНГ (массив; значение)
Минимальное значение выборки	=МИН (число 1; число 2; …) =МИНА (значение1; значение 2; …)
Максимальное значение выборки	=МАКС (число 1; число 2; …) =МАКСА (значение1; значение 2; …)
Объем выборки	=СЧЕТ (значение1; значение 2; …)

Литература

1 Айвазян, С. А. Прикладная статистика : Основы моделирования и первичная обработка данных [Текст] : справочное издание / С. А. Айвазян, И. С. Енюков, Л. Д. Мешалкин; под общей ред. С. А. Айвазяна. – М. : Финансы и статистика, 1983. – 471 с.

2 Булдык, Г. М. Теория вероятностей и математическая статистика [Текст] : учеб. пособие для вузов / Г. М. Булдык. – Мн. : Вышэйшая школа, 1989. – 285 с.

3 Герасимович, А. И. Математическая статистика [Текст] : учеб. пособие для вузов / А. И. Герасимович, Я. И. Матвеева. – Мн. : Вышэйшая школа, 1978. – 279 с.

4 Горелова, Г. В. Теория вероятностей и математическая статистика в примерах и задачах с применением Exell [Текст] : учеб. пособие для вузов / Г. В. Горелова, И. А. Кацко. – 3-е изд., доп. и перераб. – Ростов-на-Дону : Феникс, 2005. – 475 с.

5 Ермолаев, О. Ю. Математическая статистика для психологов [Текст] : учебник / О. Ю. Ермолаев. – М. : Флинта, 2002. – 336 с.

6 Малинковский, Ю. В. Теория вероятностей и математическая статистика : в 2 ч. Ч. 1. Теория вероятностей [Текст] : учеб. пособие для вузов / Ю. В. Малинковский. – Гомель : УО «ГГУ им. Ф. Скорины», 2004. – 354 с.

7 Малиновский, Ю. В. Теория вероятностей и математическая статистика : в 2 ч. Ч. 2. Математическая статистика [Текст] : учеб. пособие для вузов / Ю. В. Малинковский. – Гомель : УО «ГГУ им. Ф. Скорины», 2004. – 146 с.

8 Мацкевич, И. П. Высшая математика : Теория вероятностей и математическая статистика [Текст] : учебник для вузов / И. П. Мацкевич, Г. П. Свирид. – Мн. : Вышэйшая школа, 1993. – 269 с.

9 Михок, Г. Выборочный метод и статистическое оценивание [Текст] : науч. издание / Г. Михок, В. Урсяну. – М. : Финансы и статистика, 1982. – 245 с.

10 Савич, Л. К. Теория вероятностей и математическая статистика [Текст] : учеб. пособие для вузов / Л. К. Савич, И. А. Смольская; науч. ред. О. И. Лаврова. – М. : Адукацыя і выхаванне, 2006. – 208 с.

Учебное издание

ХАРЛАМОВА Валентина Ивановна

ТЕОРИЯ ВЕРОЯТНОСТЕЙ И МАТЕМАТИЧЕСКАЯ СТАТИСТИКА

ПЕРВИЧНАЯ ОБРАБОТКА СТАТИСТИЧЕСКИХ ДАННЫХ

ПРАКТИЧЕСКОЕ ПОСОБИЕ

для студентов университета

В авторской редакции

Лицензия № 02330/0549481 от 14.05.09.
Подписано в печать 11.12.09. Формат $60\times84\ ^{1}/_{16}$.
Бумага офсетная. Гарнитура «Таймс». Усл. печ. л. 6,51.
Уч.-изд. л. 7,12. Тираж 25 экз. Заказ № 427

Отпечатано с оригинала-макета на ризографе
учреждения образования
«Гомельский государственный университет
имени Франциска Скорины».
Лицензия № 02330/0150450 от 03.02.09.
246019, г. Гомель, ул. Советская, 104

Министерство образования Республики Беларусь

Учреждение образования
«Гомельский государственный университет
имени Франциска Скорины»

В. И. Харламова

ТЕОРИЯ ВЕРОЯТНОСТЕЙ И МАТЕМАТИЧЕСКАЯ СТАТИСТИКА

ПЕРВИЧНАЯ ОБРАБОТКА СТАТИСТИЧЕСКИХ ДАННЫХ

Гомель
УО «ГГУ им. Ф. Скорины»
2009