Г.И. Фалин, А.И. Фалин

ИЗБРАННЫЕ ГЛАВЫ ОПИСАТЕЛЬНОЙ СТАТИСТИКИ

Москва
МАКС Пресс
2011

МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
имени М.В.Ломоносова

Факультет вычислительной математики и кибернетики

Г.И.Фалин, А.И.Фалин

ИЗБРАННЫЕ ГЛАВЫ ОПИСАТЕЛЬНОЙ СТАТИСТИКИ

Учебное пособие

Москва
МАКС Пресс
2011

УДК 519.2 (075.3)
ББК 22.172я729
Ф19

Рецензенты:
доктор физико-математических наук, профессор кафедры общей математики факультета вычислительной математики и кибернетики МГУ им.М.В.Ломоносова И.С.Ломов

доктор физико-математических наук, профессор кафедры математической статистики факультета вычислительной математики и кибернетики МГУ им.М.В.Ломоносова В.Ю.Королёв

Фалин Г.И., Фалин А.И.

Избранные главы описательной статистики. – М.: МАКС Пресс, 2011 –131 с., ил.

ISBN 978-5-317-03814-4

Книга посвящена более глубокому изучению как стандартных понятий описательной статистики, изучаемых в средней школе (среднее значение, медиана, дисперсия, гистограмма), так и сопутствующих понятий (функция распределения, квартили и т.д.). При этом описательная статистика излагается как математическая наука с определениями, теоремами, строгими доказательствами, использующая обычный алгебраический язык формул. Теория иллюстрируется большим числом задач. Книга базируется на опыте преподавания статистики в школах Великобритании.

Для студентов младших курсов, учителей математики, учащихся старших классов.

УДК 519.2 (075.3)
ББК 22.172я729

ISBN 978-5-317-03814-4

Оглавление

Предисловие

В настоящее время в курс математики общеобразовательной школы вводятся элементы статистики и теории вероятностей. В хороших школьных учебниках (например, в [1], [2]) теория вероятностей излагается как математический предмет, с определениями и теоремами (другой вопрос, в состоянии ли школьники понять содержательную сторону теории вероятностей). Со статистикой (под которой, в сущности, понимается только описательная статистика) дело обстоит хуже. Она либо совершенно не затрагивается, либо излагается на примитивном уровне. Достаточно сказать, что в некоторых школьных учебниках по математике разделы, посвящённые статистике, не содержат ни одной формулы и ни одного утверждения. Фактически ничего не рассказывается об общих понятиях статистики; вряд ли цитаты из художественных произведений могут заменить академическое изложение. Соответственно и задачи ГИА по теории вероятностей и статистике сводятся к простейшим арифметическим действиям, не затрагивая суть предмета.

Всё это создаёт плохую основу для последующего изучения статистики и теории вероятностей в высшей школе. Поэтому мы решили написать небольшую книгу и показать, что описательная статистика – это математическая наука с определениями, теоремами, строгими доказательствами, излагаемыми на обычном алгебраическом языке формул.

При этом мы руководствовались опытом преподавания статистики в школах Великобритании – страны с высоким уровнем образования и давними традициями изучения статистики в школах. В Великобритании, как и во многих других странах, в школьные программы изучения математики и статистики давно включены такие темы как стандартизация набора, квартили, древовидная диаграмма, функция распределения и т.д.

Излагаемая теория иллюстрируется большим числом задач. Они составлены с использованием вариантов экзаменов, которые сдают выпускники школ Великобритании для получения общего сертификата о среднем образовании (GCSE). Хотя практически все задачи являются авторскими, мы старались сохранить дух британских экзаменов, включая уровень сложности задач и их практическую направленность.

Изложенные теоретические факты и их простые применения будут полезны студентам младших курсов ВУЗов, приступающим к изучению статистики и теории вероятностей, и учителям математики. Книгу вполне можно изучать в 9 классе, в классах с углублённым изучением математики.

Г.И.Фалин, д.ф.м.н., проф.
кафедра теории вероятностей
механико-математический факультет
МГУ им.М.В.Ломоносова

А.И.Фалин, к.ф.м.н., доцент
кафедра общей математики
факультет ВМК
МГУ им.М.В.Ломоносова

31 августа 2011 г.

1. Основные понятия статистики

1.1 Данные, информация и статистика

Каждый день люди, предприятия, организации, правительственные органы вынуждены *принимать решения*.

Пример 1.1. Экзаменационная комиссия Московского университета после письменного экзамена по математике должна *принять решение*, какую оценку поставить абитуриенту Пете Сидорову: «5», «4», «3» или «2». Для этого экзаменационная комиссия должна располагать *информацией* об уровне подготовки Пети по математике; является ли этот уровень отличным, хорошим, удовлетворительным или неудовлетворительным. Чтобы получить эту информацию, ей нужны определённые *данные*: в рассматриваемой ситуации – это письменная работа, написанная Петей. Чтобы *собрать* эти *данные*, университет проводит экзамен. Важно подчеркнуть, что сама по себе экзаменационная работа не содержит никакой информации. Она лишь являются основой для последующего анализа ситуации. Экзаменационная комиссия проверяет письменную работу Пети, анализирует и интерпретирует его записи, соотносит работу с Программой по математике для поступающих в МГУ. Только после этого экзаменационная комиссия принимает объективное решение. Для принятия решения исходные данные (т.е. письменная работа) и результаты анализа должны быть *представлены* в стандартной, понятной всем членам комиссии, наглядной форме. Письменную работу представляет Петя в виде рукописного текста на русском языке, с использованием общепринятой математической символики на трёх листах формата А3 и дополнительного титульного листа формата А3, на котором указана личная информация. Результаты анализа данных (экзаменационных работ) членами экзаменационной комиссии обычно оформляются в виде таблиц, где указан шифр работы (экзаменационная комиссия не знает, чью работу она проверяет) и уровень решения каждой задачи (обычно это символы $+, \pm, \mp, -$ с дополнительными комментариями).

Пример 1.2. Приёмная комиссия (не путать с экзаменационной!) Московского университета в конце июля должна *принять решение* о величине проходного балла; на первый курс зачисляют абитуриентов, которые сдали все экзамены на положительные оценки и имеют сумму баллов по всем экзаменам бóльшую или равную, чем проходной балл (на самом деле из-за ЕГЭ процедура запутаннее; кроме того, для простоты мы исключаем возможность полупроходного балла). Для этого приёмная комиссия должна располагать *информацией* о результатах вступительных экзаменов. Эту информацию

приёмная комиссия получает из экзаменационных ведомостей – это исходные *данные* для её работы. После этого приёмная комиссия

- исключает из списка абитуриентов тех, кто не смог успешно пройти вступительные испытания, проводимые МГУ;
- для каждого абитуриента находит общую сумму его оценок по всем установленным экзаменам;
- упорядочивает список абитуриентов по убыванию суммарного балла (при равных суммарных баллах порядок определяется специальными правилами; кроме того, законодательство и Правила приёма дают преимущество нескольким категориям абитуриентов, например, победителям Всероссийской олимпиады соответствующего профиля);
- обрезает этот упорядоченный список абитуриентов на уровне, определяемом планом приёма на первый курс.

Сумма баллов, набранная последним абитуриентом в списке зачисленных, и будет проходным баллом (конечно, мы сильно упростили схему работы комиссии). Как для принятия решения, так и для ознакомления с ним абитуриентов и их родителей, исходные данные и результаты анализа должны быть *представлены* в понятной и наглядной форме (табличной).

Эти два примера очень похожи. И в первом, и во втором примерах мы использовали понятия «данные» и «информация». Сейчас мы попробуем определить эти понятия.

Данные *– это сведения, факты, графические образы и т.п., которые используются людьми или организациями в качестве основы для рассуждений, анализа ситуаций, принятия решений и т.п.*

Данные всегда характеризуют какое-то качество, свойство (или набор качеств) некоторого объекта, явления и т.п. Данные не следует отождествлять с информацией.

Информация *– это данные, которые определённым образом обработаны, интерпретированы и имеют некоторый смысл и ценность для кого-то (отличный от самого факта наличия данных).*

Например, результаты анализа крови какого-то человека являются данными, но информацией они являются только для врача, который понимает их смысл (применяя свои медицинские знания и опыт). Для обычного человека эти данные не являются информацией.

На основе информации приобретается ***знание*** об изучаемом объекте или явлении, которое помогает в ***принятии решений***.

Но два приведённых выше примера и сильно различаются. Рассмотрим данные, с которыми мы имели дело в этих примерах, и характер их анализа немного подробнее.

В Примере 1.1 экзаменационная комиссия имеет дело с *изолированными данными* (письменной работой по математике), характеризующими отдельный изолированный объект (Петю Сидорова). Чтобы проанализировать эти данные и извлечь из них информацию, необходимо учитывать содержательный смысл

данных (например, понимать, что зачёркивание в записи «$2y + \cancel{x} = 4 + \cancel{x}$» буквы x – не исправление ошибки, а математическая операция – приведение подобных членов) и обладать профессиональными знаниями и большим опытом в определённой предметной области (математике). Информация, содержащаяся в этих данных (работе Пети Сидорова), никак не зависит от данных, полученных для других аналогичных объектов (т.е. работ других абитуриентов).

В Примере 1.2 приёмная комиссия имеет дело с большим *набором* данных (оценками за экзамены), характеризующими большую *группу* однородных объектов (абитуриентов). Эта группа называется *генеральной* (от латинского generalis – главный) *совокупностью*. Информация, которую приёмная комиссия извлекает из этих данных, связана именно с этим обстоятельством, а не с содержательным смыслом этих данных. Действительно, оценка, которую получил конкретный абитуриент на университетском экзамене по математике, может многое сказать члену экзаменационной комиссии об уровне его математических знаний, умении размышлять и т.д. Однако эта информация совершенно неинтересна приёмной комиссии. Её интересует лишь место, которое занимает абитуриент в упорядоченном списке всех претендентов. Чтобы извлечь эту информацию, имеющиеся данные относительно этого абитуриента следует рассматривать не изолированно, а как элемент большой группы однородных данных (относительно всех абитуриентов). Такие данные называются статистическими. Проходной балл, который является результатом анализа совокупности экзаменационных оценок, *является характеристикой всей генеральной совокупности (группы абитуриентов)*. Такая информация называется *статистической*.

В зависимости от того, кто собирает данные и с какой целью, одни и те же данные могут быть, а могут и не быть, статистическими. Например, данные о доходах жителей некоторого города, которые собирает налоговая инспекция, не являются статистическими, хотя жители и образуют большую однородную группу. Дело в том, что анализ данных о доходе человека не зависит от доходов других людей, а результат этого анализа (сумма налога, подлежащая уплате) относится к этому человеку, а не ко всем жителям города. Но эти же данные могут представлять интерес для администрации города, чтобы получить информацию об уровне жизни горожан. Администрацию не интересует доход каждого конкретного человека; интерес представляет, например, доля жителей с доходом ниже некоторого уровня (бедных) и доля жителей с доходом выше некоторого уровня (богатых). Эти величины характеризуют социальную обстановку в городе и могут служить основой для принятия важных решений об адресной социальной помощи и уровне местных налогов. Поскольку они характеризует всю совокупность жителей (в данном случае это и будет генеральная совокупность), данные о доходах в этом случае будут статистическими.

В зависимости от целей, с которыми собираются данные, выделяются и другие их виды. Например, данные из расписания движения поездов о номере поезда, маршруте, времени отправления и т.д. являются *справочными*.

Резюмируя можно сказать, что

Статистические данные *– это сведения, факты, графические образы и т.п. относительно большой группы однородных объектов, которые используются людьми или организациями в качестве основы для рассуждений, анализа ситуаций, принятия решений и т.п. относительной этой группы как целого.*

Статистическое исследование*, анализируя статистические данные, имеет своей целью установление свойств всей группы как целого.*

Генеральная совокупность *– это вся группа объектов, свойства которой изучаются в ходе статистического исследования.*

Генеральная совокупность, как и всякое множество, может быть задана перечислением всех своих элементов или указанием общего свойства, которое объединяет объекты в эту совокупность. Если изучаемая группа состоит из небольшого числа объектов (например, школьный класс, в котором учится 20-40 учеников), то все эти объекты можно явно указать без всякого труда. При выставлении оценок по ЕГЭ по математике в 2011 году изучаемая группа содержала около 740 тысяч объектов (школьников), но использование современной компьютерной техники для сбора и хранения данных позволяет явно указать все объекты и в этом случае. В ряде случаев трудно даже подсчитать общее число объектов в генеральной совокупности (хотя совершенно ясно, какие объекты в неё входят). Например, при изучении рыбных запасов в озере Байкал генеральная совокупность включает всех рыб, живущих в этом озере. Ясно, что подсчитать их число точно практически невозможно (даже если не принимать в расчёт непрерывное изменение этого числа). Тем не менее, существуют статистические методы, которые позволяют решить эту задачу с разумной точностью. С ещё более трудной проблемой сталкивается автомобильная компания, которая хочет понять, какие цвета предпочитают потенциальные покупатели новой модели автомобиля. В этом примере невозможно сразу точно сказать, какие люди входят в генеральную совокупность.

При статистическом исследовании чрезвычайно важно ясно понимать, о какой генеральной совокупности идёт речь, т.е. какая группа объектов исследуется. Без этого нельзя правильно организовать сбор и обработку данных, правильно интерпретировать полученные результаты и выработать рекомендации по их практическому применению.

Наука, которая изучает проблемы сбора, представления, анализа данных относительно больших групп однородных объектов и явлений, а также разрабатывает методы получения на их основе обоснованных выводов относительно этой группы как целого, называется ***статистикой****.*

Ниже, говоря о данных, мы всегда будем иметь в виду статистические данные. В математической статистике содержательный смысл данных, как правило, не играет никакой роли; обычно это просто *наборы чисел*. В специальных разделах статистики (медицинская статистика, финансовая статистика и т.д.) для получения практически значимых результатов необходимо в определённой мере принимать в расчёт соответствующую (медицинскую, финансовую и т.д.) информацию, содержащуюся в данных.

В заключение ещё раз подчеркнём (несколько упрощая ситуацию), что статистика изучает только общие свойства группы однородных объектов как целого. Если вас интересуют индивидуальные члены генеральной совокупности сами по себе, статистика бесполезна. Допустим, например, что вы хотите приобрести автомобиль и перед вами в автосалоне стоит два автомобиля – один марки «А», второй – «Б». У них примерно одинаковые технические характеристики и цена. Вы знаете, что в течение трёх лет существенного ремонта требует примерно 3% автомашин марки «А» и примерно 8% автомашин марки «Б». Важно понимать, что эта информация совершенно ничего не говорит о качестве тех двух конкретных автомобилей, из которых вы выбираете – она характеризует большие партии машин в целом. В нашей ситуации единственно разумным будет детальная диагностика каждого из двух предложенных автомобилей.

Даже если в результате анализа статистических данных принимается решение относительно индивидуального объекта совокупности (например, решение о том, зачислять или нет Петю Сидорова в студенты), это решение рассматривает этот индивидуальный объект не изолированно, а как элемент этой совокупности, и соотносит свойства этого индивидуального объекта со свойствами всей совокупности объектов как целого.

1.2 Виды статистических данных по способу сбора

В зависимости от источника статистические данные делят на *первичные* и *вторичные*.

Первичные данные *– это данные, которые вы собираете сами или же их собирает кто-то под вашим руководством.*

Вторичные данные *– это данные, которые берутся из существующих источников (книги, газеты, журналы, интернет и т.п.).*

В России важным источником данных об экономике, населении, разных сторонах жизни общества является официальный сайт Федеральной службы государственной статистики (http://www.gks.ru), где можно найти электронные версии большого числа статистических публикаций, в том числе *Российского статистического ежегодника* (обобщённые данные о географических и климатических условиях, населении, экономике и т.д.) и *Демографического*

ежегодника России (детальные данные о численности и возрастно-половом составе населения, рождаемости, смертности и т.д. как по России в целом, так и по отдельным регионам) разных лет выпуска.

Подробная статистика о ЕГЭ и ГИА может быть найдена на официальном информационном портале ЕГЭ (http://www.ege.edu.ru).

Сайт Московского университета им.М.В.Ломоносова (http://www.msu.ru) содержит статистические данные о вступительных экзаменах в МГУ.

Первичные и вторичные данные имеют свои достоинства и недостатки.

Вы знаете кто и как собирал первичные данные и потому знаете, в какой мере им можно доверять. Первичные данные (при правильном планировании метода их получения)

- содержат все сведения, которые нужны вам для вашего специфического исследования,
- не содержат ненужных фактов,
- вы можете так собирать эти данные, чтобы последующая их обработка была простой.

Всё это – важные достоинства первичных данных. Из недостатков можно отметить

- необходимость затрат времени и денег на планирование, организацию и сбор первичных данных,
- часто невозможность или неудобство их получения.

Получение вторичных данных обычно

- не требует затрат времени и денег (если только вы не обращаетесь к специальным коммерческим базам данных),
- вы можете получить данные относительно бóльшего числа объектов, эти данные могут характеризовать бóльшее число свойств этих объектов,
- вторичные данные, как правило, уже представлены в удобной для последующего анализа форме (например, таблиц),
- часто они являются результатом определённого статистического анализа.

Всё это – важные достоинства вторичных данных. Основные недостатки:

- вы не знаете, в какой мере можно доверять вторичным данным (это особенно относится к данным, полученным из интернета); более того, часто вторичные данные собираются или обрабатываются так, чтобы можно было обосновать вывод, выгодный тому, кто занимался этой работой,
- поскольку они собирались до начала вашего исследования, они могут и не отражать текущие свойства изучаемых объектов (если они достаточно быстро меняются с течением времени),
- часто они не относятся в точности к той генеральной совокупности, которая вас интересует и/или не отражают в точности интересующие вас свойства изучаемых объектов,

- часто они обработаны или представлены не так, как нужно вам для вашего специфического исследования.

1.3 Статистические переменные и наборы и их виды

1.3.1 Числовые переменные и наборы

Статистические данные всегда характеризуют какое-то качество, свойство (или набор качеств) объектов, образующих генеральную совокупность. Предположим, например, что генеральная совокупность состоит из трёх учеников, Пети, Маши, Саши, а интересующее нас свойство этих «объектов» – оценка по математике за первую четверть. Вообще говоря, эта оценка меняется от ученика к ученику. Поэтому оценка ученика как свойство неопределённого ученика из рассматриваемой группы является *переменной величиной*. Поскольку эта переменная появляется в ходе статистического исследования, её называют *статистической переменной*. Мы для краткости будем опускать в дальнейшем слово «статистическая» и говорить просто о «переменной». В нашем примере переменную можно обозначить каким-нибудь символом, например, W. Точное *значение* этой *переменной* определяется тем, о каком конкретно из учащихся идёт речь. Если, например, Петя получил за четверть оценку «4», Маша – «4», Саша – «3», то этот факт можно выразить следующим образом: $W(\text{Петя}) = 4$, $W(\text{Маша}) = 4$, $W(\text{Саша}) = 3$. Поскольку значения нашей переменной W являются числами, её называют *числовой* (или *количественной*) переменной, а соответствующий *набор значений переменной* [4;4;3] – *числовым набором*. Разделять элементы набора друг от друга можно не точкой с запятой «;», а запятой «,», если только вы не перепутаете её с десятичной запятой. В статистике обычно работают только с набором значений переменной и фактически отождествляют статистическую переменную и набор её значений. Но важно понимать, что числовые наборы в статистике – это не какие-то произвольные, неизвестно откуда появившиеся, группы чисел. Любой числовой набор в статистике всегда является результатом измерения какого-то свойства объектов определённой генеральной совокупности, а каждое число из набора относится к *одному* объекту этой генеральной совокупности. Эти замечания относятся и к нечисловым переменным, о которых мы будем говорить позже.

Школьник, знакомый с основами теории множеств в объёме школьного курса, без труда поймёт, что статистическая числовая переменная – это отображение множества изучаемых объектов (генеральной совокупности) в множество действительных чисел.

В элементарных школьных учебниках по статистике (см., например, [9]) часто используют термин «данные» вместо термина «переменная»; в более серьёзных

пособиях для школьников (см., например, [8]) эти термины употребляются именно так, как принято в современной статистике, т.е. в соответствии с данными выше определениями.

Понятие «набор» нуждается в дополнительных комментариях. В рассматриваемом нами примере для статистического исследования успеваемости в группе не играет роли, кто конкретно из учеников получил какую оценку из набора [4;4;3]. Если бы Маша получила «3», а «Саша» – «4», то это никак бы не повлияло на общую оценку успеваемости учеников этой группы школьников (хотя, несомненно, обрадовало бы родителей Саши и очень расстроило бы родителей Маши). С этой точки зрения наборы [4;4;3] и [4;3;4] следует считать одинаковыми. Неформально, можно сказать, что

Числовой (статистический) набор – *это конечная неупорядоченная последовательность чисел.*

В принципе несложно дать и аккуратное математическое определение числового статистического набора (как класса последовательностей одинаковой длины, отличающихся друг от друга перестановкой элементов), но это совершенно излишне в элементарном школьном курсе.

Чтобы не усложнять обозначения, мы будем обозначать набор значений переменной тем же символом, что и саму переменную (как правило, заглавной буквой латинского алфавита). Таким образом, в рассматриваемом примере можно писать: набор $W=[4;4;3]$.

В математике запись {4;3} обычно означает множество из элементов 4 и 3, а запись (4;4;3) – *упорядоченную* последовательность из трёх элементов 4, 4 и 3. По этой причине мы решили употреблять для обозначения *неупорядоченного* набора [4;4;3] квадратные скобки.

Ещё раз обратим внимание на то, что {4;3}={3;4} (элементы множества не упорядочены), запись {4;4;3} не имеет смысла (в множестве каждый элемент указывается только один раз; с некоторой натяжкой её можно было бы отождествить с записью {4;3}), $(4;4;3)\neq(4;3;4)$, а [4;4;3]=[4;3;4].

Отметим, что в некоторых статистических исследованиях порядок, в котором расположены анализируемые значения, очень важен. В этих случаях мы будем указывать числа, образующие набор, в круглых скобках. Рассмотрим, например, следующую таблицу, в которой указана средняя температура в некотором регионе в июле за последние 8 лет:

год	2003	2004	2005	2006	2007	2008	2009	2010
температура	21,2°	22,1°	23,2°	22,6°	22,9°	23,6°	24,8°	24,6°

Если мы хотим подсчитать среднюю температуру за 8 лет, то порядок чисел в наборе [21,2; 22,1; 23,2; 22,6; 22,9; 23,6; 24,8; 24,6] не играет никакой роли. Но если мы хотим понять, происходит ли изменение климата, то порядок очень важен (в данном регионе явно видна тенденция роста среднемесячной

температуры в июле) и чтобы подчеркнуть, что рассматривается упорядоченный набор, мы будем записывать его в круглых скобках: (21,2; 22,1; 23,2; 22,6; 22,9; 23,6; 24,8; 24,6).

Резюмируя можно сказать, что любой числовой набор в статистике появляется как результат измерения определённого свойства, характеристики и т.п. объектов из некоторой генеральной совокупности, т.е., в нашей терминологии, определённой переменной. Поэтому когда мы говорим о среднем значении, дисперсии, функции распределения и т.д. числового набора мы фактически говорим о среднем значении, дисперсии, функции распределения и т.д. этой переменной.

1.3.2 Дискретные и непрерывные числовые переменные

Числовые переменные в статистике, в свою очередь, делятся на два типа: *дискретные* и *непрерывные*.

Числовая переменная называется ***дискретной*** (от латинского discretus – разделённый, прерывистый), *если её значения ясно отделены друг от друга.*

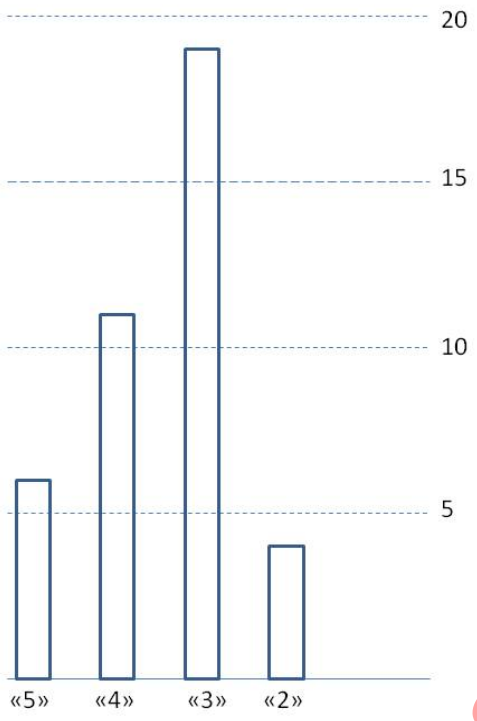
С формальной математической точки зрения дискретные переменные – это переменные, значения которых могут быть перенумерованы (что возможно, в частности, если количество значений конечно).

В рассмотренном выше примере переменная W является дискретной, т.к. может принимать только 4 значения: 2, 3, 4, 5.

Первый шаг в статистической обработке дискретных переменных заключается в вычислении числа повторений каждого значения. Как правило, подсчитывают и (относительную) частоту повторения каждого значения. Результаты представляют в табличной или графической (всевозможные диаграммы) форме. Например, если из n=40 учеников некоторого класса оценку «5» получили t_5=6 человек, оценку «4» – t_4=11 человек, оценку «3» – t_3=19 человек, оценку «2» – t_2=4 человека, то эти данные можно оформить в виде следующей таблицы:

Значение оценки k	Количество учащихся с данной оценкой t_k	Доля $f(k)=\frac{t_k}{n}$
5	6	0,15
4	11	0,275
3	19	0,475
2	4	0,10
Всего:	40	1

Столбиковая диаграмма, соответствующая этой таблице, изображена ниже; она даёт общее представление о характере распределения значений.



На этой диаграмме для каждого возможного значения k рассматриваемой переменной мы отметили абсолютное число повторений этого значения (обозначенное выше через t_k). С равным успехом можно было бы указывать (относительную) частоту $f(k)=\frac{t_k}{n}$ повторения каждого значения. Соответствующая диаграмма отличается от уже построенной только масштабом по оси ординат. Имея в виду дальнейшее изучение теории вероятностей, мы будем обычно отмечать именно частоты. Набор относительных частот $f(k)$, который указывает, какие значения и насколько часто принимает переменная, мы будем называть *распределением* этой переменной. Если нужно явно указать, к какой переменной относится распределение, мы будем добавлять соответствующий нижний индекс, так что, например, $f_X(k)$ – распределение переменной X, $f_Y(k)$ – распределение переменной Y и т.д.

Для дискретных переменных столбики отделены друг от друга небольшим промежутком. Другие виды диаграмм (особенно популярны круговые диаграммы) обычно дают только приблизительное визуальное впечатление о характере распределения и чаще всего используются только для украшения статьи, доклада и пр.

Числовая переменная называется ***непрерывной****, если её значением может быть любое действительное число из некоторого промежутка (по крайней мере, теоретически).*

Непрерывными являются все переменные, которые выражают физические характеристики объекта: длину, массу, объём, температуру, время и т.д.

При физических измерениях мы всегда определяем значения длины, массы и т.д. приближённо, с точностью, зависящей от вида используемого прибора, так

что результат всегда выражается десятичной дробью с конечным числом знаков после запятой. При подходящем выборе единицы измерения результат будет выражается целым числом. Примерно так же обстоит дело и с денежными суммами – они всегда выражаются целым числом копеек. Поэтому с практической точки зрения непрерывных переменных нет вовсе. Однако теоретически измеряемые величины могут выражаться любым положительным действительным числом.

Важно учесть и ещё одно соображение. Дискретные переменные характеризуются тем, что их значения ясно отделены друг от друга. Но разница между 456 руб. 89 коп. и 457руб. 23коп. настолько мала, что с практической точки зрения вряд ли разумно считать эти суммы различными. Если денежные суммы отмечать точками на числовой оси, то даже при выборе единицы масштаба в 1 руб. разница между соседними точками (величиной 0,01) будет настолько мала, что эти дискретные точки сольются в непрерывную линию. Это соображение показывает, что в ряде случаев формально дискретные переменные иногда удобно считать непрерывными.

Для непрерывных переменных любое конкретное значение малоинтересно. Например, при анализе доходов населения совершенно неважно, сколько человек имели в прошлом году суммарный доход в размере 327851 руб. 37 коп. Действительно интересно, например, сколько человек имели годовой доход меньше, чем 150 тыс. руб., т.е. количество малообеспеченных, сколько человек имели годовой доход от 300 тыс. руб. до 600 тыс. руб., т.е. количество достаточно обеспеченных, и т.д. Таким образом, для непрерывных переменных разумно обращать внимание лишь на классы значений, попадающих в тот или иной промежуток. Поэтому первый шаг в статистической обработке непрерывных переменных заключается в разбиении диапазона возможных значений на несколько промежутков (их называют интервалами группировки) и вычислении числа значений, попадающих в каждый интервал. Интервалы группировки должны быть *непересекающимися* (нельзя, чтобы какое-то значение можно было бы отнести к двум разным интервалам) и *исчерпывающими* (т.е. каждое теоретически возможное значение должно быть отнесено в какой-то интервал).

По поводу интервалов группировки необходимо сделать важное замечание, связанное с процедурой округления значений непрерывных переменных. Предположим, например, что мы измеряем время T на дорогу до школы для 10 учеников и получили следующие значения (в минутах): 18; 23; 29; 18; 37; 25; 25; 21; 18; 27. Если время округлялось до целых минут отбрасыванием секунд, то значение 18 означает, что точное время лежит в промежутке $18 \le T < 19$. Если же время округлялось до целых минут по обычному правилу (до ближайшего целого), то значение 18 означало бы, что точное значение времени на дорогу до школы удовлетворяет двойному неравенству $17{,}5 \le T < 18{,}5$. Если мы собираем в один класс все значения от 10 до 19 включительно, то при первом способе округления этот класс характеризуется неравенством $10 \le T < 20$, а при втором – неравенством $9{,}5 \le T < 19{,}5$. В каждом случае длина интервала группировки

равна 10, но сами интервалы разные. В частности, в первом случае центр интервала (он важен для расчёта статистических характеристик по сгруппированным значениям) равен 15, а во втором этим центром будет точка 14,5. Чтобы избежать недоразумений, лучше точно описывать классы группировки двойными неравенствами. В школьных учебниках на процедуру округления числовых данных вообще не обращают внимания, что недопустимо при аккуратном статистическом анализе непрерывных переменных.

Общее визуальное представление о характере распределения значений непрерывной переменной дают гистограмма и функция распределения (она определена также и для дискретных переменных). Им будут посвящены разделы 8 и 9.

1.3.3 Качественные переменные и наборы

Предположим теперь, что интересующее нас свойство ученика – его любимый предмет. Как и четвертная оценка по математике, вообще говоря, этот предмет меняется от ученика к ученику. Поэтому любимый предмет ученика как свойство неопределённого ученика из рассматриваемого класса является *переменной величиной* или короче, *переменной*. Эту переменную можно обозначить каким-нибудь символом, например, F. Точное *значение* этой *переменной* определяется тем, о каком конкретно из учащихся идёт речь. Если, например, Петя больше всего любит историю, Маша – математику, Саша – физику, то этот факт можно выразить следующим образом:

$$F(\text{Петя}) = \text{история}, \; F(\text{Маша}) = \text{математика}, \; F(\text{Саша}) = \text{физика}.$$

Для статистического исследования интересов учащихся не играет роли, кто из учеников любит какой предмет. Важно знать лишь общий *набор* значений рассматриваемой переменной, т.е. любимых предметов школьников. Этим набором является неупорядоченная последовательность из трёх слов: [история, математика, физика].

Поскольку значения нашей переменной F не являются числами, а описывают общее качество рассматриваемого объекта, её называют *нечисловой* или *качественной* переменной, а соответствующий набор – *нечисловым* набором.

Неформально, можно сказать, что

Нечисловой (статистический набор) *– это просто конечная неупорядоченная последовательность слов, букв и т.д.*

Определение значения качественной переменной обычно заключается в анализе для каждого объекта свойства, которое описывает эта переменная, и отнесении этого объекта к одному из нескольких *классов* (в статистике эти классы называют *категориями*). Классы должны быть *взаимно исключающими* (ни один объект не может быть отнесён к двум классам) и *исчерпывающими* (классы должны покрывать все возможности). Этот процесс часто субъективен и

потому по сравнению с количественными переменными статистический анализ качественных переменных труднее, а выводы менее надёжны.

Часто для удобства сбора и обработки информации значения качественных переменных кодируют с помощью чисел. Мы могли бы, например, присвоить математике код 1, физике код 2, истории код 3. Тогда соответствующий набор [история, математика, физика] принял бы вид: [3;1;2]. Однако мы не будем называть его числовым, т.к. обычные действия над числами (сравнение, сложение и т.д.) в этой ситуации не имели бы никакого смысла. Чтобы подчеркнуть это обстоятельство, значения такой переменной F и саму переменную F называют *номинальными* (от латинского nomen – имя).

Обычные школьные оценки «2», «3», «4», «5» в сущности являются кодами для качественных оценок «неудовлетворительно», «удовлетворительно», «хорошо», «отлично». Эти нечисловые оценки мы можем совершенно точно *линейно упорядочить*: «отлично» лучше, чем «хорошо», «хорошо» лучше, чем «удовлетворительно», «удовлетворительно» лучше, чем «неудовлетворительно». Такие переменные называют *порядковыми* или *ординальными* (от латинского ordinatus – расположенный в порядке). Числовые коды оценок соответствуют этому порядку. Поэтому числовые оценки лучше считать качественными ординальными. Рассматривать их как количественные в строгом смысле этого слова можно только с определённой натяжкой: разница между «3» и «2» вовсе не равнозначна разнице между «5» и «4» и её вообще нельзя измерить числом. Суммирование оценок при определении «средней» оценки или «общего» балла можно оправдать только простотой процедуры. Особенно нелепо это при суммировании оценок по разным экзаменам, что приводит, например, к выводу, что 90 баллов по ЕГЭ и 50 баллов по независимому экзамену, проводимому МГУ (сумма баллов равна 140), лучше, чем 60 баллов по ЕГЭ и 75 баллов по независимому экзамену, проводимому МГУ (сумма баллов равна 135).

При статистическом исследовании чрезвычайно важно ясно понимать тип анализируемой переменной, так как от этого зависит сам характер этого исследования (как представлять данные, какие характеристики вычислять и т.д.)

При изложении описательной статистики мы будем говорить о произвольных наборах, образованных перечислением значений некоторой переменной (которые характеризуют некоторое свойство предметов определённой совокупности). Такой набор мы обычно будем обозначать той же заглавной буквой латинского алфавита, что и соответствующую переменную, а значения переменной для индивидуальных объектов из совокупности – соответствующей прописной буквой с нижним индексом, указывающим на номер объекта в совокупности при некоторой их нумерации.

Например, в приведённых выше примерах можно было бы писать: набор W – это неупорядоченная последовательность $[w_1, w_2, w_3]$, состоящая из элементов $w_1 = 4, w_2 = 4, w_3 = 3$, набор F – это неупорядоченная последовательность $[f_1, f_2, f_3]$, состоящая из элементов $f_1 =$ история, $f_2 =$ математика, $f_3 =$ физика.

1.4 Вариационный ряд

Если элементы числового статистического набора упорядочить по возрастанию, то получившаяся упорядоченная по возрастанию элементов версия набора является последовательностью в обычном смысле и называется *вариационным рядом*. Например, для набора [4;3;4] вариационный ряд имеет вид: (3;4;4).

В математической статистике k-й член вариационного ряда набора $[x_1,x_2,\ldots,x_n]$ обычно обозначают $x_{(k)}$; он является определённой функцией от всего этого набора: $x_{(k)}=v_k(x_1,\ldots,x_n)$. В частности, $x_{(1)}=\min(x_1,x_2,\ldots,x_n)$ – наименьшее число набора, $x_{(n)}=\max(x_1,x_2,\ldots,x_n)$ – наибольшее число набора.

Если мы сразу будем рассматривать не набор, а его вариационный ряд, то для его записи мы будем использовать обычное обозначение для последовательности (с круглыми скобками): $X=(x_1,x_2,\ldots,x_n)$.

Хотя данное выше определение вариационного ряда применимо и к качественному ординальному набору, содержательная теория существует только для числовых наборов.

1.5 Двумерные переменные

При изучении зависимости одного свойства объекта от другого (например, веса человека от его роста) или взаимозависимости двух разных свойств объектов (например, оценки по ЕГЭ по математике и оценки независимого университетского экзамена по математике) соответствующие переменные нужно рассматривать не изолированно, а в неразрывной связи друг с другом. Такие переменные называют *двумерными*. Значения двумерной переменной являются парами, а соответствующий набор является неупорядоченной последовательностью пар из чисел, слов, букв и т.д. (в зависимости от того, числовыми или качественными являются рассматриваемые свойства). Например, при изучении зависимости веса человека от его роста набор мог бы выглядеть так: [(64кг,158см), (68кг,173см),…].

Более формально, если объекты некоторой генеральной совокупности характеризуются двумя переменными, X и Y, то для каждого объекта i соответствующие значения этих переменных, x_i и y_i, должны рассматриваться как неразрывная пара (x_i,y_i). Ещё раз подчеркнём, что каждое значение (x_i,y_i) двумерной переменной относится к *одному* объекту какой-то генеральной сововупности. Если переменные X и Y описывают объекты двух *разных* генеральных совокупностей, состоящих из одинакового числа n объектов, то можно формально образовать из наборов $[x_1,\ldots,x_n]$ и $[y_1,\ldots,y_n]$ значений этих переменных набор $[(x_1,y_1),\ldots,(x_n,y_n)]$ пар (x_i,y_i). Но этот набор не будет иметь

никакого содержательного смысла. В частности, к нему нельзя применять теорию статистической зависимости, которую мы опишем ниже. Подробно мы будем говорить о двумерных переменных в разделе 10.

1.6 Задачи

Задача 1.1. Приёмная комиссия механико-математического факультета МГУ для принятия решения о зачислении абитуриентов использует данные о результатах каждого абитуриента по ЕГЭ по русскому языку, математике, физике и дополнительному экзамену по математике, проводимому экзаменационной комиссией МГУ. Какие из этих данных первичные, а какие вторичные? В чём проявляется различие между первичными и вторичными данными в этой ситуации?

Решение. Оценки (тестовые баллы) по ЕГЭ являются вторичными данными, а результаты проводимых МГУ независимых вступительных испытаний – первичными. Оценки по ЕГЭ приёмная комиссия получает без затраты каких-либо ресурсов, в то время как организация и проведение независимых экзаменов требует много времени, материальных и человеческих ресурсов (составлением экзаменационных вариантов, проведением экзаменов, проверкой работ занимается большое число профессоров и сотрудников МГУ). Однако результаты независимого университетского экзамена вызывают гораздо больше доверия у преподавателей МГУ (хотя по Правилам приёма в МГУ учитываются наравне с результатами ЕГЭ).

Задача 1.2. Учитель математики ставит ученику Пете Сидорову оценки в дневник и в классный журнал. Являются ли эти данные об успеваемости статистическими или нет? Если да, то какую генеральную совокупность они характеризуют?

Решение. Оценки в дневнике характеризуют работу Пети на уроке по математике в конкретный день. Их цель – регулярно и своевременно информировать его родителей и потому данные об оценках в дневнике не являются статистическими. Те же самые оценки в школьном журнале используются для выставления оценок за четверть. Обычно четвертная оценка – среднее арифметическое всех оценок ученика по предмету. Хотя оценки в журнале также характеризуют работу Пети по математике в конкретный день, их основная цель – получить общее представление об успеваемости Пети за рассматриваемую четверть и потому данные об оценках из классного журнала можно считать статистическими. Генеральная совокупность, которую характеризуют эти оценки – это весь набор уроков по математике за четверть. Отметим, что Петя может использовать оценки из дневника для того, чтобы оценить свою успеваемость за прошедшую часть четверти и, в случае необходимости, предпринять необходимые усилия для улучшения ожидаемой

оценки за четверть. В этом случае оценки из дневника следует рассматривать как статистические данные.

Задача 1.3. Данные об учениках вашего класса, которые вы собираете в ходе статистического исследования, включают сведения о возрасте (переменная A), числе братьев/сестёр (переменная F), месте проведения отдыха летом (город, деревня, море – переменная P). Определите тип каждой из этих переменных.

Решение. Переменная A является количественной (или, что то же самое, числовой) и непрерывной. Хотя обычно люди считают свой возраста целыми годами, возраст увеличивается непрерывно и в конце 40 минутного урока вы на 40 минут старше, чем в его начале. Переменная F также является числовой, но дискретной. Переменная P является качественной номинальной.

Задача 1.4. Готовясь к ЕГЭ по математике, ученик ежедневно записывает число решённых задач (переменная N) и время, которое он потратил на их решение (переменная T). Какая из этих числовых переменных является непрерывной, а какая – дискретной?

Решение. Переменная T является непрерывной, переменная N – дискретной.

Задача 1.5. Готовясь к ЕГЭ по математике, ученик ежедневно решает по одной задаче и записывает время, которое он потратил на её решение (переменная T), и раздел математики (алгебра, геометрия, тригонометрия, анализ, статистика), к которой можно отнести эту задачу (переменная S). Какая из этих переменных является числовой, а какая – качественной? Какие сложности могут возникнуть с определением значения качественной переменной в данном случае?

Решение. Переменная T является количественной (числовой) непрерывной, переменная S – качественной номинальной. Для решения сложной задачи часто приходится использовать разные разделы математики, которые к тому же могут и не совпадать с типом задачи (например, уравнение может решаться с помощью исследования функции). В этом случае однозначно определить, к какой теме следует отнести задачу, довольно тяжело и ответ на этот вопрос субъективен.

Задача 1.6. Владелец небольшого магазина по продаже мужских рубашек после каждой покупки записывает цвет купленной рубашки (переменная C) и её размер (переменная S). Охарактеризуйте эти переменные. Зачем ему могут быть нужны эти данные? Являются ли они первичными или вторичными?

Решение. Переменная C является качественной номинальной, переменная S – качественной ординальной. Поскольку эти переменные фиксируются одновременно, речь идёт о двумерной переменной. Используя эти данные, владелец магазина может установить, сколько рубашек определённого цвета и

размера он продаёт за неделю и правильно планировать свои закупки (если он закупает товар раз в неделю). Эти данные являются первичными.

Задача 1.7. Кондитерская фабрика разработала рецепт нового вида печенья. Чтобы принять окончательное решение о начале массового производства, в большом магазине всем желающим предлагали попробовать это печенье и оценить его вкус по пятибалльной шкале: «1» означает «ужасный вкус», «2» – «скорее плохой вкус, чем хороший», «3» – «печенье как печенье», «4» – «хороший вкус», «5» – «восхитительный вкус». В результате были получены следующие данные:

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

Являются эти данные первичными или вторичными? Охарактеризуйте тип измеряемой переменной.

Решение. Данные являются первичными. Измеряемая переменная – вкус печенья. Она качественная ординальная.

Задача 1.8. Кондитерская фабрика решила выпускать новый вид печенья и разработала для него четыре варианта упаковки. Чтобы принять окончательное решение о виде упаковки, в большом магазине всем желающим бесплатно предлагали попробовать это печенье и в обмен на это указать упаковку, которая нравится им больше всего. В результате были получены следующие данные:

вид упаковки	красные розы	полевые цветы	голубое небо с облаками	геометрический рисунок
число ответов	12	16	63	9

Являются эти данные первичными или вторичными? Какая переменная измерялась в ходе этого исследования? Охарактеризуйте её тип.

Решение. Данные являются первичными. Измеряемая переменная – внешний вид упаковки, который больше всего нравится человеку. Она качественная номинальная. Хотя в таблице стоят числа, указывающие, сколько раз потенциальные покупатели отдали предпочтение тому или иному виду упаковки, это не означает, что измерялась эта числовая переменная. Эти числа появились при табличном способе представления полученных данных о виде упаковки. Отсюда, в частности, следует, что вычисление среднего арифметического чисел из таблицы – совершенно бессмысленная операция.

Задача 1. 9. Для того, чтобы установить справедливую плату за страховку от пожара в Москве специалисты страховой компании решили выяснить, насколько часто происходят пожары в жилых помещениях в разных административных округах города. На сайте Министерства по чрезвычайным ситуациям они нашли следующие данные о числе пожаров в 2010 году:

округ	**пожаров**	**округ**	**пожаров**
центральный	654	юго-западный	503
северо-восточный	740	западный	616
восточный	842	северо-западный	441
юго-восточный	583	северный	434
южный	516	г.Зеленоград	151

Являются эти данные первичными или вторичными? Какая переменная измерялась в ходе этого исследования? Охарактеризуйте её тип. Объясните, почему на основе только этих данных нельзя принять обоснованное решение? Какие ещё данные имело бы смысл собрать?

Решение. Данные являются вторичными (тем не менее, нет никаких сомнений в их достоверности). Измеряемая переменная – число пожаров в административном округе. Она количественная дискретная. Плата за страховку устанавливается не для всего округа, а для отдельной квартиры. Поэтому нужно знать общее число квартир в этих округах. Полезно также иметь данные о типах домов (вид перекрытий, тип плит на кухнях т.д.). Возможно, что риск пожара зависит от числа жителей. Для анализа этой проблемы нужны данные о числе жителей в разных округах.

Задача 1.10. При изучении радиоактивности во многих учебниках приводятся данные о результатах эксперимента, проведённого английскими учёными Резерфордом, Чедвиком и Эллисом (см. таблицу). В ходе их опыта специальный прибор регистрировал число частиц, испускаемых небольшим количеством радиоактивного вещества за фиксированный короткий промежуток времени 7,5 сек. Они провели 2608 наблюдений за одним и тем же кусочком вещества и выяснили, что это число меняется от опыта к опыту (хотя промежуток времени фиксирован и с веществом ничего не происходит).

число частиц	число опытов	число частиц	число опытов
0	57	6	273
1	203	7	139
2	383	8	45
3	525	9	27
4	532	10 и больше	16
5	408	Всего:	2608

Являются эти данные первичными или вторичными? Какая переменная измерялась в ходе этого исследования? Охарактеризуйте её тип.

Решение. Данные являются вторичными (тем не менее, они абсолютно достоверны). Измеряемая переменная – число частиц, зафиксированных счётчиком за 7,5 сек.. Она количественная дискретная.

Задача 1.11. К какой из переменных, рассмотренных в задачах 2-5, может относиться набор [3;0;1;1;2]? Является ли этот набор числовым? Найдите вариационный ряд для этого набора, если он числовой.

Решение. Этот набор может являться значениями переменной F (число братьев/сестёр) из задачи 2; значениями переменной N (число задач, решённых за день) из задачи 3; значениями переменной S (раздел математики) из задачи 4, если основные разделы математики закодированы, например, так: алгебра – «0», геометрия – «1», тригонометрия – «2», анализ – «3», статистика – «4»; значениями переменной C (цвет рубашки) из задачи 5, при соответствующем кодировании цвета. В двух первых случаях набор является дискретным числовым; в двух последних – качественным номинальным. Вариационным рядом будет последовательность (0;1;1;2;3).

Задача 1.12. В рекламном проспекте «Кипр-08», изданном одной из туристических компаний, в разделе «Полезная информация» приведена следующая таблица о погоде на Кипре

Месяц	Дождливых дней		Месяц	Дождливых дней
Январь	13		Июль	Нет
Февраль	9		Август	Нет
Март	7		Сентябрь	1
Апрель	5		Октябрь	4
Май	2		Ноябрь	6
Июнь	нет		Декабрь	11

и написано: «На Кипре 330 солнечных дней в году!»

(1) Используя данные, приведённые в таблице, подсчитайте общее число дождливых дней на Кипре в году.

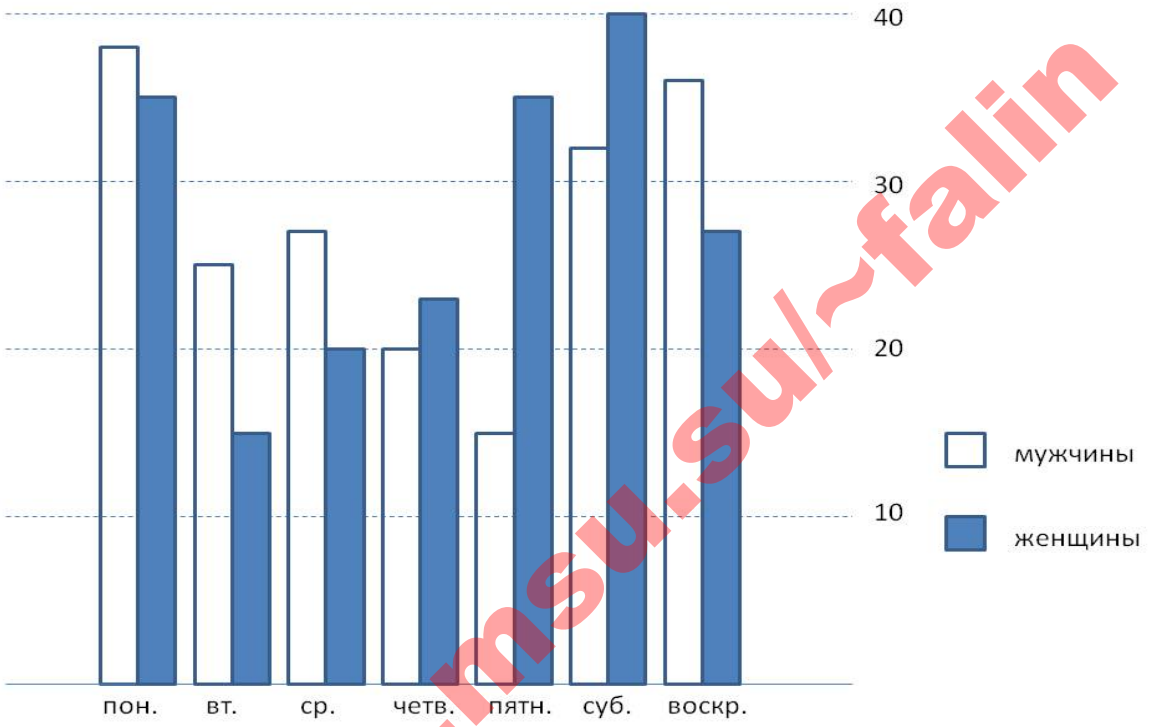
(2) Проанализируйте полученный результат.

Решение. (1) Общее число дождливых дней в году на Кипре равно сумме всех чисел, стоящих во втором и четвёртом столбцах таблицы:

$$13+9+7+5+2+1+4+6+11=58.$$

(2) Этот результат означает, что число солнечных дней на Кипре равно $365-58=307$ (для обычного года) или $366-58=308$ (для високосного года). В каждом случае мы считаем, что день на Кипре либо дождливый, либо солнечный, хотя, возможно, и с небольшой облачностью (на этом острове не бывает дней, когда всё небо затянуто облаками, а дождей нет). Эти данные противоречат рекламному объявлению: «На Кипре 330 солнечных дней в году!», так что доверять «полезной информации» от туристической компании нельзя. При выборе месяца для поездки имеет смысл найти альтернативные источники информации.

Задача 1.13. Владелец газетного киоска, решил выяснить, кто чаще покупает газеты – мужчины или женщины. С этой целью он на протяжении недели проводил статистическое исследование и фиксировал общее число газет, купленных за день мужчинами, и общее число газет, купленных за день женщинами. Результаты этого исследования показаны на столбиковой диаграмме:



(1) Сколько газет купили женщины в понедельник?

(2) Сколько газет купили мужчины в пятницу?

(3) Представьте данные о продажах в виде таблицы.

(4) Владелец киоска думает, что женщины покупают газеты чаще, чем мужчины. Подтверждают ли данные о продажах это мнение?

Решение. (1) В понедельник женщины купили 35 газет. (2) В пятницу мужчины купили 15 газет. (3) Следующая таблица соответствует диаграмме:

	пон.	вт.	ср.	четв.	пятн.	суббота	воскр.
Мужчины	38	25	27	20	15	32	36
Женщины	35	15	20	23	35	40	27

(4) За неделю мужчины купили 193 газеты, а женщины – 195. Хотя женщины и купили на 2 газеты больше, чем мужчины, по отношению к общему числу газет, купленных женщинами, эта разница составляет лишь $\frac{2}{195} \approx 0{,}01 = 1\%$. Столь малое относительное отличие можно объяснить не большим интересом женщин к чтению газет, а какими-то случайными факторами. Не исключено, что на следующей неделе мужчины купят больше газет, чем женщины.

Задача 1.14. В классе 40 учеников. Во время урока физкультуры они разбиты на три группы: школьники из первой группы играют в баскетбол, из второй – в футбол, а третья группа занимается лёгкой атлетикой. Данные о числе школьников в этих группах содержатся в следующей круговой диаграмме.

С помощью транспортира определите число школьников в каждой группе.

Решение. Угол сектора диаграммы, соответствующего второй группе (футбол), очевидно, равен 180°. Он составляет $\frac{180^\circ}{360^\circ}=\frac{1}{2}$ развёрнутого угла. Поэтому в футбол играет половина всех школьников, т.е. 20 человек. Угол сектора диаграммы, соответствующего первой группе (баскетбол), можно измерить только с помощью транспортира. Он равен 36°, что составляет $\frac{36^\circ}{360^\circ}=\frac{1}{10}$ развёрнутого угла. Поэтому в футбол играет десятая часть всех школьников, т.е. 4 человека. Оставшиеся школьники, $40-20-4=16$ человек, занимаются лёгкой атлетикой. Это же число можно получить и из нашей диаграммы. Угол сектора диаграммы, соответствующего третьей группе (лёгкая атлетика), равен 144°. Он составляет $\frac{144^\circ}{360^\circ}=\frac{2}{5}$ развёрнутого угла. Поэтому в футбол играет $\frac{2}{5}$ всех школьников, т.е. $\frac{2}{5}\times 40=16$ человек.

Задача 1.15. По заданию учителя в ноябре школьник проводил метеорологические наблюдения и, в частности, записывал температуру воздуха на улице. Часть из его результатов приведена в таблице.

день	3.11	4.11	5.11	6.11	7.11	8.11	9.11	10.11	11.11	12.11
температура	–5°	–4°	–5°	–7°	–8°	–9°	–6°	–10°	–6°	–2°

(1) Какой день из указанных в таблице был самым холодным?
(2) Какова была температура воздуха на улице в этот день?
(3) Каким является это значение температуры в ряду значений температур – наибольшим или наименьшим?
(4) Найдите размах температур за период с 3 по 12 ноября.

Решение. Самым холодным было 10 ноября, когда температура была равна -10°. Это значение температуры является наименьшим в ряду чисел –5, –4, –5, –7, –8, –9, –6, –10, –6, –2.

Самым «тёплым» было 12 ноября, когда температура была равна -2°. Размах набора чисел – это разность между наибольшим и наименьшим числами из этого набора. В нашем случае размах равен (–2) – (–10) = 8 (градусов).

Задача 1.16. Торговая компания хочет понять, сколько денег тратят её покупатели за один визит в магазин. Первые 32 покупателя пробили чеки на следующие суммы (в рублях): 108; 54; 62; 74; 40; 38; 85; 92; 64; 25; 80; 143; 50; 63; 38; 79; 155; 28; 61; 83; 62; 42; 76; 47; 70; 83; 35; 192; 140; 52; 64;88

Компанию не интересует точная сумма S, указанная в чеке; для неё покупки делятся на мелкие ($10 \le S < 50$), средние ($50 \le S < 100$), крупные ($100 \le S < 200$); при этом компания имеет в виду, что никто из её покупателей не тратит меньше 10 рублей и (за крайне редким исключением) больше 200 рублей.

(1) Заполните следующую таблицу:

Покупка	Встретилось в списке	Всего
Мелкая		
Средняя		
Крупная		

При этом сначала в графе «Встретилось в списке» отмечайте каждую покупку чёрточкой, формируя из них квадрат с диагоналями: (таким образом, фигура ∟ символизирует 2 покупки, фигура □ – 4, а фигура ☒– 6 покупок).

(2) Какие покупки, мелкие, средние или крупные, делаются чаще всего?

(3) Что можно сказать о среднем размере покупки на основе данных этой таблицы (не используя исходные данные о точной сумме каждой покупки)?

(4) Оцените средний размер покупки, считая, что размер каждой покупки равен среднему значению соответствующего диапазона (т.е. 30 руб. для мелких, 75 руб. для средних, 150 руб. для крупных). Сравните его с точным средним размером покупки.

Решение. (1) Таблица с результатами подсчётов выглядит следующим образом:

Покупка	Встретилось в списке	Всего
Мелкая	☒∟	8
Средняя	☒☒☒ |	19
Крупная	◪	5

(2) Из этой таблицы ясно видно, что наиболее распространены средние покупки.

(3) Про сумму S одной мелкой покупки мы знаем лишь то, что $10 \le S < 50$. Про общую сумму $S_{\text{м}}$, потраченную на 8 мелких покупок, мы можем сказать лишь то, что $80 \le S_{\text{м}} < 400$. Аналогично, относительно общей суммы $S_{\text{с}}$, потраченной на все 19 средних покупок мы можем сказать лишь то, что $950 \le S_{\text{с}} < 1900$, а про общую сумму $S_{\text{к}}$, потраченную на все 5 крупных покупок, мы можем сказать лишь то, что $500 \le S_{\text{к}} < 1000$. Складывая три этих неравенства, для общих расходов $S_{\text{общ.}} = S_{\text{м}} + S_{\text{с}} + S_{\text{к}}$ мы получим двойное неравенство: $1530 \le S_{\text{общ.}} < 3300$. Средняя сумма одной покупки равна $\frac{S_{\text{общ.}}}{32}$. На основе данных таблицы мы можем утверждать лишь то, что эта величина находится в пределах от $\frac{1530}{32} \approx 47$ руб. 81 коп. до $\frac{3300}{32} \approx 103$ руб. 13 коп.

(4) Если считать, что размер каждой покупки равен среднему значению соответствующего диапазона (т.е. 30 руб. для мелких, 75 руб. для средних, 150 руб. для крупных), то общая сумма, израсходованная покупателями может быть оценена как $30 \cdot 8 + 75 \cdot 19 + 150 \cdot 5 = 2415$ (руб.). Поэтому средний размер покупки приближённо равен 75 руб. 47 коп. Если использовать полные исходные данные о расходах покупателей, то $S_{\text{общ.}} = 2373$. Значит, средняя сумма одной покупки равна $\frac{S_{\text{общ.}}}{32} \approx 74$ руб. 16 коп., так что точность приближения очень высока.

2. Экстремальные свойства среднего значения и медианы

2.1 Введение

В этом разделе мы покажем, что наиболее важные в описательной статистике меры положения набора чисел, среднее значение и медиана, могут быть охарактеризованы как значения, при которых достигаются наименьшие значения определённых функций. Эти функции имеют смысл «расстояния» от некоторой точки на числовой прямой до набора как целого, что позволяет естественно определить среднее значение и медиану числового набора и глубже понять их свойства. Подобные свойства математического объекта, которые характеризуют этот объект как решение некоторой экстремальной задачи (т.е. задачи исследования некоторой функции на наибольшее/наименьшее значение) в математике называют экстремальными свойствами этого объекта.

2.2 Среднее значение

Пусть $X=[x_1,\ldots,x_n]$ – набор из n чисел и a – некоторая точка на числовой оси. Разность x_i-a называется отклонением i-го числа из набора X от a. Соответственно, сумма $(x_1-a)+\ldots+(x_n-a)=(x_1+\ldots+x_n)-na$ является суммарным отклонением всех чисел набора X от a, а её модуль $\left|(x_1+\ldots+x_n)-na\right|=\left|na-(x_1+\ldots+x_n)\right|$ – абсолютным суммарным отклонением. Это абсолютное суммарное отклонение можно интерпретировать как «расстояние» от точки a до набора $X=[x_1,\ldots,x_n]$ как целого. Отметим, что поскольку сумма не зависит от порядка слагаемых, неважно, какую версию (неупорядоченного) набора мы рассматриваем.

Рассмотрим теперь абсолютное суммарное отклонение как функцию f(a) от параметра a:

$$\mathrm{f}(a)=\left|na-(x_1+\ldots+x_n)\right|.$$

Определение. Значение M_X переменной a, при котором достигается наименьшее значение функции f(a), назовём средним значением набора чисел $X=[x_1,\ldots,x_n]$.

Таким образом, мы определяем среднее значение как значение вспомогательной переменной a, при котором абсолютное суммарное отклонение (своеобразное «расстояние» от точки a до набора $X=[x_1,\ldots,x_n]$ как целого) достигает наименьшего значения.

Из этого определения ещё не ясно, существует ли среднее значение, а если существует, то однозначно ли оно определено и по какой формуле может быть вычислено. Ответ на эти вопросы даёт следующая теорема.

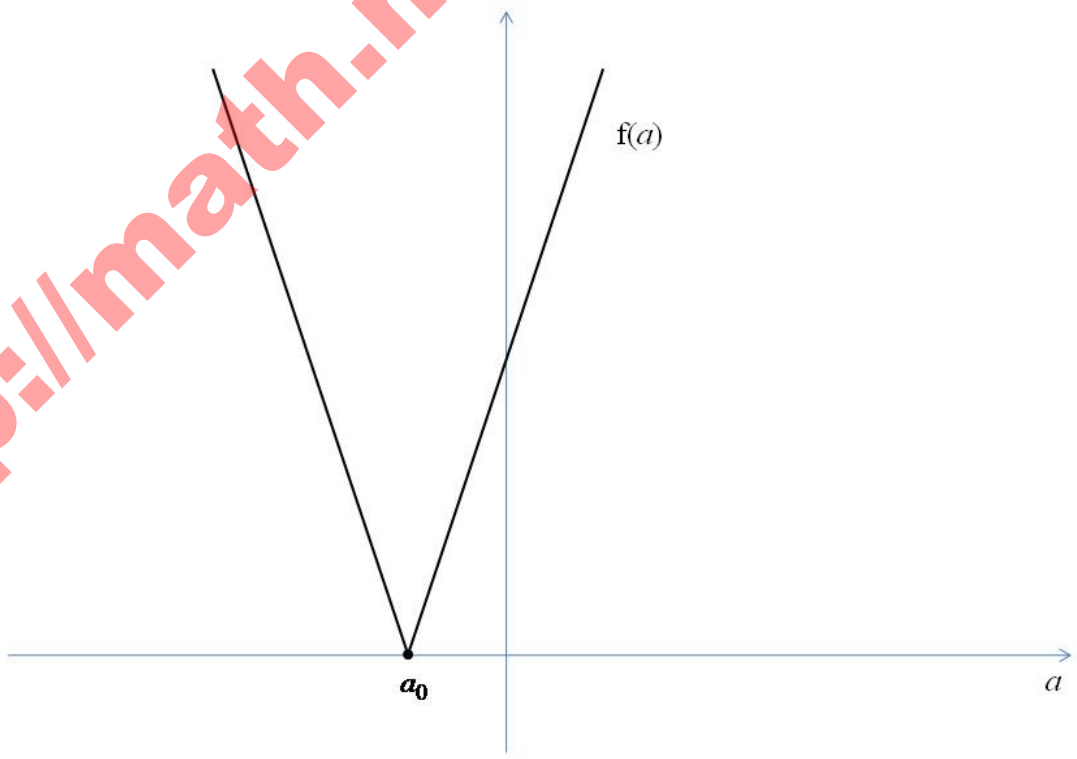
Теорема 2.1. *Для любого набора чисел* $X=\left[x_1,\ldots,x_n\right]$ *среднее значение* M_X *существует, однозначно определено и может быть вычислено по формуле:*

$$M_X=\frac{x_1+\ldots+x_n}{n}.$$

Таким образом, среднее значение набора чисел совпадает со средним арифметическим этих чисел.

Доказательство. Функция f(a) неотрицательна при всех значениях переменной a (как модуль некоторого выражения). При этом значение 0 достигается тогда и только тогда, когда выражение под знаком модуля равно 0: $na-\left(x_1+\ldots+x_n\right)=0$. Это уравнение (относительно переменной a) имеет и притом единственный корень $a_0=\frac{x_1+\ldots+x_n}{n}$. Этот корень и является единственным значением, при котором достигается наименьшее значение функции f(a).□

Рис.2.1



Справедливость теоремы 2.1 очевидна из рис.2.1, на котором показан примерный вид графика функции f(a) (при $a\ge a_0$ эта функция совпадает с

линейной функцией $y = na - (x_1 + \ldots + x_n)$, а при $a \le a_0$ – с линейной функцией $y = -na + (x_1 + \ldots + x_n)$. Проведённые рассуждения, кроме того, показывают, что среднее значение M_X набора чисел можно определить и как значение вспомогательной переменной a, при котором абсолютное суммарное отклонение (своеобразное «расстояние» от точки a до набора $X = [x_1, \ldots, x_n]$ как целого) равно 0.

Отметим, что среднее значение набора чисел может появляться и как решение других экстремальных задач. Рассмотрим, например, в качестве меры «расстояния» от точки a до набора $X = [x_1, \ldots, x_n]$ суммарное квадратичное отклонение $\mathrm{h}(a) = (x_1 - a)^2 + \ldots + (x_n - a)^2$. Поскольку сумма не зависит от порядка слагаемых, неважно, какую версию (неупорядоченного) набора мы рассматриваем.

Раскрывая скобки мы можем записать формулу, которая определяет функцию $\mathrm{h}(a)$, в виде:

$$\mathrm{h}(a) = na^2 - 2(x_1 + \ldots + x_n)a + (x_1^2 + \ldots + x_n^2).$$

Таким образом, $\mathrm{h}(a)$ является квадратичной функцией с положительным старшим коэффициентом. Следовательно, $\mathrm{h}(a)$ имеет наименьшее значение, которое достигается в той же точке $a_0 = \dfrac{x_1 + \ldots + x_n}{n} \equiv M_X$, что и наименьшее значение функции $\mathrm{f}(a)$. При этом само наименьшее значение функции $\mathrm{h}(a)$ равно

$$\mathrm{h}(M_X) = (x_1 - M_X)^2 + \ldots + (x_n - M_X)^2 = nD_X,$$

где $D_X \equiv \dfrac{(x_1 - M_X)^2 + \ldots + (x_n - M_X)^2}{n} \equiv M_{(X - M_X)^2}$ – дисперсия набора $X = [x_1, \ldots, x_n]$.

Замечание. Пусть N – количество различных чисел в основном наборе, а $y_1 < \ldots < y_N$ – сами эти числа, занумерованные в порядке возрастания. Пусть, далее, натуральное число t_{y_i} указывает, сколько раз число y_i встречается в наборе (ясно, что $t_{y_1} + \ldots + t_{y_N} = n$). Группируя вместе одинаковые числа основного набора $X = [x_1, \ldots, x_n]$, мы получим, что сумма $x_1 + \ldots + x_n$ равна $t_{y_1} \cdot y_1 + \ldots + t_{y_N} \cdot y_N$. Соответственно, среднее значение M_X равно:

$$M_X = \frac{x_1 + \ldots + x_n}{n} = \frac{t_{y_1} \cdot y_1 + \ldots + t_{y_N} \cdot y_N}{n} = y_1 \cdot \frac{t_{y_1}}{n} + \ldots + y_N \cdot \frac{t_{y_N}}{n}$$
$$= y_1 \cdot f_X(y_1) + \ldots + y_N \cdot f_X(y_N).$$

Эти преобразования можно записать и немного короче, если использовать символ суммирования $\sum$:

$$M_X = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}\sum_{k=1}^{N} t_{y_k} y_k = \sum_{k=1}^{N} y_k \frac{t_{y_k}}{n} = \sum_{k=1}^{N} y_k f_X(y_k).$$

Полученная формула удобна для расчёта среднего значения в случае, когда данные представлены в сгруппированной форме и известно распределение $f_X(y_1),\ldots,f_X(y_N)$ значений переменной.

Пример. После письменного экзамена по математике на механико-математический факультет МГУ в 1968 году выяснилось, что 57,1% абитуриентов получили оценку «2», 34,6% – оценку «3», 6,6% – оценку «4», 1,7% – оценку «5». Определите среднюю оценку.

Решение. В рассматриваемом примере генеральной совокупностью является группа из всех абитуриентов, принимавших участие в экзамене (их было 3035 человек). Изучаемая переменная – это оценка по математике. Возможные значения этой переменной: 2, 3, 4, 5. Распределение: $f(2)=0{,}571$, $f(3)=0{,}346$, $f(4)=0{,}066$, $f(5)=0{,}017$. Поэтому среднее значение есть:

$$M = 2\cdot f(2) + 3\cdot f(3) + 4\cdot f(4) + 5\cdot f(5) = 2{,}529 \approx 2{,}5.$$

Итак, средняя оценка за экзамен – «2,5».

Подобные расчёты удобно оформлять в виде следующей таблицы из трёх строк, занумерованных (1), (2), (3):

(1)		Оценка, k	2	3	4	5		Сумма всех чисел строки
(2)		Частота, $f(k)$	0,571	0,346	0,066	0,017		1,000
(3)		(1)x(2)	1,142	1,038	0,264	0,085		2,529

В строке (1) указаны возможные значения анализируемой переменной, в строке (2) под каждым значением указана его частота, в строке (3) указаны попарные произведения $k\cdot f(k)$. В последнем столбце подсчитана сумма всех чисел из соответствующей строки. Отметим, что сумма всех частот обязательно равна 1. Тем не менее, мы рекомендуем подсчитывать эту сумму, чтобы выловить возможные ошибки. В правом нижнем углу стоит искомое среднее значение.

Аналогичную формулу можно получить и для дисперсии:

$$D_X \equiv \frac{(x_1 - M_X)^2 + \ldots + (x_n - M_X)^2}{n} = \frac{t_{y_1}(y_1 - M_X)^2 + \ldots + t_{y_N}(y_N - M_X)^2}{n}$$

$$= (y_1 - M_X)^2 \cdot \frac{t_{y_1}}{n} + \ldots + (y_N - M_X)^2 \cdot \frac{t_{y_N}}{n}$$

$$= (y_1 - M_X)^2 \cdot f_X(y_1) + \ldots + (y_N - M_X)^2 \cdot f_X(y_N).$$

Её вывод также можно записать немного короче, если использовать символ суммирования $\sum$:

$$D_X \equiv \frac{1}{n}\sum_{i=1}^{n}\left(x_i - M_X\right)^2 = \frac{1}{n}\sum_{k=1}^{N} t_{y_k}\left(y_k - M_X\right)^2 = \sum_{k=1}^{N}\left(y_k - M_X\right)^2 \frac{t_{y_k}}{n}$$
$$= \sum_{k=1}^{N}\left(y_k - M_X\right)^2 f_X(y_k).$$

2.3 Медиана

В предыдущем пункте (при определении среднего значения) мы рассматривали в качестве меры «расстояния» от точки a до набора $X = [x_1,\ldots,x_n]$ как целого абсолютное суммарное отклонение $\mathrm{f}(a) = \left|(x_1 - a) + \ldots + (x_n - a)\right|$. С равным основанием в качестве меры этого «расстояния» можно взять сумму абсолютных отклонений чисел из набора X от точки a: $\mathrm{g}(a) = |x_1 - a| + \ldots + |x_n - a|$. Поскольку эта сумма не зависит от порядка слагаемых, неважно, какую версию (неупорядоченного) набора мы рассматриваем.

Определение. Значение μ_x переменной a, при котором достигается наименьшее значение функции $\mathrm{g}(a)$, назовём медианой набора чисел $X = [x_1,\ldots,x_n]$.

Таким образом, мы определяем медиану как значение вспомогательной переменной a, при котором сумма абсолютных отклонений (слегка изменённый по сравнению с предыдущим пунктом вариант «расстояния» от точки a до набора $X = [x_1,\ldots,x_n]$ как целого) достигает наименьшего значения. В этом смысле медиана (как мера положения чисел из основного набора на числовой оси) ничем не лучше и не хуже, чем среднее значение.

Как и при определении среднего значения, из этого определения ещё не ясно, существует ли медиана, а если существует, то однозначно ли она определена и по какой формуле может быть вычислена. Ответ на эти вопросы даёт нижеследующая Теорема 2.2.

Теорема 2.2. *Для любого набора чисел $X = [x_1,\ldots,x_n]$ медиана μ_X существует.*

Если количество n чисел в основном наборе $X = [x_1,\ldots,x_n]$ – нечётно, т.е. $n = 2k-1$ для некоторого натурального k, то медиана однозначно определена и может быть вычислена по формуле: $\mu_X = v_k$, где $v_k \equiv x_{(k)}$ – k-й член вариационного ряда (т.е. μ_X – это член вариационного ряда, стоящий точно посередине).

Если количество n чисел в основном наборе $X = [x_1,\ldots,x_n]$ – чётно, т.е. $n = 2k$ для некоторого натурального k, то медиана, хотя и существует, вообще говоря, не определена однозначно. В качестве медианы может быть

взято любое число из отрезка $[v_k; v_{k+1}]$, *где* $v_k \equiv x_{(k)}$ *и* $v_{k+1} \equiv x_{(k+1)}$ – *соответственно k-й и* $(k+1)$-й *члены вариационного ряда.*

Доказательство. Пусть N – количество различных чисел в основном наборе, $y_1 < \ldots < y_N$ – сами эти числа, занумерованные в порядке возрастания, натуральное число t_{y_i} указывает, сколько раз число y_i встречается в наборе. Чтобы немного упростить дальнейшие формулы, будем вместо t_{y_i} писать t_i (ниже, явно это оговаривая, мы также будем использовать это сокращение). Тогда вариационный ряд $\boldsymbol{v} = (v_1, \ldots, v_n)$ имеет вид:

$$\left(\underbrace{y_1, \ldots, y_1}_{t_1}, \underbrace{y_2, \ldots, y_2}_{t_2}, \ldots, \underbrace{y_N, \ldots, y_N}_{t_N} \right). \qquad (2.1)$$

Ясно, что $t_1 + \ldots + t_N = n$, $t_1 y_1 + \ldots + t_N y_N = x_1 + \ldots + x_n = n \cdot M_x$.

Поскольку сумма не зависит от порядка слагаемых, формулу, которая определяет функцию g(a), можно записать в виде:

$$\mathrm{g}(a) = t_1 |y_1 - a| + \ldots + t_N |y_N - a|. \qquad (2.2)$$

Отметим числа $y_1, \ldots, y_N$ на числовой оси. В результате мы получим $N+1$ промежутков:

$$(-\infty; y_1], [y_1; y_2], \ldots, [y_{N-1}; y_N], [y_N; +\infty). \qquad (2.3)$$

На промежутке $a \in (-\infty; y_1]$ все N выражений $y_1 - a, \ldots, y_N - a$ больше или равны 0. Поэтому на этом промежутке все модули в правой части формулы (2.2) раскроются со знаками «+», т.е. (2.2) примет вид: $\mathrm{g}(a) = A_0 a + B_0$, где $A_0 = -(t_1 + \ldots + t_N) = -n$, $B_0 = t_1 y_1 + \ldots + t_N y_N = x_1 + \ldots + x_n = n \cdot M_x$. Таким образом, на рассматриваемом промежутке g(a) является линейной функцией с отрицательным угловым коэффициентом.

На промежутке $a \in [y_N; +\infty)$ все N выражений $y_1 - a, \ldots, y_N - a$ меньше или равны 0. Поэтому на этом промежутке все модули в правой части формулы (2.2) раскроются со знаками «–», т.е. (2.2) примет вид: $\mathrm{g}(a) = A_N a + B_N$, где $A_N = t_1 + \ldots + t_N = n$, $B_N = -(t_1 y_1 + \ldots + t_N y_N) = -(x_1 + \ldots + x_n) = -n \cdot M_x$. Таким образом, на рассматриваемом промежутке g(a) является линейной функцией с положительным угловым коэффициентом.

Если $a \in [y_l; y_{l+1}]$, $l = 1, \ldots, N-1$, то выражения $y_1 - a, \ldots, y_l - a$ (их число равно l) отрицательны или равны 0, а выражения $y_{l+1} - a, \ldots, y_N - a$ (их равно $N - l$) положительны (или равны 0). Поэтому на промежутке $y_l \le a \le y_{l+1}$ первые l модулей в правой части формулы (2.2) раскроются со знаками «–», а последующие $N - l$ модулей – со знаками «+», т.е. $\mathrm{g}(a) = A_l a + B_l$, где

$$A_l = (t_1 + \ldots + t_l) - (t_{l+1} + \ldots + t_N) = 2(t_1 + \ldots + t_l) - n,$$

$$B_l = -(t_1 y_1 + \ldots + t_l y_l) + (t_{l+1} y_{l+1} + \ldots + t_N y_N) = nM_x - 2(t_1 y_1 + \ldots + t_l y_l).$$

Итак, на каждом из N+1 промежутков (2.3) функция g(α) будет линейной. Поэтому её график будет состоять из частей соответствующих прямых, т.е. из двух лучей, соединённых ломаной из $N-1$ звеньев. Угловые коэффициенты $A_0, A_1, \ldots, A_N$ этих кусков графика (будем их называть звеньями графика) равны (слева направо):

$$-n, -n + 2t_1, -n + 2(t_1 + t_2), \ldots, -n + 2(t_1 + \ldots + t_{N-1}) = n - 2t_N, n. \qquad (2.4)$$

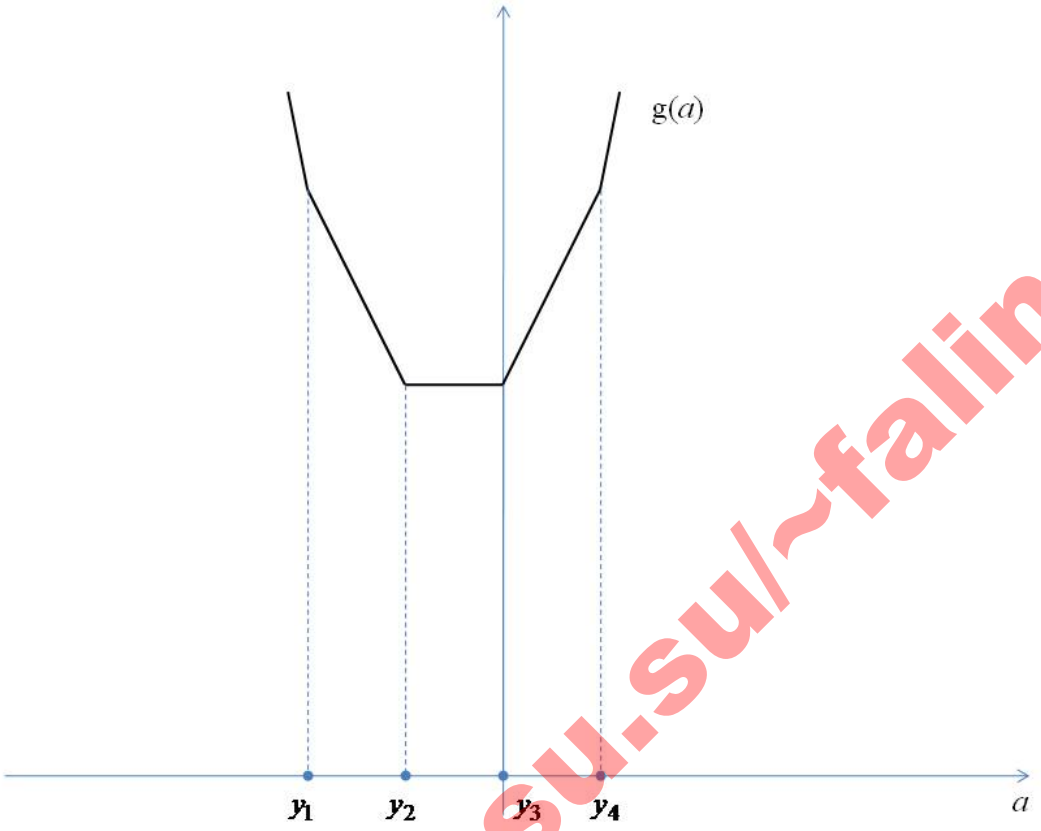
Рассмотрим последовательность (2.4). Она состоит из $N+1$ целых чисел и является монотонно возрастающей. Логически возможны два случая.

Случай 1 – в последовательности (2.4) встречается число 0. Это означает, что для некоторого l верно равенство: $n = 2(t_1 + \ldots + t_l)$. Тогда в последовательности (2.4) вначале стоят l отрицательных чисел, затем число 0, после которого стоят $(N-l)$ положительных чисел. Следовательно,

- первые l звеньев графика функции g(α) будут иметь отрицательные угловые коэффициенты, т.е. будут убывать;
- (l+1)-е звено (оно соответствует $\alpha \in [y_l; y_{l+1}]$) будет отрезком горизонтальной прямой;
- последние $N-l$ звеньев будут иметь положительные угловые коэффициенты, т.е. будут возрастать.

Типичный график такого рода изображён на рис. 2.2.

Рис. 2.2



Поэтому функция g(a) имеет наименьшее значение, которое достигается в бесконечном числе точек отрезка $\left[y_l; y_{l+1}\right]$.

Охарактеризуем теперь этот отрезок в терминах членов вариационного ряда. Пусть $k = t_1 + \ldots + t_l$. Тогда равенство $n = 2\left(t_1 + \ldots + t_l\right)$ примет вид n=2k, так что n – чётное число. Кроме того, запись (2.1) влечёт, что в вариационном ряде на месте $t_1 + \ldots + t_l = k$ стоит число y_l, а сразу за ним стоит число y_{l+1}. С другой стороны, число, которое стоит в вариационном ряде на месте k, мы обозначили v_k. Итак, $y_l = v_k$. Следующее за ним число – это v_{k+1}; оно, как мы отмечали, равно y_{l+1}: $v_{k+1} = y_{l+1}$. Поэтому отрезок $\left[y_l; y_{l+1}\right]$ – это отрезок $\left[v_k; v_{k+1}\right]$.

Случай 2 – в последовательности (2.4) число 0 не встречается, т.е. для некоторого l верны неравенства:

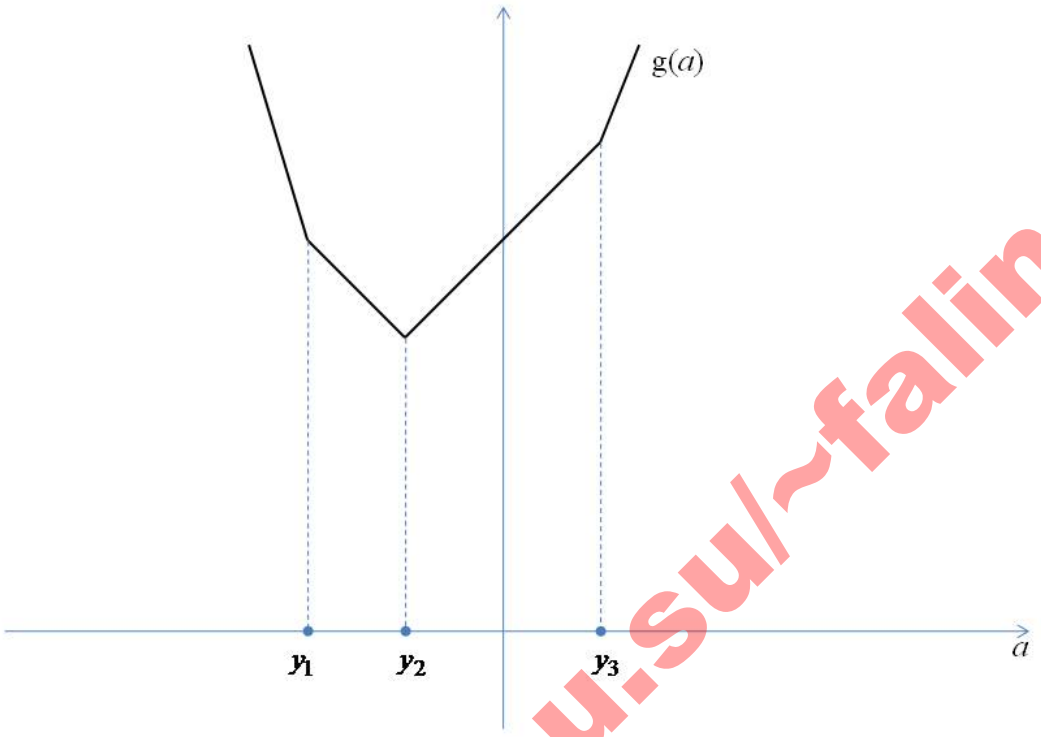
$$-n + 2(t_1 + \ldots + t_{l-1}) < 0\,,\ -n + 2(t_1 + \ldots + t_l) > 0\,. \qquad (2.5)$$

Это означает, что в последовательности (2.4) вначале стоит l отрицательных чисел, а затем – ровно $\left(N - l + 1\right)$ положительных чисел. Следовательно,

- первые l звеньев графика функции g(a) будут иметь отрицательные угловые коэффициенты, т.е. будут убывать;
- последние $N - l + 1$ звеньев будут иметь положительные угловые коэффициенты, т.е. будут возрастать.

Типичный график такого рода изображён на рис. 2.3.

Рис. 2.3



Поэтому функция g(a) имеет наименьшее значение, которое достигается в единственной точке, стоящей на границе l-го и (l+1)-го звеньев графика, т.е. в точке y_l.

Охарактеризуем теперь эту точку в терминах членов вариационного ряда.

Если число n – нечётное, т.е. $n = 2k-1$ для некоторого k, то неравенства (2.5) примут вид: $t_1 + \ldots + t_{l-1} < k - \frac{1}{2}$, $t_1 + \ldots + t_l > k - \frac{1}{2}$, что равносильно двойному неравенству: $t_1 + \ldots + t_{l-1} + 1 \le k \le t_1 + \ldots + t_l$. Тогда запись (2.1) влечёт, что k-й член вариационного ряда равен y_l: $v_k = y_l$. Таким образом, наименьшее значение функции g(a) достигается в единственной точке v_k.

Если число n – чётное, т.е. $n = 2k$ для некоторого k, то неравенства (2.5) примут вид: $t_1 + \ldots + t_{l-1} < k$, $t_1 + \ldots + t_l > k$, что равносильно двойному неравенству: $t_1 + \ldots + t_{l-1} + 1 \le k \le t_1 + \ldots + t_l - 1$. Тогда запись (2.1) влечёт, что и k-й, и (k+1)-й члены вариационного ряда равны y_l. Поэтому отрезок $\left[v_k; v_{k+1}\right]$ выродится в точку y_l и, хотя наименьшее значение функции g(a) достигается в единственной точке y_l, мы можем сказать, что оно достигается во всех точках отрезка $\left[v_k; v_{k+1}\right]$.

Резюмируя проведённые рассуждения, мы получим требуемый результат.□

Ещё раз подчеркнём, что если следовать определению медианы, данному выше, то в случае $n=2k$ медиана (вообще говоря) не определена однозначно. В качестве медианы может быть взято любое число из отрезка $[v_k;v_{k+1}]$, где $v_k \equiv x_{(k)}$ и $v_{k+1} \equiv x_{(k+1)}$ – соответственно k-й и $(k+1)$-й члены вариационного ряда. Чтобы избавиться от этой неоднозначности, в описательной статистике условились в этом случае в качестве медианы рассматривать среднее арифметическое чисел v_k и v_{k+1}, т.е. середину отрезка $[v_k;v_{k+1}]$.

2.4 Дополнительные замечания

Можно рассматривать и многие другие естественные меры «расстояния» от точки a до набора $X=[x_1,\ldots,x_n]$ как целого. В качестве ещё одного примера возьмём наибольшее абсолютное отклонение $\mathrm{F}(a)=\max\left(|x_1-a|,\ldots,|x_n-a|\right)$. Поскольку наибольшее значение набора чисел не зависит от их порядка, неважно, какую версию (неупорядоченного) набора мы рассматриваем.

Функция $\mathrm{F}(a)$ достигает наименьшего значения при $a=\dfrac{\min(x_1,\ldots,x_n)+\max(x_1,\ldots,x_n)}{2}$. Эта величина (среднее арифметическое экстремальных значений) является ещё одной интересной статистической характеристикой положения набора. Она является серединой отрезка, на котором располагаются все числа данного набора. Хотя используется эта характеристика положения редко, в ряде случаев она полезна. Соответствующее наименьшее значение функции $\mathrm{F}(a)$ равно $\dfrac{\max(x_1,\ldots,x_n)-\min(x_1,\ldots,x_n)}{2}=\dfrac{1}{2}R_X$, где $R_X=\max(x_1,\ldots,x_n)-\min(x_1,\ldots,x_n)$ – размах набора.

Таким образом, с математической точки зрения выбор характеристики положения набора, в сущности, заключается в выборе меры «расстояния» от точки a до набора $X=[x_1,\ldots,x_n]$ как целого. При этом каждой мере положения естественно соответствует своя мера рассеивания: среднему значению – дисперсия или среднее квадратичное отклонение, медиане – среднее абсолютных значений отклонений чисел набора от медианы, среднему арифметическому экстремальных значений – размах. В конечном счёте выбор меры положения и меры рассеивания определяется характером исходных данных (природой чисел исследуемого набора) и целями статистического исследования. В следующем разделе мы подробнее обсудим эти вопросы.

Мы использовали для обозначения среднего значения набора $X=[x_1,\ldots,x_n]$ символ M_X, для обозначения дисперсии – символ D_X, медианы – μ_X и т.д. Часто зависимость статистической характеристики от набора указывают не нижним

индексом, а аргументом: вместо M_X пишут $M(X)$, вместо D_X пишут $D(X)$, вместо μ_X пишут $\mu(X)$ и т.д.

3. О мерах положения числового набора

3.1 Введение

Положение числового набора может быть описано тремя величинами: средним арифметическим значением, медианой, модой. Это естественно приводит к ряду вопросов:

- Почему нельзя обойтись одной характеристикой?
- Какая из них лучше?
- Каковы их достоинства и недостатки?
- Когда нужно использовать ту или иную характеристику?

В этом разделе мы попытаемся дать ответы на эти вопросы.

Начнём с одного важного общего соображения.

Как мы видели в разделе 1, описательная статистика имеет дело с большими наборами чисел, которые описывают определённую реальную ситуацию. Таким набором может быть, например, строка оценок по математике, полученных учеником (назовём его Петя) за четверть: 4, 2, 5, 2, 3, 5, 4, 5, 5, 2 (всего 10 оценок). В конце четверти учитель должен выставить итоговую оценку. Но что такое «итоговая» оценка? Если мы попытаемся объяснить значение этого термина, мы можем сказать примерно следующее: это одно число (желательно из списка 2;3;4;5), которое «в целом» характеризует успеваемость Пети. В сущности, это расширенное определение лишь заменяет слово «итоговая» на два слова: «в целом», и ничего не добавляет к точному смыслу этого термина. Единственное, что мы можем сказать совершенно точно по поводу «итоговой» оценки, которая «в целом» характеризует успеваемость, так это то, что это должно быть одно число.

Следует сразу признать, что **найти такую оценку, которая бы содержала в себе <u>всю</u> информацию об успеваемости Пети, НЕЛЬЗЯ,** так же, как нельзя одним словом или короткой фразой описать характер человека, впечатления от кинофильма и т.д.

Например, про Петю директор школы мог бы сказать, что «Петя – плохой ученик, у него часто бывают двойки; на ЕГЭ могут быть проблемы», в то время как на призывной комиссии в военкомате могли бы отметить, что «Петя – хороший парень, крепкий; отличный солдат будет». Таким образом, короткая характеристика Пети зависит от того, какую цель ставит перед собой человек, дающий эту характеристику.

Примерно так же **нет и не может быть** «самой хорошей», «самой правильной», «самой объективной» характеристики положения числового набора. Выбор характеристики положения зависит от природы данных, целей статистического исследования и других соображений.

3.2 Среднее значение

Среднее значение числового набора используют прежде всего в тех случаях, когда важна общая сумма всех значений набора.

Например, для страховой компании чрезвычайно важна общая сумма S выплат за определённый промежуток времени (скажем, месяц) по всем договорам, заключённым этой компанией. Если эти суммарные выплаты больше, чем активы компании, то компания не сможет выполнить все свои обязательства и разорится. Общие потери S формируются из выплат по отдельным договорам: $S = x_1 + \ldots + x_n$, где n – общее число страховых случаев за рассматриваемый промежуток времени, а x_i – сумма страхового возмещения по i-му страховому случаю. Сколько наступило страховых случаев и какая сумма была выплачена по конкретному договору, страховой компании совершенно безразлично. Иными словами, наборы $X = [x_1, \ldots, x_n]$ и $Y = [y_1, \ldots, y_k]$ с равными суммами своих элементов для страховой компании фактически равнозначны (даже, если $n \neq k$).

С другой стороны, как правило, на число страховых случаев, n, влияют одни факторы, а на величины страховых возмещений, x_i, совсем другие. В качестве примера рассмотрим страхование автомобиля от повреждения при дорожно-транспортном происшествии. Вероятность попасть в аварию зависит от многих факторов, но прежде всего от возраста застрахованного: она велика для молодых людей (по причине лихачества) и для пожилых людей (из-за замедленной реакции). Большое значение имеет также погода – в дождливые дни, в туман, гололёд и т.д. число аварий заметно возрастает. Однако величина расходов на ремонт автомобиля в основном определяется маркой автомобиля. Чтобы явно учесть эти обстоятельства удобно переписать формулу $S = x_1 + \ldots + x_n$ для общих потерь в виде: $S = nM_X$, где $M_X = \dfrac{x_1 + \ldots + x_n}{n}$ – средние выплаты в расчёте на один страховой случай.

Оценив на основании прошлого опыта (как своего, так и других компаний) значения n и M_X для разных условий, страховая компания может оценить примерные значения n и M_X в наступающем месяце. Тогда по формуле $S = nM_X$ компания сможет оценить размер предстоящих выплат и если он слишком велик, принять какие-то меры по приведению в соответствие своих активов и ожидаемых обязательств. Конечно, в реальности эта процедура оценки возможного ущерба выглядит гораздо сложнее.

Подобным же образом, владельца магазина интересует прежде всего общая сумма S денег, израсходованная покупателями в его магазине за определённый промежуток времени. Эта сумма может быть записана в виде: $S = x_1 + \ldots + x_n$, где n – общее число покупателей, посетивших магазин за рассматриваемый промежуток времени, а x_i – стоимость товаров в i-м чеке.

Если в каком-то месяце продажи упали, важно понять, уменьшилось ли число покупателей или же они стали меньше тратить (от этого будут зависеть действия по улучшению ситуации). Для этого нужно смотреть на величину $S = x_1 + \ldots + x_n$ как на произведение $S = nM_X$, *где* $M_X = \dfrac{x_1 + \ldots + x_n}{n}$ – средняя сумма, израсходованная одним покупателем.

В этих примерах среднее значение является своеобразной «производительностью» (застрахованного в примере со страховой компанией, покупателя в примере с магазином). Именно в случаях, когда природа анализируемой переменной и цель исследования диктуют важность подобной «производительности», и следует использовать среднее значение.

Когда мы вводим среднее (арифметическое) значение в качестве меры положения набора, мы фактически говорим, что для нас исходный набор $X = [x_1, \ldots, x_n]$ равносилен набору $X^* = [M_X, \ldots, M_X]$ из n одинаковых чисел; равносилен в том смысле, что оба набора имеют одну и ту же сумму своих элементов: $x_1 + \ldots + x_n = \underbrace{M_X + \ldots + M_X}_{n \text{ слагаемых}}$. Если эту равносильность понимать шире, то можно сказать, что наборы эквивалентны, т.к. оба набора приводят к одному и тому же результату в рамках рассматриваемой задачи. Однако в ряде ситуаций таким эквивалентным набором будет не набор из средних арифметических, а набор из величин, которые определяются по более сложным формулам. Скажем, в финансовой математике эти формулы тесно связаны со средними геометрическими.

В качестве иллюстрации рассмотрим инвестора, который может вложить определённую сумму денег S_0 рублей в n-летний инвестиционный проект, который за первый год принесёт прибыль в размере $i_1\%$ годовых, за второй – $i_2\%$, и т.д., за n-й год – $i_n\%$. При этом прибыль не изымается, а каждый раз вкладывается в дальнейшее развитие дела и потому зарабатывает проценты в будущем. В этой ситуации

- к концу первого года средства инвестора, вложенные в проект вырастут до суммы $S_1 = S_0 + \dfrac{i_1}{100} S_0 = \left(1 + \dfrac{i_1}{100}\right) S_0$;
- к концу второго года средства инвестора, вложенные в этот проект вырастут до суммы $S_2 = S_1 + \dfrac{i_2}{100} S_1 = \left(1 + \dfrac{i_1}{100}\right)\left(1 + \dfrac{i_2}{100}\right) S_0$;

и т.д.

- к концу n-го года средства инвестора, вложенные в этот проект вырастут до суммы $S_n = \left(1 + \dfrac{i_1}{100}\right)\left(1 + \dfrac{i_2}{100}\right)\ldots\left(1 + \dfrac{i_n}{100}\right) S_0$.

Именно сумму S_n получит инвестор по истечении n-летнего периода.

Предположим теперь, что инвестору предлагают альтернативный n-летний проект, который каждый год даёт постоянную прибыль i% годовых (как и раньше, прибыль инвестируется, чтобы приносить дальнейший доход). Математически эта ситуация описывается полученными выше формулами при $i_1 = \ldots = i_n \equiv i$. Поэтому по истечении n-летнего периода инвестор получит сумму $S_n^* = \left(1 + \frac{i}{100}\right)^n S_0$. Оба проекта эквивалентны с точки зрения финансовых интересов инвестора, если

$$\left(1 + \frac{i_1}{100}\right)\left(1 + \frac{i_2}{100}\right)\ldots\left(1 + \frac{i_n}{100}\right)S_0 = \left(1 + \frac{i}{100}\right)^n S_0,$$

откуда для «средней» (т.е. эквивалентной) величины процентов i мы имеем:

$$i = 100\left(\sqrt[n]{\left(1 + \frac{i_1}{100}\right)\left(1 + \frac{i_2}{100}\right)\ldots\left(1 + \frac{i_n}{100}\right)} - 1\right).$$

Величина $k = \left(1 + \frac{i}{100}\right)$ в финансовой математике называется коэффициентом накопления, соответствующим процентной ставке i%. С использованием этого понятия приведённая выше формула для «средней» величины процентов i превратится в следующую формулу для «среднего» коэффициента накопления $k = 1 + \frac{i}{100}$: $k = \sqrt[n]{k_1 \ldots k_n}$. Таким образом, «средний» коэффициент накопления является средним геометрическим коэффициентов накопления за все n лет.

Из ситуаций, в которых использование среднего значения может создать ложное представление о положении числового набора, наиболее важная связана с описанием доходов группы людей.

Рассмотрим следующий пример. В небольшом торговом предприятии работает 10 человек. За апрель они получили следующие зарплаты (в тысячах рублей): 11, 11, 15, 15, 15, 15, 18, 20, 20, 160. В соответствии с этими данными, «средняя» зарплата равна 30 (тысяч рублей). Это значение «средней» зарплаты можно считать хорошей характеристикой доходов работников, например, для налоговой инспекции, которая на этой основе сможет оценить размер налоговых поступлений. Однако, нельзя не отметить, что

- два работника (самые низко оплачиваемые, наверное уборщицы или грузчики) получают только около трети от этой «средней» зарплаты;
- четыре работника получают только половину от этой «средней» зарплаты;
- три работника получают около двух третей от этой «средней» зарплаты.

Таким образом, заработок 9 из 10 работников существенно ниже подсчитанного среднего значения и для них месячная зарплата в 30 тысяч рублей, видимо, малореальная мечта. Но у одного человека (надо полагать – это хозяин

предприятия) зарплата примерно в 5 раз превышает подсчитанное среднее значение (этот человек получает больше, чем все остальные работники вместе).

Таким образом, в рассмотренном примере среднее арифметическое значение зарплаты даёт совершенно искажённую социальную картину доходов работников и потому здесь нельзя использовать среднее значение в качестве меры положения набора.

После этого общего описания ситуаций, в которых оправдано использование среднего значения, отметим ещё несколько его достоинств и недостатков.

Важное достоинство среднего значения M_X числового набора $X=[x_1,\ldots,x_n]$ заключается в том, что оно чувствительно к любому изменению значений чисел, входящих в этот набор: если какое-то из чисел, входящих в набор, например, x_1, увеличится на величину Δ (так что в новом наборе $(x_1+\Delta,x_2,\ldots,x_n)$ первое число будет равно $x_1+\Delta$), то среднее значение нового набора будет равно

$$\frac{x_1+\Delta+x_2+\ldots+x_n}{n}=\frac{x_1+x_2+\ldots+x_n}{n}+\frac{\Delta}{n}=M_X+\frac{\Delta}{n}, \qquad (3.1)$$

т.е. увеличится на величину $\frac{\Delta}{n}$.

Однако это достоинство превращается в недостаток, если исходный ряд чисел содержит одно или несколько значений, сильно отличающихся от других чисел в одну сторону (такие нетипичные значения называют выбросами – именно с выбросом мы имели дело в рассмотренном выше примере с зарплатами). Выбросы могут сильно повлиять на значение среднего значения и исказить представление о положении набора. Как показывает соотношение (3.1), влияние изменения отдельного значения набора на среднее уменьшается с ростом n. Поэтому увеличение размера набора может дать более объективную характеристику его положения.

Отметим, кроме того, что, как правило, среднее значение не является одним из чисел исходного набора. В частности, это может затруднить интерпретацию среднего в случае, когда природа данных такова, что числа набора могут принимать лишь конечное число определённых значений (как, например, школьная оценка, которая может быть только 2, 3, 4 или 5).

3. 3 Медиана

Медиана менее чувствительна к изменению значений числового набора, чем среднее значение. Если среднее значение реагирует на каждое изменение этих значений, то медиана – лишь на некоторые. В частности, медиана не искажается выбросами, что является её важным достоинством при описании несимметричных наборов, когда небольшое число значений резко отличается от остальных. Скажем, в рассмотренном выше примере с зарплатами медиана равна $\frac{15+15}{2}=15$ (в этом наборе $n=10$ чисел и потому медиана – среднее

арифметическое пятой и шестой по величине зарплат). Конечно, эта сумма даёт гораздо более объективное представление о распределении доходов работников. Если бы наибольшая зарплата увеличилась со 160 (тысяч рублей) до 260 (тысяч рублей), то среднее значение увеличилось бы до 40 (тысяч рублей), что создало бы ложное впечатление о росте доходов работников. Однако медиана осталась бы неизменной, что отражает неизменность доходов большинства сотрудников.

Если значения числового набора расположены на числовой оси приблизительно симметрично, то среднее значение набора и его медиана практически совпадают.

Как и среднее значение, медиана не всегда является одним из чисел набора (что в ряде случаев может затруднить её интерпретацию).

В отличие от среднего значения, вычисление медианы обычно проще. Если числа набора упорядочены и известно, чётно или нет их количество, то медиана находится практически мгновенно. Вычисление среднего значения для больших наборов предполагает суммирование большого числа слагаемых, что относительно часто приводит к ошибкам даже при использовании калькулятора или компьютера (из-за ошибок при вводе данных).

Значение медианы возрастает, когда её рассматривают как вторую квартиль Q_2 и в дополнение к ней вычисляют нижнюю квартиль Q_1 и верхнюю квартиль Q_3. Неформально говоря, квартили Q_1, Q_2, Q_3 делят исходный упорядоченный набор на 4 равные части. Иначе говоря, нижняя квартиль – это медиана первой половины исходного набора, а верхняя квартиль – медиана второй половины исходного набора. Уточнить это неформальное определение нижней и верхней квартилей можно несколькими способами, которые обычно приводят к разным результатам (хотя и не очень сильно отличающимся); общепринятого определения квартилей в описательной статистике нет. Подробнее мы будем говорить о квартилях в разделе 6.

3.4 Мода

Напомним, что для числового набора мода – это такое значение, которое встречается чаще всего. Из трёх основных мер положения числового набора, среднего значения, медианы, моды, мода является наименее интересной (если не сказать – совсем неинтересной).

Во многих случаях мода создаёт ложное впечатление о положении чисел набора. Рассмотрим, например, строку оценок по математике, полученных Петей за четверть: 4, 2, 5, 2, 3, 5, 4, 5, 5, 2 (всего 10 оценок). Оценка «2» встречается 3 раза, оценка «3» – 1 раз, оценка «4» – 2 раза, оценка «5» – 4 раза. Поэтому мода равна «5». Но вряд ли кто-нибудь скажет, что Петя – отличник. Среднее арифметическое значение его оценок равно 3,7, что, видимо, лучше отражает успеваемость Пети. При хорошем к нему отношении за четверть можно было бы поставить «4», тем более, что медиана равна «4».

Часто мода не определена однозначно, что может привести к абсурдным выводам (если моду рассматривать как меру положения). Например, если бы список оценок Пети немного изменился: 4, 2, 5, 2, 3, 5, 2, 5, 5, 2 (вместо одной из четвёрок он получил двойку), то и оценка «5», и оценка «2» встречались бы по 4 раза (а «3» и «4» – по одному разу). В этом случае мода принимает два значения: «2» и «5». Как интерпретировать эти данные? Отличник Петя или двоечник?

Для непрерывных данных обычно почти все числа x_i различны. В этом случае говорить о моде вообще не имеет смысла. Правда, после подходящей группировки данных может оказаться, что в один из интервалов попадает больше чисел, чем в другие. В этом случае вместо моды вводят понятие модального класса. Иначе говоря, для группированных значений непрерывной переменной модальный класс – это та группа значений, которая содержит больше чисел исходного набора, чем любая другая группа.

Если бы не потенциальная неоднозначность в определении моды и отмеченные выше проблемы с её интерпретацией, определённым достоинством моды можно было бы считать то, что она всегда является одним из значений набора.

Пожалуй, единственное преимущество моды заключается в том, что она может быть найдена и для нечисловых наборов. Предположим, что в киоске было продано 236 пачек ванильного мороженого, 95 пачек шоколадного мороженого и 143 пачки сливочного мороженого. Эту информацию можно описать набором

$$\left(\underbrace{\text{в},\text{в},\ldots,\text{в}}_{\text{236 раз}}, \underbrace{\text{ш},\text{ш},\ldots,\text{ш}}_{\text{95 раз}}, \underbrace{\text{с},\text{с},\ldots,\text{с}}_{\text{143 раза}} \right).$$

Его мода, которая указывает на самый популярный вид мороженого, равна «в», т.е. «ванильное».

3.5 Дополнительные задачи

Задача 3.1. Каждую субботу школьник ездит в гости к своей бабушке, которая живёт на другом конце города. На протяжении трёх месяцев (13 недель) он записывал время в пути и получил следующий ряд чисел (это время в минутах): 48, 62, 71, 53, 58, 51, 121, 58, 49, 60, 62, 54, 59.

(1) Подсчитайте среднее время в пути и суммарное абсолютное отклонение результатов наблюдений от этого среднего.

(2) Подсчитайте медиану времени в пути и суммарное абсолютное отклонение результатов наблюдений от этой медианы.

(3) Какое число, среднее значение или медиана, по вашему мнению, лучше характеризует длительность поездки? Почему?

Решение. (1) Среднее время в пути равно 62 мин. Поэтому суммарное абсолютное отклонение результатов наблюдений от среднего равно 136.

(2) Чтобы найти медиану, упорядочим результаты наблюдений по возрастанию: 48, 49, 51, 53, 54, 58, 58, 59, 60, 62, 62, 71, 121, и возьмём число,

стоящее в этом ряду точно посередине. Это будет седьмое по порядку число, т.е. число 58. Суммарное абсолютное отклонение результатов наблюдений от найденной медианы равно 122. Как и следовало ожидать из Теоремы 2.2, для медианы суммарное абсолютное отклонение от результатов наблюдений меньше, чем для среднего.

(3) Для всех поездок, за исключением одной, время в пути было около 1 часа (когда немного меньше, когда немного больше). Но одна поездка заняла 121 мин., т.е. более 2 часов. Столь большое значение, видимо, явилось следствием каких-то маловероятных случайных событий и, скорее всего, нетипично. Если его удалить из списка анализируемых значений, то для нового ряда из 12 чисел среднее значение будет примерно равно 57 мин.,
т.е. будет практически совпадать с медианой исходного набора из 13 наблюдений. Однако, поскольку это резко выделяющееся, экстремальное, значение было включено в расчёт среднего, это привело к увеличению среднего на 5 мин. Медиана менее чувствительна к наличию экстремальных значений и потому в данной ситуации точнее характеризует набор в целом.

Задача 3.2. В классе 32 ученика. В следующей таблице приведены суммарные данные об их годовых оценках по алгебре и по русскому языку:

		Русский язык		
		3	**4**	**5**
Алгебра	**3**	8	2	3
	4	1	2	11
	5	1	1	3

(1) Найдите среднее значение и медиану оценок по математике и по русскому языку.

(2) На родительском собрании классный руководитель сказала, что успеваемость по русскому языку лучше, чем успеваемость по математике. Аргументируйте этот вывод с помощью полученных результатов о средних значениях и медианах.

Решение. (1) Суммируя данные по строкам таблицы, мы получим, что по алгебре 13 учеников имели оценку «3», 14 – оценку «4», 5 – оценку «5». Поэтому средняя оценка по алгебре равна (примерно) 3,75.

Если выписать оценки всех учеников по алгебре в порядке возрастания, то мы получим: 3,3,3,3,3,3,3,3,3,3,3,3,3,4,4,4,4,4,4,4,4,4,4,4,4,4,5,5,5,5,5.
Эта последовательность состоит из 32 чисел. Посередине, на местах 16 и 17, стоят числа 4 и 4. Их среднее арифметическое равно 4 – это и будет медиана.

Этот же вывод можно получить и проще. Так как общее число оценок «3» равно 13, а затем идёт 14 оценок «4», на местах 16 и 17 будут стоять две четвёрки, так что их среднее арифметическое (т.е. медиана) равно 4.

Суммируя данные по столбцам таблицы, мы получим, что по русскому языку 10 учеников имели оценку «3», 5 – оценку «4», 17 – оценку «5». Поэтому средняя оценка по русскому языку равна (примерно) 4,22.

Для оценок по русскому языку вариационный ряд имеет вид: 3,3,3,3,3,3,3,3,3,3,4,4,4,4,4,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5. Посередине, на местах 16 и 17, стоят числа 5 и 5. Их среднее арифметическое равно 5 – это и будет медиана оценок по русскому языку.

(2) По алгебре как среднее значение, так и медиана оценок меньше, чем соответствующая величина по русскому языку. Кроме того, наивысшую оценку «5» по алгебре имеет только 5 учеников, в то время как по русскому языку наивысшую оценку имеет 17 учеников. Поэтому вывод классного руководителя о том, что успеваемость по русскому языку лучше, чем успеваемость по математике, вполне аргументирован.

Задача 3.3. В классе 12 мальчиков и 10 девочек. Учительница задала каждому ученику 20 задач на сложение двузначных чисел в уме. Набор X содержат данные о числе задач, решённых мальчиками, а набор Y – о числе задач, решённых мальчиками: X=[15;12;8;16;14;11;17;7;16;20;15;14], Y=[17;15;14;16;13;17;16;12;14;16].

Подсчитайте среднее значение и размах числа правильно решённых задач одним школьником (отдельно для мальчиков и девочек).

Можно ли с помощью этих результатов определённо сказать, кто лучше считает в уме – мальчики или девочки?

Решение. Среднее число правильно решённых задач одним мальчиком (девочкой) равно 13,75 (соответственно, 15).

Для мальчиков среднее число правильно решённых задач одним школьником равно 13,75, наибольшее число правильно решённых задач равно 20, а наименьшее равно 7. Поэтому размах равен 20 – 7 = 13.

Для девочек среднее число правильно решённых задач одним школьником равно 15, наибольшее число правильно решённых задач равно 17, а наименьшее равно 12. Поэтому размах равен 17 – 12 = 5.

Таким образом, для девочек среднее число правильно решённых задач одним человеком больше, чем для мальчиков. Кроме того, значения размахов показывают, что результаты девочек разбросаны гораздо меньше, т.е. более стабильны. Поэтому разумно сделать вывод, что девочки этого класса лучше считают в уме, чем мальчики.

4. Линейные преобразования числовых наборов

4.1 Введение

В принципе вычисление основных статистических характеристик числового набора $X=\left[x_1,\ldots,x_n\right]$ является несложной арифметической задачей. Однако, если числа $x_1,\ldots,x_n$ – «некрасивые», то для наборов большой длины вычисление таких характеристик как среднее и дисперсия даже с помощью калькулятора потребует относительно много усилий. Применение для обработки информации электронных таблиц Microsoft Office Excel не решает все проблемы (ошибки при вводе данных всё ещё возможны). В статистике известен приём, который позволяет упростить вычисления в случае, когда данные имеют линейную структуру. В школьной учебной литературе этот вопрос затронут лишь вскользь и без доказательств (см., например, [4], глава III, раздел 16). Цель этого раздела – рассмотреть его подробнее. С линейными преобразованиями также тесно связано специальное преобразование наборов, называемое *стандартизацией*; это преобразование полезно при сравнении разных числовых наборов.

4.2 Линейные преобразования наборов данных

Пусть $X=\left[x_1,\ldots,x_n\right]$ – набор из n чисел. Предположим, что все числа этого набора могут быть получены из чисел некоторого нового, вспомогательного, набора $U=\left[u_1,\ldots,u_n\right]$ применением одного и того же линейного преобразования $x=au+b$, т.е. при некоторой нумерации элементов этих наборов верны числовые равенства:

$$x_1=au_1+b, x_2=au_2+b,\ldots,x_n=au_n+b. \qquad (4.1)$$

Эту систему из n равенств можно записать и короче: $x_i=au_i+b$, где $i=1,2,\ldots,n$ (т.е. индекс i принимает последовательно все значения от 1 до n).

Параметр a называется *масштабным коэффициентом* (или *масштабным параметром*; используют также и термин «*нормирующий параметр*»), а параметр b – *параметром сдвига* (или, короче, *сдвигом*; используют также и термин «*центрирующий параметр*»).

Например, если дан основной набор $x_1 = 105, x_2 = 110, x_3 = 95$, то числа этого набора можно записать так: $x_1 = 100 + 5 = 100 + 5 \cdot 1$, $x_2 = 100 + 10 = 100 + 5 \cdot 2$, $x_3 = 100 - 5 = 100 + 5 \cdot (-1)$. Иначе говоря, $x_i = 100 + 5u_i$, где $u_1 = 1, u_2 = 2, u_3 = -1$.

Чтобы в будущем исключить случаи, которые не представляют интереса для приложений, а лишь усложняют формулировки утверждений, мы будем считать, что масштабный коэффициент a отличен от 0.

Линейные преобразования наборов данных обладают важными свойствами:

Свойство 1 (симметричность) Если набор $X = [x_1, \ldots, x_n]$ получен из набора $U = [u_1, \ldots, u_n]$ линейным преобразованием $x = au + b$, то набор $U = [u_1, \ldots, u_n]$ может быть получен из набора $X = [x_1, \ldots, x_n]$ линейным преобразованием $u = \frac{1}{a} \cdot x + \left(-\frac{b}{a}\right)$.

Обратим внимание на то, что коэффициент $\frac{1}{a}$ не равен нулю.

Свойство 2 (транзитивность) Если набор $X = [x_1, \ldots, x_n]$ получен из набора $U = [u_1, \ldots, u_n]$ линейным преобразованием $x = a'u + b'$, набор $U = [u_1, \ldots, u_n]$, в свою очередь, получен из набора $V = [v_1, \ldots, v_n]$ линейным преобразованием $u = a''v + b''$, то набор $X = [x_1, \ldots, x_n]$ может быть получен из набора $V = [v_1, \ldots, v_n]$ линейным преобразованием $x = a'a''v + (a'b'' + b')$.

Обратим внимание на то, что коэффициент $a'a''$ не равен нулю (так как отличны от нуля коэффициенты a' и a'').

Свойство 3 (рефлексивность) Любой набор $X = [x_1, \ldots, x_n]$ можно получить из самого себя линейным преобразованием $ax + b$, с коэффициентами $a = 1$, $b = 0$.

В высшей алгебре связь между парами объектов (любой природы), обладающую свойствами симметричности, транзитивности и рефлексивности называют *отношением эквивалентности*. Эквивалентные объекты в некотором смысле можно считать неотличимыми друг от друга (или различными представлениями одного и того же объекта). В нашем случае, наборы, отличающиеся друг от друга линейным преобразованием, в сущности задают один и тот же набор данных, но измеренных в разных единицах измерения.

Для приложений к описательной статистике важны свойства, связывающие средние значения (медианы, дисперсии, стандартные отклонения, размахи) исходного и преобразованного наборов данных.

4.3 Среднее значение и линейные преобразования наборов данных

Теорема 4.1. *Пусть* M_X *– среднее значение набора* $X=[x_1,\ldots,x_n]$, *а* M_U *– среднее значение набора* $U=[u_1,\ldots,u_n]$, *связанного с набором* X *соотношениями (4.1). Тогда*

$$M_X = aM_U + b. \qquad (4.2)$$

Иначе говоря, среднее значение M_X *основного набора* X *получается из среднего значения* M_U *вспомогательного набора* U *с помощью того же линейного преобразования* $x = au + b$, *которое преобразует вспомогательный набор в основной.*

Доказательство. Следующая цепочка формул связывает величины M_X и M_U между собой:

$$M_X = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{(au_1 + b) + (au_2 + b) + \ldots + (au_n + b)}{n}$$
$$= \frac{a(u_1 + u_2 + \ldots + u_n) + nb}{n} = a\frac{u_1 + u_2 + \ldots + u_n}{n} + b = aM_U + b.$$

С помощью символа суммирования эти преобразования можно записать проще:

$$M_X = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}\sum_{i=1}^{n}(au_i + b) = \frac{1}{n}\left(a\sum_{i=1}^{n} u_i + nb\right) = a\frac{1}{n}\sum_{i=1}^{n} u_i + b = aM_U + b.$$

***Следствие* 4.1.** Пусть $X=[x_1,\ldots,x_n]$ – некоторый набор чисел, а M_X – его среднее значение. Рассмотрим набор $U=[u_1,\ldots,u_n]$, полученный из набора $X=[x_1,\ldots,x_n]$ применением следующего линейного преобразования:

$$u_i = x_i - M_X \Leftrightarrow x_i = M_X + u_i. \qquad (4.3)$$

Величины u_i показывают отклонения величин x_i от их среднего значения. Поскольку среднее значение набора чисел в некотором смысле является его «центром», преобразование (4.3) называют *центрированием* исходного набора.

Операция центрирования соответствует следующим значениям нормирующего и центрирующего параметров a и b: $a=1$, $b=M_X$. Поэтому для центрированного набора общее соотношение (4.2) примет вид: $M_X = M_U + M_X$ или, что то же самое, $M_U = 0$.

Таким образом, среднее значение (а вместе с ним и обычная сумма) отклонений величин x_i от их среднего значения обязательно равна нулю.

Последнее свойство без доказательства отмечается в школьных учебниках (см., например, [4], гл. III, раздел 13, или [3], гл. 5, раздел 5.3).

Часто можно найти такое линейное преобразование исходного набора, что новый набор выглядит проще. Тогда найти среднее значение для нового набора легче, чем для исходного. Полученное общее соотношение (4.2) между средними значениями исходного и нового наборов позволяет найти среднее значение исходного набора меньшими вычислительными усилиями.

Рассмотрим примеры применения этого свойства (для простоты изложения мы ограничиваемся наборами небольшой длины).

Пример 4.1. Пусть дан набор $X=\left[2007;2011;2003;2006;1998\right]$, так что

$$x_1=2007, x_2=2011, x_3=2003, x_4=2006, x_5=1998.$$

Введём новый набор $U=\left[u_1,u_2,u_3,u_4,u_5\right]$ по формуле: $u_i=x_i-2000$. Иначе говоря, представим числа основного набора в виде $x_i=2000+u_i\equiv au_i+b$, где $a=1,\ b=2000$. Тогда $u_1=7, u_2=11, u_3=3, u_4=6, u_5=-2$. Среднее значение этого набора найти гораздо легче:

$$M_U=\frac{7+11+3+6+(-2)}{5}=\frac{25}{5}=5.$$

Поскольку $x_i=u_i+2000$, аналогичное соотношение связывает средние значения: $M_X=M_U+2000$. Поэтому для среднего значения исходного набора мы имеем: $M_X=M_U+2000=2005$.

Пример 4.2. Пусть дан набор

$$x_1=2{,}91, x_2=3{,}07, x_3=3{,}02, x_4=3{,}05, x_5=3{,}01, x_6=3{,}03.$$

Введём набор $U=\left[u_1,u_2,u_3,u_4,u_5,u_6\right]$ по формуле: $u_i=100\left(x_i-3\right)$. Иначе говоря, представим числа основного набора в виде $x_i=3+0{,}01u_i\equiv au_i+b$, где $a=0{,}01, b=3$. Тогда $u_1=-9, u_2=7, u_3=2, u_4=5, u_5=1, u_6=3$. Среднее значение этого набора найти гораздо легче:

$$M_U=\frac{-9+7+2+5+1+3}{6}=\frac{9}{6}=1{,}5.$$

Поскольку $x_i=0{,}01u_i+3$, аналогичное соотношение связывает средние значения: $M_X=0{,}01M_U+3$. Поэтому для среднего значения исходного набора мы имеем: $M_X=0{,}01\cdot 1{,}5+3=3{,}015$.

4.4 Дисперсия и линейные преобразования наборов данных

Теорема 4.2. *Пусть* D_X *– дисперсия набора* $X=[x_1,\ldots,x_n]$, *а* D_U *– дисперсия набора* $U=[u_1,\ldots,u_n]$, *связанного с набором X соотношениями (4.1). Тогда*

$$D_X = a^2 D_U. \qquad (4.4)$$

Доказательство. Следующая цепочка формул связывает величины D_X и D_U между собой:

$$\begin{aligned} D_X &= \frac{(x_1-M_X)^2+(x_2-M_X)^2+\ldots+(x_n-M_X)^2}{n} \\ &= \frac{(au_1+b-aM_U-b)^2+(au_2+b-aM_U-b)^2+\ldots+(au_n+b-aM_U-b)^2}{n} \\ &= \frac{a^2(u_1-M_U)^2+a^2(u_2-M_U)^2+\ldots+a^2(u_n-M_U)^2}{n} \\ &= a^2\frac{(u_1-M_U)^2+(u_2-M_U)^2+\ldots+(u_n-M_U)^2}{n} = a^2 D_U. \end{aligned}$$

С помощью символа суммирования эти преобразования можно записать проще:

$$D_X = \frac{1}{n}\sum_{i=1}^{n}(x_i-M_X)^2 = \frac{1}{n}\sum_{i=1}^{n}(au_i+b-aM_U-b)^2 = a^2\frac{1}{n}\sum_{i=1}^{n}(u_i-M_U)^2 = a^2 D_U.$$

Следствие 4.2. Из (4.4) для средних квадратичных отклонений мы имеем:

$$\sigma_X = |a|\,\sigma_U. \qquad (4.5)$$

В частности, если масштабный множитель a равен 1 или –1, то $D_X = D_U$, $\sigma_X = \sigma_U$.

Пример 4.3. Пусть дан набор

$$x_1 = 2007, x_2 = 2011, x_3 = 2003, x_4 = 2006, x_5 = 1998.$$

Как мы видели в Примере 4.1, числа этого набора можно представить в виде $x_i = u_i + 2000$, где $u_1 = 7, u_2 = 11, u_3 = 3, u_4 = 6, u_5 = -2$. Среднее значение нового набора мы нашли в Примере 4.1: $M_U = 5$. Теперь для его дисперсии мы имеем:

$$D_U = \frac{(7-5)^2+(11-5)^2+(3-5)^2+(6-5)^2+(-2-5)^2}{5} = \frac{94}{5} = 18{,}8.$$

Соответственно, $\sigma_U = \sqrt{18{,}8} \approx 4{,}336$. Поскольку масштабный множитель $a=1$, для исходного набора мы имеем: $D_X = D_U = 18{,}8$, $\sigma_X = \sigma_U \approx 4{,}336$.

Пример 4.4. Пусть дан набор

$$x_1 = 2{,}91, x_2 = 3{,}07, x_3 = 3{,}02, x_4 = 3{,}05, x_5 = 3{,}01, x_6 = 3{,}03.$$

Как мы видели в Примере 4.2 числа этого набора можно представить в виде $x_i = 3 + 0{,}01u_i$, где $u_1 = -9, u_2 = 7, u_3 = 2, u_4 = 5, u_5 = 1, u_6 = 3$. Среднее значение нового набора мы нашли в Примере 4.2: $M_U = 1{,}5$. Теперь для его дисперсии мы имеем:

$$D_U = \frac{(-9-1{,}5)^2 + (7-1{,}5)^2 + (2-1{,}5)^2 + (5-1{,}5)^2 + (1-1{,}5)^2 + (3-1{,}5)^2}{6}$$

$$= \frac{155{,}50}{6} \approx 25{,}92.$$

Отсюда $\sigma_U \approx 5{,}091$, так что $\sigma_X = 0{,}01\sigma_U \approx 0{,}0591$, а $D_X = 0{,}01^2 D_U \approx 0{,}002592$.

Следствие 4.3. Пусть $X = [x_1, \ldots, x_n]$ – некоторый набор чисел, M_X – его среднее значение, D_X – дисперсия, а $\sigma_X = \sqrt{D_X}$ – среднее квадратичное отклонение. Предположим, что дисперсия отлична от нуля и рассмотрим набор $U = [u_1, \ldots, u_n]$, полученный из набора $X = [x_1, \ldots, x_n]$ применением следующего линейного преобразования:

$$u_i = \frac{x_i - M_x}{\sigma_x} \Leftrightarrow x_i = \sigma_x u_i + M_x. \qquad (4.6)$$

Преобразование (4.6) соответствует следующим значениям нормирующего и центрирующего параметров a и b: $a = \sigma_X$, $b = M_X$. Поэтому для набора $u_1, u_2, \ldots, u_n$ общие соотношения (4.2) и (4.4) примут вид: $M_X = \sigma_X M_U + M_X$, $D_X = \sigma_X^2 D_U$. Поскольку $\sigma_X^2 = D_X$, эти соотношения равносильны тому, что $M_U = 0$, $D_U = 1$. Набор с нулевым средним и единичной дисперсией называется стандартным. Соответственно, преобразование (4.6) называют *стандартизацией* исходного набора. Это преобразование играет важную роль в теории вероятностей; в следующем пункте мы рассмотрим её применение для сравнения разных наборов.

4.5 Применение стандартизации числовых наборов

В статистике часто возникает задача сравнения чисел из разных наборов. Эти наборы могут быть получены в результате измерения некоторых объектов с помощью разных процедур и разных единиц измерения. Поэтому прямое сравнение чисел из таких наборов обычно не имеет смысла. Для решения этой

задачи в статистике разработаны различные методы. Простейшие из них основаны на введённой в предыдущем пункте стандартизации числового набора. Хотя часто стандартизация набора не является наилучшим подходом к задаче сравнения чисел из разных наборов, она очень проста как с методической, так и с вычислительной точек зрения.

Стандартизованный, т.е. центрированный и нормированный, набор $Z = \left(\frac{x_1 - M_X}{\sigma_X}, \ldots, \frac{x_n - M_X}{\sigma_X} \right)$ измеряет величины отклонений чисел x_i от их среднего не в абсолютных значениях, а по отношению к стандартному отклонению (которое тем самым выбирается в качестве новой единицы измерения величин x_i). Стандартизованные наборы могут использоваться для того, чтобы сравнивать разные наборы данных, полученные в результате измерения одних и тех же объектов с помощью разных процедур и разных единиц измерения.

Чтобы проиллюстрировать приведённые выше общие рассуждения, проанализируем гипотетические результаты ЕГЭ по математике для шести школьников, приведённые в следующей таблице (максимально возможное число баллов равно 100):

Имя школьника	Петя	Коля	Таня	Маша	Андрей	Оля
Исходная оценка x_i	68	72	74	80	90	96

`

Сумма всех шести оценок равна 480, так что средняя арифметическая оценка равна $M_X = \frac{480}{6} = 80$. Отклонения оценок от этого среднего (центрированные оценки за экзамен) приведены в следующей таблице:

Имя школьника	Петя	Коля	Таня	Маша	Андрей	Оля
Центрированная оценка $x_i - M_X$	–12	–8	–6	0	10	16

Центрированные оценки позволяют легко определить, хуже или лучше результат школьника, чем средний результат: если центрированная оценка отрицательна, то результат хуже среднего, а если положительна – то лучше. Например, центрированная оценка Андрея равна 10. Поэтому его результат выше среднего (по рассматриваемой группе из 6 школьников).

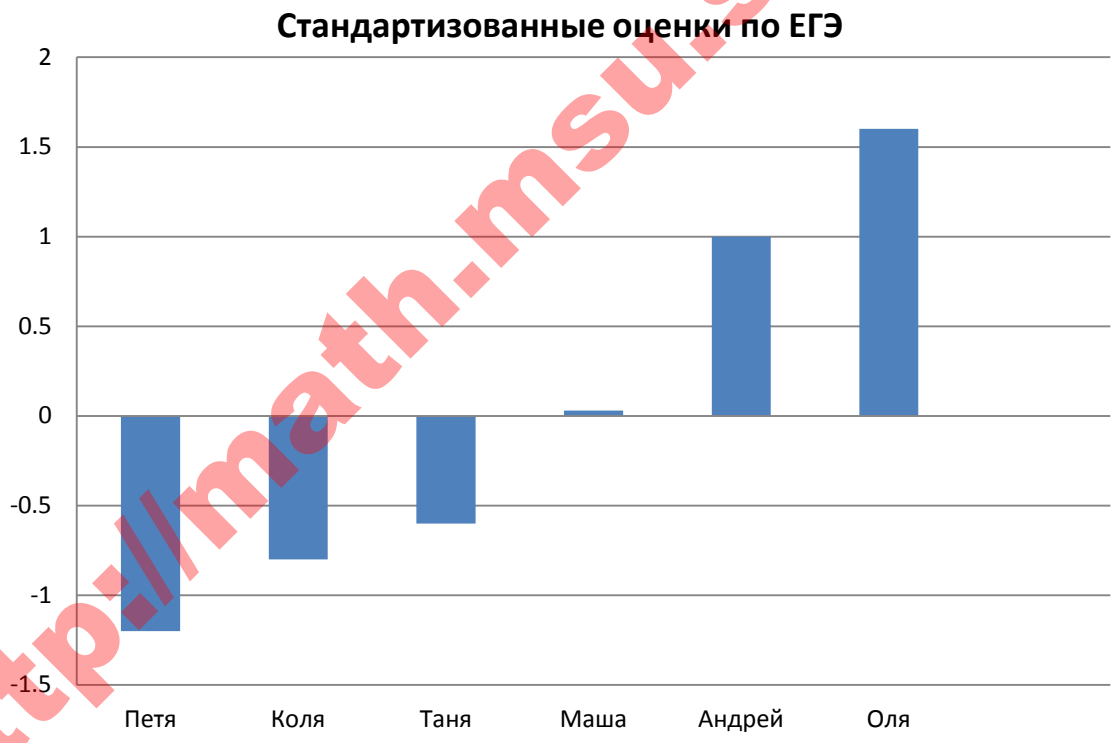
Сумма квадратов отклонений оценок от среднего равна 600, так что дисперсия исходного набора оценок (которая по определению равна среднему значению квадратов отклонений оценок от среднего) есть $D_X = \frac{600}{6} = 100$.

Соответственно, стандартное отклонение для исходного набора оценок равно $\sigma_X = \sqrt{100} = 10$. Теперь мы может подсчитать стандартизованные оценки за ЕГЭ:

Имя школьника	Петя	Коля	Таня	Маша	Андрей	Оля
Стандартизованная оценка $\frac{x_i - M_X}{\sigma_X}$	–1,2	–0,8	–0,6	0	1	1,6

Эти стандартизованные оценки позволяют, например, не просто сказать, что Оля лучше всех сдала ЕГЭ, но и количественно описать её положение в рассматриваемой группе из 6 школьников с учётом степени разброса всех оценок.

На приведённой ниже диаграмме эти стандартизованные оценки представлены графически, в виде столбиковой диаграммы. Вид этой диаграммы характеризует экзамен с точки зрения его сложности и критериев оценок (для данной группы школьников).



Предположим теперь, что эти ребята решили поступать в университет, который проводит дополнительный экзамен по математике, и получили оценки, которые приведены в следующей таблице (максимально возможное число баллов равно 100):

Имя школьника	Петя	Коля	Таня	Маша	Андрей	Оля
Исходная оценка y_i	81	80	80	80	90	87

Если зачисление проводится по результатам ЕГЭ и дополнительного вступительного экзамена, то обычно складывают оценки по двум экзаменам и затем отбирают столько лучших абитуриентов, сколько имеется мест. В нашем случае суммарные оценки приведены в следующей таблице.

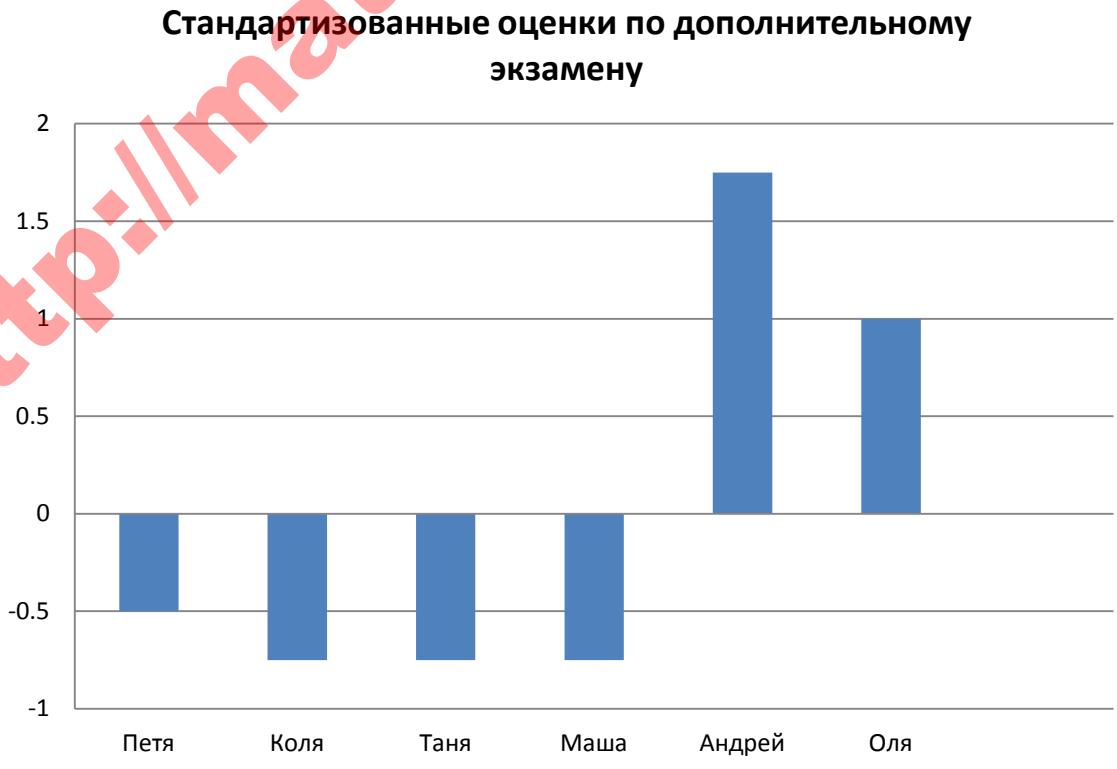
Имя школьника	Петя	Коля	Таня	Маша	Андрей	Оля
Сумма оценок	149	152	154	160	180	183

Поэтому, если, например, эти шесть абитуриентов претендуют на одно место, то зачислена в университет будет Оля, которая опередила ближайшего конкурента, Андрея, на 3 балла. Хотя Оля сдала дополнительный экзамен на 3 балла хуже, чем Андрей, преимущество в 6 баллов по результатам ЕГЭ обеспечило ей более высокий суммарный балл.

Проанализируем теперь эту ситуацию с помощью стандартизованных оценок. Нетрудно подсчитать, что для дополнительного экзамена средняя оценка равна $M_Y = \frac{498}{6} = 83$, а стандартное отклонение равно $\sigma_Y = 4$. Теперь мы может подсчитать стандартизованные оценки за дополнительный экзамен:

Имя школьника	Петя	Коля	Таня	Маша	Андрей	Оля
Стандартизованная оценка $\frac{y_i - M_Y}{\sigma_Y}$	–0,5	–0,75	–0,75	–0,75	1,75	1

На приведённой ниже столбиковой диаграмме эти стандартизованные оценки представлены графически.



Сопоставляя общий вид диаграмм для ЕГЭ и дополнительного экзамена, мы ясно видим, что они сильно отличаются друг от друга. Поэтому с точки зрения сложности и критериев оценок (для данной группы школьников) эти экзамены совершенно различны и, следовательно, сопоставлять первичные баллы не очень разумно. Переходя к стандартизованным оценкам, мы водим более объективные показатели уровня знаний школьников ***данной группы*** на ***конкретном экзамене***.

Для суммарных стандартизованных оценок мы имеем:

Имя школьника	Петя	Коля	Таня	Маша	Андрей	Оля
Сумма стандартизованных оценок	–1,7	–1,55	–1,35	–0,75	2,75	2,6

По данным этой таблицы лучшая суммарная оценка у Андрея и поэтому именно он, а не Оля, должен быть зачислен в университет.

Попробуем теперь разобраться, с чем связано преимущество Андрея; ведь по итогам первого экзамена он проигрывал Оле 6 баллов, а на втором экзамене смог опередить её лишь на 3 балла. Дело в том, что для результатов ЕГЭ стандартное отклонение равно 10, в то время как для результатов дополнительного экзамена стандартное отклонение равно 4. Поэтому большое отклонение оценки Оли по ЕГЭ от оценки Андрея по ЕГЭ (величина этого отклонения равна 6) можно объяснить просто большими колебаниями оценок по ЕГЭ ($\sigma_X = 10$ в два с половиной раза больше, чем $\sigma_Y = 4$). В то же время значение $\sigma_Y = 4$ говорит о том, что на дополнительном экзамене разброс оценок гораздо меньше. Поэтому преимущество Андрея на втором экзамене (он опередил Олю на 3 балла) имеет гораздо бóльшую значимость, чем преимущество Оли на ЕГЭ (где она опередила Андрея на 6 баллов).

Отметим, что если бы наши 6 абитуриентов претендовали не на одно, а на бóльшее число мест, то (для рассмотренных наборов первичных баллов) новая методика зачисления, основанная на суммировании стандартизованных оценок, дала бы тот же результат, что и традиционная методика суммирования абсолютных значений оценок.

Решение о том, какой способ определения суммарной оценки следует применять при решении вопроса о зачислении в университет, лежит вне математики. В реальности приёмные комиссии просто суммируют оценки за разные экзамены. Однако в подобных ситуациях, когда характер экзаменов и принципы выставления оценок сильно отличаются, статистика рекомендует использовать стандартизованные оценки или другие аналогичные процедуры.

Следующая задача предлагалась британским экзаменационным центром Edexcel в июне 2008 г. на выпускном школьном экзамене по курсу GCSE

Statistics ([27], задача №6; для удобства восприятия перевод оригинального английского условия задачи немного отредактирован).

Задача 4.1. В конце четверти в классе были проведены тесты по статистике и математике. Максимальное число баллов за каждый тест равно 100. Анализируя результаты этих тестов учитель установил, что

- по статистике средний балл равен 52, а стандартное отклонение равно 15,
- по математике средний балл равен 45, а стандартное отклонение равно 12.

(1) Прокомментируйте эти результаты.

Джон заработал 55 баллов по статистике и 48 баллов по математике.

(2) Подсчитайте стандартизованные оценки Джона по этим тестам.

(3) В каком предмете его успехи выше? Аргументируйте ваш ответ.

Решение. (1) Если мы посмотрим на результаты всего класса, то можно отметить, что средняя оценка по статистике выше средней оценки по математике. Это означает, что математика – более трудный предмет или тест по математике состоял из более сложных задач. Возможно, впрочем, что школьники меньше внимания уделяли математике или проходили сложные темы, которые трудно было хорошо понять. Различие в разбросе оценок также указывает на то, что степень усвоения материала школьниками и/или уровень сложности тестов по этим двум предметам различны. Поэтому сопоставление абсолютных значений оценок по эти предметам нельзя считать разумным. Простейший способ учесть отмеченные выше обстоятельства связан с использованием стандартизованных оценок.

(2) Стандартизованная оценка Джона по статистике равна

$$\frac{\text{первичный балл по статистике} - \text{средний балл по классу}}{\text{стандартное отклонение баллов учеников от среднего}} = \frac{55-52}{15} = 0,2.$$

Стандартизованная оценка Джона по математике равна

$$\frac{\text{первичный балл по математике} - \text{средний балл по классу}}{\text{стандартное отклонение баллов учеников от среднего}} = \frac{48-45}{12} = 0,25.$$

(3) Поскольку стандартизованная оценка Джона по математике больше, чем его стандартизованная оценка по статистике, мы должны признать, что по математике успехи Джона выше.

На первый взгляд этот вывод противоречит данным тестов: ведь по статистике Джон заработал на 7 баллов больше, чем по математике. Однако, как мы отмечали, тест по математике оказался для всего класса сложнее теста по статистике. Поэтому оценивая успехи Джона нужно принимать в расчёт общую успеваемость в классе. Отклонение оценки Джона по математике от средней по классу оценки по этому предмету равно $48-45=3$, т.е. такое же, как и отклонение оценки Джона по статистике от средней по классу оценки по

статистике ($55-52=3$). Но для всего набора оценок по статистике стандартное отклонение больше, чем стандартное отклонение для всего набора оценок по математике. Поэтому более высокий балл Джона по статистике может быть связан и с бóльшими колебаниями оценок по статистике. Чтобы нивелировать это различие, нужно измерять отклонения оценки по предмету от средней оценки не абсолютным числом баллов, а по отношению к типичному отклонению, одной из мер которого является стандартное отклонение.

Следующая задача взята из британского школьного учебника по статистике [9] (стр. 112; для удобства восприятия перевод оригинального английского условия задачи немного отредактирован).

Задача 4.2. В соревнованиях по фигурному катанию на льду принимали участие 20 спортсменов. Каждый из них должен был исполнить 2 танца – обязательный и произвольный. За каждый танец участник может получить максимум 50 баллов. По результатам соревнований оказалось, что для обязательного танца средняя оценка по группе равна 37,25, а стандартное отклонение оценки от средней равно 5,07. Для произвольного танца средняя оценка по группе равна 41,7, а стандартное отклонение оценки от средней равно 3,2.

(1) Участник А получил за исполнение обязательного танца 40 баллов, а за исполнение произвольного танца – 44 балла. Найдите соответствующие стандартизованные оценки и их сумму.

(2) Участник В получил за исполнение обязательного танца 43 балла, а за исполнение произвольного танца – 41 балл. Найдите соответствующие стандартизованные оценки и их сумму.

(3) Кто из этих двух участников показал лучший общий результат? Аргументируйте ваш ответ.

(4) Стандартизованная оценка участника С за исполнение обязательного танца равна 0,15. Сколько первичных баллов он получил за этот танец?

Решение.

(1) Стандартизованная оценка участника А за исполнение обязательного танца равна

$$\frac{\text{первичный балл участика А} - \text{средний балл}}{\text{стандартное отклонение}} = \frac{40-37{,}25}{5{,}07} \approx 0{,}542.$$

Стандартизованная оценка участника А за исполнение произвольного танца равна

$$\frac{\text{первичный балл участика А} - \text{средний балл}}{\text{стандартное отклонение}} = \frac{44-41{,}7}{3{,}2} \approx 0{,}719.$$

(2) Стандартизованная оценка участника В за исполнение обязательного танца равна

$$\frac{\text{первичный балл участика В} - \text{средний балл}}{\text{стандартное отклонение}} = \frac{43-37{,}25}{5{,}07} \approx 1{,}134.$$

Стандартизованная оценка участника В за исполнение произвольного танца равна

$$\frac{\text{первичный балл участика В} - \text{средний балл}}{\text{стандартное отклонение}} = \frac{41-41{,}7}{3{,}2} \approx -0{,}219.$$

(3) Сумма стандартизованных оценок участника А приближённо равна 1,26. Для участника В эта сумма приближённо равна 0,92. Поэтому следует признать, что участник А опередил участника В (несмотря на то, что простая сумма первичных баллов у этих участников одна и та же; она равна 84).

(4) Из формулы $z_i = \frac{x_i - M_x}{\sigma_x}$ мы имеем: $x_i = M_X + z_i\sigma_X$. Поэтому первичный балл участника С за исполнение обязательного танца равен $37{,}25 + 0{,}15 \times 5{,}07 \approx 38$.

4.6 Медиана и размах и линейные преобразования наборов данных

Теорема 3. *Пусть* μ_X *и* R_X *– соответственно медиана и размах набора* $X = [x_1, \ldots, x_n]$, *а* μ_U *и* R_U *– соответственно медиана и размах набора* $U = [u_1, \ldots, u_n]$, *связанного с набором X соотношениями (4.1). Тогда*

$$\mu_X = a\mu_U + b \tag{4.7}$$

$$R_X = |a| \cdot R_U \tag{4.8}$$

Доказательство. Упорядочим числа основного набора X по возрастанию, т.е. образуем вариационный ряд $x_{(1)} \le x_{(2)} \le \ldots \le x_{(n)}$. Вычитая из всех чисел этого ряда одно и то же число b (параметр сдвига), мы получим цепочку неравенств

$$x_{(1)} - b \le x_{(2)} - b \le \ldots \le x_{(n)} - b. \tag{4.9}$$

Дальнейшие рассуждения будут зависеть от знака масштабного коэффициента a.

1 случай. Если $a>0$, то деление всех чисел цепочки неравенств (4.9) на a даст:

$$\frac{x_{(1)} - b}{a} \le \frac{x_{(2)} - b}{a} \le \ldots \le \frac{x_{(n)} - b}{a}.$$

Поэтому, если $u^*_{(k)} = \frac{x_{(k)} - b}{a}$ – член набора $U = [u_1, \ldots, u_n]$, соответствующий k-му члену вариационного ряда $x_{(1)} \le x_{(2)} \le \ldots \le x_{(n)}$, то верны неравенства:

$$u^*_{(1)} \le u^*_{(2)} \le \ldots \le u^*_{(n)},$$

т.е. на самом деле число $u^*_{(k)}$ является k-м членом вариационного ряда $u_{(1)} \le u_{(2)} \le \ldots \le u_{(n)}$, соответствующего набору U: $u^*_{(k)} = u_{(k)}$. Отсюда следует, что для набора U размах равен

$$R_U \equiv u_{(n)} - u_{(1)} = u^*_{(n)} - u^*_{(1)} \equiv \frac{x_{(n)} - b}{a} - \frac{x_{(1)} - b}{a} = \frac{1}{a}\left(x_{(n)} - x_{(1)}\right) = \frac{1}{a} R_X,$$

что (в рассматриваемом случае $a>0$) равносильно равенству (4.8).

Для того, чтобы связать между собой медианы μ_X и μ_U исходного и нового наборов, необходимо дополнительно отдельно рассмотреть случаи: (1) n – нечётное число и (2) n – чётное число.

(1) Если $n = 2k-1$ – нечётное число, то в соответствии с определением медианы мы имеем: $\mu_U = u_{(k)} = u^*_{(k)} = \frac{x_{(k)} - b}{a} = \frac{\mu_X - b}{a}$, что равносильно (4.7).

(2) Если $n=2k$ – чётное число, то в соответствии с определением медианы мы имеем:

$$\mu_U = \frac{u_{(k)} + u_{(k+1)}}{2} = \frac{u^*_{(k)} + u^*_{(k+1)}}{2} = \frac{\frac{x_{(k)} - b}{a} + \frac{x_{(k+1)} - b}{a}}{2}$$

$$= \frac{x_{(k)} + x_{(k+1)} - 2b}{2a} = \frac{\frac{x_{(k)} + x_{(k+1)}}{2} - b}{a} = \frac{\mu_X - b}{a},$$

и потому равенство (4.7) верно и в этом случае.

2 случай. Если $a<0$, то почленное деление (4.9) на a изменит знаки неравенств на противоположные:

$$\frac{x_{(1)} - b}{a} \geq \frac{x_{(2)} - b}{a} \geq \ldots \geq \frac{x_{(n)} - b}{a}.$$

Поэтому, если $u^*_{(k)} = \frac{x_{(k)} - b}{a}$ – член набора $U = \left[u_1, \ldots, u_n\right]$, соответствующий k-му члену вариационного ряда $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$, то верны неравенства:

$$u^*_{(1)} \geq u^*_{(2)} \geq \ldots \geq u^*_{(n)} \Leftrightarrow u^*_{(n)} \leq u^*_{(n-1)} \leq \ldots \leq u^*_{(1)}$$

т.е. на самом деле число $u^*_{(k)}$ является $(n-k+1)$-м членом вариационного ряда $u_{(1)} \leq u_{(2)} \leq \ldots \leq u_{(n)}$, соответствующего набору U: $u^*_{(k)} = u_{(n-k+1)}$. Если индекс $n-k+1$ обозначить какой-нибудь новой буквой, скажем, l, то $k = n-l+1$ и это соотношение примет вид:

$$u_{(l)} = u^*_{(n-l+1)} \ . \qquad (4.10)$$

Отсюда следует, что в случае $a<0$ размах набора U равен

$$R_U \equiv u_{(n)} - u_{(1)} = u^*_{(1)} - u^*_{(n)} \equiv \frac{x_{(1)} - b}{a} - \frac{x_{(n)} - b}{a} = -\frac{1}{a}\left(x_{(n)} - x_{(1)}\right) = \frac{1}{(-a)} R_X,$$

что (в рассматриваемом случае $a<0$) равносильно равенству (4.8).

Для того, чтобы связать между собой медианы μ_X и μ_U исходного и нового наборов, необходимо, как и ранее, дополнительно отдельно рассмотреть случаи: (1) n – нечётное число и (2) n – чётное число.

(1) Если $n=2k-1$ – нечётное число, то, используя определение медианы и соотношение (4.10), мы имеем: $\mu_U = u_{(k)} = u^*_{(n-k+1)} = \dfrac{x_{(k)}-b}{a} = \dfrac{\mu_X - b}{a}$, так что равенство (4.7) верно и в этом случае.

(2) Если $n=2k$ – чётное число, то, используя определение медианы и соотношение (4.10), мы имеем:

$$\mu_U = \frac{u_{(k)}+u_{(k+1)}}{2} = \frac{u^*_{(n-k+1)}+u^*_{(n-(k+1)+1)}}{2} = \frac{\dfrac{x_{(k+1)}-b}{a}+\dfrac{x_{(k)}-b}{a}}{2}$$

$$= \frac{x_{(k)}+x_{(k+1)}-2b}{2a} = \frac{\dfrac{x_{(k)}+x_{(k+1)}}{2}-b}{a} = \frac{\mu_X - b}{a},$$

и потому равенство (4.7) верно и в этом случае.

Замечание. Соотношения (4.7) и (4.8) полностью аналогичны соотношениям (4.2) и (4.5). Это вполне соответствует тому, что медиана, как и среднее, является мерой положения чисел набора, а размах, как и стандартное отклонение, является мерой разброса.

Пример 4.5. Пусть дан набор

$$x_1 = 2{,}91, x_2 = 3{,}07, x_3 = 3{,}02, x_4 = 3{,}05, x_5 = 3{,}01, x_6 = 3{,}03.$$

Как мы видели в Примерах 4.2 и 4.4, линейное преобразование $x = 3 + 0{,}01u$ с параметрами $a = 0{,}01$, $b = 3$ приводит к набору $U = \left[-9;7;2;5;1;3\right]$.

Вариационный ряд для исходного набора выглядит так:

$$x_{(1)} = 2{,}91, x_{(2)} = 3{,}01, x_{(3)} = 3{,}02, x_{(4)} = 3{,}03, x_{(5)} = 3{,}05, x_{(6)} = 3{,}07.$$

Соответственно $\min\limits_{1\le i\le 6} x_i = x_{(1)} = 2{,}91$, $\max\limits_{1\le i\le 6} x_i = x_{(6)} = 3{,}07$ и потому размах $R_X = x_{(6)} - x_{(1)} = 3{,}07 - 2{,}91 = 0{,}16$.

Вариационный ряд для набора U выглядит так:

$$u_{(1)} = -9, u_{(2)} = 1, u_{(3)} = 2, u_{(4)} = 3, u_{(5)} = 5, u_{(6)} = 7.$$

Соответственно $\min\limits_{1\le i\le 6} u_i = u_{(1)} = -9$, $\max\limits_{1\le i\le 6} u_i = u_{(6)} = 7$ и потому размах $R_U = u_{(6)} - u_{(1)} = 7 - (-9) = 16$.

Как и следовало ожидать на основе формулы (4.8), величины R_X и R_U связаны соотношением: $R_X = 0{,}01R_U$ (напомним, что в рассматриваемом примере масштабный параметр a равен 0,01).

Для медиан мы имеем: $\mu_X = \dfrac{x_{(3)} + x_{(4)}}{2} = \dfrac{3{,}02 + 3{,}03}{2} = 3{,}025$, $\mu_U = \dfrac{u_{(3)} + u_{(4)}}{2} = \dfrac{2+3}{2} = 2{,}5$. Как и следовало ожидать на основе формулы (4.7), величины μ_X и μ_U связаны соотношением: $\mu_X = 0{,}01\mu_U + 3$ (напомним, что в рассматриваемом примере масштабный параметр a равен 0,01, а параметр сдвига $b = 3$).

4.7 Дополнительные задачи

Задача 4.3. В магазине продают семь видов сока в литровых упаковках по следующим ценам (в рублях):

$$x_1 = 28{,}90, x_2 = 34{,}90, x_3 = 29{,}90, x_4 = 32{,}90, x_5 = 39{,}90, x_6 = 24{,}90, x_7 = 38{,}90.$$

Определите среднюю цену литрового пакета сока, стандартное отклонение от этой цены, а также медиану и размах цен.

Решение. Все цены оканчиваются записью «,90» – это обычный приём розничной торговли, который позволяет создать впечатление, что товар дешевле, чем на самом деле. Например, для потребителя, в сущности, нет никакой разницы между 40 руб. и 39,90 руб. Однако психологически потребитель относит товар ценой 39,90 руб. в категорию 30 рублёвых товаров и может купить его, в то время как указание цены 40 рублей автоматически переводит товар в категорию 40 рублёвых товаров, что может быть неприемлемо для покупателя.

Имея в виду это соображение, введём округлённые цены: $u_1 = x_1 + 0{,}10 = 29$, $u_2 = x_2 + 0{,}10 = 35$, $u_3 = x_3 + 0{,}10 = 30$, $u_4 = x_4 + 0{,}10 = 33$, $u_5 = x_5 + 0{,}10 = 40$, $u_6 = x_6 + 0{,}10 = 25$, $u_7 = x_7 + 0{,}10 = 39$.

Значения этих округлённых цен колеблются вокруг числа 30. Имея в виду это соображение, введём ещё один набор $V = \left[v_1, v_2, \ldots, v_7\right]$ по формуле: $u_i = 30 + v_i$, так что $V = \left[-1;5;0;3;10;-5;9\right]$. Значения набора X связаны со значениями набора V формулой: $x_i = u_i - 0{,}1 = 30 + v_i - 0{,}1 = v_i + 29{,}90$.

Среднее значение набора V легко подсчитать в уме: $M_V = 3$. Теперь для дисперсии набора V мы имеем (все вычисления можно проделать без калькулятора): $D_V = \dfrac{178}{7} \approx 25{,}43$. Соответственно, $\sigma_V = \sqrt{D_V} \approx 5{,}04$.

Чтобы подсчитать медиану и размах, образуем вариационный ряд вспомогательного набора V – это будет упорядоченный набор $\left(-5;-1;0;3;5;9;10\right)$. Размах этого набора равен $R_V = 10 - (-5) = 15$. Далее, количество чисел в наборе V (как и в исходном наборе X) равно 7, т.е.

является нечётным числом. Поэтому медиана набора V – это четвёртый член соответствующего вариационного ряда, т.е. $\mu_V = 3$.

Теперь легко подсчитать статистические характеристики основного набора. Средние значения и медианы связаны тем же линейным соотношением $x = v + 29{,}90$, которое связывает числа наборов X и V: $M_X = M_V + 29{,}90 = 32{,}90$, $\mu_X = \mu_V + 29{,}90 = 32{,}90$, а поскольку масштабный коэффициент при преобразовании набора V в набор X равен 1, их меры разброса совпадают: $\sigma_X = \sigma_V \approx 5{,}04$, $R_X = R_V = 15$.

Задача 4.4. В витрине обувного магазина выставлены шесть пар ботинок по следующим ценам (в рублях):

$$x_1 = 3490, x_2 = 3990, x_3 = 5990, x_4 = 1990, x_5 = 4490, x_6 = 990.$$

Определите среднюю цену, стандартное отклонение от этой цены, а также медиану и размах цен.

Решение. Введём округлённые цены $u_i = x_i + 10$. В результате мы получим следующий набор чисел: $U = \left[3500,\ 4000,\ 6000,\ 2000,\ 4500,\ 1000\right]$.

Все эти числа кратны 500. Имея в виду это соображение, введём новые переменные v_i по формуле: $u_i = 500v_i$. В результате мы получим следующий набор чисел: $V = \left[7,\ 8,\ 12,\ 4,\ 9,\ 2\right]$. Значения основного набора связаны со значениями этого набора соотношением $x_i = u_i - 10 = 500v_i - 10$.

Среднее значение набора V легко подсчитать в уме: $M_V = 7$.

Теперь для дисперсии набора V мы имеем (все вычисления можно проделать без калькулятора): $D_V = \frac{64}{6} \approx 10{,}667$, так что $\sigma_V = \sqrt{D_V} \approx 3{,}266$.

Чтобы подсчитать медиану и размах цен, образуем вариационный ряд вспомогательного набора V – это будет упорядоченный набор (2;4;7;8;9;12). Размах этого набора равен $R_V = 12 - 2 = 10$. Далее, количество чисел как в наборе V (как и в исходном наборе X) равно 6, т.е. является чётным числом. Поэтому медиана набора V – это среднее арифметическое третьего и четвёртого членов соответствующего вариационного ряда, т.е. $\mu_V = 7{,}5$.

Теперь легко подсчитать статистические характеристики основного набора. Средние значения и медианы связаны тем же линейным соотношением $x = 500v - 10$, которое связывает числа наборов X и V: $M_X = 500M_V - 10 = 3490$, $\mu_X = 500\mu_V - 10 = 3740$, а поскольку масштабный коэффициент при преобразовании набора V в набор X равен 500, их меры разброса есть: $\sigma_X = 500\sigma_V \approx 1633$, $R_X = 500R_V = 5000$.

Задача 4.5. На метеорологической станции ежедневно в полдень фиксируется температура воздуха. Наблюдения в течение одной недели в

январе дали следующие значения (в градусах Цельсия): $-7,\ -9,\ -7,\ -5,\ -9,\ -10,\ -9$. Определите среднюю температуру на этой неделе, стандартное отклонение от этой средней температуры, а также медиану и размах температур.

Решение. Пусть x_i – температура в i-й день недели, i=1,…,7, так что, например, $x_1 = -7, x_7 = -9$. Чтобы не работать с отрицательными числами, введём абсолютные значения температур u_i по формуле: $u_i = -x_i$. В результате мы получим следующий набор чисел: U=(7;9;7;5;9;10;9). Используя круглые скобки, мы подчёркиваем, что набор является упорядоченным, т.е. мы сохраняем информацию о днях недели.

На глаз можно оценить, что эти числа колеблются вокруг значения 7 или 8. Имея в виду это соображение, введём новые переменные v_i по формуле: $v_i = u_i - 8$ (с равным успехом можно было бы использовать формулу $v_i = u_i - 7$; мы рекомендуем читателю самостоятельно проделать расчёты, используя это преобразование, и сравнить окончательные числовые значения статистических характеристик основного набора с полученными ниже). В результате мы получим следующий набор чисел: $V = (-1;1;-1;-3;1;2;1)$. Значения основного набора связаны со значениями этого вспомогательного набора соотношением $x_i = -v_i - 8$.

Среднее значение набора V легко подсчитать в уме: $M_V = 0$.

Теперь для дисперсии набора V мы имеем (все вычисления можно проделать без калькулятора): $D_V = \frac{18}{7} \approx 2{,}57$, так что $\sigma_V = \sqrt{D_V} \approx 1{,}60$.

Чтобы подсчитать медиану и размах температур, образуем вариационный ряд вспомогательного набора V – это будет набор $(-3;-1;-1;1;1;1;2)$. Размах этого набора равен $R_V = 2 - (-3) = 5$. Далее, количество чисел в наборе V (как и в исходном наборе X) равно 7, т.е. является нечётным числом. Поэтому медиана набора V – это четвёртый член соответствующего вариационного ряда, т.е. $\mu_V = 1$.

Теперь легко подсчитать статистические характеристики основного набора. Средние значения и медианы связаны тем же линейным соотношением $x = -v - 8$, которое связывает числа наборов X и V: $M_X = -M_V - 8 = -8$, $\mu_X = -\mu_V - 8 = -9$, а поскольку значение масштабного параметра линейного преобразования, связывающего наборы V и X равно (-1), их меры разброса совпадают: $\sigma_X = \sigma_V \approx 1{,}60$, $R_x = R_v = 5$.

5. Неравенства для статистических характеристик

5.1 Неравенства Чебышева для статистических характеристик числового набора

Неравенства Чебышева – это группа однотипных по внешнему виду и методу доказательства неравенств для вероятностей выхода случайной величины за некоторые границы. Они названы в честь великого русского математика и механика Панфутия Львовича Чебышева (1821-1894), который в 1866 году доказал наиболее важное из этих неравенств и, что самое главное, применил его для простого доказательства одного из наиболее значимых результатов теории вероятностей, так называемого закона больших чисел. В этом пункте мы расскажем о естественных аналогах нескольких неравенств Чебышева для обычных числовых наборов (о неравенстве Чебышева в теории вероятностей можно прочитать в замечательном пособии [6], §26). Эти неравенства не требуют использования языка теории вероятностей и их доказательства совершенно элементарны.

Итак, пусть $X = [x_1, \ldots, x_n]$ – набор n значений некоторой переменной X. Обозначим через $N(X \geq a)$ количество чисел набора, удовлетворяющих неравенству $x \geq a$, а через $\nu(X \geq a) = \dfrac{N(X \geq a)}{n}$ – долю таких чисел.

Чтобы проиллюстрировать эти обозначения, рассмотрим набор X из 100 чисел:

27, 52, 43, 38, 47, 8, 21, 40, 32, 53, 45, 54, 35, 28, 40, 18, 31, 45, 24, 30, 37, 15, 39, 34, 48, 25, 30, 7, 32, 12, 26, 35, 48, 19, 33, 26, 17, 30, 42, 22, 53, 28, 42, 36, 23, 10, 34, 46, 16, 29, 35, 52, 41, 32, 21, 39, 55, 25, 29, 8, 36, 44, 26, 55, 34, 19, 42, 54, 27, 10, 45, 20, 31, 50, 18, 9, 41, 14, 38, 40, 23, 49, 33, 15, 24, 46, 36, 28, 32, 37, 51, 20, 29, 47, 33, 27, 41, 22, 39, 40.

Этот набор мы взяли из учебника [3] , стр. 251; числа набора – это время (в минутах), которое 100 выбранных наугад учеников гипотетической школы тратят на дорогу в школу. Ниже мы несколько раз будем использовать его в разных разделах в иллюстративных целях.

Если $a = 45$, то, как нетрудно подсчитать, ровно 20 чисел из этого набора удовлетворяют неравенству $x \geq 45$. Поэтому для рассматриваемого набора $N(X \geq 45) = 20$, а $\nu(X \geq 45) = 0{,}2$. Величина $N(X \geq 45)$ показывает, сколько учеников тратит на дорогу достаточно большое время (по меньшей мере три четверти часа) и намного лучше характеризует набор, чем стандартные характеристики: среднее значение, медиана и дисперсия. Ещё интереснее величина $\nu(X \geq 45)$, которая говорит, что 20% учеников слишком долго (по меньшей мере три четверти часа) добираются до школы.

Если $a = 6$, то, как нетрудно определить, все числа набора удовлетворяют неравенству $x \geq 6$. Поэтому для рассматриваемого набора $N(X \geq 6) = 100$, а $\nu(X \geq 6) = 1$.

Если $a = 56$, то, как нетрудно определить, ни одно число набора не удовлетворяет неравенству $x \geq 56$. Поэтому для рассматриваемого набора $N(X \geq 56) = 0$, $\nu(X \geq 56) = 0$.

Первое неравенство Чебышева (его иногда называют неравенством Маркова), которое мы сейчас докажем, устанавливает связь между этими новыми характеристиками числового набора и основной мерой положения чисел набора, средним значением M_x.

Теорема 5.1 (неравенство Чебышева). *Для произвольного набора* $X = [x_1, \ldots, x_n]$ *неотрицательных чисел и любого положительного числа* a *справедливо неравенство* $\nu(X \geq a) \leq \dfrac{M_X}{a}$.

Доказательство. Разобьём сумму $S = x_1 + \ldots + x_n$ всех чисел набора на две суммы: $S = S' + S''$. В первую включим те члены набора, которые меньше, чем заданное число a, а во вторую – те члены набора, которые больше или равны a.

Поскольку наш набор состоит из неотрицательных чисел, сумма $S' \geq 0$. Если в наборе нет чисел меньших a, мы положим $S' = 0$, так что в любом случае верно неравенство $S' \geq 0$.

Вторая сумма состоит из $N(X \geq a)$ чисел набора, каждое из которых не меньше a. Поэтому $S'' \geq a \cdot N(X \geq a)$. Если в наборе нет чисел бóльших или равных, чем a (т.е. $N(X \geq a) = 0$), мы положим $S'' = 0$, так что в любом случае верно неравенство $S'' \geq a \cdot N(X \geq a)$.

Теперь мы имеем:

$$M_X = \frac{x_1 + \ldots + x_n}{n} = \frac{S' + S''}{n} \geq \frac{0 + a \cdot N(X \geq a)}{n} = a \cdot \nu(X \geq a),$$

что равносильно доказываемому неравенству Чебышева.

Метод, применённый при доказательстве Теоремы 5.1 позволяет получить большое число других неравенств. В качестве примера мы докажем следующее утверждение, которое связывает медиану μ_X, среднее значение M_X, наименьшее $\min(x_1, \ldots, x_n)$ и наибольшее $\max(x_1, \ldots, x_n)$ числа набора.

Теорема 5.2. *Если набор* $X=\left[x_1,\ldots,x_n\right]$ *состоит из нечётного количества чисел, то*

$$\frac{2n}{n+1}\cdot M_X-\frac{n-1}{n+1}\cdot\max\left(x_1,\ldots,x_n\right)\le\mu_X\le\frac{2n}{n+1}\cdot M_X-\frac{n-1}{n+1}\cdot\min\left(x_1,\ldots,x_n\right).$$

Если же количество чисел в наборе чётно, то

$$\frac{2n}{n+2}\cdot M_X-\frac{n-2}{n+2}\cdot\max\left(x_1,\ldots,x_n\right)\le\mu_X\le\frac{2n}{n+2}\cdot M_X-\frac{n-2}{n+2}\cdot\min\left(x_1,\ldots,x_n\right).$$

Доказательство. Будем считать, что числа набора упорядочены по возрастанию: $x_1\le x_2\le\ldots\le x_n$. Тогда $\min\left(x_1,\ldots,x_n\right)=x_1$, $\max\left(x_1,\ldots,x_n\right)=x_n$. Порядок, в котором расположены эти числа не влияет на величину их суммы и, соответственно, на среднее значение набора, а также на экстремальные значения.

Предположим, что $n=2k+1$, т.е. набор состоит из нечётного количества чисел. Тогда медиана μ_X является (k+1)-м в порядке возрастания членом набора: $\mu_X=x_{k+1}$. Запишем сумму S всех чисел набора в виде: $S=\left(x_1+\ldots+x_k\right)+x_{k+1}+\left(x_{k+2}+\ldots+x_{2k+1}\right)$.

Все слагаемые в сумме $x_1+\ldots+x_k$ первых k членов набора больше или равны, чем x_1 (напомним, что мы упорядочили числа набора по возрастанию), член x_{k+1} в точности равен медиане μ_X, а все слагаемые в сумме $x_{k+2}+\ldots+x_{2k+1}$, включающей k последних членов набора, больше или равны, чем μ_X. Поэтому $S\ge kx_1+(k+1)\mu_X$. Поскольку $S=nM_X$, а $k=\frac{n-1}{2}$, мы получим: $\mu_X\le\frac{2n}{n+1}\cdot M_X-\frac{n-1}{n+1}\cdot\min\left(x_1,\ldots,x_n\right)$.

Для оценки с другой стороны отметим, что все слагаемые в сумме $x_1+\ldots+x_k$ первых k членов набора меньше или равны, чем μ_X, член x_{k+1} в точности равен медиане μ_X, а все слагаемые в сумме $x_{k+2}+\ldots+x_{2k+1}$, включающей k последних членов набора меньше или или равны, чем x_n. Поэтому $S\le kx_n+(k+1)\mu_X$. Поскольку $S=nM_X$, а $k=\frac{n-1}{2}$, мы получим: $\mu_X\ge\frac{2n}{n+1}\cdot M_X-\frac{n-1}{n+1}\cdot\max\left(x_1,\ldots,x_n\right)$.

Предположим теперь, что $n=2k$, т.е. набор состоит из нечётного количества чисел. Тогда медиана μ_X является средним арифметическим k-го (k+1)-го в порядке возрастания членов набора: $\mu_X=\frac{x_k+x_{k+1}}{2}$. Запишем сумму S всех чисел набора в виде: $S=\left(x_1+\ldots+x_{k-1}\right)+\left(x_k+x_{k+1}\right)+\left(x_{k+2}+\ldots+x_{2k}\right)$.

Все слагаемые в сумме $x_1+\ldots+x_{k-1}$ первых $k-1$ членов набора больше или равны, чем x_1, сумма x_k+x_{k+1} в точности равна $2\mu_X$, а все слагаемые в сумме $x_{k+2}+\ldots+x_{2k}$, включающей $k-1$ последних членов набора больше или равны, чем μ_X (эти слагаемые не меньше x_{k+1}, которое, в свою очередь, не меньше, чем среднее арифметическое $\frac{x_k+x_{k+1}}{2}=\mu_X$). Поэтому $S\geq(k-1)x_1+(k+1)\mu_X$. Поскольку $S=nM_X$, а $k=\frac{n}{2}$, мы получим: $\mu_X\leq\frac{2n}{n+2}\cdot M_X-\frac{n-2}{n+2}\cdot\min\left(x_1,\ldots,x_n\right)$.

Для оценки с другой стороны отметим, что все слагаемые в сумме $x_1+\ldots+x_{k-1}$ первых $k-1$ членов набора меньше или равны, чем μ_X, сумма x_k+x_{k+1} в точности равна $2\mu_X$, а все слагаемые в сумме $x_{k+2}+\ldots+x_{2k}$, включающей $k-1$ последних членов набора меньше или равны, чем x_n. Поэтому $S\leq(k-1)x_n+(k+1)\mu_X$. Поскольку $S=nM_X$, а $k=\frac{n}{2}$, мы получим: $\mu_X\geq\frac{2n}{n+2}\cdot M_X-\frac{n-2}{n+2}\cdot\max\left(x_1,\ldots,x_n\right)$.

Замечание 5.1. Утверждение теоремы 5.2 неулучшаемо в том смысле, что существуют наборы, для которых оба доказанных двойных неравенства превращаются в точные равенства.

Например, если в случае нечётного $n=2k+1$ взять набор из k чисел 0 и $k+1$ чисел 1 (тогда медиана равна 1, среднее значение равно $\frac{k+1}{n}=\frac{n+1}{2n}$, наименьшее значение равно 0, наибольшее значение равно1), то в точное равенство превратится правая часть первого неравенства. Если же взять набор из $k+1$ чисел 0 и k чисел 1 (тогда медиана равна 0, среднее значение равно $\frac{k}{n}=\frac{n-1}{2n}$, наименьшее значение равно 0, наибольшее значение равно1), то в точное равенство превратится левая часть первого неравенства.

В случае чётного $n=2k$ можно взять набор из $k-1$ чисел 0 и $k+1$ чисел 1 (тогда медиана равна 1, среднее значение $\frac{k+1}{n}=\frac{n+2}{2n}$, наименьшее значение равно 0, наибольшее значение равно1), то в точное равенство превратится правая часть второго неравенства. Если же взять набор из $k+1$ чисел 0 и $k-1$ чисел 1 (тогда медиана равна 0, среднее значение равно $\frac{k-1}{n}=\frac{n-2}{2n}$, наименьшее значение равно 0, наибольшее значение равно1), то в точное равенство превратится левая часть второго неравенства.

Замечание 5.2. Если набор $X=\left[x_1,\ldots,x_n\right]$ состоит из *неотрицательных* чисел, то правые части доказанных неравенств влекут, что $\mu_X \le \frac{2n}{n+1}\cdot M_X$, если n нечётно, и $\mu_X \le \frac{2n}{n+2}\cdot M_X$, если n чётно. Дроби $\frac{n}{n+1}$ и $\frac{n}{n+2}$ строго меньше 1. Поэтому для любого набора из неотрицательных чисел, исключая вырожденный случай, когда набор состоит из одних нулей, верно строгое неравенство $\mu_X < 2M_X$.

С помощью теоремы 5.1 можно легко получить ещё одно неравенство. Обычно в учебниках по теории вероятностей именно его (точнее, его вероятностный аналог) называют неравенством Чебышева.

Теорема 5.3 (неравенство Чебышева для абсолютных отклонений от среднего). *Для произвольного числового набора* $X=\left[x_1,\ldots,x_n\right]$ *и любого положительного числа* a *справедливо неравенство* $\nu\left(\left|X-M_X\right|\ge a\right)\le\frac{D_X}{a^2}$, *т.е. доля тех чисел набора, которые отклоняются от среднего значения на величину* a *или больше, не превосходит* $\frac{D_X}{a^2}$.

Доказательство. Рассмотрим новый набор Y из n чисел, поученных из чисел исходного набора по формуле $y=\left(x-M_X\right)^2$:

$$y_1=\left(x_1-M_X\right)^2, y_2=\left(x_2-M_X\right)^2,\ldots,y_n=\left(x_n-M_X\right)^2.$$

Эти числа, очевидно, неотрицательны и потому к ним применимо первое неравенство Чебышева, доказанное в Теореме 5.1. Если в качестве числа a мы возьмём число a^2, то мы получим неравенство

$$\nu(Y\ge a^2)\le\frac{M_Y}{a^2}. \qquad (5.1)$$

Поскольку $y=\left(x-M_X\right)^2$, неравенство $y\ge a^2$ равносильно неравенству $\left(x-M_X\right)^2\ge a^2$, или, после извлечения квадратного корня, неравенству $\left|x-M_X\right|\ge a$. Соответственно, количество чисел набора $Y=\left[y_1,\ldots,y_n\right]$, удовлетворяющих неравенству $y\ge a^2$, совпадает с количеством чисел исходного набора $X=\left[x_1,\ldots,x_n\right]$, удовлетворяющих неравенству $\left|x-M_X\right|\ge a$:

$$N\left(Y\ge a^2\right)=N\left(\left|X-M_X\right|\ge a\right)\Leftrightarrow\nu\left(Y\ge a^2\right)=\nu\left(\left|X-M_X\right|\ge a\right).$$

Кроме того, среднее значение M_Y, равное по определению $\frac{y_1+y_2+\ldots+y_n}{n}$, может быть переписано как $\frac{\left(x_1-M_X\right)^2+\left(x_2-M_X\right)^2+\ldots+\left(x_n-M_X\right)^2}{n}$, т.е. на самом

деле равно дисперсии D_X исходного набора. Заменяя в неравенстве (5.1) $\nu\left(Y \ge a^2\right)$ на $\nu\left(\left|X - M_X\right| \ge a\right)$, а M_Y на D_X, мы и получим требуемый результат.

Чтобы проиллюстрировать применение доказанных неравенств Чебышева, рассмотрим набор, приведённый в начале этого пункта. Нетрудно подсчитать (хотя это довольно утомительно), что $M_X = 32{,}67$, $D_X = 150{,}9811$. Поэтому в случае a=45 первое неравенство Чебышева примет вид: $0{,}2 \le \frac{32{,}67}{45}$ или, что то же самое, $0{,}2 \le 0{,}726$. Чтобы проиллюстрировать второе неравенство Чебышева возьмём a=15. Неравенству $\left|x - M_X\right| \ge 15 \Leftrightarrow x \le 17{,}67$ или $x \ge 47{,}67$ удовлетворяет 25 чисел набора. Поскольку $D_X = 150{,}9811$, второе неравенство Чебышева примет вид $0{,}25 \le \frac{150{,}9811}{15^2}$ или, что то же самое, $0{,}25 \le 0{,}671027$. Конечно, эти числовые неравенства истинны. Но нельзя не отметить, что они довольно «грубые». Истинное значение неравенств Чебышева проявляется в теоретических рассмотрениях теории вероятностей и прежде всего при доказательстве знаменитого «закона больших чисел» (в связи с этим отметим, что $1 - \nu(X \ge a)$ фактически является эмпирической функцией распределения). Сам Чебышев доказал неравенство, называемое его именем, не в терминах описательной статистики, а на языке теории вероятностей, в связи с изучением предельного поведения среднего арифметического большого числа независимых случайных величин.

Сейчас мы рассмотрим одно применение неравенства Чебышева для анализа важной теоретической проблемы описательной статистики.

В школьных учебниках про среднее значение M_X обычно говорят, что числа набора *группируются* вокруг среднего значения, а дисперсия D_X характеризует величину рассеивания данных вокруг среднего арифметического. Однако обычные школьные определения среднего значения и дисперсии не дают никаких оснований для подобных утверждений. В особенности это относится к дисперсии, определение которой содержит смутные рассуждения о необходимости возведения в квадрат отклонений $x_i - M_X$ для того, чтобы «стереть» знаки у этих отклонений (на самом деле для этого естественно просто взять модули отклонений, но как же тогда объяснить смысл введения дисперсии?).

Неравенство Чебышева позволяет немного прояснить этот вопрос. Прежде всего, отметим следующее соображение: естественно говорить, что числа набора *группируются* вокруг среднего значения, если лишь небольшая их доля более-менее сильно отклоняется от среднего значения, т.е. если a – предельно допустимое абсолютное отклонение, то число $\nu\left(\left|X - M_X\right| \ge a\right)$ достаточно мало. Теорема 5.3 для любого набора оценивает это число дробью $\frac{D_X}{a^2}$. Число a по

смыслу рассматриваемой проблемы фиксировано. Поэтому при малой дисперсии можно говорить, что числа набора группируются вокруг среднего значения. В этом смысле дисперсия может рассматриваться как мера рассеивания.

Ещё один аргумент в пользу этого утверждения может быть получен следующим образом. Пусть предельно допустимое отклонение от среднего (число a) равно $3\sigma_x$, где $\sigma_X = \sqrt{D_X}$ – среднее квадратичное отклонение. Тогда второе неравенство Чебышева примет вид: $\nu\left(\left|X - M_X\right| \geq 3\sigma_X\right) \leq \frac{1}{9}$, т.е. для *любого* набора количество чисел, отклоняющихся от среднего значения больше (нестрого), чем на три средних квадратичных отклонения, не превосходит $\frac{1}{9} \approx 11\%$ от общего количества чисел в наборе, т.е. по меньшей мере 89% всех чисел *любого* набора расположены в интервале $\left(M_X - 3\sigma_X; M_X + 3\sigma_X\right)$. Это одна из форм знаменитого правила «трёх сигм». Для конкретных наборов эта доля может быть не 89%, а гораздо выше. Например, для набора, приведённого в начале этого пункта, как мы уже отмечали, $M_X = 32{,}67$, $D_X = 150{,}9811$. Поэтому $\sigma_X \approx 12{,}29$, так что интервал $\left(M_X - 3\sigma_X; M_X + 3\sigma_X\right) = \left(-4{,}19; 69{,}53\right)$. В него попадает 100% чисел набора.

Если предельно допустимое отклонение от среднего характеризовать не абсолютным значением (числом a), а достаточно малой долей $\varepsilon > 0$ от среднего значения, т.е. считать, что $a = \varepsilon \cdot \left|M_X\right|$, то (при $M_X \neq 0$) Теорема 5.3 примет вид: $\nu\left(\left|X - M_X\right| \geq \varepsilon \cdot \left|M_X\right|\right) \leq \frac{1}{\varepsilon^2} \cdot \frac{D_X}{M_X^2}$. В этом случае мы можем говорить, что числа набора группируются вокруг среднего значения, если мало число $\frac{D_X}{M_X^2}$ или, что то же самое, мало число $\frac{\sigma_X}{\left|M_X\right|}$. Дробь $\frac{\sigma_X}{\left|M_X\right|}$ называется коэффициентом вариации – это ещё одна важная мера разброса чисел набора.

Ещё один подход к определению меры положения числового набора, при котором естественно появляются медиана, среднее арифметическое и дисперсия, а также проясняется связь между средним арифметическим и дисперсией, был описан в разделе 2. Но истинный смысл введения понятий среднего значения и дисперсии можно объяснить только с помощью предельных теорем теории вероятностей (закона больших чисел и, главным образом, центральной предельной теоремы).

5.2 Неравенство, связывающее среднее значение, медиану и стандартное отклонение

Среднее значение и медиана описывают положение чисел набора на числовой оси, а среднее квадратичное отклонение характеризует разброс чисел вокруг среднего значения. Хотя эти статистические характеристики числового набора определяются совершенно независимо друг от друга, между ними имеется интересная связь. Следующая теорема является аналогом для числовых наборов известного неравенства из теории вероятностей. Идейно она связана с теоремой 5.2.

Теорема 5.4. *Для произвольного числового набора справедливо неравенство*

$$|\mu_X - M_X| \le \sigma_X. \qquad (5.2)$$

Доказательство. В выражении $|\mu_X - M_X|$ заменим M_X на $\frac{x_1 + \ldots + x_n}{n}$:

$$|\mu_X - M_X| = \left|\mu_X - \frac{x_1 + \ldots + x_n}{n}\right| = \frac{\left|(x_1 - \mu_X) + \ldots + (x_n - \mu_X)\right|}{n}.$$

Поскольку модуль суммы меньше или равен суммы модулей,

$$|\mu_X - M_X| \le \frac{|x_1 - \mu_X| + \ldots + |x_n - \mu_X|}{n}.$$

Теорема 2.2 об экстремальном свойстве медианы влечёт, что $|x_1 - \mu_X| + \ldots + |x_n - \mu_X| \le |x_1 - M_X| + \ldots + |x_n - M_X|$, откуда мы получим:

$$|\mu_X - M_X| \le \frac{|x_1 - M_X| + \ldots + |x_n - M_X|}{n}. \qquad (5.3)$$

В правой части этого неравенства стоит среднее арифметическое чисел $y_1 = |x_1 - M_X|, \ldots, y_n = |x_n - M_X|$. Для его оценки нам понадобится классическое неравенство о среднем арифметическом и среднем квадратичном.

Лемма. *Если* $M_X = \frac{x_1 + \ldots + x_n}{n}$ *– среднее арифметическое, а* $K_X = \sqrt{\frac{x_1^2 + \ldots + x_n^2}{n}}$ *– среднее квадратичное чисел* $x_1, \ldots, x_n$, *то* $|M_X| \le K_X$, *причём знак равенства в этом неравенстве достигается тогда и только тогда, когда все числа* $x_1, \ldots, x_n$ *равны между собой.*

Чтобы доказать Лемму, рассмотрим дисперсию набора $X = [x_1, \ldots, x_n]$. После раскрытия скобок и перегруппировки членов, мы можем переписать формулу, которая определяет дисперсию в виде:

$$D_X = \frac{x_1^2 + \ldots + x_n^2 - 2M_X\left(x_1 + \ldots + x_n\right) + nM_X^2}{n} = \frac{x_1^2 + \ldots + x_n^2 - 2nM_X \cdot M_X + nM_X^2}{n}$$

$$= \frac{x_1^2 + \ldots + x_n^2 - nM_X^2}{n} = K_X^2 - M_X^2.$$

Поскольку дисперсия неотрицательна, можно утверждать, что $K_x^2 \geq M_x^2$. Извлекая из обеих частей квадратные корни и используя неотрицательность среднего квадратичного мы получим требуемое неравенство. Равенство $|M_X| = K_X$ равносильно равенству 0 дисперсии, что, в свою очередь, равносильно тому, что все числа $x_1,\ldots,x_n$ равны своему среднему, т.е. равны между собой.

Теперь, чтобы закончить доказательство неравенства (5.2), достаточно применить к правой части неравенства (5.3) неравенство о среднем арифметическом и среднем квадратичном: среднее квадратичное чисел $y_1 = |x_1 - M_X|,\ldots,y_n = |x_n - M_X|$ – это в точности среднее квадратичное отклонение σ_X. □

Замечание 5.2. Полученная в ходе доказательства Леммы формула $D_X = K_X^2 - M_X^2$ интересна и сама по себе.

Замечание 5.3. Если $M_X \neq 0$, то неравенство (5.2) можно записать в виде: $\left|\frac{\mu_X}{M_X} - 1\right| \leq \frac{\sigma_X}{|M_X|}$, где $\frac{\sigma_X}{|M_X|}$ – введённый ранее коэффициент вариации.

Замечание 5.4. Рассматриваемое для *произвольного* числового набора, неравенство (5.2) неулучшаемо в том смысле, что существует набор (из одинаковых чисел), для которого это неравенство выполняется в виде равенства.

Более интересна другая интерпретация термина «неулучшаемо»: для любой положительной константы $C < 1$ существует набор, для которого справедливо неравенство $|\mu_X - M_X| > C \cdot \sigma_X$. Действительно, если $C \in (0;1)$ задано, то возьмём число k настолько большим, что дробь $\frac{1}{k+1}$ меньше, чем $1 - C^2$ (это число, очевидно, положительно), и рассмотрим набор $X = \left[\underbrace{0,\ldots,0}_{k \text{ раз}},1,\underbrace{1,\ldots,1}_{k \text{ раз}}\right]$ из $n = 2k+1$ чисел. Его медиана μ_X равна 1, среднее значение M_X равно $\frac{k+1}{2k+1}$, а дисперсия D_X равна $\frac{k(k+1)}{(2k+1)^2}$. Для этого набора неравенство $|\mu_X - M_X| > C \cdot \sigma_X$ примет вид: $\frac{1}{k+1} < 1 - C^2$, т.е. будет истинно (в силу выбора числа k).

6. Квартили

6.1 Определение квартилей

В разделе 3 мы рассказали о различных мерах положения числового набора (среднем значении, медиане, моде) и провели их сравнительный анализ. В связи с этим мы упомянули, что более точно охарактеризовать положение и разброс чисел набора можно с помощью так называемых квартилей. В этом разделе мы подробно расскажем о квартилях и связанных с ними понятиях. Всюду в этом разделе мы будем считать, что рассматриваемые наборы упорядочены по возрастанию (т.е. мы будем рассматривать только вариационные ряды).

Начнём с того, что кратко повторим определение медианы числового набора. Неформально медиана μ_X числового набора $X=(x_1,\ldots,x_n)$, *упорядоченного по возрастанию*, определяется как такое число, слева и справа от которого лежит одно и то же количество чисел набора. Точное определение длиннее:

- если набор состоит из нечётного количества чисел $n=2k-1$, то его медиана – это число x_k с номером k;
- если же набор состоит из чётного количества чисел $n=2k$, то его медиана – это число $\frac{x_k+x_{k+1}}{2}$, лежащее посредине отрезка $[x_k;x_{k+1}]$.

Если считать, что среднее арифметическое чисел x_k и x_{k+1} (т.е. k-го и (k+1)-го членов набора) – это число $x_{k+\frac{1}{2}}$ с «номером» $\left(k+\frac{1}{2}\right)$, то оба случая можно объединить в один, сказав, что медиана – это $\frac{n+1}{2}$–е в порядке возрастания число основного набора.

Значение медианы возрастает, когда её рассматривают как вторую квартиль Q_2 и в дополнение к ней вычисляют первую квартиль Q_1 (её обычно называют нижней) и третью квартиль Q_3 (её обычно называют верхней). Неформально говоря, квартили Q_1,Q_2,Q_3 делят исходный упорядоченный набор на 4 (примерно) равные части. Иначе говоря, нижняя квартиль – это медиана первой половины исходного набора, а верхняя квартиль – медиана второй половины исходного набора. Уточнить это неформальное определение нижней и верхней квартилей можно несколькими способами, которые обычно приводят к разным результатам (хотя и не очень сильно отличающимся) – общепринятого определения квартилей в описательной статистике нет (подробнее по этому поводу см., например, [14,15]). Скажем, в британском школьном учебнике [8], подготовленным экзаменационным центром OCR (Oxford, Cambridge and Royal

Society of Arts), принято следующее определение (именно это определение мы будем использовать дальше).

Определение 1. *Чтобы найти квартили числового набора,*

1. *Нужно упорядочить числа исходного набора по возрастанию. Если некоторые числа набора повторяются, то они стоят одной группой, т.е. учитываются в наборе нужное количество раз.*
2. *Если набор содержит чётное количество чисел, то нужно разделить эту упорядоченную версию исходного набора на две равные (по числу элементов) половины. Медиана первой половины – это нижняя квартиль, а медиана второй половины – верхняя.*
3. *Если набор содержит нечётное количество чисел, то нужно найти медиану и вычеркнуть её из набора (так что останется чётное количество чисел). После этого нужно оставшиеся числа разделить на две равные (по числу элементов) половины. Медиана первой половины – это нижняя квартиль, а медиана второй половины – верхняя.*

Рассмотрим, например, набор 11, 11, 15, 15, 15, 15, 18, 20, 20, 160. Он состоит из 10 чисел, которые уже упорядочены по возрастанию. Первая половина – это набор 11, 11, 15, 15, 15. Он состоит из 5 чисел. Поэтому его медиана – третье по счёту число, т.е. 15 – это и будет нижняя квартиль Q_1. Вторая половина – это набор 15, 18, 20, 20, 160. Он также состоит из 5 чисел. Поэтому его медиана – третье по счёту число, т.е. 20 – это и будет верхняя квартиль Q_3. Вторая квартиль Q_2 – это медиана исходного набора, т.е. среднее арифметическое пятого и шестого чисел исходного набора: $Q_2 \equiv \mu_X = \frac{15+15}{2} = 15$.

Возьмём теперь набор, содержащий нечётное количество чисел, например, 2, 7, 6, 2, 11, 8, 9, 4, 3 (n=9). После упорядочивания по величине мы получим набор: 2, 2, 3, 4, 6, 7, 8, 9, 11. Медиана этого набора (она будет второй квартилью Q_2) – это пятое по порядку число, т.е. 6. Вычёркивая медиану, мы получим набор из 8 чисел: 2, 2, 3, 4, 7, 8, 9, 11. Первая половина – это набор 2, 2, 3, 4. Он состоит из 4 чисел. Поэтому его медиана – среднее арифметическое второго и третьего чисел, т.е. 2,5 – это и будет нижняя квартиль Q_1. Вторая половина – это набор 7, 8, 9, 11. Он также состоит из 4 чисел. Поэтому его медиана – среднее арифметическое второго и третьего чисел, т.е. 8,5 – это и будет верхняя квартиль Q_3.

Обратим внимание на следующее обстоятельство. В рассмотренном только что примере ниже нижней квартили $Q_1 = 2{,}5$ находится два числа из рассматриваемого набора: $x_1 = 2$ и $x_2 = 2$. Эти два числа составляют $\frac{2}{9} \approx 22\%$ от общего количества чисел набора. Выше верхней квартили $Q_3 = 8{,}5$ находится тоже два числа из рассматриваемого набора: $x_8 = 9$ и $x_9 = 11$. Эти два числа

составляют $\frac{2}{9}\approx 22\%$ от общего количества чисел набора. «Центральная» часть набора, которая состоит из чисел, лежащих между нижней и верхней квартилями, содержит 5 чисел, что составляет $\frac{5}{9}\approx 56\%$ от общего количества чисел набора. Поэтому фразу «квартили делят набор на 4 равные части» нельзя понимать буквально.

Ещё один распространённый способ определения квартилей связан с определением медианы как $\frac{n+1}{2}$-го в порядке возрастания числа основного набора. В соответствии с этим определением (назовём его **Определение 2**),

i-я квартиль Q_i *– это* $\left(i\cdot\frac{n-1}{4}+1\right)$*-е в порядке возрастания число основного набора. Иначе говоря, нижняя квартиль* Q_1 *– это* $\frac{n+3}{4}$*-е в порядке возрастания число основного набора, медиана* Q_2 *– это (как и следовало ожидать)* $\frac{n+1}{2}$*-е в порядке возрастания число основного набора, а верхняя квартиль* Q_3 *– это* $\frac{3n+1}{4}$*-е в порядке возрастания число основного набора:* $Q_1=x_{\frac{n+3}{4}}$, $Q_2=x_{\frac{n+1}{2}}$, $Q_3=x_{\frac{3n+1}{4}}$. *При этом, по определению, если «номер»* $\frac{n+3}{4}$, $\frac{n+1}{2}$ *или* $\frac{3n+1}{4}$ *имеет вид*

- $k+\frac{1}{2}$, *где* k *– некоторое натуральное число, то* $x_{k+\frac{1}{2}}=\frac{1}{2}x_k+\frac{1}{2}x_{k+1}$;
- $k+\frac{1}{4}$, *где* k *– некоторое натуральное число, то* $x_{k+\frac{1}{4}}=\frac{3}{4}x_k+\frac{1}{4}x_{k+1}$;
- $k+\frac{3}{4}$, *где* k *– некоторое натуральное число, то* $x_{k+\frac{3}{4}}=\frac{1}{4}x_k+\frac{3}{4}x_{k+1}$.

Этот способ определения значений для дробных «номеров», когда значение x_{k+t} (число k – натуральное, а число t лежит между 0 и 1) определяется по формуле $x_{k+t}=(1-t)x_k+tx_{k+1}$, называется линейной интерполяцией. Отметим, что в рассматриваемой ситуации число k называется целой частью числа $k+t$, а число t – дробной.

Применим это новое определение квартилей к рассмотренным ранее наборам.

Набор 11, 11, 15, 15, 15, 15, 18, 20, 20, 160 состоит из n=10 чисел. Его нижняя квартиль Q_1 равна $Q_1=x_{\frac{13}{4}}=x_{3+\frac{1}{4}}=\frac{3}{4}x_3+\frac{1}{4}x_4=\frac{3}{4}\cdot 15+\frac{3}{4}\cdot 15=15$ (напомним, что в

соответствии с первым определением мы получили то же значение нижней квартили). Верхняя квартиль Q_3 равна $Q_3 = x_{\frac{31}{4}} = x_{7+\frac{3}{4}} = \frac{1}{4}x_7 + \frac{3}{4}x_8 = \frac{1}{4}\cdot 18 + \frac{3}{4}\cdot 20 = 19{,}5$. Хотя первое определение дало для верхней квартили значение 20, отличие не очень большое.

Набор 2, 2, 3, 4, 6, 7, 8, 9, 11 состоит из n=9 чисел. Нижняя квартиль Q_1 равна $Q_1 = x_{\frac{12}{4}} = x_3 = 3$ (определение 1 дало для нижней квартили значение 2,5). Верхняя квартиль Q_3 равна $Q_3 = x_{\frac{28}{4}} = x_7 = 8$ (определение 1 дало для верхней квартили значение 8,5). Как и в первом примере, оба определения дают для квартилей близкие значения.

Термин «квартиль» (англ.: quartile) происходит от латинского слова quartus (в средневековой латыни: quartilis) – «четвёртый». От этой латинской основы произошли и слова: «квартал» – четвёртая часть года (три месяца), «квартет» – музыкальный ансамбль из четырёх исполнителей. В статистике термин «квартиль» появился в 1879 г. и обычно связывается с именем английского учёного Ф.Гальтона (F.Galton, 1822–1911).

6.2 Применение компьютеров для вычисления квартилей

Второе определение квартилей реализовано в электронных таблицах Microsoft Office Excel, где для подсчёта квартилей можно использовать стандартную функцию КВАРТИЛЬ.

Вернемся к ранее рассмотренному набору чисел 11, 11, 15, 15, 15, 15, 18, 20, 20, 160. Введём эти данные в ячейки с адресами A1, A2, …, A10.

Чтобы найти первую квартиль введём в ячейку A11 формулу =КВАРТИЛЬ(A1:A10,1) (в зависимости от версии Excel для разделения параметров в функциях используется или запятая, или точка с запятой.) Значение функции (т.е. первой квартили) равно 15.

Чтобы найти вторую квартиль (медиану), введём в ячейку A12 формулу =КВАРТИЛЬ(A1:A10,2). Значение второй квартили на этом наборе также равно 15.

Для нахождения третьей квартили введём в ячейку A13 формулу =КВАРТИЛЬ(A1:A10,3). Её значение равно 19,5.

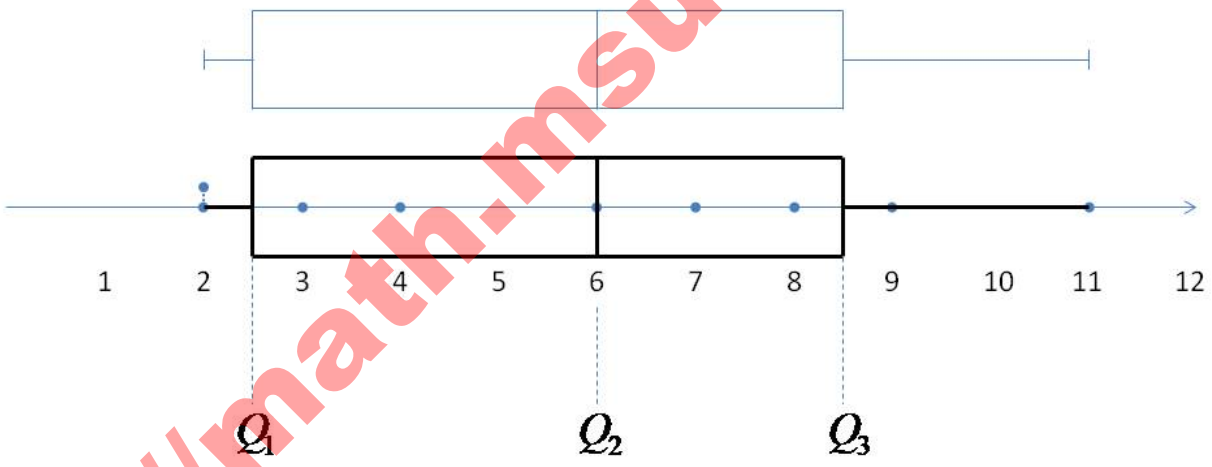
Отметим, что для функции КВАРТИЛЬ совершенно неважно, упорядочены или нет числа набора. Кроме того, эта функция позволяет найти наибольшее и наименьшее числа набора: формула КВАРТИЛЬ(A1:A10,0) даст наименьшее число набора, а формула КВАРТИЛЬ(A1:A10,4) – наибольшее.

6.3 Квартильная диаграмма

Обычно квартили изображают графически с помощью *квартильной диаграммы* (английский термин «box and whisker plot», «box and whisker diagram»; буквально – «ящик с усами»).

Чтобы нарисовать квартильную диаграмму, на числовой прямой нужно отметить квартили и нарисовать два смежных прямоугольника («ящика») одинаковой высоты или, что то же самое, нарисовать прямоугольник на отрезке $[Q_1;Q_3]$ и разделить его на два прямоугольника вертикальным отрезком, проходящим через медиану (см. рис.6.1). Высоту прямоугольника можно взять любой, хотя есть варианты квартильных диаграмм, когда высота несёт определённую смысловую нагрузку [20].

Рис.6.1



Кроме того, на числовой оси нужно отметить экстремальные значения исходного числового набора (т.е. наибольшее и наименьшее числа) и соединить их горизонтальными отрезками («усами») с серединами соответствующих вертикальных отрезков (проведённых через нижнюю и верхнюю квартили); концы этих отрезков обычно отмечают небольшими вертикальными отрезками.

Ситуация, изображённая на рис.6.1, соответствует набору 2, 7, 6, 2, 11, 8, 9, 4, 3, рассмотренному выше. Как было показано, в соответствии с Определением 1 (которое мы используем как основное) для этого набора $Q_1 = 2{,}5$, $Q_2 = 6$, $Q_3 = 8{,}5$.

Для большей наглядности мы отметили и числа исходного набора. Обратим внимание на число 2. Оно появилось в рассматриваемом наборе два раза. Мы отразили это обстоятельство двумя точками – одна из них стоит на числовой оси, а вторая чуть выше её.

Часто получившуюся конфигурацию приподнимают над осью абсцисс (мы сделали это на рис.6.1) – это удобно для сравнения разных наборов, а иногда размещают вертикально.

Таким образом, квартильная диаграмма в наглядной форме показывает положение чисел набора на числовой оси:

- медиана указывает на «среднее» значение набора;
- ширина «ящиков» показывает разброс (примерно) 50% центральных, наиболее характерных значений;
- длина «усов» показывает насколько сильно выброшены в сторону (примерно) 25% наименьших и (примерно) 25% наибольших чисел набора;
- сравнивая между собой ширину левого и правого «ящиков», а также длины левого и правого «усов», можно судить о том, насколько несимметричным является рассматриваемый набор.

Графическое описание набора данных с помощью квартильной диаграммы было предложено в 1970 г. известным американским статистиком, профессором университета Принстона, Джоном Тьюки (John Tukey, 1915–2000). Этот метод стал общепринятым после опубликования в 1977 г. его книги [11].

6.4 Интерквартильный размах

С помощью верхней и нижней квартилей определяют важную меру рассеивания набора чисел – *интерквартильный размах* (соответствующий английский термин *interquartile range* переводят и как *межквартильный размах*). По определению, интерквартильный размах (его обычно обозначают *IQR*) – это разность $Q_3 - Q_1$: $IQR = Q_3 - Q_1$. Интерквартильный размах показывает, насколько разбросаны 50% «центральных» значений рассматриваемого набора чисел. Это понятие было введено в 1882 г. Ф.Гальтоном.

6.5 Выбросы

Выбросы – это числа, которые сильно отличаются от остальных чисел набора и в ситуации, которую описывает рассматриваемый набор чисел, являются необычными. Рассмотрим, например, следующий набор из 10 чисел: 63, 68, 62, 59, 64, 62, 67, 65, 94, 64 (это могут быть оценки ЕГЭ по математике для группы школьников). На рис. 6.2 эти числа изображены на числовой оси. Видно, что 9

оценок из 10 стоят плотной группой, а одна оценка, 94, стоит далеко справа. Если бы мы сравнивали эту группу школьников с другой, в которой математика преподаётся по другой методике (например, чтобы понять, какая из двух методик лучше), то эту оценку было бы неразумно принимать в расчёт, т.к. столь высокая оценка, видимо, связана с математической одарённостью ученика и мало зависит от методики преподавания математики в школе. В рассматриваемой ситуации число 94 и будет выбросом.

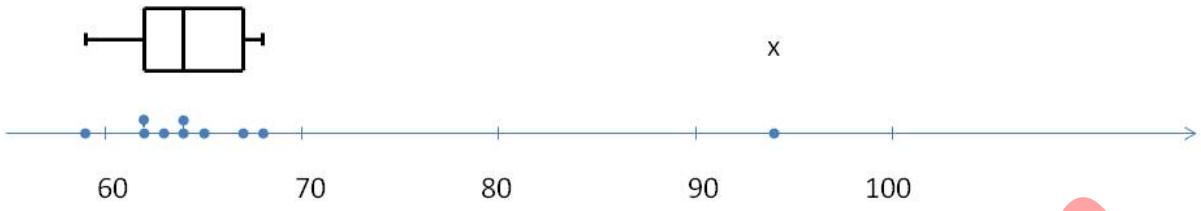
Однозначного ответа на вопрос, насколько далеко должно лежать число от основной массы значений набора, чтобы его можно было считать выбросом, нет. Джон Тьюки [11] предложил следующий подход, который является общепринятым в описательной статистике:

- числа набора, лежащие между нижней и верхней квартилями (т.е. в «ящике» квартильной диаграммы), являются наиболее характерными;
- числа, которые отклоняются от нижней и верхней квартилей не больше, чем на полтора интерквартильных размаха, т.е. удовлетворяют неравенству $f_* \le x < Q_1$, где $f_* = Q_1 - 1{,}5 \cdot IQR$, или неравенству $Q_3 < x \le f^*$, где $f^* = Q_3 + 1{,}5 \cdot IQR$, не столь характерны, но должны учитываться как значимые, т.к. их отклонения от типичных значений не очень большие;
- те числа, которые отклоняются от нижней и верхней квартилей больше, чем на полтора интерквартильных размаха, т.е. удовлетворяют неравенству $x < f_*$ или неравенству $x > f^*$, следует считать выбросами.

Чтобы наглядно отразить выбросы, квартильную диаграмму немного модифицируют, именно, если есть выбросы, то рисуют «усы» не до минимального и максимального чисел набора, а лишь до чисел набора, наиболее удалённых от соответствующих квартилей, но не дальше, чем на $1{,}5 \cdot IQR$. Таким образом, «усы» не переходят через барьеры f_* и $f*$. Числа набора, которые не попадают на эту модифицированную квартильную диаграмму (т.е. выбросы), отмечают особо. Часто для этого используют не точки, а маленькие крестики.

Для рассматриваемого нами иллюстративного набора такая модифицированная диаграмма изображена на рис.6.2 (в нашем случае медиана равна 64, нижняя квартиль равна 62, верхняя квартиль равна 67, интерквартильный размах равен 5, имеется только один выброс, равный 94).

Рис.6.2



Если в анализируемом наборе есть выбросы, необходимо разобраться, почему они появились. Иногда выбросы являются следствием ошибок при сборе данных. В этом случае их нужно исправить или, если это невозможно, исключить выбросы из набора. Но чаще выбросы – это верные значения. В этом случае нужно особенно тщательно проанализировать причину их появления. Обычно это позволяет получить важные выводы о реальной ситуации, которую описывает анализируемый набор данных. В частности, необходимо понять, могут ли подобные экстремальные значения появиться в других подобных ситуациях.

Имея в виду проведённые выше рассуждения, уместно сделать несколько замечаний по поводу интерквартильного размаха:

- Важное достоинство интерквартильного размаха заключается в том, что его значение не зависит от выбросов.
- С другой стороны, оставшиеся 50% чисел набора игнорируются – это, конечно, недостаток этой меры рассеивания.
- Поэтому применять интерквартильный размах имеет смысл в тех случаях, когда выбросы не являются типичными значениями, характерными для ситуации, описываемой набором, который мы изучаем, и их следует игнорировать. Вообще, важно понимать, что математика – это только аппарат для исследования реальной ситуации. Если не учитывать природу изучаемого объекта, то формальные математические вычисления могут привести нас не к самым разумным выводам.

6.6 Асимметрия набора

Как мы уже отмечали, с помощью квартилей можно понять, насколько несимметричным является рассматриваемый набор.

Если медиана находится точно посередине между нижней и верхней квартилью (т.е. левый и правый «ящики» на квартильной диаграмме имеют одинаковые основания), то из (примерно) 50% «центральных» значений

рассматриваемого набора чисел примерно половина лежит в левом «ящике», а вторая половина – в равном ему по размеру правом. Иначе говоря, эти «центральные» значения в целом расположены вокруг медианы симметрично.

Если нижняя квартиль находится от медианы дальше, чем верхняя (т.е. левый «ящик» на квартильной диаграмме больше правого), то левая половина «центральных» значений рассматриваемого набора чисел разбросана больше, чем правая. Иначе говоря, эти «центральные» значения в целом расположены вокруг медианы несимметрично, со скосом в левую сторону. Такой набор называют отрицательно асимметричным.

Если же верхняя квартиль находится от медианы дальше, чем нижняя (т.е. правый «ящик» на квартильной диаграмме больше левого), то правая половина «центральных» значений рассматриваемого набора чисел разбросана больше, чем левая. Иначе говоря, эти «центральные» значения в целом расположены вокруг медианы несимметрично, со скосом в правую сторону. Такой набор называют положительно асимметричным.

Асимметрию набора можно описать и количественно. В статистике для этого введено несколько разных величин (примерно так же, как положение набора можно описывать средним значением, медианой и модой). Например, можно использовать отношение $\lambda = \frac{Q_2 - Q_1}{Q_3 - Q_2}$ (при $Q_3 \neq Q_2$). Говорят, что точка Q_2 делит отрезок $[Q_1; Q_3]$ в отношении λ. Число λ показывает, во сколько раз левый «ящик» квартильной диаграммы шире правого:

- если λ=1, то эти ящики имеют одинаковую ширину и распределение является симметричным в смысле данного выше определения;
- если λ >1, то «левый» ящик шире правого и распределение является отрицательно асимметричным в смысле данного выше определения;
- если λ <1, то «правый» ящик шире левого и распределение является положительно асимметричным в смысле данного выше определения.

Единственный недостаток этого, в общем-то естественного определения, заключается в том, что для симметричного распределения, т.е. с нулевой асимметрией, коэффициент равен 1, для положительно асимметричного распределения коэффициент меньше 1, для отрицательно асимметричного распределения коэффициент больше 1. Хотелось бы ввести такую меру, чтобы для симметричного распределения, т.е. с нулевой асимметрией, она была бы равна 0, для положительно асимметричного распределения была бы больше 0 (т.е. положительна), для отрицательно асимметричного распределения была бы меньше 0 (т.е. отрицательна). Имея это в виду, в качестве меры асимметрии рассматривают число $k = \frac{1-\lambda}{1+\lambda}$. Его обычно называют *квартильный коэффициент* или *асимметрия Баули* (A.L.Bowley, 1869–1957 – английский статистик и экономист). Через квартили квартильный коэффициент выражается следующим образом:

$$k = \frac{Q_3 + Q_1 - 2Q_2}{Q_3 - Q_1} \equiv \frac{\text{верхняя квартиль} + \text{нижняя квартиль} - 2 \times \text{медиана}}{\text{верхняя квартиль} - \text{нижняя квартиль}}.$$

Квартильный коэффициент определён только для наборов, у которых верхняя квартиль не совпадает с нижней, т.е. «центральные» 50% значений набора разбросаны по отрезку $[Q_3; Q_1]$ ненулевой длины.

Возможные значения квартильного коэффициента лежат на отрезке $[-1; +1]$. Действительно, двойное неравенство $-1 \le \frac{Q_3 + Q_1 - 2Q_2}{Q_3 - Q_1} \le 1$ равносильно двум неравенствам: $-Q_3 + Q_1 \le Q_3 + Q_1 - 2Q_2$ и $Q_3 + Q_1 - 2Q_2 \le Q_3 - Q_1$, которые после приведения подобных членов сводятся к двойному неравенству $Q_1 \le Q_2 \le Q_3$, которое, очевидно, истинно.

Если квартильный коэффициент равен 0, то это означает, что $Q_3 + Q_1 - 2Q_2 = 0$, т.е. $Q_3 - Q_2 = Q_2 - Q_1$. Но $Q_3 - Q_2$ – это ширина правого ящика, а $Q_2 - Q_1$ – левого. Поэтому равенство $Q_3 - Q_2 = Q_2 - Q_1$ означает симметрию набора в смысле определения, данного в начале этого пункта.

При изменении параметра λ от 0 до $+\infty$ функция $k(\lambda) = \frac{1-\lambda}{1+\lambda}$ монотонно убывает от 1 до -1. Таким образом, значение квартильного коэффициента, близкое к -1, равносильно тому, что значение λ велико, что, в свою очередь, означает, что левый «ящик» много шире правого, т.е. большую отрицательную асимметрию. Значение квартильного коэффициента, близкое к $+1$, равносильно тому, что значение λ очень мало, что, в свою очередь, означает, что правый «ящик» много шире левого, т.е. большую положительную асимметрию.

Особо подчеркнём, что при обсуждении симметрии или асимметрии числового набора мы принимали в расчёт только (примерно) 50% «центральных», наиболее типичных чисел, игнорируя числа, которые лежат ниже нижней квартили или выше верхней. В статистике есть и другие меры асимметрии, которые используют все числа набора. Наиболее простым из них (он использует только основные меры положения и рассеивания) является коэффициент асимметрии Пирсона:

$$3\frac{M_X - \mu_X}{\sigma_X} = 3\frac{\text{среднее значение} - \text{медиана}}{\text{стандартное отклонение}}.$$

Он равен нулю для наборов, у которых среднее значение совпадает с медианой, и только для них. Таким образом, этот подход считает характерным свойством симметричных наборов равенство среднего значения и медианы. В силу Теоремы 5.4 коэффициент асимметрии Пирсона меняется в пределах от -3 до $+3$.

7. Древовидная диаграмма

7.1 Введение

Напомним, что основным объектом статистики являются наборы данных (обычно, числовых). Эти наборы

- либо берут из существующих источников, например, из газет или интернета – такие данные называют вторичными,
- либо получают в результате опроса с помощью какого-нибудь вопросника/анкеты или в результате эксперимента – такие данные называют первичными.

Первичные данные обычно не имеют сколь-нибудь явно выраженной структуры и производят впечатление хаотического набора чисел.

Статистический анализ начинается с того, что эти данные представляют в каком-нибудь организованном виде. Например, упорядочивают по возрастанию, разбивают на группы и т.д. Это служит основой для дальнейшего статистического анализа, в ходе которого вычисляют статистические характеристики исследуемого набора (среднее значение, медиану, дисперсию и т.д.), представляют данные в графическом виде (например, рисуют столбиковую диаграмму, что позволяет визуально выявить характерные особенности распределения чисел набора и т.д.

В этом разделе мы расскажем об удобном способе представления небольших (содержащих не более сотни чисел) наборов первичных данных с помощью так называемой *древовидной диаграммы*. В статистической литературе на английском языке древовидную диаграмму называют stem-and-leaf diagram. Буквально этот термин можно было бы перевести как «диаграмма из стеблей с листьями». В информатике при представлении данных методом, используемым в этих диаграммах, говорят о *дереве*. По этой причине мы решили использовать термин *древовидная диаграмма*. Следует, однако, подчеркнуть, что общепринятого термина для этого объекта в отечественной статистической литературе нет.

Древовидная диаграмма вошла в число стандартных инструментов описательной статистики после опубликования известным американским статистиком, профессором университета Принстона, Джоном Тьюки (John Tukey, 1915–2000) в 1977 книги [11]. Их большая популярность в 70-е—80-е годы прошлого века была связана с тем, что в то время большинство принтеров для компьютеров работали как печатные машинки, т.е. могли печатать лишь строки из цифр, букв и символов. Для рисования графиков и чертежей использовались специальные дорогостоящие устройства – плоттеры, которые работали под управлением мощных (по меркам того времени) и дорогих компьютеров. Древовидная диаграмма похожа на таблицу со строками переменной длины и легко может быть напечатана даже на пишущей машинке.

В сущности, для получения древовидной диаграммы не нужно никаких технических средств. Она позволяет быстро, в любой обстановке, с помощью ручки и листа бумаги представить небольшие наборы данных в «графическом» виде и провести их простейший статистический анализ – с этой точки зрения древовидная диаграмма идеально подходят для обучения элементам описательной статистики в классе.

Умение представить исходную информацию в удобном для решения задачи виде относится к общешкольным умениям. Об этом же говорится в Федеральном государственном образовательном стандарте основного общего образования (от 17 декабря 2010г); в предметные результаты освоения основной образовательной программы основного общего образования (раздел «Математика. Алгебра. Геометрия. Информатика») входит «… 13) формирование умений формализации и структурирования информации, умения выбирать способ представления данных в соответствии с поставленной задачей — таблицы, схемы, графики, диаграммы, с использованием соответствующих программных средств обработки данных».

7.2 Построение древовидной диаграммы

Принцип построения древовидной диаграммы мы продемонстрируем на примере следующего набора чисел: 9, 43, 34, 5, 28, 16, 18, 23, 45, 52, 23, 38, 25, 29, 31, 16, 27, 30, 22, 36, 12, 5, 17, 25, 31. Начнём с того, что создадим таблицу следующего вида:

		числа с данным количеством десятков
десятки	0	
	1	
	2	
	3	
	4	
	5	

Затем станем последовательно просматривать числа нашего набора и вписывать очередное число в соответствующую строку. Беглый анализ числового ряда показал, что все числа в наборе меньше 60. Это и определило количество строк в таблице. В результате мы получим таблицу:

		числа с данным количеством десятков
десятки	**0**	9, 5, 5
	1	16, 18, 16, 12, 17
	2	28, 23, 23, 25, 29, 27, 22, 25
	3	34, 38, 31, 30, 36, 31
	4	43, 45
	5	52

Посмотрим повнимательнее на какую-нибудь строку этой таблицы, например,

1	16, 18, 16, 12, 17

Все числа в этой строке начинаются с цифры 1, причём этот факт явно указан в первом столбце. Поэтому в строке чисел 16, 18, 16, 12, 17 можно указывать только число единиц, т.е. записать строку

1	16, 18, 16, 12, 17

в виде:

1	6, 8, 6, 2, 7

Соответственно, вся таблица с числами рассматриваемого набора может быть записана в виде:

		единицы
десятки	**0**	9, 5, 5
	1	6, 8, 6, 2, 7
	2	8, 3, 3, 5, 9, 7, 2, 5
	3	4, 8, 1, 0, 6, 1
	4	3, 5
	5	2

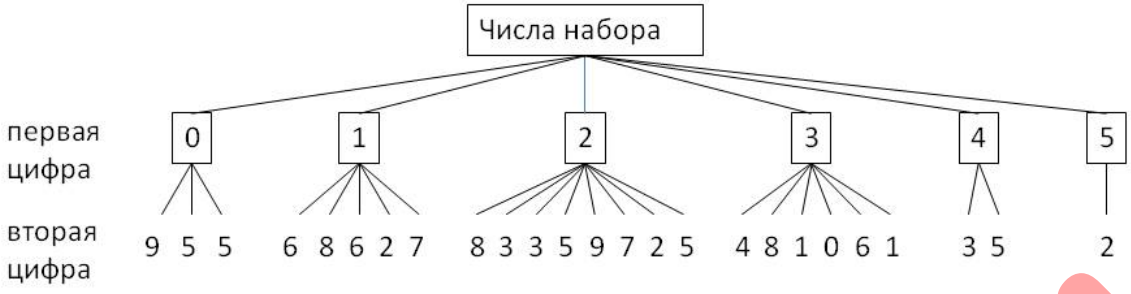
Обычно эту таблицу оформляют так:

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

Эту таблицу и назовем *древовидной диаграммой*. Дополнительно в ней обязательно указывают, что, например, запись **5**|2 означает число 52. Правило восстановления чисел анализируемого набора, представленных древовидной диаграммой, называется *ключом*. В дальнейшем будем писать «ключ: **5**|2 = 52». В принципе в ключе может быть указано любое число, не обязательно входящее в набор.

В информатике описанный выше способ структурирования данных представляют в виде *дерева*, изображённого на рисунке 7.1.

Рис.7.1



В древовидной диаграмме элементы столбца слева от вертикальной линии (в рассматриваемом примере – это цифры десятков), называются *узлами*, а цифры в строках справа от вертикальной линии (в рассматриваемом примере – это цифры единиц), называются *листьями*. Термины «узел» и «лист» используются здесь в том же смысле, что и в теории структур данных в информатике («лист» – это узел дерева самого нижнего уровня или узел, не имеющий потомков). Про листья, которые расположены в той же строке, что и некоторый узел, говорят, что они принадлежат этому узлу (находятся в этом узле). Например, узлу 4 принадлежат листья 3 и 5. Можно также сказать, что в узле 4 расположены листья 3 и 5.

В англоязычной статистической литературе часто термином «stem» называют весь столбец чисел слева от вертикальной линии (т.е., в нашей терминологии, весь набор узлов). В этом контексте слово «stem» можно перевести как «стебель» (в английском языке слово «stem» означает длинную, тонкую основную часть чего-либо).

Листья, принадлежащие одному узлу, упорядочивают (это несложно сделать, т.к. последовательность цифр-листьев обычно легко охватить взглядом), что даёт упорядоченную древовидную диаграмму:

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

Ключ: **5** | 2 = 52

Обычно, когда говорят о древовидной диаграмме, имеют в виду именно подобную упорядоченную версию.

Кроме того, поскольку все листья являются цифрами, можно опустить запятые (или любые другие разделители):

0	559
1	26678
2	23355789
3	011468
4	35
5	2

Ключ: **5** | 2 = 52

Это и будет окончательная версия древовидной диаграммы. Иногда, впрочем, в дополнительном столбце справа в круглых скобках указывают количество листьев в соответствующем узле (т.е. чисел исходного набора с заданной цифрой десятков):

0	559	(3)
1	26678	(5)
2	23355789	(8)
3	011468	(6)
4	35	(2)
5	2	(1)

Ключ: **5** | 2 = 52

В построенной только что древовидной диаграмме узлы упорядочены по возрастанию сверху вниз. С равным успехом их можно было бы расположить в порядке возрастания снизу вверх. Тогда наша таблица приняла бы вид:

5	2	(1)
4	35	(2)
3	011468	(6)
2	23355789	(8)
1	26678	(5)
0	559	(3)

Ключ: **5** | 2 = 52

Принцип, который лежит в основе построения древовидной диаграммы, широко используется при построении словарей. Посмотрим, например, на следующий список из четырёх английских слов (в скобках дан перевод на

русский язык): book (книга), bookcase (книжный шкаф), bookmark (закладка), bookshop (книжный магазин).

Во всех четырёх словах видна общая основа – **book**. Поэтому в англо-русском словаре все четыре слова переводятся в рамках одной статьи, которая выглядит примерно так (мы приводим слегка сокращённый вариант соответствующей статьи из Collins Pocket Russian Dictionary, HarperCollins Publishers, 2008):

book книга; ~case книжный шкаф; ~mark закладка; ~shop книжный магазин.

Для экономии места *заглавное слово* **book** указывается только один раз, в начале словарной статьи. В словосочетании (например, bookcase) заглавное слово заменяется знаком ~ (тильда). Таким образом, запись «~case» означает «bookcase».

7.3 Связь древовидной диаграммы и гистограммы

Конструирование древовидной диаграммы фактически заключается в группировании данных в соответствии с числом десятков, т.е. в промежутки $0 \le x \le 9$, $10 \le x \le 19$, …

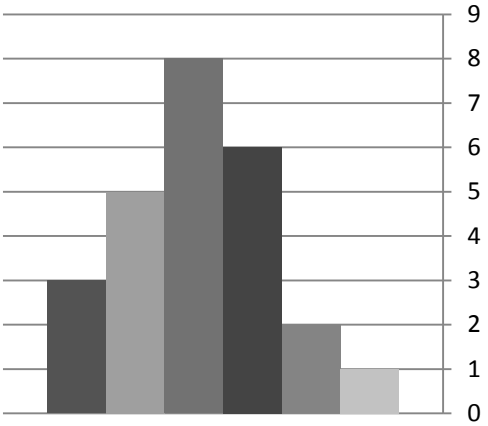
Если листья из одного узла заключить в прямоугольник, то мы получим следующую картинку:

0	559
1	26678
2	23355789
3	011468
4	35
5	2

Нетрудно видеть, что по сути дела на ней изображена повёрнутая на 90° (по часовой стрелке) гистограмма (если анализируется набор значений непрерывной переменной). Это сходство ещё больше усилится, если узлы расположить в горизонтальной строке, а листья – в вертикальных столбцах, т.е. повернуть древовидную диаграмму на 90° против часовой стрелки.

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

Для сравнения ниже приведена гистограмма для рассматриваемого набора чисел, построенная с помощью MS Excel:



Хорошо видно, что визуально гистограмма даёт то же представление о характере распределения чисел в исходном наборе, что и древовидная диаграмма. Но важное преимущество древовидной диаграммы по сравнению с гистограммой заключается в том, что она явно указывает все индивидуальные значения анализируемого числового набора.

Древовидную диаграмму обычно создают, если анализируемый набор содержит несколько десятков чисел. Если же чисел меньше, то древовидная диаграмма не имеет смысла, т.к. выявить какие-то особенности распределения чисел в маленьком наборе нельзя. С другой стороны, если количество членов набора превышает сотню, то древовидная диаграмма становится очень громоздкой и трудно обозримой (ведь мы должны указывать каждое индивидуальное значение) – в этом случае лучше использовать гистограмму.

7.4 Дополнительные примеры

В рассмотренном выше примере мы имели дело с набором двузначных чисел. Но описанную методику представления данных можно применить и к более сложным ситуациям.

Пример 7.1. Рассмотрим набор 1,52; 1,43; 1,57; 1,62; 1,40; 1,49; 1,61; 1,53; 1,64; 1,59; 1,61; 1,44; 1,48; 1,57, 1,48 (числа набора могут представлять рост группы школьников в метрах). Этот набор можно представить с помощью следующей древовидной диаграммы:

14	034889	(6)
15	23779	(5)
16	1124	(4)

Ключ: **16** | 1 = 1,61 (м)=161 (см)

7.5 Сдвоенная древовидная диаграмма

Предположим, что в примере 1 речь идёт о росте 15 девочек определённого возраста. Пусть, далее, набор 1,72; 1,63; 1,76; 1,48; 1,67; 1,57; 1,67; 1,71; 1,55; 1,63; 1,59; 1,70; 1,60; 1,51; 1,65 содержит информацию о росте 15 мальчиков того же возраста. Эти данные можно представить с помощью следующей древовидной диаграммы:

14	8	(1)
15	1579	(4)
16	033577	(6)
17	0126	(4)

Ключ: **14** | 8 = 1,48 (м)=148 (см)

Если мы хотим сравнить данные о росте мальчиков с данными о росте девочек, удобно одну из соответствующих древовидных диаграмм, например, только что созданную, отобразить симметрично относительно столбца узлов, т.е. записать в виде:

(1)	8	**14**
(4)	1579	**15**
(6)	033577	**16**
(4)	0126	**17**

Ключ: 8 | **14** = 1,48 (м)=148 (см)

а после этого присоединить к соответствующей диаграмме для второго из двух сравниваемых наборов:

мальчики			девочки	
(1)	8	**14**	034889	(6)
(4)	1579	**15**	23779	(5)
(6)	033577	**16**	1124	(4)
(4)	0126	**17**		

В этом примере ключ можно было бы описать так:
Ключ: 8 | **14** = 1,48 (м)=148 (см) из первого набора , **14** | 3 = 1,43 (м)=143 (см) из второго набора.

Полученная диаграмма и называется *сдвоенной* (английский термин «back-to-back»; буквально – «спина к спине») *древовидной диаграммой*. Из неё хорошо видны различия в распределении чисел двух рассматриваемых наборов; например, ясно, что мальчики в целом выше девочек примерно на 10 см.

7.6 Использование древовидной диаграммы для подсчёта статистических характеристик набора

С помощью древовидной диаграммы можно находить различные статистические характеристики этого набора. Мы продемонстрируем это на примере набора, заданного следующей древовидной диаграммой:

0	559	(3)
1	26678	(5)
2	23355789	(8)
3	011468	(6)
4	35	(2)
5	2	(1)

Ключ: **5** | 2 = 52

7.6.1 Наибольшее и наименьшее значения

Древовидная диаграмма фактически представляет изучаемый числовой набор в упорядоченном (по возрастанию) виде. Поэтому наименьшее и наибольшее число набора находятся мгновенно: наименьшее значение определяется первым листом в первом узле, а наибольшее – последним листом в последнем узле. В рассматриваемом примере наименьшее число определяется кодом 0 | 5 = 5, а наибольшее – кодом 5 | 2 = 52.

7.6.2 Медиана

Вычисление медианы с помощью древовидной диаграммы удобно производить, если для каждого узла указано количество принадлежащих ему листьев. Прежде всего, складывая эти количества листьев в каждом узле, мы найдём общее количество чисел в наборе. В рассматриваемом примере нетрудно подсчитать, что набор содержит $n = 25$ чисел. Поэтому его медиана – это 13-е в порядке возрастания число набора. Складывая последовательно числа в круглых скобках, начиная с верхней строки, мы видим, что первые два узла содержат 8 членов набора, а первые три – уже 16 членов. и определяется кодом 2 | 5, т.е равна 25.

7.6.3 Среднее значение

Подсчёт среднего значения в рассматриваемом примере начнём с определения общего количества чисел в наборе n, которое совпадает с общим количеством листьев. В древовидной диаграмме для каждого узла количество

листьев явно указано в круглых скобках. Суммируя эти числа, мы найдём n: $n = 3+5+8+6+2+1 = 25$.

Сумму $x_1 + \ldots + x_n$ всех членов набора можно найти, подсчитав для каждого узла сумму значений его «листьев», а затем сложить все эти суммы. Например, для узла 1 сумму значений чисел-листьев S_1 удобно находить так:

$$S_1 \equiv 12+16+16+17+18 = (10+2)+(10+6)+(10+6)+(10+7)+(10+8)$$
$$= 10\cdot 5 + (2+6+6+7+8) = 50+29 = 79.$$

Нетрудно сообразить, что для нахождения S_i достаточно значение узла умножить на количество листьев (число, указанное в круглых скобках), а затем к результату прибавить сумму всех листьев, так что:

$$S_0 = 0\cdot 3 + (5+5+9) = 19,$$
$$S_1 = 10\cdot 5 + (2+6+6+7+8) = 79,$$
$$S_2 = 20\cdot 8 + (2+3+3+5+5+7+8+9) = 202,$$
$$S_3 = 30\cdot 6 + (0+1+1+4+6+8) = 200,$$
$$S_4 = 40\cdot 2 + (3+5) = 88,$$
$$S_5 = 50\cdot 1 + (2) = 52.$$

Тогда сумма всех членов набора равна 19+79+202+200+88+52=640, а среднее значение равно $\frac{640}{25} = 25{,}6$.

Медиана этого набора, как мы видели, равна 25 и практически не отличается от среднего значения.

7.7 Задачи

Задача 7.1 В классе учится 32 школьника. В один из дней они потратили в школьном буфете следующие суммы (в рублях): 68; 54; 62; 74; 40; 38; 45; 52; 64; 25; 50; 43; 50; 63; 38; 49; 55; 28; 61; 43; 62; 42; 36; 47; 60; 63; 35; 52; 40; 52; 64;58. Постройте древовидную диаграмму для этого набора и с её помощью вычислите медиану и размах.

Решение. Используя описанную выше процедуру, мы получим следующую древовидную диаграмму:

2	58	(2)
3	5688	(4)
4	00233579	(8)
5	00222458	(8)
6	012233448	(9)
7	4	(1)

Ключ: **2** | 5 = 25 (руб.)

Наименьшее число набора закодировано первым листом в первом узле: min=25. Наибольшее число набора закодировано последним листом в последнем узле: max=74. Поэтому размах равен $74-25=49$.

Пусть x_i – i-е в порядке возрастания число рассматриваемого набора. Поскольку рассматриваемый набор содержит чётное количество элементов $n=32$, медиана μ_X является средним арифметическим 16-го и 17-го в порядке возрастания чисел набора: $\mu_X=\frac{x_{16}+x_{17}}{2}$. Первые три узла содержат 14 листьев (2 + 4 + 8). Поэтому 16-й и 17-й элементы расположены в четвёртом узле (первый лист четвёртого узла соответствует 15-му элементу набора, второй – 16-му, а третий – 17-му), так что медиана равна $\frac{50+52}{2}=51$.□

Задача 7.2. Приводимая ниже древовидная диаграмма содержит результаты школьных соревнований в беге:

3	4	(1)
4	348	(3)
5	0134589	(7)
6	399	(3)
7	8	(1)

Ключ: **3** | 6 = 36(сек)

Опишите характер распределения результатов соревнований. Найдите медиану и среднее значение этого набора. Сравните их между собой и прокомментируйте результат.

Решение. Из древовидной диаграммы видно, что большая часть результатов лежит в промежутке 50 сек. $\le x \le$ 59 сек. Иначе говоря, эта группа результатов является единственным и ярко выраженным *модальным классом*.

Остальные значения разбросаны вокруг этого класса более-менее *симметрично*. В узле 6 содержится столько же листьев, сколько и в узле 4, а в узле 7 – столько же, сколько в узле 3. Более того, внутри центральных узлов (узлы 4,5,6) представлены как маленькие, так и большие значения листьев.

Пусть x_i – i-е в порядке возрастания число рассматриваемого набора. Поскольку рассматриваемый набор содержит нечётное количество элементов $n=15$, медиана μ_X является 8-м в порядке возрастания числом набора: $\mu_X=x_8$. Первые два узла содержат 4 листа; нетрудно видеть, что 8-й элемент находится в третьем узле. Первый лист третьего узла соответствует 5-му элементу набора, второй – 6-му, и т.д. Восьмому элементу набора соответствует 4-й лист. Поэтому медиана набора закодирована как 5 | 4, что означает 54 сек.

Чтобы найти среднее значение набора, сначала подсчитаем суммы листьев для каждого узла:

$$S_3 = 30 \cdot 1 + 4 = 34,$$
$$S_4 = 40 \cdot 3 + (3 + 4 + 8) = 135,$$
$$S_5 = 50 \cdot 7 + (0 + 1 + 3 + 4 + 5 + 8 + 9) = 380,$$
$$S_6 = 60 \cdot 3 + (3 + 9 + 9) = 201,$$
$$S_7 = 70 \cdot 1 + 8 = 78.$$

Тогда сумма всех членов нового набора равна 828, а среднее значение равно $\frac{828}{15} = 55{,}2$ (сек).

Практически идентичные значения среднего значения и медианы связаны с большой симметрией набора.

Симметрия влечёт ещё один важный результат. Если середину отрезка, соответствующего модальному классу (50 сек. $\le x \le$ 59 сек.), рассматривать как меру положения набора, то мы получим значение 54,5 сек., которое практически совпадает со значениями медианы и среднего.

Отметим также, что набор имеет единственную моду – значение 69 сек. (оно встречается чаще других – два раза). Но это значение не входит в модальный класс и сильно отличается как от медианы, так и от среднего значения. Поэтому в данном случае мода не может рассматриваться в качестве хорошей меры положения набора.□

Задача 7.3. Для набора из 100 чисел, введённого в начале пункта 5.1, постройте древовидную диаграмму и с её помощью охарактеризуйте распределение чисел набора.

Решение. Используя описанную выше процедуру, построим древовидную диаграмму, при этом в качестве узлов возьмем цифры десятков. Легко видеть, что все числа меньше 60, поэтому в диаграмме только 6 строк:

0	7889	(4)
1	002455678899	(12)
2	00112233445566777888999	(24)
3	000112222333444555666778899 9	(28)
4	0000111222345556677889	(22)
5	0122334455	(10)

Ключ: **5** | 0 = 50 мин.

Из неё видно, что бóльшая часть чисел набора (28 из 100, т.е. 28%) лежит в промежутке $30 \le x \le 39$. Иначе говоря, эта группа результатов является единственным *модальным классом*.

Остальные значения распределены вокруг этого класса с небольшим скосом в сторону меньших значений (такое распределение называют отрицательно асимметричным). Действительно, выше узла 3 расположены узлы 2, 1 и 0, содержащие 24, 12 и 4 листьев соответственно (им соответствует 40 чисел

исходного набора). Ниже него расположены узлы 4 и 5, содержащие 22 и 10 листьев соответственно (им соответствует 32 числа исходного набора).

Как мы покажем позже, этот вывод подтверждается и более подробным статистическим анализом, основанным на квартилях.

Пусть x_i – i-е в порядке возрастания число рассматриваемого набора. Поскольку рассматриваемый набор содержит чётное количество элементов $n=100$, медиана μ_X является средним арифметическим членов с номерами 50 и 51. Первые три узла содержат 40 листьев; им соответствует 40 элементов исходного набора. Поэтому 50-й и 51-й элементы закодированы в четвёртом узле. Первый лист из этого узла соответствует 41-му элементу набора, второй – 42-му, и т.д. Поэтому 50-му элементу набора соответствует 10-й лист из этого узла, а 51-му – 11-й, так что $\mu_X = \frac{33+33}{2} = 33$.

Чтобы найти нижнюю и верхнюю квартили, разделим (в соответствии с определением квартилей) упорядоченную версию исходного набора на две равные по числу элементов половины, из 50 чисел каждая. Нижняя квартиль Q_1 – это медиана первой половины: $Q_1 = \frac{x_{25}+x_{26}}{2} = \frac{24+24}{2} = 24$. Верхняя квартиль Q_3 – это медиана второй половины: $Q_3 = \frac{x_{75}+x_{76}}{2} = \frac{41+42}{2} = 41{,}5$.

Теперь мы можем количественно описать асимметрию набора с помощью квартильного коэффициента $k = \frac{Q_3+Q_1-2Q_2}{Q_3-Q_1} = \frac{41{,}5+24-66}{41{,}5-24} \approx -0{,}03$. Малое отрицательное значение квартильного коэффициента как раз означает небольшой скос распределения в сторону меньших значений.

Тот факт, что распределение практически симметрично, означает, в частности, что среднее значение должно мало отличаться от медианы. Поэтому в качестве меры положения набора можно взять число 33. Этот вывод согласуется и с отмеченным наличием единственного модального класса $30 \le x \le 39$: если в качестве меры положения взять середину этого класса, мы получим число 34,5, которое весьма близко к медиане. Отметим, что среднего значение набора равно 32,67, т.е., как и следовало ожидать, мало отличается от медианы, причём в меньшую сторону (это является следствием отрицательной асимметрии).

Точное вычисление среднего значения в рассматриваемом примере довольно утомительное занятие даже с использованием компьютера (ввод 100 чисел требует не только времени, но и внимания). Однако, очень легко получить оценку среднего.

Все числа, закодированные листьями узла 0, больше или равны, чем 0; все числа, закодированные листьями узла 1, больше или равны, чем 10; все числа, закодированные листьями узла 2, больше или равны, чем 20, и т.д. Поэтому сумма всех чисел набора больше или равна, чем

$0\cdot 4+10\cdot 12+20\cdot 24+30\cdot 28+40\cdot 22+50\cdot 10=2820$, так что среднее значение набора больше или равно, чем 28,2. Чтобы найти границу с другой стороны, нужно левые границы для всех чисел увеличить на 9. В результате левая граница для среднего также увеличится на 9 и составит 37,2. Имея в виду, что внутри каждого узла листья распределены более-менее равномерно (исключая крайние узлы 0 и 5), в качестве приближённого значения среднего значения анализируемого набора можно взять среднее арифметическое этих границ $\frac{28,2+37,2}{2}=32,7$ (это число можно получить и проще, прибавляя к левой границе число 4,5). Найденное приближение для среднего значения набора практически не отличается от его точного значения, равного 32,67.

Проведённое исследование можно немного уточнить, если каждый узел **s** разбить на два: $\mathbf{s}^{(0)}$ и $\mathbf{s}^{(5)}$; в первый собрать листья 0,1,2,3,4, а во второй – листья 5,6,7,8,9. В результате мы получим следующую модифицированную древовидную диаграмму:

$\mathbf{0}^{(0)}$		(0)
$\mathbf{0}^{(5)}$	7889	(4)
$\mathbf{1}^{(0)}$	0024	(4)
$\mathbf{1}^{(5)}$	55678899	(8)
$\mathbf{2}^{(0)}$	0011223344	(10)
$\mathbf{2}^{(5)}$	55666777888999	(14)
$\mathbf{3}^{(0)}$	000112222333444	(15)
$\mathbf{3}^{(5)}$	5556667788999	(13)
$\mathbf{4}^{(0)}$	000011122234	(12)
$\mathbf{4}^{(5)}$	5556677889	(10)
$\mathbf{5}^{(0)}$	01223344	(8)
$\mathbf{5}^{(5)}$	55	(2)

Ключ: **5** | 0 = 50(мин)

Как и раньше, хорошо видно, что набор приближённо является симметричным и имеет один модальный класс, который характеризуется неравенством $30\le x\le 34$. Середина этого класса, число 32, может рассматриваться как хорошая приближённая мера положения набора.□

8. О гистограмме и её свойствах

8.1 Введение

Гистограмма – это широко распространённый в статистике способ графического представления значений непрерывной переменной. В школьных учебниках гистограмме уделяется большое внимание, но при этом рассматриваются только простейшие их виды (с интервалами группировки постоянной длины), не показывается их связь с другими способами статистического описания (например, функцией распределения) и основными статистическими характеристиками (скажем, средним), не проводится ясное различие между столбиковой диаграммой и гистограммой. Последнее является следствием недостаточного внимания к основным понятиям статистики, изложенным в разделе 1. Например, в [13] считается, что «существует особый тип данных – сгруппированные данные», а «гистограмма является частным случаем столбиковой диаграммы для специальным образом сгруппированных исходных наборов чисел». На самом деле различие между гистограммой и столбиковой диаграммой гораздо глубже и связано с природой анализируемой переменной: если эта переменная является дискретной, то распределение её значений представляется столбиковой диаграммой, а если непрерывной – то гистограммой.

8.2 Определение гистограммы

Пусть $X = [x_1, \ldots, x_n]$ – набор (вообще говоря, неупорядоченный) значений некоторой непрерывной переменной X. Выберем некоторый промежуток в качестве множества возможных значений анализируемой переменной. Этот промежуток, конечно, должен содержать все числа x_i, так что он шире отрезка $\left[\min(x_1, \ldots, x_n); \max(x_1, \ldots, x_n)\right]$. Поскольку для непрерывной переменной отдельные значения малоинтересны, совершенно неважно, входят или нет в этот промежуток его граничные точки. Для определённости мы будем рассматривать промежутки вида $[a;b)$.

Предположим, что промежуток возможных значений $[a;b)$ разбит точками $c_0 = a < c_1 < \ldots < c_{k-1} < c_k = b$ на k непересекающихся промежутков (*классов* или *интервалов группировки*): $[c_0;c_1), [c_1;c_2), \ldots, [c_{k-1};c_k)$. Обратим внимание на то, что мы не предполагаем равенство длин $\Delta_i = c_i - c_{i-1}$ этих интервалов.

Пусть n_i – количество значений, попавших в i-й класс, $f_i = \dfrac{n_i}{n}$ – (относительная) частота попадания в i-й класс, $f_i^* = \dfrac{n_i}{n \cdot \Delta_i}$ – *нормированная* частота попадания в i-й класс.

Определение 8.1. Функция $h_n(x)$ действительного аргумента x, которая на промежутке $[a;b)$ возможных значений анализируемой переменной определена формулой

$$h_n(x) = f_i^* \equiv \frac{f_i}{\Delta_i}, \text{ если } c_{i-1} < x < c_i,$$

называется *гистограммой*.

В граничных точках интервалов группировки гистограмму можно определять произвольным способом, например, считать, что $h_n(c_{i-1}) = \frac{f_i}{\Delta_i}$, а можно вообще не определять. Удобно также считать, что вне промежутке $[a;b)$ возможных значений анализируемой переменной гистограмма тождественно равна 0.

График гистограммы является *кусочно-постоянным* и *разрывным*. Он получается если каждый интервал группировки $[c_0;c_1),[c_1;c_2),\ldots,[c_{k-1};c_k)$ поднять на высоту, равную нормированной частоте $f_i^* \equiv \frac{f_i}{\Delta_i}$ попадания в этот интервал. Поэтому график гистограммы состоит из k конечных отрезков длиной $\Delta_1,\ldots,\Delta_k$ и двух бесконечных лучей: левый луч идёт из точки $c_0 \equiv a$ на высоте 0 (совпадая с осью абсцисс) до $-\infty$, правый луч идёт из точки $c_k \equiv b$ на высоте 0 (совпадая с осью абсцисс) до $+\infty$. Характер разрывов в граничных точках интервалов группировки зависит от того, как определять значение гистограммы в этих точках.

Гистограмма $h_n(x)$ зависит не только от количества чисел в наборе, но и от самих этих чисел и классов группировки. Поэтому следовало бы как-нибудь отразить это обстоятельство в обозначениях, например, писать $h_n(x \mid x_1,\ldots,x_n;c_0,c_1,\ldots,c_k)$ вместо $h_n(x)$. Однако, это сильно усложняло бы обозначения и поэтому мы будем отмечать зависимость функции $h_n(x)$ от чисел набора и классов группировки только указанием общего количества чисел в наборе. По ходу изложения всегда будет понятно, о каком наборе и о каких классах идёт речь, но если возникает опасность путаницы, лучше использовать это более длинное обозначение.

Во многих учебниках, и не только школьных, рассматривают лишь равные интервалы группировки, а под гистограммой понимают ступенчатую *фигуру*, которая образована прямоугольниками, построенными на интервалах группировки как на основаниях; высота прямоугольника, построенного на промежутке $[c_{i-1};c_i)$, равна количеству значений n_i из соответствующего класса или частоте $f_i = \frac{n_i}{n}$. В отличие от этого, мы

- считаем, что гистограмма – это *функция*;

• используем нормированные частоты $f_i^* = \dfrac{n_i}{n \cdot \Delta_i}$ вместо абсолютных количеств n_i или частот $f_i = \dfrac{n_i}{n}$;

• допускаем интервалы группировки переменной длины.

Проиллюстрируем теперь наше определение гистограммы на конкретном примере. Рассмотрим набор из 100 чисел, введённый в начале пункта 5.1; числа набора – это время T (в минутах), которое 100 выбранных наугад учеников гипотетической школы тратят на дорогу в школу.

Вначале для этого конкретного числового набора построим гистограмму в соответствии с обычной процедурой.

Для построения гистограммы сгруппируем, как это сделано в учебнике [3], эти значения в 7 классов: 3–10, 11–18, 19–26, 27–34, 35–42,43–50, 51–58. В учебнике [3] приняты немного другие обозначения для классов, скажем, класс 11–18 обозначается 11–19, но говорится, что «значение, оказавшееся на границе двух интервалов, будем считать лежащим в правом промежутке». Вместо этой оговорки мы просто точнее описали все классы (сохраняя смысл подхода [3]). В результате мы получим следующую таблицу (для удобства последующего сравнения двух способов построения гистограммы мы подсчитали и нормированные частоты):

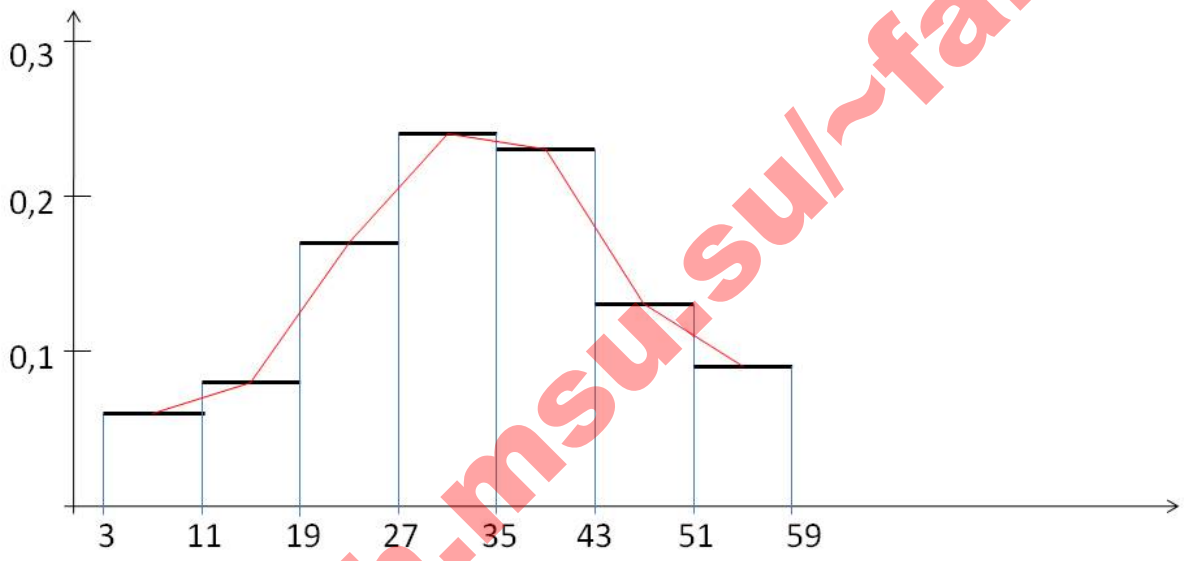
Время (мин)	Количество чисел в классе, n_i	Относительная частота, $f_i = \dfrac{n_i}{n}$	Нормированная частота, $f_i^* = \dfrac{n_i}{n\Delta_i}$
3–10	6	0,06	0,00750
11–18	8	0,08	0,01000
19–26	17	0,17	0,02125
27–34	24	0,24	0,03000
35–42	23	0,23	0,02875
43–50	13	0,13	0,01625
51–58	9	0,09	0,01125
Всего	**100**	**1**	

По поводу интервалов группировки необходимо сделать важное замечание. К сожалению, мы не знаем, как собиралась анализируемые данные – кто-то уже собрал их за нас (как мы определили в разделе 1, такие данные называют *вторичными*). В частности, мы не знаем использовавшуюся процедуру округления. Для определённости будем считать, что время округлялось до целых минут отбрасыванием секунд. Поэтому, например, запись 11–18 означает, что точное значение времени на дорогу до школы удовлетворяет двойному неравенству $11 \le T < 19$ – именно это двойное неравенство, а вовсе не запись 11–18 или 11–19, точно характеризует интервал группировки. Таким образом, результаты измерений сгруппированы по интервалам длиной 8 минут.

Если бы время округлялось до целых минут по обычному правилу (до ближайшего целого), то запись 11–18 означала бы, что точное значение времени на дорогу до школы удовлетворяет двойному неравенству $10{,}5 \le T < 18{,}5$. Таким образом, хотя длина интервала группировки осталась прежней, сам интервал изменился. Чтобы избежать недоразумений, лучше точно описывать классы группировки двойными неравенствами. В школьном учебнике на процедуру округления числовых данных вообще не обращают внимания, что недопустимо при аккуратном статистическом анализе непрерывных переменных.

Теперь мы можем построить гистограмму. Она изображена на рис.8.1.

Рис.8.1



Чтобы нагляднее показать характер гистограммы, на этом же рисунке мы нарисовали ломаную линию, которая соединяет между собой середины горизонтальных отрезков, из которых состоит гистограмма. Эта ломаная в англоязычной литературе по статистике называется frequency polygon, что при буквальном переводе на русский язык означает «ломаная линия, образованная частотами».

В школьном учебнике [3] используется термин «полигон». В принципе он допустим; в соответствии с Математическим энциклопедическим словарём (Гл.редактор Ю.В.Прохоров. – М., Сов. энциклопедия, 1988), «*Полигон* – ломаная линия, составленная из конечного числа прямоугольных отрезков (звеньев). Под полигоном также понимают замкнутую ломаную линию, т.е. многоугольник.» Нельзя, однако, не отметить, что:

(1) в современном русском языке «полигон» обычно означает «участок суши или моря, предназначенный для испытания оружия, боевых средств и техники и для боевой подготовки войск» (Советский энциклопедический словарь, М., Сов. энциклопедия, 1980);

(2) в современной математике «полигон» обычно означает «непустое множество с моноидом операторов» (Математическая энциклопедия/ Гл. ред. И.М.Виноградов. – М., Советская энциклопедия, 1984, том 4);

(3) если математики и допускают использование термина «полигон», то лишь как эквивалент терминов «многоугольник» или «ломая линия» *в самом общем их смысле*.

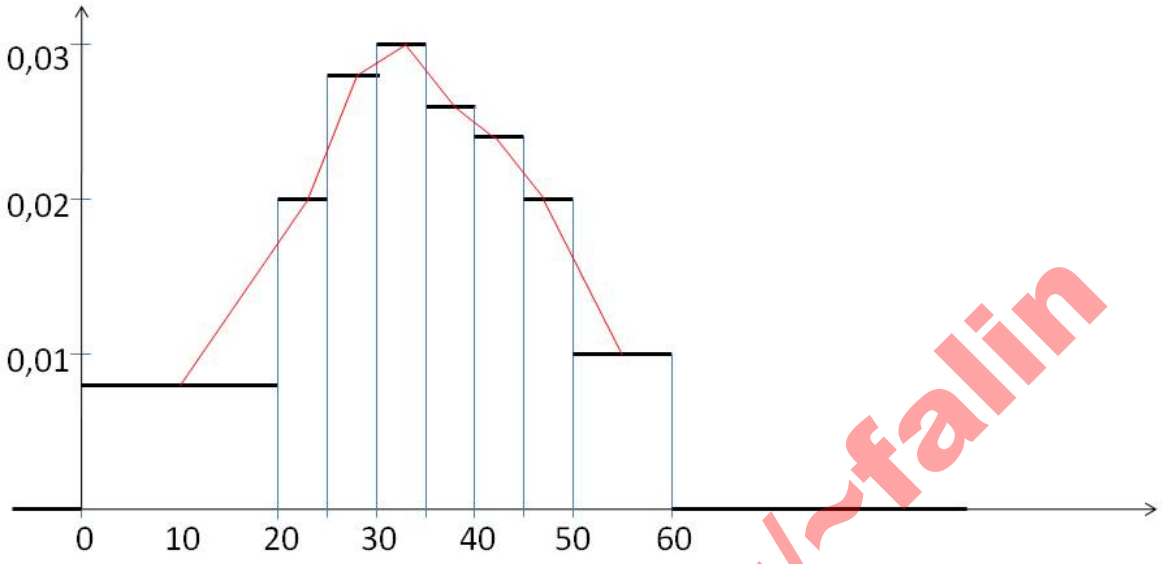
Отметим, кстати, что ломаная частот обычно строится только для гистограмм и не употребляется для столбиковых диаграмм (как рекомендуется в [3]; см., например, [8], стр.16, или [9], стр. 63). Иначе говоря, ломаная частот является специфическим графическим представлением распределения *непрерывных* переменных. Как правило, гистограмму и ломаную частот не строят на одном рисунке; мы сделали это лишь для экономии места. Строить ломаную частот можно сразу, без предварительного построения гистограммы. Для этого для каждого класса $[c_{i-1};c_i)$ нужно вычислить его центр $\frac{c_{i-1}+c_i}{2}$, отметить на координатной плоскости точку $\left(\frac{c_{i-1}+c_i}{2};f_i\right)$ и соединить все эти точки отрезками.

Теперь применим нашу процедуру построения гистограммы. Прежде всего, в качестве диапазона возможных значений времени на дорогу до школы возьмём более широкий промежуток $0<T<60$ (мин). Беглый взгляд на исходные данные показывает, что большинство школьников тратит на дорогу от 20 до 50 минут. Этот промежуток мы разобьём на 6 равных интервалов группировки, длиной 5 минут каждый. В каждый из них попадает около 10 чисел набора. В интервалы $0<T<20$ и $50<T<60$ попадает 16 и 10 чисел соответственно. Поэтому мы не будем дробить их на более мелкие промежутки. В результате мы получим следующую таблицу:

Интервал группировки	Длина интервала, Δ_i (мин)	Количество чисел в классе, n_i	Нормированная частота, $f_i^* = \frac{n_i}{n\Delta_i}$
$0<T<20$	20	16	0.008
$20\le T<25$	5	10	0.02
$25\le T<30$	5	14	0.028
$30\le T<35$	5	15	0.03
$35\le T<40$	5	13	0.026
$40\le T<45$	5	12	0.024
$45\le T<50$	5	10	0.02
$50\le T<60$	10	10	0.01
Всего	**60**	**100**	

Теперь мы можем построить гистограмму. Она изображена на рис.8.2 (обратим внимание на то, что масштабы по осям не такие, как на рис.8.1).

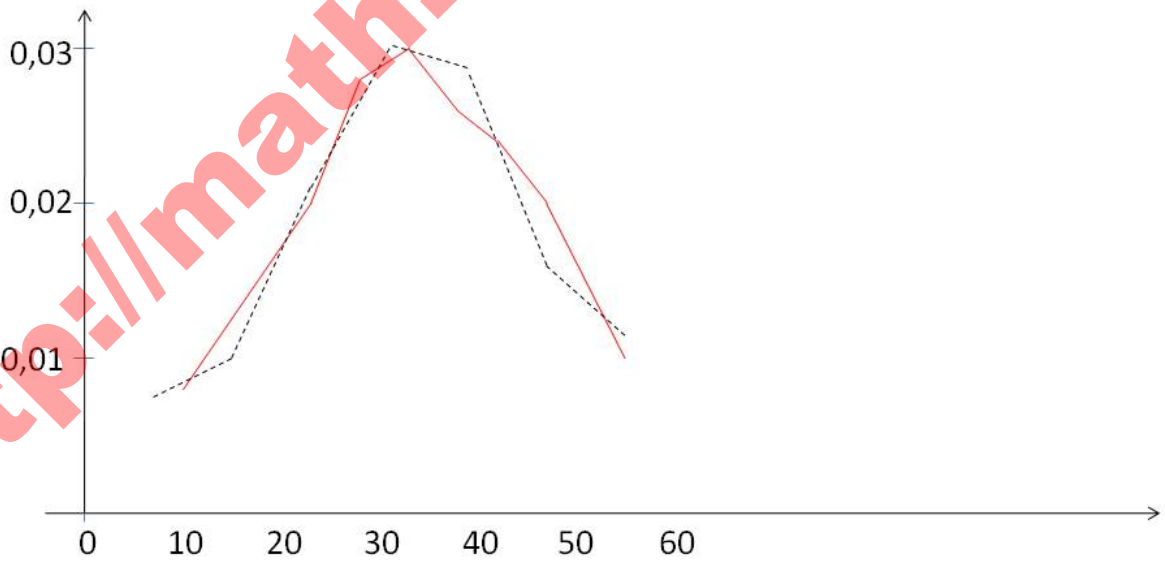
Рис.8.2



Как и на рис.8.1, чтобы нагляднее показать характер распределения, мы изобразили на рис.8.2 и ломаную частот.

Нетрудно видеть, что в рассматриваемом конкретном случае обе гистограммы дают примерно одно и то же визуальное представление о характере распределения анализируемой переменной.

Рис.8.3



Более того, как видно из рис.8.3, где мы изобразили ломаные частот для первого способа с нормированными частотами (штрихованная линия) и второго способа (непрерывная линия), гистограммы практически идентичны. Это связано с большой симметрией набора.

Для асимметричных распределений с длинными «хвостами» (а на практике часто встречаются именно такие распределения) рассматривать, как это обычно рекомендуется в учебниках по статистике, интервалы группировки равной длины неразумно. Чтобы продемонстрировать это, рассмотрим настоящий, а не «игрушечный» пример (он взят из монографии [12], стр. 64, посвящённой применению статистических методов в страховании).

В 1977 году в США были зафиксированы 40 катастрофических ураганов, каждый из которых привёл к убыткам, превышающим 1,5 миллиона долларов. Величины потерь L в миллионах долларов, округлённые до целого числа миллионов, приведены ниже (для удобства последующего анализа значения потерь упорядочены по возрастанию): 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 3, 3, 3, 3, 4, 4, 4, 5, 5, 5, 5, 6, 6, 6, 6, 8, 8, 9, 15, 17, 22, 23, 24, 24, 25, 27, 32, 43.

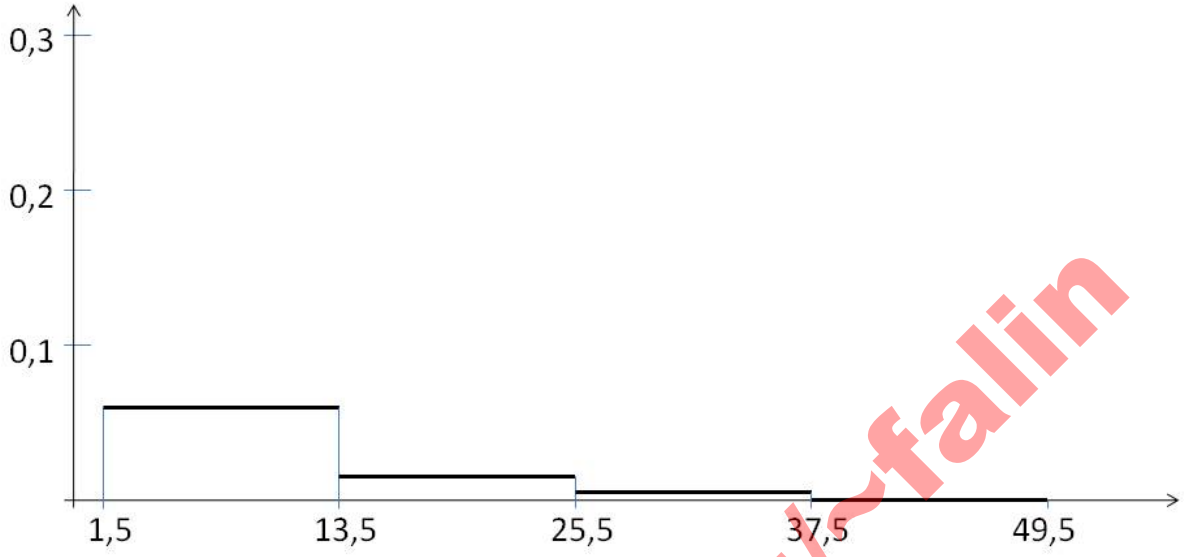
Правильный статистический анализ подобных данных о потерях критически важен для обеспечения финансовой устойчивости страховых компаний. Им занимаются *актуарии* – специалисты по математическим и статистическим расчётам в страховании.

В соответствии с принятой процедурой округления, например, значение 2 означает, что реальные потери L находились в промежутке $1{,}5 \le L < 2{,}5$. Чаще всего в списке встречается число 2, но его нельзя называть модой, т.к. в описательной статистике этот термин употребляется только в случае наборов значений дискретной переменной. В нашей ситуации числу 2 соответствует целый класс значений (1,5;2,5). Поэтому говорят, что распределение имеет единственный и ярко выраженный *модальный класс* $1{,}5 \le L < 2{,}5$.

Среднее значение потерь равно $M_L = 9{,}225$, медиана $\mu_L = 5$, среднее квадратичное (стандартное) отклонение $\sigma_L \approx 10{,}108$. Поэтому коэффициент асимметрии Пирсона $3\dfrac{M_L - \mu_L}{\sigma_L} \equiv 3\dfrac{\text{среднее значение} - \text{медиана}}{\text{стандартное отклонение}}$ равен примерно 1,25 (максимальное значение равно 3), что свидетельствует о значительной положительной асимметрии набора (это, впрочем, ясно видно и из беглого взгляда на приведённый ряд значений потерь).

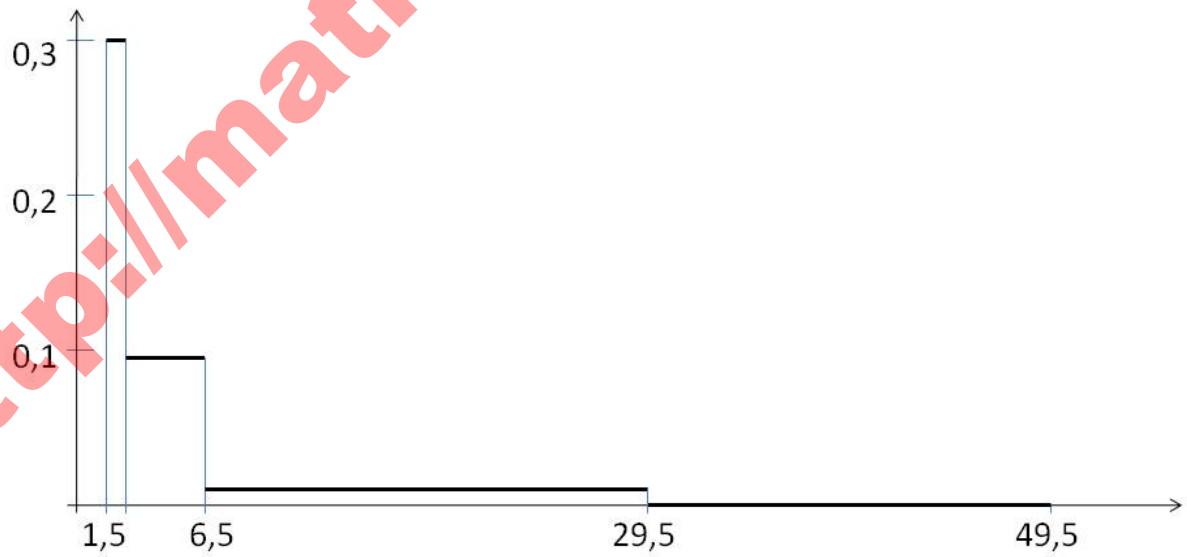
Выберем в качестве диапазона возможных потерь промежуток (1,5;49,5) длиной 48 и разобьём его на 4 промежутка одинаковой длины 12: $1{,}5 < L < 13{,}5$; $13{,}5 \le L < 25{,}5$; $25{,}5 \le L < 37{,}5$; $37{,}5 \le L < 49{,}5$ с центрами в точках 7,5; 19,5; 31,5; 43,5 соответственно. Количество чисел набора, попадающих в эти классы, равно 30, 7, 2, 1 соответственно, так что относительные частоты равны 0,75; 0,175; 0,05; 0,025 соответственно, а нормированные частоты равны 0,0625; 0,0146; 0,0042; 0,0021 соответственно. Построенная по этим данным гистограмма приведена на рис.8.4. Для удобства сравнения гистограмм мы и в случае равных по длине интервалов группировки строим гистограмму по нормированным частотам.

Рис.8.4



Теперь разобьём, как сделано в [12], промежуток (1,5;49,5) на 4 промежутка: $1,5<L<2,5$; $2,5\le L<6,5$; $6,5\le L<29,5$; $29,5\le L<49,5$ длиной 1, 4, 23, 20 соответственно с центрами в точках 2; 4,5; 18; 39,5 соответственно. Количество чисел набора, попадающих в эти классы, равно 12, 15, 11, 2 соответственно, так что нормированные частоты равны 0,3000; 0,0938; 0,0120; 0,0025 соответственно. Построенная по этим данным гистограмма приведена на рис.8.5.

Рис.8.5

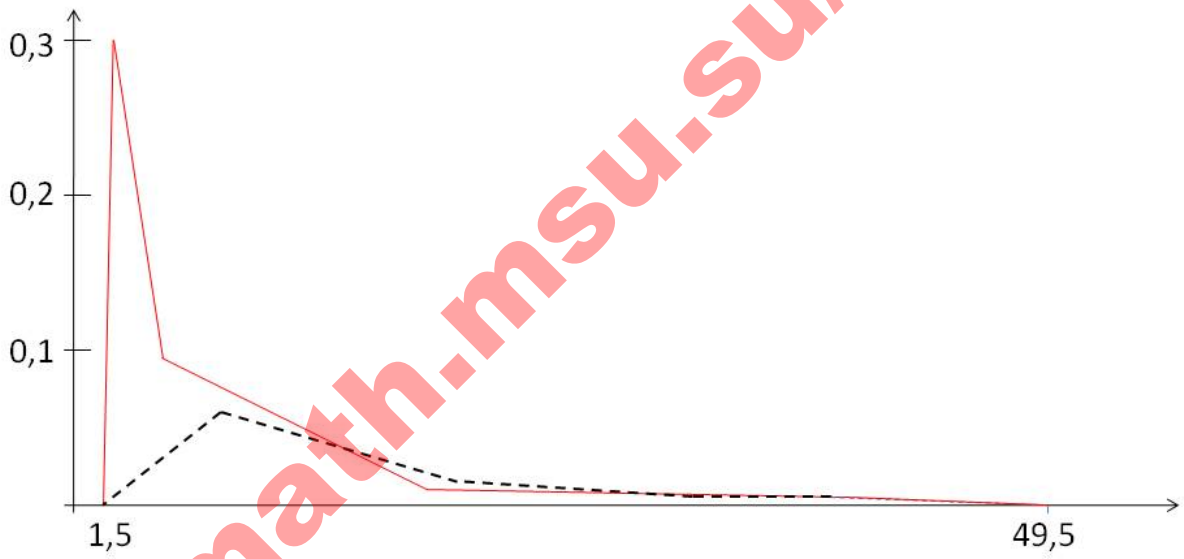


Как видно из рис. 8.4 и 8.5, гистограммы совершенно различны (на этих рисунках масштаб один и тот же). Особенно хорошо отличие видно на рис.8.6, где мы изобразили ломаные частот для первого способа группировки (с интервалами группировки равной длины; штрихованная линия) и второго

способа (с интервалами группировки разной длины; непрерывная линия). На этом рисунке мы соединили точки, соответствующие крайним интервалам группировки с точками 1,5 и 49,5 на оси абсцисс, чтобы подчеркнуть, что гистограмма равна 0 вне промежутка (1,5;49,5).

Если подсчитать среднее значение набора по сгруппированным данным, заменяя все значения из класса его центром, то первый способ группировки даст значение 11,7, а второй – значение 9,2125. Напомним, что точное среднее значение потерь равно $M_L = 9{,}225$ (мы игнорируем тот факт, что на самом деле мы работаем с округлёнными значениями потерь). Таким образом, даже это простое соображение показывает, что второй способ группировки лучше. Более тонкий и сложный статистический анализ (детали можно найти в [12], стр. 94-95, 116-118) показывает, что только подход, основанный на интервалах группировки переменной длины, адекватно моделирует реальную ситуацию.

Рис.8.6



8.3 Свойства гистограммы

В этом пункте мы установим несколько важных общих свойств гистограммы $h_n(x)$.

Как и в начале предыдущего пункта, рассмотрим набор $X = [x_1, \ldots, x_n]$ значений некоторой непрерывной переменной X и разобьём промежуток возможных значений $[a;b)$ точками $c_0 = a < c_1 < \ldots < c_{k-1} < c_k = b$ на k классов: $[c_0;c_1), [c_1;c_2), \ldots, [c_{k-1};c_k)$. Пусть $\Delta_i = c_i - c_{i-1}$ – длина i-го промежутка группировки, n_i – количество значений, попавших в i-й класс, $f_i = \frac{n_i}{n}$ – (относительная)

частота попадания в i-й класс, $f_i^* = \dfrac{n_i}{n \cdot \Delta_i}$ – нормированная частота попадания в i-й класс, а гистограмма (функция действительного аргумента x) определена формулой

$$h_n(x) = \begin{cases} f_i^*, & \text{если } c_{i-1} \le x < c_i,\ i = 1,\ldots,k; \\ 0, & \text{если } x < a \equiv c_0 \text{ или } x \ge b \equiv c_k. \end{cases}$$

Прежде всего отметим, что непосредственно из определения гистограммы следует

Свойство 1. *При всех значениях аргумента x верно неравенство*

$$h_n(x) \ge 0. \qquad (8.1)$$

Свойство 2. *Площадь фигуры, ограниченной снизу осью абсцисс, а сверху гистограммой, равна* 1.

Доказательство. Поскольку вне промежутка $[a;b)$ возможных значений переменной X гистограмма тождественно равна 0, фигура, о которой идёт речь, может быть разбита на k прямоугольников; i-й прямоугольник, $i = 1,\ldots,k$, построен на i-м интервале группировки $[c_{i-1};c_i)$ как на основании (так что длина основания равна Δ_i), а его высота равна нормированной частоте $f_i^* = \dfrac{n_i}{n \cdot \Delta_i}$. Поэтому площадь S этой фигуры равна сумме площадей всех k прямоугольников:

$$S = \Delta_1 \cdot f_1^* + \ldots + \Delta_k \cdot f_k^* = \frac{n_1}{n} + \ldots + \frac{n_k}{n} = \frac{n_1 + \ldots + n_k}{n}.$$

Но $n_1 + \ldots + n_k$ – это общее количество чисел в наборе, так что S=1, как мы и утверждаем.

Поскольку площадь фигуры, ограниченной снизу осью абсцисс, а сверху графиком некоторой функции, равна интегралу от этой функции, доказанное свойство можно записать в эквивалентной форме как равенство:

$$\int_{-\infty}^{+\infty} h_n(x)dx = 1. \qquad (8.2)$$

Отметим, что, так как гистограмма тождественно равна 0 вне промежутка $[a;b)$ возможных значений переменной X, пределы интегрирования в интеграле в левой части (8.2) можно изменить с $-\infty$ и $+\infty$ на a до b соответственно.

Дословное повторение рассуждений, проведённых при доказательстве Свойства 2, позволяет доказать, что площадь фигуры, ограниченной снизу осью абсцисс, сверху гистограммой, а справа вертикальной прямой, проходящей через точку x на оси абсцисс,

- равна 0, если $x \le a$;
- даётся формулой: $f_1 + \ldots + f_{i-1} + f_i \dfrac{x - c_{i-1}}{c_i - c_{i-1}}$, если $x \in \left[c_{i-1};c_i\right)$;

- равна 1, если $x>b$.

Но по этим же формулам вычисляется непрерывное приближение $F_n^*(x)$ функции распределения набора по сгруппированным значениям (подробно о функции распределения мы будем говорить в разделе 9). Поэтому верно равенство

$$\int_{-\infty}^{x} h_n(u)du = F_n^*(x). \qquad (8.3)$$

Теперь мы установим важную формулу, позволяющую подсчитать с помощью гистограммы среднее значение набора. Напомним, что среднее значение набора, M_X, *по определению* даётся формулой: $M_X = \frac{x_1 + \ldots + x_n}{n}$. Если числа x_i сгруппированы в классы и $M_x^{(i)}$ – среднее арифметическое n_i чисел из i-го интервала группировки $\left[c_{i-1};c_i\right)$, то эту формулу можно переписать в виде:

$$M_X = \frac{M_X^{(1)} \cdot n_1 + \ldots + M_X^{(k)} \cdot n_k}{n} \equiv M_X^{(1)} \cdot f_1 + \ldots + M_X^{(k)} \cdot f_k.$$

Если предположить, что $M_X^{(i)} \approx \frac{c_{i-1} + c_i}{2}$, т.е. что среднее арифметическое чисел из каждого интервала группировки расположено почти посередине этого интервала, то мы получим приближённую формулу:

$$M_X \approx \frac{c_0 + c_1}{2} \cdot f_1 + \ldots + \frac{c_{k-1} + c_k}{2} \cdot f_k. \qquad (8.4)$$

Выражение в правой части этого равенства мы назовём средним значением набора, подсчитанным по сгруппированным данным, и обозначим M_X^*. С практической точки зрения можно считать, что $M_X = M_X^*$.

Свойство 3. *Среднее значение набора, подсчитанное по сгруппированным данным даётся формулой*

$$M_X^* = \int_{-\infty}^{+\infty} x h_n(x)dx. \qquad (8.5)$$

Иначе говоря, величина M_X^* равна алгебраической площади (т.е. площади с учётом того, что площадь той части фигуры, которая находится под осью абсцисс, учитывается со знаком «–») фигуры, ограниченной осью абсцисс и графиком функции $xh_n(x)$.

Доказательство. Рассмотрим функцию $xh_n(x)$. Она тождественно равна 0 вне промежутка $[a;b)$ возможных значений переменной X, а на i-м интервале группировки $[c_{i-1};c_i)$ даётся формулой: $xh_n(x) = f_i^* \cdot x$, так что соответствующий

кусок графика функции $xh_n(x)$ является отрезком прямой с угловым коэффициентом f_i^*.

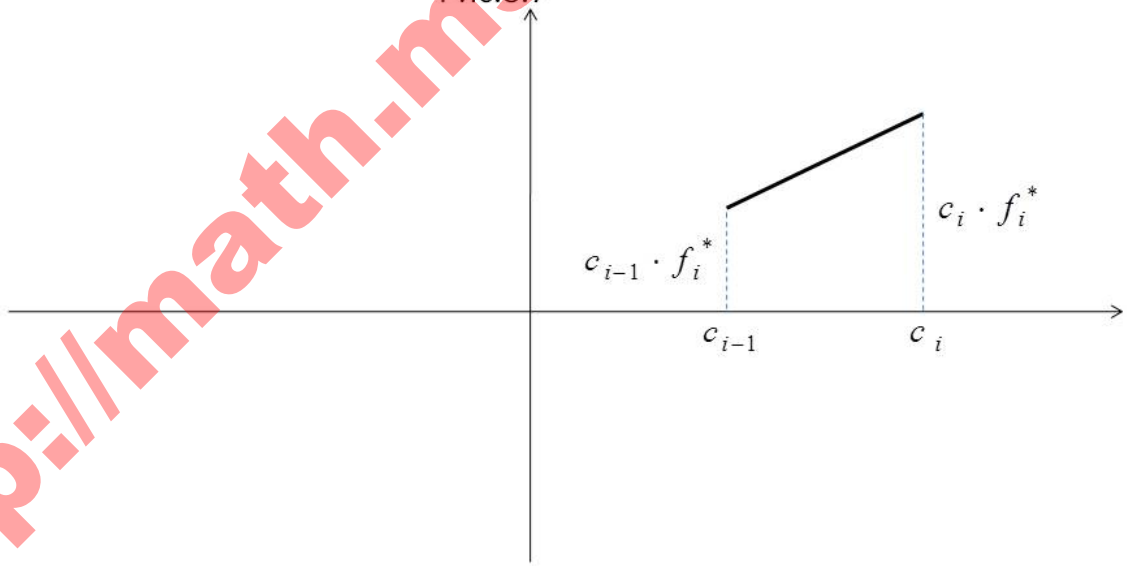
Фигура, о которой идёт речь в формулировке Свойства 3, состоит из k фигур: $\Phi_1,\ldots,\Phi_k$. Фигура Φ_i ограничена промежутком $[c_{i-1};c_i)$ на оси абсцисс, отрезком прямой $y = f_i^* \cdot x$ и двумя вертикальными прямыми, проходящими через точки c_{i-1}, c_i на оси абсцисс. Мы докажем, что алгебраическая площадь этой фигура равна $\frac{c_{i-1}+c_i}{2} \cdot f_i$, т.е. i-му слагаемому в правой части (4). Отсюда, очевидно, вытекает требуемое утверждение.

Если $f_i^* = 0$, то фигура Φ_i вырождается в отрезок, площадь которого мы считаем равной 0, так что формула $\frac{c_{i-1}+c_i}{2} \cdot f_i$ для её площади верна. Поэтому ниже будем предполагать, что $f_i^* > 0$.

Логически возможны 3 случая: (1) $c_{i-1} \ge 0$, (2) $c_i \le 0$, (3) $c_{i-1} < 0 < c_i$.

Если $c_{i-1} > 0$, т.е. промежуток $[c_{i-1};c_i)$ расположен на положительной полуоси, то фигура Φ_i лежит выше оси абсцисс и является трапецией с основаниями $c_{i-1} \cdot f_i^*$, $c_i \cdot f_i^*$ и высотой Δ_i (см. рис. 8.7).

Рис.8.7



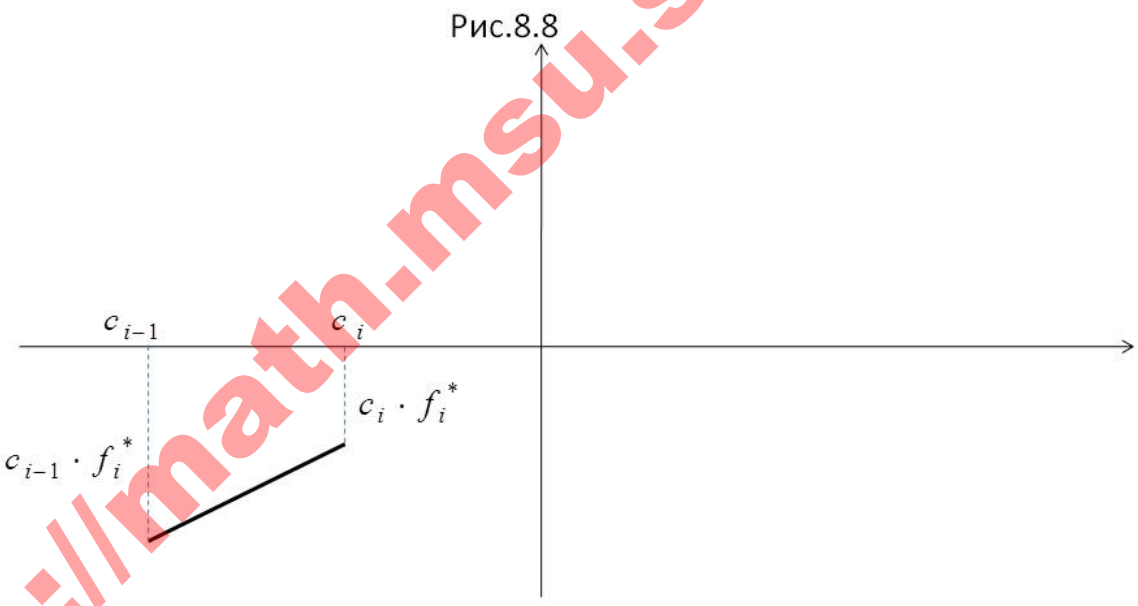
Поэтому её площадь равна $\frac{c_{i-1} \cdot f_i^* + c_i \cdot f_i^*}{2} \cdot \Delta_i = \frac{c_{i-1}+c_i}{2} \cdot f_i^* \cdot \Delta_i = \frac{c_{i-1}+c_i}{2} \cdot f_i,$

т.е. i-му слагаемому в правой части (8.4).

Если $c_{i-1}=0$, то трапеция Φ_i вырождается в прямоугольный треугольник с с основанием $c_i \cdot f_i^*$ и высотой Δ_i. Его площадь равна $\frac{1}{2} c_i \cdot f_i^* \cdot \Delta_i = \frac{c_{i-1}+c_i}{2} \cdot f_i$ (с учётом того, что $c_{i-1}=0$), т.е. опять i-му слагаемому в правой части (8.4).

Если промежуток $[c_{i-1};c_i)$ расположен на отрицательной полуоси, т.е. $c_i<0$, то фигура Φ_i опять является трапецией, но целиком лежит ниже они абсцисс. Её основания равны $(-c_{i-1})\cdot f_i^*>0$, $(-c_i)\cdot f_i^*>0$ (поскольку $c_{i-1}, c_i<0$, эти длины выражается неотрицательными числами), а высота равна Δ_i (см. рис. 8.8). Геометрически площадь этой трапеции равна $\frac{(-c_{i-1})\cdot f_i^* + (-c_i)\cdot f_i^*}{2} \cdot \Delta_i = -\frac{c_{i-1}+c_i}{2} \cdot f_i^* \cdot \Delta_i = -\frac{c_{i-1}+c_i}{2} \cdot f_i > 0$. Если (как мы условились) считать площадь такой фигуры со знаком «–», то алгебраическое значение этой площади равно $\frac{c_{i-1}+c_i}{2} \cdot f_i < 0$, т.е. опять i-му слагаемому в формуле (8.4) для подсчёта M_X по сгруппированным данным.

Рис.8.8



Если $c_i=0$, то эта трапеция Φ_i вырождается в прямоугольный треугольник с с основанием $(-c_{i-1})\cdot f_i^*$ и высотой Δ_i, лежащий ниже оси абсцисс. Его площадь с учётом знака равна $-\frac{1}{2}(-c_{i-1})\cdot f_i^* \cdot \Delta_i = \frac{c_{i-1}+c_i}{2} \cdot f_i$ (с учётом того, что $c_i=0$), т.е. опять i-му слагаемому в правой части (8.4).

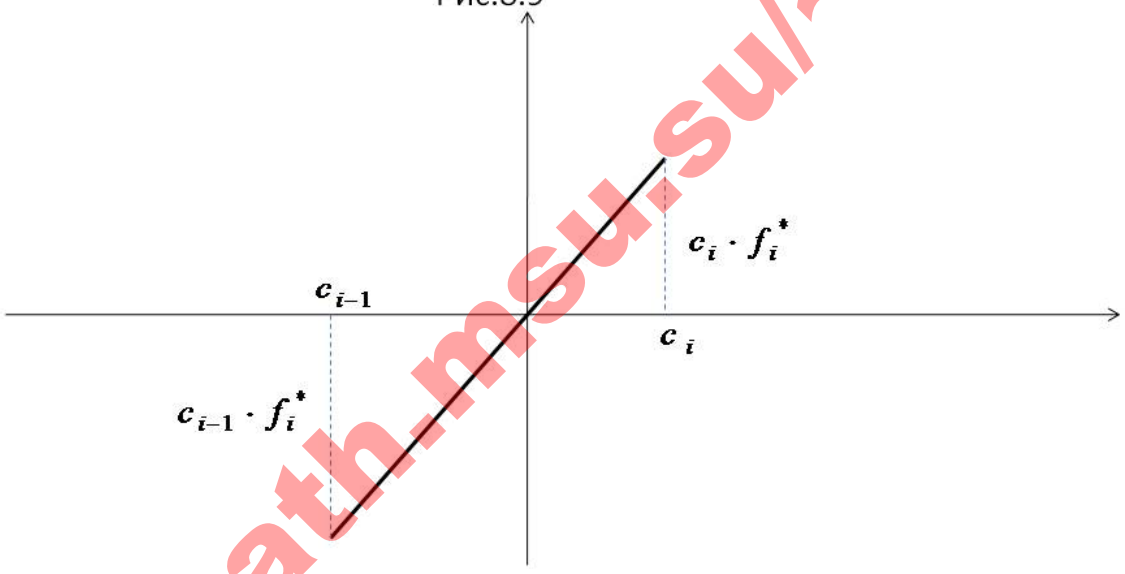
И наконец, пусть $c_{i-1}<0<c_i$. В этом случае фигура Φ_i состоит из двух треугольников (см. рис.8.9). Длина основания левого треугольника равна $(-c_{i-1})$ (поскольку $c_{i-1}<0$, эта длина выражается положительным числом), а высота

равна $(-c_{i-1}f_i^*)$ (поскольку $c_{i-1}<0$, эта длина выражается положительным числом). Поэтому (геометрически) площадь левого треугольника равна $\frac{1}{2}c_{i-1}^2 f_i^*$. Так как этот треугольник лежит под осью абсцисс, его площадь должна браться со знаком «–», т.е. алгебраически она равна $-\frac{1}{2}c_{i-1}^2 f_i^*$. Площадь правого треугольника равна $\frac{1}{2}c_i^2 f_i^*$, так что алгебраическая площадь всей фигуры Φ_i равна

$$\frac{1}{2}c_i^2 f_i^* - \frac{1}{2}c_{i-1}^2 f_i^* = \frac{c_i^2 - c_{i-1}^2}{2}\cdot f_i^* = \frac{(c_i - c_{i-1})(c_i + c_{i-1})}{2}\cdot\frac{f_i}{c_i - c_{i-1}} = \frac{c_i + c_{i-1}}{2}\cdot f_i,$$

т.е. опять i-му слагаемому в правой части (8.4).

Рис.8.9



Доказанные соотношения (8.1), (8.2), (8.3), (8.5) означают, что гистограмма, определённая в соответствии с нашим определением, обладает всеми свойствами, характерными для плотности случайной величины. Поэтому описанный подход к понятию гистограммы может служить хорошей основой для развития понятной школьникам теории непрерывных случайных величин, включая теорию нормального распределения (хорошо известно, насколько тяжело изложить эту теорию достаточно строго и в то же время понятно даже студентам нематематических специальностей).

8.4 Задачи

Задача 8.1. В 1982 г. американский физик Альберт Майкельсон провёл серию опытов по измерению скорости света. Результаты 25 из них представлены в виде гистограммы, которая изображена на рис. 8.10.

Найдите модальный класс. Сколько опытов дали результат в этом диапазоне? Сколько опытов дали результат от 299,80 (тыс.км/сек) до 299,85 (тыс.км/сек)?

Рис.8.10

Решение. Модальный класс определяется интервалом группировки $299{,}8 \le v < 299{,}9$. Соответствующее значение гистограммы равно 3,2, а длина этого интервала равна 0,1. Поэтому количество результатов, попавших в этот класс, равно $25 \cdot 3{,}2 \cdot 0{,}1 = 8$.

Если считать, что внутри этого интервала результаты распределены более-менее равномерно, то число опытов, давших для скорости света значение от 299,80 до 299,85 равно 4.

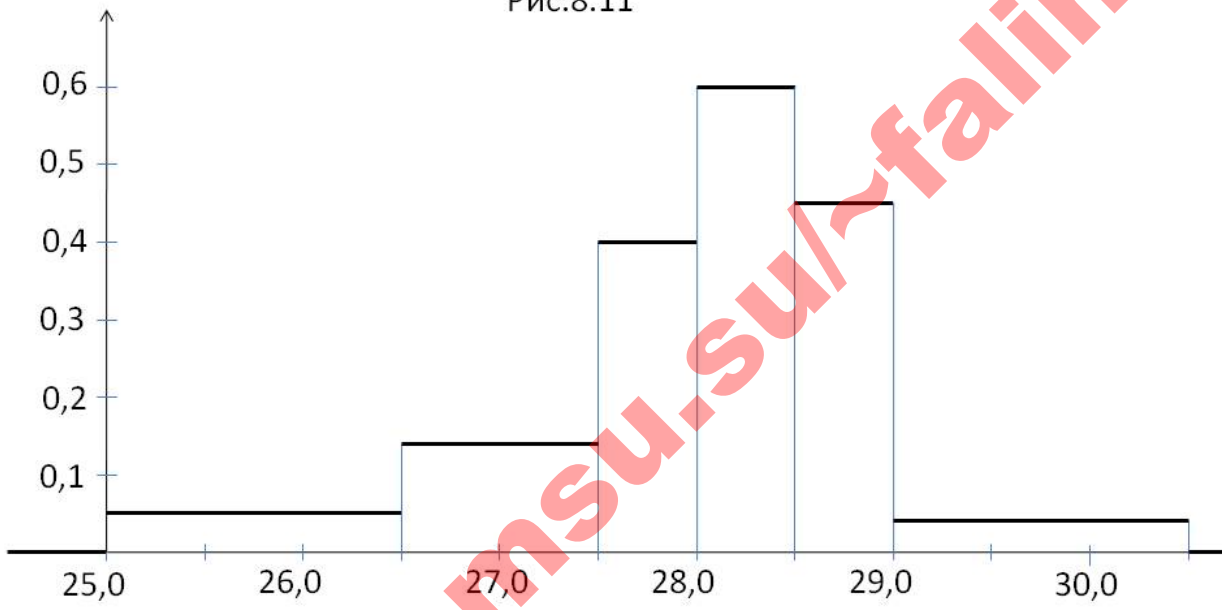
Задача 8.2. Крупная обувная фабрика заказала социологическое исследование, в ходе которого у 200 мужчин была измерена длина стопы. Результаты приведены в следующей таблице.

длина стопы (см)	число случаев	длина стопы (см)	число случаев
$25{,}0 \le L < 26{,}5$	15	$28{,}0 \le L < 28{,}5$	60
$26{,}5 \le L < 27{,}5$	28	$28{,}5 \le L < 29{,}0$	45
$27{,}5 \le L < 28{,}0$	40	$29{,}0 \le L < 30{,}5$	12

Постройте гистограмму по этим данным. Как обувная фабрика может использовать результаты исследования?

Решение. Гистограмма, построенная по этим данным, изображена на рис.8.11. Фабрика может использовать результаты исследования, чтобы определить, какую долю в общем объёме производства должна занимать обувь каждого размера (если распределение размера обуви у 200 мужчин, участвовавших в исследовании, примерно такое же, как и у мужчин тех регионов, где фабрика предполагает продавать обувь).

Рис.8.11



Задача 8.3. Школьник рисует гистограмму по сгруппированным значениям непрерывной переменной X. В класс $1{,}7 \le x < 2$ попадает 15 значений, а значение гистограммы для этого класса равно 1,0. В класс $2{,}0 \le x < 2{,}1$ попадает 10 значений. Найдите значение гистограммы для этого класса.

Решение. Используя данные относительно первого класса и формулу $f_i^* = \dfrac{n_i}{n \cdot \Delta_i}$, мы найдём общее число значений: $n = \dfrac{n_i}{f_i^* \cdot \Delta_i} = \dfrac{15}{1{,}0 \cdot 0{,}3} = 50$. Теперь из той же формулы можно найти значение гистограммы для второго класса: $f_i^* = \dfrac{n_i}{n \cdot \Delta_i} = \dfrac{10}{50 \cdot 0{,}1} = 2$.

Задача 8.4. Школьник рисует гистограмму по сгруппированным значениям непрерывной переменной U. Значение гистограммы равно 0,8 для класса $3{,}3 \le u < 3{,}5$ и 0,3 для класса $3{,}5 \le u < 3{,}8$. Затем он решил объединить эти два класса в один. Найдите значение гистограммы для этого класса.

Решение. Используя индекс 1 для величин, относящихся к первому промежутку группировки, и 2 для величин, относящихся ко второму, для искомой величины имеем:

$$f_i^* = \frac{n_1 + n_2}{n \cdot (\Delta_1 + \Delta_2)} = \frac{f_1^* \cdot n \cdot \Delta_1 + f_2^* \cdot n \cdot \Delta_2}{n \cdot (\Delta_1 + \Delta_2)} = \frac{f_1^* \cdot \Delta_1 + f_2^* \cdot \Delta_2}{\Delta_1 + \Delta_2} = 0{,}5.$$

9. Функция распределения числового набора

9.1 Введение

Функция распределения является специфическим способом графического представления данных. Это понятие играет ключевую роль в теории вероятностей и математической статистике. В теории вероятностей функция распределения *полностью* характеризует стохастическую природу *любой* случайной величины. В математической статистике аналогом функции распределения из теории вероятностей является эмпирическая функция распределения, с которой связан ряд фундаментальных результатов этой науки (достаточно упомянуть критерий Колмогорова для проверки гипотезы о виде функции распределения). В описательной статистике обычно рассматривают определённые модификации стандартной эмпирической функции распределения, которые позволяют проще и нагляднее отобразить характер распределения чисел набора на числовой оси. С помощью функции распределения можно по-новому взглянуть на многие понятия описательной статистики: среднее значение, медиану, гистограмму. Отметим также, что для статистического анализа набора значений непрерывной переменной функция распределения предпочтительнее гистограммы.

9.2 Функция распределения числового набора

Пусть $X = [x_1, \ldots, x_n]$ – набор из n действительных чисел (вообще говоря, неупорядоченный). Обозначим через $K_n(a)$ количество чисел набора, удовлетворяющих неравенству $x < a$, а через $F_n(a) = \dfrac{K_n(a)}{n}$ – долю таких чисел. Параметр a может быть любым действительным числом. Поэтому $F_n(a)$ можно рассматривать как функцию аргумента a. Эта функция и называется *функцией распределения числового набора*. Поскольку сам набор получен из какого-то опыта, эксперимента (таковым мы считаем и социологический опрос и т.п.), эту функцию в математической статистике называют *эмпирической* (т.е. полученной на основе опыта или наблюдений; от греческого empeiria – опыт). В англоязычной литературе по теории вероятностей обычно определяют эмпирическую функцию распределения $F_n(a)$ как долю чисел набора, удовлетворяющих *нестрогому* неравенству $x \le a$. Наше определение следует отечественной традиции.

Как $K_n(a)$, так и $F_n(a)$ зависят не только от количества чисел в наборе, но и от самих этих чисел. Поэтому следовало бы как-нибудь отразить это обстоятельство в обозначениях, например, писать $K_n(a \mid x_1,\ldots,x_n)$ вместо $K_n(a)$ и $F_n(a \mid x_1,\ldots,x_n)$ вместо $F_n(a)$. Однако, это сильно усложняло бы обозначения и поэтому мы будем отмечать зависимость введённых функций от чисел набора только указанием общего количества чисел в наборе. По ходу изложения всегда будет понятно, о каком наборе идёт речь, но если возникнет опасность путаницы, лучше использовать эти более длинные обозначения.

С помощью функции распределения можно легко подсчитать, сколько чисел набора попадает на тот или иной промежуток. Например, количество чисел набора, больших или равных a, равно $n - K_n(a) = n\big(1 - F_n(a)\big)$, а доля таких чисел равна $1 - F_n(a)$ (функция $\overline{F_n}(a) \equiv 1 - F_n(a)$ называется дополнительной функцией распределения), количество чисел набора, попадающих на промежуток $[a;b)$, равно $K_n(b) - K_n(a) = n\big(F_n(b) - F_n(a)\big)$, а доля таких чисел равна $F_n(b) - F_n(a)$.

Часто удобно использовать следующее представление функции распределения. Для произвольного действительного числа x введём функцию $I_x(a)$ действительного аргумента a по формуле:

$$I_x(a) = \begin{cases} 0, & \text{если } x \ge a, \\ 1, & \text{если } x < a. \end{cases}$$

Эта функция называется *индикаторной* или *индикатором*, т.к. она указывает, верно (значение индикатора равно 1) или нет (значение индикатора равно 0) неравенство $x < a$. Тогда $K_n(a) = I_{x_1}(a) + I_{x_2}(a) + \ldots + I_{x_n}(a)$ и, соответственно,

$$F_n(a) = \frac{I_{x_1}(a) + I_{x_2}(a) + \ldots + I_{x_n}(a)}{n}.$$

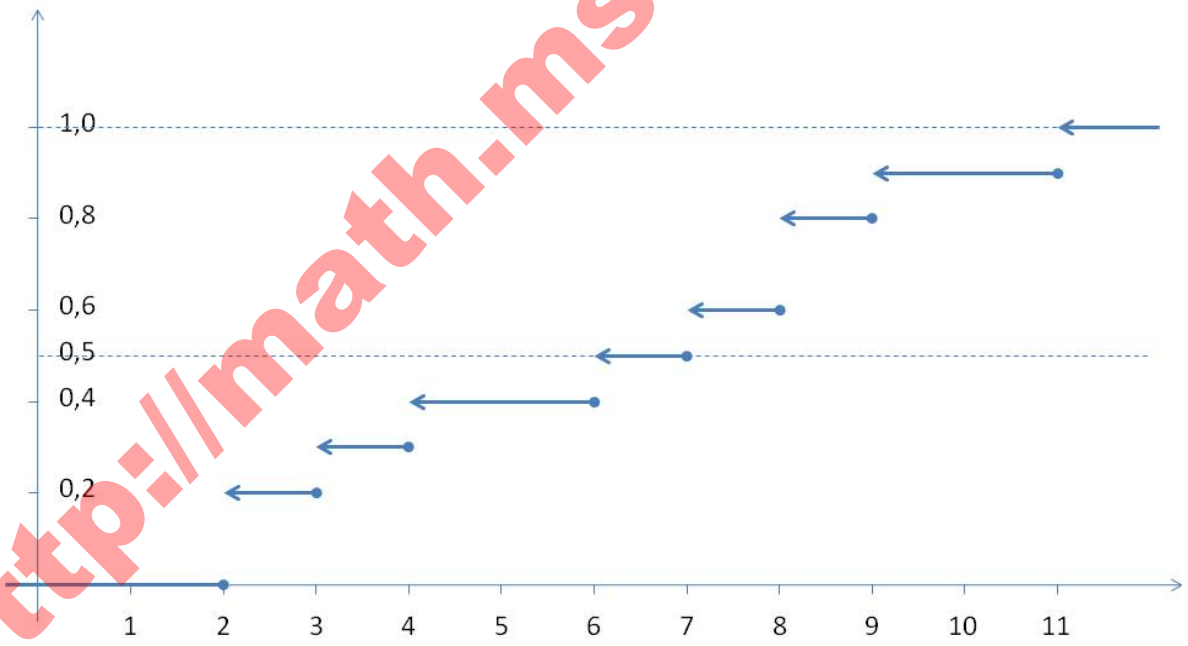
Вычисление значения $F(a)$ сильно упрощается, если числа набора упорядочены по возрастанию. Пусть N – количество различных чисел в наборе, а $y_1 < \ldots < y_N$ – сами эти числа, занумерованные в порядке возрастания. Пусть, далее, натуральное число t_i указывает, сколько раз число y_i встречается в наборе (так что вариационный ряд имеет вид (2.1)), а $f_i = \dfrac{t_i}{n}$ – относительная частота появления числа y_i. Отметим числа $y_1,\ldots,y_N$ на числовой оси. В результате мы получим $N+1$ непересекающихся промежутков: $\big(-\infty; y_1\big]$, $\big(y_1; y_2\big],\ldots,\big(y_{N-1}; y_N\big]$, $\big(y_N; +\infty\big)$. Если $a \in \big(-\infty; y_1\big]$, то $K_n\big(a\big) = 0$ и, соответственно, $F_n\big(a\big) = 0$. Если $a \in \big(y_i; y_{i+1}\big]$, $i = 1,\ldots,N-1$, то $K_n(a) = t_1 + \ldots + t_i$ и, соответственно, $F_n(a) = f_1 + \ldots + f_i$. Если $a \in \big(y_N; +\infty\big)$, то $K_n(a) = n$ и, соответственно, $F_n(a) = 1$.

Чтобы проиллюстрировать эти обозначения и понятие функции распределения, рассмотрим набор (2; 2; 3; 4; 6; 7; 8; 8; 9; 11) из n=10 чисел. В нём N=8 различных чисел: y_1=2; y_2=3; y_3=4; y_4=6; y_5=7; y_6=8; y_7=9; y_8=11.

Если, например, $a=5$, то ровно 4 числа из этого набора (это числа 2, 2, 3, 4) удовлетворяют неравенству $x<5$. Поэтому $K_{10}(5)=4$, а $F_{10}(5)=0,4$ (нижний индекс 10 указывает, как мы условились, общее количество чисел в наборе). Если a будет уменьшаться, оставаясь больше 4, то количество чисел набора, меньших a не будет меняться. Соответственно, функция $F_{10}(a)$ всё время равна 0,4. Как только a станет равным 4, число $K_{10}(a)$ уменьшится на 1. Соответственно, $F_{10}(a)$ скачком изменит своё значение на 0,3. Если a будет увеличиваться от a=5 до a=6, оставаясь меньше 6, то количество чисел набора, меньших a, не будет меняться. При a=6 это число всё ещё равно 4. Соответственно, функция $F_{10}(a)$ тождественно равна 0,4 при $a\in(4;6]$. Но как только параметр a станет больше 6 функция $F_{10}(a)$ вырастет до 0,5 и т.д.

График функции $F_{10}(a)$ для этого набора изображён на рис.9.1.

Рис.9.1



Этот график (и это типично для графика функции распределения любого числового набора) является *кусочно-постоянным* и *разрывным*. Он состоит из частей горизонтальных прямых: двух бесконечных лучей (левый луч идёт на высоте 0, а правый – на высоте 1) и конечных отрезков. Стрелки слева означают, что левые граничные точки отрезков и вершина правого луча не входят в график, а жирные точки справа подчёркивают, что правые граничные

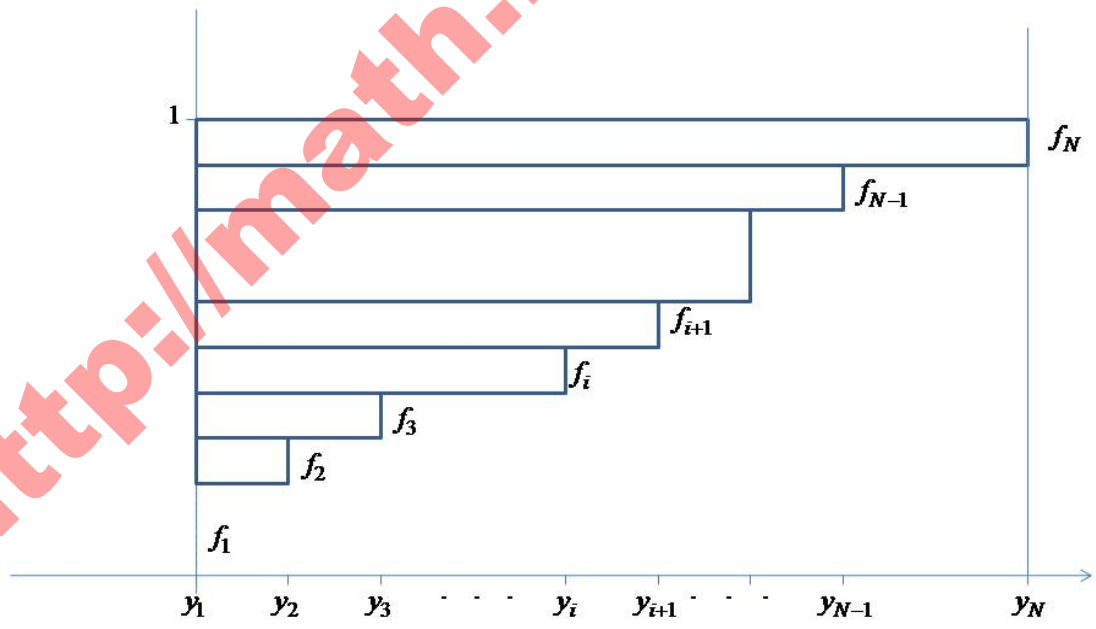
точки отрезков и вершина левого луча входят в график. Таким образом, график разрывен справа и непрерывен слева. Точки разрыва соответствуют точкам $y_1,\ldots,y_N$ на числовой оси, где расположены числа анализируемого набора. Величина скачка в точке y_i равна относительной частоте f_i появления этого числа в наборе.

9.3 Функция распределения числового набора и его среднее значение

Теорема 9.1. *Среднее значение M_X любого набора равно сумме наименьшего числа набора, $y_1=\min(x_1,\ldots,x_n)$, и площади фигуры Φ_n, ограниченной снизу графиком функции распределения $F_n(a\,|\,x_1,\ldots,x_n)$, сверху – горизонтальной прямой, проведённой на высоте 1, слева – вертикальной прямой, проведённой через точку y_1, справа – вертикальной прямой, проведённой через точку $y_N=\max(x_1,\ldots,x_n)$.*

Доказательство. Разобьём Φ_n на $N-1$ прямоугольников как показано на рис.9.2 (*i*-й прямоугольник имеет основание $y_{i+1}-y_1$ и высоту f_{i+1}).

Рис.9.2



Значит, площадь Φ_n равна $(y_2-y_1)\cdot f_2+(y_3-y_1)\cdot f_3+\ldots+(y_N-y_1)\cdot f_N$. После раскрытия скобок и перегруппировки членов мы получим:

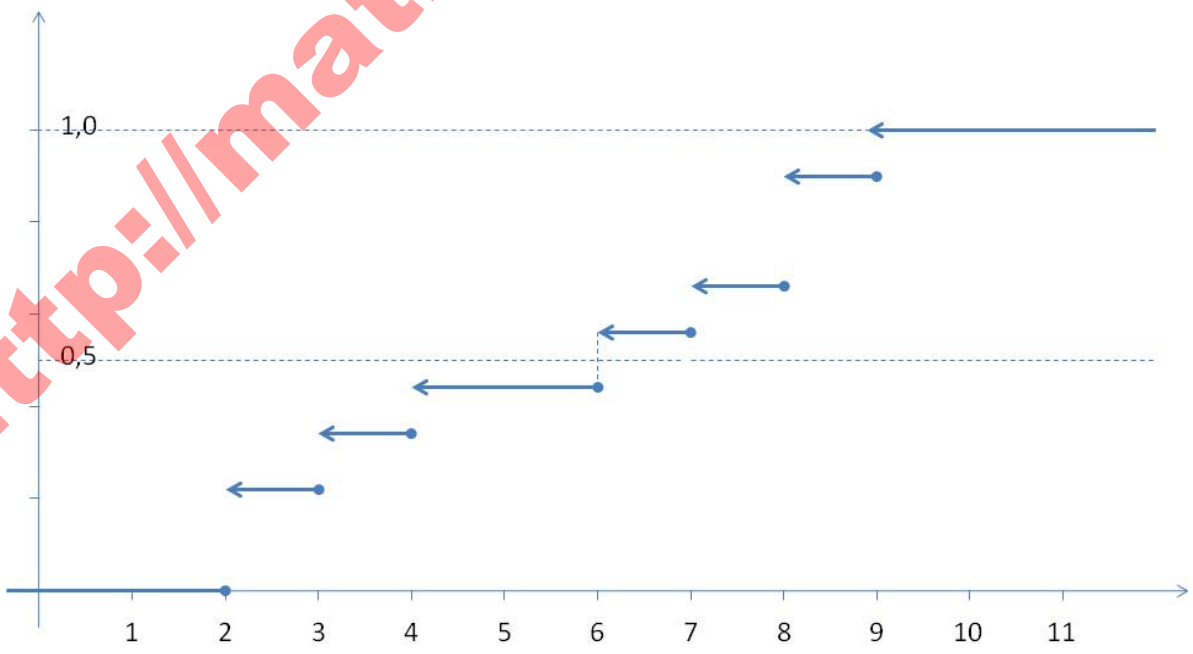
$y_2 \cdot f_2 + \ldots + y_N \cdot f_N - y_1 \cdot (f_2 + \ldots + f_N)$. Поскольку $f_2 + \ldots + f_N = 1 - f_1$, это выражение можно записать как $y_1 \cdot f_1 + \ldots + y_N \cdot f_N - y_1 = M_X - y_1$, что равносильно доказываемому утверждению.

9.4 Функция распределения набора и его медиана

Рассмотрим набор (2, 2, 3, 4, 6, 7, 8, 8, 9, 11); его функция распределения изображена на рис.9.1. Поскольку он состоит из чётного количества чисел (n=10), его медиана μ вычисляется как среднее арифметическое членов набора с номерами $k = n/2 = 5$ и $k+1 = n/2+1 = 6$ (напомним, что мы рассматриваем только наборы, упорядоченные по возрастанию) или, что то же самое, как середина отрезка $[x_5; x_6] = [6;7]$: $\mu = (x_5 + x_6)/2 = 6{,}5$. Отрезку $[x_5; x_6] = [6;7]$ на графике функции распределения соответствует горизонтальный отрезок, проведённый на высоте 0,5 (см. рис.9.1). Абсцисса середины этого отрезка равна медиане набора.

Рассмотрим теперь набор (2, 2, 3, 4, 6, 7, 8, 8, 9) из n=9 членов. Поскольку он состоит из нечётного количества чисел, его медиана μ – это член набора с номером $k = (n+1)/2 = 5$: $\mu = x_5 = 6$. График функции распределения этого набора изображён на рис.9.3.

Рис.9.3



На рис. 9.3 проведём горизонтальную прямую на высоте 0,5. Эта линия пройдёт через вертикальный отрезок, соединяющий конец одного из горизонтальных отрезков, составляющих график функции распределения, с началом следующего. Абсцисса точек этого отрезка как раз равна медиане набора.

Эти результаты вовсе не случайны.

Теорема 9.2. *Для произвольного набора* $X = [x_1, \ldots, x_n]$ *медиану можно вычислить следующим образом:*

- *нарисовать график функции распределения набора и на этом же рисунке провести горизонтальную прямую на высоте 0,5;*
- *если эта прямая пройдёт по одному из звеньев графика функции распределения (как на рис.9.1), то медиана равна абсциссе середины этого звена;*
- *если же эта прямая пройдёт через вертикальный отрезок, соединяющий два соседних звена графика функции распределения (как на рис.9.3), то медиана равна абсциссе точек этого вертикального отрезка.*

Доказательство. Вернёмся к доказательству Теоремы 2.2, где мы анализировали поведение функции

$$g(a) = |x_1 - a| + \ldots + |x_n - a| = t_1 |y_1 - a| + \ldots + t_N |y_N - a|.$$

При этом мы выделили два случая.

Случай 1 – для некоторого l верно равенство: $n = 2(t_1 + \ldots + t_l)$. Но $t_1 + \ldots + t_l$ – это значение функции $K_n(a)$ на промежутке $[y_l; y_{l+1})$. Поэтому Случай 1 равносилен тому, что на промежутке $[y_l; y_{l+1})$ функция распределения принимает значение 0,5. Иначе говоря, прямая $y = 0{,}5$ пройдёт по звену графика функции распределения, соответствующему промежутку $[y_l; y_{l+1})$. Как мы установили при доказательстве Теоремы 2.2, в этом случае функция $g(a)$ имеет наименьшее значение, которое достигается в бесконечном числе точек отрезка $[y_l; y_{l+1}]$, а медиана равна среднему арифметическому концов этого отрезка или, что то же самое, абсциссе середины соответствующего звена графика функции распределения.

Случай 2 – для некоторого l верны неравенства: $-n + 2(t_1 + \ldots + t_{l-1}) < 0$, $-n + 2(t_1 + \ldots + t_l) > 0$. Эти неравенства равносильны тому, что на промежутке $[y_{l-1}; y_l)$ функция распределения меньше 0,5, а на промежутке $[y_l; y_{l+1})$ больше 0,5. Это означает, что прямая $y = 0{,}5$ пройдёт через вертикальный отрезок, соединяющий правый конец звена, соответствующего промежутку $[y_{l-1}; y_l)$, и левый конец звена, соответствующего промежутку $[y_l; y_{l+1})$. Абсцисса точек этого вертикального отрезка равна y_l. Как мы

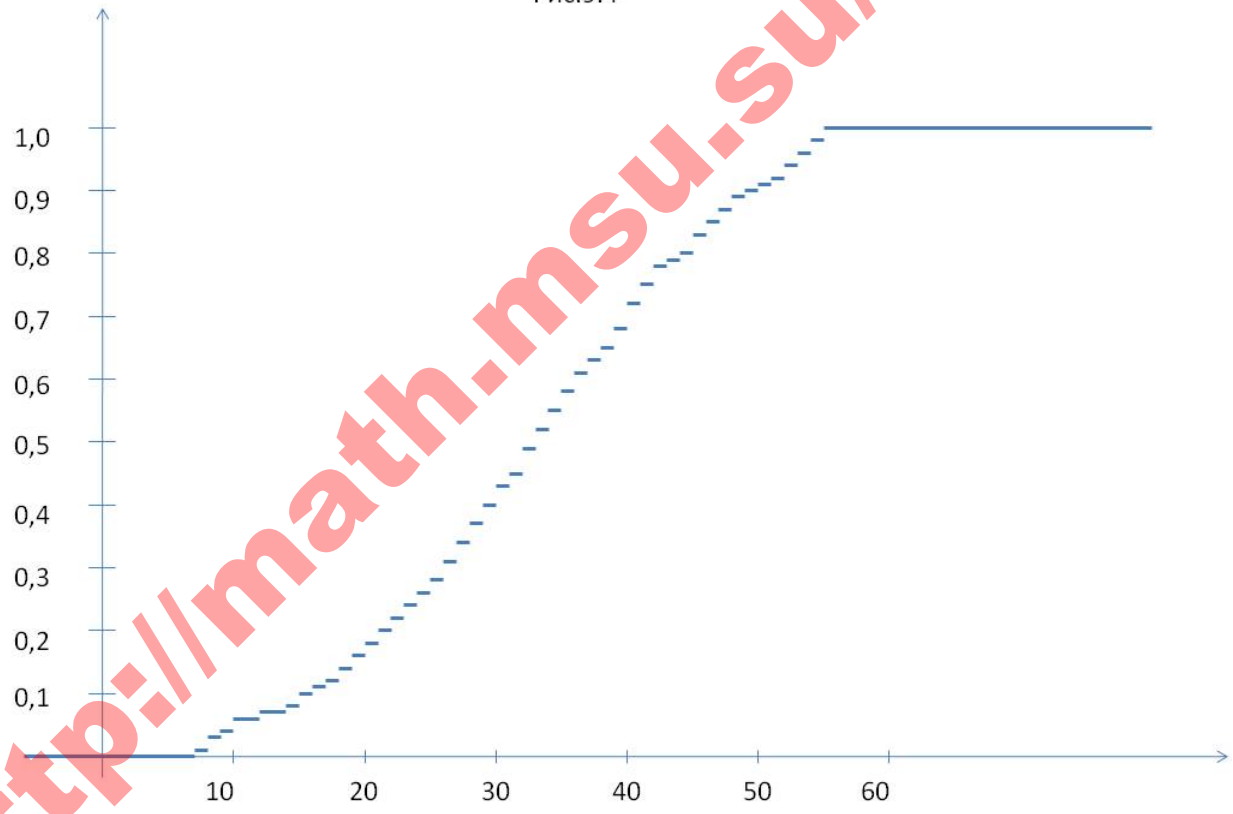
установили при доказательстве Теоремы 2.2, в этом случае функция g(a) имеет наименьшее значение, которое достигается в точке y_l, а медиана равна y_l.□

9.5 Функция распределения для больших наборов значений непрерывной переменной

Рассмотрим набор из 100 чисел, введённый в начале пункта 5.1; числа набора – это время T (в минутах), которое 100 выбранных наугад учеников гипотетической школы тратят на дорогу в школу.

График функции распределения этого набора изображён на рис.9.4.

Рис.9.4



Проанализируем этот рисунок и саму ситуацию подробнее. В ходе этого гипотетического статистического исследования измерялись затраты времени 100 школьников на дорогу до школы (в некоторый день). В рамках нашего исследования эта совокупность является генеральной.

Хотя значения переменной T являются целыми числами, при точном измерении времени теоретически её значением может быть любое положительное действительное число (из некоторого промежутка); целочисленность значений переменной T связана, конечно, с неточностью измерения времени по циферблату часов или электронному хронометру и с

округлением времени до целых минут. Поэтому переменная T является непрерывной. Но даже если бы мы считали нашу переменную дискретной, разница между 42мин. и 43мин. настолько мала, что с практической точки зрения вряд ли разумно считать эти времена различными. Если моменты времени, измеренные с точностью до 1 мин., отмечать точками на числовой оси, то даже при выборе единицы масштаба 1см=10 мин. разница между соседними точками (величиной 1мм) будет настолько мала, что эти дискретные точки практически сольются в непрерывную линию. Если же время измерять с точностью до 1 сек., то при той же единице масштаба разницу между соседними точками (величиной 1/60 мм) вообще невозможно не то, что заметить без увеличительного стекла, но даже и отобразить на листе бумаги. Соответственно, и график функции распределения $F_n(x)$, как хорошо видно на рис.9.4, практически выглядит как график некоторой непрерывной функции $y = F_n^*(x)$, которая монотонно возрастает от значения 0 при x=7 до значения 1 при x=56. Функцию $F_n^*(x)$ мы будем называть *непрерывным приближением* функции распределения $F_n(x)$

Для больших наборов значений непрерывной переменной график непрерывного приближения часто можно построить проще и быстрее, чем график эмпирической функции распределения $F_n(x)$. Это, в свою очередь, позволяет быстро решать все задачи, решаемые с помощью функции распределения. Например, долю чисел, попадающих на промежуток $[a;b)$, можно приближённо оценить как $F_n^*(b) - F_n^*(a)$. Действительно, доля таких чисел равна (в точности) $F_n(b) - F_n(a)$. Поскольку $F_n(x) \approx F_n^*(x)$, эта доля приблизительно равна $F_n^*(b) - F_n^*(a)$.

Для дальнейшего отметим важное следствие непрерывности функции $y = F_n^*(x)$, если мы используем её как точную функцию распределения. Пусть x_0 – некоторое выделенное число. Подсчитаем долю f_{x_0} значений переменной, которые в точности равны x_0. Для этого рассмотрим промежуток $\left[x_0; x_0 + \varepsilon\right)$, где ε – маленькое число. Ясно, что на этот промежуток попадает не меньше чисел, чем в точку x_0. Поэтому верно неравенство $0 \le f_{x_0} \le F_n^*(x_0 + \varepsilon) - F_n^*(x_0)$. Если мы начнём неограниченно уменьшать число ε, то, в силу непрерывности функции $y = F_n^*(x)$, разность $F_n^*(x_0 + \varepsilon) - F_n^*(x_0)$ будет стремиться к нулю. Следовательно, число f_{x_0} обязано равняться 0. Иначе говоря, предположение о непрерывности функции распределения означает, что мы игнорируем любое конкретное значение анализируемой переменной, обращая внимание лишь на классы значений, попадающих в тот или иной промежуток. Этот вывод вполне соответствует здравому смыслу: нам совершенно безразлично, сколько учеников тратят на дорогу, например, 42мин. 35,5сек.; важно, сколько учеников тратят на дорогу слишком много времени, скажем, больше, чем 45 мин., и т.д.

Непрерывное приближение $y = F_n^*(x)$ разрывной кусочно-постоянной эмпирической функции распределения должно обладать некоторыми общими свойствами.

Прежде всего, функция $y = F_n^*(x)$ должна быть строго возрастающей (в области возможных значений переменной). Действительно, для любых чисел a и b мы оцениваем долю чисел, попадающих на промежуток $[a;b)$, как разность $F_n^*(b) - F_n^*(a)$. Поскольку эта доля является неотрицательным числом, верно неравенство $F_n^*(a) \le F_n^*(b)$. Если бы было верно равенство $F_n^*(a) = F_n^*(b)$, то на промежуток $[a;b)$ вообще не попадали бы значения непрерывной переменной. Но непрерывная переменная характеризуется тем, что теоретически может принять любое значение.

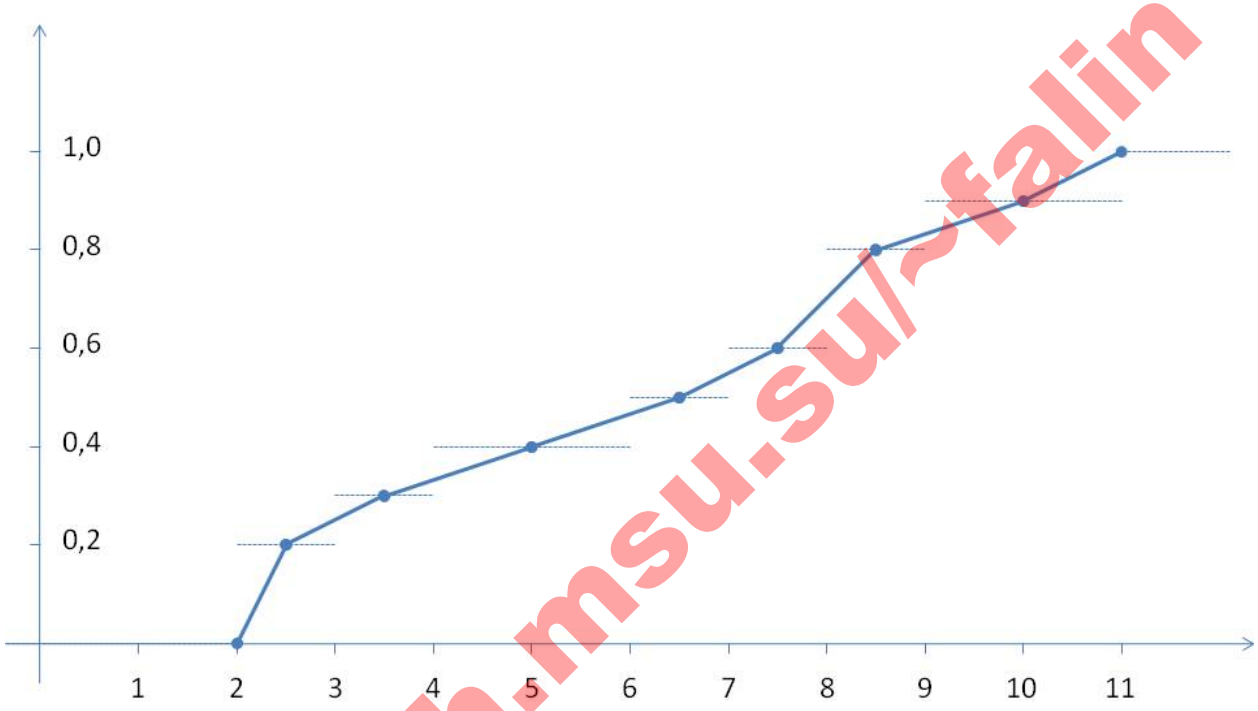
Если значения анализируемой переменной расположены на промежутке $[a;b]$, то $F_n^*(a) = 0$, $F_n^*(b) = 1$. Действительно, $F_n^*(a)$ – это доля значений, меньших, чем a. Но так как все значения лежат на $[a;b]$, таких значений нет, т.е. $F_n^*(a) = 0$. Далее, $F_n^*(b)$ – это доля значений, меньших, чем b. Так как все значения лежат на $[a;b]$, доля таких значений равна $1 - f_b = 1$. Если значения анализируемой переменной не ограничены сверху, т.е. она может, по крайней мере теоретически, принимать любое положительное значение, то условие $F_n^*(b) = 1$ заменяется условием $\lim\limits_{b \to +\infty} F_n^*(b) = 1$, которое обычно записывается как $F_n^*(+\infty) = 1$. Если значения анализируемой переменной не ограничены снизу, т.е. она может, по крайней мере теоретически, принимать любое отрицательное значение, то условие $F_n^*(a) = 0$ заменяется условием $\lim\limits_{a \to -\infty} F_n^*(a) = 0$, которое обычно записывается как $F_n^*(-\infty) = 0$.

График непрерывного приближения $y = F_n^*(x)$ разрывного кусочно-постоянного графика эмпирической функции распределения конкретного числового набора можно построить, просто проведя на глаз гладкую кривую, обладающую описанными выше свойствами, которая бы *приблизительно* совпадала с исходным кусочно-постоянным графиком функции распределения. Однако в статистике обычно применяют несколько стандартных приёмов.

Первый из них заключается в следующем. Выберем внутри каждого горизонтального участка функции распределения $F_n(x)$ по одной точке и соединим их наклонными отрезками. Как выбирать эти промежуточные точки – неважно. В частности, их абсциссы не обязаны быть равноотстоящими. Единственное, на что следует обращать внимание – соответствие общего вида графика исходной функции распределения и её приближения. Рассмотрим, например, функцию распределения $F_{10}(x)$ набора (2, 2, 3, 4, 6, 7, 8, 8, 9, 11), которая изображена на рис.9.1 и выберем для построения непрерывного приближения средние точки горизонтальных отрезков, образующих график

$F_{10}(x)$; на левом и правом горизонтальных лучах возьмём самую правую и самую левую точку соответственно (т.е., в сущности, просто наименьшее и наибольшее числа набора). Полученный непрерывный график изображён на рис.9.5 (тонкой штрихованной линией изображён исходный график функции распределения).

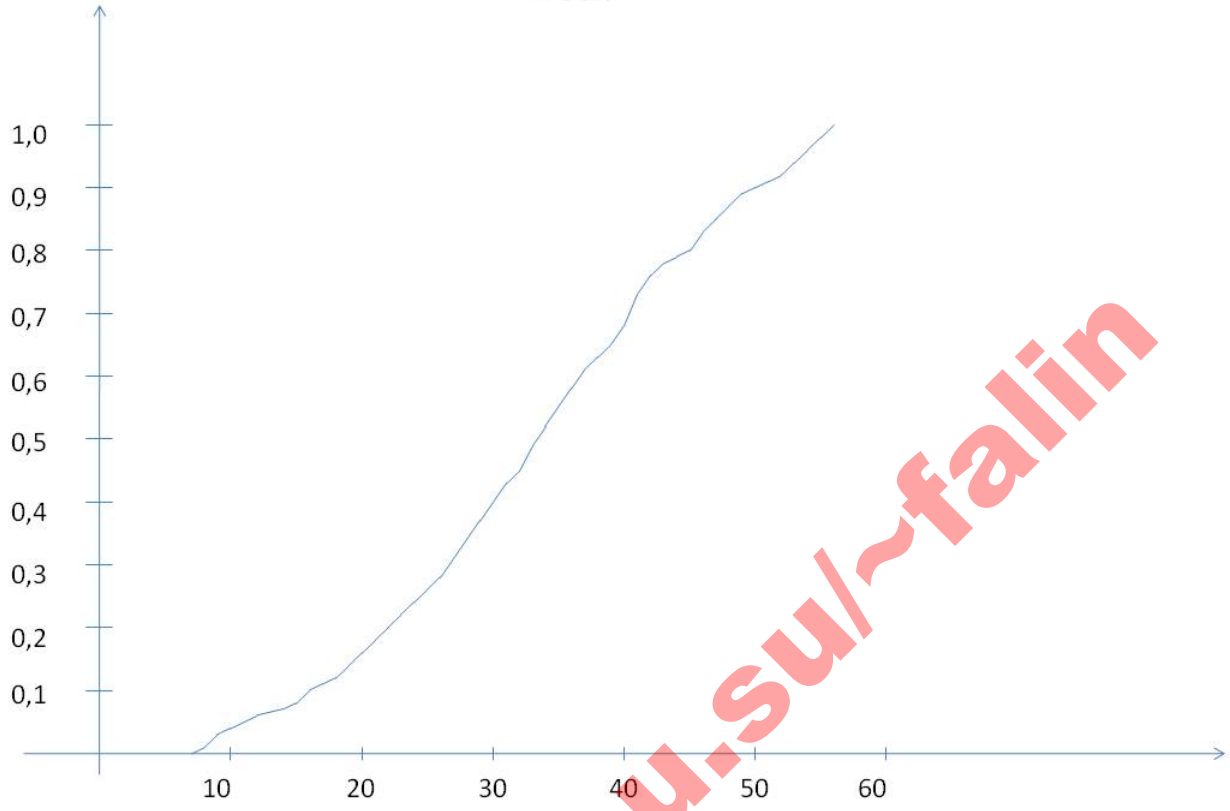
Рис.9.5



Если эту процедуру применить к функции распределения, изображённой на рис.9.4, то мы получим практически гладкую (без изломов) линию, изображённую на рис. 9.6.

Конечно, описанный способ субъективен (непрерывное приближение зависит от выбора промежуточных точек) и достаточно сложен (по сути дела получить это непрерывное приближение можно лишь построив саму функцию распределения). Однако он вполне применим при решении несложных задач на практике и, самое главное, позволяет понять саму концепцию непрерывного приближения функции распределения непрерывной переменной.

Рис.9.6



Действительно просто построить непрерывное приближение функции распределения непрерывной переменной, если исходные «сырые» данные сгруппированы в классы (интервалы группировки) и указано, сколько чисел попадает в тот или иной интервал группировки. Именно с такой ситуацией мы обычно имеем дело при статистическом исследовании непрерывных переменных.

Для рассматриваемого примера с временем на дорогу до школы исходные данные можно представить следующей таблицей:

Время (мин)	Количество чисел в классе, n_i	Относительная частота, $f_i = \frac{n_i}{n}$
0–9	4	0,04
10–19	12	0,12
20–29	24	0,24
30–39	28	0,28
40–49	22	0,22
50–59	10	0,10
Всего	**100**	**1**

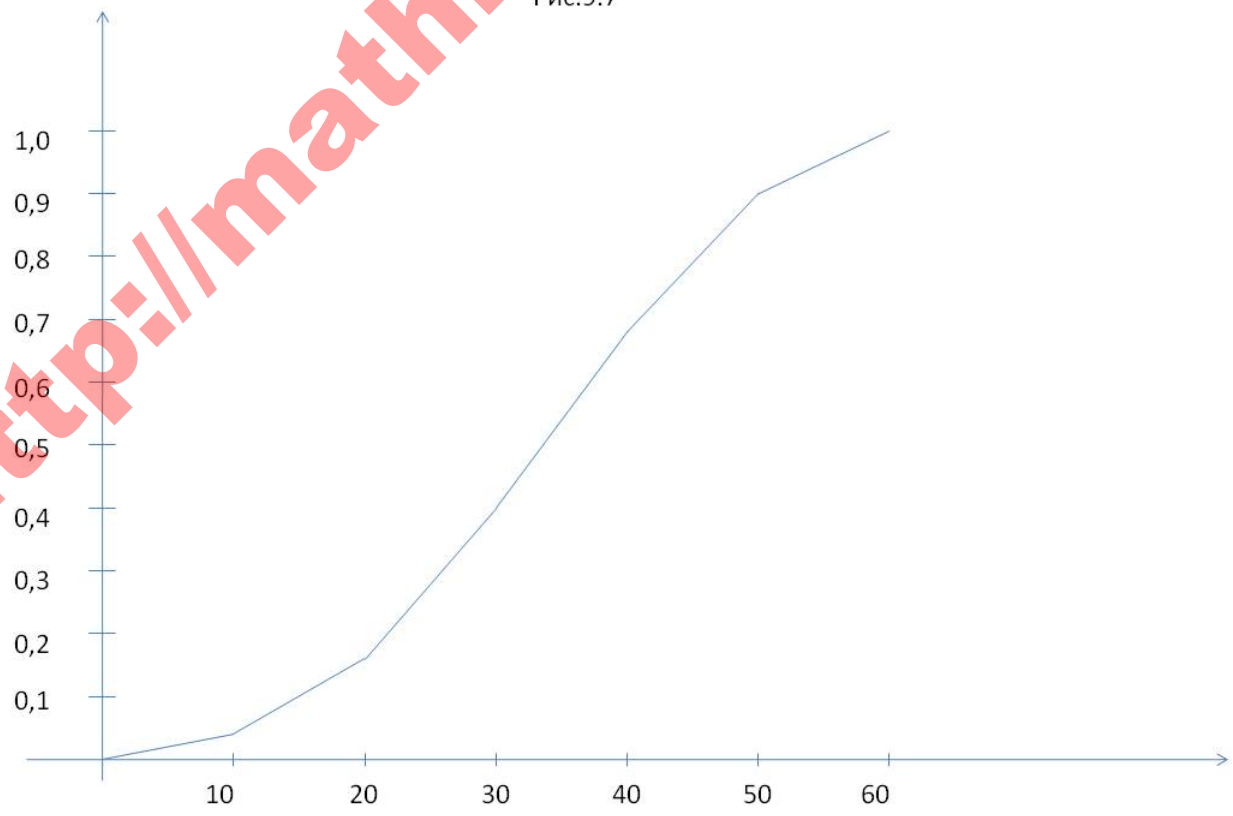
К сожалению, мы не знаем, как собиралась анализируемые данные – кто-то уже собрал их за нас (напомним, что такие данные называют *вторичными*). В частности, мы не знаем использовавшуюся процедуру округления. Для

определённости будем считать, что время округлялось до целых минут отбрасыванием секунд. Поэтому, например, запись 10–19 означает, что точное значение времени на дорогу до школы удовлетворяет двойному неравенству $10 \le T < 20$. Таким образом, результаты измерений сгруппированы по интервалам длиной 10 минут.

Для построения непрерывного приближения функции распределения возьмём в качестве узловых точек точки графика точной функции распределения, соответствующие правым границам интервалов группировки.

Если a=10, то количество чисел анализируемого набора меньших, чем a, равно 4; если a=20, то количество чисел анализируемого набора меньших, чем a, равно 16; если a=30, то количество чисел анализируемого набора меньших, чем a, равно 40 и т.д. Если a=60, то количество чисел анализируемого набора меньших, чем a, равно 100. Таким образом, графику функции распределения $F_{100}(x)$ принадлежат точки (10;0,04), (20;0,16), (30;0,40), (40;0,68), (50;0,90), (60;1). Кроме того, в число опорных точек добавим точку соответствующую левому концу a=0 первого интервала группировки; в ней, очевидно, значение функции распределения равно 0. Мы считаем, что график непрерывного приближения функции распределения, $F^*_{100}(x)$ (верхний индекс * указывает, что мы рассматриваем непрерывную функцию, приближающую разрывную функцию $F_{100}(x)$), совпадает с функцией $F_{100}(x)$ в этих точках. Соединяя все опорные точки отрезками, мы получим линию, изображённую на рис.9.7.

Рис.9.7



Сопоставляя приближения, изображённые на рис. 9.6 и 9.7, нетрудно заметить, что второй график ведёт себя гораздо регулярнее – на первом графике отчётливо видны небольшие колебания линии. Эти колебания отражают малосущественные (в силу каких-то случайных факторов) отличия в значениях переменной T для отдельных объектов нашей генеральной совокупности от некоторой средней тенденции в характере распределения переменной T, так что второй график, видимо, лучше характеризует нашу генеральную совокупность как целое.

Рассмотрим теперь произвольный набор из n значений непрерывной переменной и предположим, что эти значения сгруппированы в k классов (интервалов группировки): $[c_0;c_1),[c_1;c_2),\ldots,[c_{k-1};c_k)$. Обратим внимание на то, что мы не предполагаем равенство длин этих интервалов. Для асимметричных распределений с длинными «хвостами» (а на практике часто встречаются именно такие распределения) рассматривать, как это часто рекомендуется в учебниках по статистике, интервалы группировки равной длины неразумно. Пусть n_i – количество значений, попавших в i-й класс, $f_i = \dfrac{n_i}{n}$ – относительная частота попадания в i-й класс. На каждом интервале группировки $\left[c_{i-1};c_i\right)$ мы знаем значение непрерывного приближения $F_n^*(x)$ функции распределения $F_n(x)$ в граничных точках c_{i-1} и c_i – это сумма относительных частот для всех предшествующих интервалов:

$$F_n^*\left(c_{i-1}\right) = f_1 + \ldots + f_{i-1},\ F_n^*\left(c_i\right) = f_1 + \ldots + f_i = F_n^*\left(c_{i-1}\right) + f_i.$$

Вспоминая уравнение прямой, проходящей через две заданные точки $\left(c_{i-1};F_n^*\left(c_{i-1}\right)\right)$ и $\left(c_i;F_n^*\left(c_i\right)\right)$, мы получим, что на интервале группировки $\left[c_{i-1};c_i\right)$ значение непрерывного приближения функции распределения даётся формулой

$$F_n^*(x) = f_1 + \ldots + f_{i-1} + f_i \frac{x - c_{i-1}}{c_i - c_{i-1}}.$$

Отсюда, в частности следует, что производная непрерывного приближения функции распределения во внутренних точках интервала $\left[c_{i-1};c_i\right)$ равна $\dfrac{f_i}{\Delta_i}$, где $\Delta_i = c_i - c_{i-1}$ – длина этого интервала (в граничных точках производная, вообще говоря, не существует, т.к. угловые коэффициенты звеньев, смыкающихся в граничных точках, вообще говоря, не совпадают). Мы знаем, что формула

$$h_n(x) = \frac{f_i}{\Delta_i},\ \text{если } c_{i-1} < x < c_i$$

определяет гистограмму, построенную по сгруппированным данным. Таким образом, производная от нашего непрерывного приближения функции распределения будет гистограммой.

Применим график непрерывного приближения функции распределения по сгруппированным данным, изображённый на рис.9.7, для приближённого вычисления статистических характеристик переменной T.

Чтобы найти медиану, нужно провести горизонтальную прямую на высоте 0,5 до пересечения с графиком и найти абсциссу точки пересечения. Ясно, что линия y=0,5 пересечёт отрезок, соединяющий точки (30;0,40) и (40;0,68). Поскольку уравнение прямой, соединяющей эти точки, есть $y = 0{,}028x - 0{,}44$, для вычисления медианы μ_T имеем уравнение: $0{,}5 = 0{,}028\mu_T - 0{,}44$, откуда $\mu_T = \frac{940}{28} \approx 33{,}57$ (визуально график даёт значение $\mu_T \approx 34$). Точное значение медианы, вычисленное (по определению) как среднее арифметическое 50-го и 51-го в порядке возрастания чисел набора, равно 33, что практически не отличается от значения, полученного с помощью непрерывного приближения по сгруппированным данным.

Непосредственно из исходного набора можно подсчитать, что количество школьников, которые добираются до школы по меньшей мере 45 минут равно 20. С помощью непрерывного приближения это количество можно оценить как $100 \cdot \left(1 - F^*_{100}(45)\right)$. Поскольку точка 45 лежит на интервале группировки $40 \le T < 50$, а уравнение прямой, соединяющий точки (40;0,68) и (50;0,90), есть $y = 0{,}022x - 0{,}2$, легко подсчитать, что $F^*_{100}(45) = 0{,}79$ (визуально график даёт то же значение). Поэтому непрерывное приближение говорит, что 45 минут или больше тратит примерно 21 школьник. Опять, разница между точным значением и значением, полученным с помощью непрерывного приближения по сгруппированным данным, практически отсутствует.

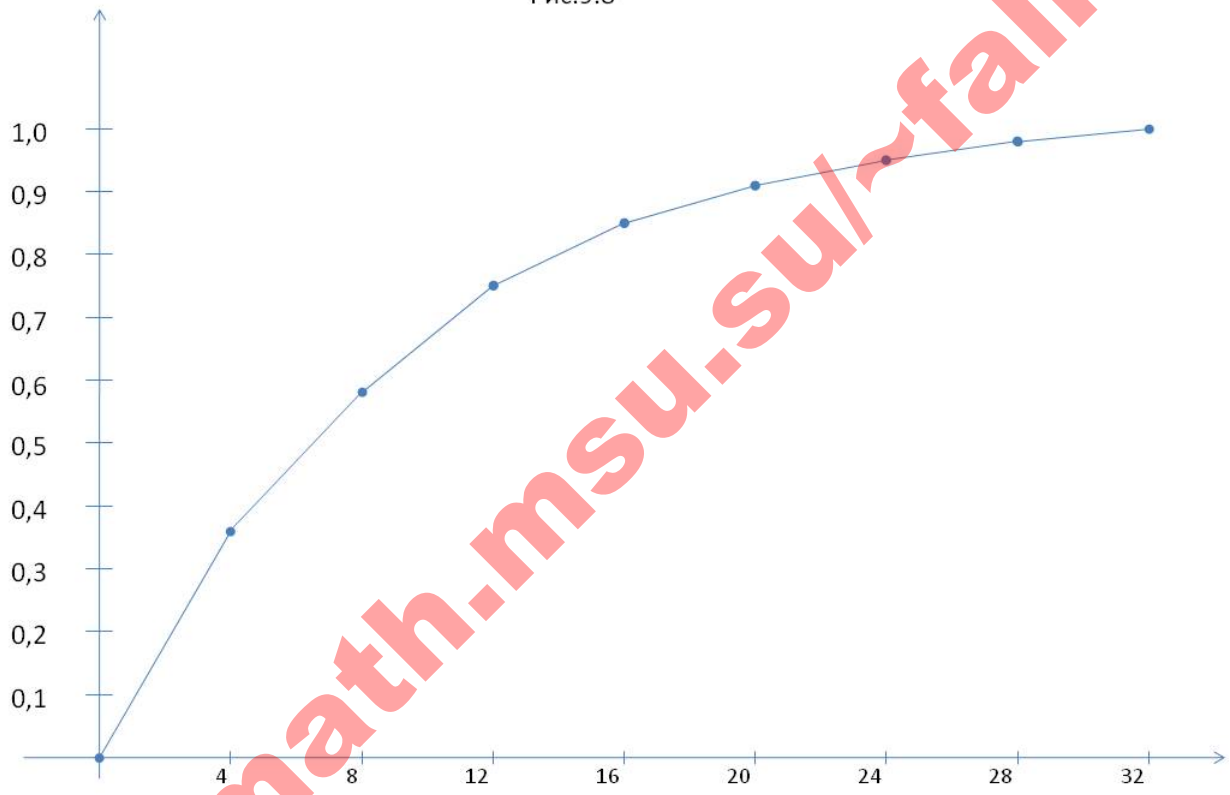
Следующий пример взят из [5], стр. 13. Для определения оптимального числа причалов изучалось распределение промежутков времени T между последовательными прибытиями судов дальнего плавания в один из портов. Результаты приведены в следующей таблице.

T (час)	0–4	4–8	8–12	12–16	16–20	20–24	24–28	28–32	**Всего**
Число случаев	67	43	30	18	11	7	5	4	185
Частоты	0,360	0,231	0,161	0,097	0,059	0,038	0,028	0,026	1,000

Запись 4–8 означает, что точное значение измеряемой переменной T удовлетворяет двойному неравенству $4(\text{час}) \le T < 8(\text{час})$. Таким образом, результаты измерений сгруппированы по интервалам длины 4 час.

Выбирая, как и в предыдущем примере, для построения непрерывного приближения функции распределения в качестве узловых точек правые границы интервалов группировки, мы получим, что графику непрерывного приближения функции распределения принадлежат точки (4;0,360), (8;0, 0,591), (12; 0,752), (16; 0,849), (20; 0,908), (24; 0,946), (28;974), (32;1). Кроме того, в число опорных точек добавим точку соответствующую левому концу a=0 первого интервала группировки; в ней, очевидно, значение функции распределения равно 0. Соединяя все опорные точки отрезками, мы получим линию, изображённую на рис.9.8.

Рис.9.8



Литература

Учебники и пособия по статистике и теории вероятностей

1. Алгебра: Учеб. для учащихся 9 кл. с углубл. изучением математики/Н.Я.Виленкин, Г.С.Сурвилло, А.С.Симонов, А.И.Кудрявцев; под ред. Н.Я.Виленкина. – 8-е изд. – М.: Просвещение, 2007.

2. Алгебра и начала математического анализа. 11 кл.: учеб. для учащихся общеобразоват. учреждений (профильный уровень)/Н.Я.Виленкин, О.С.Ивашев-Мусатов, С.И.Шварцбурд. – 14-е изд. – М.: Мнемозина, 2009.

3. Математика: алгебра. Функции. Анализ данных. Учебник для 9 кл. общеобразоват. учреждений./Г.В.Дорофеев, С.Б.Суворова, Е.А.Бунимович и др; под ред. Г.В.Дорофеева. – М.: Просвещение, 2005.

4. Тюрин Ю.Н. и др. Теория вероятностей и статистика. М.: МЦНМО: АО «Московские учебники», 2004. – 256 с.

5. Б.В.Гнеденко. Беседы о математической статистике. Изд. 2-е, испр. – М.: Книжный дом «ЛИБРОКОМ», 2010.

6.Б.В.Гнеденко, А.Я.Хинчин. Элементарное введение в теорию вероятностей. Изд. 10-е, исправленное. М., УРСС, 2003.

7. Дж.Вайнберг, Дж.Шумекер. Статистика. М.: «Статистика», 1979.

8. S. Dobbs, J. Miller. Statistics 1. Cambridge University Press, 2009.

9. A.Ballard, S.Gill, et al. GCSE Statistics. Complete Revision and Practice. Coordination Group Publications Ltd, 2010.

10. S.Procter-Green, P.Winters. AQA GCSE Statistics. Nelson Thornes, 2009.

11. Дж.Тьюки. Анализ результатов наблюдений. Разведочный анализ. М.: Мир, 1981.

12. R.V.Hogg, S.A.Klugman. Loss Distributions. John Wiley & Sons, 1984.

Методические статьи по статистике и теории вероятностей

13. Е.А.Бунимович, В.А.Булычев и др. О теории вероятностей и статистике в школьном курсе. Математика в школе, 2009, №7, стр.3-13.

14. J. Freund, B. Perles. A New Look at Quartiles of Ungrouped Data. *The American Statistician*, Vol. 41, No. 3 (Aug., 1987), pp. 200-203.

15. R.J. Hyndman, Y. Fan. Sample Quantiles in Statistical Packages. *The American Statistician*, Vol. 50, No. 4 (Nov., 1996), pp. 361-365.

16. M. Pfannkuch. Comparing box plot distributions: a teacher's reasoning. *Statistics Education Research Journal*, 2006, 5(2), pp.27-45.

17. U. Kortenkamp, K. Rolka. Using technology in the teaching and learning of box plots. *Proceedings of CERME 6*, January 28th-February 1st 2009, Lyon France. INRP 2010, pp.1070-1080.

18. E. Langford. Quartiles in Elementary Statistics. *Journal of Statistics Education, 2006,* Volume 14, Number 3.

19. A.Bakker, R.Biehler, C.Konold. Should Young Students Learn About Box Plots? *Curricular Development in Statistics Education*, Sweden, 2004, pp.163-173.

20. R. McGill, J. W. Tukey, W.A. Larsen. Variations of Box Plots. *The American Statistician*, Vol. 32, No. 1. (Feb., 1978), pp. 12-16.

Британские школьные экзамены по статистике

21. Specification. Edexcel GCSE in Statistics (2ST01). Edexcel Limited, 2008.

22. GCSE Specification. Statistics (for certification 2011 onwards). AQA, 2008.

23. Sample Assessment Materials. Edexcel GCSE in Statistics (2ST01). February 2010.

24. AQA GCSE Statistics, Foundation Tier Specimen Paper, For First Teaching 2009.

25. AQA Mathematics 2012 Specification. GCSE Foundation Tier Specimen Paper, 2010.

26. Edexcel GCSE Statistics, Foundation Tier, Paper 1F, Wednesday 18 June 2008 – Morning.

27. Edexcel GCSE Statistics, Higher Tier, Paper 1H, Wednesday 18 June 2008 – Morning.