

Краснодарский университет МВД России

А. А. Хромых

**СТАТИСТИЧЕСКИЕ МЕТОДЫ АНАЛИЗА ДАННЫХ
ДЛЯ ПОДГОТОВКИ СПЕЦИАЛИСТОВ ТЫЛОВОГО ОБЕСПЕЧЕНИЯ
ОРГАНОВ ВНУТРЕННИХ ДЕЛ**

Учебное пособие

Краснодар
2026

УДК 519.23
ББК 60.6
Х941

Одобрено
редакционно-издательским советом
Краснодарского университета
МВД России

Рецензенты:

А. Г. Карпика, кандидат технических наук, доцент (Ростовский юридический институт МВД России);

М. Б. Руденко, кандидат технических наук, профессор (Восточно-Сибирский институт МВД России).

Хромых А. А.

Х941 Статистические методы анализа данных для подготовки специалистов тылового обеспечения органов внутренних дел [Электронный ресурс] : учебное пособие / А. А. Хромых. – Электрон. дан. – Краснодар : Краснодарский университет МВД России, 2026. – 1 электрон. опт. диск.

ISBN 978-5-9266-2243-7

Рассматриваются ключевые статистические подходы, применимые для анализа и оптимизации процессов в тыловом обеспечении. Раскрываются основы статистики и методологии сбора данных. Приводятся современные методы регрессионного и корреляционного анализа с практическими примерами, ориентированными на логистику и управление запасами.

Для профессорско-преподавательского состава, курсантов, слушателей образовательных организаций МВД России и сотрудников органов внутренних дел Российской Федерации.

УДК 519.23
ББК 60.6

ISBN 978-5-9266-2243-7

© Краснодарский университет
МВД России, 2026
© Хромых А. А., 2026

Оглавление

1. Основы статистики и сбора данных	6
1.1. Основные понятия и термины статистики	6
1.1.1. Предмет и задачи статистики	7
1.1.2. Понятие выборочной совокупности	9
1.1.3. Способы отбора из генеральной совокупности	12
1.2. Методы сбора данных.....	16
1.2.1. Опросы, наблюдения, эксперименты.....	16
1.2.2. Ошибки выборки и их влияние на результаты анализа	18
1.3. Предварительная обработка данных	19
1.3.1. Пропуски, выбросы, нормализация данных.....	19
1.3.2. Группировка данных	20
1.4. Абсолютные и относительные величины	24
1.4.1. Абсолютные величины.....	24
1.4.2. Относительные величины	25
1.5. Статистические ряды распределения.....	27
1.6. Средние величины и показатели вариации	31
1.7. Визуализация данных	42
1.8. Практические примеры.....	46
1.8.1. Анализ затрат на транспортировку	46
1.8.2. Изучение распределения запасов на складе.....	48
1.9. Применение возможностей <i>Excel</i> для анализа выборочных совокупностей.....	49
1.9.1. Исследование выборочной совокупности с помощью электронных таблиц <i>Excel</i>	49
1.9.2. Вычисление числовых характеристик выборочной совокупности с помощью встроенных функций.....	55
1.9.3. Обработка статистических данных с помощью надстройки Анализ данных	57
2. Инференциальная статистика	62

2.1. Закон больших чисел и центральная предельная теорема.....	62
2.1.1. Закон больших чисел.....	62
2.1.2. Центральная предельная теорема.....	63
2.2. Статистические оценки параметров распределения	64
2.2.1. Оценки параметров	64
2.2.2. Точечные оценки математического ожидания и дисперсии	68
2.2.3. Понятие интервального оценивания параметров	70
2.2.4. Доверительные интервалы для параметров нормального распределения.....	72
2.3. Статистическая проверка гипотез	86
2.3.1. Статистическая гипотеза. Ошибки первого и второго рода.....	86
2.3.2. Проверка гипотезы о равенстве генеральной средней нормальной совокупности заданному числовому значению	90
2.3.3. Проверка гипотезы о равенстве генеральной дисперсии нормальной совокупности заданному числовому значению	94
2.3.4. Проверка гипотезы о равенстве двух средних нормальных генеральных совокупностей	96
2.3.5. Проверка гипотез о законе распределения.....	99
3. Регрессионный и корреляционный анализ	103
3.1. Понятие регрессионного и дисперсионного анализов.....	103
3.2. Построение регрессионной модели.....	104
3.3. Построение линейной модели парной регрессии с помощью надстройки Анализ данных.....	115
Заключение.....	120
Литература	122
<i>Приложение 1</i>	125
<i>Приложение 2</i>	126
<i>Приложение 3</i>	127

Введение

В современном мире эффективность работы органов внутренних дел во многом зависит от качественной организации тылового обеспечения, которое включает в себя управление материальными ресурсами, техническую поддержку, логистику и другие важные процессы. Надежное и своевременное снабжение силовых структур становится ключевым фактором обеспечения их боеспособности и оперативности.

В условиях огромного потока данных, поступающих из различных подразделений и систем, для квалифицированного анализа и принятия обоснованных управленческих решений необходимы современные статистические методы анализа данных. Они позволяют выявлять скрытые закономерности, прогнозировать потребности, оптимизировать процессы и минимизировать риски.

Данное учебное пособие предназначено для подготовки специалистов тылового обеспечения органов внутренних дел, обладающих теоретическими знаниями и практическими навыками применения статистических методов. В нем рассматриваются основы статистики, методы обработки и интерпретации данных, а также особенности анализа именно в контексте работы тыла МВД.

Освоение материала данного учебного пособия поможет будущим специалистам повысить качество планирования и контроля, оптимизировать ресурсное обеспечение, а также эффективно взаимодействовать внутри структур ведомства. Это, в конечном итоге, способствует укреплению безопасности и стабильности общества.

1. Основы статистики и сбора данных

Основы статистики и сбора данных включают изучение теоретических и практических основ сбора, обработки и анализа данных, что позволяет выявлять закономерности и делать выводы из массовых количественных данных.

1.1. Основные понятия и термины статистики

Слово «*статистика*» происходит от латинского термина *status*, который в Средние века означал «*политическое состояние*». Этот термин был введен в научный оборот немецким ученым Готфридом Ахенвалем (1719-1772) и первоначально обозначал «государствоведение».

Прежде чем принять современное значение, статистика прошла длительный путь развития, включавший три основных направления.

Первое направление основал профессор Ахенваль, который с 1746 года в Марбургском университете читал новый курс под названием «статистика», содержание которого заключалось в описании политического состояния государства. Это направление называют *описательным*.

Второе направление связано с экономической статистикой, которую заложил Вильям Пети, представитель английской школы политических арифметиков. Политические арифметики использовали числовые данные для описания состояния и динамики развития общества, раскрывая закономерности социальных явлений. Их исследования послужили основой для формирования современной статистической науки.

В первой половине XIX века появилось третье направление – статистико-математическое, представителями которого были А. Кетле, Ф. Гальтон, В. Стьюдент, Р. Фишер и другие. Они рассматривали теорию вероятностей в качестве фундаментальной основы статистики.

Развитие статистической науки, расширение сферы применения практических статистических исследований привели к изменению содержания самого понятия «статистика». В настоящее время термин «статистика» употребляется в трех значениях:

- отрасль практической деятельности по сбору, обработке, анализу и публикации массовых цифровых данных о самых различных явлениях и процессах общественной жизни;
- цифровые сведения, фактические данные, представляемые в отчетности предприятий, организаций, отраслей экономики, а также публикуемые в сборниках, периодической печати, которые являются результатом статистической работы;
- отрасль общественных наук, специальная научная дисциплина, изучаемая в учебных заведениях. Статистическая деятельность на профессиональном уровне реализуется в рамках государственной статистики (ее осуществляет Федеральная служба государственной статистики (ФСГС) и система ее учреждений, организованных по административно-территориальному признаку) и ведомственной статистики (на предприятиях, в ведомствах, министерствах и т. д.). Информация ФСГС публикуется в специальных печатных изданиях, а также в сети Интернет (<https://rosstat.gov.ru/>).

1.1.1. Предмет и задачи статистики

Статистика – отрасль знаний, наука, которая изучает количественную сторону массовых общественных явлений и процессов в неразрывной связи с их качественными характеристиками. Она исследует закономерности общественного развития, а также влияние природных, технических и производственных факторов на количественные отношения в обществе. *Предметом статистики* является изучение количественной стороны массовых явлений с учетом их содержания и условий места и времени.

Статистика является *мультидисциплиной*, так как она использует методы и принципы, заимствованные из других дисциплин. Так, в качестве теоретической базы для формирования статистической науки служат знания в области социологии и экономической теории. В рамках этих дисциплин происходит изучение законов общественных явлений. Статистика помогает произвести оценку масштаба того или иного явления, а также разработать систему методов для анализа и изучения. Статистика, несомненно, связана с матема-

тикой, так как для выявления закономерностей, оценки и анализа объекта исследования требуется ряд математических операций, методов и законов, а систематизация результатов находит отражения в виде графиков и таблиц.

Основные *цели статистики* включают формирование знаний и навыков по сбору, группировке, обработке и анализу статистических данных. *Задачи статистики* – построение статистических таблиц и графиков, расчет и интерпретация статистических показателей, способность делать обоснованные выводы на основе данных. Статистика способствует разработке теоретических моделей процессов и явлений, их исследованию и оценке.

Предмет статистики изучается посредством специфических методов, объединённых в особую методологию. Основой формирования и применения статистической методологии служит диалектический подход, подразумевающий рассмотрение общественных явлений и процессов в динамике развития, взаимосвязи и причинно-следственных отношениях.

Статистическая методология – это система приемов, способов и методов, направленных на изучение количественных закономерностей (повторяемость и порядок изменений в явлениях), проявляющихся в структуре, динамике и взаимосвязи социально-экономических явлений.

Статистические методы используются комплексно (системно). Это обусловлено сложностью процесса экономико-статистического исследования, состоящего из четырех основных стадий:

- 1) сбор первичной статистической информации;
- 2) предварительная обработка первичной информации;
- 3) анализ и интерпретация статистической информации;
- 4) моделирование и прогнозирование.

На первой стадии статистического исследования осуществляется сбор первичного статистического материала, проверяются его достоверность и полнота. С этой целью применяются методы сплошного и несплошного статистического наблюдения.

На второй стадии производятся предварительная обработка данных, подсчет групповых и общих итогов, расчет некоторых относительных показателей. Основной метод, используемый на данном этапе, – метод статистических группировок. В результате его реализации осуществляется переход от больших массивов статистических данных к компактным и удобным для анализа статистическим группировкам.

На третьей стадии проводится анализ статистической информации на основе расчета обобщающих статистических показателей: абсолютных, относительных, средних величин, вариации, структуры, взаимосвязи и динамики.

В процессе реализации четвертого этапа осуществляется моделирование взаимосвязей между социально-экономическими процессами и явлениями, строятся модели, отражающие основные тенденции динамики изучаемых показателей.

При изучении статистической информации широко применяются табличный и графический методы.

1.1.2. Понятие выборочной совокупности

Статистическая совокупность – это группа, состоящая из большого числа относительно однородных объектов, взятых вместе в известных границах пространства и времени (например, дети, появившиеся в течение какого-то года или месяца). Статистическая совокупность формируется по признакам, они делятся на *количественные* и *атрибутивные (качественные)*.

Признак называют *количественным*, если значения признака выражаются числами, например, возраст, рост, масса, объем, производительность труда, количество автомобилей и т.п.

Признак называют *качественным*, если его значения характеризуют некоторое свойство или состояние элементов совокупности. Например, профессия, квалификация, сорт, окраска, консистенция, форма, регион, пол человека и т.п.

Статистическая совокупность, состоящая из набора всех значений признака, интересующих исследователя, называется *генеральной совокупностью (популяцией)*. Другими словами, *генеральная совокупность* – вся совокупность, подлежащая изучению.

Для определенности генеральные совокупности обозначают прописными буквами латинского алфавита – $X, Y, Z, \dots$, а их значения (элементы) – $x_1, x_2, \dots, y_1, y_2, \dots$

Генеральная совокупность может быть *конечной* или *бесконечной* в зависимости от того, конечна или бесконечна совокупность составляющих ее объектов.

Безусловно, исследование всей генеральной совокупности позволяет получить наиболее полное и достоверное представление об изучаемых объектах. Однако в большинстве случаев это оказывается невозможным из-за различных ограничений – пространственных, временных, финансовых, а также из-за большого объема объектов или невозможности сохранить объект в процессе исследования. В таких ситуациях приходится обратиться к исследованию части совокупности – *выборки*, которая служит основанием для вывода о закономерностях всей генеральной совокупности.

Выборочной совокупностью (выборкой) называется совокупность случайно отобранных объектов из генеральной совокупности.

Пример. Все автомобили, выпускаемые с конвейера определенного завода за год – это генеральная совокупность. А 100 автомобилей, отобранных случайным образом с конвейера для проверки качества – это выборочная совокупность.

Пример. Если исследуется успеваемость по конкретному предмету, то все курсанты университета составляют генеральную совокупность. А 200 курсантов из разных факультетов университета, участвующих в исследовании по успеваемости, образуют выборочную совокупность.

Выборка называется *репрезентативной (представительной)*, если ее объекты достаточно хорошо отражают свойства генеральной совокупности.

Пример. Если нужно изучить отношение населения России к вопросам экологии, репрезентативная выборка будет включать людей из разных регионов, с разным возрастом, полом и уровнем дохода, пропорционально их доле в населении. Если в генеральной совокупности 52% женщин и 48% мужчин, то и в выборке будет примерно такое же соотношение, чтобы результаты можно было обобщить на всю страну.

Пример. Для оценки удовлетворенности работников в большой компании выборка формируется так, чтобы в ней были пропорционально представлены сотрудники из всех отделов, с разным стажем и уровнями ответственности. Например, если 30% сотрудников работают в отделе продаж, они должны составлять 30% выборки. Это позволит получить точную картину настроений всего коллектива.

Число N объектов генеральной совокупности и число n объектов выборочной совокупности называют *объемами* генеральной и выборочной совокупностей соответственно. При этом предполагают, что N значительно больше n .

Относительный показатель выборки – отношение объема выборочной совокупности к генеральной совокупности.

$$N_n = \frac{n}{N}.$$

Пример. При выборе из 100 единиц товара 30 имеют дефекты. Относительный показатель дефектных товаров будет:

$$N_n = \frac{30}{100} = 0,3 \text{ или } 30\%.$$

Это означает, что 30% от всей выборки оказались бракованными товарами.

Таким образом, относительный показатель позволяет сопоставить доли различных признаков между собой и выявить их структуру в выборочных данных, что важно для анализа и принятия решений.

1.1.3. Способы отбора из генеральной совокупности

На практике используются разнообразные способы выборки объектов из общей совокупности. Один из таких методов не предполагает разделения исходной совокупности на отдельные группы. Примером являются *простая случайная бесповторная* и *простая случайная повторная* выборки. Вторым методом предполагает предварительное деление генеральной совокупности на отдельные части, группы или серии, после чего отбор ведётся по определённым правилам. К этим способам относятся *типический*, *механический* и *серийный* отбор. Обычно, если не требуется предварительно оценивать однородность исследуемой совокупности, применяется простой случайный отбор, как повторный, так и бесповторный, при котором каждый элемент обладает одинаковой вероятностью попасть в выборку без предварительного деления на группы.

При *простой бесповторной выборке* выбранный объект исключается из генеральной совокупности и не может быть выбран снова.

Пример. Есть мешок с 100 одинаковыми карточками, каждую из которых пронумеровали с 1 до 100. Чтобы отобрать выборку из 10 карточек, вытягивают одну, записывают номер и больше не возвращают её в мешок. Затем вытягивают следующую карточку из оставшихся 99, и так далее, пока не соберут необходимую выборку из 10 карточек. Такой подход исключает повторение объектов в выборке и используется для точного представления генеральной совокупности без дублирования.

Пример. В микрорайоне города проживает 5000 семей. Для определения среднего размера семьи была выбрана случайная бесповторная выборка определенного числа семей (например, 200), при этом каждая семья могла попасть в выборку только один раз. Это обеспечивало репрезентативность и уникальность каждого объекта в выборке.

Пример. При телефонном опросе респондентов список номеров составляется так, чтобы каждый номер был выбран не более одного раза, исключая

повторные опросы одних и тех же лиц. Такая выборка считается бесповторной.

Пример. Из базы данных клиентов магазина случайным образом выбрали 300 клиентов для опроса об уровне удовлетворенности, при этом каждый клиент выбран один раз и не повторяется.

При *простой повторной выборке* выбранный объект после отбора возвращается обратно в генеральную совокупность и может быть выбран повторно.

Пример. Есть мешок с 100 одинаковыми карточками, каждую из которых пронумеровали с 1 до 100. Чтобы отобрать выборку из 10 карточек, вытягивают одну, записывают её номер, затем возвращают карточку обратно в мешок и перемешивают. Так повторяют 10 раз. В итоге одна и та же карточка может быть выбрана несколько раз. В этом примере каждый объект имеет равные шансы попасть в выборку при каждом отборе.

Пример. При контроле качества на конвейере случайным образом берут одну деталь, проверяют её параметры, возвращают обратно на конвейер и повторяют отбор выборки из 30 деталей. Одна и та же деталь может попасть в выборку несколько раз.

Пример. Случайно выбирают курсанты для проверки качества и количества выдаваемого продовольствия. Повторная выборка используется для регулярного мониторинга состояния питания без исключения из общего взвода.

Пример. Проводится выборочный осмотр автотранспорта на базе ремонтного депо. Автомобили для проверки выбираются случайно с возвратом. Это позволяет проверять малые объемы техники многократно в течение времени без исключения из общего парка.

Следует отметить, что в практических целях чаще используется бесповторный случайный отбор. Случайный повторный отбор обычно используется, когда необходимо создать теоретическую модель исследуемой генеральной совокупности (допустим, при изучении пропорций исследуемой со-

вокупности). Следует отметить, что если выборка составляет лишь малую часть всей совокупности, то различие между повторным и бесповторным отборами стирается.

В свою очередь, в зависимости от состава и соотношений генеральной совокупности при организации исследования применяются механический, типический, серийный (расслоенная выборка) или двухстадийный отборы.

Механический отбор – это метод выборки, при котором генеральная совокупность разбивается на равные по объему группы, и из каждой группы отбирается одна единица через регулярные интервалы (шаг отбора) на основе определённого порядка расположения объектов в совокупности (например, по месту в списке, по времени поступления или по алфавиту). Начало отсчета отбора выбирается случайно внутри первого интервала. Такой способ позволяет обеспечить равномерное представительство всех частей совокупности и упрощает процесс выборки. Механический отбор всегда бесповторный.

Пример. Для оценки успеваемости курсантов проводится 10-процентная механическая выборка. Для этого из алфавитного списка отбирается каждый десятый студент, например, 1-й, 11-й, 21-й, 31-й или 7-й, 17-й, 27-й, 37-й и т.д. Если выборка 5%, то отбору подлежит каждый двадцатый курсант.

Пример. В исследовании туристов, прибывающих в аэропорт, проводится опрос каждого десятого человека для анализа предпочтений.

Следует отметить, что механический отбор – это простой и наглядный способ отбора объектов через равные промежутки, который экономит время и снижает субъективизм при составлении выборки, особенно удобен при упорядоченных данных.

Типический отбор – это метод отбора, при котором генеральная совокупность предварительно делится на однородные по какому-либо существенному признаку группы (типические группы). Затем из каждой из этих групп выбираются объекты для выборки случайным или механическим спо-

собою, часто пропорционально численности группы в совокупности. Такой подход позволяет повысить точность и репрезентативность выборки, так как в неё попадают представители всех типических подгрупп, существенно отличающихся по выбранному признаку.

Пример. При исследовании отношения к приватизации предприятия разбиваются на группы по форме собственности (государственные, частные, смешанные), затем из каждой группы выбираются респонденты пропорционально численности предприятия.

Пример. Опрашиваются семьи, разбитые по уровню дохода (низкий, средний, высокий), берётся выборка из каждой группы для изучения потребительского поведения.

Таким образом, типический отбор обеспечивает охват разнообразных подгрупп, делая результаты исследования более достоверными и точными.

Серийный отбор – это метод выборки, при котором отбираются не отдельные единицы генеральной совокупности, а целые группы или серии, внутри которых проводится сплошное обследование всех элементов. Такой способ удобен, когда объекты изначально объединены в однородные группы (например, партии продукции, студенческие группы, бригады), и позволяет снизить затраты на поиск и обследование отдельных элементов, сохраняя при этом информативность выборки

Пример. Контроль качества продукции, когда выбираются несколько партий товара, и каждый предмет внутри партии проверяется.

Пример. Выборка курсантов из учебных групп – случайно отбирается несколько групп, и обследуются все студенты в этих группах.

Двухстадийный отбор – это метод выборки, при котором процесс отбора осуществляется в два последовательных этапа. Сначала из генеральной совокупности выбирается первая выборка (первый этап), а затем из элементов этой выборки производится дальнейший отбор для получения второй выборки (второй этап). Такой подход обычно используется для уменьшения объёма работы и повышения эффективности отбора, особенно при больших

генеральных совокупностях. Такой метод позволяет улучшить представительность выборки и оптимизировать затраты на её формирование.

Пример. В крупной компании сначала случайным образом отбираются филиалы, а затем из каждого выбранного филиала выбираются отдельные сотрудники для опроса.

Пример. В образовательном исследовании для проведения тестирования сначала случайным образом выбираются факультеты, а затем внутри каждого из них отбираются учебные взводы для непосредственного тестирования.

Отметим, что на практике часто применяется комбинированный отбор, при котором сочетаются указанные выше способы. При этом выбираемый способ отбора объектов из генеральной совокупности должен соответствовать поставленной задаче и хорошо согласовываться с сущностью изучаемого явления.

1.2. Методы сбора данных

1.2.1. Опросы, наблюдения, эксперименты

Методы сбора данных служат основой для получения необходимой информации в исследовании. Основные из них – это опросы, наблюдения и эксперименты.

Опрос – метод, при котором исследователь получает информацию прямо от участников посредством постановки вопросов. Анкетирование, интервью и фокус-группы представляют собой разные формы опросов.

Виды опросов:

1) анкетирование – участникам предлагается заполнить заранее разработанные анкеты самостоятельно или при участии интервьюера;

2) интервью – прямой контакт с участником, позволяющий задать уточняющие вопросы и получить глубокие пояснения;

3) фокус-группа – коллективное обсуждение небольшой группы людей, направленное на изучение реакции и восприятие определенных идей.

Преимуществами метода являются быстрота сбора данных, возможность охватить широкую аудиторию, а также удобство при изучении мнений, установок и самооценок респондентов. В то же время к недостаткам можно отнести риск искажений, возникающих из-за непонимания вопросов (например, некорректная формулировка), эффекта социальной желательности, т.е. когда опрашиваемые стремятся дать «правильный» с их точки зрения ответ, а также низкую мотивацию участников, что может снижать качество и достоверность полученной информации.

Наблюдение – метод, при котором исследователь пассивно фиксирует события, действия и поведение объектов без вмешательства и воздействия на них.

Исследовательское наблюдение может быть двух видов: с активным участием исследователя в изучаемой среде или без его участия, то есть с дистанционным наблюдением. Кроме того, наблюдение может быть структурированным – проводится по заранее разработанному плану и с применением конкретных критериев и форм для фиксации, либо неструктурированным, когда план фиксирования данных минимален и исследователь опирается на более свободное и интуитивное восприятие происходящего.

Преимуществами метода являются объективность данных и возможность фиксировать реальные, а не заявленные действия. Этот метод подходит для изучения процессов, поведения потребителей, изделий и т.п., однако требует внимательности и специальных условий для точной фиксации.

Эксперимент – активное вмешательство исследователя с целью создания контролируемых условий для проверки гипотез. Эксперимент позволяет изучать причинно-следственные связи, например, как изменение условий влияет на эффект. В маркетинге это, например, А/В тесты.

Эти методы часто комбинируются для повышения качества и полноты результатов исследования.

1.2.2. Ошибки выборки и их влияние на результаты анализа

Поскольку в большинстве исследований невозможно изучить всю популяцию (генеральную совокупность), то используется только ее часть – выборка. При этом неизбежно возникают ошибки выборки, которые могут привести к искажению полученных результатов и сделать выводы менее точными или неверными.

Ошибки выборки – это расхождения между характеристиками выборки и генеральной совокупности, вызванные случайными или систематическими причинами при отборе данных.

Выделяют следующие типы ошибок:

- *случайные ошибки* возникают из-за случайности отбора, когда выбранная выборка не полностью отражает все разнообразие генеральной совокупности. Их можно уменьшать за счет увеличения размера выборки;
- *систематические ошибки* возникают по причинам неправильной методики отбора, ошибочного формирования выборки или предвзятого отношения. Например, если выборка не репрезентативна, результаты будут искажены.

Источниками ошибок в выборочном исследовании могут служить следующие факторы:

- *Непредставительная выборка*: если выбранная выборка недостаточно точно отражает структуру генеральной совокупности, результаты исследования окажутся смещёнными и не будут адекватно характеризовать популяцию.
- *Несоответствие размеров выборки*: малая выборка увеличивает вероятность случайных колебаний и нестабильности оценок параметров, что снижает надежность выводов.
- *Ошибка измерения*: неправильная формулировка вопросов, ошибки в регистрации данных или неверная обработка информации приводят к систематическим искажениям результатов.

- *Самоотбор*: когда отдельные члены популяции добровольно отказываются участвовать в исследовании, это ведёт к нерепрезентативности выборки и искажению итоговых данных.

Учет данных источников ошибок важен для корректного анализа и повышения достоверности исследований.

Ошибки выборки влияют на надежность и точность анализа, могут приводить к неверным выводам и плохому фундаменту для принятия решений. Для минимизации ошибок важна правильная методология отбора и контроль качества данных.

1.3. Предварительная обработка данных

Процесс предварительной обработки данных направлен на подготовку сырых данных для дальнейшего анализа и моделирования. Обработка данных необходима для повышения качества и надежности результатов последующих этапов анализа. Качество результата во многом зависит от того, насколько тщательно проведена обработка, поскольку исходные данные часто содержат пропуски, выбросы и различия в масштабе.

1.3.1. Пропуски, выбросы, нормализация данных

Пропуски – это отсутствующие значения в наборе данных, которые могут возникать по разным причинам: ошибки при сборе данных, отказ респондентов отвечать на вопросы, технические сбои и другие. Игнорирование пропусков или некорректная их обработка может привести к искажению результатов. Существуют разные стратегии работы с пропусками: удаление строк с пропущенными значениями (эффективно при малом количестве пропусков), заполнение пропусков средними или медианными значениями, прогнозирование пропущенных данных с помощью моделей или методы интерполяции.

Выбросы – это наблюдения, которые резко отличаются от остальных данных и могут влиять на статистические выводы и модели. Причины появления выбросов могут быть естественными (редкие события) или ошибочными (ошибки ввода, сбои в оборудовании). Для выявления выбросов применяют графические методы, например, диаграммы размаха, а также стати-

стические критерии. После обнаружения выбросы можно удалить, скорректировать или оставить при условии обоснования их природы.

Нормализация – процесс преобразования данных к единому масштабу без искажения различий в диапазоне значений. Это особенно важно, когда в наборе присутствуют признаки, измеряемые в разных единицах или имеющие сильно отличающиеся диапазоны. Распространённые методы нормализации – масштабирование до диапазона или стандартизация (приведение к среднему 0 и стандартному отклонению 1). Нормализация улучшает работу многих алгоритмов машинного обучения и статистических моделей.

Таким образом, предварительная обработка данных, включающая корректное устранение пропусков, выявление и работу с выбросами, а также нормализацию, является ключевым этапом, который обеспечивает достоверность, стабильность и качество получаемых аналитических выводов и моделей.

1.3.2. Группировка данных

Важнейшим методом статистики является метод группировок. *Группировкой* называется распределение изучаемых явлений на однородные в качественном отношении группы по ряду существенных признаков. Значение группировки состоит в том, что этот метод обеспечивает обобщение данных по качественно однородным группам, представление их в компактном, оо-зримом виде. Метод группировок формирует научно обоснованную базу для дальнейшего использования иных приемов статистического исследования. Значение этого подхода подчеркнул известный русский ученый Д.П. Журавский, определяя статистику как дисциплину, занимающуюся исчислением категорий, иными словами, подсчетом показателей по классам и группам. Такое определение подчеркивает одну из ключевых особенностей статистической методологии. Потребность в проведении группировок объективна и вызвана, главным образом, существованием качественных отличий среди изучаемых явлений. Благодаря этому методу удастся решать комплексные

задачи статистического анализа. Исходя из решаемых задач выделяют три типа группировок: *типологические, структурные и аналитические*.

1. *Типологическая группировка* – с ее помощью выделяются социально-экономические типы общественных явлений. Например, группировка населения по уровню образования (начальное, среднее, высшее); группировка предприятий по формам собственности (государственные, частные, смешанные).

Пример. Группировка семей по количеству детей:

Количество детей	Доля семей (%)
Бездетные семьи	25
Один ребенок	40
Два ребенка	25
Три и более детей	10

2. Если цель группировки состоит в том, чтобы охарактеризовать состав, структуру совокупности по какому-либо признаку, то такая группировка называется *структурной*. Например, группировка населения по уровню образования, по национальности, по месту проживания.

Пример. Зависимость заработной платы сотрудников от стажа работы:

Стаж работы (лет)	Средняя зарплата (руб.)
Менее 3	30 000
От 3 до 5	45 000
От 6 до 10	60 000
Более 10	80 000

3. Если группировка проводится для того, чтобы установить взаимосвязь между явлениями и показателями, то такую группировку называют *аналитической*. Аналитическую группировку проводят по признаку – фактору.

Пример. Распределение студентов университета по факультетам:

Факультеты	Количество студентов
Экономический	300
Юридический	250
Исторический	150
Технический	200

Статистические группировки строятся как по одному, так и по нескольким признакам. Группировка по одному признаку называется *простой*. Группировка по двум или нескольким признакам называется *комбинированной*.

При формировании структурных классификаций, основанных на различительных количественных показателях, важно заранее установить число выделяемых групп и границы интервалов распределения.

Интервал представляет собой числовое значение, которое отделяет одну единицу (группу) от другой, определяя тем самым четкие рамки каждой категории.

Как правило, интервалы величин представляют собой разницу между максимальными и значениями каждого признака в группе.

Вопрос о количестве групп и крупных интервалах следует решать с учетом обстоятельств, прежде всего исходя из целей исследования, значения изучаемого признака и т.д.

Количество групп и величина интервалов между собой: чем больше образовано групп, тем меньше интервал, и наоборот. Количество групп зависит от *количества наблюдаемого объекта* и *степени колеблемости* группового признака. При небольшом распределении нельзя образовывать большое число групп, так как группы будут малочисленными.

При определении *количества групп* необходимо стремиться к тому, чтобы были учтены особенности изучаемого явления. Поэтому число групп должно быть ниже, в каждой должно быть достаточно большое количество измерений, что отвечает за соблюдение законов больших чисел. Однако в отдельных случаях проявляются интерес и малочисленные группы: новое, передовое, пока оно не станет массовым, проявляется в незначительном количестве фактов; поэтому задача статистики – выявить эти факты, изучить их.

Следовательно, при определении количества единиц в каждой группе важно учитывать не внешние признаки, а глубокое понимание природы изучаемого процесса. Количество формируемых групп существенно зависит от

степени разброса значений выбранного признака: чем сильнее варьирует признак, тем большее число групп целесообразно создавать.

Ориентировочно определить оптимальное количество групп с равными интервалами можно по формуле американского ученого Стерджесса:

$$k \approx 1 + \log_2 n,$$

где n – объем выборки.

Формула Стерджесса применима лишь тогда, когда распределение единиц по заданному признаку близко к нормальному закону распределения и используется принцип равных интервалов внутри групп. Для построения действительно репрезентативных групп обязательно учитывать природу и особенности изучаемого явления.

Интервалы могут быть *равные* и *неравные*. При исследовании экономических явлений могут применяться неравные (прогрессивно возрастающие, прогрессивно убывающие) интервалы. Так, например, по численности работающих промышленные предприятия могут быть разбиты на следующие группы: до 100 человек, 100 – 200, 200 – 300, 300 – 500, 500 – 1000, 1000 и более человек. Это объясняется тем, что количественные изменения размера признака имеют неодинаковые значения в низших и высших по размеру признака группах: изменение количества работающих на 50 – 100 человек имеет существенное значение для мелких предприятий, а для крупных – не имеет.

Группировки с равными интервалами целесообразны в тех случаях, когда вариация проявляется в сравнительно узких границах и распределение является практически равномерным (например, при группировке рабочих одной профессии по размеру заработной платы, посевов какой-либо культуры по урожайности).

Для группировок с равными интервалами величина интервала:

$$h = \frac{x_{\max} - x_{\min}}{1 + \log_2 n},$$

где $x_{\min}$, $x_{\max}$ – наименьшее и наибольшее значение признака.

1.4. Абсолютные и относительные величины

Для того чтобы кратко охарактеризовать эмпирические данные, используют количественные величины, которые представляют все данные так, что можно отказаться от отдельных, попавших под наблюдение значений признака – количественные величины, которые называются *статистическими показателями*.

К ним относятся: *абсолютные и относительные показатели, показатели центра и формы распределения, концентрации, корреляции*.

Также выделяют *индивидуальные* (относятся к отдельным единицам совокупности) и *обобщающие* (сводные) статистические показатели.

Последние получаются либо суммированием значений признака отдельных единиц совокупности (объемные показатели), либо расчетным путем по формулам (расчетные показатели).

1.4.1. Абсолютные величины

Абсолютные показатели выражают абсолютные размеры явлений, процессов и получаются в результате сводки и группировки статистической информации. Например, масса, площадь, объем, протяженность и т.д.

Все абсолютные показатели имеют определенную размерность. Единица измерения может быть выражена:

- 1) натуральными измерителями (кг);
- 2) условно-натуральными измерителями (условная консервная банка);
- 3) составными натуральными показателями (кВт/ч);
- 4) стоимостными показателями (руб., \$);
- 5) трудовыми показателями (трудодень).

Абсолютные показатели просты и наглядны, однако они не могут дать полную характеристику изучаемого явления или процесса. Проблема заключается в сопоставлении показателей, особенно это касается стоимостных единиц измерения. В этом случае появляется необходимость рассмотрения других видов статистических показателей.

1.4.2. Относительные величины

Относительные показатели представляют собой частное от деления двух статистических величин и характеризуют количественное соотношение между ними.

$$\text{Относительный показатель} = \frac{\text{Сравниваемая величина}}{\text{База сравнения}}.$$

Если база сравнения равна единице, то относительный показатель является *коэффициентом*. Если она равна ста, то относительный показатель выражается в *процентах* (%), если она равна тысяче – в *промилле* (‰), десяти тысячам – в *продецимилле* (‰_{000}).

Виды относительных показателей

1. Выполнение договорных обязательств:

$$\text{Макет показателя} = \frac{\text{Фактический уровень}}{\text{Уровень предусмотренный договором}}.$$

2. Относительный показатель структуры:

$$\text{Макет показателя} = \frac{\text{Показатель, характеризующий часть совокупности}}{\text{Показатель по всей совокупности}}.$$

3. Относительный показатель сравнения показывает соотношение одних и тех же данных, но относящихся к разным объектам наблюдения:

$$\text{Макет показателя} = \frac{\text{Показатель, характеризующий объект } A}{\text{Показатель, характеризующий объект } B}.$$

4. Относительная величина координации показывает, во сколько раз сравниваемая часть совокупности больше или меньше другой части совокупности, выбранной в качестве базы сравнения (например, отношение мужской части населения к женской или наоборот):

$$\text{Макет показателя} = \frac{\text{Показатель } i - \text{й части совокупности}}{\text{Показатель части совокупности, выбранной в качестве базы сравнения}}.$$

5. Относительный показатель интенсивности характеризует распределение изучаемого явления в присущей ему среде (например, плотность населения):

$$\text{Макет показателя} = \frac{\text{Показатель, характеризующий явление}}{\text{Показатель, характеризующий среду распространения явления}}$$

6. Цепные и базисные показатели динамики характеризуют развитие явления во времени по сравнению с предыдущим периодом времени (цепные) или периодом времени, выбранным в качестве базисного (базисные):

$$\text{Макет показателя} = \frac{\text{Текущий, показатель}}{\text{Предыдущий (базисный) показатель}}$$

Пример. Вычислить цепные и базисные показатели динамики (период 1- базисный).

Период	Величина показателя
1	200
2	250
3	300

Решение. По условию задачи в качестве базисного показателя берем период 1.

1) Базисные относительные величины (в процентах) находим как отношение величины каждого показателя к величине базисного значения, т.е.
200 – 100%;

250 составляет 125% от 200;

300 составляет 150% от 200.

2) Цепные относительные величины (в процентах) находим как отношение величины текущего показателя к предыдущему, т.е.

250 составляет 125% от 200;

300 составляет 120% от 250.

Все результаты заносим в таблицу.

Период	Величина показателя	Базисная относительная величина	Цепная относительная величина
1	200	100%	—
2	250	125%	125%
3	300	150%	120%

Темп роста – отношение текущего показателя к показателю, выбранному за базу сравнения:

$$TP = \frac{250}{200} \cdot 100\% = 125\%.$$

Темп прироста – темп роста минус единица (100%):

$$ТП = 125\% - 100\% = 25\%.$$

1.5. Статистические ряды распределения

После того, как выполнена группировка признака и определены границы групп, строится ряд распределения.

Статистический ряд распределения представляет собой упорядоченное распределение единиц изучаемой совокупности на группы по определенному варьирующему признаку. Он характеризует состав (структуру) изучаемого явления, позволяет судить об однородности совокупности, закономерности распределения и границах варьирования единиц совокупности.

Ряды распределения, построенные по *атрибутивным* признакам (в порядке возрастания или убывания наблюдаемых значений), называются атрибутивными. Примером атрибутивных рядов могут служить распределения населения по полу, занятости, национальности, профессии и т.д.

Ряды распределения, построенные по количественному признаку, называются *вариационными*. Например, распределение населения по возрасту, рабочих – по стажу работы, заработной плате и т.д.

Вариационные ряды распределения состоят из двух элементов: *вариантов* и *частот*.

Числовые значения количественного признака в вариационном ряду распределения называются *вариантами*. Они могут быть положительными и отрицательными, абсолютными и относительными. Так, при группировке предприятий по результатам хозяйственной деятельности варианты положительные (прибыль) и отрицательные (убыток) числа.

Частоты (вес варианта) – это численности отдельных вариантов или каждой группы вариационного ряда, т.е. это числа, показывающие, как часто встречаются те или иные варианты в ряду распределения и обозначаются n_i ($i=1, 2, \dots, k$), где i – номер варианта.

Сумма всех частот называется *объемом совокупности* и определяет число элементов всей совокупности.

$$\sum_{i=1}^k n_i = n - \text{объем выборки.}$$

Частоты (относительная частота или доля) – это частоты, выраженные в виде относительных величин (долях единиц или процентах).

$$\omega_i = \frac{n_i}{n} \quad (i=1, 2, \dots, k),$$

где k – число вариантов, n – объем выборки.

Сумма частостей равна единице или 100%.

$$\sum_{i=1}^k \omega_i = 1, \quad (i=1, 2, \dots, k).$$

Частость ω_i является статистической вероятностью появления варианта x_i . Она является аналогом вероятности p_i появления значения x_i случайной величины X .

Замена частот частостями позволяет сопоставлять вариационные ряды с разным числом наблюдений.

Вариационные ряды в зависимости от характера вариации подразделяются на *дискретные* и *интервальные*. Дискретные вариационные ряды основаны на дискретных (прерывных) признаках, имеющих только целые значения (например, тарифный разряд рабочих, число детей в семье), на дискретных признаках, представленных в виде интервалов; интервальные – на непрерывных признаках (имеющих любые значения, в том числе и дробные).

При наличии достаточно большого количества вариантов значений признака первичный ряд является трудно обозримым и непосредственное рассмотрение его не дает представления о распределении единиц по значе-

нию признака в совокупности. Поэтому первым шагом в упорядочении первичного ряда является его *ранжирование*, т.е. расположение всех вариантов в возрастающем (или убывающем) порядке.

Накопленная частота $n_i^{\text{нак}}$ показывает, сколько наблюдалось вариантов со значением признака, меньшим x . Отношение накопленной частоты $n_i^{\text{нак}}$ к общему числу наблюдений n называется *накопленной частотой* $\omega_i^{\text{нак}}$.

Дискретный статистический ряд:

x_i	x_1	x_2	...	x_k
n_i	n_1	n_2	...	n_k

x_i	x_1	x_2	...	x_k
ω_i	ω_1	ω_2	...	ω_k

Интервальный статистический ряд:

$x_i - x_{i+1}$	$x_1 - x_2$	$x_2 - x_3$	...	$x_k - x_{k+1}$
n_i	n_1	n_2	...	n_k

Переход от дискретного к интервальному вариационному ряду

При переходе от дискретного вариационного ряда к интервальному необходимо весь интервал наблюдаемых значений разбивается на k частичных интервалов $[x_1, x_2)$, $[x_2, x_3)$, ..., $[x_k, x_{k+1}]$ одинаковой длины h . Далее подсчитывают количество попаданий наблюдений в эти интервалы. Эти числа принимают за частоты n_i . Согласно формуле Стерджеса рекомендуемое число интервалов разбиений $k \approx 1 + \log_2 n$, а длины частичных интервалов

$h = \frac{x_{\max} - x_{\min}}{1 + \log_2 n}$, n – объем выборки. Если шаг окажется дробным, то за длину

интервала берут ближайшее целое число или ближайшую простую дробь.

За начало первого интервала рекомендуется брать величину

$x_{\text{нач}} = x_{\min} - \frac{h}{2}$, а конец последнего должен удовлетворять условию

$x_{\text{кон}} - h \leq x_{\max} < x_{\text{кон}}$. Промежуточные интервалы получают, прибавляя к концу предыдущего интервала шаг h .

Пример. Органами МВД за отчетный период зарегистрировано следующее количество преступлений: 169, 179, 165, 156, 179, 178, 160, 154, 183,

155, 153, 167, 186, 163, 155, 157, 175, 170, 166, 159, 173, 182, 167, 171, 169, 179, 165, 156, 179, 158, 171, 175, 173, 164, 172. Построить интервальный статистический ряд.

Решение. Ранжируем полученные данные: 153, 154, 155, 155, 156, 157, 158, 159, 160, 163, 164, 165, 166, 167, 167, 169, 170, 171, 171, 172, 173, 173, 175, 175, 178, 179, 179, 182, 183, 186.

Объем выборки: $n = 30$.

Наименьшее значение варианты: $x_{\min} = 153$.

Наибольшее значение варианты: $x_{\max} = 186$.

По формуле Стерджеса находим число интервалов разбиений:
 $k \approx 1 + \log_2 30 \approx 5,9069 \approx 6$.

Длина частичных интервалов: $h = \frac{186-153}{1+\log_2 30} \approx \frac{33}{6} \approx 5,5 \approx 6$.

Начальное значение: $x_{нач} = 153 - \frac{6}{2} = 150$.

Составляем интервальный статистический ряд.

$x_i - x_{i+1}$	150–156	156–162	162–168	168–174	174–180	180–186
n_i	4	5	6	7	5	3

Переход от интервального к дискретному вариационному ряду

Иногда интервальный статистический ряд, для простоты исследований, условно заменяют дискретным. В этом случае срединное значение i -го интервала принимают за вариант x_i , а соответствующую интервальную частоту n_i – за частоту этого варианта.

Пример. Дан интервальный вариационный ряд:

$x_i - x_{i+1}$	150–156	156–162	162–168	168–174	174–180	180–186
n_i	4	5	6	7	5	3

Построить дискретный вариационный ряд.

Решение. Найдем середины частичных интервалов, получим числа: 153, 159, 165, 171, 177, 183. Запишем дискретный вариационный ряд.

x_i	153	159	165	171	177	183
n_i	4	5	6	7	5	3

1.6. Средние величины и показатели вариации

Рассмотрим дискретный статистический ряд:

x_i	x_1	x_2	$\dots$	x_k
n_i	n_1	n_2	$\dots$	n_k

1. *Размах варьирования* $R = x_{\max} - x_{\min}$.
2. *Мода* M_o – вариант, имеющий наибольшую частоту.
3. *Медиана* M_e – вариант, приходящийся на середину ранжированной выборки.

Пусть n – объем выборки. Тогда:

- 1) если $n=2k$, то есть ряд имеет четное число членов, то $M_e = \frac{x_k + x_{k+1}}{2}$;
- 2) если $n=2k+1$, то есть ряд имеет нечетное число членов, то $M_e = x_{k+1}$.

Пример. Органами МВД зарегистрирована подростковая преступность в возрасте: 14, 16, 12, 13, 14, 15, 14, 16, 14, 15, 15, 16, 14, 15, 13, 14, 15, 15, 16, 13.

Составить дискретный ряд распределения числа преступлений по возрасту. Найти объем, размах, моду и медиану выборки.

Решение. Ранжируем исходные данные: 12, 13, 13, 13, 14, 14, 14, 14, 14, 14, 14, 15, 15, 15, 15, 15, 16, 16, 16, 16.

Объем выборки: $n=20$.

Наименьшее значение варианты: $x_{\min} = 12$.

Наибольшее значение варианты: $x_{\max} = 16$.

Размах: $R = x_{\max} - x_{\min} = 16 - 12 = 4$.

Мода: $M_0 = 14$, так как вариант $x_i = 14$ встречается наибольшее количество раз $n_i = 7$.

Так как $n=20$, то медиана $M_e = \frac{x_{10} + x_{11}}{2} = \frac{14 + 14}{2} = \frac{28}{2} = 14$.

Вариационный ряд:

x_i	12	13	14	15	16
n_i	1	3	7	5	4

Рассмотрим интервальный статистический ряд:

$x_i - x_{i+1}$	$x_1 - x_2$	$x_2 - x_3$	...	$x_k - x_{k+1}$
n_i	n_1	n_2	...	n_k

1. *Мода*. Для интервального вариационного ряда вначале находят интервал группировки, в котором содержится медиана, путем подсчета накопленных частот или накопленных относительных частот. Внутри медианного интервала медиана определяется по формуле:

$$M_e = x_k + \frac{0,5n - n_{k-1}}{n_{M_e}} h_{M_e},$$

где x_k – нижняя граница медианного интервала; h_{M_e} – ширина медианного интервала; n_{k-1} – накопленная частота интервала, предшествующему медианному; n_{M_e} – частота медианного интервала.

Пример. Найти медиану для интервального вариационного ряда.

Превышение разрешенной скорости движения (км/ч)	20-30	30-40	40-50	50-60	Больше 60
Количество нарушений	50	32	26	11	5

Решение.

Вычислим объем выборки: $n=124$, $n/2=62$.

Определяем, на каком частичном интервале достигается значение $n/2=62$: 30-40 – медианный интервал.

$x_k = 30$; $h_{M_e} = 10$; $n_{k-1} = 50$; $n_{M_e} = 32$.

$$M_e = 30 + \frac{0,5 \cdot 124 - 50}{32} \cdot 10 = 33,75.$$

Т.е. количество нарушений со скоростью ниже 34 км/ч и со скоростью выше 34 км/ч будет одинаковое.

2. *Медиана*. Для интервального вариационного ряда вначале находят интервалы группировки с наибольшей частотой (модальный интервал). Внутри модального интервала мода определяется по формуле:

$$M_o = x_k + \frac{f_k - f_{k-1}}{2f_k - (f_{k-1} + f_{k+1})}h,$$

где x_k – нижняя граница модального интервала; f_k – частота модального интервала; f_{k-1} – частота предмодального интервала; f_{k+1} – частота послемодального интервала.

Пример. Для интервального статистического ряда определить моду и медиану.

Возрастные группы	До 20 лет	20-25	25-30	30-35	35-40	40-45	Свыше 45
Число курсантов (слушателей)	346	872	1054	781	212	121	76

Решение.

Вычислим объем выборки: $n = 3462$.

Модальный интервал в возрастной группе: 25-30.

$x_k = 25$; $f_k = 1054$; $f_{k-1} = 872$; $f_{k+1} = 781$.

$$M_o = 25 + \frac{1054 - 872}{2 \cdot 1054 - (872 + 781)} \cdot 5 = 27.$$

Модальный возраст слушателей – 27 лет.

Медианный интервал в возрастной группе: 25-30.

$$M_e = 25 + \frac{0,5 \cdot 3462 - 1218}{1054} \cdot 5 = 27,4.$$

Половина слушателей в возрасте до 27,4 года, а другая половина – свыше 27,4 года.

Пусть дано статистическое распределение выборки объема n :

x_i	x_1	x_2	...	x_m
n_i	n_1	n_2	...	n_m

Выборочным средним $\bar{x}_e$ называется среднее арифметическое всех значений выборки:

$$\overline{x}_e = \frac{1}{n} \sum_{i=1}^m x_i n_i .$$

Так как $\omega_i = \frac{n_i}{n}$, то для вычисления выборочной средней можно использовать формулу:

$$\overline{x}_e = \sum_{i=1}^m x_i \omega_i .$$

В случае интервального вариационного ряда в качестве x_i берут середины интервалов, а n_i – соответствующие им частоты.

Выборочной дисперсией D_e называется среднее арифметическое квадратов отклонений значений выборки от выборочного среднего $\overline{x}_e$:

$$D_e = \frac{1}{n} \sum_{i=1}^m (x_i - \overline{x}_e)^2 \cdot n_i .$$

Выборочная дисперсия может быть вычислена по формуле:

$$D_e = \overline{x_e^2} - (\overline{x_e})^2, \text{ где } \overline{x_e^2} = \frac{1}{n} \sum_{i=1}^m x_i^2 n_i .$$

Так как $\omega_i = \frac{n_i}{n}$, то для вычисления выборочной дисперсии можно использовать формулу:

$$D_e = \sum_{i=1}^m (x_i - \overline{x}_e)^2 \cdot \omega_i .$$

Выборочное среднее квадратическое отклонение вычисляется по формуле:

$$\sigma_e = \sqrt{D_e} .$$

Особенность σ_e состоит в том, что оно измеряется в тех же единицах, что и данные выборки.

При решении практических задач используется *исправленная выборочная дисперсия* S^2 :

$$S^2 = \frac{n}{n-1} D_{\epsilon}.$$

Замечание. Формулы для вычисления выборочной дисперсии и исправленной дисперсии отличаются только знаменателями. При достаточно большом объеме выборки n выборочная D_{ϵ} и исправленная S^2 дисперсии мало отличаются друг от друга, поэтому на практике исправленной дисперсией пользуются, если $n \leq 30$.

Исправленное среднее квадратическое отклонение S определяется по формуле:

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

Пример. Дан статистический ряд распределения частот:

x_i	1	3	4	5	6
n_i	3	5	7	2	3

Вычислить среднюю выборочную ($\bar{x}_{\epsilon}$), выборочную дисперсию (D_{ϵ}), выборочное среднее квадратическое отклонение (σ_{ϵ}), исправленную дисперсию (S^2), исправленное среднее квадратическое отклонение (S).

Решение. Объем выборки: $n = 20$.

Выборочная средняя:

$$\bar{x}_{\epsilon} = \frac{1}{n} \sum_{i=1}^m x_i n_i = \frac{1 \cdot 3 + 3 \cdot 5 + 4 \cdot 7 + 5 \cdot 2 + 6 \cdot 3}{20} = 3,7.$$

Выборочная дисперсия:

$$\overline{x_{\epsilon}^2} = \frac{1}{n} \sum_{i=1}^m x_i^2 n_i = \frac{1^2 \cdot 3 + 3^2 \cdot 5 + 4^2 \cdot 7 + 5^2 \cdot 2 + 6^2 \cdot 3}{20} = 15,9.$$

$$D_{\epsilon} = \overline{x_{\epsilon}^2} - (\bar{x}_{\epsilon})^2 = 15,9 - (3,7)^2 = 2,21.$$

Выборочное среднее квадратическое отклонение:

$$\sigma_{\epsilon} = \sqrt{D_{\epsilon}} = \sqrt{2,21} \approx 1,4866.$$

Исправленная дисперсия:

$$S^2 = \frac{n}{n-1} D_{\epsilon} = \frac{20}{20-1} \cdot 2,21 \approx 2,3263.$$

Исправленное среднее квадратическое отклонение:

$$S = \sqrt{S^2} = \sqrt{2,3263} \approx 1,5252 .$$

Еще одним показателем вариационного ряда является безразмерная характеристика – *коэффициент вариации*, равный процентному отношению среднего квадратического отклонения к средней арифметической:

$$\tilde{v} = \frac{S}{\bar{x}_B} \cdot 100\% \quad (\bar{x} \neq 0) .$$

Данный статистический показатель показывает относительный разброс данных вокруг среднего значения и чем ниже его значение, тем более однородны данные, а высокий коэффициент (особенно выше 33%) указывает на значительную неоднородность информации.

Замечание. Если коэффициент вариации менее 10%, то данные однородны, т.е. наблюдается незначительная степень рассеивания данных. Значение в пределах 10%-20% указывают на среднюю степень разброса. Если $\tilde{v}$ принимает значение в интервале 20%-33%, то наблюдается значительная степень рассеивания. При значениях коэффициента вариации более 33% необходима корректировка выборки, т.е. выборка неоднородна.

Пример. Имеются показатели прибыли пяти предприятий отрасли А (в млн рублей):

Предприятие	Прибыль
A1	50
A2	60
A3	70
A4	80
A5	90

Вычислить коэффициент вариации.

Решение.

Среднее арифметическое:

$$\bar{x}_B = \frac{50 + 60 + 70 + 80 + 90}{5} = \frac{350}{5} = 70 .$$

Дисперсия:

$$S^2 = \frac{(50-70)^2 + (60-70)^2 + (70-70)^2 + (80-70)^2 + (90-70)^2}{5-1} = \frac{1000}{4} = 250.$$

Стандартное отклонение:

$$S = \sqrt{S^2} = \sqrt{250} \approx 15,81.$$

Коэффициент вариации:

$$\tilde{v} = \frac{15,81}{70} \cdot 100\% \approx 22,59\%.$$

Таким образом, коэффициент вариации прибыли предприятий группы А составляет примерно 22,59%. Это означает, что прибыль предприятий колеблется около среднего показателя примерно на четверть своей величины.

Пример. Имеются показатели себестоимости продукции шести заводов отрасли Б (в тыс. руб./ед.):

Завод	Себестоимость
В1	10
В2	12
В3	14
В4	16
В5	18
В6	20

Вычислить коэффициент вариации.

Решение.

Среднее арифметическое:

$$\bar{x}_B = \frac{10 + 12 + 14 + 16 + 18 + 20}{6} = \frac{90}{6} = 15.$$

Дисперсия:

$$S^2 = \frac{(10-15)^2 + (12-15)^2 + (14-15)^2 + (16-15)^2 + (18-15)^2 + (20-15)^2}{6-1} = \frac{70}{5} = 14.$$

Стандартное отклонение:

$$S = \sqrt{S^2} = \sqrt{14} \approx 3,74.$$

Коэффициент вариации:

$$\tilde{v} = \frac{3,74}{15} \cdot 100\% \approx 24,93\%.$$

Итак, коэффициент вариации себестоимости продукции заводов группы Б равен примерно 24,93%, что свидетельствует о средней величине колебаний себестоимости вокруг среднего значения.

Сравнивая результаты двух рассмотренных примеров, можно сделать вывод, что обе группы демонстрируют среднюю изменчивость показателей, хотя группа Б чуть более нестабильна.

Средняя выборочная и выборочная дисперсия являются частными случаями *моментов статистического распределения*.

Начальным выборочным моментом порядка k называется среднее арифметическое k -х степеней всех значений выборки:

$$\tilde{v}^k = \frac{\sum_{i=1}^m x_i^k \cdot n_i}{n} \quad \text{или} \quad \tilde{v}^k = \sum_{i=1}^m x_i^k \cdot \omega_i.$$

Замечание. Средняя арифметическая является начальным выборочным моментом первого порядка: $\tilde{v}_1 = \frac{1}{n} \sum_{i=1}^m x_i^1 \cdot n_i = \overline{x_g}$.

Центральным выборочным моментом порядка k называется среднее арифметическое k -х степеней отклонений наблюдаемых значений выборки от выборочного среднего $\overline{x_g}$:

$$\tilde{\mu}^k = \frac{\sum_{i=1}^m (x_i - \overline{x_g})^k \cdot n_i}{n} \quad \text{или} \quad \tilde{\mu}^k = \sum_{i=1}^m (x_i - \overline{x_g})^k \cdot \omega_i.$$

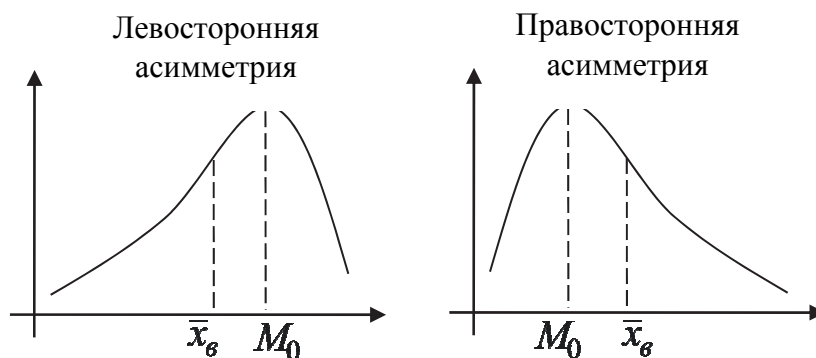
Замечание. Центральный выборочный момент первого порядка равен нулю: $\tilde{\mu}_1 = \frac{1}{n} \sum_{i=1}^m (x_i - \overline{x_g})^1 \cdot n_i = 0$, так как $\sum_{i=1}^m (x_i - \overline{x_g}) \cdot n_i = 0$.

Замечание. Выборочная дисперсия является центральным выборочным моментом второго порядка: $\tilde{\mu}_2 = \frac{1}{n} \sum_{i=1}^m (x_i - \overline{x_g})^2 \cdot n_i = D_g$.

Выборочным коэффициентом асимметрии называется число $\tilde{A}$, вычисляемое по формуле: $\tilde{A} = \frac{\tilde{\mu}_3}{\sigma_s^3}$.

Выборочный коэффициент асимметрии служит для характеристики асимметрии полигона вариационного ряда. Если $\tilde{A} = 0$, то распределение имеет симметричную форму, т.е. варианты равноудаленные от x , имеют одинаковую частоту (средняя арифметическая, мода и медиана совпадают). Если это равенство нарушается, то распределение ассиметрично.

Если $\tilde{A} < 0$, то наблюдается положительная (правосторонняя) асимметрия, т.е. более пологий «спуск» полигона наблюдается слева. Если $\tilde{A} > 0$, то наблюдается отрицательная (левосторонняя) асимметрия, т.е. более пологий «спуск» полигона наблюдается справа.

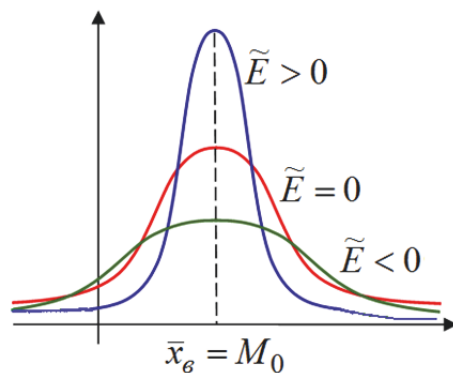


Выборочным коэффициентом эксцесса или коэффициентом крутости называется число $\tilde{E}$, вычисляемое по формуле:

$$\tilde{E} = \frac{\tilde{\mu}^4}{\sigma_s^4} - 3.$$

Выборочный коэффициент эксцесса служит для сравнения на «крутость» выборочного распределения с нормальным распределением.

Коэффициент эксцесса для случайной величины, распределенной по нормальному закону, равен нулю.



Поэтому за стандартное значение выборочного коэффициента эксцесса принимают $\tilde{E} = 0$. Если $\tilde{E} > 0$ ($\tilde{E} < 0$), то полигон вариационного ряда имеет более крутую (пологую) вершину по сравнению с нормальной кривой.

Пример. Вычислить коэффициент асимметрии и эксцесс по данным представленным в таблице:

№	Значения	Частота
1	94-100	3
2	100-106	7
3	106-112	11
4	112-118	20
5	118-124	28
6	124-130	19
7	130-136	10
8	136-142	2

Решение. Объем выборки: $n=100$.

Составим дискретный вариационный ряд распределения частот:

№	x_i	n_i
1	97	3
2	103	7
3	109	11
4	115	20
5	121	28
6	127	19
7	133	10
8	139	2

Выборочная средняя:

$$\bar{x}_6 = \frac{1}{n} \sum_{i=1}^m x_i n_i = \frac{1}{20} \cdot (97 \cdot 3 + 103 \cdot 7 + 109 \cdot 11 + 115 \cdot 20 + 121 \cdot 28 + 127 \cdot 19 + 133 \cdot 10 + 139 \cdot 2)$$

$$+133 \cdot 10 + 139 \cdot 2) = 119,2.$$

Выборочная дисперсия:

$$\overline{x^2} = \frac{1}{n} \sum_{i=1}^m x_i^2 n_i = \frac{1}{20} \cdot (97^2 \cdot 3 + 103^2 \cdot 7 + 109^2 \cdot 7 + 115^2 \cdot 20 + 121^2 \cdot 28 + \\ + 127^2 \cdot 19 + 133^2 \cdot 10 + 139^2 \cdot 2) = 14296,1.$$

$$D_{\epsilon} = \overline{x^2} - (\overline{x})^2 = 14296,1 - (119,2)^2 = 87,48.$$

Выборочное среднее квадратическое отклонение:

$$\sigma_{\epsilon} = \sqrt{D_{\epsilon}} = \sqrt{87,48} \approx 9,353.$$

Центральный выборочный момент 3-го порядка:

$$\tilde{\mu}^3 = \frac{1}{n} \sum_{i=1}^m (x_i - \overline{x})^3 \cdot n_i = \frac{1}{100} \cdot ((97 - 119,2)^3 \cdot 3 + (103 - 119,2)^3 \cdot 7 + \\ + (109 - 119,2)^3 \cdot 7 + (115 - 119,2)^3 \cdot 20 + (121 - 119,2)^3 \cdot 28 + \\ + (127 - 119,2)^3 \cdot 19 + (133 - 119,2)^3 \cdot 10 + (139 - 119,2)^3 \cdot 2) \approx -247,54$$

Центральный выборочный момент 4-го порядка:

$$\tilde{\mu}^4 = \frac{1}{n} \sum_{i=1}^m (x_i - \overline{x})^4 \cdot n_i = \frac{1}{100} \cdot ((97 - 119,2)^4 \cdot 3 + (103 - 119,2)^4 \cdot 7 + \\ + (109 - 119,2)^4 \cdot 7 + (115 - 119,2)^4 \cdot 20 + (121 - 119,2)^4 \cdot 28 + \\ + (127 - 119,2)^4 \cdot 19 + (133 - 119,2)^4 \cdot 10 + (139 - 119,2)^4 \cdot 2) \approx 20767,8$$

Коэффициент асимметрии:

$$\tilde{A} = \frac{\tilde{\mu}_3}{\sigma_{\epsilon}^3} = \frac{-247,54}{9,353^3} = -0,3025.$$

Экссесс:

$$\tilde{E} = \frac{\tilde{\mu}^4}{\sigma_{\epsilon}^4} - 3 = \frac{20767,8}{9,353^4} - 3 = -0,2862.$$

Так как коэффициент асимметрии $\tilde{A}$ отрицателен, то наблюдается левосторонняя асимметрия. Поскольку эксцесс $\tilde{E}$ отрицательный, то рассмат-

риваемое распределение имеет пологую вершину по сравнению с нормальной кривой.

1.7. Визуализация данных

Наиболее распространенными способами графического представления вариационных рядов являются *полигон*, *гистограмма* и *кумулята*.

Полигон частот или *относительных частот* служит для изображения дискретного вариационного ряда и представляет собой ломаную линию, отрезки которой соединяют точки с координатами $(x_i; n_i)$ или $(x_i; \omega_i)$, $i = 1, 2, \dots, k$. Полигон относительных частот является аналогом многоугольника распределения дискретной случайной величины в теории вероятностей.

Гистограмма частот или *относительных частот* служит только для изображения интервальных вариационных рядов и представляет собой ступенчатую фигуру, состоящую из прямоугольников с основаниями, равными частичным интервалам значений признака $h_i = x_{i+1} - x_i$, $i = 1, 2, \dots, k$, и высотами, равными частотам n_i или относительным частотам ω_i . Если соединить середины верхних оснований прямоугольников отрезками прямой, то можно получить полигон того же распределения.

Кумулятивная кривая (кумулята) – кривая накопленных частот (частостей). Для дискретного вариационного ряда кумулята представляет ломаную линию, соединяющую точки $(x_i; n_i^{нак})$ или $(x_i; \omega_i^{нак})$, $i = 1, 2, \dots, k$. Для интервального вариационного ряда ломаная начинается с точки, абсцисса которой равна началу первого интервала, а ордината – накопленной частоте (частости), равной нулю. Другие точки этой ломаной соответствуют концам интервалов.

Пример. В результате некоторого эксперимента получены следующие данные:

2, 4, 9, 5, 4, 8, 3, 6, 7, 2,
8, 7, 5, 3, 7, 4, 4, 3, 4, 6.

Найти размах варьирования R , моду M_0 , медиану M_e . Построить полигон и кумюляту частот и относительных частот.

Решение. Объем выборки: $n = 20$.

Размах: $R = 7$.

Мода: $M_0 = 4$.

Медиана: $M_e = \frac{4+5}{2} = 4,5$.

Ранжируем исходные данные: 2, 2, 3, 3, 3, 4, 4, 4, 4, 4, 5, 5, 6, 6, 7, 7, 7, 8, 8, 9.

Составим дискретный вариационный ряд.

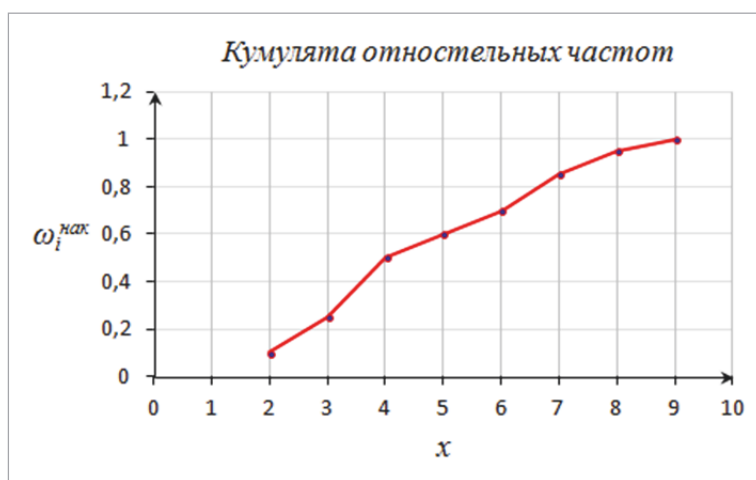
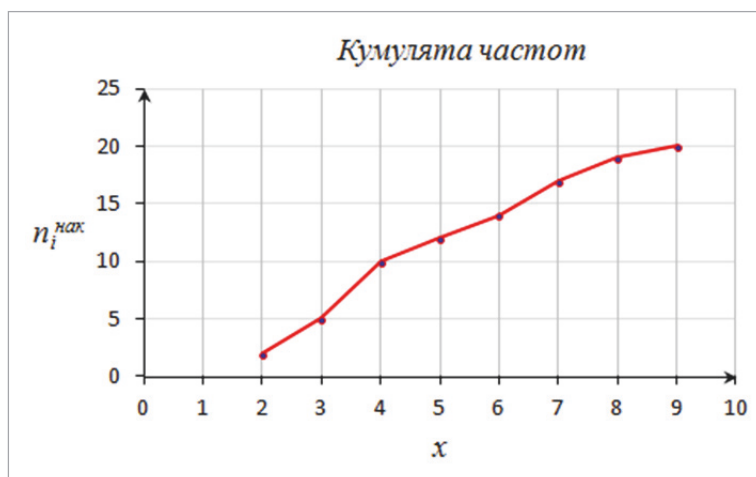
x_i	2	3	4	5	6	7	8	9
n_i	2	3	5	2	2	3	2	1

Найдем относительные частоты, накопленные частоты и накопленные относительные частоты. Составим таблицу:

x_i	2	3	4	5	6	7	8	9
n_i	2	3	5	2	2	3	2	1
ω_i	0,1	0,15	0,25	0,1	0,1	0,15	0,1	0,05
$n_i^{\text{нак}}$	2	5	10	12	14	17	19	20
$\omega_i^{\text{нак}}$	0,1	0,25	0,5	0,6	0,7	0,85	0,95	1

На основе полученной таблицы строим полигон частот и относительных частот, кумюляту частот и относительных частот.





Пример. Дан интервальный вариационный ряд:

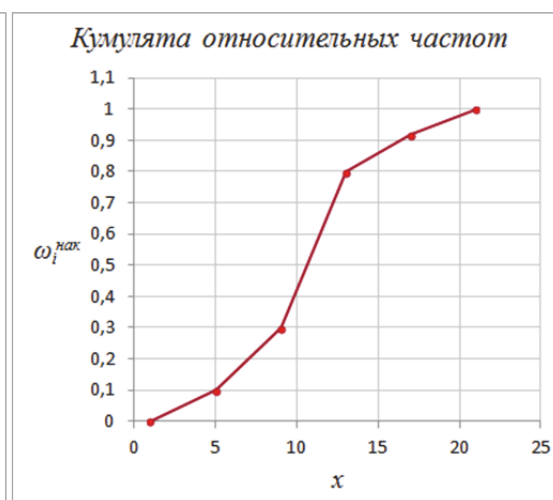
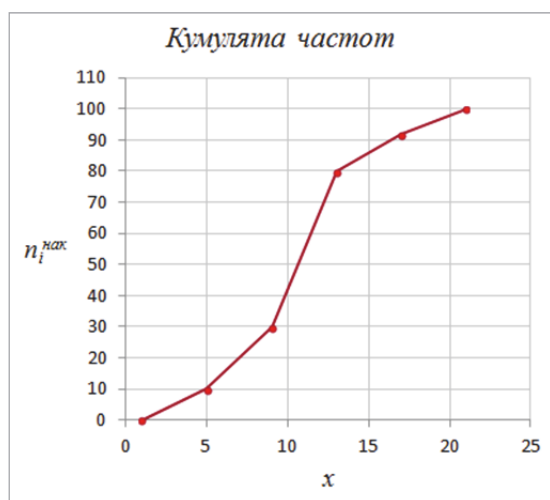
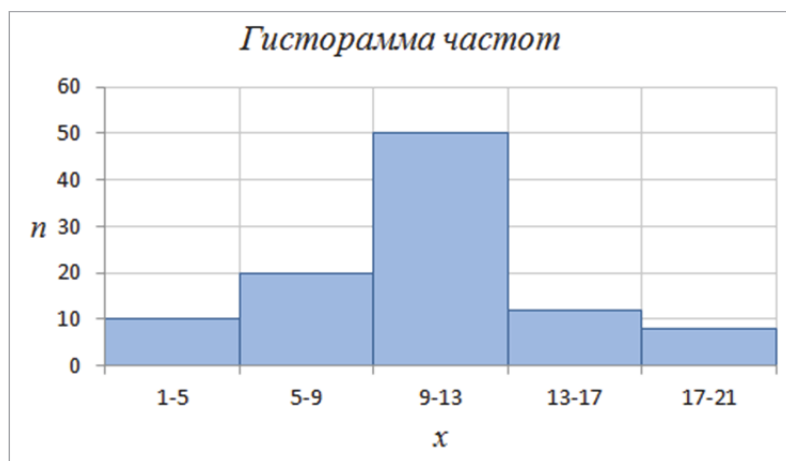
$x_i - x_{i+1}$	1 – 5	5 – 9	9 – 13	13 – 17	17 – 21
n_i	10	20	50	12	8

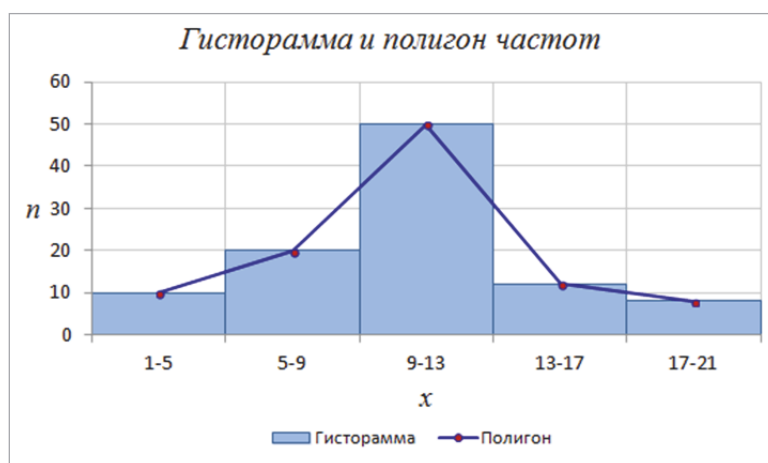
Построить гистограмму и кумуляту частот и относительных частот.

Решение. Найдем относительные частоты, накопленные частоты и накопленные относительные частоты. Составим таблицу:

$x_i - x_{i+1}$	1 – 5	5 – 9	9 – 13	13 – 17	17 – 21
n_i	10	20	50	12	8
ω_i	0,1	0,2	0,5	0,12	0,08
$n_i^{\text{нак}}$	10	30	80	92	100
$\omega_i^{\text{нак}}$	0,1	0,3	0,8	0,92	1

На основе полученной таблицы строим гистограмму частот и относительных частот, кумуляту частот и относительных частот.





1.8. Практические примеры

1.8.1. Анализ затрат на транспортировку

Транспортировка является одним из важнейших элементов логистической системы, обеспечивающего своевременное перемещение материальных ресурсов и сотрудников. Затраты на транспортировку включают расходы на топливо, техническое обслуживание транспортных средств, оплату труда водителей и прочие издержки. Эффективный контроль и оптимизация этих расходов позволяет снизить общие затраты и повысить эффективность службы материально-технического снабжения любого подразделения.

Пример. Рассчитать среднюю стоимость одного километра пробега автомобиля УАЗ Патриот, используемого для доставки ГСМ подразделениям полиции.

Исходные данные:

- Стоимость топлива: 50 руб./литр.
- Средний расход топлива: 12 л/100 км.
- Средняя заработная плата водителя: 30 тыс. рублей/мес.
- Количество рабочих часов в месяце: 180 ч.
- Среднее количество километров за рабочий день: 100 км/день.

Решение:

1. Рассчитаем месячный пробег автомобиля:

Месячный пробег = $100 \text{ км/день} \times 22 \text{ рабочих дня} = 2200 \text{ км}$.

2. Определим ежемесячные расходы на топливо:

$$\begin{aligned} \text{Расходы на топливо} &= \frac{\text{Средний расход топлива}}{100} \times \text{Месячный пробег} \times \\ &\times \text{Стоимость топлива} = \frac{12}{100} \times 2200 \times 50 = 13200 \text{ руб.} \end{aligned}$$

3. Вычислим среднемесячную зарплату водителя на километр пробега:

$$\text{Заработная плата на километр} = \frac{\text{Зарплата водителя}}{\text{Месячный пробег}} = \frac{30000}{2200} = 13,64 \text{ руб./км.}$$

4. Итоговая средняя стоимость километра пробега:

$$\begin{aligned} \text{Общая стоимость километра} &= \frac{\text{Расходы на топливо} + \text{Зарплата водителя}}{\text{Месячный пробег}} = \\ &= \frac{13200 + 30000}{2200} \approx 19,64 \text{ руб./км.} \end{aligned}$$

Таким образом, средняя стоимость транспортировки на автомобиле УАЗ Патриот составляет примерно 19,64 рубля за каждый пройденный километр.

Пример. Подразделение ОВД осуществляет доставку товаров питания сотрудникам в два удаленных участка города с интервалом в три дня. Необходимо определить оптимальный маршрут доставки, минимизирующий транспортные расходы.

Имеются следующие исходные данные:

- Расстояние от склада до пункта А: 30 км.
- Расстояние от склада до пункта Б: 40 км.
- Общее расстояние маршрута: 70 км.
- Тариф на перевозку груза автомобилем Газель Некст: 15 руб./км.

Возможны варианты:

1. Поездка дважды за неделю: сначала в пункт А, потом отдельно в пункт Б.
2. Одновременная доставка в оба пункта за одну поездку.

Решение:

Вариант 1 (две отдельные поездки):

$$\text{Общий километраж} = 30 \times 2 + 40 \times 2 = 140 \text{ км.}$$

$$\text{Общие расходы} = 140 \text{ км.} \times 15 \text{ руб./км.} = 2100 \text{ руб.}$$

Вариант 2 (одна поездка):

Общий километраж = $30 + 40 = 70 \text{ км.}$

Общие расходы = $70 \text{ км.} \times 15 \text{ руб./км.} = 1050 \text{ руб.}$

Оптимальным решением становится вариант второй – совместная доставка грузов в оба пункта, позволяющая сократить затраты почти вдвое.

Эти расчеты показывают важность грамотного планирования перевозок и учета всех факторов, влияющих на итоговую стоимость транспортировки.

1.8.2. Изучение распределения запасов на складе

Распределение запасов на складе играет ключевую роль в эффективном управлении материальными ресурсами. Правильное распределение обеспечивает бесперебойное обеспечение подразделений необходимыми материалами и оборудованием, предотвращает дефицит и излишки продукции, снижает риск порчи и устаревания запасов.

Для изучения распределения запасов используются статистические методы, позволяющие выявить закономерности размещения материалов, оценить степень равномерности распределения, обнаружить аномалии и диспропорции.

Рассмотрим наиболее распространенные методы анализа распределения запасов на примере конкретной ситуации.

Пример. Предположим, на центральном складе органа внутренних дел хранятся запасы трех видов униформы: зимняя форма, демисезонная форма и летняя форма. Требуется проверить, равномерно ли распределены запасы каждой категории относительно общего объема хранения.

Данные по запасам формы:

Категория	Запасы
Зимняя форма	200 шт.
Демисезонная форма	300 шт.
Летняя форма	500 шт.

Решение.

Используем коэффициент вариации для оценки однородности распределения:

$$\tilde{v} = \frac{s}{\bar{x}} \times 100\%,$$

где $\bar{x}$ – среднее значение, s – стандартное отклонение.

Вычислим среднее арифметическое и стандартное отклонение запасов:

$$\bar{x} = \frac{200 + 300 + 500}{3} = 333,3 \text{ шт.}$$

Стандартное отклонение:

$$s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n}} = \sqrt{\frac{(200 - 333,3)^2 + (300 - 333,3)^2 + (500 - 333,3)^2}{3}} \approx 152,75$$

Коэффициент вариации:

$$\tilde{v} = \frac{152,75}{333,3} \times 100\% = 45,8\%.$$

Высокий показатель коэффициента вариации свидетельствует о неравномерности распределения запасов форм одежды по категориям.

Использование статистических методов позволяет эффективно анализировать и контролировать распределение запасов на складах подразделений ОВД, обеспечивая надежность функционирования служб материального обеспечения и рациональное использование бюджетных средств.

1.9. Применение возможностей *Excel* для анализа выборочных совокупностей

1.9.1. Исследование выборочной совокупности с помощью электронных таблиц *Excel*

Для вычисления числовых характеристик выборочной совокупности удобно применять электронные таблицы *Excel*. Продемонстрируем это на конкретном примере.

Пример. Зарегистрированы данные о лицах, совершивших кражи:

Возраст	Число выявленных лиц, совершивших преступление
10-14	25
14-18	17
18-22	19
22-26	21
26-30	18
30-34	30

34-38	16
38-42	20
42-46	16
46-50	23

На основании этих данных необходимо вычислить объем выборки n , размах R , среднюю выборочную $\bar{x}_e$, выборочную дисперсию D_e , выборочное среднее квадратическое отклонение σ_e , исправленную дисперсию S^2 , исправленное среднее квадратическое отклонение S , начальные ν_k ($k=1, 2, 3, 4$) и центральные моменты μ_k ($k=1, 2, 3, 4$), асимметрию $\tilde{A}$, эксцесс $\tilde{E}$.

Решение. На рабочем листе книги *Excel* составим дискретный вариационный ряд распределения частот.

	A	B	C
1	№	x_i	n_i
2	1	2	3
3	1	12	25
4	2	16	17
5	3	20	19
6	4	24	21
7	5	28	18
8	6	32	30
9	7	36	16
10	8	40	20
11	9	44	16
12	10	48	23

Для нахождения всех числовых характеристик выборочной совокупности дополним таблицу вспомогательными столбцами.

	A	B	C	D	E	F	G	H	I	J	K
1	№	x_i	n_i	w_i	$x_i w_i$	x_i^2	$x_i^2 w_i$	x_i^3	$x_i^3 w_i$	x_i^4	$x_i^4 w_i$
2	1	2	3	4	5	6	7	8	9	10	11
3	1	12	25	0,121951	1,463415	144	17,56098	1728	210,7317	20736	2528,780488
4	2	16	17	0,082927	1,326829	256	21,22927	4096	339,6683	65536	5434,692683
5	3	20	19	0,092683	1,853659	400	37,07317	8000	741,4634	160000	14829,26829
6	4	24	21	0,102439	2,458537	576	59,00488	13824	1416,117	331776	33986,80976
7	5	28	18	0,087805	2,458537	784	68,83902	21952	1927,493	614656	53969,79512
8	6	32	30	0,146341	4,682927	1024	149,8537	32768	4795,317	1048576	153450,1463
9	7	36	16	0,078049	2,809756	1296	101,1512	46656	3641,444	1679616	131091,9805
10	8	40	20	0,097561	3,902439	1600	156,0976	64000	6243,902	2560000	249756,0976
11	9	44	16	0,078049	3,434146	1936	151,1024	85184	6648,507	3748096	292534,322
12	10	48	23	0,112195	5,385366	2304	258,4976	110592	12407,88	5308416	595578,3805
13	сумма		205	1	29,77561		1020,41		38372,53		1533160,273

	L	M	N	O	P	Q	R	S
1	$x_i - x_{\bar{e}}$	$(x_i - x_{\bar{e}})w_i$	$(x_i - x_{\bar{e}})^2$	$(x_i - x_{\bar{e}})^2 w_i$	$(x_i - x_{\bar{e}})^3$	$(x_i - x_{\bar{e}})^3 w_i$	$(x_i - x_{\bar{e}})^4$	$(x_i - x_{\bar{e}})^4 w_i$
2	12	13	14	15	16	17	18	19
3	-17,7756	-2,16776	315,9723	38,53321	-5616,6	-684,951	99838,5	12175,43
4	-13,7756	-1,14237	189,7674	15,73681	-2614,16	-216,784	36011,68	2986,334
5	-9,77561	-0,90603	95,56255	8,857016	-934,182	-86,5827	9132,2	846,399
6	-5,77561	-0,59165	33,35767	3,417127	-192,661	-19,736	1112,734	113,9874
7	-1,77561	-0,15591	3,15279	0,27683	-5,59812	-0,49154	9,940085	0,872788
8	2,22439	0,325521	4,947912	0,724085	11,00609	1,610647	24,48183	3,582707
9	6,22439	0,485806	38,74303	3,023847	241,1518	18,8216	1501,023	117,153
10	10,22439	0,997501	104,5382	10,19884	1068,839	104,277	10928,23	1066,168
11	14,22439	1,110196	202,3333	15,79187	2878,068	224,6297	40938,76	3195,22
12	18,22439	2,044688	332,1284	37,26319	6052,838	679,0988	110309,3	12376,16
13		0		133,8228		19,89203		32881,31

Легко заметить, что сумма столбца 3 соответствует объему выборки,

$$\text{т.е. } n = \sum_{i=1}^k n_i = 205.$$

Сумма столбца 5 соответствует значению средней выборочной и начальному моменту первого порядка, т.е. $\bar{x}_e = \sum_{i=1}^m x_i \omega_i = 29,77561$ и

$$\nu_1 = \sum_{i=1}^m x_i \cdot \omega_i = 29,77561.$$

Сумма столбца 7 определяет начальный момент второго порядка, т.е.

$$\nu_2 = \sum_{i=1}^m x_i^2 \cdot \omega_i = 1020,41.$$

Сумма столбца 9 определяет начальный момент третьего порядка, т.е.

$$\nu_3 = \sum_{i=1}^m x_i^3 \cdot \omega_i = 38372,53.$$

Сумма столбца 11 соответствует значению начального момента четвертого порядка, т.е. $\nu_4 = \sum_{i=1}^m x_i^4 \cdot \omega_i = 1533160.$

Сумма столбца 13 определяет значение центрального момента первого порядка, т.е. $\mu_1 = \sum_{i=1}^m (x_i - \bar{x}_e) \cdot \omega_i = 0.$

Сумма столбца 15 соответствует значению выборочную дисперсию и центральному моменту второго порядка, т.е. $D_e = \sum_{i=1}^m (x_i - \bar{x}_e)^2 \cdot \omega_i = 133,8228$ и

$$\mu_2 = \sum_{i=1}^m (x_i - \bar{x}_e)^2 \cdot \omega_i = 133,8228.$$

Сумма столбца 17 определяет значение центрального момента третьего порядка: $\mu_3 = \sum_{i=1}^m (x_i - \bar{x}_e)^3 \cdot \omega_i = 19,89203$.

Сумма столбца 19 соответствует значению центральный момент четвертого порядка: $\mu_4 = \sum_{i=1}^m (x_i - \bar{x}_e)^4 \cdot \omega_i = 32881,31$.

	A	B	C	D	E	F	G	H	I	J	K
1	№	x_i	n_i	w_i	$x_i w_i$	x_i^2	$x_i^2 w_i$	x_i^3	$x_i^3 w_i$	x_i^4	$x_i^4 w_i$
2	1	2	3	4	5	6	7	8	9	10	11
3	1	12	25	0,121951	1,463415	144	17,56098	1728	210,7317	20736	2528,780488
4	2	16	17	0,082927	1,326829	256	21,22927	4096	339,6683	65536	5434,692683
5	3	20	19	0,092683	1,853659	400	37,07317	8000	741,4634	160000	14829,26829
6	4	24	21	0,102439	2,458537	576	59,00488	13824	1416,117	331776	33986,80976
7	5	28	18	0,087805	2,458537	784	68,83902	21952	1927,493	614656	53969,79512
8	6	32	30	0,146341	4,682927	1024	149,8537	32768	4795,317	1048576	153450,1463
9	7	36	16	0,078049	2,809756	1296	101,1512	46656	3641,444	1679616	131091,9805
10	8	40	20	0,097561	3,902439	1600	156,0976	64000	6243,902	2560000	249756,0976
11	9	44	16	0,078049	3,434146	1936	151,1024	85184	6648,507	3748096	292534,322
12	10	48	23	0,112195	5,385366	2304	258,4976	110592	12407,88	5308416	595578,3805
13	сумма		205	1	29,77561		1020,41		38372,53		1533160,273
14											
15											
16											

	L	M	N	O	P	Q	R	S
1	$x_i - x_e$	$(x_i - x_e) w_i$	$(x_i - x_e)^2$	$(x_i - x_e)^2 w_i$	$(x_i - x_e)^3$	$(x_i - x_e)^3 w_i$	$(x_i - x_e)^4$	$(x_i - x_e)^4 w_i$
2	12	13	14	15	16	17	18	19
3	12	1,463415	144	17,56098	1728	210,7317	20736	2528,78
4	16	1,326829	256	21,22927	4096	339,6683	65536	5434,693
5	20	1,853659	400	37,07317	8000	741,4634	160000	14829,27
6	24	2,458537	576	59,00488	13824	1416,117	331776	33986,81
7	28	2,458537	784	68,83902	21952	1927,493	614656	53969,8
8	32	4,682927	1024	149,8537	32768	4795,317	1048576	153450,1
9	36	2,809756	1296	101,1512	46656	3641,444	1679616	131092
10	40	3,902439	1600	156,0976	64000	6243,902	2560000	249756,1
11	44	3,434146	1936	151,1024	85184	6648,507	3748096	292534,3
12	48	5,385366	2304	258,4976	110592	12407,88	5308416	595578,4
13		29,77561		1020,41		38372,53		1533160
14								
15								
16								

Остальные числовые характеристики не сложно найти по известным формулам:

– размах: $R = x_{\max} - x_{\min} = 48 - 12 = 36$;

– выборочное среднее квадратическое отклонение: $\sigma_{\epsilon} = \sqrt{D_{\epsilon}} = 11,56818$;

– исправленную дисперсию: $S^2 = \frac{n}{n-1} D_{\epsilon} = 134,4788$;

– исправленное среднее квадратическое отклонение: $S = \sqrt{S^2} = 11,5965$;

– асимметрия: $\tilde{A} = \frac{\tilde{\mu}_3}{\sigma_{\epsilon}^3} = \frac{19,89203}{11,56818^3} = 0,012849$;

– эксцесс: $\tilde{E} = \frac{\tilde{\mu}_4}{\sigma_{\epsilon}^4} - 3 = \frac{32881,31}{11,56818^4} - 3 = -1,16393$.

Пример. При статистическом обследовании семей получены сведения о возрасте детей, являющихся студентами ВУЗов:

Дети студенты ВУЗов по возрасту, лет (x)	Число семей, (n)
17	18
18	46
19	58
20	12
21	23
22	29
23	35
24	19

Необходимо рассчитать показатели вариации среднего возраста детей-студентов ВУЗов.

Решение. Составим и заполним таблицу:

x_i	n_i	$x_i n_i$	$x_i - \bar{x}_{cp}$	$(x_i - \bar{x}_{cp})^2$	$(x_i - \bar{x}_{cp})^2 n_i$
17	18	306	-3,242	10,508	189,1513
18	46	828	-2,242	5,0251	231,1532
19	58	1102	-1,242	1,5417	89,42069
20	12	240	-0,242	0,0584	0,700833
21	23	483	0,7583	0,5751	13,2266
22	29	638	1,7583	3,0917	89,66035
23	35	805	2,7583	7,6084	266,2941
24	19	456	3,7583	14,125	268,3763
Σ	240	4858			1147,983

Общее количество семей: $n=240$.

1. Средний возраст детей-студентов определим по формуле средней выборочной:

$$\bar{x}_e = \frac{1}{n} \sum_{i=1}^m x_i n_i = \frac{4858}{240} = 20,24 \text{ (года)}.$$

2. Вычислим отклонения индивидуального возраста студента от среднего $(x_i - \bar{x}_e)$, а также квадрат отклонений $(x_i - \bar{x}_e)^2$. Далее вычислим дисперсию и среднее квадратическое отклонение:

$$\sigma_e^2 = \frac{1}{n} \sum_{i=1}^m (x_i - \bar{x}_e)^2 n_i = \frac{1147,983}{240} = 4,78 \text{ (года)};$$

$$\sigma_e = \sqrt{\sigma_e^2} = 2,19 \text{ (года)}.$$

Таким образом, индивидуальное значение возраста студента в среднем отклоняется на 2,19 года.

4. Вычислим коэффициент вариации:

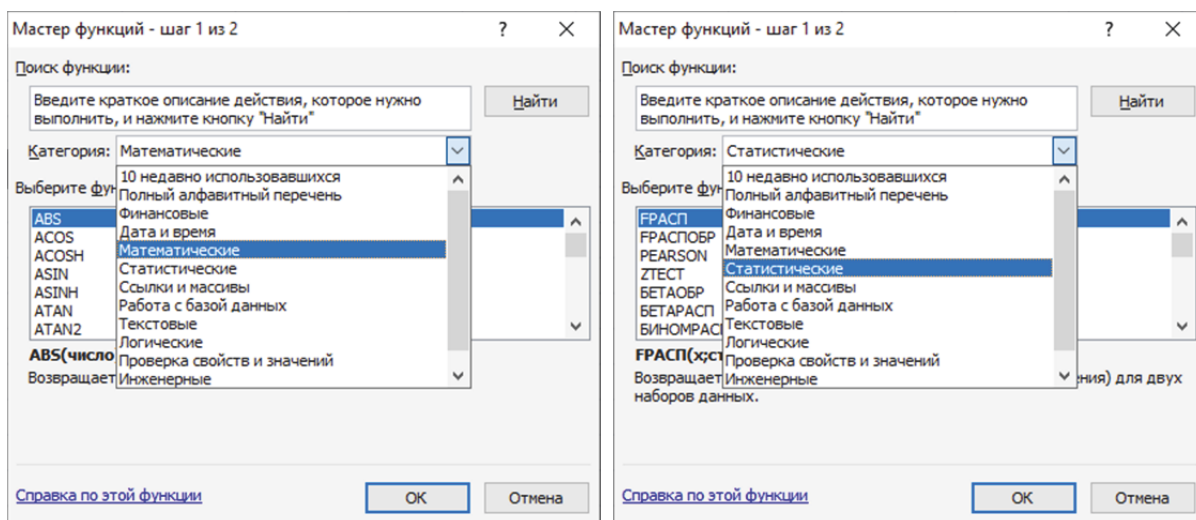
$$\tilde{v} = \frac{\sigma_e(x)}{\bar{x}_e} \cdot 100\% = \frac{2,19}{20,24} \cdot 100\% = 10,8\%.$$

Полученное значение показывает, что в результате статистического наблюдения исследуемая статистическая совокупность распределения детей-студентов ВУЗов в семьях является однородной, т.к. $\tilde{v} = 10,8\% < 33\%$.

1.9.2. Вычисление числовых характеристик выборочной совокупности с помощью встроенных функций

Важным инструментом обработки данных в электронных таблицах являются встроенные функции.

Основными разделами библиотеки табличного процессора *Excel* являются *Статистические* и *Математические функции*. Эти разделы предназначены для решения некоторых наиболее востребованных задач теории вероятностей и математической статистики.



Рассмотрим некоторые функции, применяемые для вычисления числовых характеристик выборочной совокупности:

ДИСП(число1;число2; ...) –оценивает дисперсию по выборке;

КВАДРОТКЛ(число1;число2;...) – возвращает сумму квадратов отклонений точек данных от их среднего;

КОРЕНЬ(число) – возвращает положительное значение квадратного корня;

МАКС(число1;число2; ...) – возвращает наибольшее значение в списке аргументов;

МЕДИАНА(число1;число2;...) – возвращает медиану заданных чисел;

МИН(число1;число2; ...) – возвращает наименьшее значение в списке аргументов;

МОДА(число1;число2; ...) – возвращает значение моды набора данных;

СРЗНАЧ(число1; число2; ...) – возвращает среднее арифметическое своих аргументов;

СТАНДОТКЛОН(число1; число2; ...) – оценивает стандартное отклонение по выборке;

СУММ(число1;число2; ...) – суммирует аргументы;

СУММПРОИЗВ(массив1;массив2;массив3; ...) – перемножает соответствующие элементы заданных массивов и возвращает сумму произведений;

СЧЕТ(значение1; значение2; ...) – подсчитывает количество чисел в списке аргументов;

ЧАСТОТА(массив_данных;массив_интервалов) – возвращает распределение частот в виде вертикального массива;

ЭКСЦЕСС(число1;число2; ...) – возвращает эксцесс множества данных.

Пример. В результате некоторого эксперимента получены следующие статистические данные: 183, 158, 171, 173, 158, 185, 168, 180, 168, 178, 169, 170, 183, 170, 173, 174, 178, 180, 180, 183, 155, 183, 178, 188, 190, 170.

Необходимо с помощью статистических функций Excel вычислить объем выборки n , размах R , моду M_o , медиану M_e , среднюю выборочную $\bar{x}_e$, выборочную дисперсию D_e , выборочное среднее квадратическое отклонение σ_e , исправленную дисперсию S^2 , исправленное среднее квадратическое отклонение S , эксцесс $\tilde{E}$.

Решение. Запишем исходные данные на лист книги Excel и, применяя встроенные функции, вычислим все числовые характеристики.

	A	B
1	№	x_i
2	1	183
3	2	158
4	3	171
5	4	173
6	5	158
7	6	185
8	7	168
9	8	180
10	9	168
11	10	178
12	11	169
13	12	170
14	13	183
15	14	170
16	15	173
17	16	174
18	17	178
19	18	180
20	19	180
21	20	183
22	21	155
23	22	183
24	23	178
25	24	188
26	25	190
27	26	170

	D	E	F	G	H	I
1						
2	Объем выборки:				N=	=СЧЁТ(B2:B27)
3	Наименьшая варианта:				$x_{\min}$ =	=МИН(B2:B27)
4	Наибольшая варианта:				$x_{\max}$ =	=МАКС(B2:B27)
5	Размах выборки:				R=	=МАКС(B2:B27)-МИН(B2:B27)
6	Мода:				M_0 =	=МОДА(B2:B27)
7	Медиана:				M_e =	=МЕДИАНА(B2:B27)
8	Средняя выборочная:				$\bar{x}_B$ =	=СРЗНАЧ(B2:B27)
9	Исправленная дисперсия:				S^2_B =	=ДИСП(B2:B27)
10	Исправленное ср. квадр. отклонение:				S_B =	=СТАНДОТКЛОН(B2:B27)
11	Выборочная дисперсия:				D_B =	=(I2-1)/I2*I9
12	Среднее квадратическое отклонение:				σ_B =	=КОРЕНЬ(I11)
13	Сумма всех вариантов:				Σx_i =	=СУММ(B2:B27)
14	Экспесс:				E_s =	=ЭКСПЕСС(B2:B27)

В результате получим следующие данные:

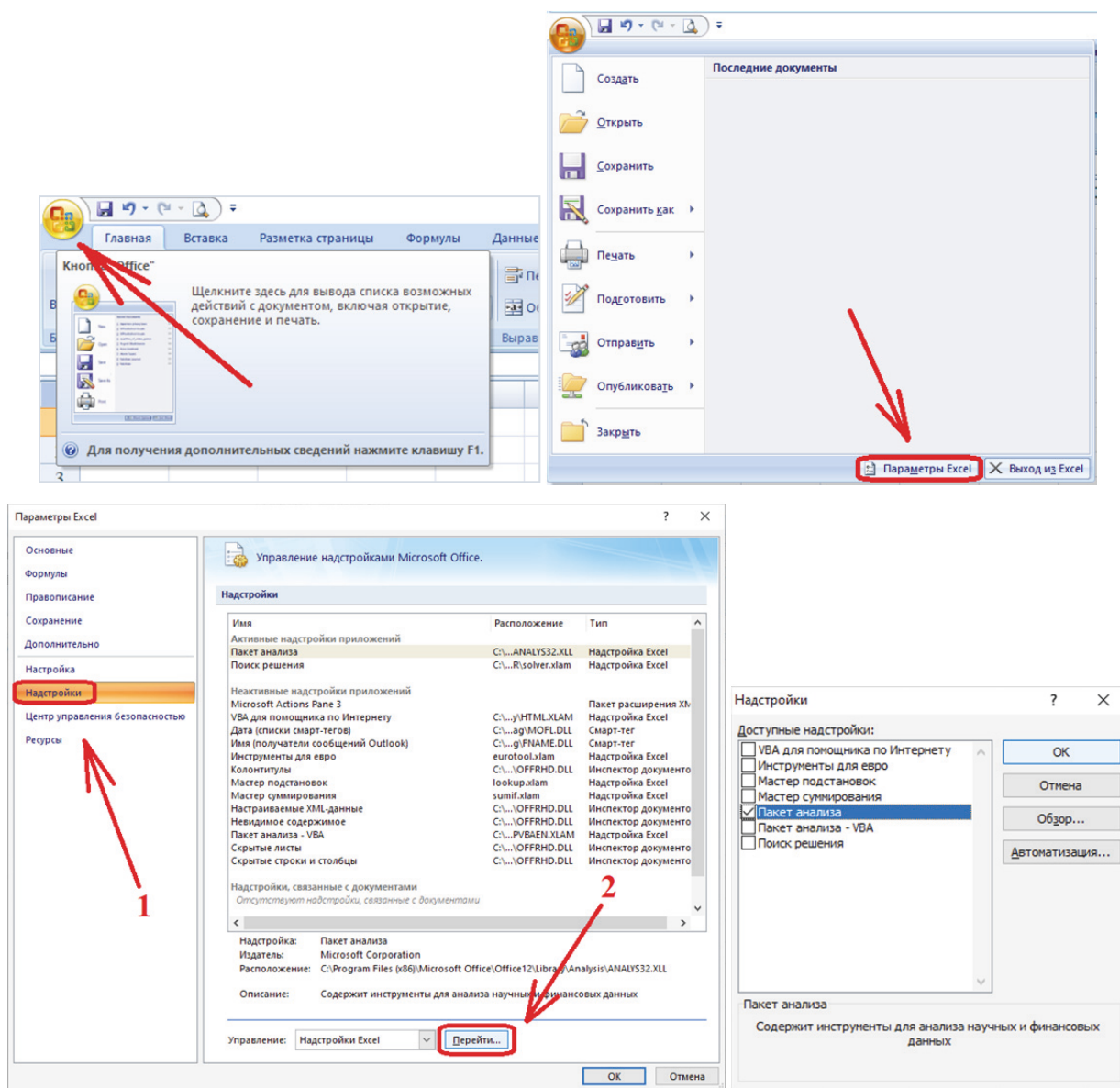
	D	E	F	G	H	I
1						
2	Объем выборки:				N=	26
3	Наименьшая варианта:				$x_{\min}$ =	155
4	Наибольшая варианта:				$x_{\max}$ =	190
5	Размах выборки:				R=	35
6	Мода:				M_0 =	183
7	Медиана:				M_e =	176
8	Средняя выборочная:				$\bar{x}_B$ =	174,8461538
9	Исправленная дисперсия:				S^2_B =	82,53538462
10	Исправленное ср. квадр. отклонение:				S_B =	9,084898712
11	Выборочная дисперсия:				D_B =	79,36094675
12	Среднее квадратическое отклонение:				σ_B =	8,908476118
13	Сумма всех вариантов:				Σx_i =	4546
14	Экспесс:				E_s =	-0,132939724

1.9.3. Обработка статистических данных с помощью надстройки Анализ данных

Наиболее распространенными средствами анализа данных являются статистические процедуры *Пакета анализа*. Они обладают большими возможностями, чем статистические функции. С их помощью можно решать более сложные задачи обработки статистических данных и выполнять более тонкий анализ этих данных.

Для доступа к процедурам *Пакета анализа* необходимо выполнить следующие действия:

- зайти в параметры *Excel*;
- перейти в надстройки;
- установить флажок *Пакет анализа*;
- нажать ОК.



Пример. В результате некоторого эксперимента получены следующие статистические данные: 183, 158, 171, 173, 158, 185, 168, 180, 168, 178, 169, 170, 183, 170, 173, 174, 178, 180, 180, 183, 155, 183, 178, 188, 190, 170.

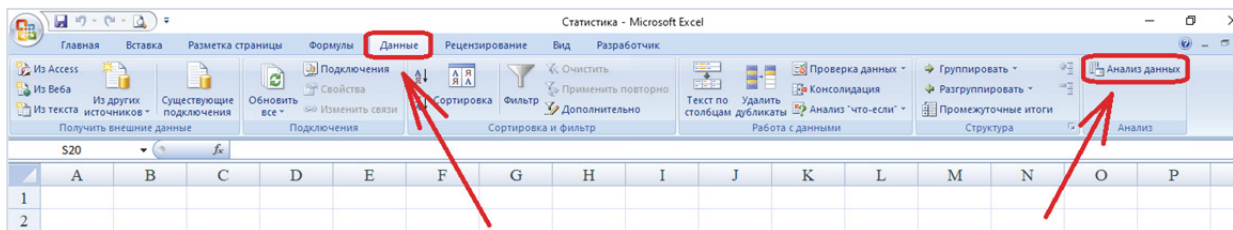
1) Исследовать статистические данные с помощью надстройки Анализ данных Excel.

2) Построить интервальный вариационный ряд распределения частот, разбив исходные данные на интервалы равной длины $h=5$ и $x_{нач} = 155$.

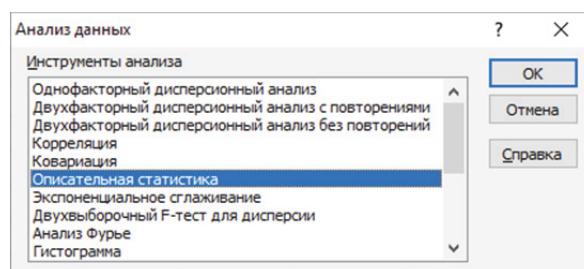
Решение. 1) Запишем исходные данные на лист книги *Excel*.

	A	B
1	№	x_i
2	1	183
3	2	158
4	3	171
5	4	173
6	5	158
7	6	185
8	7	168
9	8	180
10	9	168
11	10	178
12	11	169
13	12	170
14	13	183
15	14	170
16	15	173
17	16	174
18	17	178
19	18	180
20	19	180
21	20	183
22	21	155
23	22	183
24	23	178
25	24	188
26	25	190
27	26	170

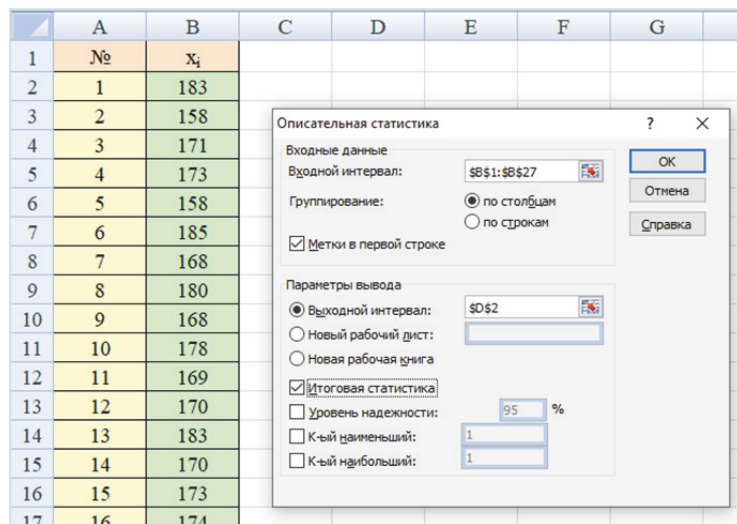
В меню *Данные* выбираем *Анализ данных*.



В появившемся окне выбираем *Описательная статистика*.



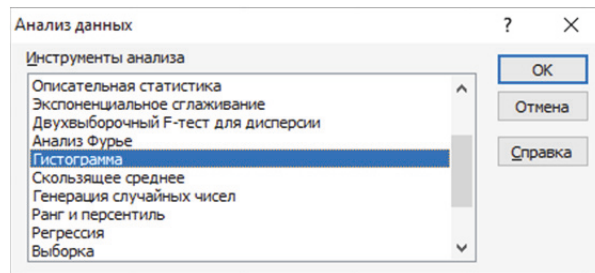
Заполняем диалоговое окно.



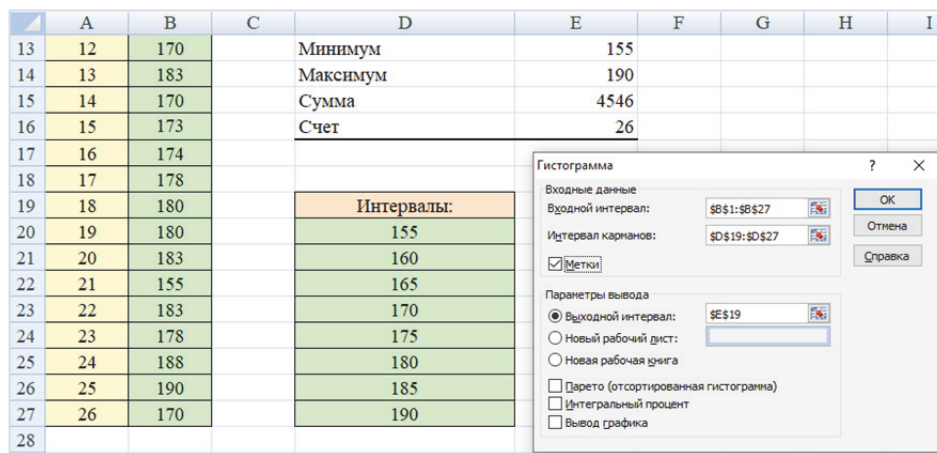
В качестве *Входной интервал* указываем столбец с исходными статистическими данными (включая название столбца); *Выходной интервал* – любая свободная ячейка рабочего листа; устанавливаем флажки – *Метки в первой строке* и *Итоговая статистика*. В результате получаем данные:

	A	B	C	D	E
1	№	x_i			
2	1	183		$\bar{x}$	
3	2	158			
4	3	171		Среднее	174,8461538
5	4	173		Стандартная ошибка	1,781695224
6	5	158		Медиана	170
7	6	185		Мода	183
8	7	168		Стандартное отклонение	9,084898712
9	8	180		Дисперсия выборки	82,53538462
10	9	168		Экссесс	-0,132939724
11	10	178		Асимметричность	-0,529111747
12	11	169		Интервал	35
13	12	170		Минимум	155
14	13	183		Максимум	190
15	14	170		Сумма	4540
16	15	173		Счет	20

Для построения интервального вариационного ряда воспользуемся процедурой *Гистограмма* надстройки *Анализ данных*.



В качестве *Входной интервал* указываем столбец с исходными статистическими данными (включая название столбца); *Интервал карманов* – столбец «*Интервалы:*» (включая название столбца); *Выходной интервал* – любая свободная ячейка; устанавливаем флажок –*Метки*.



В результате получаем следующие значения:

	D	E	F
19	Интервалы:	Интервалы:	Частота
20	155	155	1
21	160	160	2
22	165	165	0
23	170	170	6
24	175	175	4
25	180	180	6
26	185	185	5
27	190	190	2
28		Еще	0

Таким образом, интервальный вариационный ряд имеет вид:

≤ 155	(155;160]	(160;165]	(165;170]	(170;175]	(175;180]	(180;185]	(185;190]
1	2	0	6	4	6	5	2

2. Инференциальная статистика

2.1. Закон больших чисел и центральная предельная теорема

Законы больших чисел и центральная предельная теорема (ЦПТ) являются фундаментальными результатами теории вероятности, позволяющими изучать поведение сумм большого числа случайных величин.

2.1.1. Закон больших чисел

Закон больших утверждает, что при увеличении количества экспериментов средние значения случайных величин стремятся к своему ожидаемому значению. Существует несколько вариантов закона больших чисел, наиболее известные из которых: теорема Бернулли и закон больших чисел Колмогорова.

Теорема Бернулли. Пусть производится серия независимых испытаний с двумя возможными исходами («успех» и «неудача»), причем вероятность успеха в каждом испытании постоянна и равна p . Тогда частота появления события стремится к его вероятности при большом числе испытаний.

Т.е. $\frac{S_n}{n} \rightarrow p$ при $n \rightarrow \infty$, где S_n – число успеха в первых n испытаниях.

Например, монета подбрасывается многократно. Вероятность выпадения орла равняется $p=1/2$. Согласно закону больших чисел, при большом числе бросков отношение числа выпавших орлов к общему числу бросаний приближенно равно $1/2$.

Закон больших чисел Колмогорова. Пусть имеется последовательность независимых одинаково распределенных случайных величин $X_1, X_2, \dots, X_n$, каждая из которых имеет конечное математическое ожидание $M(X_i) = a$ и конечную дисперсию $D(X_i) = \sigma^2 < \infty$. Тогда выполняется условие:

$$\lim_{n \rightarrow \infty} P\left(\left|\frac{1}{n} \sum_{i=1}^n X_i - a\right| \geq \varepsilon\right) = 0.$$

Т.е. среднее арифметическое последовательности сходится к среднему значению случайной величины при неограниченном росте n .

Пример. Предположим, что завод выпускает лампочки, и известно, что средняя продолжительность горения одной лампочки составляет 1000 часов. Необходимо проверить, насколько близко среднее время горения к этому значению, если собрано 100 образцов лампочек, и суммарное время горения составило 102000 часов?

Решение.

Количество лампочек: $n=100$, общее время горения: $S_n=102000$ ч.

Средняя продолжительность горения:

$$\frac{S_n}{n} = \frac{102000}{100} = 1020 \text{ ч.}$$

Т.е. среднее отличается от ожидаемого значения (1000 часов) незначительно, однако согласно закону больших чисел, при дальнейшем увеличении размера выборки расхождение должно уменьшиться ещё сильнее.

Таким образом, средняя продолжительность горения близка к заявленной производителем длительности.

2.1.2. Центральная предельная теорема

Центральная предельная теорема описывает распределение суммы независимых случайных величин при условии, что их количество велико. Эта теорема гласит, что сумма большого числа независимых одинаково распределенных случайных величин приблизительно подчиняется нормальному распределению независимо от вида исходного распределения каждой отдельной случайной величины.

Теорема. Пусть дана последовательность независимых одинаково распределенных случайных величин $Y_1, Y_2, \dots, Y_n$ с математическим ожиданием $M(Y_i) = a$ и дисперсией $D(Y_i) = \sigma^2$. Рассмотрим сумму: $S_n = Y_1 + Y_2 + \dots + Y_n$.

Нормированная величина:

$$Z_n = \frac{S_n - na}{\sigma\sqrt{n}}.$$

При $n \rightarrow \infty$ эта величина асимптотически распределяется нормально с параметрами $(0,1)$, то есть $N(0,1)$.

Это упрощает многие расчеты, поскольку нормальное распределение обладает известными свойствами и легко обрабатывается.

Пример. Игральный кубик бросают 100 раз подряд. Какова вероятность того, что сумма очков окажется больше 350?

Решение.

Вероятность выпадения любого числа на кубике равна $1/6$, а среднее значение очка при одном броске:

$$a = M(X) = 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + \dots + 6 \cdot \frac{1}{6} = 3,5.$$

Дисперсия одного броска:

$$\sigma^2 = D(X) = \frac{1}{6} \left((1-3,5)^2 + (2-3,5)^2 + \dots + (6-3,5)^2 \right) = \frac{35}{12} \approx 2,917$$

Используя центральную предельную теорему, строим нормированную величину:

$$Z_n = \frac{S_n - na}{\sigma\sqrt{n}} = \frac{350 - 100 \cdot 3,5}{\sqrt{2,917} \cdot \sqrt{100}} \approx -\frac{5}{5,4} \approx -0,93.$$

Необходимо найти вероятность того, что сумма очков превысит 350, т.е.:

$$P(S_n > 350) = P(Z > -0,93) \approx \Phi(0,93) \approx 0,8238.$$

Таким образом, вероятность того, что сумма очков при 100 бросках игрального кубика превысит 350, примерно равна 82,38%.

2.2. Статистические оценки параметров распределения

2.2.1. Оценки параметров

Задачей статистики является *описание характера распределения* некоторого признака в генеральной совокупности на основании изучения этого признака у некоторой части совокупности (выборки), полученной в результате случайного отбора.

При этом распределение относительных частот в выборке рассматривается как *эмпирическое приближение* к теоретическому распределению вероятностей в генеральной совокупности.

Выяснение закона распределения по данным выборки составляет *главную проблему математической статистики*, так как на основании закона распределения изучаемого признака можно решать задачи по анализу и предсказанию результатов массового процесса.

На практике часто теоретический закон распределения случайной величины в генеральной совокупности известен (или построено его приближенное аналитическое представление), т.е. известно, что закон распределения принадлежит к тому или иному семейству (нормальный закон, закон Пуассона и т.д.), зависящему от одного или нескольких параметров.

Если бы точные значения параметров были известны, например, μ и σ при нормальном законе, λ при законе Пуассона, то и закон распределения был бы полностью определен. Поэтому именно для определения этих параметров и проводится само *статистическое исследование*.

Пусть изучается случайная величина X с законом распределения, зависящим от одного или нескольких параметров. Требуется по выборке $X_1, X_2, \dots, X_n$, полученной в результате n наблюдений (опытов), оценить неизвестный параметр θ .

$X_1, X_2, \dots, X_n$ — это случайные величины, где X_1 — результат первого наблюдения, X_2 — второго и т.д., причем случайные величины X_i , $i = 1, 2, \dots, n$ имеют такое же распределение, что и случайная величина X . Конкретная выборка $x_1, x_2, \dots, x_n$ — это значения (реализация) независимых случайных величин $X_1, X_2, \dots, X_n$.

Статистической оценкой $\tilde{\theta}$ параметра θ теоретического распределения называют его приближенное значение, зависящее от данных выбора.

Оценка $\tilde{\theta}$ представляет собой значение некоторой функции результатов наблюдений над случайной величиной X :

$$\tilde{\theta} = \tilde{\theta}(X_1, X_2, \dots, X_n).$$

Функцию результатов наблюдений называют *выборочной функцией* или *статистикой*, а ее значение в приближенном равенстве – *оценкой*.

Можно сказать, что оценка $\tilde{\theta}$ параметра θ – это *статистика*, которая близка к истинному значению θ .

Оценка $\tilde{\theta}$ является случайной величиной, так как является функцией независимых случайных величин $X_1, X_2, \dots, X_n$. Если произвести другую выборку, то функция примет другое значение.

Если число опытов (наблюдений) невелико, то замена неизвестного параметра θ его оценкой $\tilde{\theta}$, например математического ожидания средним арифметическим, приводит к ошибке. Это ошибка в среднем тем больше, чем меньше число опытов.

К оценке любого параметра предъявляется ряд требований (*свойства статистических оценок*), которым она должна удовлетворять, чтобы быть «близкой» к истинному значению параметра: *свойства несмещенности, состоятельности, эффективности*.

Оценка $\tilde{\theta}$ параметра θ называется *несмещенной*, если математическое ожидание оценки равно оцениваемому параметру, т.е. $M(\tilde{\theta}) = \theta$.

В противном случае оценка называется *смещенной*.

Если $M(\tilde{\theta}) \rightarrow \theta$, то оценка $\tilde{\theta}$ называется *асимптотически несмещенной*.

Требование несмещенности является особенно важным при малом числе наблюдений (опытов).

Пусть $\tilde{\theta}$ – статистическая оценка неизвестного параметра θ теоретического распределения. Допустим, что по выборке объема n найдена оценка $\tilde{\theta}_1$.

Повторим опыт, т.е. извлечем из генеральной совокупности другую выборку того же объема и по ее данным найдем оценку $\tilde{\theta}_2$. Повторяя опыт

множественно, получим числа $\tilde{\theta}_1, \tilde{\theta}_2, \dots, \tilde{\theta}_k$, которые будут различны между собой. Таким образом, оценку $\tilde{\theta}$ можно рассматривать как случайную величину, а числа $\tilde{\theta}_1, \tilde{\theta}_2, \dots, \tilde{\theta}_k$ – как ее возможные значения.

Предположим, что оценка $\tilde{\theta}$ дает приближенное значение θ с избытком. В этом случае каждое найденное по данным выборки число $\tilde{\theta}_i$ ($i = 1, 2, \dots, k$) больше истинного значения θ . Следовательно, и математическое ожидание (среднее значение) случайной величины θ больше, чем $\tilde{\theta}$, т.е. $M(\tilde{\theta}) > \theta$. Очевидно, что если $\tilde{\theta}$ дает оценку с недостатком, то $M(\tilde{\theta}) < \theta$.

Таким образом, использование статистической оценки, математическое ожидание которой не равно оцениваемому параметру, привело бы к систематическим (одного знака) ошибкам.

По этой причине естественно потребовать, чтобы математическое ожидание оценки $\tilde{\theta}$ было равно оцениваемому параметру. Хотя соблюдение этого требования не устранит ошибок (одни значения $\tilde{\theta}$ больше, а другие меньше $\tilde{\theta}$), однако ошибки разных знаков будут встречаться одинаково часто. Иными словами, соблюдение требований $M(\tilde{\theta}) = \theta$ гарантирует от получения систематических ошибок.

Оценка $\tilde{\theta}$ параметра θ называется *состоятельной*, если она сходится по вероятности к оцениваемому параметру, т.е. для любого $\delta > 0$ выполнено условие:

$$\lim_{n \rightarrow \infty} P(|\tilde{\theta} - \theta| < \delta) = 1.$$

Это означает, что с увеличением объема выборки оценка параметра θ приближается к истинному его значению.

Свойство состоятельности обязательно для любого правила оценивания (несостоятельные оценки не используются).

Состоятельность оценки $\tilde{\theta}$ часто может быть установлена с помощью следующей теоремы.

Теорема 1. Если оценка $\tilde{\theta}$ параметра θ является несмещенной и $D(\theta) \rightarrow 0$ при $n \rightarrow \infty$, то $\tilde{\theta}$ – состоятельная оценка.

Несмещенная оценка $\tilde{\theta}$ параметра θ называется *эффективной*, если она имеет наименьшую дисперсию среди всех возможных несмещенных оценок параметра θ , т.е. оценка $\tilde{\theta}$ эффективна, если ее дисперсия минимальна.

На практике не всегда удастся удовлетворить всем перечисленным выше требованиям (несмещенность, состоятельность, эффективность). Поэтому приходится довольствоваться оценками, не обладающими сразу всеми тремя свойствами.

2.2.2. Точечные оценки математического ожидания и дисперсии

Оценки параметров подразделяются на *точечные* и *интервальные*. Точечная оценка параметра θ определяется одним числом $\tilde{\theta}$. При выборке малого объема точечная оценка может значительно отличаться от оцениваемого параметра, т.е. приводить к грубым ошибкам.

Интервальная оценка определяется двумя числами $\tilde{\theta}_1$ и $\tilde{\theta}_2$ – концами интервала, внутри которого содержится неизвестный параметр θ . Интервальные оценки позволяют установить *точность* и *надежность оценок*.

Пусть изучается случайная величина X с математическим ожиданием $a = M(X)$ и дисперсией $D(X)$. Оба параметра неизвестны.

Статистика, используемая в качестве приближенного значения неизвестного параметра генеральной совокупности, называется ее *точечной оценкой*. Т.е. точечная оценка характеристики генеральной совокупности – это число, определяемое по выборке.

Теорема 2. Пусть $X_1, X_2, \dots, X_n$ – выборка из генеральной совокупности и $M(X_i) = M(X) = a$, $D(X_i) = D(X)$. Тогда выборочное среднее $\bar{x}_g$ – *несмещенная* и *состоятельная* оценка математического ожидания $M(X)$.

При нормальном распределении случайной величины X оценка $\bar{x}_e$ будет и эффективной. На практике во всех случаях в качестве оценки математического ожидания используется среднее арифметическое, т.е. $\bar{x}_e$.

Обратимся к изучению дисперсии.

Рассмотрим равенство:

$$M(D_e) = \frac{n-1}{n} D(X).$$

Из данного равенства следует, что $M(D_e) \neq D(X)$, т.е. выборочная дисперсия является смещенной оценкой дисперсии $D(X)$.

Поэтому выборочную дисперсию исправляют, умножая ее на множитель $\left(\frac{n-1}{n}\right)$, получая формулу:

$$S^2 = \frac{n-1}{n} D_e.$$

Теорема 3. Пусть $X_1, X_2, \dots, X_n$ – выборка из генеральной совокупности и $M(X_i) = M(X) = a$, $D(X_i) = D(X)$. Тогда исправленная выборочная дисперсия S^2 – несмещенная и состоятельная оценка дисперсии $D(X)$.

Докажем несмещенность оценки S^2 .

$$M(S^2) = M\left(\frac{n-1}{n} D_e\right) = \frac{n-1}{n} M(D_e) = \frac{n-1}{n} \cdot \frac{n-1}{n} D(X) = D(X).$$

Следовательно, по определению S^2 – несмещенная оценка $D(X)$.

В контексте сказанного, имеют место следующие теоремы.

Теорема 4. Относительная частота появления события A в n независимых испытаниях является несмещенной, состоятельной и эффективной оценкой неизвестной вероятности $p = P(A)$ этого события (p – вероятность наступления события A в каждом испытании).

Теорема 5. Эмпирическая функция распределения выборки $F_n(x)$ является *несмещенной* и *состоятельной* оценкой функции распределения $F(x)$ случайной величины X .

Пример. Монету подбрасывают n раз. Вероятность выпадения орла при каждом подбрасывании равна p . В ходе опыта монета выпала орлом n_A раз. Необходимо доказать несмещенность оценки $\tilde{\theta} = \frac{n_A}{n}$ вероятности $\theta = p$ выпадения орла в каждом опыте.

Доказательство. Число успехов (n_A) имеет распределение Бернулли. Тогда $M(n_A) = np$; $D(n_A) = npq$. Так как

$$M(\tilde{\theta}) = M\left(\frac{n_A}{n}\right) = \frac{1}{n} M(n_A) = \frac{1}{n} np = p = \theta.$$

Следовательно, оценка $\tilde{\theta} = \frac{n_A}{n}$ является несмещенной.

2.2.3. Понятие интервального оценивания параметров

Точечные оценки неизвестного параметра θ хороши в качестве первоначальных результатов обработки наблюдений. Их недостаток – неизвестно с какой точностью они дают оцениваемый параметр. Для выборок небольшого объема в этом случае может наблюдаться большое расхождение между θ и $\tilde{\theta}$. Кроме того, при решении практических задач часто требуется определить надежность этих оценок. В этом случае возникает задача о приближении параметра θ не одним числом, а целым интервалом $(\tilde{\theta}_1; \tilde{\theta}_2)$.

Оценка неизвестного параметра называется *интервальной*, если она определяется двумя числами – концами интервала.

Пусть найденная по данным выборки статистическая характеристика $\tilde{\theta}$ служит оценкой неизвестного параметра θ . Будем считать θ постоянным числом (θ может быть и случайной величиной). Очевидно, что чем меньше абсолютная величина разности $|\theta - \tilde{\theta}|$, тем точнее $\tilde{\theta}$ определяет параметр θ .

Другими словами, если $\delta > 0$ и $|\theta - \tilde{\theta}| < \delta$, то чем меньше δ , тем оценка точнее. Таким образом, положительное число δ характеризует *точность оценки*.

Однако статистические методы не позволяют категорически утверждать, что оценка $\tilde{\theta}$ удовлетворяет неравенству $|\theta - \tilde{\theta}| < \delta$, можно лишь говорить о вероятности γ , с которой это неравенство осуществляется.

Надежностью (доверительной вероятностью) оценки называют вероятность $\gamma = 1 - \alpha$ с которой осуществляется неравенство $|\theta - \tilde{\theta}| < \delta$. Обычно надежность оценки задается наперед, причем в качестве γ берут число, близкое к единице. Наиболее часто задают надежность, равную 0,95; 0,99 и 0,999.

Значение α – уровень значимости, принимающее значения 0,05; 0,01; 0,001.

Пусть выполняется условие

$$P(|\theta - \tilde{\theta}| < \delta) = \gamma,$$
$$P(\tilde{\theta} - \delta < \theta < \tilde{\theta} + \delta) = \gamma.$$

Вероятность того, что интервал $(\tilde{\theta} - \delta; \tilde{\theta} + \delta)$ включает в себе (покрывает) неизвестный параметр θ , равна γ .

Доверительным называют интервал $(\tilde{\theta} - \delta; \tilde{\theta} + \delta)$, который покрывает неизвестный параметр с заданной надежностью γ . Метод доверительных интервалов разработал американский статистик Ю. Нейман, исходя из идей английского статистика Р. Фишера.

2.2.4. Доверительные интервалы для параметров нормального распределения

Доверительные интервалы для оценки математического ожидания при известной дисперсии

Пусть количественный признак X генеральной совокупности распределен нормально, причем среднее квадратическое отклонение σ этого распределения известно.

Требуется оценить неизвестное математическое ожидание a по выборочной средней $\bar{x}_g$. Необходимо найти доверительные интервалы, покрывающие параметр a с надежностью γ .

Будем рассматривать выборочную среднюю $\bar{x}_g$ как случайную величину $\bar{X}$ ($\bar{x}_g$ изменяется от выборки к выборке) и выборочные значения признака $x_1, x_2, \dots, x_n$ – как одинаково распределенные независимые случайные величины $X_1, X_2, \dots, X_n$ (эти числа также изменяются от выборки к выборке).

Другими словами, математическое ожидание каждой из этих величин равно a и среднее квадратическое отклонение – σ .

Если случайная величина X распределена нормально, то выборочная средняя $\bar{x}_g$, найденная по независимым наблюдениям, также распределена нормально. Параметры распределения $\bar{X}$:

$$M(\bar{X}) = a; \sigma(\bar{X}) = \frac{\sigma}{\sqrt{n}}.$$

Данные равенства получены из следующих рассуждений.

Рассмотрим n взаимно независимых случайных величин $X_1, X_2, \dots, X_n$, которые имеют одинаковые распределения, а следовательно, и одинаковые характеристики (математическое ожидание, дисперсию и др.). Наибольший интерес представляет изучение числовых характеристик среднего арифметического этих величин $\bar{X}$.

Рассмотрим следующие положения, которые устанавливают связь между числовыми характеристиками среднего арифметического и соответствующими характеристиками каждой отдельной величины.

1. Математическое ожидание среднего арифметического n одинаково распределенных взаимно независимых случайных величин равно математическому ожиданию a каждой из величин:

$$M(\bar{X}) = a.$$

Доказательство.

$$M(\bar{X}) = M\left(\frac{X_1 + X_2 + \dots + X_n}{n}\right) = \frac{M(X_1) + M(X_2) + \dots + M(X_n)}{n} = \frac{na}{n} = a.$$

2. Дисперсия среднего арифметического n одинаково распределенных взаимно независимых случайных величин в n раз меньше дисперсии D каждой из величин:

$$D(\bar{X}) = \frac{D}{n}.$$

Доказательство.

$$D(\bar{X}) = D\left(\frac{X_1 + X_2 + \dots + X_n}{n}\right) = \frac{D(X_1) + D(X_2) + \dots + D(X_n)}{n^2} = \frac{nD}{n^2} = \frac{D}{n}.$$

3. Среднее квадратическое отклонение среднего арифметического n одинаково распределенных взаимно независимых случайных величин в $\sqrt{n}$ раз меньше среднего квадратического отклонения каждой из величин:

$$\sigma(\bar{X}) = \frac{\sigma}{\sqrt{n}}.$$

Доказательство.

$$\sigma(\bar{X}) = \sqrt{D(\bar{X})} = \sqrt{\frac{D}{n}} = \frac{\sqrt{D}}{\sqrt{n}} = \frac{\sigma}{\sqrt{n}}.$$

Вернемся к задаче нахождения доверительных интервалов для оценки математического ожидания нормального распределения при известном σ (покрывающие параметр a с надежностью γ).

Пусть выполняется следующее соотношение:

$$P(|X - a| < \delta) = \gamma.$$

Из теории случайных величин следует, что вероятность того, что отклонение случайной величины X , распределенной по нормальному закону, от математического ожидания a не превысит величину $\delta > 0$ (по абсолютной величине), равна:

$$P(|X - a| < \delta) = 2\Phi\left(\frac{\delta}{\sigma}\right),$$

где $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_0^x e^{-\frac{t^2}{2}} dt$ – функция Лапласа.

Заменим X на $\bar{X}$ и σ на $\sigma(\bar{X})$, получим:

$$P(|\bar{X} - a| < \delta) = 2\Phi\left(\frac{\delta\sqrt{n}}{\sigma}\right) = 2\Phi(t), \text{ где } t = \frac{\delta\sqrt{n}}{\sigma}.$$

Из последнего выражения выразим δ , тогда получим, что $\delta = \frac{t\sigma}{\sqrt{n}}$.

Подставим полученное выражение в последнее соотношение, тогда:

$$P\left(|\bar{X} - a| < \frac{t\sigma}{\sqrt{n}}\right) = 2\Phi(t).$$

Учитывая, что вероятность P задана и равна γ окончательно получим (выборочную среднюю вновь обозначим через $\bar{x}_e$):

$$P\left(\bar{x}_e - \frac{t\sigma}{\sqrt{n}} < a < \bar{x}_e + \frac{t\sigma}{\sqrt{n}}\right) = 2\Phi(t) = \gamma.$$

С надежностью γ можно утверждать, что доверительный интервал $\left(\bar{x}_e - \frac{t\sigma}{\sqrt{n}}; \bar{x}_e + \frac{t\sigma}{\sqrt{n}}\right)$ покрывает неизвестный параметр a . Точность оценки

$$\delta = \frac{t\sigma}{\sqrt{n}}.$$

Число t определяется из равенства $2\Phi(t) = \gamma$, или $\Phi(t) = \frac{\gamma}{2}$. По таблице функции Лапласа находим аргумент t , которому соответствует значение функции Лапласа, равное $\frac{\gamma}{2}$.

Замечание. Оценку $|\bar{X} - a| < \frac{t\sigma}{\sqrt{n}}$ называют *классической*. Из формулы

$\delta = \frac{t\sigma}{\sqrt{n}}$, определяющей точность классической оценки, можно сделать сле-

дующие выводы:

1. При *возрастании* объема выборки n число δ убывает и, следовательно, точность оценки *увеличивается*.

2. Увеличение надежности оценки $\Phi(t) = \frac{\gamma}{2}$ приводит к увеличению t ($\Phi(t)$ – возрастающая функция), следовательно, и к возрастанию δ .

Другими словами, увеличение надежности классической оценки влечет за собой уменьшение ее точности.

Пример. Случайная величина X имеет нормальное распределение с известным средним квадратическим отклонением $\sigma = 3$. Найти доверительные интервалы для оценки неизвестного математического ожидания a по выборочным средним $\bar{x}_g$, если объем выборки $n = 36$ и задана надежность оценки $\gamma = 0,95$.

Решение. Найдем t . Из соотношения $2\Phi(t) = 0,95$ получим $\Phi(t) = 0,475$. По таблице функции Лапласа (Приложение 1) находим значение $t = 1,96$.

Найдем точность оценки: $\delta = \frac{t\sigma}{\sqrt{n}} = \frac{1,96 \cdot 3}{\sqrt{36}} = 0,98$.

Тогда доверительный интервал примет вид: $(\bar{x}_g - 0,98; \bar{x}_g + 0,98)$.

Например, если $\bar{x}_g = 3$, то доверительный интервал имеет следующие границы: $(3 - 0,98; 3 + 0,98)$ или $(2,02; 3,98)$.

Таким образом, значение неизвестного параметра a , согласующиеся с данными выборки, удовлетворяют неравенству: $2,02 < a < 3,98$.

Ошибочным является утверждение, что $P(2,02 < a < 3,98) = 0,95$.

Действительно, так как a – постоянная величина, то либо она заключена в найденном интервале (тогда событие $2,02 < a < 3,98$ достоверно и его ве-

роятность равна единице), либо в нем не заключена (в этом случае событие $2,02 < a < 3,98$ невозможно и его вероятность равна нулю).

Другими словами, доверительную вероятность не следует связывать с оцениваемым параметром; она связана лишь с границами доверительного интервала, которые изменяются от выборки к выборке.

Надежность $\gamma = 0,95$ указывает на то, что если произведено достаточно большое число выборок, то 95% из них определяет такие доверительные интервалы, в которых параметр действительно заключен, лишь в 5% случаев он может выйти за границы доверительного интервала.

Пример. На заводе производят детали. Известно, что стандартное отклонение длины детали составляет $\sigma = 2$ мм. Из партии случайным образом отобрано $n = 100$ деталей. Среднее значение выборки составило $\bar{x}_g = 50$ мм. Построить 95% доверительный интервал для оценки истинной средней длины детали.

Решение. Определяем параметры задачи: $\bar{x}_g = 50$, $\sigma = 2$, $n = 100$, $\gamma = 0,95$.

Найдем критическое значение t . Из соотношения $2\Phi(t) = 0,95$ получим $\Phi(t) = 0,475$. По таблице функции Лапласа (Приложение 1) находим значение $t = 1,96$.

Найдем точность оценки:
$$\delta = \frac{t\sigma}{\sqrt{n}} = \frac{1,96 \cdot 2}{\sqrt{100}} = 0,392.$$

Тогда доверительный интервал примет вид: $(\bar{x}_g - 0,392; \bar{x}_g + 0,392)$, $(49,608; 50,392)$.

Замечание. Если требуется оценить математическое ожидание с наперед заданной точностью δ и надежностью γ , то минимальный объем выборки n , который обеспечит эту точность, находят по формуле:

$$n = \frac{t^2 \sigma^2}{\delta^2} \quad (\text{так как } \delta = \frac{t\sigma}{\sqrt{n}}).$$

Пример. Исследователь хочет оценить средний вес продукта с точностью $\pm 0,5$ кг. при 95% доверительной вероятности. Известно, что $\sigma = 3$ кг. Какой объем выборки необходим?

Решение. Определяем параметры задачи: $\sigma = 3$, $\gamma = 0,95$, $t = 1,96$, $\delta = 0,5$.

Подставим данные в формулу для нахождения объема выборки:

$$n = \frac{t^2 \sigma^2}{\delta^2} = \left(\frac{1,96 \cdot 3}{0,5} \right)^2 = 138,298.$$

Таким образом, необходимо $n=139$ наблюдений (округляем в большую сторону).

Некоторые замечательные статистические распределения

Наряду с нормальным распределением, которое занимает центральное место в теории и практике вероятностно-статистических исследований, в математической статистике необходимы и некоторые другие важные распределения, которые используются для построения разного рода статистических оценок и в других целях.

К таким распределениям в первую очередь относятся распределения «хи- квадрат» (χ^2), Стьюдента и Фишера.

Распределение «хи-квадрат» (Пирсона)

Случайная величина используется при интервальном оценивании параметров распределений, при статистической проверке гипотез и др.

Пусть $X_1, X_2, \dots, X_n$ – последовательность независимых случайных величин, причем математическое ожидание каждой из них равно нулю, а среднее квадратическое отклонение – единице ($X_i \in N(0;1)$, $i = 1, 2, \dots, n$). Тогда сумма квадратов этих величин $\chi^2 = X_1^2 + X_2^2 + \dots + X_n^2$ распределена по закону χ^2 с $k = n$ степенями свободы (указывает число слагаемых).

Если же эти величины связаны одним линейным соотношением, например $\sum X_i = n\bar{X}$, то число степеней свободы $k = n - 1$.

Плотность распределения «хи квадрат»:
$$f(x) = \begin{cases} 0, & x \leq 0 \\ \frac{e^{-x/2} x^{(k/2)-1}}{2^{k/2} \Gamma\left(\frac{k}{2}\right)}, & x > 0 \end{cases}$$

$\Gamma(x) = \int_0^{\infty} t^{x-1} e^{-t} dt$ – гамма-функция; в частности $\Gamma(n+1) = n!$

Из последних соотношений видно, что «хи квадрат» определяется одним параметром – числом степеней свободы k . С увеличением числа степеней свободы распределение медленно приближается к нормальному.

Под *числом степеней свободы* в статистике понимают разность между числом опытов и количеством коэффициентов модели, вычисленных по результатам этих экспериментов независимо друг от друга.

Гамма-функция – математическая функция, которая была введена Л. Эйлером, а своим обозначением обязана Лежандру. Среди основных областей применения гамма-функции – математический анализ, теория вероятностей, комбинаторика, статистика, атомная физика, астрофизика, гидродинамика, сейсмология.

В частности, гамма-функция распространяет понятие факториала на дробные, отрицательные и комплексные значения аргумента.

В статистике широко используется гамма-распределение, частными случаями которого являются экспоненциальное распределение и распределение «хи-квадрат».

Распределение Стьюдента

Как и распределение χ^2 , распределение Стьюдента играет в основном роль вспомогательного вычислительного средства и используется при построении доверительного интервала для математического ожидания, проверке гипотезы о равенстве математических ожиданий двух нормальных случайных величин и т.д.

Оно впервые описано в 1908 г. английским статистиком В. Госсетом, который опубликовал свою научную работу под псевдонимом Стьюдент.

Пусть Z – нормальная случайная величина, причем $M(Z) = 0$, $\sigma(Z) = 1$, а V – независимая от Z величина, которая распределена по закону с k степенями свободы. Тогда величина

$$T = \frac{Z}{\sqrt{\frac{V}{k}}}$$

имеет распределение, которое называют t -распределением или распределением Стьюдента с k степенями свободы.

Итак, отношение нормированной нормальной величины к квадратному корню из независимой случайной величины, распределенной по закону «хи квадрат» с k степенями свободы, деленной на k , распределено по закону Стьюдента с k степенями свободы.

С возрастанием числа степеней свободы распределение Стьюдента быстро *приближается к нормальному*.

Распределение Фишера

Пусть X и Y – независимые случайные величины, распределенные по закону χ^2 со степенями свободы k_1 и k_2 . Тогда величина

$$F = \frac{X/k_1}{Y/k_2}$$

имеет распределение Фишера со степенями свободы k_1 и k_2 .

Таким образом, распределение Фишера определяется двумя параметрами – числами степеней свободы.

При больших значениях k_1 и k_2 это распределение приближается к нормальному.

Плотность распределение Фишера:

$$f(x) = \begin{cases} 0, & x \leq 0 \\ C_0 \frac{x^{(k_1-2)/2}}{(k_2 + k_1 x)^{(k_1+k_2)/2}}, & x > 0 \end{cases}, \quad C_0 = \frac{\Gamma\left(\frac{k_1+k_2}{2}\right) k_1^{k_1/2} k_2^{k_2/2}}{\Gamma(k_1/2) \Gamma(k_2/2)},$$

$$T_n^2 = F_{1,n}$$

Замечание. T_n – случайная величина, имеющая распределение Стьюдента с числом степеней свободы n , $F_{1,n}$ – случайная величина, имеющая распределение Фишера с числами степеней свободы 1 и n .

Распределение Фишера используется при проверке статистических гипотез и для него составлена таблица критических точек, как правило, при уровнях значимости $\alpha = 0,1$, $\alpha = 0,05$, $\alpha = 0,01$.

Доверительные интервалы для оценки математического ожидания при неизвестной дисперсии

Пусть количественный признак X генеральной совокупности распределен нормально, причем среднее квадратическое отклонение σ этого распределения неизвестно. Требуется оценить неизвестное математическое ожидание a с помощью доверительных интервалов.

Найдем такое число δ , чтобы выполнялось соотношение:
 $P(|X - a| < \delta) = \gamma$.

Введем случайную величину, имеющая распределение Стьюдента с $k = n - 1$ степенями свободы:

$$T = \frac{(\bar{X} - a)}{\frac{S}{\sqrt{n}}}$$

где S – исправленное среднее квадратическое отклонение; n – объем выборки; $\bar{X}$ – выборочная средняя.

$$S = \sqrt{\frac{n}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2}.$$

Перейдем от случайной величины $\bar{X}$ к случайной величине T , получим:

$$P\left(\left|\frac{\bar{X} - a}{\frac{S}{\sqrt{n}}}\right| < \frac{\delta}{\frac{S}{\sqrt{n}}}\right) = \gamma \Leftrightarrow P\left(|T| < \frac{\delta \sqrt{n}}{S}\right) = \gamma \Leftrightarrow P(|T| < t_\gamma) = \gamma,$$

$$t_{\gamma} = \frac{\delta \sqrt{n}}{S} \Rightarrow \delta = t_{\gamma} \frac{S}{\sqrt{n}}.$$

Следовательно, доверительный интервал принимает вид:

$$P\left(\bar{X} - t_{\gamma} \frac{S}{\sqrt{n}} < a < \bar{X} + t_{\gamma} \frac{S}{\sqrt{n}}\right) = \gamma$$

Значит, интервал $\left(\bar{X} - t_{\gamma} \frac{S}{\sqrt{n}}; \bar{X} + t_{\gamma} \frac{S}{\sqrt{n}}\right)$ покрывает неизвестный параметр a с надежностью γ , т.е. является доверительным интервалом для неизвестного математического ожидания случайной величины X .

Значение t_{γ} находится по таблице распределения Стьюдента (Приложение 2) в зависимости от доверительной вероятности γ и числа степеней свободы $(n - 1)$.

Пример. Количественный признак X генеральной совокупности распределен нормально. При выборке объема $n = 16$ найдены выборочная средняя $\bar{x}_g = 20,2$ и исправленное среднее квадратическое отклонение $s = 0,8$. Оценить неизвестное математического ожидания при помощи доверительного интервала с надежностью $\gamma = 0,95$.

Решение. По таблице (Приложение 2) найдем t_{γ} для $\gamma = 0,95$ и $n = 16$. Получим, $t_{\gamma} = 2,13$.

Найдем доверительные границы интервала $\left(\bar{X} - t_{\gamma} \frac{S}{\sqrt{n}}; \bar{X} + t_{\gamma} \frac{S}{\sqrt{n}}\right)$ с надежностью $\gamma = 0,95$. Получаем, что неизвестный параметр a заключен в доверительном интервале $(19,774; 20,626)$.

При неограниченном возрастании объема выборки n распределение Стьюдента стремится к нормальному. Поэтому практически при $n > 30$ можно вместо распределения Стьюдента пользоваться нормальным распределением.

Для малых выборок замена распределения нормальным приводит к грубым ошибкам, а именно к неоправданному сужению доверительного интервала, т.е. к повышению точности оценки.

Например, если $n = 5$ и $\gamma = 0,99$, то, пользуясь распределением Стюдента, найдем $t_\gamma = 4,6$, а используя функцию Лапласа, найдем $t_\gamma = 2,58$, т.е. доверительный интервал в последнем случае окажется более узким, чем найденный по распределению Стюдента.

Пример. По данным девяти независимых равноточных (одинаковые дисперсии) измерений физической величины найдены среднее арифметическое результатов отдельных измерений $\bar{x}_g = 42,319$ и исправленное среднее квадратическое отклонение $s = 5$. Оценить истинное значение измеряемой величины с надежностью $\gamma = 0,95$.

Решение. Истинное значение измеряемой величины равно ее математическому ожиданию. Поэтому задача сводится к оценке математического ожидания (при неизвестном σ) при помощи доверительного интервала:

$$\left(\bar{X} - t_\gamma \frac{S}{\sqrt{n}}; \bar{X} + t_\gamma \frac{S}{\sqrt{n}} \right).$$

Для $\gamma = 0,95$ и $n = 9$, $t_\gamma = 2,31$. Найдем доверительные интервалы:

$$\left(\bar{X} - t_\gamma \frac{S}{\sqrt{n}}; \bar{X} + t_\gamma \frac{S}{\sqrt{n}} \right), (38,469; 46,169).$$

Таким образом, с надежностью $\gamma = 0,95$ истинное значение измеряемой величины заключено в доверительном интервале: $38,469 < a < 46,169$.

Пример. Измерили время реакции (в мс.) у 16 курсантов: $\bar{x}_g = 245$ мс., $s = 32$ мс. с доверительная вероятность равной 95%.

Решение. По таблице (Приложение 2) найдем t_γ для $\gamma = 0,95$ и $n = 16$. Получим, $t_\gamma = 2,13$.

Найдем доверительные границы интервала $\left(\bar{X} - t_\gamma \frac{S}{\sqrt{n}}; \bar{X} + t_\gamma \frac{S}{\sqrt{n}} \right)$ с надежностью $\gamma = 0,95$. Получаем, что неизвестный параметр a заключен в

доверительном интервале $(227,96; 262,04)$. Следовательно, с доверительной вероятностью 95% истинное среднее время реакции генеральной совокупности курсантов находится в интервале от 227,96 до 262,04 мс.

Пример. Выборка результатов стрельбы (очки) 10 сотрудников: 85, 92, 78, 88, 95, 81, 89, 93, 87, 90. Построить 99% доверительный интервал для среднего результата.

Решение. По условию задачи имеем: $n = 10$, $\gamma = 0,99$. Вычислим числовые характеристики: $\bar{x}_e = 87,8$, $s = 5,036$. Т.к. $\gamma = 0,95$, то $t_\gamma = 3,25$. Доверительный интервал имеет вид: $(82,62; 92,98)$.

Доверительные интервалы для оценки среднего квадратического отклонения нормального распределения

Пусть количественный признак X генеральной совокупности распределен нормально. Требуется оценить неизвестное генеральное среднее квадратическое отклонение σ по исправленному выборочному среднему квадратическому отклонению s . Необходимо найти доверительные интервалы, покрывающие параметр σ с заданной надежностью γ .

По определению доверительного интервала:

$$P(|\sigma - s| < \delta) = \gamma \Leftrightarrow P(s - \delta < \sigma < s + \delta) = \gamma.$$

Для того, чтобы можно было пользоваться готовой таблицей, преобразуем двойное неравенство:

$$s\left(1 - \frac{\delta}{s}\right) < \sigma < s\left(1 + \frac{\delta}{s}\right).$$

Введем обозначение: $q = \frac{\delta}{s}$. Следовательно, $s(1 - q) < \sigma < s(1 + q)$

Замечание. Типичная для подобных задач ситуация, когда $\delta < s$, т.е. $q < 1$. Считаем, что последнее выполняется, тогда:

$$\frac{1}{s(1 + q)} < \frac{1}{\sigma} < \frac{1}{s(1 - q)}$$

Умножим все части неравенства на $s\sqrt{n-1} > 0$, получим

$$\frac{\sqrt{n-1}}{(1+q)} < \frac{s\sqrt{n-1}}{\sigma} < \frac{\sqrt{n-1}}{(1-q)}.$$

Введем случайную величину, которая распределена по закону χ^2 с $(n-1)$ степенями свободы. Это распределение не зависит от оцениваемого параметра σ , зависит лишь от объема выборки n :

$$\chi^2 = \frac{s^2(n-1)}{\sigma^2} \Rightarrow \chi = \frac{s\sqrt{n-1}}{\sigma}.$$

Получим окончательные формулы для доверительного интервала

$$\frac{s\sqrt{n-1}}{\chi_2} < \sigma < \frac{s\sqrt{n-1}}{\chi_1}.$$

$$\chi_1^2 = \chi^2\left(\frac{1+\gamma}{2}; n-1\right), \quad \chi_2^2 = \chi^2\left(\frac{1-\gamma}{2}; n-1\right).$$

Определяются по таблице «хи квадрат» распределения.

При построении доверительного интервала для дисперсии достаточно возвести все члены последнего неравенства в квадрат.

Пример. Пусть по выборке объема 20 найдено $s = 0,03$. Требуется с надежностью $\gamma = 0,9$ построить доверительный интервал для σ .

Решение. По таблице для $\gamma = 0,9$ и $n = 20$ находим:

$$\chi_1^2 = \chi^2\left(\frac{1+0,9}{2}; 20-1\right) = \chi^2(0,95; 19) = 10,1;$$

$$\chi_2^2 = \chi^2\left(\frac{1-0,9}{2}; 20-1\right) = \chi^2(0,05; 19) = 30,1;$$

Доверительный интервал:

$$\frac{s\sqrt{n-1}}{\chi_2} < \sigma < \frac{s\sqrt{n-1}}{\chi_1},$$

$$\frac{0,03\sqrt{20-1}}{\sqrt{30,1}} < \sigma < \frac{0,03\sqrt{20-1}}{\sqrt{10,1}},$$

$$0,024 < \sigma < 0,041.$$

Пример. На заводе измеряли диаметр деталей (мм). Выборка из 16 деталей дала результаты: выборочное среднее квадратическое отклонение $s=0,8$ мм. Требуется найти 95% доверительный интервал для σ .

Решение. По условию задачи имеем: $n=16$, $s=0,8$, $\gamma=0,95$, значит, $\alpha=0,05$. Число степеней свободы: $k=n-1=15$. По таблице «хи квадрат» распределения для $\gamma=0,95$ и $n=16$ находим:

$$\chi_1^2 = \chi^2\left(\frac{1+0,95}{2}; 16-1\right) = \chi^2(0,975; 15) = 6,26;$$

$$\chi_2^2 = \chi^2\left(\frac{1-0,95}{2}; 16-1\right) = \chi^2(0,025; 15) = 27,49.$$

Доверительный интервал:

$$\frac{s\sqrt{n-1}}{\chi_2} < \sigma < \frac{s\sqrt{n-1}}{\chi_1},$$

$$\frac{0,8\sqrt{16}}{\sqrt{27,49}} < \sigma < \frac{0,8\sqrt{16}}{\sqrt{6,26}},$$

$$0,61 < \sigma < 1,28.$$

Оценка точности измерений

В теории ошибок принято точность измерений (точность прибора) характеризовать с помощью среднего квадратического отклонения σ случайных ошибок измерений. Для оценки σ используют «исправленное» среднее квадратическое отклонение s .

Поскольку обычно результаты измерений взаимно независимы, имеют одно и то же математическое ожидание (истинное значение измеряемой величины) и одинаковую дисперсию (в случае равноточных измерений), то выше изложенная теория применима для оценки точности измерений.

Пример. По 15 равноточным измерениям найдено «исправленное» среднее квадратическое отклонение $s = 0,12$. Найти точность измерений с надежностью $\gamma = 0,99$.

Решение. Точность измерений характеризуется средним квадратическим отклонением σ случайных ошибок, поэтому задача сводится к отыска-

нию доверительного интервала, покрывающего σ с заданной надежностью $\gamma = 0,99$.

$$\chi_1^2 = \chi^2\left(\frac{1+0,99}{2}; 15-1\right) = \chi^2(0,995; 14) = 4,1;$$

$$\chi_2^2 = \chi^2\left(\frac{1-0,99}{2}; 14\right) = \chi^2(0,005; 14) = 31,32;$$

Доверительный интервал:

$$\frac{s\sqrt{n-1}}{\chi_2} < \sigma < \frac{s\sqrt{n-1}}{\chi_1};$$

$$\frac{0,12\sqrt{15-1}}{\sqrt{31,32}} < \sigma < \frac{0,12\sqrt{15-1}}{\sqrt{4,1}};$$

$$0,08 < \sigma < 0,22.$$

2.3. Статистическая проверка гипотез

2.3.1. Статистическая гипотеза. Ошибки первого и второго рода

Распространенной практической задачей, связанной с использованием статистических методов, является принятие решения на основе имеющейся выборки: подтвердить или отклонить гипотезу относительно всей генеральной совокупности (или случайной величины).

Процедура сопоставления высказанного предположения (гипотезы) с выборочными данными называется *проверкой гипотез*.

Задача статистической проверки гипотез заключается в следующем: относительно некоторой генеральной совокупности высказывается та или иная гипотеза H . Из этой генеральной совокупности извлекается выборка. Требуется указать правило, при помощи которого можно было бы по выборке решить вопрос о том, следует ли отклонить гипотезу H или принять ее.

Статистической называют гипотезу о виде неизвестного распределения или о параметрах известных распределений.

Гипотеза называется *параметрической*, если в ней содержится некоторое утверждение о значении параметра известного распределения. Гипотезу,

в которой сформулированы предположения относительно вида распределения, называют *непараметрической*.

Например, статистическими являются гипотезы:

- 1) генеральная совокупность распределена по закону Пуассона – сделано предположение о виде неизвестного распределения;
- 2) дисперсии двух нормальных совокупностей равны между собой – сделано предположение о параметрах двух известных распределений.

Наряду с выдвинутой гипотезой рассматривают и противоречащую ей гипотезу. Если выдвинутая гипотеза будет отвергнута, то имеет место противоречащая гипотеза. По этой причине эти гипотезы целесообразно различать.

Нулевой (основной) называют выдвинутую гипотезу H_0 .

Конкурирующей (альтернативной) называют гипотезу H_1 , которая противоречит нулевой.

Например, если нулевая гипотеза состоит в предположении, что математическое ожидание a нормального распределения равно 10, то конкурирующая гипотеза, может состоять в предположении, что $a \neq 10$, т.е. если $H_0 : a = 10$, то $H_1 : a \neq 10$.

Различают гипотезы *простые* и *сложные*. *Простой* называют гипотезу, содержащую только одно предположение. *Сложной* называют гипотезу, которая состоит из конечного или бесконечного числа простых гипотез.

Например, гипотеза H_0 , состоящая в предположении, что математическое ожидание случайной величины X равно a_0 , т.е. $M(X) = a_0$, является простой. В качестве альтернативной гипотезы можно рассматривать одну из следующих гипотез: $H_1 : M(X) > a_0$ (сложная), $H_1 : M(X) < a_0$ (сложная), $H_1 : M(X) \neq a_0$ (сложная) или $H_1 : M(X) \neq a_1$ (простая).

Выдвинутая гипотеза может быть правильной или неправильной, поэтому возникает необходимость ее проверки. Правило, по которому решают о принятии или отклонении гипотезы H_0 (соответственно отклонении или принятии H_1), называют *статистическим критерием*.

В зависимости от вида распределения критерий обозначают через U (если распределен по нормальному закону), T (если распределен по закону Стьюдента), F (если распределен по закону Фишера-Снедекора) и т.д.

В целях общности обозначим статистический критерий через K . Наблюдаемым значением критерия называют значение критерия, вычисленное по выборкам, т.е. $K_{набл.} = K(x_1, x_2, \dots, x_n)$.

После выбора критерия множество его значений разделяются на два подмножества: область S отклонения нулевой гипотезы (*критическая область*) и область $\bar{S}$ принятия этой гипотезы (*область принятия гипотезы*).

Если значение x критерия попало в область S , то нулевая гипотеза H_0 отвергается и принимается альтернативная гипотеза H_1 . Если точка x попала в область $\bar{S}$, то принимается гипотеза H_0 , а H_1 – отвергается.

При проверке гипотезы может быть принято неправильное решение, т.е. могут быть допущены ошибки двух родов.

Ошибка первого рода состоит в том, что отвергается нулевая гипотеза H_0 , когда на самом деле она верна. Таким образом, будет принята альтернативная гипотеза H_1 .

Вероятность ошибки первого рода обозначают через α : $\alpha = P\{H_1 / H_0\}$, где $P\{H_1 / H_0\}$ – вероятность того, что будет принята гипотеза H_1 , если на самом деле верна гипотеза H_0 .

Чем меньше α , тем ниже вероятность отклонить верную гипотезу. Допустимую ошибку первого рода обычно задают заранее.

Число α называют *уровнем значимости*. Обычно для α используют стандартные значения: 0,1; 0,05; 0,01 и др.

Ошибка второго рода состоит в том, что отвергается альтернативная гипотеза H_1 , когда на самом деле она верна. Таким образом, будет принята основная гипотеза H_0 .

Вероятность ошибки второго рода обозначают через β : $\beta = P\{H_0 / H_1\}$, где $P\{H_0 / H_1\}$ – вероятность того, что будет принята гипотеза H_0 , если на самом деле верна гипотеза H_1 .

Гипотеза H_0	Отвергается	Принимается
верна	Ошибка 1 рода	Правильное решение
неверна	Правильное решение	Ошибка 2 рода

Правильные решения также подразделяют на два вида. Если будет принята гипотеза H_0 , тогда как и на самом деле в генеральной совокупности она верна, то вероятность такого решения будет равна:

$$1 - \alpha = P\{H_0 / H_0\}.$$

Может быть принята гипотеза H_1 , тогда как и на самом деле она верна. Вероятность решения будет равна:

$$1 - \beta = P\{H_1 / H_1\}.$$

Число $1 - \beta$ называют *мощностью критерия*. Чем больше мощность критерия, тем вероятность ошибки второго рода ниже.

Вероятности ошибок 1-го и 2-го рода определяются выбором критической области. При фиксированном объеме выборки уменьшение ошибки 1-го рода α неизбежно приводит к увеличению ошибки 2-го рода β и наоборот.

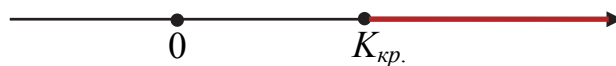
Лишь при увеличении объема выборки возможно одновременное уменьшение вероятностей α и β .

Поскольку критерий K – одномерная случайная величина, то все ее возможные значения принадлежат некоторому интервалу, поэтому критическую область и область принятия гипотезы можно *разделить точками*.

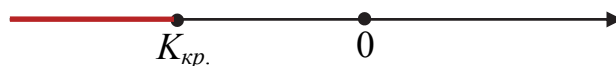
Критическими точками $K_{кр.}$ называют точки, отделяющие критическую область от области принятия гипотезы.

Различают *одностороннюю (правостороннюю и левостороннюю)* и *двустороннюю* критические области.

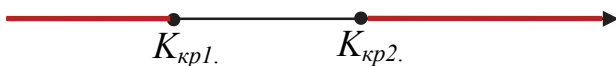
Правосторонней называют критическую область, определяемую неравенством $K > K_{кр.}$, при этом $P(K > K_{кр.}) = \alpha$.



Левосторонней называют критическую область, определяемую неравенством $K < K_{кр.}$, при этом $P(K < K_{кр.}) = \alpha$.



Двусторонней называют критическую область, определяемую неравенствами $K < K_{кр1.}$ и $K > K_{кр2.}$, при этом $P(K > K_{кр2.}) = P(K < K_{кр1.}) = \frac{\alpha}{2}$.



Рассмотрим способы проверки некоторых наиболее часто встречающихся гипотез.

2.3.2. Проверка гипотезы о равенстве генеральной средней нормальной совокупности заданному числовому значению

Пусть генеральная совокупность X распределена нормально, причем имеются основания предполагать, что генеральная средняя этой совокупности $\bar{x}$ равна некоторому значению a .

Предположим, что *дисперсия генеральной совокупности* $D = \sigma^2$ известна (например, может быть найдена теоретически или вычислена по выборке большого объема).

Кроме того, по произведенной выборке объема n найдена выборочная средняя $\bar{x}$. Требуется по выборочной средней при заданном уровне значимости α проверить нулевую гипотезу $H_0 : \bar{x} = a$.

В качестве критерия проверки нулевой гипотезы принимают случайную величину:

$$U = \frac{(\bar{x} - a)\sqrt{n}}{\sigma},$$

которая имеет нормальное распределение.

Для того, чтобы при заданном уровне значимости α проверить нулевую гипотезу $H_0 : \bar{x} = a$ нужно вычислить наблюдаемое значение критерия:

$$U_{набл} = \frac{(\bar{x}_e - a)\sqrt{n}}{\sigma}.$$

1. При конкурирующей гипотезе $H_1 : \bar{x} \neq a$ критическую точку $U_{кр.}$ находим по таблице функции Лапласа из условия $\Phi(U_{кр.}) = \frac{1-\alpha}{2}$.

Если $|U_{набл.}| < U_{кр.}$, то нет оснований отвергнуть нулевую гипотезу H_0 . В противном случае нулевую гипотезу отвергают.

2. При конкурирующей гипотезе $H_1 : \bar{x} > a$ критическую точку $U_{кр.}$ правосторонней критической области находим по таблице функции Лапласа из условия $\Phi(U_{кр.}) = \frac{1-2\alpha}{2}$.

Если $U_{набл.} < U_{кр.}$, то нет оснований отвергнуть нулевую гипотезу H_0 . В противном случае нулевую гипотезу отвергают.

3. При конкурирующей гипотезе $H_1 : \bar{x} < a$ критическую точку $U_{кр.}$ левосторонней критической области находим по таблице функции Лапласа из условия $\Phi(U_{кр.}) = \frac{1-2\alpha}{2}$.

Если $U_{набл.} > -U_{кр.}$, то нет оснований отвергнуть нулевую гипотезу H_0 . В противном случае нулевую гипотезу отвергают.

Пример. Из нормальной генеральной совокупности с известной дисперсией $\sigma^2 = 25$ извлечена выборка объема $n = 36$ и по ней найдено $\bar{x} = 17$. При уровне значимости $\alpha = 0,05$ проверить нулевую гипотезу $H_0 : a = 20$ при конкурирующей гипотезе а) $H_1 : a \neq 20$, б) $H_1 : a < 20$.

Решение. а) Вычисляем значение статистики:

$$|U_{набл}| = \frac{|\bar{x}_e - a|\sqrt{n}}{\sigma} = \frac{|17 - 20| \cdot 6}{5} = 3,6.$$

Так как критерий двусторонний, то $U_{кр.}$ находим из соотношения

$$\Phi(U_{кр.}) = \frac{1-\alpha}{2} = 0,475. \text{ По таблице Лапласа находим } U_{кр.} = 1,96.$$

Так как $|U_{набл.}| > U_{кр.}$, то гипотеза $H_0 : a = 20$ – отклоняется.

б) Вычисляем значение статистики:

$$U_{набл.} = \frac{(\bar{x}_e - a)\sqrt{n}}{\sigma} = \frac{(17 - 20) \cdot 6}{5} = -3,6.$$

Так как критерий левосторонний, то $U_{кр.}$ находим из соотношения

$$\Phi(U_{кр.}) = \frac{1-2\alpha}{2} = 0,45. \text{ По таблице Лапласа находим } U_{кр.} = 1,64.$$

Так как $U_{набл.} < -U_{кр.}$ ($-3,6 < -1,64$), то гипотеза $H_0 : a = 20$ отклоняется в пользу гипотезы $H_1 : a < 20$.

Пример. Автоматическая линия разливает сок в пакеты по 1 литру. Известно, что стандартное отклонение объема разливки $\sigma = 15$ мл. Для проверки качества взяли выборку из $n = 36$ пакетов и получили среднее значение $\bar{x} = 993$ мл. При уровне значимости $\alpha = 0,05$ проверить, правильно ли настроена линия.

Решение.

Формулируем гипотезы:

$H_0 : a = 1000$ мл (линия настроена правильно),

$H_1 : a \neq 1000$ мл (линия настроена неправильно).

По условию задачи дисперсия σ известна. Уровень значимости $\alpha = 0,05$.

Вычисляем наблюдаемое значение критерия (статистики):

$$|U_{набл.}| = \frac{|\bar{x}_e - a|\sqrt{n}}{\sigma} = \frac{|993 - 1000| \cdot \sqrt{36}}{15} = 2,8.$$

Так как критерий двусторонний, то $U_{кр.}$ находим из соотношения:

$$\Phi(U_{кр.}) = (1-\alpha)/2 = 0,475. \text{ По таблице Лапласа находим } U_{кр.} = 1,96.$$

Так как $|U_{набл.}| > U_{кр.}$, то гипотеза $H_0 : a = 1000$ отклоняется (значение попало в критическую область).

Таким образом, на уровне значимости 0,05 отвергаем гипотезу H_0 . Есть статистически значимые основания полагать, что автоматическая линия разливает сок не по 1 литру. Средний объем отличается от заявленного.

Предположим теперь, что дисперсия генеральной совокупности $D = \sigma^2$ неизвестна, а известна только ее исправленная выборочная оценка s^2 .

При выборке небольшого объема (менее 30 наблюдений) нахождение критической точки по таблице функции Лапласа может привести к существенной погрешности.

В таком случае в качестве критерия проверки нулевой гипотезы принимают случайную величину:

$$T = \frac{(\bar{x}_e - a)\sqrt{n}}{s},$$

которая имеет распределение Стьюдента с $(n-1)$ степенями свободы.

Для того, чтобы при заданном уровне значимости α проверить нулевую гипотезу $H_0: \bar{x} = a$ нужно вычислить наблюдаемое значение критерия:

$$T_{\text{набл}} = \frac{(\bar{x}_e - a)\sqrt{n}}{s}.$$

1. При конкурирующей гипотезе $H_1: \bar{x} \neq a$ критическую точку $T_{\text{кр.}}(\alpha, n-1)$ находим по таблице критических точек распределения Стьюдента при $(n-1)$ степенях свободы и вероятности α .

Если $|T_{\text{набл.}}| < T_{\text{кр.}}$, то нет оснований отвергнуть нулевую гипотезу. В противном случае нулевую гипотезу H_0 – отвергают.

2. При конкурирующей гипотезе $H_1: \bar{x} > a$ критическую точку $T_{\text{кр.}}(2\alpha, n-1)$ правосторонней критической области находим по таблице критических точек распределения Стьюдента при $(n-1)$ степенях свободы и вероятности 2α .

Если $T_{\text{набл.}} < T_{\text{кр.}}$, то нет оснований отвергнуть нулевую гипотезу. В противном случае нулевую гипотезу H_0 – отвергают.

3. При конкурирующей гипотезе $H_1: \bar{x} < a$ критическую точку $T_{кр.}(2\alpha, n-1)$ левосторонней критической области находим по таблице критических точек распределения Стьюдента при $(n-1)$ степенях свободы и вероятности 2α .

Если $T_{набл.} > -T_{кр.}$, то нет оснований отвергнуть нулевую гипотезу. В противном случае нулевую гипотезу H_0 – отвергают.

2.3.3. Проверка гипотезы о равенстве генеральной дисперсии нормальной совокупности заданному числовому значению

Пусть генеральная совокупность X распределена нормально, причем имеются основания предполагать, что генеральная дисперсия этой совокупности равна некоторому заданному значению σ_0^2 .

Пусть по выборке объема n получена исправленная выборочная дисперсия s^2 . Требуется по исправленной выборочной дисперсии при заданном уровне значимости α проверить нулевую гипотезу $H_0: \sigma^2 = \sigma_0^2$.

Для проверки гипотезы H_0 используют случайную величину $\frac{(n-1)s^2}{\sigma^2}$, которая имеет распределение χ^2 с $(n-1)$ степенями свободы.

Для того, чтобы при заданном уровне значимости α проверить нулевую гипотезу $H_0: \sigma^2 = \sigma_0^2$, нужно вычислить наблюдаемое значение критерия:

$$\chi_{набл}^2 = \frac{(n-1)s^2}{\sigma^2}.$$

1. При конкурирующей гипотезе $H_1: \sigma^2 \neq \sigma_0^2$ критическая область является двусторонней. Критические точки находим по таблице критических точек распределения χ^2 с $(n-1)$ степенями свободы: левую критическую

точку при вероятности $1 - \frac{\alpha}{2}$, а правую критическую точку при вероятности $\frac{\alpha}{2}$.

Таким образом, получим точки $\chi_{лев.кр}^2 \left(1 - \frac{\alpha}{2}, n - 1\right)$, $\chi_{пр.кр}^2 \left(\frac{\alpha}{2}, n - 1\right)$.

Если $\chi_{лев.кр}^2 < \chi_{набл}^2 < \chi_{пр.кр}^2$, то нет оснований отвергнуть нулевую гипотезу. В противном случае нулевую гипотезу отвергают.

2. При конкурирующей гипотезе $H_1: \sigma^2 > \sigma_0^2$ критическую точку $\chi_{кр}^2(\alpha, n - 1)$ правосторонней критической области находим по таблице критических точек распределения χ^2 с $(n - 1)$ степенями свободы при вероятности α .

Если $\chi_{набл}^2 < \chi_{кр}^2$, то нет оснований отвергнуть нулевую гипотезу. В противном случае нулевую гипотезу отвергают.

3. При конкурирующей гипотезе $H_1: \sigma^2 < \sigma_0^2$ критическую точку $\chi_{кр}^2(1 - \alpha, n - 1)$ левосторонней критической области находим по таблице критических точек распределения χ^2 с $(n - 1)$ степенями свободы при вероятности $1 - \alpha$.

Если $\chi_{набл}^2 > \chi_{кр}^2$, то нет оснований отвергнуть нулевую гипотезу. В противном случае нулевую гипотезу отвергают.

Пример. Из нормально распределенной совокупности извлечена выборка объема $n = 21$, для которой вычислено значение $s^2 = 10,3$. Требуется при уровне значимости $\alpha = 0,05$ проверить нулевую гипотезу $H_0: \sigma^2 = 9$ при альтернативной гипотезе $H_1: \sigma^2 > 9$.

Решение. Вычисляем значение статистики:

$$\chi_{набл}^2 = \frac{(n - 1)s^2}{\sigma^2} = \frac{20 \cdot 10,3}{9} \approx 22,9.$$

Критерий правосторонний, поэтому находим критическую точку $\chi_{кр}^2(\alpha, n-1) = 31,4$.

Так как $\chi_{набл}^2 < \chi_{кр}^2$ ($22,9 < 31,4$), то нет оснований отвергнуть нулевую гипотезу.

2.3.4. Проверка гипотезы о равенстве двух средних нормальных генеральных совокупностей

Пусть генеральные совокупности X_1 и X_2 имеют нормальное распределение, причем генеральные средние этих совокупностей $\bar{x}_1$ и $\bar{x}_2$ неизвестны.

По произведенным выборкам объемов n_1 и n_2 найдены выборочные средние $\bar{x}_{г1}$ и $\bar{x}_{г2}$.

Предполагаем, что дисперсии обеих генеральных совокупностей известны и равны σ_1^2 и σ_2^2 соответственно. Требуется при заданном уровне значимости α проверить нулевую гипотезу $H_0 : \bar{x}_1 = \bar{x}_2$.

В качестве критерия проверки нулевой гипотезы принимают случайную величину:

$$U = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\sigma_1^2 / n_1 + \sigma_2^2 / n_2}},$$

которая имеет нормальное распределение.

Для того, чтобы при заданном уровне значимости α проверить нулевую гипотезу $H_0 : \bar{x}_1 = \bar{x}_2$, нужно вычислить наблюдаемое значение критерия:

$$U_{набл} = \frac{\bar{x}_{г1} - \bar{x}_{г2}}{\sqrt{\sigma_1^2 / n_1 + \sigma_2^2 / n_2}}.$$

1. При конкурирующей гипотезе $H_1 : \bar{x}_1 \neq \bar{x}_2$ критическую точку $U_{кр}$ находим по таблице функции Лапласа из условия $\Phi(U_{кр}) = \frac{1-\alpha}{2}$ Критическая область в этом случае является двусторонней.

Если $|U_{набл.}| < U_{кр.}$, то нет оснований отвергнуть нулевую гипотезу. В противном случае нулевую гипотезу отвергают.

2. При конкурирующей гипотезе $H_1: \bar{x}_1 > \bar{x}_2$ критическую точку $U_{кр.}$ правосторонней критической области находим по таблице функции Лапласа из условия $\Phi(U_{кр.}) = \frac{1-2\alpha}{2}$.

Если $U_{набл.} < U_{кр.}$, то нет оснований отвергнуть нулевую гипотезу. В противном случае нулевую гипотезу отвергают.

3. При конкурирующей гипотезе $H_1: \bar{x}_1 < \bar{x}_2$ критическую точку $U_{кр.}$ левосторонней критической области находим по таблице функции Лапласа из условия $\Phi(U_{кр.}) = \frac{1-2\alpha}{2}$.

Если $U_{набл.} > -U_{кр.}$, то нет оснований отвергнуть нулевую гипотезу. В противном случае нулевую гипотезу отвергают.

Приведенное выше правило можно также применять, если генеральные совокупности не имеют нормального распределения и их дисперсии неизвестны, а выборки имеют большой объем (не менее 30 каждая) и независимы. В этом случае наблюдаемое значение критерия:

$$U_{набл.} = \frac{\bar{x}_{г1} - \bar{x}_{г2}}{\sqrt{s_1^2 / n_1 + s_2^2 / n_2}}.$$

Предположим теперь, что дисперсии обеих генеральных совокупностей неизвестны, а известны только их исправленные выборочные оценки s_1^2 и s_2^2 , а выборки имеют небольшой объем (меньше 30). Предполагается, что дисперсии двух генеральных совокупностей одинаковы.

Если же нет оснований считать дисперсии одинаковыми, то, прежде чем сравнивать средние, следует проверить гипотезу о равенстве генеральных дисперсий, пользуясь критерием Фишера-Снедекора.

В этом случае в качестве критерия проверки нулевой гипотезы принимают случайную величину:

$$T = \frac{\bar{x}_1 - \bar{x}_2}{s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}},$$

где $s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$,

имеющую распределение Стьюдента с $n_1 + n_2 - 2$ степенями свободы.

Для того, чтобы при заданном уровне значимости α проверить нулевую гипотезу $H_0 : \bar{x}_1 = \bar{x}_2$, нужно вычислить наблюдаемое значение критерия:

$$T_{набл} = \frac{\bar{x}_{e1} - \bar{x}_{e2}}{s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}.$$

1. При конкурирующей гипотезе $H_1 : \bar{x}_1 \neq \bar{x}_2$ критическую точку $T_{кр.}(\alpha, n_1 + n_2 - 2)$ находим по таблице критических точек распределения Стьюдента при $n_1 + n_2 - 2$ степенях свободы и вероятности α .

Если $|T_{набл.}| < T_{кр.}$, то нет оснований отвергнуть нулевую гипотезу. В противном случае нулевую гипотезу отвергают.

2. При конкурирующей гипотезе $H_1 : \bar{x}_1 > \bar{x}_2$ критическую точку $T_{кр.}(2\alpha, n_1 + n_2 - 2)$ правосторонней критической области находим по таблице критических точек распределения Стьюдента при $n_1 + n_2 - 2$ степенях свободы и вероятности 2α .

Если $T_{набл.} < T_{кр.}$, то нет оснований отвергнуть нулевую гипотезу. В противном случае нулевую гипотезу отвергают.

3. При конкурирующей гипотезе $H_1 : \bar{x}_1 < \bar{x}_2$ критическую точку $T_{кр.}(2\alpha, n_1 + n_2 - 2)$ левосторонней критической области находим по таблице критических точек распределения Стьюдента при $n_1 + n_2 - 2$ степенях свободы и вероятности 2α .

Если $T_{набл.} > -T_{кр.}$, то нет оснований отвергнуть нулевую гипотезу. В противном случае нулевую гипотезу отвергают.

Пример. Для двух независимых выборок объемов $n_1 = 6$ и $n_2 = 8$ из нормально распределенных совокупностей известны значения выборочных средних и исправленных выборочных дисперсий: $\bar{x}_1 = 6,58$, $\bar{x}_2 = 8,56$, $s_1^2 = 1,32$, $s_2^2 = 1,1$.

Требуется при уровне значимости $\alpha = 0,05$ проверить нулевую гипотезу $H_0 : \bar{x}_1 = \bar{x}_2$ при альтернативной гипотезе $H_1 : \bar{x}_1 \neq \bar{x}_2$.

Решение. Вычисляем значение статистики:

$$T_{набл} = \frac{\bar{x}_{e1} - \bar{x}_{e2}}{s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}, s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} = 1,19, T_{набл} = \frac{6,58 - 8,56}{1,1 \sqrt{\frac{1}{6} + \frac{1}{8}}} = -3,36.$$

Значение $T_{кр.}$ находим по таблице критических точек распределения Стьюдента при $n_1 + n_2 - 2 = 12$ степенях свободы и вероятности $\alpha = 0,05$: $T_{кр.} = 2,179$.

Так как $|T_{набл.}| > T_{кр.}$, то гипотеза H_0 – отвергается.

2.3.5. Проверка гипотез о законе распределения

Во многих случаях закон распределения изучаемой случайной величины неизвестен, но есть основания предположить, что он имеет вполне определенный вид: нормальный, биномиальный или какой-либо другой.

Пусть необходимо проверить гипотезу H_0 о том, что случайная величина X подчиняется определенному закону распределения, заданному функцией распределения $F_0(x)$ т.е. $H_0 : F_x(x) = F_0(x)$. Под альтернативной гипотезой H_1 будем понимать в данном случае то, что просто не выполнена основная $H_1 : F_x(x) \neq F_0(x)$.

Для проверки гипотезы о распределении случайной величины X проведем выборку, которую оформим в виде статистического ряда:

x_i	x_1	x_2	$\dots$	x_m
n_i	n_1	n_2	$\dots$	n_m

Требуется сделать заключение: согласуются ли результаты наблюдений с высказанным предположением. Для этого используем *критерий согласия*.

Критерием согласия называется статистический критерий проверки гипотезы о предполагаемом законе неизвестного распределения. Он используется для проверки согласия предполагаемого вида распределения с опытными данными на основании выборки.

Существуют различные критерии согласия: Пирсона, Колмогорова, Фишера, Смирнова и др.

Критерий Пирсона (χ^2)

Критерий согласия Пирсона – наиболее часто употребляемый критерий для проверки простой гипотезы о законе распределения.

Для проверки гипотезы H_0 поступают следующим образом.

Разбивают всю область значений случайной величины X на m интервалов $\Delta_1, \Delta_2, \dots, \Delta_m$. Затем, используя формулу $P\{\alpha \leq X \leq \beta\} = F_0(\beta) - F_0(\alpha)$, подсчитывают вероятности p_i ($i=1,2,\dots,m$) попадания случайной величины X (т.е. наблюдения) в интервал Δ_i .

Теоретическое число значений случайной величины X , попавших в интервал Δ_i , рассчитывается по формуле np_i . Таким образом, имеем статистический ряд распределения случайной величины X и теоретический ряд распределения:

Δ_1	Δ_2	$\dots$	Δ_m
$n'_1 = np_1$	$n'_2 = np_2$	$\dots$	$n'_m = np_m$

Если эмпирические частоты (n_i) сильно отличаются от теоретических ($n'_i = np_i$), то проверяемую гипотезу H_0 следует отвергнуть; в противном случае – принять.

В качестве меры расхождения между эмпирическими и теоретическими частотами Пирсон предложил величину:

$$\chi^2 = \sum_{i=1}^m \frac{n_i^2}{np_i} - n.$$

Согласно теореме Пирсона, при $n \rightarrow \infty$ статистика имеет χ^2 -распределение с $k = m - r - 1$ степенями свободы, где m – число групп (интервалов) выборки, r – число параметров предполагаемого распределения.

В частности, если предполагаемое распределение нормально, то оценивают два параметра (μ и σ), поэтому число степеней свободы $k = m - 3$.

Правило применения критерия χ^2 сводится к следующему:

1. Вычисляем $\bar{x}_1$ – выборочное значение статистики критерия.
2. Выбрав уровень значимости α критерия, по таблице χ^2 распределения находим критическую точку.
3. Если $\chi_{набл.}^2 \leq \chi_{\alpha, k}^2$, то гипотеза H_0 не противоречит опытным данным. Если $\chi_{набл.}^2 > \chi_{\alpha, k}^2$, то гипотеза H_0 – отвергается.

Необходимым условием применения критерия Пирсона является наличие в каждом из интервалов не менее 5 наблюдений. Если в отдельных интервалах их меньше, то число интервалов надо уменьшить путем объединения (укрупнения) соседних интервалов.

Пример. Обследовано 100 обработанных деталей. Отклонения от заданного размера приведены в таблице:

$[x_i, x_{i+1})$	$[-3, -2)$	$[-2, -1)$	$[-1, 0)$	$[0, 1)$	$[1, 2)$	$[2, 3)$	$[3, 4)$	$[4, 5)$
n_i	3	10	15	24	25	13	7	3

Проверить при уровне значимости $\alpha = 0,01$ гипотезу H_0 о том, что отклонения от проектного размера подчиняются нормальному закону распределения.

Решение. Число наблюдений в крайних интервалах меньше 5. Для выполнения условия применения критерия Пирсона объединим их с соседними. Получим следующий ряд распределения:

$[x_i, x_{i+1})$	$[-3, -1)$	$[-1, 0)$	$[0, 1)$	$[1, 2)$	$[2, 3)$	$[3, 5)$
n_i	13	15	24	25	13	10

Случайную величину (отклонение) обозначим через X . Для вычисления вероятностей необходимо вычислить параметры, определяющие нормальный закон распределения (a и σ). Их оценки вычислим по выборке:

$$\bar{x}_1 = \frac{1}{100}(-2 \cdot 13 + (-0,5) \cdot 15 + \dots + 4 \cdot 10) = 0,885 \approx 0,9,$$

$$D_g = \frac{1}{100}(4 \cdot 13 + 0,25 \cdot 15 + \dots + 16 \cdot 10) - (0,885)^2 \approx 2,81,$$

$$\sigma \approx 1,7.$$

Находим вероятности p_i . Так как случайная величина $X \sim N(a, \sigma)$ определена на всей числовой оси $(-\infty, +\infty)$, то крайние интервалы в ряде распределения заменяем, соответственно, на $(-\infty, -1)$ и $(3, +\infty)$. Тогда

$$\begin{aligned} p_1 &= p\{-\infty < X < -1\} = \left\langle p(x_1 \leq X \leq x_2) = \Phi\left(\frac{x_2 - a}{\sigma}\right) - \Phi\left(\frac{x_1 - a}{\sigma}\right) \right\rangle = \\ &= \Phi_0\left(\frac{-1 - 0,9}{1,7}\right) - \Phi_0(-\infty) = \left\langle \Phi_0(-x) = -\Phi_0(x); \Phi_0(\infty) = \frac{1}{2} \right\rangle = \Phi_0(\infty) - \Phi_0(1,12) = \\ &= \frac{1}{2} - \Phi_0(1,12) = 0,1314. \end{aligned}$$

Рассуждая аналогичным образом, получаем

$$p_2 = 0,1667, \quad p_3 = 0,2258, \quad p_4 = 0,2183, \quad p_5 = 0,1503, \quad p_6 = 0,1075.$$

$[x_i, x_{i+1})$	$[-3, -1)$	$[-1, 0)$	$[0, 1)$	$[1, 2)$	$[2, 3)$	$[3, 5)$
n_i	13	15	24	25	13	10
$n'_i = np_1$	13,14	16,67	22,58	21,83	15,03	10,75

Вычисляем $\chi^2_{набл}$:

$$\chi^2_{набл} = \sum_{i=1}^6 \frac{n_i^2}{np_i} - n = \left(\frac{13^2}{13,14} + \frac{15^2}{16,67} + \dots + \frac{10^2}{10,75} \right) - 100 \approx 1,045.$$

Найдем число степеней свободы. По выборке рассчитаны два параметра, следовательно, $r = 2$. Количество интервалов $m = 6$. Значит, $k = 6 - 2 - 1 = 3$.

При $\alpha = 0,01$ и $k = 3$, по таблице χ^2 распределения находим $\chi^2_{0,013} = 11,3$.

Итак, $\chi^2_{набл.} < \chi^2_{\alpha,k}$, то нет оснований отвергнуть проверяемую гипотезу.

3. Регрессионный и корреляционный анализ

Корреляционный и регрессионный анализы являются фундаментальными методами статистики, используемыми для изучения взаимосвязей между переменными и количественной оценкой этих связей. Корреляционный анализ направлен на выявление и измерение степени тесноты статистической связи между двумя и более признаками, без необходимости строгой функциональной зависимости, выявляя, как изменение одной переменной связано с изменением другой. В свою очередь, регрессионный анализ позволяет не только определить наличие связи, но и формализовать эту зависимость в виде уравнения, что дает возможность прогнозировать значения зависимой переменной на основе значений независимых факторов. Эти методы широко применяются в различных областях науки и практики для описания, моделирования и прогноза процессов и явлений.

3.1. Понятие регрессионного и дисперсионного анализов

В естественных науках часто идет речь о функциональной зависимости, когда каждому значению одной переменной соответствует вполне определенное значение другой переменной.

Однако на практике между переменными величинами существуют зависимости, когда каждому значению одной переменной может соответствовать множество значений другой переменной, имеющее определенное распределение. Такая зависимость называется *статистической*.

Статистические связи между переменными можно изучать методами *корреляционного и регрессионного анализа*.

Задача корреляционного анализа – выявление связи между случайными переменными и оценка ее тесноты.

Задача регрессионного анализа – установление формы и изучение зависимости между переменными. В регрессионном анализе рассматривается зависимость случайного результативного признака y от неслучайных факторных признаков $x_1, x_2, \dots, x_n$. В случае единственного факторного признака x уравнение взаимосвязи двух переменных может быть представлено в виде:

$$y = \varphi(x) + \varepsilon,$$

где ε – случайная величина, причем предполагается, что $M(\varepsilon)=0$, а дисперсия постоянна и не зависит от x . Переменную ε называют *возмущающей переменной* или просто *возмущением*.

В зависимости от вида функции $\varphi(x)$ различают следующие виды регрессий: линейную, гиперболическую, показательную, степенную, логарифмическую, параболическую и т.д.

Предположим, что для оценки параметров регрессии взята выборка, содержащая n пар значений (x_i, y_i) , где $i=1, 2, \dots, n$. Оценкой предложенных выше уравнений регрессии являются выборочные уравнения регрессии:

1. Линейное $\hat{y} = a_0 + a_1 x$;
2. Гиперболическое $\hat{y} = a_0 + \frac{a_1}{x}$;
3. Показательное $\hat{y} = a_0 a_1^x$;
4. Степенное $\hat{y} = a_0 x^{a_1}$;
5. Логарифмическое $\hat{y} = a_0 + a_1 \ln x$;
6. Параболическое $\hat{y} = a_0 + a_1 x + a_2 x^2$,

где параметры a_0, a_1, a_2 являются точечными оценками соответствующих параметров исходного уравнения и могут быть найдены на основе *метода наименьших квадратов*.

3.2. Построение регрессионной модели

Сущность метода наименьших квадратов заключается в нахождении параметров модели a_0, a_1 , при которых минимизируется сумма квадратов отклонений эмпирических (фактических) значений результативного признака от теоретических, полученных по выборочному уравнению регрессии:

$$S = \sum_{i=1}^n (y_i - \hat{y})^2 \rightarrow \min.$$

Для линейной модели:

$$S = \sum_{i=1}^n (y_i - a_0 - a_1 x_i)^2 \rightarrow \min.$$

Функция двух переменных $S(a_0, a_1)$ может достигнуть экстремума в том случае, когда ее частные производные равны нулю. Вычисляя эти частные производные, получим систему уравнений для нахождения параметров a_0, a_1 линейного уравнения регрессии $\hat{y} = a_0 + a_1 x$.

Частные производные:

$$\begin{aligned} \frac{\partial S}{\partial a_0} &= 2 \sum_{i=1}^n (y_i - a_0 - a_1 x_i) \cdot (y_i - a_0 - a_1 x_i)' = 2 \sum_{i=1}^n (a_0 + a_1 x_i - y_i); \\ \frac{\partial S}{\partial a_1} &= 2 \sum_{i=1}^n (y_i - a_0 - a_1 x_i) \cdot (y_i - a_0 - a_1 x_i)' = 2 \sum_{i=1}^n (y_i - a_0 - a_1 x_i) \cdot (-x_i) = \\ &= 2 \sum_{i=1}^n (a_0 + a_1 x_i - y_i) \cdot x_i. \end{aligned}$$

Таким образом, система нормальных уравнений имеет вид:

$$\begin{cases} \frac{\partial S}{\partial a_0} = 2 \sum_{i=1}^n (a_0 + a_1 x_i - y_i) = 0 \\ \frac{\partial S}{\partial a_1} = 2 \sum_{i=1}^n (a_0 + a_1 x_i - y_i) x_i = 0 \end{cases}$$

После упрощения, получаем систему:

$$\begin{cases} \sum_{i=1}^n a_0 + \sum_{i=1}^n a_1 x_i = \sum_{i=1}^n y_i \\ \sum_{i=1}^n a_0 x_i + \sum_{i=1}^n a_1 x_i^2 = \sum_{i=1}^n y_i x_i \end{cases}$$

или

$$\begin{cases} na_0 + a_1 \sum x_i = \sum y_i \\ a_0 \sum x_i + a_1 \sum x_i^2 = \sum x_i y_i \end{cases}$$

В случае, когда возмущающая переменная ε имеет нормальное распределение, коэффициенты a_0, a_1 , полученные методом наименьших квадратов для линейной регрессии, являются несмещенными и эффективными оценками параметров α_0, α_1 исходного уравнения.

Для параболического уравнения $\hat{y} = a_0 + a_1x + a_2x^2$ регрессии система уравнений для нахождения параметров a_0, a_1, a_2 имеет вид:

$$\begin{cases} na_0 + a_1 \sum x_i + a_2 \sum x_i^2 = \sum y_i \\ a_0 \sum x_i + a_1 \sum x_i^2 + a_2 \sum x_i^3 = \sum x_i y_i \\ a_0 \sum x_i^2 + a_1 \sum x_i^3 + a_2 \sum x_i^4 = \sum x_i^2 y_i \end{cases}$$

Замечание. Все предложенные выше виды нелинейных регрессий (кроме параболической) могут быть сведены к линейной путем какой-либо замены переменной.

Для гиперболической регрессии вводится переменная $x' = 1/x$, для логарифмической регрессии $x' = \ln x$, уравнения показательной и степенной регрессии предварительно логарифмируют.

Для показательного уравнения получаем: $\ln \hat{y} = \ln a_0 + x \ln a_1$

Делаем замены: $\hat{y}' = \ln \hat{y}, a'_0 = \ln a_0, a'_1 = \ln a_1$.

Таким образом, получаем линейное уравнение регрессии:
 $\hat{y}' = a'_0 + a'_1 x$

Найдя коэффициенты этого уравнения методом наименьших квадратов, получим коэффициенты исходного уравнения: $a_0 = e^{a'_0}; a_1 = e^{a'_1}$.

Для степенного уравнения: $\ln \hat{y} = \ln a_0 + a_1 \ln x$

Делаем замены: $x' = \ln x, \hat{y}' = \ln \hat{y}, a'_0 = \ln a_0$.

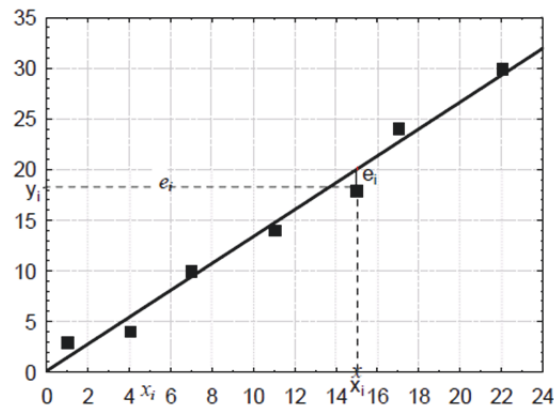
Таким образом, получаем линейное уравнение регрессии:
 $\hat{y}' = a'_0 + a_1 x'$

Найдя коэффициенты этого уравнения методом наименьших квадратов, вычислим и коэффициент исходного уравнения: $a_0 = e^{a'_0}$.

Регрессионную модель удобно представлять графически. Рассмотрим выборку $n=7$.

X	1	4	7	11	15	17	22
Y	3	4	10	14	18	24	30

Отложим на координатной плоскости точки $P_i(x_i, y_i)$, $(i=1, 2, \dots, n)$. Полученный график называется диаграммой рассеивания. Видим, что точки группируются около некоторой прямой $y = a_0 + a_1 x$.



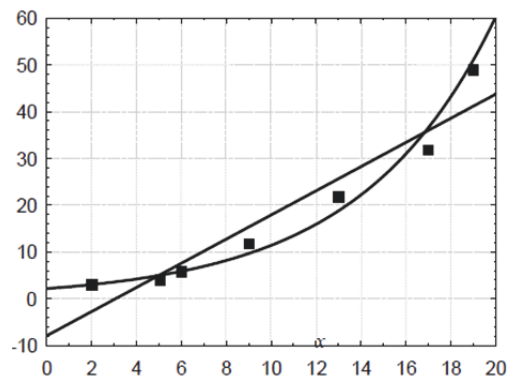
Найдя коэффициенты этой прямой $a_0=0,876$; $a_1=1,284$ на этом же графике проведем линию, соответствующую уравнению регрессии. Отклонение точки P_i от оцениваемой линии, измеренное по вертикали, будет равно: $\varepsilon_i = (y_i - a_0 - a_1 x_i)$.

Тогда функцию $S(a_0, a_1)$, определенную $S = \sum (y_i - a_0 - a_1 x_i)^2$.

Можно представить как $S = \sum \varepsilon_i^2$.

Таким образом, если фактические данные точно лежали бы на некоторой прямой, то $S=0$, а чем дальше точки отклоняются от линии регрессии, тем больше значение S .

Построив диаграмму рассеяния, можно подобрать вид уравнения регрессии.



По одним и тем же экспериментальным данным были получены линейная и экспоненциальная зависимости, причем экспериментальные точки располагаются ближе именно к экспоненциальной кривой.

Однако по графику можно только приближенно сделать вывод о качестве той или иной модели. Существуют способы более точной оценки адекватности (значимости) уравнения регрессии.

Под проверкой значимости уравнения регрессии понимается оценка соответствия предложенной математической модели, отражающей связь между переменными, реально наблюдаемым экспериментальным данным.

Для проверки значимости уравнения регрессии на уровне значимости α вычисляют наблюдаемое значение случайной величины:

$$F_{набл} = \frac{\sigma_y^2(n-2)}{\sigma_{ост}^2},$$

где остаточная дисперсия и дисперсия уравнения регрессии находятся по формулам:

$$\sigma_{ост}^2 = \frac{\sum (y_i - \hat{y}_i)^2}{n-1}; \quad \sigma_y^2 = \frac{\sum (\bar{y} - \hat{y}_i)^2}{n-1}.$$

Значение $F_{набл}$ показывает, насколько лучше уравнение регрессии оценивает значение результативного признака по сравнению с его средней.

Далее находим критическое значение критерия F по таблице критических точек распределения Фишера при $k_1=1$, $k_2=n-2$ степенях свободы и уровне значимости α . Если $F_{набл} > F(\alpha; 1; n-2)$, то уравнение регрессии признается значимым, в противном случае уравнение регрессии признается незначимым, т.е. статистически подтверждается отсутствие линейной связи между факторным и результативным признаком.

По расположению точек на диаграмме рассеяния не всегда можно принять окончательное решение о виде уравнения регрессии. Если теоретические соображения не могут подсказать точного решения, то следует сделать расчеты по двум или более уравнениям.

Предпочтение следует отдать той модели, для которой *меньше величина остаточной дисперсии* (меньше величина остаточной суммы квадратов S).

Однако, при незначительных расхождениях в остаточных дисперсиях, следует остановиться на более простом уравнении.

Рассмотрим более подробно линейное уравнение регрессии. Коэффициент a_1 показывает величину изменения результативного признака при изменении факторного признака на единицу.

Следовательно, коэффициент a_1 зависит от единиц измерения переменных. Поэтому в качестве универсального показателя тесноты связи между величинами x и y используется *выборочный линейный коэффициент корреляции*:

$$r = a_1 \frac{s_x}{s_y},$$

где s_x и s_y – средние квадратические отклонения соответствующих признаков (факторного и результативного). Они вычисляются по формулам:

$$s_x^2 = \frac{\sum (x_i - \bar{x}_e)^2}{n-1}; s_y^2 = \frac{\sum (y_i - \bar{y}_e)^2}{n-1},$$

где $\bar{x}_e$, $\bar{y}_e$ – выборочные средние значения.

Линейный коэффициент корреляции изменяется в пределах:

$$-1 \leq r \leq 1.$$

Если $r > 0$, то связь между переменными x и y прямая, если $r < 0$, то связь между переменными x и y обратная. В зависимости от того, насколько $|r|$ приближается к 1, различают связь слабую, умеренную и сильную.

При $|r| < 0,2$ связь между переменными практически отсутствует. При $|r|=1$ связь между x и y функциональная, т.е. наблюдаемые значения располагаются точно на прямой.

Силу связи можно определить по шкале Чеддока:

Величина $ r_{xy} $	Характер связи
0-0,3	Практически отсутствует
0,3-0,5	Слабая
0,5-0,7	Умеренная
0,7-1	Сильная

Выборочный коэффициент корреляции r является оценкой генерального коэффициента корреляции ρ , который определяется по формуле:

$$\rho = \frac{M(xy) - M(x)M(y)}{\sigma_x \sigma_y}.$$

Величина ρ характеризует тесноту связи между переменными x и y в генеральной совокупности. Однако в практических исследованиях о тесноте корреляционной зависимости между переменными судят фактически не по величине генерального коэффициента корреляции ρ , а по величине его выборочного аналога r .

Пусть вычисленное значение $r \neq 0$. Возникает вопрос, объясняется ли это действительно существующей линейной связью между переменными x и y или является следствием случайности отбора переменных в выборку. Обычно в этих случаях проверяется гипотеза H_0 об отсутствии линейной корреляционной связи между переменными, т.е. $H_0: \rho = 0$ при альтернативной гипотезе $H_1: \rho \neq 0$.

Для проверки этой гипотезы на уровне значимости α вычисляют наблюдаемое значение критерия:

$$T_{набл} = \frac{|r|}{\sqrt{1-r^2}} \sqrt{n-2}.$$

Критическое значение критерия $T(1-\alpha, n-2)$ находят по таблице критических точек распределения Стьюдента для числа степеней свободы $n-2$ и уровня значимости α . Если $T_{набл} < T(\alpha, n-2)$, то гипотеза H_0 принимается, в противном случае гипотеза H_0 отвергается, т.е. коэффициент корреляции признается существенно отличающимся от нуля.

Средняя ошибка аппроксимации – среднее отклонение расчетных значений от фактических в процентах:

$$\bar{A} = A_{cp.} = \frac{1}{n} \sum \left| \frac{y_i - \hat{y}_i}{y_i} \right| \cdot 100\%.$$

Ошибка аппроксимации $A_{cp.}$ в пределах 8-10% свидетельствует о хорошем подборе модели к исходным данным.

Пример. Исследуем связь между объемом поставляемых товаров и затратами на их доставку. Объем поставок обозначим как X , а стоимость доставки – как Y .

Исходные данные:

Номер поставки	Объём поставок, X (тонны)	Стоимость доставки, Y (рублей)
1	10	500
2	15	600
3	20	700
4	25	800
5	30	900

Решение.

Определим наличие связи методом корреляционного анализа. Вычислим коэффициент линейной корреляции Пирсона, который покажет силу и направление связи между X и Y .

Формула коэффициента корреляции Пирсона:

$$r = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2 \sum (Y_i - \bar{Y})^2}}.$$

Вычислим средние значения:

$$\bar{X} = \frac{10 + 15 + 20 + 25 + 30}{5} = 20,$$

$$\bar{Y} = \frac{500 + 600 + 700 + 800 + 900}{5} = 700.$$

Для упрощения вычислений, составим вспомогательную таблицу.

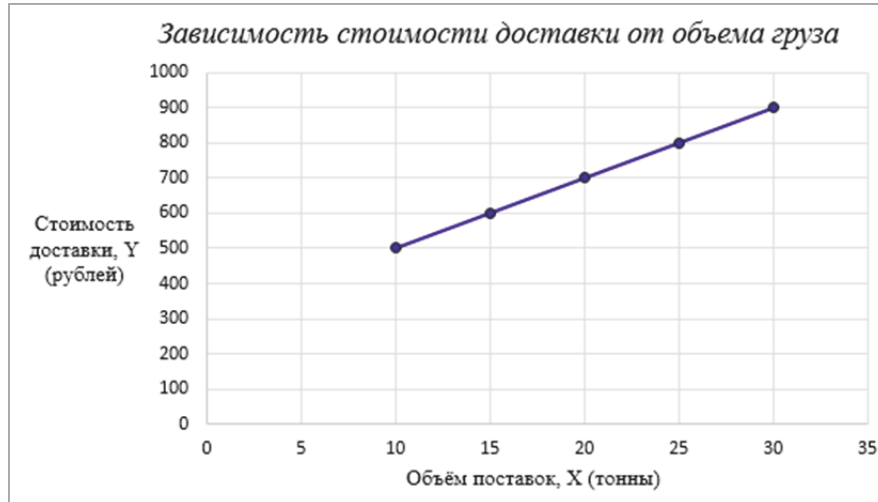
X_i	Y_i	$X_i - \bar{X}$	$Y_i - \bar{Y}$	$(X_i - \bar{X})(Y_i - \bar{Y})$	$(X_i - \bar{X})^2$	$(Y_i - \bar{Y})^2$
10	500	-10	-200	2000	100	40000
15	600	-5	-100	500	25	10000
20	700	0	0	0	0	0
25	800	5	100	500	25	10000
30	900	10	200	2000	100	40000
сумма				5000	250	100000

Вычислим коэффициент корреляции:

$$r = \frac{5000}{\sqrt{250 \cdot 100000}} = \frac{5000}{5000} = 1.$$

Коэффициент корреляции $r = 1$, что говорит о наличии сильной положительной линейной связи между объемом поставок и стоимостью доставки.

Изобразим графически зависимость между объемом поставляемых товаров и затратами на их доставку.



Графическое представление данных подтверждает наличие сильной, прямой линейной зависимости между объемом поставляемых товаров и затратами на их доставку.

Далее проведем регрессионный анализ, чтобы выразить зависимость стоимости доставки (Y) от объема поставок (X). Используем метод наименьших квадратов.

Уравнение простой линейной регрессии выглядит так:

$$Y = a + bX$$

Параметры модели находятся по формулам:

$$b = r \frac{S_Y}{S_X}, \quad a = \bar{Y} - b\bar{X}.$$

Параметр наклона линии регрессии (b): $b = 1 \cdot \frac{158,11}{7,91} \approx 20$.

Свободный член (a): $a = 700 - 20 \cdot 20 = 300$.

Регрессионное уравнение: $Y = 300 + 20X$.

Коэффициент наклона $b = 20$ показывает, что при увеличении объема поставок на одну тонну приводит к увеличению стоимости доставки на 20 рублей.

Свободный член $a = 300$ указывает на то, что даже при отсутствии груза (объем поставок равен нулю) существует фиксированная минимальная стоимость доставки, составляющая 300 рублей.

Полученное уравнение регрессии $Y = 300 + 20X$ позволит точнее планировать расходы на транспортировку и минимизировать финансовые потери

Пример. Компания занимается доставкой грузов автомобильным транспортом по различным маршрутам разной протяженности. Необходимо оценить зависимость времени доставки груза от пройденного расстояния.

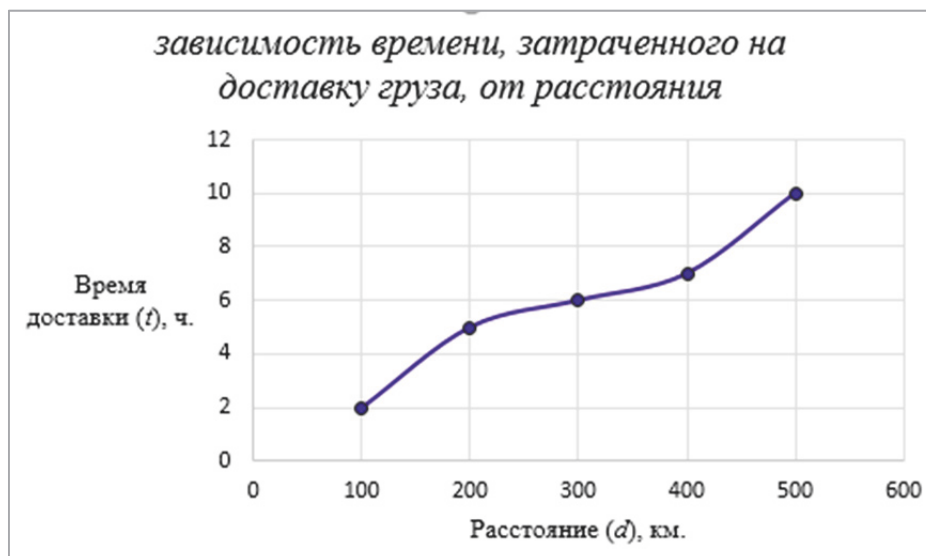
Имеются следующие экспериментальные данные по нескольким доставкам, представленные в таблице:

Расстояние (d), км.	Время доставки (t), ч.
100	2
200	5
300	6
400	7
500	10

Решение.

Необходимо определить линейную зависимость вида $t = ad + b$, где a и b – коэффициенты регрессии.

Изобразим графически зависимость времени, затраченного на доставку груза, от расстояния.



На основе графического представления данных, можно сделать вывод, что наблюдается прямая линейная зависимость времени, затраченного на доставку груза, от расстояния.

Найдем параметры регрессионного уравнения методом наименьших квадратов.

Для упрощения вычислений, составим вспомогательную таблицу.

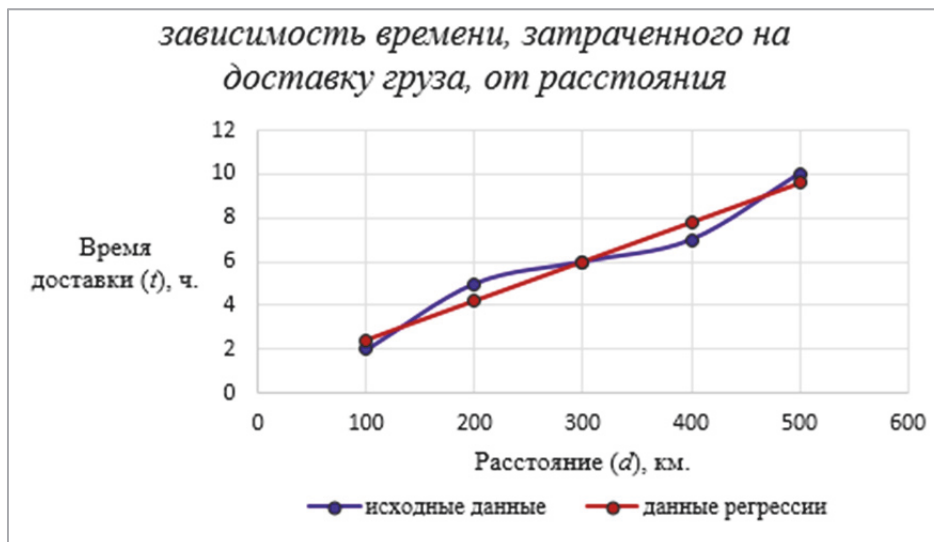
№	X	Y	$X_i - X_{cp}$	$Y_i - Y_{cp}$	$(X_i - X_{cp})(Y_i - Y_{cp})$	$(X_i - X_{cp})^2$	$(Y_i - Y_{cp})^2$
1	10	500	-10	-200	2000	100	40000
2	15	600	-5	-100	500	25	10000
3	20	700	0	0	0	0	0
4	25	800	5	100	500	25	10000
5	30	900	10	200	2000	100	40000
сумма	100	3500	0	0	5000	250	100000

Параметр наклона линии регрессии (b):.

Параметр наклона линии регрессии (a): $a = 1 \cdot \frac{158,114}{2,915} \approx 0,0184$ ч/км.

Свободный член (b): $b=0,6$.

Уравнение зависимости принимает вид: $t=0,018d+0,6$.



Коэффициент наклона $a = 0,018$ означает, что на каждые дополнительные 100 километров пути доставка увеличивается примерно на 1 час. Свободный член $b = 0,6$ показывает минимальное базовое время доставки независимо от расстояния (например, подготовка автомобиля, оформление документов и т.п.).

Проверим адекватность построенной модели, используя найденные коэффициенты для новых точек данных. Например, для расстояния $d = 600$ км расчетное время:

$$t_{расч} = 0,018 \cdot 600 + 0,6 = 11,4 \text{ часов.}$$

При наличии реальных данных по доставке на расстояние 600 км, можем сравнить рассчитанное значение с фактическим временем и оценить точность нашей модели.

Таким образом, получили простую и понятную формулу для прогнозирования времени доставки в зависимости от расстояния, основанную на методе наименьших квадратов.

Корреляционный и регрессионный анализ являются важными статистическими методами, используемыми для изучения взаимосвязей между переменными. Корреляционный анализ позволяет выявить наличие и степень зависимости между величинами, количественно измеряемую с помощью коэффициента корреляции, который показывает силу и направление связи. В свою очередь, регрессионный анализ не только определяет наличие связи, но и формализует её в виде математической модели (уравнения регрессии), позволяющей количественно оценить влияние одной или нескольких независимых переменных на зависимую, а также осуществлять прогнозирование значений зависимой переменной.

Таким образом, корреляционный анализ служит основой для выявления взаимосвязей, а регрессионный – инструментом для их более глубокого анализа и практического использования в задачах прогнозирования и оценки влияния факторов.

3.3. Построение линейной модели парной регрессии с помощью надстройки Анализ данных

Для построения регрессионных моделей можно воспользоваться процедурой *Регрессия* надстройки *Пакет анализа*. Рассмотрим пример.

Задача. По двенадцати регионам Южного Федерального округа известны значения двух признаков:

№ региона	1	2	3	4	5	6	7	8
y_i	6,9	9,1	9,5	7,6	6,0	5,8	9,6	6,7
x_i	25,8	59,6	122,7	29,5	13,0	15,1	129,0	24,0

где x – среднедушевой прожиточный минимум в день одного трудоспособного (у.е.); y – среднедневная заработная плата (у.е).

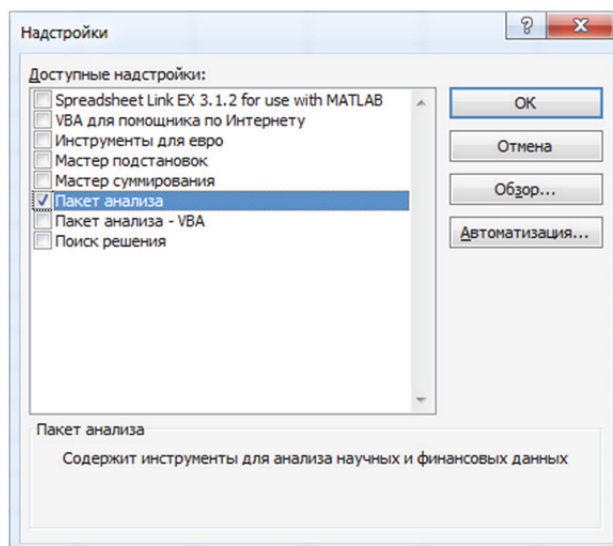
Требуется решить задачу с помощью *Регрессии* надстройки *Анализ данных* табличного редактора *MSExcel*.

Решение.

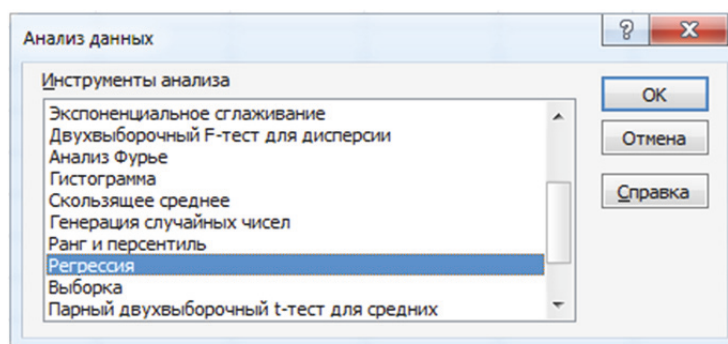
1. Откроем новую книгу *MSExcel*.
2. На листе 1 создадим и заполним следующую таблицу:

	A	B	C
1	№	x_i	y_i
2	1	25,8	6,9
3	2	59,6	9,1
4	3	121,7	9,5
5	4	29,5	7,6
6	5	13	6
7	6	15,1	5,8
8	7	129	9,6
9	8	24	6,7

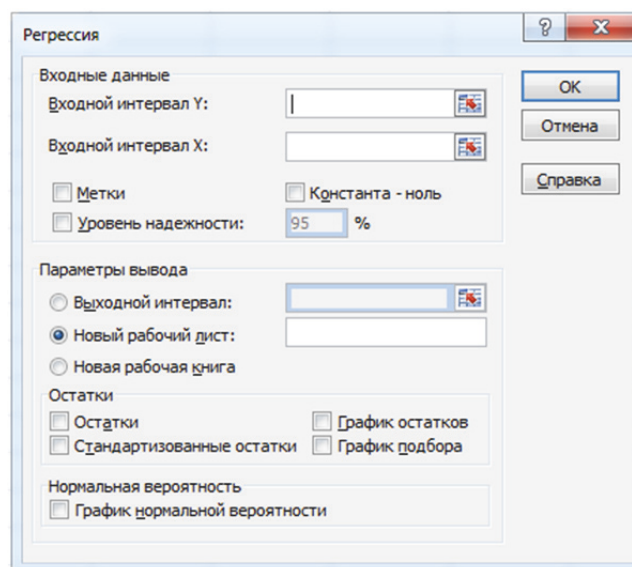
3. В надстройках *MSExcel* установим *Пакет анализа*.



4. В главном меню выбираем *Данные* → *Анализ данных* → *Регрессия*. Нажимаем *ОК*.



5. Заполняем диалоговое окно:



Входной интервал Y– диапазон, содержащий данные результативного признака;

Входной интервал X – диапазон, содержащий данные факторов независимого признака;

Метки – флажок, который указывает, содержит ли первая строка название столбцов или нет;

Константа – ноль – флажок, указывающий на наличие или отсутствие свободного члена в уравнении;

Выходной интервал – достаточно указать левую верхнюю ячейку будущего диапазона;

Новый рабочий лист – можно задать произвольное имя нового листа.

Регрессия

Входные данные

Входной интервал Y:

Входной интервал X:

☒ Метки ☐ Константа - ноль

☒ Уровень надежности: %

Параметры вывода

☒ Выходной интервал:

☐ Новый рабочий лист:

☐ Новая рабочая книга

Остатки

☒ Остатки ☐ График остатков

☐ Стандартизованные остатки ☐ График подбора

Нормальная вероятность

☐ График нормальной вероятности

OK Отмена Справка

Нажимаем **OK**.

6. Получаем следующие результаты регрессионного анализа:

	A	B	C	D	E	F	G	H	I
11	ВЫВОД ИТОГОВ								
12									
13	Регрессионная статистика								
14	Множественный R	0,911456497							
15	R-квадрат	0,830752946							
16	Нормированный R-квадрат	0,802545104							
17	Стандартная ошибка	0,691260333							
18	Наблюдения	8							
19									
20	Дисперсионный анализ								
21		df	SS	MS	F	Значимость F			
22	Регрессия	1	14,07295491	14,07295491	29,45113414	0,001622236			
23	Остаток	6	2,867045088	0,477840848					
24	Итого	7	16,94						
25									
26		Коэффициенты	Стандартная ошибка	t-статистика	P-Значение	Нижние 95%	Верхние 95%	Нижние 95,0%	Верхние 95,0%
27	Y-пересечение	6,086978636	0,377733088	16,11449677	3,6304E-06	5,162699067	7,011258205	5,162699067	7,011258205
28	x1	0,029935769	0,005516193	5,426889914	0,001622236	0,016438133	0,043433406	0,016438133	0,043433406
29									
30	ВЫВОД ОСТАТКА								
31									
32	Наблюдение	Предсказанное y1	Остатки						
33	1	6,859321489	0,040678511						
34	2	7,871150497	1,228849503						
35	3	9,730161782	-0,230161782						
36	4	6,970083836	0,629916164						
37	5	6,476143639	-0,476143639						
38	6	6,539008755	-0,739008755						
39	7	9,948692899	-0,348692899						
40	8	6,805437104	-0,105437104						

Параметры $a_0 = 6,087$ и $a_1 = 0,0299$ уравнения регрессии находим из столбца «Коэффициенты», т.е. получаем линейное регрессионное уравнение:

$$\hat{y} = 6,087 + 0,0299x.$$

Значения линейной модели (прогнозные значения) $y(x)$ находим из столбца, «Предсказанное y_1 ».

Линейный коэффициент парной корреляции $r_{xy} = 0,9115$ находим из

столбца «Регрессионная статистика», «Множественный R ».

Коэффициент детерминации $R^2 = 0,8308$ находим из столбца «Регрессионная статистика», « R - квадрат».

Корень квадратный из остаточной дисперсии $S_{ост.} = 0,6913$ находим из столбца «Регрессионная статистика», «Стандартная ошибка».

Значение $F_{факт} = 29,451$ находим из столбца « F ».

Значения t -статистики $t_a = 16,114$, $t_b = 5,426$ находим из столбца « t -статистика».

Доверительные интервалы $\gamma_{a\min} = 5,1615$, $\gamma_{a\max} = 7,0124$, $\gamma_{b\min} = 0,0164$, $\gamma_{b\max} = 0,0435$ находим из столбцов «Нижние 95%», «Верхние 95%».

Заключение

Настоящее учебное пособие представляет собой систематизированное изложение основ статистических методов анализа данных, необходимых для обработки и интерпретации информации в исследовательской работе, управленческой практике и повседневной профессиональной деятельности. Материал пособия охватывает широкий спектр вопросов, начиная от базовых понятий статистики и заканчивая продвинутыми техниками статистического моделирования и проверки гипотез.

В первом разделе были рассмотрены основы статистики и способы сбора данных, включая принципы формирования выборочных совокупностей, методы первичного анализа данных, работу с абсолютными и относительными показателями, формирование рядов распределения и визуализацию данных. Особое внимание уделялось предварительной обработке данных, способам выявления ошибок и аномалий, а также принципам группировки и нормализации данных.

Во втором разделе подробно описаны инструменты inferential statistics, позволяющие на основе ограниченных выборочных данных делать обоснованные выводы о характеристиках генеральной совокупности. Были детально раскрыты важнейшие положения закона больших чисел и центральной предельной теоремы, представлены подходы к точечным и интервальным оценкам параметров распределения, рассмотрены методики проверки статистических гипотез различного типа, в том числе гипотезы о средних значениях и дисперсиях.

Третий раздел посвящен регрессионному и корреляционному анализу, необходимым для выявления и количественной оценки взаимосвязей между различными факторами и признаками. Читатель получил представление о методах построения регрессионных моделей, выявлении взаимосвязей между несколькими переменными и оценке точности прогнозов.

Практические примеры, представленные в учебном пособии, наглядно продемонстрировали возможности и ограничения статистических инстру-

ментов, подчеркнули важность правильного подхода к сбору и подготовке данных, выбору адекватных методов анализа и правильной интерпретации результатов.

Литература

1. Гмурман В.Е. Теория вероятностей и математическая статистика. Учеб. пособие для вузов. Изд. 7-е, стер. – М.: Высш. шк., 1999. – 479 с.
2. Горошко, И.В. Математические методы исследования социальных систем: курс лекций / И.В. Горошко, Б.А. Торопов, Ш.Х. Гонов. М.: Академия управления МВД России, 2019. 80 с. ISBN 978-5-906942-95-1. Текст: непосредственный + Текст: электронный. Режим доступа: <http://lib.krdumvd.ru/elib/5765>, требуется авторизация.
3. Долгова, В.Н. Статистика: учебник и практикум для бакалавров / В.Н. Долгова, Т.Ю. Медведева. 2-е изд., перераб. и доп. М.: Юрайт, 2014. 627 с. (Бакалавр. Базовый курс). ISBN 978-5-9916-2946-1. Текст: непосредственный.
4. Кондрашенкова, Н.Е. Математическая статистика: учебное пособие / Н.Е. Кондрашенкова. – Санкт-Петербург: ИЭО СПбУТУиЭ, 2012. – 207 с.
5. Логинова, С.Л. Общая теория статистики: конспект лекций: учеб. пособие для вузов [Гриф УМО] / С.Л. Логинова; Рос. гос. проф.-пед. ун-т. – Екатеринбург: Издательство РГППУ, 2011. – 89 с.
6. Математическая статистика: Методические разработки и контрольные задания / Авт.-сост.: С.В. Плотникова. Тамбов: Изд-во Тамб. гос. техн. ун-та, 2005. 52 с.
7. Михайленко, Е.В. Автоматизация обработки табличных данных для сбора и анализа социальноэкономических показателей сотрудниками подразделений ОВД: учеб. пособие / Е.В. Михайленко, А.А. Хромых. Краснодар: Краснодарский университет МВД России, 2021. 148 с. ISBN 978-5-9266-1725-9. Текст: непосредственный + Текст: электронный. Режим доступа: <http://lib.krdumvd.ru/elib/6113>, требуется авторизация.
8. Михайленко, Е.В. Парная регрессия и корреляция: лекция / Е.В. Михайленко. – Краснодар: Краснодарский университет МВД России, 2008. – Б. ц. – Текст: непосредственный + Текст: электронный.

9. Михайленко, Е.В. Прикладная математика: курс лекций / Е.В. Михайленко, А.А. Хромых. Краснодар: Краснодарский университет МВД России, 2014. 192 с. ISBN 978-5-9266-0766-3. Текст: непосредственный + Текст: электронный. Режим доступа: <http://lib.krdumvd.ru/elib/4843>, требуется авторизация.

10. Михайленко, Е.В. Прикладная математика: учеб. пособие / Е.В. Михайленко. Электрон. текстовые дан. Краснодар: Краснодарский университет МВД России, 2018. эл. опт. диск (CD-ROM). ISBN 978-5-9266-1371-8: Текст: непосредственный + Текст: электронный. Режим доступа: <http://lib.krdumvd.ru/elib/5552>, требуется авторизация.

11. Михайленко, Е.В. Теория вероятностей и математическая статистика: учеб. пособие / Е.В. Михайленко, М.А. Жукова, Ф.Ф. Бараненко. Краснодар: Краснодарский университет МВД России, 2011. 142 с. Текст: непосредственный + Текст: электронный. Режим доступа: <http://lib.krdumvd.ru/elib/4624>, требуется авторизация.

12. Старостенко, И.Н. Парная регрессия и корреляция: учеб. пособие / И.Н. Старостенко, А.А. Хромых. Электрон. текстовые дан. Краснодар: Краснодарский университет МВД России, 2017. эл. опт. диск (CD-ROM). ISBN 978-5-9266-1243-8. Текст: электронный. Режим доступа: <http://lib.krdumvd.ru/elib/6350>, требуется авторизация.

13. Старостенко, И.Н. Элементы статистической обработки данных: учеб. пособие / И.Н. Старостенко, А.А. Хромых. Краснодар: Краснодарский университет МВД России, 2020. 74 с. ISBN-978-5-9266-1613-9. Текст: непосредственный + Текст: электронный. Режим доступа: <http://lib.krdumvd.ru/elib/470>, требуется авторизация.

14. Торопов, Б.А. Математические методы исследования социальных систем: учеб. пособие / Б.А. Торопов, Э.Ф. Болтачев, В.В. Баранов. М.: Академия управления МВД России, 2020. 80 с. ISBN 978-5-907187-30-6. Текст: непосредственный + Текст: электронный. Режим доступа: <http://lib.krdumvd.ru/elib/6045>, требуется авторизация.

15. Торопов, Б.А. Статистические методы принятия управленческих решений: сб. задач (задачник) / Б.А. Торопов, Ш.Х. Гонов. М.: Академия управления МВД России, 2019. 76 с. ISBN 978-5-906942-84-5. Текст: непосредственный + Текст: электронный. Режим доступа: <http://lib.krdumvd.ru/elib/5897>, требуется авторизация.

ТАБЛИЦА ЗНАЧЕНИЙ ФУНКЦИИ $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_0^x e^{-\frac{z^2}{2}} dz$

0,00	0,0000	0,41	0,1591	0,82	0,2939	1,23	0,3907
0,01	0,0040	0,42	0,1628	0,83	0,2967	1,24	0,3925
0,02	0,0080	0,43	0,1664	0,84	0,2995	1,25	0,3944
0,03	0,0120	0,44	0,1700	0,85	0,3023	1,26	0,3962
0,04	0,0160	0,45	0,1736	0,86	0,3051	1,27	0,3980
0,05	0,0199	0,46	0,1772	0,87	0,3078	1,28	0,3997
0,06	0,0239	0,47	0,1808	0,88	0,3106	1,29	0,4015
0,07	0,0279	0,48	0,1844	0,89	0,3133	1,30	0,4032
0,08	0,0319	0,49	0,1879	0,90	0,3159	1,31	0,4049
0,09	0,0359	0,50	0,1915	0,91	0,3186	1,32	0,4066
0,10	0,0398	0,51	0,1950	0,92	0,3212	1,33	0,4082
0,11	0,0438	0,52	0,1985	0,93	0,3238	1,34	0,4099
0,12	0,0478	0,53	0,2019	0,94	0,3264	1,35	0,4115
0,13	0,0517	0,54	0,2054	0,95	0,3289	1,36	0,4131
0,14	0,0557	0,55	0,2088	0,96	0,3315	1,37	0,4147
0,15	0,0596	0,56	0,2123	0,97	0,3340	1,38	0,4162
0,16	0,0636	0,57	0,2157	0,98	0,3365	1,39	0,4177
0,17	0,0675	0,58	0,2190	0,99	0,3389	1,40	0,4192
0,18	0,0714	0,59	0,2224	1,00	0,3413	1,41	0,4207
0,19	0,0753	0,60	0,2257	1,01	0,3438	1,42	0,4222
0,20	0,0793	0,61	0,2291	1,02	0,3461	1,43	0,4236
0,21	0,0832	0,62	0,2324	1,03	0,3485	1,44	0,4251
0,22	0,0871	0,63	0,2357	1,04	0,3508	1,45	0,4265
0,23	0,0910	0,64	0,2389	1,05	0,3531	1,46	0,4279
0,24	0,0948	0,65	0,2422	1,06	0,3554	1,47	0,4292
0,25	0,0987	0,66	0,2454	1,07	0,3577	1,48	0,4306
0,26	0,1026	0,67	0,2486	1,08	0,3599	1,49	0,4319
0,27	0,1064	0,68	0,2517	1,09	0,3621	1,50	0,4332
0,28	0,1103	0,69	0,2549	1,10	0,3643	1,51	0,4345
0,29	0,1141	0,70	0,2580	1,11	0,3665	1,52	0,4357
0,30	0,1179	0,71	0,2611	1,12	0,3686	1,53	0,4370
0,31	0,1217	0,72	0,2642	1,13	0,3708	1,54	0,4382
0,32	0,1255	0,73	0,2673	1,14	0,3729	1,55	0,4394
0,33	0,1293	0,74	0,2703	1,15	0,3749	1,56	0,4406
0,34	0,1331	0,75	0,2734	1,16	0,3770	1,57	0,4418
0,35	0,1368	0,76	0,2764	1,17	0,3790	1,58	0,4429
0,36	0,1406	0,77	0,2794	1,18	0,3810	1,59	0,4441
0,37	0,1443	0,78	0,2823	1,19	0,3830	1,60	0,4452
0,38	0,1480	0,79	0,2852	1,20	0,3849	1,61	0,4463
0,39	0,1517	0,80	0,2881	1,21	0,3869	1,62	0,4474
0,40	0,1554	0,81	0,2910	1,22	0,3883	1,63	0,4484

Продолжение приложения 1

1,64	0,4495	1,88	0,4699	2,24	0,4875	2,72	0,4967
1,65	0,4505	1,89	0,4706	2,26	0,4881	2,74	0,4969
1,66	0,4515	1,90	0,4713	2,28	0,4887	2,76	0,4971
1,67	0,4525	1,91	0,4719	2,30	0,4893	2,78	0,4973
1,68	0,4535	1,92	0,4726	2,32	0,4898	2,80	0,4974
1,69	0,4545	1,93	0,4732	2,34	0,4904	2,82	0,4976
1,70	0,4554	1,94	0,4738	2,36	0,4909	2,84	0,4977
1,71	0,4564	1,95	0,4744	2,38	0,4913	2,86	0,4979
1,72	0,4573	1,96	0,4750	2,40	0,4918	2,88	0,4980
1,73	0,4582	1,97	0,4756	2,42	0,4922	2,90	0,4981
1,74	0,4591	1,98	0,4761	2,44	0,4927	2,92	0,4982
1,75	0,4599	1,99	0,4767	2,46	0,4931	2,94	0,4984
1,76	0,4608	2,00	0,4772	2,48	0,4934	2,96	0,4985
1,77	0,4616	2,02	0,4783	2,50	0,4938	2,98	0,4986
1,78	0,4625	2,04	0,4793	2,52	0,4941	3,00	0,49865
1,79	0,4633	2,06	0,4803	2,54	0,4945	3,20	0,49931
1,80	0,4641	2,08	0,4812	2,56	0,4948	3,40	0,49966
1,81	0,4649	2,10	0,4821	2,58	0,4951	3,50	0,49975
1,82	0,4656	2,12	0,4830	2,60	0,4953	3,60	0,499841
1,83	0,4664	2,14	0,4838	2,62	0,4956	3,80	0,499928
1,84	0,4671	2,16	0,4846	2,64	0,4959	4,00	0,499968
1,85	0,4678	2,18	0,4854	2,66	0,4961	4,50	0,499997
1,86	0,4686	2,20	0,4861	2,68	0,4963	5,00	0,499997
1,87	0,4693	2,22	0,4868	2,70	0,4965		

Приложение 2

ТАБЛИЦА ЗНАЧЕНИЙ $t_{\gamma} = t(\gamma, n)$

n	γ			n	γ		
	0,95	0,99	0,999		0,95	0,99	0,999
5	2,78	4,60	8,61	20	2,093	2,861	3,883
6	2,57	4,03	6,86	25	2,064	2,797	3,745
7	2,45	3,71	5,96	30	2,045	2,756	3,659
8	2,37	3,50	5,41	35	2,032	2,720	3,600
9	2,31	3,36	5,04	40	2,023	2,708	3,558
10	2,26	3,25	4,78	45	2,016	2,692	3,527
11	2,23	3,17	4,59	50	2,009	2,679	3,502
12	2,20	3,11	4,44	60	2,001	2,662	3,464
13	2,18	3,06	4,32	70	1,996	2,649	3,439
14	2,16	3,01	4,22	80	1,001	2,640	3,418
15	2,15	2,98	4,14	90	1,987	2,633	3,403
16	2,13	2,95	4,07	100	1,984	2,627	3,392
17	2,12	2,92	4,02	120	1,980	2,617	3,374
18	2,11	2,90	3,97	∞	1,960	2,576	3,291
19	2,10	2,88	3,92				

КРИТИЧЕСКИЕ ТОЧКИ РАСПРЕДЕЛЕНИЯ χ^2

Число степеней свободы k	Уровень значимости α					
	0,01	0,025	0,05	0,95	0,975	0,99
1	6,6	5,0	3,8	0,0039	0,00098	0,00016
2	9,2	7,4	6,0	0,103	0,051	0,020
3	11,3	9,4	7,8	0,352	0,216	0,115
4	13,3	11,1	9,5	0,711	0,484	0,297
5	15,1	12,8	11,1	1,15	0,831	0,554
6	16,8	14,4	12,6	1,64	1,24	0,872
7	18,5	16,0	14,1	2,17	1,69	1,24
8	20,1	17,5	15,5	2,73	2,18	1,65
9	21,7	19,0	16,9	3,33	2,70	2,09
10	23,2	20,5	18,3	3,94	3,25	2,56
11	24,7	21,9	19,7	4,57	3,82	3,05
12	26,2	23,3	21,0	5,23	4,40	3,57
13	27,7	24,7	22,4	5,89	5,01	4,11
14	29,1	26,1	23,7	6,57	5,63	4,66
15	30,6	27,5	25,0	7,26	6,26	5,23
16	32,0	28,8	26,3	7,96	6,91	5,81
17	33,4	30,2	27,6	8,67	7,56	6,41
18	34,8	31,5	28,9	9,39	8,23	7,01
19	36,2	32,9	30,1	10,1	8,91	7,63
20	37,6	34,2	31,4	10,9	9,59	8,26
21	38,9	35,5	32,7	11,6	10,3	8,90
22	40,3	36,8	33,9	12,3	11,0	9,54
23	41,6	38,1	35,2	13,1	11,7	10,2
24	43,0	39,4	36,4	13,8	12,4	10,9
25	44,3	40,6	37,7	14,6	13,1	11,05
26	45,6	41,9	38,9	15,4	13,8	12,2
27	47,0	43,2	40,1	16,2	14,6	12,9
28	48,3	44,5	41,3	16,9	15,3	13,6
29	49,6	45,7	42,6	17,7	16,0	14,3
30	50,9	47,0	43,8	18,5	16,8	15,0

Учебное издание

Хромых Анна Алексеевна

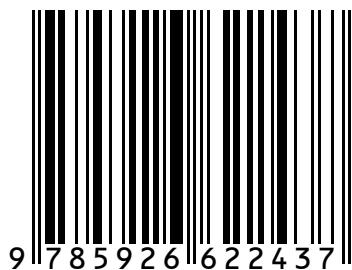
**СТАТИСТИЧЕСКИЕ МЕТОДЫ АНАЛИЗА ДАННЫХ
ДЛЯ ПОДГОТОВКИ СПЕЦИАЛИСТОВ ТЫЛОВОГО ОБЕСПЕЧЕНИЯ
ОРГАНОВ ВНУТРЕННИХ ДЕЛ**

Учебное пособие

В авторской редакции

Компьютерная верстка *Г. А. Артемовой*

ISBN 978-5-9266-2243-7



Подписано в печать 03.04.2026.

Авт. л. 3,6. Заказ 438.

Краснодарский университет МВД России
350005, г. Краснодар, ул. Ярославская, 128.