Григоренко В. В.
Назина Н. Б.

Статистические методы и модели управления

Учебно-методическое пособие

В. В. Григоренко, Н. Б. Назина

СТАТИСТИЧЕСКИЕ МЕТОДЫ И МОДЕЛИ УПРАВЛЕНИЯ

Учебно-методическое пособие

Электронное текстовое издание

Санкт-Петербург
Наукоемкие технологии
2026

© Григоренко В. В., Назина Н. Б., 2026
ISBN 978-5-00271-143-7

УДК 519.2(075.8)
ББК 22.172я73
Г83

Рецензенты:

Светлана Александровна Лысенкова,
Бюджетное учреждение высшего образования
Ханты-Мансийского автономного округа – Югры
«Сургутский государственный университет»;

Тарас Владимирович Гавриленко,
Бюджетное учреждение высшего образования
Ханты-Мансийского автономного округа – Югры
«Сургутский государственный университет»

Г83 Григоренко В. В., Назина Н. Б. Статистические методы и модели управления [Электронный ресурс]: учебно-методическое пособие / В. В. Григоренко, Н. Б. Назина – СПб.: Наукоемкие технологии, 2026. – 99 с. – URL: https://publishing.intelgr.com/archive/Statisticheskie-metodi-i-modeli-upravleniya.pdf.

ISBN 978-5-00271-143-7

Учебно-методическое пособие разработано в соответствии с ФГОС ВО по направлениям 09.03.02 и 09.04.02 «Информационные системы и технологии» и соответствующими рабочими программой дисциплины «Статистические методы и модели управления» для студентов бакалавриата и магистратуры всех форм обучения. Представленное учебно-методическое пособие актуально для подготовки научно-квалификационных работ по научной специальности Информационные технологии и телекоммуникации, в том числе 2.3.1. Системный анализ, управление и обработка информации, статистика.

В пособии приведены такие разделы дисциплины как: статистика как наука, ее предмет и история развития; основы статистических методов и первичная обработка результатов наблюдений; генеральная совокупность и выборка; описательные характеристики распределений; интервальные оценки и доверительные интервалы; статистические гипотезы и критерии их проверки; дисперсионный анализ; корреляционный анализ; регрессионный анализ. Особое внимание уделено формированию у обучающихся практических умений по выбору адекватных методов обработки и анализу данных, выполнению вычислений и содержательной интерпретации полученных результатов, в ходе освоения таких дисциплин как «Математический анализ», «Теория вероятностей и математическая статистика», «Методы оптимальных решений», «Математическое моделирование».

УДК 519.2(075.8)
ББК 22.172я73

ISBN 978-5-00271-143-7 © Григоренко В. В., Назина Н. Б., 2026

Учебное издание

Григоренко Виолетта Вячеславовна
Назина Нина Борисовна

Статистические методы и модели управления

Учебно-методическое пособие

Электронное текстовое издание

Издается в авторской редакции
Верстка *Ю. Н. Сафонкина*
Главный редактор *В. М. Коровин*

Подписано к использованию 15.06.2026.
Объем издания – 3,3 Мб.

Издательство «Наукоемкие технологии»
ООО «Корпорация «Интел Групп»
https://publishing.intelgr.com
E-mail: publishing@intelgr.com
Тел.: +7 (812) 945-50-63
Интернет-магазин издательства
https://shop.intelgr.com/

ISBN 978-5-00271-143-7



ОГЛАВЛЕНИЕ

Введение 5
1. Статистика как наука: предмет и история возникновения 6
2. Основы статистических методов 12
2.1. Основные понятия и методы статистического описания 12
2.2. Генеральная совокупность и выборка 13
2.3. Первичная обработка результатов эксперимента 15
2.4. Графическое представление рядов распределения 17
2.5. Оценка характеристик генеральной совокупности по выборке 19
2.6. Оценка характеристик положения 20
2.7. Оценка характеристик рассеяния 23
2.8. Оценка характеристик формы распределения 25
2.9. Эмпирическая функция распределения 25
2.10. Эмпирическая плотность распределения 26
2.11. Принципы статистического оценивания. Классификация оценок 27
3. Интервальные оценки. Понятие доверительного интервала 34
3.1. Интервальная оценка среднего при известном среднеквадратическом отклонении 36
3.2. Определение размера выборки 38
3.3. Интервальная оценка для доли (биномиального параметра) 39
3.4. Интервальная оценка генерального среднего при неизвестном среднеквадратическом отклонении 40
3.5. Интервальная оценка для дисперсии генеральной совокупности 42
4. Статистические гипотезы 48
4.1. Статистический критерий принятия решения. Проверка статистических гипотез 50
4.2. Использование критерия Стьюдента. Гипотеза о равенстве среднего значения числу 53
4.3. Использование критерия $\chi 2$ Пирсона. Гипотеза о соответствие эмпирического закона распределения нормальному 54
4.4. Использование критерия $\chi 2$ Пирсона. Гипотеза о равенстве средних двух выборок 56
4.5. Гипотеза о равенстве долей (биномиальных параметров) 57
5. Дисперсионный анализ 65
5.1. Однофакторный дисперсионный анализ для несвязанных выборок 65
5.2. Дисперсионный анализ для связанных выборок 68
6. Корреляционный анализ 74
6.1. Понятие корреляционной связи 74
6.2. Коэффициент корреляции Пирсона 76
6.3. Корреляция ранговых переменных 78
6.3.1. Нахождение коэффициента ранговой корреляции Спирмена 80
6.3.2. Нахождение коэффициента ранговой корреляции Кендалла 81
7. Регрессионный анализ 85
Список литературы 99

ВВЕДЕНИЕ

Настоящее учебно-методическое пособие «Статистические методы и модели управления» разработано для обеспечения учебного процесса по направлениям подготовки 09.03.02 «Информационные системы и технологии» (бакалавриат) и 09.04.02 «Информационные системы и технологии» (магистратура), также учебно-методическое пособие актуально для подготовки научно-квалификационных работ по научной специальности Информационные технологии и телекоммуникации, в том числе 2.3.1. Системный анализ, управление и обработка информации, статистика.

Структура пособия включает теоретический блок, содержащий справочный материал, определения, базовые утверждения и формулы, а также практический блок, в который входят примеры решения типовых задач, контрольные вопросы и задания для самостоятельного выполнения.

Целью пособия является формирование у студентов системы компетенций в области сбора, систематизации и анализа статистических данных, необходимых для решения задач системного анализа и управления техническими и социально-экономическими объектами. Методические материалы составлены таким образом, чтобы обеспечить последовательное освоение инструментария математической статистики: от основ первичной обработки данных и расчета описательных характеристик до построения сложных корреляционно-регрессионных моделей.

Содержание пособия охватывает основные разделы учебной дисциплины, предусмотренные образовательной программой: статистика как наука, ее предмет и история развития; основы статистических методов и первичная обработка результатов наблюдений; генеральная совокупность и выборка; описательные характеристики распределений; интервальные оценки и доверительные интервалы; статистические гипотезы и критерии их проверки; дисперсионный анализ; корреляционный анализ; регрессионный анализ. Особое внимание уделено формированию у обучающихся практических умений по выбору адекватных методов обработки данных, выполнению вычислений и содержательной интерпретации полученных результатов.

Актуальность пособия определяется возрастающей ролью статистических методов в задачах анализа, моделирования и управления сложными системами. Владение современным статистическим аппаратом необходимо будущим специалистам в области информационных систем и технологий для решения прикладных задач, связанных с обработкой экспериментальных данных, оценкой параметров, проверкой гипотез, выявлением зависимостей между признаками и построением прогностических моделей. Освоение этих методов способствует развитию аналитического мышления, исследовательской культуры и навыков принятия обоснованных решений на основе данных.

Издание предназначено для использования на аудиторных занятиях и в ходе самостоятельной подготовки студентов всех форм обучения, а также может быть полезно магистрантам и аспирантам при обработке результатов научно-исследовательских работ.

1. СТАТИСТИКА КАК НАУКА: ПРЕДМЕТ И ИСТОРИЯ ВОЗНИКНОВЕНИЯ

Статистика зародилась одновременно с государственностью как инструмент учета населения и ресурсов. Слово «статистика» имеет латинское происхождение и связано с термином «*status*», означающим «состояние дел». В научный оборот это понятие ввёл немецкий учёный Готфрид Ахенвалль в 1746 году, хотя теоретические основы заложил еще У. Петти в рамках «политической арифметики». Он предложил заменить название университетского курса «Государствоведение» на «Статистика», что стало началом формирования статистики как самостоятельной науки и учебной дисциплины. Однако статистический учёт возник значительно раньше. Ещё в Древнем Китае проводились переписи населения, в Древнем Риме учитывали имущество граждан, а также сравнивали военный потенциал различных государств.

Как современная общественная наука, статистика исследует количественную сторону массовых социально-экономических явлений в неразрывной связи с их качественным содержанием. Основной целью является выявление закономерностей общественного развития в конкретных условиях места и времени, учитывая влияние природных и технических факторов.

Статистика – отрасль знаний, в которой излагаются общие вопросы сбора, измерения и анализа массовых статистических (количественных или качественных) данных; изучение количественной стороны массовых общественных явлений в числовой форме.

Методология и этапы статистического исследования

Изучение объекта проходит через три последовательные стадии, формирующие статистическую методологию:

- **Статистическое наблюдение:** первичный сбор и регистрация данных о каждой единице совокупности.
- **Сводка и группировка:** систематизация собранной информации. На этом этапе данные делятся на однородные группы, подсчитываются итоги и вычисляются обобщающие показатели (средние и относительные величины).
- **Статистический анализ:** заключительный этап, на котором с помощью специальных методов (индексов, рядов динамики) выявляются скрытые закономерности, делаются выводы и составляются прогнозы, сопровождаемые графиками и таблицами.

Задачи и организация современной статистики в РФ

Главная задача статистики в России – обеспечение органов власти и общества достоверной информацией о социально-экономическом положении страны для принятия эффективных управленческих решений. В условиях рыночной экономики Россия перешла к системе национальных счетов (СНС) и активно использует выборочные методы наблюдения. Возглавляет эту деятельность Федеральная служба государственной статистики (Росстат), которая координирует единую систему органов от федерального до районного уровней, обеспечивая научно обоснованную методологию учета.

Развитие статистики в историческом контексте

История статистики связана с практическими нуждами государства. С древнейших времён она использовалась для учёта населения, скота, земли и имущества. В Древнем Китае ещё в V веке до н. э. проводились статистические работы, связанные с численностью населения, возрастным и половым составом, а также доходностью земли и торговлей.

На Руси учётно-статистические работы также появились очень рано. Сначала они носили имущественный характер и были связаны с налогообложением. Для этого использовались различные единицы обложения: земля, двор, «дым», орудия труда. Все сведения заносились в специальные книги – писцовые, дозорные, межевые и другие.

Термин «статистика» в научном смысле впервые употребил немецкий учёный Готфрид Ахенваль. В России развитие статистической науки связано с именами К. Ф. Германа, К. И. Арсеньева, Н. И. Надеждина, Д. А. Журавского и других исследователей, которые заложили основы политико-экономического направления.

Вклад отечественных и зарубежных учёных в развитие статистики

Большое значение в истории статистики имела земская статистика. Она внесла серьёзный вклад в совершенствование методов сбора и обработки данных, а также в развитие практической статистической работы.

В XIX веке статистические методы стали применяться не только в общественных, но и в естественных науках. Это привело к возникновению математической статистики. Существенный вклад в её развитие внесли Ф. Гальтон, К. Пирсон, Р. Фишер, В. Госсет (Стьюдент). Особое значение имеют работы Р. Фишера, который развил теорию статистических гипотез и дисперсионный анализ.

В России важнейшую роль в становлении математической статистики сыграли П. Л. Чебышёв, А. А. Марков и А. М. Ляпунов. Их труды стали основой русской школы теории вероятностей. В советский период значительный вклад в развитие статистической науки внесли С. Г. Струмилин, В. С. Немчинов, Б. С. Ястремский, А. Я. Боярский, Л. С. Козинец, Д. Ц. Урланис и другие учёные.

Статистические методы

Статистические методы – это способы анализа данных, полученных в ходе статистического исследования. К ним относят как методы прикладной статистики, которые можно использовать в самых разных сферах науки и хозяйственной деятельности, так и специальные статистические методы, применимые только в определенных областях. Например, это статистический приемочный контроль, статистическое регулирование технологических процессов, методы оценки надежности и проведения испытаний, а также планирование экспериментов.

Статистические методы анализа данных широко применяются практически во всех видах деятельности. Они необходимы в тех случаях, когда нужно сделать выводы о совокупности объектов или субъектов, обладающей внутренней неоднородностью.

В области статистических методов анализа данных можно выделить три основных направления научной и практической работы.

Первое связано с разработкой и изучением методов общего назначения, не зависящих от конкретной сферы применения.

Второе предполагает создание и исследование статистических моделей реальных процессов и явлений с учетом особенностей определенной области.

Третье направлено на применение статистических методов и моделей при анализе конкретных данных для решения практических задач, например, при проведении выборочных обследований.

Методы статистической обработки данных можно разделить на элементарные статистики, многомерные методы статистической обработки данных и data mining.

К элементарным статистикам относятся:

1. Описательные статистики (Среднее, мода, медиана, стандартное отклонение, дисперсия, размах, процентиль, доверительный интервал).

2. Параметрические и непараметрические критерии сравнения средних для зависимых и независимых выборок. Основная задача параметрических и непараметрических критериев, это во-первых проверить статистическую гипотезу, а во-вторых сравнить средние значения.

3. Таблицы сопряженности с критерием Хи-квадрат – позволяют определять связь между несколькими номинативными признаками, сравнить распределения между собой.

4. Корреляционный анализ – позволяет изучить положительные и отрицательные зависимости между переменными.

5. Однофакторный дисперсионный анализ используется для сравнения средних значений. Частный случай однофакторного дисперсионного анализа, это критерий Т-Стьюдента, отличие лишь в том, что в группирующей переменной дисперсионного анализа допускается не 2 градации фактора, а больше.

К многомерным методам относится очень большой класс статистических методов:

1. Многофакторный и многомерный дисперсионный анализ. В отличии от простого дисперсионного анализа, многофакторный дисперсионный анализ в качестве зависимой переменной может иметь несколько факторов, а значит и проверять больше гипотез.

Есть формула, которая определяет сколько гипотез может проверить многофакторный дисперсионный анализ: $2^p - 1$, p – это степень. Например, высчитаем сколько гипотез сможет проверить двухфаторный дисперсионный анализ $2^2 - 1 = 4 - 1 = 3$. Трёхфакторный дисперсионный анализ $2^3 - 1 = 8 - 1 = 7$. То есть 2 в кубе минус 1. Итого 7 гипотез может проверить за раз многофакторный дисперсионный анализ.

2. Множественный регрессионный анализ является предсказательным методом. Позволяет вычислить значение зависимой переменной зная лишь значения независимых переменных.

3. Факторный анализ. Факторный анализ объединяет в факторы те переменные, которые максимально связаны. Основное предназначение факторного анализа – это снижение размерности и изучение структуры интересующего нас явления.

4. Кластерный анализ. Похож на факторный анализ. Кластерный анализ объединяет в группы переменные или объекты, которые являются схожими по каким-либо параметрам.

5. Дискриминантный анализ. Похож на множественный регрессионный анализ. Только если в роли зависимой переменной в множественном регрессионном анализе является метрическая переменная, то в дискриминантном анализе, она измеряется в номинативной шкале, таким образом мы можем классифицировать какие-то количественные переменные по классам. Классический пример дискриминантного анализа это классификация по длине и ширине лепестков и чашелистиков ирисов к их сортам.

6. Многомерное шкалирование. Данный метод похож на факторный анализ, так как его основная задача – это выявление структуры множества объектов.

7. Продолжение многомерных методов является класс методов относящимся к data mining операциям. Основная их задача обнаруживать неизвестные данные, то есть «добыча» данных. Основаны на алгоритмах обучения.

8. Нейронные сети позволяют моделировать очень сложные зависимости. Основные их задачи – это классификация, прогнозирование.

9. Анализ временных рядов – основное назначение – предсказание будущего временного ряда, основываясь уже на имеющихся данных.

10. Деревья классификаций – очень удобный и наглядный метод для классификации групп, а также на основе этой классификации предсказание поведения потребителя или заемщика

11. Text mining – мощный метод для анализа текстовых массивов данных.

12. Machine Learning – это раздел искусственного интеллекта, позволяет строить модели способные обучаться. Классический пример Machine Learning это Байесовское оценивание.

Статистический анализ данных это огромный раздел прикладной математики, который нашел свое применение в разных областях научной деятельности, будь то точные науки или будь это гуманитарные науки.

Основная цель – на базе имеющихся данных выявлять закономерности. Казалось, как анализ статистических данных мог помогать психологам, задача которых общаться и консультировать людей. Но психология бывает не только в консультировании, есть также экспериментальная психология в которой исходя из названия проводятся эксперименты. В данном контексте психология является естественно-научной дисциплиной, а там, где есть эксперимент, там важно убедиться, что его результаты не случайны и будут воспроизведены на любой выборке из генеральной совокупности. Анализ статистических данных в психологии для этого и нужен. Вообще наибольший размах своей деятельности анализ статистических данных получил в области экономики, финансов и маркетинга. Анализ статистических данных используется для того, чтобы прогнозировать рост цен или поведения покупателя. Банки используют анализ статистических данных, чтобы, например, проверить кредитоспособность заемщика. Однако без математического бэкграунда анализировать статистические данные не представляется возможным, потому что в анализ статистических данных входит огромное количество методов математической статистики: От простых описательных статистик до data mining анализа. Соответственно и интерпретация результатов одних и тех же методов математической статистики разнится в зависимости от сферы исследования. Одного нажатия на кнопки программы для анализа статистических данных типа spss или statistica является недостаточным, необходимо понимать, что стоит за каждой кнопкой, а для этого надо глубоко понимать смысл каждого статистического метода. Вообще любой анализ статистических данных должен начинаться с определением шкалы измерения. Так как к разным шкалам измерения подходят разные методы. Например, нельзя сравнивать средние значения по переменным, относящимся к номинативной шкале. Средние значения сравниваются только для переменных измеренных в количественной или ранговых шкала. Причем для каждой из этих шкал свой критерий Т-Стьюдента можно использовать для количественных шкал, а сравнивать переменные измеренные в ранговых шкалах лучше с помощью критерия Манна-Уитни. Путаница в

шкалах может привести к неточным результатам. Наша практика показывает, что студенты часто путают шкалы. Хотя там ничего сложного нет.

Рассмотрим классические примеры шкал по Стивенсону:

Неметрические шкалы

1. Шкала наименований или номинативная группирует испытуемых на какие-то классы. Классический пример номинативной шкалы это пол. Обычно делают 2 градации: 1 – мужской; 2 – женский. Классы можно делить, как угодно. Например, можно группы людей разделить на 3 класса по предпочтению автомобилей: первые любят седаны, вторые - внедорожники и кроссоверы, третьи – спорткары. И так далее. Сколько исследователю нужно классов настолько он может разбить любые объекты.
2. Ранговая шкала (ординальная или порядковая). Основная характеристика в том, что она показывает степень выраженности признака. Например, самый популярный маршрут для путешествий до совсем непопулярный. Или ранг человек: есть рядовой, а есть генерал. Есть топ менеджеры, а есть младшие менеджеры. И так далее.

Метрические шкалы

1. Интервальная, она же метрическая, она же количественная**.** В этой шкале числа отражают какой признак, у кого он больше и насколько больше. В данной шкале ноль не является показателем отсутствия признака. Например, температура по Цельсию.
2. Абсолютная шкала – по сути тоже метрическая, только в отличии от интервальной, ноль является показателем отсутствия признака. Например, не может быть роста в 0 см или веса в 0 кг. На практике не всегда легко отличить ранговую шкалу от количественной.

Анализ статистических данных требует сугубо научного подхода. Ошибка может стоить компаниям много денег, а ошибка в анализе статистических данных в медицине может привезти к необратимым последствиям.

Связь статистики с другими дисциплинами

Статистика носит мультидисциплинарный характер, поскольку ее методологический аппарат тесно переплетен с целым рядом других научных областей.

В качестве фундаментальной теоретической основы статистики выступают экономическая теория и социология. Именно эти дисциплины позволяют понять природу и законы развития общественных процессов, в то время как статистика предоставляет конкретный инструментарий для цифровой оценки их масштабов и структурного анализа.

Неразрывная связь существует также между статистикой и математикой. Математический аппарат является необходимым средством для проведения расчетов, проверки гипотез и выявления скрытых закономерностей. Кроме того, методы математического моделирования и визуализации позволяют систематизировать полученные результаты и представлять их в наглядной форме в виде графических схем и аналитических таблиц.

КОНТРОЛЬНЫЕ ВОПРОСЫ

1. Что изучает статистика как наука?
2. Каково происхождение термина «статистика»?
3. Кто ввёл понятие «статистика» в научный оборот?
4. Какой вклад в развитие статистики внёс У. Петти?
5. В каких древних государствах появились первые статистические учёты?
6. Какие основные этапы включает статистическое исследование?
7. Что такое статистическое наблюдение?
8. Что понимают под сводкой и группировкой статистических данных?
9. В чём заключается статистический анализ?
10. Какова главная задача статистики в современной России?
11. Какова роль Росстата в системе государственной статистики?
12. Какие основные направления развития статистики выделяются в истории науки?
13. Чем отличаются элементарные и многомерные статистические методы?
14. Какие шкалы измерения используются в статистике?
15. С какими научными дисциплинами статистика тесно связана?

2. ОСНОВЫ СТАТИСТИЧЕСКИХ МЕТОДОВ

2.1. Основные понятия и методы статистического описания

Типы статистических данных

Статистические данные представляют собой наблюдаемые или измеряемые значения одного или нескольких признаков обследуемой совокупности объектов.

Различают **количественные и качественные признаки**. Значения количественных признаков могут быть непрерывными (вес, суточная производительность, годовой объем производства) или дискретными (количество детей, год выпуска продукции, количество холодильников, проданных в течение дня).

Примерами **качественных признаков** являются, например, пол, семейное положение, цвет кожи, качество товара. В свою очередь качественные признаки в зависимости от вида данных делятся на **номинальные** (классификационные) и **ординальные** (порядковые). Говорят, также, что соответствующие качественные признаки измеряются в номинальной или порядковой шкале. Разница между этими шкалами состоит в следующем.

Признак, измеряемый **в номинальной шкале**, принимает одно значение из конечного числа заведомо установленных градаций. Примерами признаков, измеряемых в номинальной шкале, являются пол (мужской, женский), цвет, марка автомобиля, тип строительного материала, классификация животных и т. п. Статистические данные, измеряемые в номинальных шкалах, представляются в виде таблиц, в которых приводятся частоты появления той или иной градации признака. Часто номинальные данные появляются при обработке социологических опросов. Например, может представлять интерес вопрос о предпочтительности той или иной группы избирателей к каждому из кандидатов на пост президента России. В этом случае результаты опроса могут быть представлены в виде прямоугольной таблицы, содержащей m строк (по числу кандидатов) и n столбцов (по числу групп избирателей), в каждой клетке которой записывается число избирателей данной группы, отдавших предпочтение соответствующему кандидату. Таблицы таких данных называются **таблицами сопряженности** размера m x n.

Значения качественных признаков, измеряемых **в ординальной шкале**, могут быть упорядочены. Примерами таких признаков являются тестовые баллы и школьные оценки, качество условий жизни (плохое, удовлетворительное, хорошее, очень хорошее), сила ветра, оцениваемая по шкале Бофорта (штиль, слабый ветер, умеренный ветер, свежий ветер, шторм и т. д.). Для признаков, измеряемых в ординальных шкалах, операции сложения и вычитания не имеют смысла. Так, нельзя сказать, что студент, получивший на экзамене «пять» по статистике знает предмет на одну единицу лучше, чем студент, получивший по этому предмету «четыре», поскольку для знаний не существует единицы измерения. Однако можно сказать, что первый студент знает статистику лучше, чем второй.

Для представления значений ординальных признаков в числовой форме используется следующий способ. Все значения признака записываются в порядке возрастания в виде ряда. Каждому значению поставим в соответствие натуральное число, равное его номеру в ряду. Это число называется **рангом**. Например, качество условий жизни (плохое, удовлетворительное, хорошее, очень хорошее) будет представлено рангами 1,2, 3, 4. Для ординальных признаков, представленных в виде рангов, разработаны специальные статистические методы,

позволяющие измерять степень близости признаков (например, ранговая корреляция), проверять гипотезы о виде распределения, проводить дисперсионный анализ.

Для данных, представленных в номинальной шкале, также не определены операции сложения и вычитания. Эти данные (в отличие от ординальных признаков) не могут быть упорядочены и, следовательно, оцифрованы с помощью рангов. Применяя специальные статистические методы для номинальных признаков, можно проверить гипотезы о независимости признаков и о принадлежности двух или нескольких выборок к одному виду. Для оцифровки номинальных признаков используются числовые метки. Выбор меток в зависимости от цели статистического анализа может проводиться по различным критериям.

2.2. Генеральная совокупность и выборка

Множество всех обследуемых объектов называется **генеральной совокупностью**. Если это множество содержит небольшое число элементов, то возможно полное обследование всех его элементов.

Однако в большинстве случаев в силу того, что генеральная совокупность имеет очень много элементов либо ее элементы труднодоступны, либо по другим причинам обследуется некоторая часть генеральной совокупности – **выборка**.

В этом случае основные характеристики генеральной совокупности (их называют статистиками: среднее, дисперсия, среднеквадратическое отклонение и т. д.) **оцениваются** (**т. е. определяются приближенно**) по выборке.

Соответствующие статистики называются «**выборочное среднее**», «**выборочная дисперсия**» и т. д. Очевидно, что не всякая выборка правильно отражает свойства генеральной совокупности. Например, нельзя судить о среднем душевом доходе населения по выборке, составленной из доходов служащих финансовых компаний. Выборка должна давать правильное, неискаженное представление о генеральной совокупности, или, как говорят, быть **репрезентативной**. Если свойства генеральной совокупности заранее неизвестны, то, за неимением лучшего, следует использовать **простой случайный выбор**. Это означает, что все элементы генеральной совокупности должны иметь равные шансы попасть в выборку.

Например, при выяснении мнения всех студентов университета по какому-либо вопросу, выборка, составленная из студентов первого курса, не будет репрезентативной.

Процедуру случайного выбора можно организовать, например, так. Запишем фамилии всех студентов на отдельные карточки, которые затем тщательно перетасуем, и из всего множества карточек отберем нужное количество. Ответы выбранных таким способом студентов составят репрезентативную выборку. Если требуется, чтобы в выборке были представлены элементы различных групп, составляющих генеральную совокупность, используется процедура **типического отбора**. Так, если студенты первого курса составляют 15% всех студентов университета, то и в выборке они должны составлять 15%. В некоторых случаях необходимо учитывать не только курс, но и специализацию студентов, если это может повлиять на результаты опроса.

Как правило, статистические данные в силу ошибок измерений, влияния внешней среды, присущей индивидуумам случайной изменчивости, имеют разброс. Рассмотрим, например, результаты выборочного контроля партии расфасованной продукции. С большой вероятностью можно сказать, что в выборке не найдется ни одной пары пакетов, имеющих один и тот же вес. **Основная задача статистики состоит в получении осмысленных**

заключений именно из такого типа данных, т. е. данных, подверженных случайной изменчивости.

Математическая модель статистических данных содержит детерминированную и случайную составляющие. В простейшей модели случайная компонента – это случайная величина. Например, математическая модель данных, представляющих вес расфасованной продукции, есть сумма двух величин: номинального веса m пакета (детерминированная компонента) и отклонения истинного веса пакета от номинального. Это отклонение (имеющее случайный характер) можно рассматривать как сумму очень большого числа случайных факторов, всегда имеющих место в производственных условиях (износ оборудования, влажность и температура продукта и др.). Следовательно, в силу центральной предельной теоремы отклонение от номинального веса – это случайная величина, имеющая нормальное распределение $N(0, \sigma^2)$ с нулевым математическим ожиданием (m) и некоторой дисперсией σ^2.

В связи с этим генеральная совокупность определяется как множество значений случайной величины X, представляющей модель данных. В рассмотренном примере это бесконечная генеральная совокупность, имеющая нормальное распределение $N(m, \sigma^2)$.

Выборка определяется следующим образом. Пусть случайная величина X в случайном эксперименте E наблюдается n раз в предположении, что условия проведения эксперимента, а следовательно, и распределение наблюдаемой случайной величины X не изменяются от эксперимента к эксперименту. Этот новый составной эксперимент связан с n-мерной случайной величиной – случайным вектором $(X_1, X_2, \dots, X_n)$, где X_j – случайная величина, соответствующая j-му эксперименту. Очевидно, $X_j, j = 1, \dots, n$ – независимые в совокупности случайные величины, каждая из которых имеет тот же закон распределения, что и случайная величина X.

Распределением генеральной совокупности называется закон распределения случайной величины X, а случайный вектор $(X_1, X_2, \dots, X_n)$ – **выборочным вектором**. Числа $x_1, x_2, \dots, x_n$, получаемые на практике при n-кратном повторении эксперимента E в неизменных условиях, представляют собой реализацию выборочного вектора $(X_1, X_2, \dots, X_n)$, и называются **выборкой** $x_1, x_2, \dots, x_n$ **объема** n.

Выборку $x_1, x_2, \dots, x_n$ при необходимости можно рассматривать как точку выборочного пространства, т. е. множества, на котором задано распределение вектора $(X_1, X_2, \dots, X_n)$.

Аналогично определяется выборка, когда случайный эксперимент E связан с несколькими случайными величинами. Например, выборка объема n из двумерной генеральной совокупности есть последовательность $(x_1, y_1), \dots, (x_n, y_n)$ пар значений случайных величин X и Y, принимаемых ими при n независимых повторениях случайного эксперимента E.

Таким образом, в математической статистике выборка объема n понимается двояко: как последовательность n чисел, являющихся результатами наблюдения случайной величины X: $x_1, x_2, \dots, x_n$ и как выборочный вектор, т. е. совокупность n независимых случайных величин $(X_1, X_2, \dots, X_n)$, каждая из которых имеет одно и то же распределение, совпадающее с распределением наблюдаемой случайной величины X. При этом выборка как совокупность чисел $x_1, x_2, \dots, x_n$ есть некоторая возможная реализация выборочного вектора $(X_1, X_2, \dots, X_n)$.

В каком именно смысле используется понятие выборки в том или ином месте, должно быть ясно, во-первых, из обозначений: числа обозначаются строчными буквами, а случайные величины – прописными, а во-вторых, из контекста. Понятие выборочного вектора является

основополагающим в математической теории статистики: теории статистического оценивания, при построении доверительных интервалов, проверке статистических гипотез.

2.3. Первичная обработка результатов эксперимента

Ряды распределения

Явления, происходящие в природе, обществе очень сложны и разнообразны. Ученые изучают разные стороны этих явлений, причем каждая наука вырабатывает свои специфические методы исследования.

Для того, чтобы выработать методику тренировок с учетом индивидуальных особенностей спортсмена, необходимо проводить наблюдения за его физиологическими характеристиками, достигнутыми результатами и т.д.

Результаты наблюдений представляют в виде **таблиц, графиков, диаграмм** и **различных числовых характеристик**.

Подготовка таких материалов начинается с построения **рядов распределения** числовых данных.

Различают **вариационный (ранжированный**) и **статистический ряд**, который представляют в двух видах: статистический **дискретный** и статистический **интервальный ряды**.

Вариационный ряд – это ряд числовых значений исследуемой величины, расположенных в порядке возрастания.

$$x_1, x_2, \dots \dots \dots \dots \dots, x_n$$

$$x_1 \le x_2 \le \ \dots \dots \dots \dots \dots \le x_n$$

Статистический дискретный ряд – это таблица

x_i	x_1	…	…	…	x_k
n_i – частота	n_1	…	…	…	n_k

Статистический интервальный ряд – это таблица, где в первой строке приводятся границы интервала, n_i – частота i-того интервала, т. е. количество значений величины x, попавших в этот интервал.

x_{i-1}, x_i	$x_0\ , x_1$	x_1, x_2	…	…	x_{k-1}, x_k
n_i – частота	n_1	n_2	…	…	n_m

При построении интервального ряда количество интервалов определяют по **правилу Стерджесса:**

$$k = 1 + 3{,}31 \lg n, \tag{1}$$

где n – число данных в выборке:

$$n = \sum_{i=1}^{k} n_i, \tag{2}$$

Размах выборки:

$$R = x_{max} - x_{min}, \tag{3}$$

где x_{max} – значение максимального элемента выборки; x_{min} – значение минимального элемента выборки.

Наряду с частотами подсчитывают также **накопленные частоты** $\sum_{i=1}^{k} n_i$, относительные частоты $\frac{n_i}{n}$, накопленные относительные частоты $\sum_{i=1}^{k} \frac{n_i}{n}$.

Пример 1. Представить выборку из 55 наблюдений (табл. 1) в виде таблицы частот, рассчитать количество интервалов группировки, размах выборки, длину интервала группировки.

Таблица 1

Выборка наблюдений

18,3	15,4	17,2	19,2	23,3	18,1	21,9
15,3	16,8	13,2	20,4	16,5	19,7	20,5
14,3	20,1	16,8	14,7	20,8	19,5	20,3
19,3	17,8	16,2	15,7	22,8	21,9	12,5
10,1	21,1	18,3	20,2	23,8	16,7	20,4
13,9	19,1	18,5	20,2	23,8	16,7	20,4
19,5	17,2	19,6	17,8	21,3	17,5	19,4
17,8	13,5	17,8	11,8	18,6	19,1	

Решение:

1. Расчет количества интервалов группировки:

$$k = 1 + 3{,}31 \lg n = 1 + 3,31 * \lg 55 \approx 7$$

2. Расчет размаха выборки:

$$R = x_{max} - x_{min} = 23{,}8 + 10{,}1 = 13{,}7$$

3. Расчет длины интервала группировки:

$$b = \frac{R}{k} = 13{,}7/7 \approx 2$$

4. В качестве первого интервала удобно взять интервал 10–12. Результаты группировки сведены в таблице 2.

Таблица 2

Таблица частот для примера 1

Номер интервала	Границы интервала	Середина интервала, x_i	Частота, n_i	Накопленная частота, $\sum_{i=1}^{k} n_i$	Относительная частота, $\frac{n_i}{n}$	Накопленная относительная частота, $\sum_{i=1}^{k} \frac{n_i}{n}$
1	10-12	11	2	2	2/55=0,0364	0,0364
2	12-14	13	5	7	5/55=0,0909	0,1273
3	14-16	15	6	13	0,1019	0,2364
4	16-18	17	12	25	0,2182	0,4545
5	18-20	19	13	38	0,2364	0,6909

Номер интервала	Границы интервала	Середина интервала, x_i	Частота, n_i	Накопленная частота, $\sum_{i=1}^{k} n_i$	Относительная частота, $\frac{n_i}{n}$	Накопленная относительная частота, $\sum_{i=1}^{k} \frac{n_i}{n}$
6	20-22	21	13	51	0,2364	0,9273
7	22-24	23	4	55	0,0727	1 ,0000

2.4. Графическое представление рядов распределения

Интервальный ряд изображается столбиковой диаграммой. Основание столбиков, расположенные на оси абсцисс, являются интервалами изучаемой величины, а высоты – частотами. Такая диаграмма называется **гистограммой**. Площадь ступенчатой фигуры над графиком гистограммы равна объему выборки n. Так как значения гистограммы пропорциональны самим частотам, обычно по оси ординат откладывают значения частот n_i, а не значения $\frac{n_i}{b}$.

Аналогично определяется гистограмма относительных частот. Площадь соответствующей фигуры для нее равна единице.

При увеличении объема выборки и уменьшении интервала группировки гистограмма относительных частот становится статистическим аналогом плотности распределения $f(x)$ генеральной совокупности.

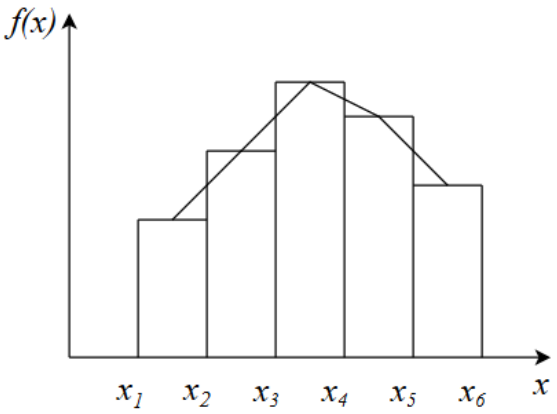


Рис. 1. Гистограмма к примеру 1

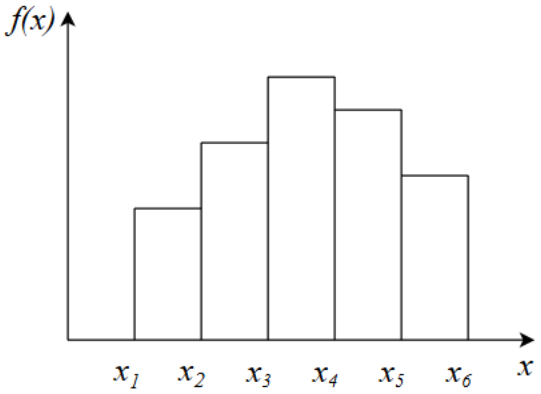


Рис. 2. Полигон распределения

Для дискретного и интервального рядов можно построить **полигон распределения** (Рис. 2), который представляет собой ломаную линию, соединяющую середины интервалов x_i и частоты $\frac{n_i}{n}$.

Распределение признака в вариационном ряду по накопленным частотам называется **кумулятой** (или **огивой**), то есть на координатную плоскость необходимо нанести накопленные частоты. Вершины ее имеют абсциссы (значение признака), совпадающие с правыми границами интервалов группировки, и ординаты, совпадающие со значениями накопленных частот для соответствующих интервалов (или относительных частот в процентах).

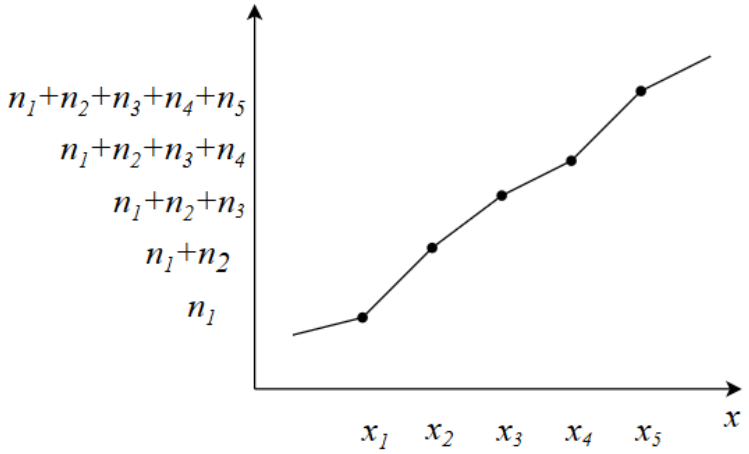


Рис. 3. Огива (кумулята)

Пример 2. Для предыдущего примера построить гистограмму, полигон и огиву.

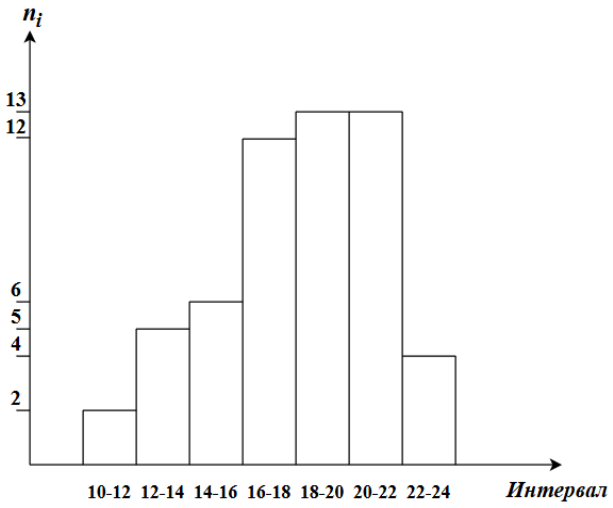


Рис. 4. Гистограмма к примеру 1

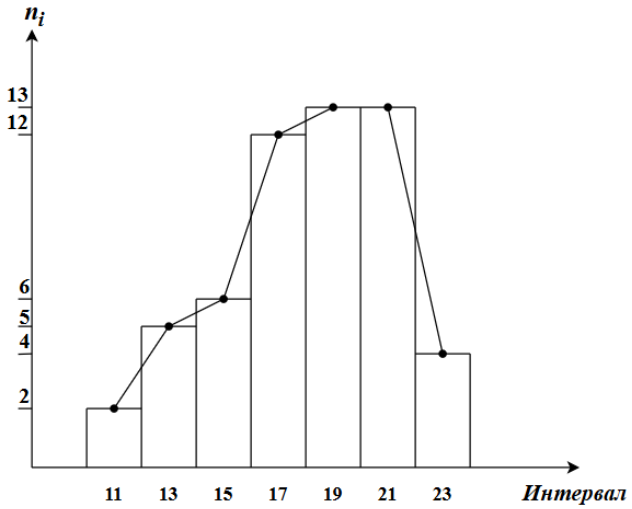


Рис. 5. Полигон частот к примеру 1

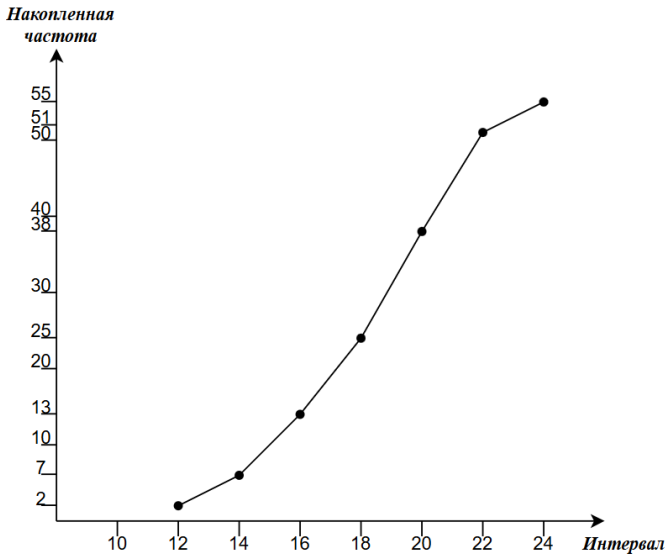


Рис. 6. Огива для примера 1

Кривая на рис.6 является статистической функцией распределения $F^*(x)$, которая приближенно представляет теоретическую функцию распределения $F(x)$ или, как говорит статистика, является оценкой для $F(x)$. Статистическая функция распределения $F^*(x)$ полностью решает задачу описания опытных данных.

Оценку для теоретической плотности вероятности $f^*(x)$ – статистическую **плотность вероятности** – можно найти по известной функции $F^*(x)$ следующим образом:

1. Аппроксимировать $F^*(x)$ аналитически, а затем аппроксимированную функцию продифференцировать.
2. Выполнить численное дифференцирование $F^*(x)$ с достаточно большим шагом.

Рассмотренные приемы обработки опытных данных используются при n порядка нескольких десятков. При больших выборках (n – несколько сотен) чаще используется другой метод обработки и вместо статистической функции распределения $F^*(x)$ находится статистическая плотность вероятности $f^*(x)$.

2.5. Оценка характеристик генеральной совокупности по выборке

Представление данных в виде гистограмм и полигонов дает информацию о распределении генеральной совокупности. Соответствующие гипотезы проверяются по статистическим критериям. Однако часто требуется охарактеризовать генеральную совокупность некоторыми количественными показателями, которые определяют положение центра распределения, рассеяние (разброс) и асимметрию, что дает возможность сравнить одну совокупность данных с другой. По выборке можно определить приближенные значения (оценки) этих числовых характеристик, которые называются **выборочными характеристиками**.

Пусть $x_1, x_2, \dots, x_n$ – выборка объема n из генеральной совокупности.

Выборочный **начальный момент** *r*-того порядка определяется как соответствующий момент выборочного распределения и вычисляется по формуле:

$$a_r^* = \frac{1}{n}\sum_{i=1}^{N} x_i^r, r = 1,2, \dots l. \quad (4)$$

Выборочный начальный момент первого порядка обычно называют средним арифметическим и обозначают $\bar{x}$:

$$\bar{x} = a_1^* = \frac{1}{n}\sum_{i=1}^{n} x_i. \quad (5)$$

Среднее арифметическое выборки является в определенном смысле «хорошей» оценкой математического ожидания m генеральной совокупности для симметричных распределений.

Выборочный центральный момент *r*-того порядка определяется по формуле:

$$\mu_r^* = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^r. \quad (6)$$

Если выборка $x_1, x_2, \dots, x_n$ представлена в виде статистического ряда, причем частота появления элемента $x_i, i = 1, 2, \dots k$ в выборке равна n_i, а $\sum_{i=1}^{n} n_i = n$, то выборочные начальные и центральные моменты *r*-того порядка вычисляются по формулам:

$$a_r^i = \frac{1}{n}\sum_{i=1}^{r} n_i x_i^r , \quad (7)$$

$$\mu_r^* = \frac{1}{n}\sum_{i=1}^{n} n_i (x_i - \bar{x})^r, r = 1, 2, \dots l. \quad (8)$$

Для выборки большого объема представленной в виде **таблицы частот группированной** выборки (см. пример 1, табл. 1), выборочные начальные и центральные моменты рассчитываются по формулам (7) и (8), в которых в качестве $x_i, i = 1, 2, \dots k$ принимается среднее значение i-го интервала группировки, а n_i – частота попадания в i-тый интервал.

Начальные выборочные моменты a_r^* связаны с центральными выборочными моментами μ_r^*, следующими соотношениями:

$$\mu_2^* = a_2^* - \bar{x}^2$$

$$\mu_3^* = a_3^* - 3a_2^*\bar{x} + \bar{x}^3$$

$$\mu_4^* = a_4^* - 4a_3^*\bar{x} + 6a_2^*\bar{x}^2 - 3\bar{x}^4 \text{ и т. д.}$$

2.6. Оценка характеристик положения

Наиболее распространенными оценками характеристик положения являются **среднее арифметическое выборки (выборочное среднее), выборочная медиана** и **выборочная мода**. В дальнейшем будем опускать термин «выборочная», имея, однако, в виду, что любая оценка, вычисляемая по выборке, является всего лишь приближенным значением соответствующей характеристики генеральной совокупности.

Для выборки объема n: $x_1, x_2, \dots, x_n$ **среднее арифметическое** равно:

$$x_{\text{ср}} = \frac{\sum_{i=1}^{n} x_i}{n}. \quad (9)$$

Если выборка представлена в виде статистического ряда, где n, есть частота элемента x_i , то среднее арифметическое вычисляется по формуле:

$$x_{\text{ср}} = \frac{\sum_{i=1}^{n} x_i * n_i}{\sum_{i=1}^{n} n_i}. \quad (10)$$

По этой же формуле вычисляется среднее арифметическое и для сгруппированной выборки (интервального ряда), причем в качестве $\tilde{x}_i$ берется середина интервала:

$$x_{\text{ср}} = \frac{\sum_{i=1}^{n} \tilde{x}_i * n_i}{\sum_{i=1}^{n} n_i}, \quad (11)$$

где $\tilde{x}_i$ – середина i-того интервала.

Последняя формула может применяться как в случаях, когда в каждом интервале содержится по крайней мере несколько десятков реализаций.

Модой d называется элемент выборки, имеющий **наибольшую частоту**.

Пусть $x^{(1)}, x^{(2)}, \dots, x^{(n)}$ – вариационный ряд выборки.

Медианой называется число h, которое **делит вариационный ряд на две части**, содержащие равное количество элементов. Если объем выборки n – нечетное число ($n = 2 * k + 1$), то $h = x^{k+1}$, т. е, h является **средним** элементом вариационного ряда. Если же $n = 2k$, то $h = \frac{1}{2}(x^{(k)} + x^{(k+1)}$.

Пример 3. Определить среднее, моду и медиану для выборки 5,6,8,2,3,1,1,4.

Решение. Представим данные в виде вариационного ряда 1,1,2,3,4,5,6,8.

Среднее значение ряда:

$$\bar{x} = \frac{1}{8}(1 + 1 + 2 + 3 + 4 + 5 + 6 + 8) = 3{,}75.$$

Все элементы входят в выборку по одному разу, кроме 1, следовательно:
Мода равна:

$$d = 1.$$

Так как $n = 8$, то медиана:

$$h = \frac{1}{2}(3 + 4) = 3{,}5.$$

Мода предоставляет важную информацию для производителей товаров, конструкторов, работников торговли. Например, производитель часов должен знать, по какой цене распродается основная часть его продукции, в магазине бытовой техники должны знать наиболее популярные марки товара и т. д.

Соотношение между средним, модой и медианой в зависимости от формы функции плотности распределения показывают рис. 7, 8, 9.

Соотношение между средним арифметическим, медианой и модой в зависимости от формы функции плотности распределения (рис. 7, 8, 9) показывают **асимметрию (скошенность) распределения**. Они описывают центр данных, но реагируют на форму распределения по-разному:

- Симметричное распределение (например, нормальное) – среднее, медиана и мода практически совпадают и находятся в центре (Рис. 7):

$$\text{Среднее арифметическое } (\bar{x}) = \text{Медиана}(h) = \text{Мода}(d), \qquad (12)$$

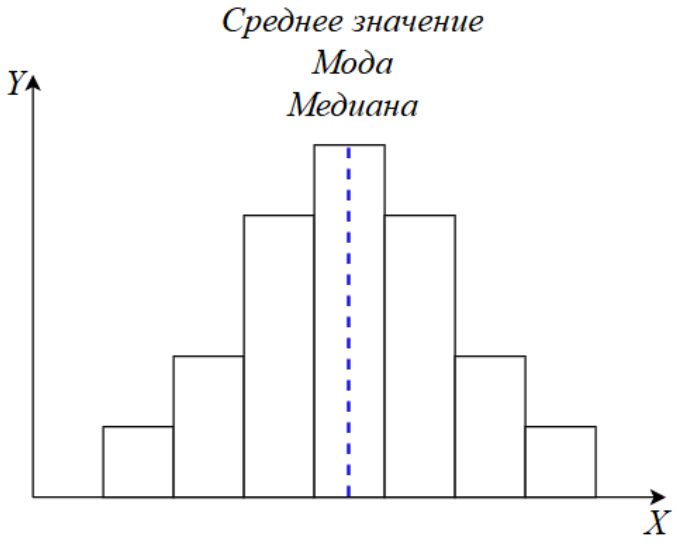


Рис. 7. Симметричное унимодальное распределение

- Правосторонняя асимметрия (положительная, «хвост» справа) – среднее значение чувствительно к экстремально большим значениям (выбросам), поэтому оно смещается вправо сильнее, чем медиана (Рис. 8):

$$\text{Мода}(d) < \text{Медиана}(h) < \text{Среднее арифметическое } (\bar{x}), \qquad (13)$$

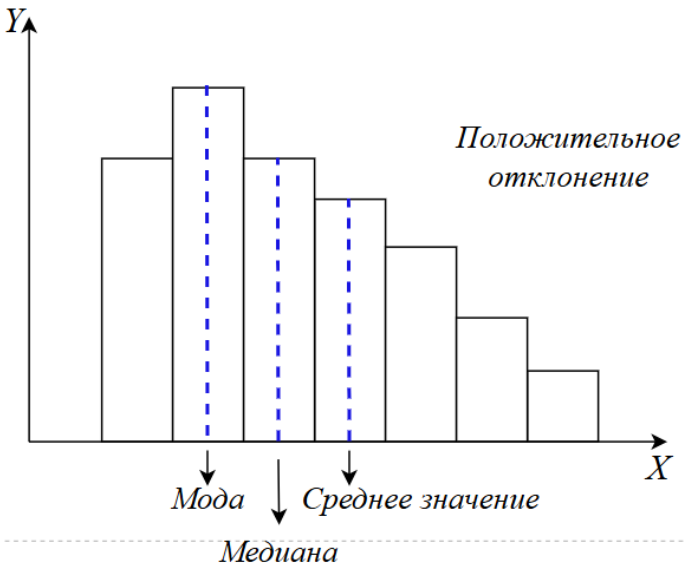


Рис. 8. Скошенное справа распределение (асимметрия > 0)

- Левосторонняя асимметрия (отрицательная, «хвост» слева) – среднее значение смещается влево под влиянием экстремально малых значений (Рис. 9):

$$\text{Среднее арифметическое } (\bar{x}) < \text{Медиана}(h) < \text{Мода}(d), \tag{14}$$

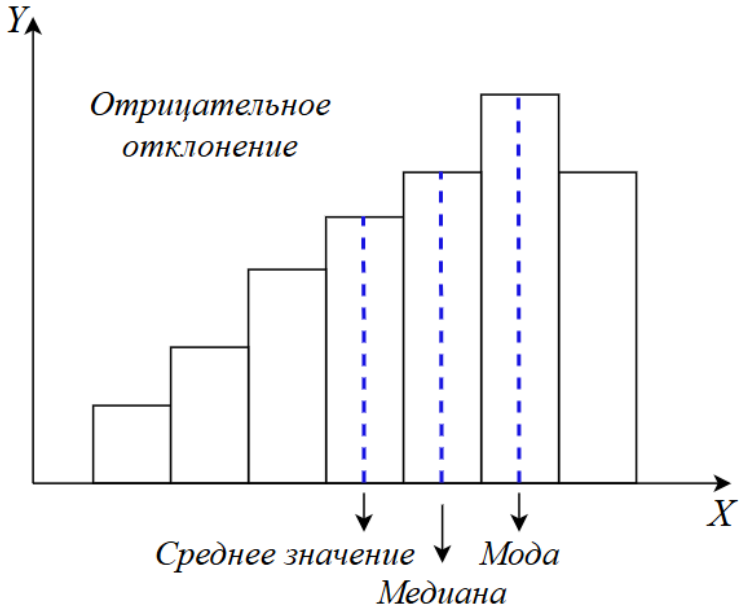


Рис. 9. Скошенное слева распределение (асимметрия < 0)

Среднее геометрическое. Пусть $x_1, x_2, \dots, x_n$ – положительные числа. Среднее геометрическое вычисляется по формуле:

$$x_g = \sqrt[n]{x_1 * x_2 * \dots * x_n}, \tag{15}$$

Среднее геометрическое применяется для характеристики данных, заданных через равные промежутки времени, и определяет среднюю долю относительных изменений. Такие

данные могут характеризовать явления роста, например, доходы по вкладам, эксплуатационные расходы, прирост населения и т. д.

Между средним геометрическим и средним арифметическим имеет место следующее соотношение:

$$\sqrt[n]{x_1 * x_2 * \dots * x_n} \leq \frac{x_1+x_2+\dots+x_n}{n}, \quad (16)$$

Причем знак равенства имеет место при $x_1 = x_2 = \dots = x_n$.

Логарифм среднего геометрического вычисляется по формуле:

$$lgx_g = \frac{\sum_{i=1}^{n} lgx_i}{n}, \quad (17)$$

Среднее гармоническое x_H вычисляется по формуле 18:

$$\frac{1}{x_H} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{x_i}, \quad (18)$$

Среднее арифметическое обладает следующими свойствами.

1. Сумма разностей наблюдаемых значений и среднего арифметического равна нулю:

$$\sum(x_i - x_{\text{ср}}) = (x_1 - x_{\text{ср}}) + (x_2 - x_{\text{ср}}) + \dots + (x_n - x_{\text{ср}}) = \sum x_i - nx_{\text{ср}} = \sum x_i - \frac{n\sum x_i}{n} = 0, \quad (19)$$

2. Если каждое наблюдаемое значение умножить на одно и тоже число, то среднее умножится на это число:

$$\frac{\sum x_i * C}{n} = \frac{C * \sum x_i}{n} = C * x_{\text{ср}}. \quad (20)$$

Среднее гармоническое используется для характеристики данных, размерность которых выражена отношением различных физических величин: км/ч; л/на 100 км и т. д. Между средним арифметическим, средним гармоническим и средним геометрическим существует следующее соотношение:

$$x_{\text{ср}} \geq x_g \geq x_H$$

2.7. Оценка характеристик рассеяния

Наиболее распространенными мерами рассеяния являются размах, средний межквартилъный размах, персентильный размах, дисперсия и среднее квадратическое отклонение.

Размах определяется как разность между максимальным и минимальным значениями выборки:

$$R = x_{max} - x_{min} = x^{(n)} - x^{(1)}. \quad (21)$$

Пусть $x^{(1)}, x^{(2)}, \dots, x^{(n)}$ – вариационный ряд выборки. Вариационный ряд делится тремя квартилями (Q_1, Q_2, Q_3) на 4 равные части, то есть Q_1 – это значение, ниже которого лежит 25% наблюдений, Q_2 – 50% наблюдений, (Q_2 равен медиане), Q_3 – 75% наблюдений.

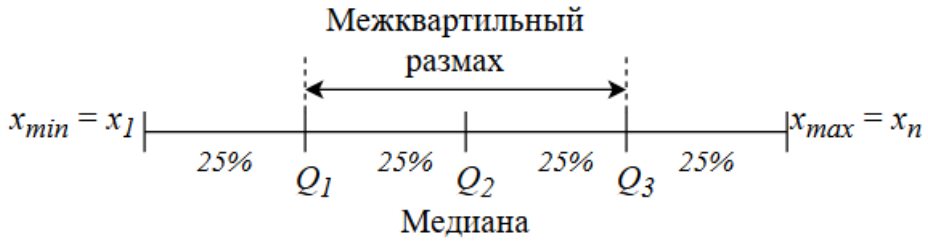


Рис. 10. Межквартильный размах

Средний межквартильный размах равен половине разности верхнего и нижнего квартилей:

$$\frac{Q_3 - Q_1}{2}. \tag{22}$$

Персентильный размах равен разности 90- и 10-го персентилей:

$$P_{90} - P_{10} = x_{0,9} - x_{0,1}. \tag{23}$$

Для вычисления характеристик генеральной совокупности по выборке также используют **выборочные квантили** различных порядков.

Выборочный квантиль x_p порядка p, $1 < p < 0$, определяется как элемент вариационного ряда выборки $x_1, x_2, \dots, x_n$ с номером $[ap] + 1$, где $[a]$ целая часть a.

В статистической практике используют ряд квантилей, имеющих специальные названия:

Квартили: Q_1, Q_2, Q_3 – квантили порядков 0,25; 0,50; 0,75;

Децили: $D_1, D_2, \dots, D_9$ – квантили порядков 0,10; 0,20; ...; 0,90;

Персентили: $P_1, P_2, \dots, P_{99}$ – квантили порядков 0,01; 0,02; ...; 0,99.

Для сгруппированных выборок оценки выборочных квантилей можно определить по *огиве* (см. пример 1 и рис. 6.).

Дисперсия для выборки негруппированных данных определяется по формуле:

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}. \tag{24}$$

Для сгруппированных данных – по формуле (25):

$$\sigma^2 = \frac{\sum_{i=1}^{n} n_i(\tilde{x}_i - \bar{x})^2}{n-1} = \frac{\sum_{i=1}^{n} n_i(\tilde{x}_i - \bar{x})^2}{\sum n_i - 1}, \tag{25}$$

где n_i – частота; $\tilde{x}_i$ – середина интервала группировки; $\bar{x}$ – среднее значение выборки; n – количество элементов выборки.

Оценку дисперсии можно вычислить, используя второй выборочный центральный момент μ_2^*:

$$\sigma^2 = \frac{n}{n-1} * \mu_2^*. \tag{26}$$

Среднее квадратическое отклонение σ определяется как арифметический квадратный корень из дисперсии:

$$\sigma = \sqrt{\sigma^2}. \tag{27}$$

Коэффициент вариации служит мерой относительного разброса данных:

$$V = \frac{\sigma}{\bar{x}} \text{ или } C_V = \frac{\sigma}{\bar{x}} * 100\%. \quad (28)$$

Выбор той или иной характеристики рассеяния выборки неоднозначен. Наиболее просто вычисляется размах, так же просто вычисляются средний межквартильный размах и персентильный размах, причем, как и все оценки, основанные на квантилях, они оказываются более устойчивыми к поведению распределений на «хвостах» распределений и более стабильны для распределений, отклоняющихся от нормального. Ошибки в данных приводят к значительным колебаниям среднего арифметического и дисперсии, в то время как оценки, основанные на квантилях, в меньшей степени подвержены такого рода колебаниям.

2.8. Оценка характеристик формы распределения

На рисунках 7–9 приведены три вида унимодальных распределений с различной асимметрией. Асимметрия обычно измеряется **коэффициентом асимметрии S_{k1}:**

$$S_{k1} = \frac{\mu_3^*}{\sigma^3}, \quad (29)$$

где μ_3^* – выборочный центральный момент третьего порядка.

Значительно проще вычисляется другой показатель асимметрии основе квантилей распределения:

$$S_{k2} = \frac{Q_3 - Q_1 - 2Q_2}{Q_3 - Q_1}. \quad (30)$$

Коэффициент эксцесса применяется для оценки «остроты пика»:

$$K = \frac{\mu_4^*}{\sigma^4} - 3, \quad (31)$$

где μ_4^* – выборочный центральный момент четвертого порядка.

Для нормального распределения коэффициенты асимметрии и эксцесса равны нулю.

2.9. Эмпирическая функция распределения

Пусть получено статистическое распределение выборки и каждому варианту из этой выборки поставлена в соответствие его частость (число, показывающее, сколько раз определенное значение, признак или событие встречается в наборе данных (выборке). Это базовый инструмент для анализа распределения данных и понимания того, какие значения встречаются чаще, а какие реже).

Определение. *Эмпирической функцией* (функцией распределения выборки) называется функция $F^*(x)$, определяющая для каждого значения x частость события $\{X < x\}$:

$$F^*(x) = \frac{n_x}{n}, \quad (32)$$

где n – объем выборки, n_x – число наблюдений, меньших $x (x \in R)$.

При увеличении объема выборки частость события $\{X < x\}$ приближается к вероятности этого события. Эмпирическая функция $F^*(x)$ является оценкой интегральной функции $F(x)$ в теории вероятностей.

Функция $F^*(x)$ обладает теми же свойствами, что и функция $F(x)$:

1. $0 \le F^*(x) \le 1$;
2. $F^*(x)$ – неубывающая функция;
3. $F^*(-\infty) = 0, F^*(+\infty) = 1$.

Пример 4. Построить эмпирическую функцию и ее график по выборке: 2, 4, 7, 3, 1, 1, 3, 2, 7, 3.

Дискретный статистический ряд удобно записывать в виде таблице 3, предварительно рассчитав:

$\sum_{i=1}^{5} n_i = 10$;

$\sum_{i=1}^{5} p_i^* = 1$;

где x_i – ранжированная совокупность вариантов; n_i – частота; p_i^* - частость.

Таблица 3 (для примера 3)

x_i	1	2	3	4	7
n_i	2	2	3	1	2
$\frac{n_i}{n}$	$\frac{2}{10}$	$\frac{2}{10}$	$\frac{3}{10}$	$\frac{1}{10}$	$\frac{2}{10}$

$$F^*(x) = \begin{cases} 0 \text{ при } x \le 0 \\ 0{,}2 \text{ при } 0 < x \le 2 \\ 0.4 \text{ при } 2 < x \le 3 \\ 0{,}7 \text{ при } 3 < x \le 4 \\ 0{,}8 \text{ при } 4 < x \le 10 \\ 1 \text{ при } x > 10 \end{cases}$$

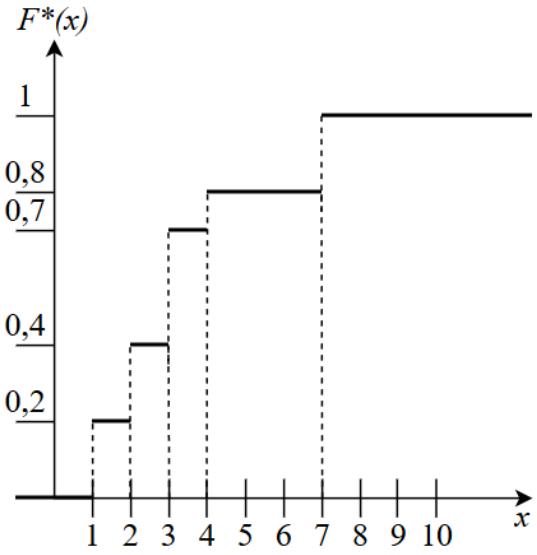


Рис. 11. Гистограмма к примеру 3

2.10. Эмпирическая плотность распределения

Для интегральной функции распределения $F(x)$ справедливо приближенное равенство: $F(x + \Delta) - F(x) \approx f(x) * \Delta x$, где $f(x)$ – дифференциальная функция распределения (функция плотности вероятности).

Потому естественно выборочным аналогом функции $f(x)$ считать функцию:

$$f^*(x) = \frac{F^*(x+\Delta x)-F^*(x)}{\Delta x}, \tag{33}$$

где $F^*(x+\Delta x)-F^*(x)$ – частость попадания наблюдаемых значений случайной величины X в интервал $[x; x+\Delta x)$.

Таким образом, значение $f^*(x)$ характеризует плотность частости на этом интервале.

Пусть наблюдаемые значения непрерывной случайной величины представлены в виде интервального вариационного ряда.

Полагая, что p_i^* – частость попадания наблюдаемых значений в интервал $[a_i;\ a_i+h)$, где h – длина частичного интервала, выборочную функцию плотности $f^*(x)$ можно задать соотношением:

$$f^*(x) = \begin{cases} 0 \text{ при } x < a_1 \\ \frac{p_i^*}{h} \text{ при } a_i \le x \le a_{i+1}, i = 1, 2, \dots, m, \\ 0 \text{ при } x > a_{m+1} \end{cases} \tag{34}$$

где a_{m+1} – конец последнего m-го интервала.

Так как функция $f^*(x)$ является аналогом распределения плотности случайной величины, площадь область под графиком этой функции равна 1.

2.11. Принципы статистического оценивания. Классификация оценок

Часто объектом статистического анализа являются результаты наблюдения случайной величины X, вид распределения которой неизвестен. Например, если имеются данные о росте студентов в возрасте 20 лет, то опыт обработки такого рода данных дает основание утверждать, что наблюдаемая случайная величина X имеет нормальное распределение $N(m, \sigma^2)$. Задача состоит в нахождении приближенных значений-оценок параметров распределения m (m – математическое ожидание) и σ^2 (σ^2 – дисперсия) по выборке. Аналогично, ставится задача об оценке параметров других распределений.

В общем случае, пусть $x_1, x_2, \dots, x_n$ – выборка наблюдений случайной величины X с известным законом распределения $F(x, \theta)$, где параметр θ неизвестен. Оценка $\tilde{\theta}_n$ параметра θ по выборке объема n вычисляется как значение некоторой функции элементов выборки (такие функции называются **статистиками)**:

$$\tilde{\theta}_n = \tilde{\theta}_n(x_1, x_2, \dots, x_n), \tag{35}$$

Как правило, существуют несколько статистик, которые можно использовать для оценки одного и того же параметра. Например, в случае если распределение генеральной совокупности симметрично относительно математического ожидания m, для оценки параметра m можно взять выборочное среднее:

$$\tilde{m}_1 = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \tag{36}$$

или выборочную медиану h:

$$\tilde{m}_1 = h, \tag{37}$$

Оценка параметра θ при помощи любой статистики дает приближенное значение параметра. Более того, значения статистик для разных выборок объема n из одной генеральной совокупности будут различны. В качестве оценки следует взять такую статистику, значения которой для различных выборок из данной генеральной совокупности были бы «в среднем» близки к истинному значению параметра. Желательно также, чтобы с увеличением объема выборки точность оценки возрастала.

Для выбора статистики, которая определяет значение оценки $\tilde{\theta}_n$ имеющее наименьший разброс около истинного значения параметра θ, полученную выборку наблюдений $x_1, x_2, \dots, x_n$ рассматривают, как одну из возможных реализаций выборочного вектора $(X_1, X_2, \dots, X_n)$, а оценку – как функцию его элементов:

$$\tilde{\theta}_n = \tilde{\theta}_n(X_1, X_2, \dots, X_n), \tag{38}$$

Распределение случайной величины θ называют **выборочным распределением оценки** (или просто распределением оценки). Лучшая оценка параметра может быть получена путем сравнения выборочных распределений нескольких подходящих статистик.

На рисунке 12 приведены эскизы графиков плотностей распределений для трех статистик $\tilde{\theta}_1, \tilde{\theta}_2, \tilde{\theta}_3$, используемых для оценки одного параметра θ. Очевидно, статистика $\tilde{\theta}_1$, предпочтительнее статистик $\tilde{\theta}_2$ и $\tilde{\theta}_3$ так как плотность $f_{\tilde{\theta}_1}$ теснее сосредоточена около истинного значения θ, чем плотности $f_{\tilde{\theta}_2}$ и $f_{\tilde{\theta}_3}$, и, следовательно, значения $\tilde{\theta}_1$, в «среднем» будут ближе к θ, чем значения $\tilde{\theta}_2$ и $\tilde{\theta}_3$.

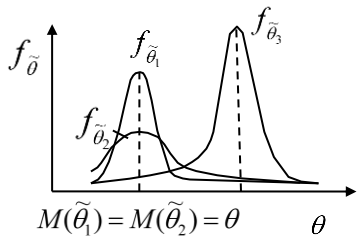


Рис. 12. Графики плотностей распределения статистик, используемых для оценки одного параметра θ

Однако часто выборочное распределение той или иной статистики получить невозможно. В этом случае для выбора наилучшей оценки используют несколько свойств, которые характеризуют качество оценок параметров и которыми должна обладать «хорошая» оценка. Это следующие свойства.

1. Состоятельность. Оценка $\tilde{\theta}_n = \tilde{\theta}_n(X_1, X_2, \dots, X_n)$ называется состоятельной оценкой параметра θ, если, при $n \to \infty$, $\tilde{\theta}_n$ сходится по вероятности к θ.

Состоятельность оценки θ во многих случаях может быть установлена с помощью следующей теоремы.

Теорема 1 (достаточное условие состоятельности). Если $M[\tilde{\theta}_n] \to \theta$ и $D[\tilde{\theta}_n] \to 0$ при $n \to \infty$, то $\tilde{\theta}_n$ – состоятельная оценка параметра θ.

2. Несмещенность. Оценка $\tilde{\theta}_n$ называется несмещенной оценкой параметр θ, если ее математическое ожидание равно оцениваемому параметру, то есть $M[\tilde{\theta}_n] = \theta$. Разность $M[\tilde{\theta}_n] - \theta$ называется **смещением,** или **систематической ошибкой.** На рисунке 12 $f_{\tilde{\theta}_3}$ –

плотность смещенной оценки $\tilde{\theta}_3$ параметра θ, смещение равно $M[\tilde{\theta}_3] = \theta$. Для несмещенных оценок систематическая ошибка оценивания равна нулю.

3. Пусть для оценки параметра θ можно использовать две несмещенные оценки $\tilde{\theta}_1$ и $\tilde{\theta}_2$. Если $D[\tilde{\theta}_1] < D[\tilde{\theta}_2]$, то говорят, что оценка $\tilde{\theta}_1$, **более эффектна,** чем оценка $\tilde{\theta}_2$. На рисунке 12 показаны плотности двух несмещенных оценок $\tilde{\theta}_1$ и $\tilde{\theta}_2$: $M[\tilde{\theta}_1] < M[\tilde{\theta}_2]$. Для этих оценок $D[\tilde{\theta}_1] < D[\tilde{\theta}_2]$ и, следовательно, оценка $\tilde{\theta}_1$ более эффективна, чем $\tilde{\theta}_2$.

В математической статистике разработаны различные методы нахождения оценок параметров распределений: метод моментов, метод максимально правдоподобия, метод наименьших квадратов и другие.

КОНТРОЛЬНЫЕ ВОПРОСЫ

1. Что понимается под статистическими данными?
2. Чем количественные признаки отличаются от качественных?
3. Какие бывают количественные признаки по характеру изменения значений?
4. Что такое номинальная шкала и какие признаки в ней измеряются?
5. Что такое ординальная шкала и чем она отличается от номинальной?
6. Что такое ранг и для чего он используется?
7. Какие статистические методы применяются для ординальных признаков?
8. Какие задачи решаются при анализе номинальных признаков?
9. Что называется генеральной совокупностью?
10. Что такое выборка и чем она отличается от генеральной совокупности?
11. Что означает репрезентативность выборки?
12. В чем заключается простой случайный отбор?
13. Что такое типический отбор и в каких случаях он применяется?
14. Что такое вариационный ряд?
15. Чем статистический дискретный ряд отличается от статистического интервального ряда?
16. Как определяется число интервалов группировки по правилу Стерджесса?
17. Что такое размах выборки и как он вычисляется?
18. Что называют гистограммой, полигоном распределения и огивой?
19. Какие характеристики положения используют для описания распределения выборки?
20. Что такое эмпирическая функция распределения и какие свойства она имеет?

ПРАКТИЧЕСКАЯ РАБОТА 1

Требуется представить выборку из 70 элементов в виде таблицы частот, рассчитать количество интервалов группировки по формуле Стерджесса, размах выборки, длину интервала группировки, а затем построить интервальный ряд. Рассчитать среднее значение выборки, моду и медиану. Построить гистограмму, полигон и огиву по данным варианта.

Вариант 1

Время обработки заявок в информационной системе, единицы измерения – минуты.
Выборка наблюдений:

12,4	15,1	14,7	16,3	18,2	13,9	17,5	19,1	20,4	14,2
15,8	16,7	17,9	18,5	13,4	14,9	15,6	16,1	17,3	18,8
19,5	20,1	21,4	22,0	13,7	14,5	15,2	16,9	17,6	18,1
19,8	20,6	21,2	22,4	12,9	13,5	14,8	15,7	16,4	17,8
18,6	19,2	20,3	21,7	22,6	13,1	14,1	15,4	16,6	17,1
18,4	19,7	20,8	21,1	22,3	12,7	13,8	14,6	15,7	16,8
17,4	18,9	19,4	20,7	21,5	22,1	13,3	15,0	17,0	18,7

Вариант 2

Суточное потребление интернет-трафика пользователями, единицы измерения – ГБ.
Выборка наблюдений:

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

Вариант 3

Время отклика сервера, единицы измерения – мс.
Выборка наблюдений:

82	95	101	110	124	88	97	106	115	129
84	92	103	112	121	86	99	108	117	126
81	94	105	111	123	89	96	107	116	128
83	91	102	114	125	87	98	109	118	127
85	93	104	113	122	90	100	119	120	130
82	112	121	86	103	99	108	117	126	95
101	110	124	88	97	106	115	129	84	92

Вариант 4

Ежедневное число обращений в службу поддержки, единицы измерения – шт.
Выборка наблюдений:

18	25	32	22	29	18	25	34	22	27
21	28	18	25	34	22	29	18	25	31
24	31	24	28	21	26	31	21	29	21
27	20	27	31	24	28	20	26	33	22
30	23	30	20	27	33	24	28	20	26
19	26	33	23	30	19	27	32	23	30
22	29	19	26	32	23	30	19	24	34

Вариант 5

Температура процессора при нагрузке, единицы измерения – °C.
Выборка наблюдений:

48,2	52,5	55,8	50,2	54,3	49,1	51,5	56,1	50,8	53,5
50,1	54,1	49,3	51,6	56,5	50,6	53,6	48,6	52,7	55,7
51,4	56,3	50,7	53,4	48,7	52,9	55,4	50,3	54,2	49,2
53,7	48,9	52,8	55,1	50,4	54,8	49,7	51,7	56,4	50,5
55,2	50,5	54,6	49,8	51,8	56,2	51,0	53,3	48,8	52,6
49,6	51,9	56,0	50,9	53,0	48,3	52,4	55,9	50,1	54,7
50,8	53,1	48,5	52,1	55,6	50,0	54,5	49,4	51,3	56,6

Вариант 6

Масса отправлений на складе, единицы измерения – кг.
Выборка наблюдений:

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

Вариант 7

Продолжительность телефонных разговоров, единицы измерения – минуты.
Выборка наблюдений:

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

Вариант 8

Уровень шума оборудования, единицы измерения – дБ.
Выборка наблюдений:

42	44	47	49	51	43	45	48	50	52
41	46	47	49	53	42	44	48	50	51
43	45	46	49	52	41	44	47	50	53
42	46	48	49	51	43	45	47	50	52
41	44	46	48	53	42	45	49	51	52
43	44	47	50	51	41	46	48	49	53
42	45	46	47	50	43	44	48	51	52

Вариант 9

Скорость загрузки файлов в сети, единицы измерения – Мбит/с.
Выборка наблюдений:

32,4	35,1	37,6	40,2	42,7	33,8	36,4	38,9	41,5	43,2
31,9	34,7	37,1	39,8	42,4	33,1	35,9	38,4	41,0	43,6
32,7	35,3	37,8	40,5	42,9	34,0	36,6	39,1	41,7	43,4
31,5	34,9	37,4	40,0	42,6	33,4	36,1	38,7	41,3	43,0
32,2	35,6	37,9	40,7	42,8	33,7	36,3	39,0	41,8	43,5
31,8	34,5	37,2	39,9	42,3	33,0	35,8	38,5	41,1	43,1
32,6	35,2	37,7	40,4	42,5	33,5	36,0	38,8	41,4	43,3

Вариант 10

Рост растений в теплице, единицы измерения – см.
Выборка наблюдений:

18,2	19,5	20,1	21,4	22,8	18,9	19,7	20,6	21,9	23,1
18,4	19,2	20,3	21,6	22,5	18,7	19,9	20,8	22,1	23,4
18,1	19,4	20,0	21,3	22,7	18,8	19,6	20,7	22,0	23,2
18,5	19,1	20,4	21,8	22,6	18,6	19,8	20,9	21,7	23,0
18,3	19,3	20,2	21,5	22,9	18,9	19,5	20,5	21,9	23,3
18,0	19,7	20,1	21,9	22,4	18,7	19,9	20,8	22,2	23,1
18,6	19,2	20,6	21,6	22,8	18,5	19,6	20,7	22,0	23,5

Вариант 11

Время выполнения SQL-запросов, единицы измерения – с.
Выборка наблюдений:

0,82	0,95	1,04	1,16	1,28	0,88	0,97	1,09	1,18	1,31
0,84	0,93	1,06	1,14	1,25	0,86	0,99	1,11	1,20	1,33
0,81	0,94	1,05	1,17	1,29	0,89	0,98	1,10	1,19	1,30
0,83	0,92	1,03	1,15	1,27	0,87	1,00	1,12	1,21	1,34
0,85	0,96	1,07	1,13	1,26	0,90	1,01	1,08	1,22	1,32
0,82	0,95	1,04	1,16	1,28	0,88	0,97	1,09	1,18	1,31
0,84	0,93	1,06	1,14	1,25	0,86	0,99	1,11	1,20	1,33

Вариант 12

Расход топлива автомобилей на маршруте, единицы измерения – л/100 км.
Выборка наблюдений:

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

3. ИНТЕРВАЛЬНЫЕ ОЦЕНКИ. ПОНЯТИЕ ДОВЕРИТЕЛЬНОГО ИНТЕРВАЛА

По выборочным данным можно приблизительно определить значение показателя генеральной совокупности. Однако у точечных оценок есть существенный недостаток: если имеется только одна выборка и по ней получено одно расчетное значение, то невозможно судить о точности результата и величине его отклонения от истинного значения. Именно этот недостаток устраняют интервальные оценки, которые позволяют определить степень надежности полученного результата.

Интервальная оценка – это определенный диапазон значений, в пределах которого оценивается параметр генеральной совокупности. При таком подходе предполагается, что истинное значение находится внутри заданного интервала. Например, записи вида $19{,}9 < m < 20{,}7$ или $m = 20{,}3 \pm 0{,}4$ являются примерами интервальных оценок.

Построение доверительного интервала основано на задании вероятности того, что истинное значение попадет в выбранный промежуток. Обычно в статистике используют доверительную вероятность 95%, то есть исследователь предполагает с такой степенью уверенности, что искомый параметр находится в указанном интервале. Если же требуется более высокая надежность, например 99% или 99,5%, то интервал становится шире. Так, для некоторого среднего интервал с доверительной вероятностью 99% может быть записан $19{,}7 < m < 20{,}9$ или $m = 20{,}3 \pm 0{,}6$.

Доверительный интервал – это интервальная оценка параметра генеральной совокупности, определяемая с помощью выборки для заданной доверительной вероятности.

Доверительная вероятность (p_0) интервальной оценки – это вероятность, что истинное (генеральное) значение характеристики попадет в заданный интервал.

Уровень значимости (α) – величина, равная вероятности, что реальное значение не попадает в заданный интервал $\alpha = 1 - p_0$.

Доверительная вероятность обычно задается в процентах, а уровень значимости – в долях от единицы. Наиболее часто используются 95% и 99% (для медицинских исследований, например) доверительные интервалы, которым соответствуют уровни – значимости $\alpha = 0{,}05$ или $\alpha = 0{,}01$. Реже встречается 90% доверительный интервал, $\alpha = 0{,}1$. При построении доверительных интервалов по известному распределению оценки пользуются одним и тем же приемом, рассмотренным на примере нормального распределения.

Пусть есть некоторая случайная величина z, подчиняющаяся нормальному закону распределения $N(0, 1)$. Необходимо найти такое значение $z_{\frac{\alpha}{2}}\ \alpha = 1 - p_0$, при котором вероятность $P(-z_{\frac{\alpha}{2}} \le z \le +z_{\frac{\alpha}{2}})$ равна p_0, при $p_0 = 90\%, 95\%, 99\%$.

Решение. Следует отметить, что данная задача является обратной по отношению к рассмотренным выше. Так как ранее по заданному интервалу значений случайной величины определялась вероятность попадания ее в некоторый интервал. В этой же задаче задана вероятность, а интервал, в который она попадает необходимо найти. На рисунке 13 представлен график функции плотности распределения вероятностей случайной величины z.

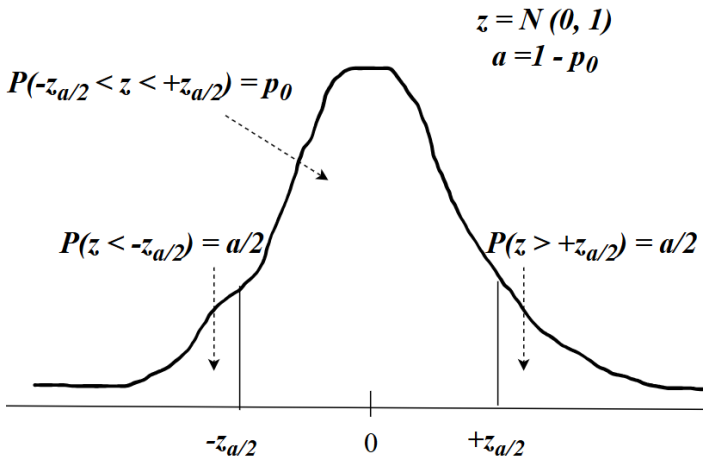


Рис. 13. Нахождение интервала по заданной вероятности

Согласно основному свойству плотности распределения площадь фигуры под графиком в интервале $-z_{\frac{\alpha}{2}} \leq z \leq +z_{\frac{\alpha}{2}}$ равняется p_0. Величина $\alpha = 1 - p_0$ представляет вероятность того, что величина z не попадет на этот интервал и будет находится на одном из «хвостов» распределения в области $z < -z_{\frac{\alpha}{2}}$ или $z_{\frac{\alpha}{2}} < z$. Так как мы имеет дело с симметричной функцией плотности распределения вероятностей, вероятности этих двух событий равны между собой и равны $\frac{\alpha}{2}$ (Рис. 13). Этот факт одновременно служит пояснением обозначения $z_{\frac{\alpha}{2}}$, использованного для границ интервала. Согласно функции Лапласа F, представим:

$$P\left(z \leq -z_{\frac{\alpha}{2}}\right) = P\left(z \geq z_{\frac{\alpha}{2}}\right) = 0{,}5 - F\left(z_{\frac{\alpha}{2}}\right) = \frac{\alpha}{2} \rightarrow F\left(z_{\frac{\alpha}{2}}\right) = 0{,}5 - \frac{\alpha}{2}, \quad (39)$$

С помощью таблицы функции Лапласа находим значение аргумента $z_{\frac{\alpha}{2}}$ функции, такое, чтобы ее значение при этом приближенно равнялось $(0{,}5 - \frac{\alpha}{2})$:

$$p_0 = 90\%;\ \alpha = 1 - p_0 = 0{,}1;\ \frac{\alpha}{2} = 0{,}05; F\left(z_{\frac{\alpha}{2}}\right) = 0{,}5 - \frac{\alpha}{2} = 0{,}45;\ z_{\frac{\alpha}{2}} = 1{,}64;$$

$$p_0 = 95\%;\ \alpha = 1 - p_0 = 0{,}05;\ \frac{\alpha}{2} = 0{,}025; F\left(z_{\frac{\alpha}{2}}\right) = 0{,}5 - \frac{\alpha}{2} = 0{,}475;\ z_{\frac{\alpha}{2}} = 1{,}96;$$

$$p_0 = 99\%;\ \alpha = 1 - p_0 = 0{,}01;\ \frac{\alpha}{2} = 0{,}005; F\left(z_{\frac{\alpha}{2}}\right) = 0{,}5 - \frac{\alpha}{2} = 0{,}495;\ z_{\frac{\alpha}{2}} = 2{,}58;$$

Таким образом искомые доверительные интервалы равны:

$$P = (-1{,}64 \leq z \leq +1{,}64) = 90\%$$

$$P = (-1{,}96 \leq z \leq +1{,}96) = 95\%$$

$$P = (-2{,}58 \leq z \leq +2{,}658) = 99\%$$

По полученным доверительным оценкам проанализируем выборочную среднюю величину $\bar{x}$. Можно сказать, что эта величина приближенно подчиняется нормальному распределения $N(m; \frac{\sigma}{\sqrt{n}})$, где m – математическое ожидание (или генеральное среднее), σ – среднеквадратическое отклонение (генеральное отклонение). Пользуясь преобразованием

общего нормального распределения к стандартному, получаем, что величина $\frac{\bar{x}-m}{\frac{\alpha}{\sqrt{n}}}$ имеет распределение $N(0,1)$. По результатам задачи получаем:

$$P\left(-z_{\frac{\alpha}{2}} \leq \frac{\bar{x}-m}{\frac{\sigma}{\sqrt{n}}} \leq +z_{\frac{\alpha}{2}}\right) = p_0, \quad (40)$$

Умножаем части неравенства на $\frac{\sigma}{\sqrt{n}}$:

$$P\left(-z_{\frac{\alpha}{2}} * \frac{\sigma}{\sqrt{n}} \leq \bar{x} - m \leq +z_{\frac{\alpha}{2}} * \frac{\sigma}{\sqrt{n}}\right) = p_0, \quad (41)$$

Далее прибавляем m получаем окончательно:

$$P\left(m - z_{\frac{\alpha}{2}} * \frac{\sigma}{\sqrt{n}} \leq \bar{x} \leq m + z_{\frac{\alpha}{2}} * \frac{\sigma}{\sqrt{n}}\right) = p_0, \quad (42)$$

Таким образом, для вероятности $p_0 = 1 - \alpha$ выполняется следующее соотношение:

$$m - z_{\frac{\alpha}{2}} * \frac{\sigma}{\sqrt{n}} \leq \bar{x} \leq m + z_{\frac{\alpha}{2}} * \frac{\sigma}{\sqrt{n}}, \quad (43)$$

Соотношение (43) не является интервальной оценкой, так как в нем положение известной выборочной величины $\bar{x}$ выражается через неизвестную величину математического ожидания m генеральной совокупности и генеральное отклонение σ.

3.1. Интервальная оценка среднего при известном среднеквадратическом отклонении

Построим доверительный интервал для среднего в случае, когда значение среднего квадратического отклонения σ известно. Запишем формулу (43) для вероятности $p_0 = 95\%, z_{\frac{\alpha}{2}} = 1{,}96$:

$$m - 1{,}96 * \frac{\sigma}{\sqrt{n}} \leq \bar{x} \leq m + 1{,}96 * \frac{\sigma}{\sqrt{n}}$$

Это значит, что для выборки из n элементов выборочное среднее с 95% вероятностью отклонится от истинного значения m не более чем на $1{,}96 * \frac{\sigma}{\sqrt{n}}$. Взаимное положение выборочных средних относительно генерального изображено на рисунке 14.

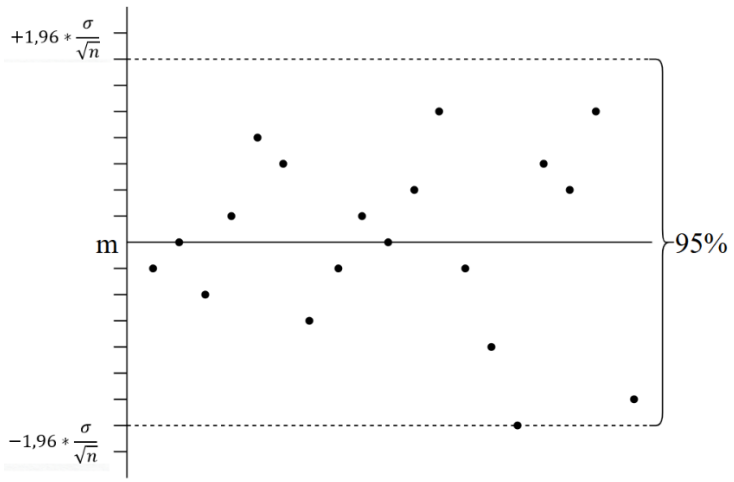


Рис. 14. 95% доверительные интервалы для выборочных средних

Формулу (43) можно представить в другом виде. Если мы знаем конкретное значение $\bar{x}$, то вокруг него 95-процентный доверительный интервал можно записать в виде:

$$\bar{x} = \pm 1,96 * \frac{\sigma}{\sqrt{n}}, \tag{44}$$

На рисунке 15 представлены 95% интервальные оценки для различных выборочных средних.

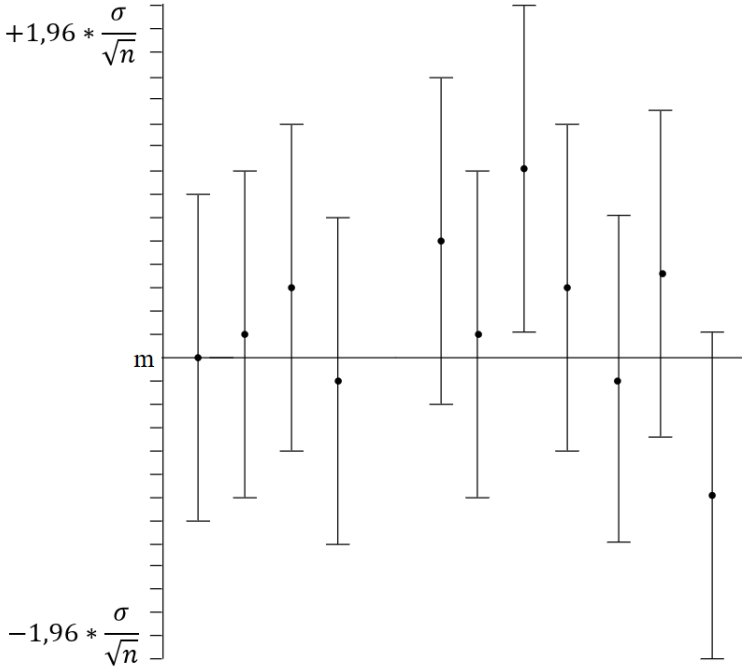


Рис. 15. 95% интервальные оценки для различных выборочных средних

В 95% всех случаев доверительные интервалы, построенные по выборочным средним, захватывают истинное значение. Но 5% случаев генеральное среднее не захватывается. Таким образом, в частном случае $p_0 = 95\%$ интервальная оценка имеет вид:

$$\bar{x} - 1,96 * \frac{\sigma}{\sqrt{n}} \leq m \leq \bar{x} + 1,96 * \frac{\sigma}{\sqrt{n}}, \tag{45}$$

В общем виде формула доверительной вероятности имеет вид:

$$\bar{x} - z_{\frac{\alpha}{2}} * \frac{\sigma}{\sqrt{n}} \leq m \leq \bar{x} + z_{\frac{\alpha}{2}} * \frac{\sigma}{\sqrt{n}}, \tag{46}$$

Максимальная ошибка оценки E рассчитывается по формуле (47):

$$E = z_{\frac{\alpha}{2}} * \frac{\sigma}{\sqrt{n}}, \tag{47}$$

Полученная оценка включает в себя значение среднего квадратического отклонения генеральной совокупности σ. Существует два случая для нахождения этого значения:

1. Из данных предшествующих наблюдений (экспериментов), предполагая, что в прошлых итерациях экспериментов величина отклонения не изменилась, то есть $\sigma = \sigma_0$, особенно при больших объемах выборок. В других случаях неизменной полагают величину коэффициента вариации $V = \frac{\sigma_0}{\bar{x}_0} * 100\%$**,** а текущее отклонение оценивают как $\sigma = \frac{V * \bar{x}_0}{100\%}$.

2. В зависимости от закона распределения. Если распределение близко к нормальному, то размах вариации выборки (3) примерно в 6 раз больше среднеквадратического отклонения, то есть $R \approx 6 * \sigma$. В таком случае, зная значения минимального и максимального элемента выборки, можно оценить величину среднеквадратического отклонения σ:

$$\sigma \approx \frac{R}{6} = \frac{x_{max} - x_{min}}{6}, \quad (48)$$

Пример 5. При обслуживании 40 клиентов банка, специалист аналитического отдела определил, что среднее время обработки и решения вопроса клиента составляет 16,2 минут. Из предшествующих наблюдений другого специалиста известно, что среднеквадратическое отклонение σ этой величины составляет 3 минуты. Построить 95% доверительный интервал для среднего времени обработки и решения вопросов всех клиентов банка.

Алгоритм решения заключается в следующем:

1. Рассчитываем среднюю ошибку выборки по формуле (49):

$$\frac{\sigma}{\sqrt{n}} = \frac{3}{\sqrt{40}} \approx 0{,}47, \quad (49)$$

2. Используем уровень значимости для $p_0 = 95\%$: $\alpha = 1 - p_0 = 1 - 0.95 = 0{,}05$ и $z_{\frac{\alpha}{2}} = 1{,}96$.

3. Рассчитываем максимальную ошибку выборки по формуле (47):

$$E = z_{\frac{\alpha}{2}} * \frac{\sigma}{\sqrt{n}} \approx 1{,}96 * 0{,}47 \approx 0{,}92$$

4. Применяя общую формулу (46), получаем:

$$m \in \left[\bar{x} - z_{\frac{\alpha}{2}} * \frac{\sigma}{\sqrt{n}};\ \bar{x} + z_{\frac{\alpha}{2}} * \frac{\sigma}{\sqrt{n}}\right] = [16{,}2 - 0{,}92; 16{,}2 + 0{,}92] = [15{,}28; 17{,}12] = 16{,}2 \pm 0{,}92$$

3.2. Определение размера выборки

Другим важным аспектом, неразрывно связанным с формированием доверительных интервалов, является расчет необходимого объема выборки. На этапе планирования эксперимента исследователю часто нужно понять: какое количество объектов необходимо обследовать, чтобы погрешность результата (при выбранном уровне надежности) оставалась в строго определенных пределах.

Например, при изучении среднего времени, которое пользователи тратят на мобильное приложение, необходимо заранее установить предельную ошибку, которая считается допустимой (например, 5 секунд, 30 секунд или 1 минута), а также выбрать уровень доверительной вероятности p_0 – ту степень уверенности, с которой мы хотим получить результат (допустим, 90%, 95% или 99%). Если нам известно стандартное отклонение σ в генеральной совокупности, расчет минимально достаточного размера выборки выполняется на основе формулы максимальной ошибки (47).

В формуле (47) нам необходимо перенести $\sqrt{n}$ в одну сторону, все остальное – в другую, затем возводим обе части в квадрат, получим:

$$\sqrt{n} = z_{\frac{\alpha}{2}} * \frac{\sigma}{E}, \quad (50)$$

И извлекаем корень:

$$n = \left(z_{\frac{\alpha}{2}} * \frac{\sigma}{E}\right)^2, \quad (51)$$

Формула (51) есть искомая оценка размера выборки. В случае необходимости она округляется в большую сторону до ближайшего целого.

Пример 6. Руководитель отдела кадров компании хочет оценить, сколько часов в месяц в среднем сотрудники тратят на внутреннее обучение. При этом он хочет получить результат с точностью до 1 часа и быть уверенным в нем на 99%. Сколько сотрудников необходимо опросить, если по данным прошлых лет известно, что среднее квадратическое отклонение этого показателя составляет 5 часов?

Решение. По условиям задачи $\sigma = 5$ часов, $E = 1$ час, доверительная вероятность $p_0 = 99\%, \alpha = 0{,}01, z_{\frac{\alpha}{2}} = 2{,}58$. По формуле (51) определения размера выборки получаем:

$$n = \left(z_{\frac{\alpha}{2}} * \frac{\sigma}{E}\right)^2 = \left(2{,}58 * \frac{5}{1}\right)^2 \approx 167 \text{ человек}$$

Таким образом, отдел кадров должен опросить не менее 167 человек.

При анализе точечных оценок учитывался коэффициент поправки на конечный объем выборки. Применение этого коэффициента позволяет корректировать расчет средней и предельной ошибок выборки следующим образом:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} * \sqrt{1 - \frac{n}{N}}; E = z_{\frac{\alpha}{2}} * \frac{\sigma}{\sqrt{n}} * \sqrt{1 - \frac{n}{N}}$$

Конечность выборки легко учесть и в формуле расчета длины:

$$E = z_{\frac{\alpha}{2}} * \frac{\sigma}{\sqrt{n}} * \sqrt{1 - \frac{n}{N}} = z_{\frac{\alpha}{2}} * \sigma * \sqrt{\frac{1}{n} - \frac{1}{N}}, \quad (52)$$

$$n = \frac{z_{\frac{\alpha}{2}}^2 * \sigma^2 * N}{E^2 * N + z_{\frac{\alpha}{2}}^2 * \sigma^2}$$

3.3. Интервальная оценка для доли (биномиального параметра)

Пусть исследуемый признак у элементов генеральной совокупности имеет альтернативный характер, то есть каждый объект либо обладает данным признаком, либо не обладает им. Тогда всю генеральную совокупность можно описать через долю единиц p, имеющих этот признак. В задачах вероятностного характера эта доля p выступает как параметр биномиального распределения. Если параметры генеральной совокупности неизвестны, то долю можно оценить по выборке с помощью несмещенной точечной оценки $p \approx \omega = \frac{k}{n}$, где n – объем выборки, ω – выборочная доля.

Интервальная оценка параметра p строится на том, что выборочная доля является средним значением случайных величин с одинаковым распределением $B(1, p)$, одинаковым математическим ожиданием p и среднеквадратическим отклонением $\sqrt{p(1-p)}$. Поэтому при достаточно большом объеме выборки можно использовать центральную предельную теорему

и считать, что выборочная доля ω приближенно подчиняется нормальному закону распределения с параметрами: $m_\omega = p, \sigma_\omega = \sqrt{\frac{p(1-p)}{n}}$.

После замены истинной дисперсии ее оценкой $\sigma_\omega = \sqrt{\frac{\omega(1-\omega)}{n}}$ получаем формулу доверительного интервала:

$$\omega - z_{\frac{\alpha}{2}}\sqrt{\frac{\omega(1-\omega)}{n}} < p < \omega + z_{\frac{\alpha}{2}}\sqrt{\frac{\omega(1-\omega)}{n}}, \quad (53)$$

Пример 7. Опрос 100 туристов показал, что 30% из них выбирают отдых в экскурсионный туризм. Построить 95% интервал для доли всех опрошенных, которые выбирают интеллектуальный отдых.

Решение. Количество элементов выборки $n = 100$, выборочная доля $\omega = 0{,}3$ при значении доверительной вероятности $p_0 = 99\%$, и уровня значимости $\alpha = 0{,}01, z_{\frac{\alpha}{2}} = 2{,}58$.

Подставим значения в формулу доверительного интервала (53), получим:

$$0{,}3 - 1{,}96 * \sqrt{\frac{0{,}3(1-0{,}3)}{100}} < p < 0{,}3 + 1{,}96 * \sqrt{\frac{0{,}3(1-0{,}3)}{100}}$$

$$0{,}3 - 0{,}0898 < p < 0{,}3 + 0{,}0898$$

$$02102 < p < 0{,}3898$$

Значение доверительного интервала для доли всех туристов, которые выбирают интеллектуальный отдых равно: $[21{,}02; 38{,}98]\%$.

3.4. Интервальная оценка генерального среднего при неизвестном среднеквадратическом отклонении

Среднеквадратическое отклонение генеральной совокупности σ должно быть известно, так как от него зависит интервальная оценка для генерального среднего значения (математического ожидания в общем случае) $m \in \bar{x} \pm z_{\frac{\alpha}{2}} * \frac{\sigma}{\sqrt{n}}$.

В случае, когда среднеквадратическое отклонение генеральной совокупности σ неизвестно, его заменяют на несмещенную оценку:

$$s \approx \sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

Построение любого доверительного интервала основано на законе распределения значений оценки, полученных по различном выборкам, относительно истинного значения. Так при известном значении $\sigma, \frac{\bar{x}-m}{\frac{\sigma}{\sqrt{n}}}$ – есть функция нормального распределения с параметрами $N(0, 1)$. А при неизвестном значении σ определить закон распределения случайной величины удается только в частном случае – когда значения признака в генеральной совокупности приближенно подчиняются нормальному распределению.

Свойство. Если значения x в генеральной совокупности приближенно описываются законом нормального распределения, то величина $\frac{\bar{x}-m}{\frac{\sigma}{\sqrt{n}}}$ подчиняется распределению Стьюдента (число степеней свободы $k = n - 1$).

Приведем стандартные свойства распределения Стьюдента:

1. Распределение Стьюдента является семейством кривых, зависящим от числа степеней свободы k.
2. Имеет колоколообразную форму и обладает осевой симметрией относительно оси ординат Oy.
3. Кривая распределения приближается к оси абсцисс Ox, но никогда не касается ее.
4. Среднее значение $\bar{x}$, медиана h и мода d совпадают и равны 0, дисперсия и среднеквадратическое отклонение больше 1.
5. При увеличении числа степеней свободы k распределение приближается к нормальному, а при $k > 30$ совпадает с ним практически полностью.

Число степеней свободы k – это показатель, который встречается в распределении Стьюдента и других статистических расчетах. По сути, это количество данных в выборке, которые могут свободно принимать любые значения при условии, что основные итоговые показатели (например, среднее арифметическое $\bar{x}$) уже определены и должны оставаться неизменными.

Пусть, например, в выборке с 5 элементами, среднее значение $\bar{x} = 10$: $\frac{1}{5} = (x_1 + x_2 + x_3 + x_4 + x_5) = 10$. Тогда из пяти элементов можно произвольно менять любые четыре, но пятое будет автоматически вычисляться из условия постоянства среднего. Отсюда следует, что число степеней свободы равно $k = 4$. Так если варьируются x_1, x_2, x_3, x_4, то $x_5 = 10 * 5 - (x_1 + x_2 + x_3 + x_4)$. Если же для этой выборки дополнительно зафиксировать и величину выборочного отклонения, то число степеней свободы $k = 3$. На рисунке 16 представлены кривые плотности распределения вероятностей при $k = 3$ и $k = 20$.

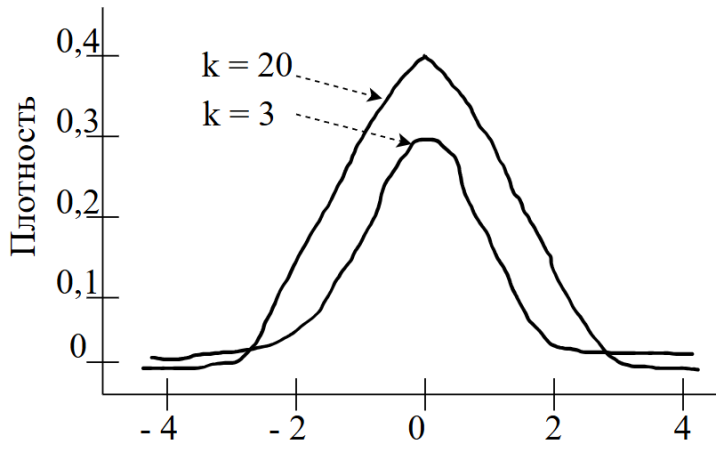


Рис. 16. Распределение Стьюдента

Для некоторой заданной выборки и доверительной вероятности p_0, на основе распределения Стьюдента, вид интервальной оценки следующий:

$$m \in \left[\bar{x} - t_{\frac{\alpha}{2}} * \frac{\sigma}{\sqrt{n}};\ \bar{x} + t_{\frac{\alpha}{2}} * \frac{\sigma}{\sqrt{n}}\right], \tag{54}$$

В этом случае критические значения $t_{\frac{\alpha}{2}}$, которые определяются из условия $P\left(t > t_{\frac{\alpha}{2}}\right) = P\left(t < -t_{\frac{\alpha}{2}}\right) = \frac{\alpha}{2}$, где $\alpha = 1 - p_0$, находятся по таблице распределения Стьюдента. Нужная строка таблицы определяется числом степеней свободы, а нужная колонка – по уровню значимости $\frac{\alpha}{2}$. Для длины выборки $n > 30$ критическое значение $t_{\frac{\alpha}{2}}$ находится по таблице нормального распределения $t_{\frac{\alpha}{2}} \approx z_{\frac{\alpha}{2}}$.

Например, найдем доверительные интервалы для выборки из $n = 11$ элементов и доверительной вероятности $p_0 = 95\%$. Вычислим число степеней свободы $k = n - 1 = 11 - 1 = 10, \frac{\alpha}{2} = \frac{1}{2}(1 - 0{,}95) = 0{,}025$. По таблице распределения найдем $t_{\frac{\alpha}{2}} = 2{,}23$, тогда 95% доверительный интервал будем равен $[-2{,}23; 2{,}23]$. Для сравнения, для стандартного нормального распределения 95% доверительный интервал более узкий $[-1{,}96; 1{,}96]$.

Пример 8. На предприятии провели контроль партии деталей и проверили 10 изделий. Установлено, что средняя величина контролируемого параметра составляет 8 условных единиц (у.е.) при среднеквадратическом отклонении 2 у.е. Построить 99% доверительный интервал для генерального среднего, предполагая, что контролируемый параметр приближенно описывается законом нормального распределения.

Решение. Будем предполагать, что генеральная совокупность распределена нормально, поэтому для построения доверительного интервала можно использовать распределение Стьюдента. Так как размер выборки $n = 10$, число степеней свободы $k = n - 1 = 10 - 1 = 9$, доверительная вероятность $p_0 = 99\%$, уровень значимости $\alpha = 0{,}01$, найдем находим $t_{\frac{\alpha}{2}} = 3{,}25$. Доверительный интервал равен:

$$m \in \left[\bar{x} - t_{\frac{\alpha}{2}} * \frac{\sigma}{\sqrt{n}};\ \bar{x} + t_{\frac{\alpha}{2}} * \frac{\sigma}{\sqrt{n}}\right] = 8 \pm 3{,}25 * \frac{2}{\sqrt{10}} \approx 8 \pm 2{,}05 \approx [6; 10]$$

Вывод: контролируемый параметр с 99% вероятностью лежит в интервале $[6; 10]$ у.е.

3.5. Интервальная оценка для дисперсии генеральной совокупности

Точечной оценкой для дисперсии D генеральной совокупности является величина:

$$D \approx s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2, \quad (55)$$

Интервальная оценка для дисперсии основывается на следующем утверждении:

Свойство. Если значения x в генеральной совокупности приближенно описываются нормальным распределением, то величина $\frac{(n-1)*s^2}{D}$ подчинена распределению χ^2 (хи-квадрат) с числом степеней свободы $k = n - 1$.

Свойства распределения хи-квадрат χ^2 (рис. 17):

1. Распределение χ^2 является семейством кривых, зависящим от числа степеней свободы.
2. Распределение χ^2 принимает только положительные значения и положительно-асимметрично.
3. Для числа степеней свободы $k > 100$ распределение становится приближенно-симметричным.

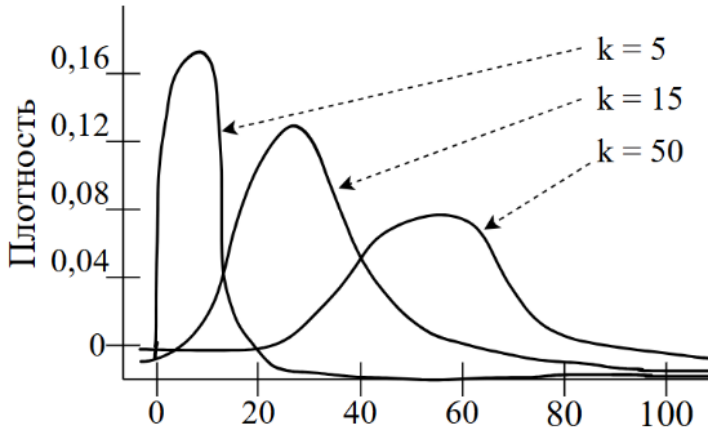


Рис. 17. Распределение хи-квадрат χ^2

При построении доверительного интервала для распределения χ^2 в силу его асимметричности (при заданном уровне значимости) необходимо находить два критических (граничных) значения – нижнее и верхнее, для двух интервалов:

$$P\left(\chi^2 < \chi^2_{\frac{\alpha}{2},\text{нижнее}}\right) = \frac{\alpha}{2} \text{ и } P\left(\chi^2 > \chi^2_{\frac{\alpha}{2},\text{верхнее}}\right) = \frac{\alpha}{2}, \quad (56)$$

Для того, чтобы определить эти границы используется таблица критических значений для распределения χ^2. Строка таблицы определяется числом степеней свободы k, нижнее критическое соответствует колонке $1 - \frac{\alpha}{2}$, верхнее $\frac{\alpha}{2}$.

Интервальная оценка для дисперсии представлена формулой (57):

$$\frac{(n-1)*s^2}{\chi^2_{\frac{\alpha}{2},\text{верхнее}}} < D < \frac{(n-1)*s^2}{\chi^2_{\frac{\alpha}{2},\text{нижнее}}}, \quad (57)$$

Пример 9. Контрольная лаборатория завода по производству косметических средств проводила проверку точности дозирования активного компонента в креме. В ходе исследования была произведена случайная выборка из 20 упаковок продукции. По результатам измерений выборочное среднее квадратическое отклонение массы компонента составило 1 мг. Требуется построить доверительный интервал для дисперсии данного параметра при уровне доверия 0,95, исходя из предположения о нормальном распределении исследуемого признака.

Решение:

1. По условию объем выборки $n = 20$, следовательно, число степеней свободы $k = n - 1 = 20 - 1 = 19$.
2. Выборочное отклонение $s = 1$, значит, исправленная выборочная дисперсия $s^2 = 1^2 = 1$.
3. Для построения интервальной оценки при доверительной вероятности $p_0 = 95\%$, уровень значимости $\alpha = 0{,}05$ используются критические точки распределения χ^2 с 19 степенями свободы.
4. Нижняя и верхняя границы интервала рассчитываются по формуле:

$$P\left(\chi^2 < \chi^2_{\frac{\alpha}{2},\text{нижнее}}\right) = 8{,}9 \text{ и } P\left(\chi^2 > \chi^2_{\frac{\alpha}{2},\text{верхнее}}\right) = 32{,}9$$

5. Используя соответствующие табличные значения χ^2, вычисляются границы, в которых с вероятностью 95% находится истинное значение генеральной дисперсии:

$$\frac{(n-1)*s^2}{\chi^2_{\frac{\alpha}{2},\text{верхнее}}} = \frac{(20-1)*1^2}{32{,}9} < D < \frac{(n-1)*s^2}{\chi^2_{\frac{\alpha}{2},\text{нижнее}}} = \frac{(20-1)*1^2}{8{,}9}$$

$$0{,}58 < D < 2{,}13 \text{ и } 0{,}76 < \sigma < 1{,}46$$

Дисперсия содержания активного компонента в креме находится в интервале $[0{,}76; 1{,}46]$ мг.

КОНТРОЛЬНЫЕ ВОПРОСЫ

1. Что называется интервальной оценкой параметра?
2. Чем интервальная оценка отличается от точечной?
3. Что такое доверительный интервал?
4. Что означает доверительная вероятность?
5. Как обозначают доверительную вероятность?
6. Что такое уровень значимости?
7. Как связан уровень значимости с доверительной вероятностью?
8. Какие значения доверительной вероятности чаще всего используют?
9. Какие уровни значимости соответствуют 90%, 95% и 99% доверия?
10. Почему при росте доверительной вероятности интервал становится шире?
11. Как изменится точность интервальной оценки при увеличении объема выборки в 4 раза?
12. Какое распределение (Стьюдента или нормальное) используется для построения интервала, если объем выборки мал, а дисперсия генеральной совокупности неизвестна?
13. Что такое стандартная нормальная случайная величина?
14. Зачем при построении доверительных интервалов используют нормальное распределение?
15. Что такое квантиль распределения и как он используется при определении границ доверительного интервала?
16. Почему доверительный интервал строится с учетом двух хвостов распределения?
17. Какую роль играет функция Лапласа в расчетах доверительных интервалов?
18. Как строится доверительный интервал для математического ожидания при известной дисперсии?
19. Что такое предельная ошибка оценки?
20. Как влияет объем выборки на ширину доверительного интервала?

ПРАКТИЧЕСКАЯ РАБОТА 2

Во всех вариантах для задачи 1 задано выборочное среднее, объем выборки и известно среднеквадратическое отклонение. Нужно построить 95% и 99% доверительные интервалы для среднего значения генеральной совокупности. Для задачи 2 необходимо построить доверительный интервал либо для генерального среднего значения (распределение Стьюдента), либо для генеральной дисперсии (распределение χ^2).

Вариант 1

Задача 1. При обследовании 36 сотрудников предприятия среднее время выполнения одного задания составило 24,6 мин. Из предыдущих наблюдений известно, что среднеквадратическое отклонение этого показателя равно 4,2 мин. Построить 95% и 99% доверительные интервалы для среднего времени выполнения задания всеми сотрудниками.

Задача 2. В лаборатории пищевого предприятия изучали массу порций готового продукта. Случайным образом было отобрано 16 упаковок. По результатам измерений выборочное среднее составило 248 г, а выборочное среднее квадратическое отклонение – 6 г. Предполагая, что распределение массы близко к нормальному, построить 95%-й доверительный интервал для генерального среднего.

Вариант 2

Задача 1. При анализе 49 измерений средняя масса упаковки составила 1,82 кг. Известно, что среднеквадратическое отклонение равно 0,18 кг. Построить 95% и 99% доверительные интервалы для среднего значения массы упаковки.

Задача 2. На складе бытовой техники исследовали время зарядки аккумуляторов новой модели устройства. Было проверено 12 приборов. Выборочное среднее время зарядки составило 3,8 ч, выборочное среднее квадратическое отклонение – 0,5 ч. Считая распределение признака нормальным, построить 99%-й доверительный интервал для генерального среднего времени зарядки.

Вариант 3

Задача 1. По результатам обследования 64 студентов средний балл по дисциплине составил 72,4. Среднеквадратическое отклонение известно и равно 8,5. Построить 95% и 99% доверительные интервалы для среднего балла всех студентов.

Задача 2. В учебном центре анализировали результаты тестирования слушателей. По выборке из 20 человек средний балл оказался равным 74, а выборочное среднее квадратическое отклонение составило 8 баллов. Принимая распределение результатов близким к нормальному, построить 95%-й доверительный интервал для генерального среднего балла.

Вариант 4

Задача 1. При оценке 100 деталей средняя длина детали составила 125,3 мм. Из предыдущих партий известно, что среднеквадратическое отклонение равно 6,4 мм. Построить 95% и 99% доверительные интервалы для среднего значения длины детали.

Задача 2. На мебельной фабрике контролировали длину деревянных заготовок. Было случайно отобрано 18 заготовок. Выборочное среднее значение длины составило 152,4 см, выборочное среднее квадратическое отклонение – 2,7 см. При предположении о нормальности распределения построить 90%-й доверительный интервал для генерального среднего значения длины.

Вариант 5

Задача 1. В ходе наблюдений за 81 посетителем среднее время ожидания обслуживания составило 9,8 мин. Среднеквадратическое отклонение известно и равно 2,7 мин. Построить 95% и 99% доверительные интервалы для среднего времени ожидания.

Задача 2. В сервисном центре исследовали продолжительность ремонта смартфонов. По выборке из 25 заказов среднее время ремонта составило 5,6 дня, а выборочное среднее квадратическое отклонение – 1,4 дня. Предполагая нормальность распределения, построить 95%-й доверительный интервал для генерального среднего времени ремонта.

Вариант 6

Задача 1. При изучении 25 образцов материала средняя прочность на разрыв составила 560 МПа. Среднеквадратическое отклонение равно 45 МПа. Построить 95% и 99% доверительные интервалы для среднего значения прочности.

Задача 2. В клинике проводили наблюдение за уровнем гемоглобина у пациентов после курса лечения. По результатам выборки из 14 пациентов выборочное среднее составило 128 г/л, а выборочное среднее квадратическое отклонение – 9 г/л. Считая распределение признака близким к нормальному, построить 99%-й доверительный интервал для генерального среднего уровня гемоглобина.

Вариант 7

Задача 1. По данным 72 измерений средняя температура воздуха в помещении составила 23,1 °C. Известно, что среднеквадратическое отклонение равно 1,6 °C. Построить 95% и 99% доверительные интервалы для среднего значения температуры.

Задача 2. На фармацевтическом предприятии изучали разброс массы таблеток одной серии. Случайным образом было отобрано 20 таблеток. Выборочное среднее квадратическое отклонение массы составило 1,2 мг. Предполагая нормальность распределения массы, построить 95%-й доверительный интервал для генеральной дисперсии.

Вариант 8

Задача 1. При исследовании 50 клиентов банка средняя сумма ежемесячного платежа составила 14,7 тыс. руб. Среднеквадратическое отклонение равно 3,1 тыс. руб. Построить 95% и 99% доверительные интервалы для среднего размера платежа.

Задача 2. В типографии исследовали изменчивость толщины бумажных листов. По выборке из 24 листов выборочное среднее квадратическое отклонение толщины составило 0,08 мм. При предположении о нормальном распределении признака построить 99%-й доверительный интервал для генеральной дисперсии толщины.

Вариант 9

Задача 1. На основе 90 наблюдений средний расход топлива автомобиля составил 7,4 л на 100 км. Известно, что среднеквадратическое отклонение равно 0,9 л. Построить 95% и 99% доверительные интервалы для среднего расхода топлива.

Задача 2. На автозаправочной станции контролировали разброс объема топлива, отпускаемого одной колонкой. Было выполнено 15 контрольных измерений. Выборочное

среднее квадратическое отклонение составило 0,16 л. Считая распределение признака нормальным, построить 95%-й доверительный интервал для генеральной дисперсии.

Вариант 10

Задача 1. При контроле качества 121 изделия средняя доля брака составила 2,8%. Среднеквадратическое отклонение показателя равно 0,7%. Построить 95% и 99% доверительные интервалы для среднего уровня брака.

Задача 2. В швейном цехе изучали изменчивость длины швов на готовых изделиях. Для анализа отобрали 18 изделий. Выборочное среднее квадратическое отклонение длины шва составило 0,9 см. Предполагая нормальность распределения, построить 90%-й доверительный интервал для генеральной дисперсии длины шва.

Вариант 11

Задача 1. В результате измерений 30 пациентов среднее систолическое давление составило 128,5 мм рт. ст. Среднеквадратическое отклонение известно и равно 11,2 мм рт. ст. Построить 95% и 99% доверительные интервалы для среднего давления.

Задача 2. В энергетической лаборатории исследовали колебания напряжения в электрической сети. По результатам 22 наблюдений выборочное среднее квадратическое отклонение составило 3,5 В. Принимая распределение признака нормальным, построить 95%-й доверительный интервал для генеральной дисперсии напряжения.

Вариант 12

Задача 1. При исследовании 75 рекламных кампаний средняя конверсия составила 4,6%. Среднеквадратическое отклонение равно 1,4%. Построить 95% и 99% доверительные интервалы для среднего значения конверсии.

Задача 2. На молочном заводе анализировали разброс жирности продукции в одной партии. По выборке из 17 упаковок выборочное среднее квадратическое отклонение жирности составило 0,24%. Считая распределение признака близким к нормальному, построить 99%-й доверительный интервал для генеральной дисперсии.

4. СТАТИСТИЧЕСКИЕ ГИПОТЕЗЫ

При проведении исследований часто нужно ответить на вопрос, верно ли то или иное предположение о генеральной совокупности. Например, определить, повышается ли температура Земли, снижает ли медикаментозное лечение давление, лучше ли новый метод обучения, предпочитают ли покупатели новую коллекцию или уменьшают ли подушки безопасности тяжесть травм.

Такие задачи решаются в разделе статистики, который называется проверкой статистических гипотез. Он позволяет по данным выборки судить об истинности утверждений о генеральной совокупности.

Статистической гипотезой называют научное предположение о тех или иных статистических законах распределения рассматриваемых случайных величин, которое может быть проверено на основе выборки.

Гипотезы бывают:

1. Простая. Выдвигаемая гипотеза сводится к утверждению о том, что значение некоторого неизвестного параметра генеральной совокупности в точности равно заданной величине
2. Сложная. Состоит из конечного или бесконечного множества простых гипотез, при этом указывается некоторая область вероятных значений параметра.
3. Параметрическая. Гипотеза о параметрах генеральной совокупности
4. Непараметрическая. Гипотеза о распределениях.
5. Основная. Гипотеза об отсутствии различий. Обозначается H_0, содержит число 0: $X_1 * X_2 = 0$, где X_1, X_2 – составляемые значения признаков. Это то, что мы хотим опровергнуть, если перед нами стоит задача доказать значимость различий.
6. Альтернативная. Гипотеза о значимости различий. Обозначается H_1. Это то, что мы хотим доказать.

Типы гипотез:

1. Гипотезы о виде закона распределения исследуемой случайной величины.
2. Гипотезы о числовых значениях параметров исследуемой генеральной совокупности.
3. Гипотезы об общем виде модели, описывающей статистическую зависимость между признаками.
4. Гипотезы об однородности двух или нескольких выборок, или некоторых характеристик анализируемых совокупностей.

Пример 10. Тренер хочет проверить, влияет ли новая программа тренировок на среднюю частоту сердечных сокращений у спортсменов. Так как известно, что среднее значение пульса в норме составляет 82 удара в минуту, гипотезы формулируются следующим образом:

$$H_0: m = 82;\ H_1: m \neq 82$$

Основная гипотеза: средний пульс не изменился и равен 82 ударам в минуту. Альтернативная гипотеза: средний пульс изменился и не равен 82 ударам в минуту. Основная гипотеза утверждает, что среднее значение пульса осталось прежним, а альтернативная – что оно изменилось. Такая альтернативная гипотеза называется двусторонней, поскольку изменение может быть, как в сторону увеличения, так и в сторону уменьшения.

Общие формулировки трех типов рассмотренных гипотез (гипотез о среднем) представлены в таблице 4.

Таблица 4

Формулировки гипотез различного типа

Проверка правосторонней гипотезы	Проверка левосторонней гипотезы	Проверка двусторонней гипотезы
$H_0 : m = m_0$	$H_0 : m = m_0$	$H_0 : m = m_0$
$H_1 : m > m_0$	$H_1 : m < m_0$	$H_1 : m \neq m_0$

Чтобы правильно составить гипотезы, исследователю необходимо перевести вербальные предположения на язык математических формул и стандартных статистических соотношений. В таблице 5 приведен ряд выражений, часто встречающихся в задачах на проверку гипотез, и соответствующие им математические символы.

Таблица 5

Выражения, часто встречающиеся при проверке гипотез

Типы утверждений	
$>$	$<$
больше, чем длинней, чем шире, чем выше, чем лучше, чем эффективнее, чем дороже, чем	меньше, чем короче, чем уже, чем ниже, чем хуже, чем менее эффективно, чем дешевле, чем
$=$	$\neq$
равен точно соответствует точно совпадает не изменился	не равен отличается не совпадает изменился
$\geq$	$\leq$
равен точно соответствует точно совпадает	не равен отличается не совпадает

Так как проверка статистических гипотез осуществляется на основании выборочных данных, решения относительно нулевой гипотезы H_0 имеют вероятностный характер.

Следовательно, такое решение сопровождается некоторой, хотя возможно и очень малой, вероятностью ошибочного заключения как в ту, так и в другую сторону. То есть в небольшой доле случаев нулевая гипотеза H_0 может оказаться отвергнутой, в то время как в действительности в генеральной совокупности она является справедливой.

При проверке гипотез с точки зрения правильности принятого решения может осуществиться один из четырех случаев, представленных в таблице 6. С одной стороны, гипотеза H_0 в действительности, в генеральной совокупности, может быть или не быть верной

(столбец). С другой стороны, на основании выборочной совокупности может быть выбрано одно из двух решений – отклонить или принять гипотезу H_0 (строка). На «пересечении» этих возможностей и возникают 4 случая, в двух из которых принятое решение оказывается правильным, а в двух - ошибочным.

Таблица. 6

Возможные результата проверки гипотез

	H_0 верна	H_0 не верна
H_0 отклоняется	Ошибка I рода	Правильное решение
H_0 принимается	Правильное решение	Ошибка II рода

Если основная гипотеза верна, но мы ее отклоняем, тогда совершается ошибка I рода. В примере 10 это случай, когда в среднем по генеральной совокупности новая программа тренировок не вызывает серьезных изменений пульса у спортсменов, но, в силу случайного совпадения, значительно изменит его у тех из них, кто попал в выборку. В этом случае тренер отклонит основную гипотезу, тогда как на самом деле она верна. Такую ошибку **называют ошибкой 1-го рода, а ее вероятность – уровнем значимости и обозначают α.**

В какой-то небольшой доле случаев нулевая гипотеза H_0 принимается, в то время как на самом деле в генеральной совокупности она ошибочна, а справедлива альтернативная гипотеза H_1. Такую ошибку называют **ошибкой 2-го рода**. **Вероятность ошибки 2-го рода обозначается как β. Вероятность $1-\beta$ называют мощностью критерия**.

То есть, может сложиться и так, что у спортсменов в выборке новая программа тренировок существенных изменений пульса не вызовет, тогда как на самом деле, если бы были рассмотрены все спортсмены в генеральной совокупности, были бы выявлены его существенное снижение или увеличение. Однако, базируясь на выборке, тренер принимает решение – гипотезу H_0 принять (считать, что новая программа тренировок не оказывает влияния на пульс), тем самым совершая ошибку II рода.

При фиксированном объеме выборки можно выбрать по своему усмотрению величину вероятности только одной из ошибок α или β. Увеличение вероятности одной из них приводит к снижению другой.

Принято задавать вероятность ошибки 1-го рода α – уровень значимости. Как правило, пользуются некоторыми стандартными значениями уровня значимости α: 0,1; 0,05; 0,025; 0,01; 0,005; 0,001. В технических исследования обычно используется 0,05 или 5% уровень значимости. В медицинских (в том числе биологических, психологических) исследованиях – 0,01 или 1%.

В отношении свойств генеральной совокупности могут выдвигаться некоторые гипотезы о величине средней, дисперсии, характере распределения, форме и тесноте связи между переменными.

4.1. Статистический критерий принятия решения. Проверка статистических гипотез

Проверка статистических гипотез осуществляется с помощью **статистического критерия**. **Статистический критерий** – это строгое математическое правило, по которому принимается или отвергается та или иная статистическая гипотеза. Построение критерия

представляет собой выбор подходящей функции от результатов наблюдений (ряда эмпирически полученных значений признака), которая служит для выявления меры расхождения между эмпирическими значениями и гипотетическими.

Конкретная функциональная форма критерия строится для каждого типа гипотез в отдельности. Основной характеристикой критерия служит его **закон распределения**. Интерпретация критерия основана на понятии критических значений и критических областей.

Критическое значение – значение критерия, разделяющее критическую и некритическую область. Величина критического значения определяется уровнем значимости α.

Критическая область – диапазон значений критерия, при которых отличия выборочного и тестируемого значения считаются значимыми, и основная гипотеза отклоняется.

Некритическая область – диапазон значений критерия, при которых отличия выборочного и тестируемого значения полагаются случайными, и основная гипотеза не отклоняется.

Проанализируем выше сказанное на основании гипотез о среднем, из примера 9, и предположим, что величина среднего квадратического отклонения генеральной совокупности σ известна. В этом случае критерием для принятия решения является величина $z = \frac{\bar{x} - m_0}{\frac{\sigma}{\sqrt{n}}}$, где $\bar{x}$ – выборочное среднее; m_0 – предполагаемое значение генеральной средней. Для различных выборок значения z будут приближенно подчиняться закону нормального распределения $N(0; 1)$, а критическими значениями для двусторонних (односторонних) гипотез являются величины $z_{\frac{\alpha}{2}}(z_\alpha)$, определяемые по уровню значимости α.

Возьмем уровень значимости $\alpha = 0{,}01$ и осуществим проверку гипотез в примере 10. На рисунке 18 представлены критические значения, расположения критических и некритических областей. Если выборочное значение далеко отклоняется от предполагаемого, $m_0 = 82$ в большую или меньшую сторону, то значение критерия попадет в критическую область $|z| > 2{,}58$, и основная гипотеза H_0 будет отклонена в пользу альтернативной H_1. Какова вероятность при таком решении совершить ошибку первого рода, то есть отклонить основную гипотезу, когда она верна?

Предположим, что нулевая гипотеза верна и истинное значение генерального среднего равно 82. Тогда критерий z имеет нормальное распределение и вероятность его попадания в критическую область $|z| > 2{,}58 = z_{\frac{\alpha}{2}}, \alpha = 0{,}01$. Так как при попадании в критическую область основная гипотеза отклоняется, вероятность совершить ошибку первого рода также равна α. Следовательно, критическая область построена правильно.

Существует особенность, связанная с односторонними гипотезами. При проверке, например, правосторонней гипотезы критическая область слева отсутствует, то есть даже для выборочных средних намного меньших, чем m_0, все равно будет принята основная гипотеза $H_0: m = m_0$. При истинном среднем $m = m_0$ вероятность уклониться от m_0 далеко влево мала, но при любом из значений $m > m_0$ вероятность будет еще меньше. Для односторонних гипотез удобнее вести рассуждение относительно альтернативной гипотезы: если утверждение $H_1: m > m_0$, m больше m_0 – отклоняется, то принимается противоположное ему утверждение, m не больше m_0, $m \leq m_0$.

Использование одной и той же числовой характеристики при проверке гипотез и построении доверительных интервалов не случайно: между этими методами существует глубокая математическая связь.

На практике теоретические гипотезы и эмпирические наблюдения неразрывно связаны. В зависимости от последовательности действий можно выделить два подхода:

1. Если сначала выдвигается утверждение, которое нужно подтвердить данными, используется классическая проверка гипотез.
2. Если сначала собраны данные, то на их основе строится доверительный интервал. Это позволяет сразу увидеть, какие теоретические значения согласуются с экспериментом, а какие – нет.

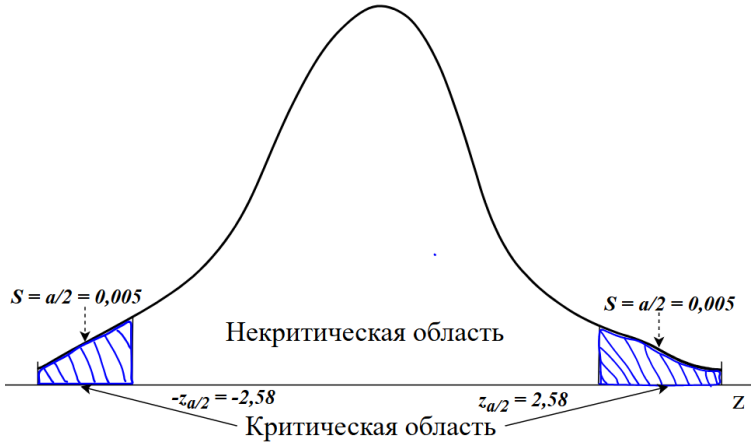


Рис. 18. Проверка двусторонней гипотезы

На рисунках 19 и 20 представлены примеры проверки правосторонней и левосторонних гипотез соответственно.



Рис. 19. Проверка правосторонней гипотезы

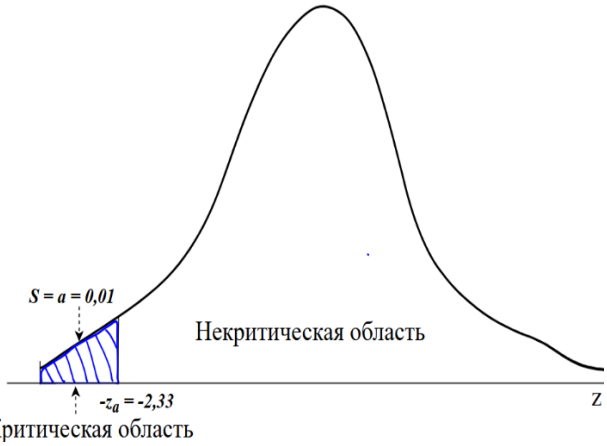


Рис. 20. Проверка левосторонней гипотезы

Приведем схему рассуждений при проверке гипотез:

1. $H_0: m = m_0;\ H_1: m \neq m_0 \rightarrow |z| > z_{\frac{\alpha}{2}}$: Да $- H_0$ отклоняется; Нет $- H_0$ принимается.
2. $H_0: m = m_0;\ H_1: m > m_0 \rightarrow |z| > z_{\alpha}$: Да $- H_0$ отклоняется; Нет $- H_0$ принимается.
3. $H_0: m = m_0;\ H_1: m < m_0 \rightarrow |z| < -z_{\alpha}$: Да $- H_0$ отклоняется; Нет $- H_0$ принимается.

Алгоритм проверки гипотез заключается в следующем:

1. По условиям задачи выбрать тип гипотез, сформулировать основную и альтернативную гипотезы.
2. Задать уровень значимости и найти критические значения для критерия, соответствующего типу гипотез.
3. Рассчитать значение критерия по данной выборке.
4. Принять решение, отклоняется или нет основная гипотеза.
5. Сделать вывод.

Пример 11. Социальный работник проверяет утверждение, что среднее время (в неделю), которое дети дошкольного возраста проводят за игрой в телефоне, меньше, чем среднее время по стране, составляющее $m = 29{,}4$ часа при среднеквадратическом отклонении $\sigma = 2$ часа. Для проверки опрошено 25 родителей, у которых дети в среднем проводят за гаджетом $\bar{x} = 27$ часов. Проверить утверждение при уровне значимости $\alpha = 0{,}01$.

Решение: Применяем левостороннюю гипотезу о среднем $\mathrm{H}_0\colon \mathrm{m} = 29{,}4;\ \mathrm{H}_1\colon \mathrm{m} < 29{,}4$. Для проверки гипотезы о среднем применяется z-критерий, для левосторонней гипотезы его критическое значение составляет $\mathrm{z}_\alpha = -2{,}33$. Находим расчетное значение критерия z по выборке:

$$z = \frac{\bar{x} - m_0}{\frac{\sigma}{\sqrt{n}}};\ m_0 = 29{,}4;\ \sigma = 2; n = 25;\ \bar{x} = 27;\ z = \frac{27 - 29{,}4}{\frac{2}{\sqrt{25}}} = -6$$

После анализа полученных результатов принимаем решение: так как $-6 < -2{,}33$, то значение критерия находится в критической области и гипотеза H_0 отклоняется, то есть принимается альтернативная гипотеза H_1.

Вывод. По предоставленным данным от родителей детей дошкольного возраста в среднем они проводят меньше времени за использованием мобильного телефона, чем в стране в целом.

Статистические критерии делятся на:

Критерии значимости. Правило проверки статистических гипотез, основанное на свойствах распределения меры отклонения эмпирической функции распределения выборки при одной гипотезе от эмпирической функции распределения при др. гипотезе.

Критерии согласия. Проверка на согласие подразумевает проверку факта о том, что исследуемая случайная величина подчиняется предполагаемому закону. Критерии согласия можно также воспринимать, как критерии значимости.

Критерии на однородность. При проверке на однородность случайные величины исследуются на факт взаимного соответствия их законов распределения (подчиняются ли эти величины одному и тому же закону). Используются в факторном (дисперсионном) анализе для определения наличия зависимостей.

Разберем на примерах применение различных статистических критериев для решения практических задач.

4.2. Использование критерия Стьюдента. Гипотеза о равенстве среднего значения числу

Пример 12. Производитель кофе утверждает, что средний вес пачки кофе составляет 100 г. Из партии извлечена выборка и взвешена. Вес каждой пачки – 98, 104, 97, 97, 101, 100,

99, 101, 99, 98. Не противоречит ли это утверждению производителя? Доверительная вероятность 99%. Вес пачек кофе распределен нормально.

Решение:

1. Вычислим показатели выборки:

$$\bar{x} = \frac{1}{n}\sum x_i n_i = \frac{994}{10} = 99{,}4$$

2. Вычислим дисперсию:

$$\sigma^2 \approx 4{,}71$$

3. Вычислим среднеквадратическое отклонение:

$$\sigma \approx 2{,}171$$

4. Расчеты в таблице:

x_i	n_i	$x_i n_i$	$(x_i - \bar{x})^2 * n_i$
97	2	194	11,52
98	2	196	3,92
99	2	198	0,32
100	1	100	0,36
101	2	202	5,12
104	1	104	21,16
Σ	10	994	42,4

5. Введем нулевую гипотезу $H_0: a = 100$ при конкурирующей гипотезе $H_1: a \neq 100$. Если $T_{\text{эмп}} < t_{\text{кр}}$, то принимаем H_0, в противном случае H_1.

6. Вычисляем наблюдаемое значение критерия:

$$T_{\text{эмп}} = \frac{(\bar{x} - a)}{\sigma}\sqrt{n} = \frac{(99{,}4 - 100)}{2{,}171}\sqrt{100} \approx -0{,}87$$

7. По таблице критических точек распределения Стьюдента найдем критическую точку по уровню значимости α = 0,01 и числу степеней свободы k = 9, откуда:

$$t_{\text{кр}} = 3{,}25$$

Вывод: Так как $T_{\text{эмп}} < t_{\text{кр}}$, то есть $-0{,}87 < 3{,}25$, то нулевую гипотезу о равенстве среднего веса кофе 100 г можно принять. Гипотеза принимается.

4.3. Использование критерия χ^2 Пирсона. Гипотеза о соответствие эмпирического закона распределения нормальному

Пример 13. Используя критерий Пирсона, при уровне значимости $\alpha = 0{,}05$ проверить, согласуется ли гипотеза о нормальном распределении генеральной совокупности X по результатам выборки:

1. Построим таблицу со вспомогательными расчетами:

x_i	n_i	$x_i * n_i$	$(\bar{x} - x_i)^2 * n_i$
0,3	7	2,1	6,461
0,5	9	4,5	5,209
0,7	28	19,6	8,805
0,9	27	24,3	3,514
1,1	30	33	0,775
1,3	26	33,8	0,040
1,5	21	31,5	1,202
1,7	25	42,5	4,823
1,9	22	41,8	8,990
2,1	9	18,9	6,339
2,3	5	11,5	5,400
Σ	209	263,5	51,558

2. Вычислим показатели выборки:

$$\bar{x} = \frac{1}{n}\sum x_i n_i = \frac{1}{209}263{,}5 \approx 1{,}261$$

$$\sigma^2 \approx 0{,}248$$

$$\sigma \approx 0{,}498$$

3. Выдвинем гипотезу H_0: распределение генеральной совокупности X подчинено нормальному закону с параметрами $m = 1{,}261$ и $\sigma = 0{,}498$. Проверим эту гипотезу по критерию Пирсона при уровне значимости $\alpha = 0{,}05$.

4. Рассчитаем теоретические частоты n_i^0 по формуле 58:

$$n_i^0 = \frac{nh}{\sigma}\varphi(u_i), \quad (58)$$

где $u_i = \frac{x_i - \bar{x}}{\sigma}$, $h = 0{,}2$ – шаг между вариантами, $\varphi(u_i) = \frac{1}{\sqrt{2\pi}} e^{\frac{-u^2}{2}}$

5. Вычисления в таблице:

x_i	u_i	$\varphi(u_i)$	n_i^0	n_i	$\frac{(n_i - n_i^0)^2}{n_i^0}$
0,3	-1,930	0,062	5,204	7	0,620
0,5	-1,528	0,124	10,422	9	0,194
0,7	-1,126	0,212	17,762	28	5,901
0,9	-0,725	0,307	25,70	27	0.060
1,1	-0,323	0,379	31,793	30	0,101

x_i	u_i	$\varphi(u_i)$	n_i^0	n_i	$\frac{(n_i - n_i^0)^2}{n_i^0}$
1,3	0,079	0,398	33,390	26	1,636
1,5	0,481	0,355	29,842	21	2,620
1,7	0,882	0,270	22,697	25	0,234
1,9	1,284	0,175	14,689	22	3,638
2,1	1,686	0,096	8,090	9	0,102
2,3	2,087	0,045	3,792	5	0,385
Σ					15,491

6. Наблюдаемое значение критерия вычислим по формуле и сразу рассчитаем значение:

$$\chi^2_{\text{эмп}} = \sum_{i=1}^{11} \frac{(n_i - n_i^0)^2}{n_i^0} = 15{,}491$$

7. По таблице критических значений $\chi^2_{\text{кр}}$ при уровне значимости $\alpha = 0{,}05$ и числе степеней свободы $k = l - 3 = 11 - 3 = 8$ найдем $\chi^2_{\text{кр}} \approx 15{,}507$. Так как $\chi^2_{\text{эмп}} < \chi^2_{\text{кр}}$ нулевую гипотезу можно принять при заданном уровне значимости.

4.4. Использование критерия χ^2 Пирсона. Гипотеза о равенстве средних двух выборок

Пример 14. Используя критерий «хи-квадрат» при уровне значимости α = 0,05, проверить, существует ли зависимость уровня интеллектуального развития учеников от типа школы по результатам обследования 100 региональных и 100 московских школьников.

1. Построим таблицу с исходными данными, посчитаем суммы по строке и по столбцу:

Тип школы	Уровень интеллектуального развития			
	низкий	нормальный	высокий	Σ
Региональная	25	50	25	100
Московская	52	41	7	100
Σ	77	91	32	200

2. Вычислим теоретические частоты. Для этого вычислим долю школьников с низким, нормальным и высоким развитием в выборке:

$$p_1 = \frac{77}{200} = 0{,}385$$

$$p_2 = \frac{91}{200} = 0{,}455$$

$$p_3 = \frac{32}{200} = 0{,}16$$

3. Теперь вычисляем теоретические частоты, зная значения долей и количество учебников в каждой школе (100 человек):

Тип школы	Уровень интеллектуального развития			
	низкий	нормальный	высокий	Σ
Региональная	38,5	45,5	16	100
Московская	38,5	45,5	16	100
Σ	77	91	32	200

Суммы по всем строкам и столбцам должны остаться те же, что подтверждает правильность расчетов.

4. Находим величину χ^2:

$$\chi^2_{\text{эмп}} = \frac{(25-38{,}5)^2}{38{,}5} + \frac{(52-38{,}5)^2}{38{,}5} + \frac{(50-45{,}5)^2}{45{,}5} + \frac{(41-45{,}5)^2}{45{,}5} \frac{(25-16)^2}{16} + \frac{(7-16)^2}{16} = 20{,}48$$

5. Критическое значение при числе степеней свободы $k = (v-1) * (c-1) = = 1 * 2 = 2$ и уровне значимости $\alpha = 0{,}05$, $\chi^2_{\text{кр}} = 6$.

6. Так как наблюдаемое значение больше критического, следует принять гипотезу H_1 о наличии различий между двумя эмпирическими распределениями.

Вывод: существует зависимость уровня интеллектуального развития учеников от типа школы (для региональной школы уровень развития выше).

4.5. Гипотеза о равенстве долей (биномиальных параметров)

Гипотезы о равенстве долей признака в двух генеральных совокупностях часто возникают тогда, когда нужно сравнить две группы и понять, отличаются ли они по какому-то качественному признаку.

Например, можно проверить такие утверждения:

1. Правда ли, что доля покупателей, выбирающих онлайн-заказ, выше среди жителей крупных городов, чем среди жителей небольших населённых пунктов?

2. Ниже ли доля пользователей, предпочитающих мобильные приложения, среди людей старшего возраста по сравнению с молодёжью?

3. Одинакова ли доля студентов и работающих взрослых, которые регулярно читают научно-популярные статьи?

Такие гипотезы помогают установить, существует ли различие между группами или наблюдаемая разница объясняется случайными колебаниями выборки.

Основная гипотеза утверждает, что доли единиц, обладающих признаком, p_1 и p_2, равны, $H_0: p_1 = p_2$. Альтернативная гипотеза может заключаться в том, что доли не равны $H_1: p_1 \neq p_2$. Первая доля больше второй, $H_1: p_1 > p_2$ или первая доля меньше второй, $H_1: p_1 < p_2$.

Решение относительно соотношения долей генеральных совокупностей принимается с помощью сравнения двух выборочных долей $w_1 = \frac{k_1}{n_1}$ и $w_2 = \frac{k_2}{n_2}$, где k_1, k_2 – число единиц, обладающих признаком; n_1, n_2 – размер выборок.

Насколько значимо отличаются выборочные доли в количественном эквиваленте позволяет z-критерий:

$$z = \frac{w_1 - w_2}{\sqrt{\overline{w}(1-\overline{w}) * \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}, \quad (59)$$

где $\overline{w} = \frac{k_1 + k_2}{n_1 + n_2}$.

Критерий приближенно подчиняется стандартному нормальному распределению при условии независимости выборок и выполнении соотношений $n * p > 5, n * (1-p) > 5$ для каждой из них.

Пример 15. Выборка из 50 квартир, в которых живут семьи с двумя работающими взрослыми, показала, что в 36% из них заказывают доставку продуктов. Аналогичное исследование среди 80 квартир, где работает только один взрослый, установило, что в этой группе доставкой продуктов пользуются только 25% опрошенных. Используя уровень значимости 0,01, установить, есть ли значимое различие между двумя группами.

Решение. Значения долей заданы в процентах, поэтому необходимо сначала перевести их в числовое выражение, рассчитав количество единиц, обладающих признаком, и вычислив среднюю долю:

$$w_1 = 0{,}36;\ k_1 = 0{,}36 * 50 = 18;\ w_2 = 0{,}25;\ k_2 = 0{,}25 * 80 = 20;\ \overline{w} = \frac{18 + 20}{50 + 80} = 0{,}292$$

Далее проводим ряд расчетов:

1. $H_0: p_1 = p_2;\ H_1: p_1 \neq p_2$
2. Находим двустороннее критическое значение для z-критерия при уровне значимости $\alpha = 0{,}01;\ z_{\frac{\alpha}{2}} = 2{,}58$. Критическая область $|z| > z_{\frac{\alpha}{2}}$.
3. Вычислим значение критерия по формуле (59):

$$z = \frac{w_1 - w_2}{\sqrt{\overline{w}(1-\overline{w}) * \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = z = \frac{0{,}36 - 0{,}25}{\sqrt{0{,}292 * (1 - 0{,}292) * \left(\frac{1}{50} + \frac{1}{80}\right)}} = 1{,}34$$

4. Принятие решения. Значение критерия не попадает в критическую область, так как $z = 1{,}34 > z_{\frac{\alpha}{2}} = 2{,}58$, нулевая гипотеза не отклоняется.

Вывод. Результаты не подтверждают предположение, что доли семей, пользующихся доставкой продуктов, различны.

КОНТРОЛЬНЫЕ ВОПРОСЫ

1. Дайте определение статистической гипотезы. Какие задачи решаются с помощью этого раздела статистики?
2. В чем различие между простой и сложной статистическими гипотезами? Приведите краткий пример для каждой.
3. Чем отличаются параметрические гипотезы от непараметрических?

4. Что такое основная (нулевая) гипотеза (H_0)? Почему она обычно формулируется как утверждение об «отсутствии различий»?

5. Что представляет собой альтернативная гипотеза (H_1)? Какие три типа альтернативных гипотез (по направленности) выделяют в статистике?

6. Дайте определение понятию «двусторонняя гипотеза». В каких случаях она используется (на примере с пульсом спортсменов из текста)?

7. Что такое ошибка I рода? Каким символом обозначается вероятность её совершения?

8. Что такое ошибка II рода? Как называется величина, равная $1 - \beta$, и что она характеризует?

9. Как связаны между собой вероятности ошибок I и II рода при фиксированном объеме выборки?

10. Что такое уровень значимости (α)? Какие стандартные значения обычно используются в технических и медицинских исследованиях?

11. Дайте определение статистического критерия. Какую роль он играет в проверке гипотез?

12. Разграничьте понятия «критическая область» и «некритическая область». Какое решение принимается при попадании значения критерия в каждую из них?

13. Что такое критическое значение критерия и от чего зависит его величина?

14. Опишите общий алгоритм проверки статистических гипотез (5 основных шагов).

15. В каких случаях для проверки гипотезы о среднем используется z-критерий, а в каких критерий Стьюдента (t-критерий)?

16. Что такое критерии согласия? Какую задачу они решают (на примере критерия χ^2 Пирсона)?

17. Для чего используются критерии на однородность? В каких видах анализа они находят применение?

18. Как определяется число степеней свободы (k) при проверке гипотезы о нормальном распределении по критерию Пирсона?

19. Какую формулу используют для проверки гипотезы о равенстве долей признака в двух генеральных совокупностях? Какие условия должны выполняться для её применения?

20. Какова математическая связь между проверкой гипотез и построением доверительных интервалов? Как последовательность действий (выдвижение утверждения и сбор данных) влияет на выбор метода?

ПРАКТИЧЕСКАЯ РАБОТА 3

Тема: Проверка статистических гипотез с использованием критериев Стьюдента, Пирсона и сравнение долей.

Цель работы: Овладеть навыками проверки гипотез о средних значениях, виде распределения, независимости признаков и равенстве выборочных долей.

Задача 1. – Проверка среднего значения с помощью t-критерия Стьюдента.

Задача 2. – Проверка согласия распределения с нормальным законом по критерию Пирсона.

Задача 3. – Проверка независимости признаков с помощью критерия Пирсона.

Задача 4. – Сравнение двух долей.

Вариант 1

Задача 1. На предприятии производится контроль объема бутылок питьевой воды, заявленный объем каждой бутылки должен составлять 0,5 л. Для проверки качества случайно отобрали 5 бутылок и измерили фактический объем: 0,48 л, 0,51 л, 0,49 л, 0,47 л, 0,50 л. Требуется при уровне значимости 0,05 проверить, можно ли считать, что средний объем бутылок соответствует заявленной норме.

Задача 2. В исследовании были получены данные о росте 100 саженцев одного и того же вида деревьев. Измерения сгруппированы по интервалам от 10 до 50 см. Необходимо проверить при уровне значимости 0,05, можно ли считать, что распределение роста саженцев подчиняется нормальному закону.

Задача 3. Проводилось социологическое исследование, в ходе которого изучалось, зависит ли выбор напитка от пола человека. Были опрошены 50 мужчин и 50 женщин, и для каждого респондента зафиксировано, предпочитает он чай или кофе. Требуется проверить, существует ли статистически значимая связь между полом и выбором напитка.

Задача 4. В группе А, состоящей из 100 человек, 40% респондентов курят. В группе Б, состоящей из 120 человек, курят 30% опрошенных. Необходимо при уровне значимости 0,01 проверить, является ли различие между долями курящих в двух группах статистически значимым.

Вариант 2

Задача 1. На производстве детали должны собираться в среднем за 15 минут. Для проверки этого показателя была взята случайная выборка из 6 деталей, для которых время сборки составило 14, 16, 15, 17, 13 и 15 минут. Требуется при уровне значимости 0,01 установить, соответствует ли среднее время сборки установленному нормативу.

Задача 2. На предприятии изучались ошибки при наборе текста у 150 операторов. Для каждого сотрудника было зафиксировано количество ошибок за определенный период, после чего данные были сгруппированы по интервалам. Требуется проверить, можно ли считать, что распределение количества ошибок подчиняется нормальному закону при уровне значимости 0,05.

Задача 3. Исследуется, зависит ли наличие ошибок в отчетах от стажа работы сотрудников. Для этого работников разделили на две группы: со стажем до 5 лет и со стажем более 5 лет. Для каждой группы было подсчитано число сотрудников, допускающих ошибки, и число сотрудников без ошибок. Необходимо определить, есть ли статистически значимая связь между стажем и наличием ошибок.

Задача 4. Среди 60 студентов 50% сдали тест с первого раза, а среди 80 школьников – 45%. Нужно при уровне значимости 0,05 выяснить, можно ли считать различие в долях успешно сдавших тест статистически значимым.

Вариант 3

Задача 1. Заявленная жирность творога составляет 9%. Для проверки качества были случайно отобраны 6 упаковок, в которых фактическая жирность оказалась равной 8,8%, 9,1%, 8,9%, 9,0%, 8,7% и 9,2%. Требуется при уровне значимости 0,05 проверить, можно ли считать, что средняя жирность соответствует заявленной.

Задача 2. В родильном доме в течение месяца были собраны данные о массе тела новорожденных. Всего исследовано 200 наблюдений. Данные распределены по интервалам массы тела. Необходимо проверить, можно ли считать это распределение нормальным при уровне значимости 0,05.

Задача 3. Изучается, существует ли связь между цветом автомобиля и участием в дорожно-транспортных происшествиях. В исследовании участвовали автомобили двух цветов, например, белые и черные. Для каждой группы подсчитано количество машин, попавших в ДТП, и количество машин без ДТП. Требуется проверить наличие статистически значимой зависимости между цветом автомобиля и фактом участия в ДТП.

Задача 4. В одном районе 15% жителей высказались против строительства нового объекта, а в другом районе – 20%. В каждом районе опрошено по 200 человек. Нужно при уровне значимости 0,01 проверить, есть ли статистически значимое различие между долями противников строительства.

Вариант 4

Задача 1. Срок службы лампочки по паспорту составляет 1000 часов. Для проверки были взяты 5 лампочек, и их фактический срок службы составил 950, 1020, 980, 990 и 1010 часов. Требуется при уровне значимости 0,05 установить, отличается ли средний срок службы от заявленного.

Задача 2. В одной из школ были собраны результаты ЕГЭ по математике 120 учащихся. Баллы сгруппированы по интервалам. Необходимо проверить, можно ли считать, что распределение результатов экзамена подчиняется нормальному закону.

Задача 3. Проводится исследование, связана ли особенность питания человека, например мясное или вегетарианское питание, с уровнем артериального давления. Для каждой группы питания подсчитано число людей с нормальным, повышенным и пониженным давлением. Требуется установить, есть ли связь между видом питания и уровнем давления.

Задача 4. На двух станках проверяли качество продукции. На станке №1 доля брака составила 4% из 500 изделий, а на станке №2 – 6% из 400 изделий. Требуется проверить, является ли различие в долях бракованных изделий статистически значимым.

Вариант 5

Задача 1. По технологии содержание сахара в напитке должно составлять 12 г. Из партии были случайно отобраны 5 образцов, в которых содержание сахара оказалось 11,5 г, 12,5 г, 12,0 г, 11,8 г и 12,2 г. Нужно при уровне значимости 0,01 проверить, соответствует ли среднее содержание сахара установленной норме.

Задача 2. Изучалось время реакции водителей на светофор. Всего было обследовано 180 человек, и полученные данные сгруппированы по временным интервалам. Требуется проверить при уровне значимости 0,05, можно ли считать распределение времени реакции нормальным.

Задача 3. Исследуется, связан ли выбор вуза – технического или гуманитарного профиля – с успеваемостью в школе. Для этого были собраны данные о выпускниках с разной школьной успеваемостью и их предпочтениях при выборе направления обучения. Необходимо проверить, существует ли статистически значимая зависимость между этими признаками.

Задача 4. В опросе 100 мужчин и 150 женщин выяснилось, что 60% мужчин одобряют новый закон, а среди женщин такую позицию поддерживают 70%. Требуется при уровне

значимости 0,05 проверить, существует ли значимое различие между долями одобряющих в двух группах.

Вариант 6

Задача 1. В холодильной камере температура должна составлять −18° С. Для контроля измерили температуру в 5 точках: −17° С, −19° С, −18° С, −17,5° С и −18,5° С. Необходимо при уровне значимости 0,05 проверить, соответствует ли средняя температура установленной норме.

Задача 2. Проводилось исследование глубины протектора шин после пробега 20 тысяч километров. Было обследовано 150 автомобилей, а результаты измерений сгруппированы по интервалам. Требуется проверить, подчиняется ли распределение глубины протектора нормальному закону.

Задача 3. Исследуется, существует ли связь между знаком зодиака и уровнем дохода человека, разделенного на две категории: высокий и низкий. Данные представлены в виде таблицы частот. Необходимо проверить, есть ли статистически значимая зависимость между указанными признаками.

Задача 4. В городе А безработными являются 12% жителей, а в городе Б – 15%. В каждой выборке опрошено по 300 человек. Требуется при уровне значимости 0,01 проверить, различаются ли доли безработных в двух городах статистически значимо.

Вариант 7

Задача 1. По чертежу длина болта должна составлять 50 мм. Для проверки качества была взята выборка из 7 болтов, длины которых составили 49 мм, 51 мм, 50 мм, 49 мм, 52 мм, 50 мм и 50 мм. Требуется при уровне значимости 0,05 установить, соответствует ли средняя длина болтов заданному стандарту.

Задача 2. В офисе у 200 сотрудников был измерен уровень IQ. Полученные результаты сгруппированы по интервалам. Необходимо проверить при уровне значимости 0,05, можно ли считать распределение IQ нормальным.

Задача 3. Проводится исследование, влияет ли наличие высшего образования на использование онлайн-банкинга. В выборке имеются данные о людях с высшим и без высшего образования, а также о том, пользуются ли они интернет-банком. Требуется проверить статистическую зависимость между уровнем образования и использованием онлайн-банкинга.

Задача 4. В группе владельцев iPhone 22% респондентов заявили, что купят новую модель, а среди владельцев Android – 18% готовы перейти на неё. В каждой группе опрошено по 100 человек. Нужно проверить, есть ли статистически значимое различие между долями.

Вариант 8

Задача 1. Нормативное давление в шинах автомобиля должно составлять 2,2 бар. Для контроля были измерены 4 колеса, и получены значения 2,1 бар, 2,3 бар, 2,2 бар и 2,0 бар. Требуется при уровне значимости 0,01 проверить, отличается ли среднее давление от нормы.

Задача 2. Изучалась урожайность пшеницы на различных участках поля. Всего было выполнено 100 измерений, и результаты распределены по интервалам урожайности. Необходимо проверить, можно ли считать это распределение нормальным при уровне значимости 0,05.

Задача 3. Проводится исследование, связано ли количество осадков с урожаем грибов. Для каждого района зафиксировано, был ли урожай грибов высоким или низким, и сопоставлены эти данные с количеством выпавших осадков. Требуется определить, существует ли статистическая зависимость между осадками и урожаем грибов.

Задача 4. В 2022 году 30% семей имели автомобиль, а в 2023 году – 33%. В каждой выборке исследовано по 500 семей. Нужно при уровне значимости 0,05 проверить, можно ли считать различие в долях статистически значимым.

Вариант 9

Задача 1. Вес плитки шоколада по стандарту должен составлять 90 г. Для контроля качества были взяты 5 плиток, и фактический вес составил 88 г, 91 г, 89 г, 90 г и 87 г. Требуется при уровне значимости 0,05 установить, соответствует ли средний вес заявленной норме.

Задача 2. Изучалось опоздание поездов в минутах. Было проанализировано 250 рейсов, а данные сгруппированы по интервалам времени опоздания. Необходимо проверить при уровне значимости 0,05, можно ли считать распределение опозданий нормальным.

Задача 3. Исследуется, влияет ли просмотр рекламы на покупку товара. Для этого были выделены группы людей, которые видели рекламу и не видели ее, а затем подсчитано количество совершивших покупку. Требуется проверить, существует ли статистически значимая связь между просмотром рекламы и покупкой товара.

Задача 4. В сельской местности 45% жителей поддерживают кандидата A, а в городе – 52%. В селе опрошено 200 человек, в городе – 300 человек. Необходимо при уровне значимости 0,05 проверить, есть ли значимое различие между долями поддержки.

Вариант 10

Задача 1. Кислотность почвы на участке должна быть равна pH 7,0. Было отобрано 5 проб почвы, и измеренные значения составили 6,8; 7,2; 7,0; 6,9 и 7,1. Требуется при уровне значимости 0,01 проверить, можно ли считать, что среднее значение pH соответствует норме.

Задача 2. Изучалось количество звонков в колл-центр по часам в течение 100 дней. Полученные данные сгруппированы по интервалам. Необходимо проверить, подчиняется ли распределение числа звонков нормальному закону.

Задача 3. Исследуется, связан ли тип климата с долголетием жителей. Для нескольких территорий с разным климатом были собраны данные о количестве жителей, относящихся к категории долгожителей, и количестве остальных жителей. Требуется проверить, существует ли статистически значимая связь между типом климата и долголетием.

Задача 4. В партии X 10% ламп перегорели в первый месяц эксплуатации, а в партии Y – 14%. В каждой партии было по 200 ламп. Требуется при уровне значимости 0,05 проверить, различаются ли доли отказов статистически значимо.

Вариант 11

Задача 1. Толщина бумаги по стандарту должна составлять 0,1 мм. Для контроля качества были измерены 5 листов бумаги, и получены значения 0,11 мм, 0,09 мм, 0,10 мм, 0,12 мм и 0,08 мм. Требуется при уровне значимости 0,05 проверить, соответствует ли средняя толщина бумаги установленному нормативу.

Задача 2. Изучался расход топлива на 100 км у 150 одинаковых автомобилей. Все полученные данные были распределены по интервалам значений. Необходимо проверить при уровне значимости 0,05, можно ли считать распределение расхода топлива нормальным.

Задача 3. Проводится исследование, зависит ли выбор спортивной секции ребенком от его темперамента. Для групп детей с разными типами темперамента зафиксировано, в какую секцию они ходят. Требуется установить, существует ли статистически значимая зависимость между темпераментом и выбором секции.

Задача 4. Среди 80 пенсионеров интернетом пользуются 20%, а среди 100 подростков – 98%. Требуется при уровне значимости 0,05 проверить, существует ли статистически значимое различие между долями пользователей интернета в двух группах.

Вариант 12

Задача 1. Содержание витамина С в соке по норме должно составлять 30 мг. Для проверки качества были отобраны 5 образцов, в которых содержание витамина С составило 28 мг, 32 мг, 29 мг, 31 мг и 30 мг. Требуется при уровне значимости 0,01 проверить, можно ли считать, что среднее содержание витамина С соответствует заявленной норме.

Задача 2. Изучалось время обслуживания одного клиента в банке. Для 200 клиентов были зафиксированы времена обслуживания, после чего данные сгруппировали по интервалам. Необходимо проверить, соответствует ли распределение времени обслуживания нормальному закону.

Задача 3. Исследуется, влияет ли уровень освещенности в офисе на продуктивность работников. Сотрудников разделили по условиям освещения, а затем оценили их производительность. Требуется проверить, существует ли статистически значимая связь между освещенностью и продуктивностью.

Задача 4. В выборке из 50 семей, где работают двое взрослых, 36% имеют ипотеку. В другой выборке из 80 семей, где работает только один взрослый, ипотека есть у 25% семей. Требуется при уровне значимости 0,01 проверить, различаются ли эти доли статистически значимо.

5. ДИСПЕРСИОННЫЙ АНАЛИЗ

Дисперсионный анализ, предложен Р. Фишером. Предназначен для **выявления влияния ряда отдельных факторов на результаты экспериментов**.

В основе лежит предположение о том, что одни переменные могут рассматриваться как причины (факторы, независимые переменные), а другие как следствия (зависимые переменные). Независимые переменные называют иногда регулируемыми факторами именно потому, что в эксперименте исследователь имеет возможность варьировать ими и анализировать получающийся результат.

Сущность дисперсионного анализа заключается в расчленении общей дисперсии изучаемого признака на отдельные компоненты, обусловленные влиянием конкретных факторов, и проверке гипотез о значимости влияния этих факторов на исследуемый признак. Сравнивая компоненты дисперсии друг с другом посредством F-критерия Фишера, можно определить, какая доля общей вариативности результативного признака обусловлена действием регулируемых факторов.

Исходным материалом для дисперсионного анализа служат данные исследования трех и более выборок, которые могут быть как **равными**, так и **неравными** по численности, как **связными**, так и **несвязными**.

По количеству выявляемых регулируемых факторов дисперсионный анализ может быть **однофакторным** (при этом изучается влияние одного фактора на результаты эксперимента), **двухфакторным** (при изучении влияния двух факторов) и **многофакторным** (позволяет оценить не только влияние каждого из факторов в отдельности, но и их взаимодействие)**.**

Дисперсионный анализ относится к группе **параметрических методов** и поэтому его следует применять только тогда, когда доказано, что **распределение** является **нормальным**.

5.1. Однофакторный дисперсионный анализ для несвязанных выборок

Изучается действие только одной переменной (фактора) на исследуемый признак (Рис. 21).

Исследователя интересует вопрос, как изменяется определенный признак в разных условиях действия переменной (фактора). Например, как изменяется время решения задачи при разных условиях мотивации испытуемых (низкой, средней, высокой мотивации) или при разных способах предъявления задачи (устно, письменно или в виде текста с графиками и иллюстрациями), в разных условиях работы с задачей (в одиночестве, в комнате с преподавателем, в классе). В первом случае фактором является мотивация, во втором – степень наглядности, в третьем – фактор публичности.

В данном варианте метода влиянию каждой из градаций подвергаются разные выборки испытуемых. Градаций фактора должно быть не менее **трех**.

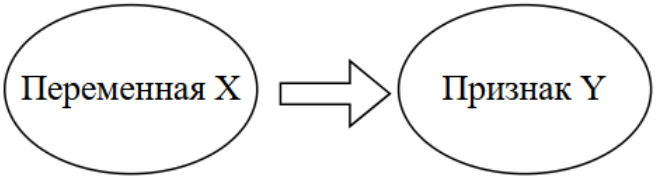


Рис. 21. Влияние переменной, *X* на признак, *Y*

Рассмотрим использование однофакторного дисперсионного анализа на примере.

Пример 16. Три различные группы из шести испытуемых получили списки из десяти слов. Первой группе слова предъявлялись с низкой скоростью –1 слово в 5 секунд, второй группе со средней скоростью – 1 слово в 2 секунды, и третьей группе с большой скоростью – 1 слово в секунду. Было предсказано, что показатели воспроизведения будут зависеть от скорости предъявления слов. Исходные данные представлены в таблице 7.

Таблица 7

Количество воспроизведенных слов

№ испытуемого	Группа 1 Низкая скорость	Группа 2 Средняя скорость	Группа 3 Высокая скорость
1	8	7	4
2	7	8	5
3	9	5	3
4	5	4	6
5	6	6	2
6	8	7	4

Далее необходимо произвести дополнительные расчеты, посчитав суммы и средние значения по столбам таблицы и вычислив общую сумму по столбам таблицы (таблица 7.1).

Таблица 7.1

Количество воспроизведенных слов

№ испытуемого	Группа 1 Низкая скорость	Группа 2 Средняя скорость	Группа 3 Высокая скорость
1	8	7	4
2	7	8	5
3	9	5	3
4	5	4	6
5	6	6	2
6	8	7	4
Суммы Σ	43	37	24
Средние $\bar{x}$	7,17	6,17	4,00
Общая сумма $\Sigma_{общ}$	104		

Формируем гипотезы:

H_0: различия в объеме воспроизведения слов между группами являются не более выраженными, чем случайные различия внутри каждой группы

H_1: различия в объеме воспроизведения слов между группами являются более выраженными, чем случайные различия внутри каждой группы.

Алгоритм решения задачи:

1. Рассчитаем $SS_{факт}$ – вариативность признака, обусловленную действием исследуемого фактора.

Часто встречающееся обозначение SS – сокращение от «суммы квадратов» (sum of squares), используется в переводных источниках (см., например: Гласс Дж., Стенли Дж., 1976).

$$SS_{\text{факт}} = \frac{\sum T_c^2}{n} - \frac{(\sum x_i)^2}{N}, \tag{60}$$

где T_c – сумма индивидуальных значений по каждому из условий. Для нашего примера 43, 37, 24 (см. табл. 7.1);

c – количество условий (градаций) фактора ($= 3$);

n – количество испытуемых в каждой группе ($= 6$);

N – общее количество индивидуальных значений ($= 18$);

$(\sum x_i)^2$ – квадрат общей суммы индивидуальных значений ($104^2 = 10816$).

По формуле 60 подсчитаем:

$$SS_{\text{факт}} = \frac{\sum T_c^2}{n} - \frac{(\sum x_i)^2}{N} = \frac{(43^2 + 37^2 + 24^2)}{6} - \frac{104^2}{18} = 31{,}44$$

2. Рассчитаем общую вариативность признака $SS_{\text{общ}}$ (формула 61):

$$SS_{\text{общ}} = \frac{\sum {x_i}^2 - (\sum x_i)^2}{N} = \frac{8^2+7^2+9^2+5^2+6^2+8^2+7^2+8^2+\cdots+2^2+4^2-102^2}{18} = 63{,}11, \tag{61}$$

3. Случайная (остаточная) величину $SS_{\text{сл}}$, обусловленную неучтенными факторами вычисляется по формуле (62):

$$SS_{\text{сл}} = SS_{\text{общ}} - SS_{\text{факт}} = 63{,}11 - 31{,}44 = 31{,}67, \tag{62}$$

4. Число степеней свободы рассчитывается для фактического, общего и остаточного значения (формула 63):

$$\begin{gathered} k_{\text{факт}} = k_1 = c - 1 = 3 - 1 = 2 \\ k_{\text{общ}} = N - 1 = 18 - 1 = 17, \\ k_{\text{сл}} = k_2 = k_{\text{общ}} - k_{\text{факт}} = 17 - 2 = 15 \end{gathered} \tag{63}$$

5. Далее необходимо вычислить «средний квадрат» или математическое ожидание суммы квадратов, усредненная величина соответствующих сумм квадратов SS:

$$MS_{\text{факт}} = \frac{SS_{\text{факт}}}{k_{\text{факт}}} = \frac{31{,}44}{2} = 15{,}72, \tag{64}$$

$$MS_{\text{сл}} = \frac{SS_{\text{сл}}}{k_{\text{сл}}} = \frac{31{,}67}{15} = 2{,}11, \tag{65}$$

6. Значение статистики критерия $F_{\text{эмп}}$ рассчитаем по формуле (66):

$$F_{\text{эмп}} = \frac{MS_{\text{факт}}}{MS_{\text{сл}}} = \frac{15{,}72}{2{,}11} = 7{,}45, \tag{66}$$

7. Определим $F_{\text{кр}}$ по статистическим таблицам для $df_1 = k_1 = 2$ и $df_2 = k_2 = 15$, табличное значение статистики равно 3,68.

8. Если $F_{\text{эмп}} < F_{\text{кр}}$, то нулевая гипотеза принимается, в противном случае принимается альтернативная гипотеза. Для нашего примера $F_{\text{эмп}} < F_{\text{кр}}$ (0,44 < 3,68), следовательно принимается основная гипотеза H_0.

Вывод: различия в объеме воспроизведения слов между группами являются более выраженными, чем случайные различия внутри каждой группы ($p<0,05$). Таким образом, скорость предъявления слов не влияет на объем их воспроизведения.

5.2. Дисперсионный анализ для связанных выборок

Метод дисперсионного анализа для связанных выборок применяется в тех случаях, когда исследуется влияние разных градаций фактора или разных условий на одну и ту же выборку испытуемых. Градаций фактора должно быть не менее трех.

В данном случае различия между испытуемыми – возможный самостоятельный источник различий. Однофакторный дисперсионный анализ для связанных выборок позволит определить, что перевешивает – тенденция, выраженная кривой изменения фактора, или индивидуальные различия между испытуемыми. Фактор индивидуальных различий может оказаться более значимым, чем фактор изменения экспериментальных условий.

Разберем применение однофакторного дисперсионного анализа для связных выборок на примере, предложенным Е. В. Сидоренко (1984 г.).

Пример 17. Группа из 5 испытуемых была обследована с помощью трех экспериментальных заданий, направленных на изучение интеллектуальной, настойчивости. Каждому испытуемому индивидуально предъявлялись последовательно три одинаковые анаграммы: четырехбуквенная, пятибуквенная и шестибуквенная. Можно ли считать, что фактор длины анаграммы влияет на длительность попыток ее решения?

В таблице 8 представлены исходные данные о длительности решения анаграмм, в секундах.

Таблица 8

Длительность решения анаграмм (сек)

Код исп-го	Условие 1 4х буквенная анаграмма	Условие 2 5ти буквенная анаграмма	Условие 3 6ти буквенная анаграмма
1	5	235	7
2	7	604	20
3	2	93	5
4	2	171	8
5	35	141	7

Произведем вспомогательные расчеты, посчитав суммы по строкам и суммы по столбцам (таблица 8.1).

Код исп-го	Условие 1 4х буквенная анаграмма	Условие 2 5ти буквенная анаграмма	Условие 3 6ти буквенная анаграмма	Суммы по строкам
1	5	235	7	247
2	7	604	20	631

Код исп-го	Условие 1 4х буквенная анаграмма	Условие 2 5ти буквенная анаграмма	Условие 3 6ти буквенная анаграмма	Суммы по строкам
3	2	93	5	100
4	2	171	8	181
5	35	141	7	183
Суммы по столбцам	51	1244	47	1342

Сформулируем два набора гипотез:

$H_0(A)$: Различия в длительности попыток решения анаграмм разной длины являются не более выраженными, чем различия, обусловленные случайными причинами.

$H_1(A)$: Различия в длительности попыток решения анаграмм разной длины являются более выраженными, чем различия, обусловленные случайными причинами.

Набор Б.

H_0(Б): Индивидуальные различия между испытуемыми являются не более выраженными, чем различия, обусловленные случайными причинами.

H_1(Б): Индивидуальные различия между испытуемыми являются более выраженными, чем различия, обусловленные случайными причинами.

Последовательность операций в однофакторном дисперсионном анализе для связанных выборок представлена ниже:

1. $SS_{\text{факт}}$ – вариативность признака, обусловленную действием исследуемого фактора по формуле (60):

$$SS_{\text{факт}} = \frac{\sum T_c^2}{n} - \frac{(\sum x_i)^2}{N} = \frac{(51^2 + 1244^2 + 47^2)}{5} - \frac{1342^2}{15} = 190405$$

где T_c – сумма индивидуальных значений по каждому из условий. Для нашего примера 51, 1244, 47 (см. табл. 8.1);

c – количество условий (градаций) фактора $(= 3)$;

n – количество испытуемых в каждой группе $(= 5)$;

N – общее количество индивидуальных значений $(= 15)$;

$(\sum x_i)^2$ – квадрат общей суммы индивидуальных значений $(= 1342^2 = 1800964)$.

2. Рассчитываем вариативность признака, обусловленную индивидуальными значения испытуемых $SS_{\text{исп}}$:

$$SS_{\text{исп}} = \frac{\sum {T_u}^2}{c} - \frac{(\sum x_i)^2}{N} = \frac{247^2 + 631^2 + 100^2 + 181^2 + 183^2}{3} - \frac{1342^2}{15} = 58409$$

где T_u – сумма индивидуальных значений по каждому испытуемому. Для нашего примера 247, 631, 100, 181, 183 (см. табл. 8.1); c – количество условий (градаций) фактора $(= 3)$; N – общее количество индивидуальных значений $(= 15)$.

3. Рассчитываем общую вариативность признака $SS_{\text{общ}}$:

$$SS_{\text{общ}} = \frac{\sum {x_i}^2 - (\sum x_i)^2}{N} = \frac{5^2 + 7^2 + 2^2 + 2^2 + 35^2 + 235^2 + 604^2 + 93^2 + \cdots + 8^2 + 7^2 - 1342^2}{15} = 88084$$

4. Случайная (остаточная) величина $SS_{\text{сл}}$, обусловленную неучтенными факторами вычисляется по формуле 62 с добавлением вариативности признака, обусловленной индивидуальными значениями испытуемых $SS_{\text{исп}}$:

$$SS_{\text{сл}} = SS_{\text{общ}} - SS_{\text{факт}} - SS_{\text{исп}} = 359642 - 190001 - 58409 = 111232$$

5. Число степеней свободы равно (63):

$$k_{\text{факт}} = k_1 = c - 1 = 3 - 1 = 2$$

$$k_{\text{общ}} = N - 1 = 15 - 1 = 14$$

$$k_{\text{исп}} = k_3 = n - 1 = 5 - 1 = 4$$

$$k_{\text{сл}} = k_2 = k_{\text{общ}} - k_{\text{факт}} - k_{\text{исп}} = 14 - 2 - 4 = 8$$

6. «Средний квадрат» или математическое ожидание суммы квадратов, усредненная величина соответствующих сумм квадратов *SS* равна (64, 65):

$$MS_{\text{факт}} = \frac{SS_{\text{факт}}}{k_{\text{факт}}} = \frac{190001}{2} = 95000$$

$$MS_{\text{сл}} = \frac{SS_{\text{сл}}}{k_{\text{сл}}} = \frac{110827}{8} = 13853{,}4$$

$$MS_{\text{исп}} = \frac{SS_{\text{исп}}}{k_{\text{исп}}} = \frac{58409}{4} = 14602{,}2$$

7. Значение статистики критерия $F_{\text{эмп}}$ рассчитаем по формуле (66):

$$F_{\text{эмп_факт}} = \frac{MS_{\text{факт}}}{MS_{\text{сл}}} = \frac{95000}{13853{,}4} = 6{,}872$$

$$F_{\text{эмп_исп}} = \frac{MS_{\text{исп}}}{MS_{\text{сл}}} = \frac{14602{,}5}{13853{,}4} = 1{,}054$$

8. Определяем $F_{\text{кр}}$ по статистическим таблицам для $df_1 = k_1 = 2$ и $df_2 = k_2 = 8$ табличное значение статистики $F_{\text{кр}_{\text{факт}}} = 4{,}46$, и для $df_3 = k_3 = 4$ и $df_2 = k_2 = 8$, $F_{\text{кр}_{\text{исп}}} = 3{,}84$.

9. $F_{\text{эмп_факт}} > F_{\text{кр}_{\text{факт}}}(6{,}872 > 4{,}46)$ – принимается альтернативная гипотеза H_1.

10. $F_{\text{эмп_исп}} < F_{\text{кр}_{\text{исп}}}(1{,}054 < 3{,}84)$ – принимается нулевая гипотеза H_0.

Вывод: различия в объеме воспроизведения слов в разных условиях являются более выраженными, чем различия, обусловленные случайными причинами ($p < 0{,}05$). Индивидуальные различия между испытуемыми являются не более выраженными, чем различия, обусловленные случайными причинами.

КОНТРОЛЬНЫЕ ВОПРОСЫ

1. Что такое дисперсионный анализ и для чего он применяется?
2. Кто предложил дисперсионный анализ?
3. В чем состоит сущность дисперсионного анализа?
4. Какие переменные в дисперсионном анализе называют независимыми, а какие зависимыми?
5. Что понимают под регулируемыми факторами?

6. Какие данные служат исходным материалом для дисперсионного анализа?

7. Какие выборки могут использоваться в дисперсионном анализе: равные или неравные, связные или несвязные?

8. По какому признаку дисперсионный анализ делится на однофакторный, двухфакторный и многофакторный?

9. Чем отличается однофакторный дисперсионный анализ от двухфакторного?

10. Что позволяет оценить многофакторный дисперсионный анализ?

11. Почему дисперсионный анализ относится к параметрическим методам?

12. При каком условии можно применять дисперсионный анализ?

13. В каких случаях используется однофакторный дисперсионный анализ для несвязанных выборок?

14. Почему в однофакторном анализе для несвязанных выборок число градаций фактора должно быть не менее трех?

15. Что показывает величина SSфакт?

16. Что показывает величина SSобщ?

17. Как определяется SSсл в однофакторном дисперсионном анализе?

18. Что означают степени свободы kфакт, kобщ и kсл?

19. Как вычисляется статистика Fэмп и с чем она сравнивается?

20. В чем состоит отличие дисперсионного анализа для связанных выборок от анализа для несвязанных выборок?

ПРАКТИЧЕСКАЯ РАБОТА 4

Инструкция к решению для всех вариантов:

1. Сформулируйте нулевую H_0 и альтернативную H_1 гипотезы.
2. Рассчитайте суммы квадратов $SS_{\text{факт}}$, $SS_{\text{общ}}$, $SS_{\text{сл}}$ и $SS_{\text{исп}}$ (для связанных выборок).
3. Определите число степеней свободы k.
4. Вычислите средние квадраты MS.
5. Найдите эмпирическое значение $F_{\text{эмп}}$.
6. Сравните с критическим значением $F_{\text{кр}}$ из таблицы Фишера и сделайте вывод.

Однофакторный дисперсионный анализ для несвязанных выборок. В этих задачах в каждой группе (градации фактора) участвуют разные люди/объекты.

Вариант 1

Тема: Сельское хозяйство.

Изучается влияние трех видов удобрений на урожайность пшеницы. Сформированы 3 группы участков земли по 25 штук в каждой ($N = 75$).

- Фактор: Тип удобрения (A, B, C).
- Задача: Сравнить средний вес зерна с участка и определить, влияет ли тип удобрения на урожайность.

Вариант 2

Тема: Образование.

Методист проверяет эффективность трех методик обучения скорочтению. Привлечены 3 группы студентов по 30 человек ($N = 90$).

- Фактор: Методика обучения (Традиционная, Игровая, Компьютерная).
- Задача: Проверить, значимы ли различия в количестве слов в минуту после курса обучения.

Вариант 3

Тема: Маркетинг.

Магазин тестирует три дизайна упаковки товара. Данные собраны по объему продаж в 80 различных точках (распределены на группы 25, 25 и 30).

- Фактор: Цвет упаковки (Красный, Синий, Зеленый).
- Задача: Выяснить, влияет ли цвет упаковки на количество покупок.

Вариант 4

Тема: Спортивная медицина.

Исследовалось время восстановления пульса после нагрузки у атлетов в зависимости от типа диеты. Опрошено 72 спортсмена (3 группы по 24 человека).

- Фактор: Тип питания (Вегетарианское, Высокобелковое, Сбалансированное).
- Задача: Влияет ли диета на скорость восстановления организма.

Вариант 5

Тема: Социальная психологи.

Изучается уровень стресса у сотрудников в зависимости от типа планировки офиса. В исследовании участвуют 90 сотрудников (по 30 в каждом типе офиса).

- Фактор: Организация пространства (Open-space, Кабинетная система, Смешанная).
- Задача: Проверить гипотезу о влиянии планировки на показатели стресса.

Вариант 6

Тема: Технический контроль.

На заводе сравнивают срок службы ламп, изготовленных на четырех разных линиях. Протестировано 100 ламп (по 25 с каждой линии).

- Фактор: Номер производственной линии (1, 2, 3, 4).
- Задача: Определить, влияет ли номер линии на долговечность изделия.

Однофакторный дисперсионный анализ для связанных выборок. В этих задачах одни и те же испытуемые проходят через все условия эксперимента.

Вариант 7

Тема: Фармакология.

Группа из 30 добровольцев принимает препарат для улучшения памяти. Замеры проводятся трижды.

- Условия: 1. До приема; 2. Через 1 час; 3. Через 5 часов.
- Задача: Выяснить, меняется ли количество запоминаемых чисел под действием препарата.

Вариант 8

Тема: Эргономика.

15 операторов работают на трех разных типах клавиатур (каждый работает на всех по очереди).

- Условия: Механическая, Мембранная, Сенсорная клавиатуры.
- Задача: Определить, влияет ли тип клавиатуры на количество совершаемых ошибок на 1000 знаков.

Вариант 9

Тема: Психофизиология.

Исследуется время реакции 20 водителей на световой сигнал при разном уровне шума в кабине.

- Условия: Тишина, Радио (тихо), Громкий шум машин.
- Задача: Вычислить, является ли уровень шума фактором, снижающим скорость реакции.

Вариант 10

Тема: Менеджмент.

25 менеджеров оценивают свою вовлеченность в работу при трех разных системах мотивации.

- Условия: Только оклад, Оклад + премия, Оклад + ДМС + обучение.
- Задача: Проверить, какая система мотивации дает статистически значимый прирост вовлеченности.

Вариант 11

Тема: Косметология.

Исследуется увлажненность кожи у 35 женщин после применения средства через разные интервалы времени.

- Условия: Сразу после нанесения, через 4 часа, через 12 часов.
- Задача: Влияет ли время с момента нанесения на показатель увлажненности.

Вариант 12

Тема: Обучение персонала 10 сотрудников службы поддержки проходят оценку качества речи (в баллах) трижды за год.

- Условия: Начало года, Середина года (после тренинга), Конец года.
- Задача: Оценить влияние фактора времени и обучения на профессиональные навыки сотрудников.

6. КОРРЕЛЯЦИОННЫЙ АНАЛИЗ

6.1. Понятие корреляционной связи

Зависимость между переменными величинами называется корреляционной, или корреляцией.

Корреляционная связь – это согласованное изменение двух признаков, отражающее тот факт, что изменчивость одного признака находится в соответствии с изменчивостью другого. **Корреляционные связи** – это вероятностные изменения, которые можно изучать только на представительных выборках методами математической статистики. Оба термина, **корреляционная связь и корреляционная зависимость**, часто используются как синонимы. Зависимость подразумевает влияние, связь – любые согласованные изменения, которые могут объясняться сотнями причин.

Корреляционные связи не могут рассматриваться как свидетельство причинно-следственной зависимости, они свидетельствуют лишь о том, что изменениям одного признака, как правило, сопутствуют определенные изменения другого.

Корреляционная зависимость – это изменения, которые вносят значения одного признака в вероятность появления разных значений другого признака.

Задача корреляционного анализа сводится к установлению направления (положительное или отрицательное) и формы (линейная, нелинейная) связи между варьирующими признаками, измерению ее тесноты, и, наконец, к проверке уровня значимости полученных коэффициентов корреляции.

Корреляционные связи различаются по типу:

По форме:

Прямолинейные (например, связь между количеством тренировок на тренажере и количеством правильно решаемых задач в контрольной сессии).

Криволинейные (например, связь между уровнем мотивации и эффективностью выполнения задачи).

По направлению:

Положительные «прямые» – при положительной корреляции коэффициент корреляции имеет положительный знак, например $r = +0{,}207$. При положительной прямолинейной корреляции более высоким значениям одного признака соответствуют более высокие значения другого, а более низким значениям одного признака - низкие значения другого.

Отрицательные «обратные» – при отрицательной корреляции - отрицательный знак, например $r = -0{,}207$. При отрицательной корреляции более высоким значениям одного признака соответствуют более низкие значения другого, а более низким значениям одного признака – более высокие значения другого.

По степени (силе или тесноте). Определяется по величине коэффициента корреляции. Сила связи не зависит от ее направленности и определяется по абсолютному значению коэффициента корреляции. Максимальное возможное абсолютное значение коэффициента корреляции $r = 1$, минимальное $r = 0$.

Общая классификация корреляционных связей (по Ивантер Э. В., Коросову А. В., 1992):

- сильная, или тесная при коэффициенте корреляции $r > 0{,}7$;
- средняя при $0{,}5 < r < 0{,}69$;

- умеренная при $0{,}3 < r < 0{,}49$;
- слабая при $0{,}2 < r < 0{,}29$;
- очень слабая при $r < 0{,}19$.

Переменные X и Y могут быть измерены в разных шкалах, именно это определяет выбор соответствующего коэффициента корреляции (таблица 9).

Таблица 9

Использование коэффициента корреляции в зависимости от типа переменных

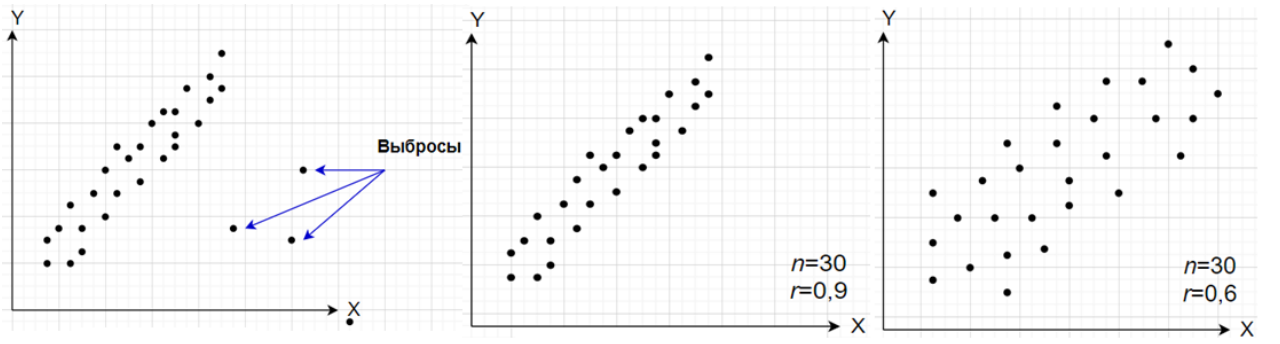
Тип шкалы		Мера связи
Переменная X	Переменная Y	
Интервальная или отношений	Интервальная или отношений	Коэффициент Пирсона
Ранговая, интервальная или отношений	Ранговая, интервальная или отношений	Коэффициент Спирмена
Ранговая	Ранговая	Коэффициент Кендалла
Дихотомическая **	Дихотомическая	Коэффициент «φ»
Дихотомическая **	Ранговая	Рангово-бисериальный
Дихотомическая **	Интервальная или отношений	Бисериальный***

* **Шкала отношений** – это шкала интервалов с естественным (не условным) нулевым значением и принятые по соглашению единицы измерений. В ней нуль характеризует естественное нулевое количество данного свойства. Например, абсолютный нуль температурной шкалы. Это наиболее совершенная и информативная шкала.

****Дихотомическая шкала** – шкала, содержащая только две категории. Пример такой шкалы: пол (мужской и женский)

*****Бисериальная корреляция** – корреляция, в которой одна переменная измеряется по градуированной или многозначной шкале, а другая дихотомической или с двумя значениями. Например, корреляция между показателями, измеряемыми по принципу «удача-неудача», и батареей показателей академической способности, состоящей из десяти отдельных шкал.

На рисунке 22 представлены типы корреляционных связей.



А. Выбросы В. Положительная корреляция С. Может быть положительная корреляция

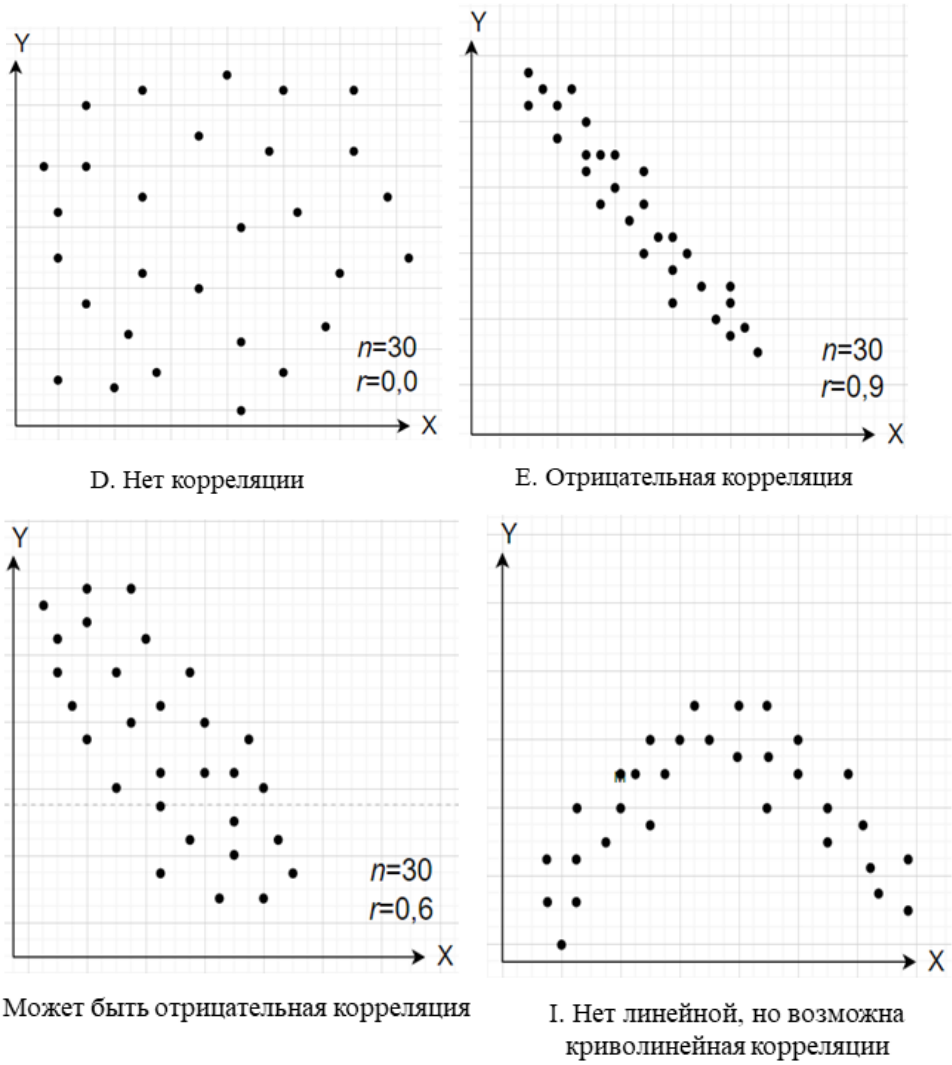


Рис. 22 Типы корреляционных связей

6.2. Коэффициент корреляции Пирсона

Термин «корреляция» был введен в науку выдающимся английским естествоиспытателем Френсисом Гальтоном в 1886 г. Однако точную формулу для подсчета коэффициента корреляции разработал его ученик Карл Пирсон.

Коэффициент характеризует наличие только линейной связи между признаками, обозначаемыми, как правило, символами X и Y. Формула расчета коэффициента корреляции построена таким образом, что, если связь между признаками имеет линейный характер, коэффициент Пирсона точно устанавливает тесноту этой связи. Поэтому он называется также коэффициентом линейной корреляции Пирсона. Если же связь между переменными X и Y не линейна, то Пирсон предложил для оценки тесноты этой связи так называемое корреляционное отношение.

Величина коэффициента линейной корреляции Пирсона не может превышать +1 и быть меньше чем -1. Эти два числа +1 и -1 – являются границами для коэффициента корреляции.

Знак коэффициента корреляции очень важен для интерпретации полученной связи.

Если знак коэффициента линейной корреляции – плюс, то связь между коррелирующими признаками такова, что большей величине одного признака (переменной) соответствует большая величина другого признака (другой переменной). Иными словами,

если один показатель (переменная) увеличивается, то соответственно увеличивается и другой показатель (переменная). Такая зависимость носит название прямо пропорциональной зависимости.

Если же получен знак минус, то большей величине одного признака соответствует меньшая величина другого. Иначе говоря, при наличии знака минус, увеличению одной переменной (признака, значения) соответствует уменьшение другой переменной. Такая зависимость носит название обратно пропорциональной зависимости.

В общем виде формула для подсчета коэффициента корреляции такова:

$$r_{xy} = \frac{\sum(x_i - \bar{x}) * (y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 * \sum(y_i - \bar{y})^2}}, \quad (67)$$

где x_i - значения, принимаемые в выборке X; y_i - значения, принимаемые в выборке Y; $\bar{x}$ – средняя по X; $\bar{y}$ - средняя по Y.

Расчет коэффициента корреляции Пирсона предполагает, что переменные X и Y распределены нормально, то есть имеют стандартное нормальное распределение.

В формуле (67) встречается величина $\sum(x_i - \bar{x}) * (y_i - \bar{y})$ при делении на n (число значений переменной X или Y) она называется **ковариацией**. Формула (67) предполагает также, что при в этом случае выборки равны $X = Y$. Число степеней свободы определяется как $k = n - 2$.

Пример 18. 10 школьникам были даны тесты на наглядно-образное и вербальное мышление. Измерялось среднее время решения заданий теста в секундах. Исследователя интересует вопрос: существует ли взаимосвязь между временем решения этих задач? Переменная X обозначает среднее время решения наглядно-образных, а переменная Y среднее время решения вербальных заданий тестов.

Решение. Представим исходные данные в виде таблицы 10. Для упрощения подсчета коэффициента корреляции, в таблицу 10 можно ввести вспомогательные столбы, характеризующие промежуточные расчеты.

Таблица 10

Исходные данные и промежуточные расчеты

№ исп-го	1	2	3	4	5	6	7
	x	y	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	$y_i - \bar{y}$	$(y_i - \bar{y})^2$	$(x_i - \bar{x})$*$(y_i - \bar{y})$
1	19	17	-16,7	278,89	-7,2	51,84	120,24
2	32	7	-3,7	13,69	-17,2	295,84	63,64
3	33	17	-2,7	7,29	-7,2	51,84	19,44
4	44	28	8,3	68,89	3,8	14,44	31,54
5	28	27	-7,7	59,29	2,8	7,84	-21,56
6	35	31	-0,7	0,49	6,8	46,24	-4,76
7	39	20	3,3	10,89	-4,2	17,64	-13,86
8	39	17	3,3	10,89	-7,2	51,84	-23,76
9	44	35	8,3	68,89	10,8	116,64	89,64
10	44	43	8,3	68,89	18,8	353,44	156,04
Σ	357	242		588,1		1007,6	416,6
$\bar{x}_{\text{общ}}$	35,7	24,2					

Рассчитываем эмпирическую величину коэффициента корреляции по формуле (67):

$$r_{xy} = \frac{416{,}6}{\sqrt{588{,}1 * 1007{,}6}} = \frac{416{,}6}{769{,}8} = 0{,}54$$

Определяем критические значения для полученного коэффициента корреляции по таблице критических значений Пирсона $r_{\text{кр}} = 0{,}72$. При нахождении критических значений для вычисленного коэффициента линейной корреляции Пирсона число степеней свободы рассчитывается как $k = n - 2 = 10 - 2 = 8$.

Так как $r_{\text{кр}} > r_{xy}$ следовательно, гипотеза H_1 отвергается и принимается гипотеза H_0, иными словами, связь между временем решения наглядно-образных и вербальных заданий теста не доказана.

6.3. Корреляция ранговых переменных

Если к количественным данным неприемлем коэффициент корреляции *r*-Пирсона, то для проверки гипотезы о связи двух переменных после предварительного ранжирования могут быть применены корреляции *r*-Спирмена или *τ*-Кендалла.

Для корректного вычисления обоих коэффициентов (Спирмена и Кендалла) результаты измерений должны быть представлены в шкале рангов или интервалов. Принципиальных отличий между этими критериями не существует, но принято считать, что коэффициент Кендалла является более «содержательным», так как он более полно и детально анализирует связи между переменными, перебирая все возможные соответствия между парами значений. Коэффициент Спирмена более точно учитывает именно количественную степень связи между переменными.

Коэффициент ранговой корреляции Спирмена является непараметрическим аналогом классического коэффициента корреляции Пирсона, но при его расчете учитываются не связанные с распределением показатели сравниваемых переменных (среднее арифметическое и дисперсия), а ранги.

Так как этот коэффициент – аналог *r*-Пирсона, то и применение его для проверки гипотез аналогично применению коэффициента *r*-Пирсона. То есть проверяемая статистическая гипотеза, порядок принятия статистического решения и формулировка содержательного вывода – те же. В компьютерных программах (SPSS, Statistica) уровни значимости для одинаковых коэффициентов r-Пирсона и r-Спирмена всегда совпадают.

Преимущество коэффициента *r*-Спирмена по сравнению с коэффициентом *r*-Пирсона – в большей чувствительности к связи. Использование этого коэффициента применимо в следующих случаях:

- наличие существенного отклонения распределения хотя бы одной переменной от нормального вида (асимметрия, выбросы);
- появление криволинейной (монотонной) связи.

Ограничением для применения коэффициента *r*-Спирмена являются:

- по каждой переменной не менее 5 наблюдений;
- коэффициент при большом количестве одинаковых рангов по одной или обеим переменным дает огрубленное значение.

Для начала разберемся с понятием ранга.

Правила ранжирования

1. Ранжирование качественных признаков. Испытуемому предлагается задание, в котором семь личностных качеств необходимо упорядочить (проранжировать) в двух столбцах: в левом столбце в соответствии с особенностями его «Я реального», а в правом столбце в соответствии с особенностями его «Я идеального». Результаты ранжирования даны в таблице 11.

Таблица 11

Результаты ранжирования

Я реальное	Качества жизни	Я идеальное
4	Ответственность	1
1	Общительность	5
3	Настойчивость	7
2	Энергичность	6
5	Жизнерадостность	4
4	Терпеливость	3
6	Решительность	2

Ранжирование в левом столбце осуществляется следующим образом: поскольку всего имеется 7 качеств, то максимальный ранг 7 приписывается качеству наиболее значимому на данный момент времени, а минимальный 1 – наименее значимому. Остальным качествам в соответствии со степенью их значимости приписываются цифры (ранги) от 6 до 2.

В правом столбце проводится ранжирование в соответствии с тем, какими качествами человек хотел бы обладать в идеале. Максимально желательному качеству ставится в соответствие наибольший ранг, а наименее желательным качествам ставятся меньшие величины рангов.

Процедура ранжирования является формальной, поэтому в зависимости от предпочтения можно проставлять величины рангов и в противоположном порядке, т. е. наиболее значимому качеству приписать ранг 1, наименее значимому – ранг 7.

2. Ранжирование количественных признаков. В результате диагностики невроза у пяти испытуемых по методике К. Хека и Х. Хесса были получены следующие баллы: 24, 25, 37, 13, 12 (таблица 12). Этому ряду чисел можно проставить ранги двумя способами:

- большему числу в ряду ставится больший ранг: в этом случае получится: 3, 4, 5, 2, 1;
- большему числу в ряду ставится меньший ранг: в этом случае получится: 3, 2,1, 4, 5.

Таблица 12

X	24	25	37	13	12
Ранг	3	4	5	2	1
Y	12	13	24	25	37
Ранг	1	2	3	4	5

6.3.1. Нахождение коэффициента ранговой корреляции Спирмена

Пример 19. Преподаватели оценили знания 12 учащихся по 100-балльной системе и выставили им следующие оценки (в первой строке указано количество баллов, выставленных первым преподавателем (А), а во второй – вторым (Б):

А	98	94	88	80	76	70	63	61	60	58	56	51
Б	99	91	93	74	78	65	64	66	52	53	48	62

Найти выборочный коэффициент ранговой корреляции Спирмена между оценками двух преподавателей.

Решение. Присвоим ранги x_i оценкам первого преподавателя А. Эти оценки расположены в убывающем порядке, поэтому их ранги равны порядковым номерам:

Оценка	98	94	88	80	76	70	63	61	60	58	56	51
Ранг (x_i)	1	2	3	4	5	6	7	8	9	10	11	12

Найдем ранги y_i . Присвоим ранги y_i оценкам второго преподавателя Б, для чего сначала расположим эти оценки в убывающем порядке и пронумеруем их:

Оценка	99	93	91	78	74	66	65	64	62	53	52	48
Ранг (y_i)	1	2	3	4	5	6	7	8	9	10	11	12

1. y_1 . Индекс $i = 1$ указывает, что рассматривается оценка учащегося, который занимает по оценке первого преподавателя первое место (оценка 98). Видно, что от второго преподавателя этот студент получил оценку 99, наивысшую, значит, ранг равен $y_1 = 1$.

2. y_2 . Индекс $i = 2$ указывает, что рассматривается оценка учащегося, который занимает по оценке первого преподавателя второе место (оценка 94). Видно, что от второго преподавателя этот студент получил оценку 91, которая в ряду занимает место 3, ранг равен $y_2 = 3$. И так далее. Аналогично вычисляем дальше и получим:

x_i	1	2	3	4	5	6	7	8	9	10	11	12
y_i	1	3	2	5	4	7	8	6	11	10	12	9

Заполняем расчетную таблицу:

x_i	1	2	3	4	5	6	7	8	9	10	11	12	
y_i	1	3	2	5	4	7	8	6	11	10	12	9	
$d_i = x_i - y_i$	0	-1	1	-1	1	-1	-1	2	-2	0	-1	3	
${d_i}^2$	0	1	1	1	1	1	1	4	4	0	1	9	$\Sigma = 24$

Находим коэффициент ранговой корреляции Спирмена:

$$\rho_{\text{В}} = 1 - \frac{6\sum d_i^2}{n^3 - n} = 1 - \frac{6 * 24}{12^3 - 12} \approx 0{,}916$$

Выборочный коэффициент ранговой корреляции Спирмена между оценками двух преподавателей $\rho_{\text{В}}$ = 0,916. Значение достаточно высоко, то есть существует взаимосвязь между оценками двух преподавателей.

6.3.2. Нахождение коэффициента ранговой корреляции Кендалла

Пример 20. Пример решения задачи на нахождение коэффициента ранговой корреляции Кендалла. Два контролера расположили 10 деталей в порядке ухудшения их качества. В итоге были получены две последовательности рангов:

x_i	y_i	R
1	1	9
2	2	8
3	4	6
4	3	6
5	6	4
6	5	4
7	7	3
8	10	0
9	9	0
10	8	0
Σ		40

Решение: $R_1 = 9$, так как в столбце y_i ниже $y_1 = 1$ есть ровно 9 значений, больших 1. $R_2 = 8$, так как в столбце y_i ниже $y_2 = 2$ есть ровно 8 значений, больших 2. Далее $R_3 = 6$, так как в столбце y_i ниже $y_3 = 4$ есть ровно 6 значений, больших 4. Аналогично заполняем дальше.

Рассчитываем коэффициент ранговой корреляции Кендалла:

$$\tau_K = \frac{4R}{n(n-1)} - 1 = \frac{4 * 40}{10(10-1)} - 1 \approx 0{,}778$$

Вывод: оценки контроллеров согласуются касательно ухудшения качества деталей.

КОНТРОЛЬНЫЕ ВОПРОСЫ

1. Что такое корреляционная связь?
2. Чем корреляционная связь отличается от причинно-следственной зависимости?
3. Какую задачу решает корреляционный анализ?
4. По каким признакам классифицируются корреляционные связи?
5. Что означает положительная корреляционная связь?
6. Что означает отрицательная корреляционная связь?
7. Чем прямолинейная связь отличается от криволинейной?

8. От чего зависит сила корреляционной связи?
9. Каковы предельные значения коэффициента корреляции?
10. Как интерпретируются значения коэффициента корреляции от 0 до 1?
11. Какие шкалы измерения переменных определяют выбор коэффициента корреляции?
12. В каких случаях применяется коэффициент Пирсона?
13. Почему коэффициент Пирсона называют коэффициентом линейной корреляции?
14. Какой вид распределения переменных предполагается при расчете коэффициента Пирсона?
15. Что такое ковариация и как она связана с коэффициентом корреляции Пирсона?
16. В каких случаях рекомендуется использовать коэффициент Спирмена?
17. В чем состоит отличие коэффициента Спирмена от коэффициента Пирсона?
18. Когда целесообразно применять коэффициент Кендалла?
19. Что такое ранжирование и для чего оно используется в корреляционном анализе?
20. Как интерпретируется высокий положительный или отрицательный коэффициент ранговой корреляции?

ПРАКТИЧЕСКАЯ РАБОТА 5

Тема: Коэффициенты корреляции Пирсона, Спирмена и Кендалла

Инструкция: Для каждого варианта выполните 3 задания. В заданиях 1–3 необходимо вычислить соответствующий коэффициент корреляции и сделать вывод о наличии и направлении связи между признаками (Задача 1. Коэффициент корреляции Пирсона; Задача 2. Коэффициент корреляции Спирмена; Задача 3. Коэффициент корреляции Кендалла).

Вариант 1

Задача 1. По данным 10 студентов исследуется связь между временем подготовки к экзамену и полученной оценкой. Время подготовки, ч: 1, 2, 2, 3, 4, 4, 5, 6, 6, 7. Оценка: 52, 55, 57, 60, 63, 66, 68, 72, 73, 75.

Задача 2. Для 8 работников определены ранги по стажу работы и по качеству выполнения задания. Ранги по стажу: 1, 2, 3, 4, 5, 6, 7, 8. Ранги по качеству: 2, 1, 4, 3, 6, 5, 8, 7.

Задача 3. Два эксперта ранжировали 7 кандидатов по уровню профессионализма. Ранжирование 1-го эксперта: A, B, C, D, E, F, G. Ранжирование 2-го эксперта: B, A, C, E, D, G, F.

Вариант 2

Задача 1. Исследуется связь между количеством часов сна и уровнем внимания у 9 учащихся. Сон, ч: 5, 6, 6, 7, 7, 8, 8, 9, 10. Внимание: 48, 51, 50, 54, 56, 59, 61, 64, 67.

Задача 2. У 10 магазинов определены места по размеру выручки и по количеству посетителей. Места по выручке: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10. Места по посетителям: 3, 1, 2, 5, 4, 7, 6, 10, 9, 8.

Задача 3. Два эксперта ранжировали 7 музыкальных произведений по выразительности. Эксперт 1: M1, M2, M3, M4, M5, M6, M7. Эксперт 2: M2, M1, M4, M3, M5, M7, M6.

Вариант 3

Задача 1. Изучается связь между количеством часов занятий спортом в неделю и уровнем выносливости. Часы занятий: 2, 3, 3, 4, 5, 5, 6, 7. Выносливость: 41, 44, 45, 49, 52, 54, 57, 60.

Задача 2. Для 9 сотрудников сравниваются ранги по опыту работы и по скорости выполнения теста. Ранги по опыту: 1, 2, 3, 4, 5, 6, 7, 8, 9. Ранги по скорости: 2, 3, 1, 5, 4, 7, 6, 9, 8.

Задача 3. Тринадцать объектов оценены двумя специалистами по степени привлекательности, но требуется провести анализ для 6 объектов. Эксперт 1: X1, X2, X3, X4, X5, X6. Эксперт 2: X2, X1, X4, X3, X6, X5.

Вариант 4

Задача 1. Исследуется связь между возрастом автомобиля и расходом топлива. Возраст автомобиля, лет: 1, 2, 3, 4, 5, 6, 7, 8. Расход топлива, л/100 км: 6.1, 6.4, 6.7, 7.0, 7.3, 7.6, 8.0, 8.3.

Задача 2. У 8 сотрудников определены ранги по уровню дисциплины и по качеству отчетности. Ранги по дисциплине: 1, 2, 3, 4, 5, 6, 7, 8. Ранги по отчетности: 1, 3, 2, 4, 6, 5, 8, 7.

Задача 3. Два врача ранжировали 7 пациентов по степени тяжести состояния. Врач 1: P1, P2, P3, P4, P5, P6, P7. Врач 2: P1, P3, P2, P4, P6, P5, P7.

Вариант 5

Задача 1. Проверяется связь между количеством выполненных задач и временем на их решение. Задачи: 4, 5, 6, 7, 8, 9, 10, 11. Время, мин: 42, 45, 47, 50, 54, 57, 60, 63.

Задача 2. Для 10 отделов фирмы определены места по прибыли и по уровню клиентской удовлетворенности. Места по прибыли: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10. Места по удовлетворенности: 2, 1, 4, 3, 6, 5, 8, 7, 10, 9.

Задача 3. Два преподавателя ранжировали 6 работ студентов по качеству оформления. Преподаватель 1: W1, W2, W3, W4, W5, W6. Преподаватель 2: W2, W1, W4, W3, W6, W5.

Вариант 6

Задача 1. Исследуется связь между количеством прочитанных страниц в день и результатом теста на понимание текста. Страниц в день: 10, 12, 14, 15, 16, 18, 20, 22. Результат теста: 58, 60, 63, 65, 68, 71, 74, 77.

Задача 2. У 9 абитуриентов сравниваются ранги по баллам ЕГЭ и по результатам внутреннего экзамена. Ранги по ЕГЭ: 1, 2, 3, 4, 5, 6, 7, 8, 9. Ранги по внутреннему экзамену: 1, 3, 2, 5, 4, 6, 8, 7, 9.

Задача 3. Два эксперта ранжировали 7 образцов по качеству упаковки. Эксперт 1: A1, A2, A3, A4, A5, A6, A7. Эксперт 2: A3, A2, A1, A4, A6, A5, A7.

Вариант 7

Задача 1. Изучается связь между уровнем физической активности и частотой пульса в покое. Часы активности в неделю: 1, 2, 2, 3, 4, 5, 6, 7. Пульс, уд/мин: 82, 80, 79, 77, 74, 72, 69, 67.

Задача 2. Для 8 товаров сравниваются ранги по цене и по объему продаж. Ранги по цене: 1, 2, 3, 4, 5, 6, 7, 8. Ранги по продажам: 2, 1, 4, 3, 5, 7, 6, 8.

Задача 3. Два сотрудника службы качества ранжировали 6 изделий по степени надежности. Сотрудник 1: I1, I2, I3, I4, I5, I6. Сотрудник 2: I2, I1, I4, I3, I6, I5.

Вариант 8

Задача 1. Проверяется связь между количеством рекламных контактов и числом заявок. Контакты: 20, 25, 30, 35, 40, 45, 50, 55. Заявки: 12, 15, 18, 21, 24, 27, 30, 33.

Задача 2. У 10 учеников сравниваются ранги по математике и по физике. Ранги по математике: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10. Ранги по физике: 2, 1, 3, 5, 4, 6, 8, 7, 10, 9.

Задача 3. Два наблюдателя ранжировали 7 фотографий по эстетической привлекательности. Наблюдатель 1: F1, F2, F3, F4, F5, F6, F7. Наблюдатель 2: F2, F1, F3, F5, F4, F7, F6.

Вариант 9

Задача 1. Исследуется связь между температурой воздуха и количеством проданных напитков. Температура, C: 15, 17, 19, 21, 23, 25, 27, 29. Продажи: 40, 45, 50, 56, 61, 67, 72, 78.

Задача 2. Для 8 проектов определены ранги по сроку выполнения и по качеству результата. Ранги по сроку: 1, 2, 3, 4, 5, 6, 7, 8. Ранги по качеству: 1, 3, 2, 5, 4, 6, 8, 7.

Задача 3. Два эксперта ранжировали 6 фильмов по зрительскому интересу. Эксперт 1: K1, K2, K3, K4, K5, K6. Эксперт 2: K2, K1, K4, K3, K6, K5.

Вариант 10

Задача 1. Изучается связь между количеством сна и уровнем концентрации внимания. Сон, ч: 4, 5, 6, 6, 7, 8, 8, 9. Концентрация: 45, 49, 52, 53, 57, 60, 62, 65.

Задача 2. У 9 сотрудников сравниваются ранги по опыту и по результатам аттестации. Ранги по опыту: 1, 2, 3, 4, 5, 6, 7, 8, 9. Ранги по аттестации: 2, 1, 4, 3, 5, 7, 6, 9, 8.

Задача 3. Два специалиста ранжировали 7 книг по полезности для обучения. Специалист 1: B1, B2, B3, B4, B5, B6, B7. Специалист 2: B2, B1, B3, B5, B4, B7, B6.

Вариант 11

Задача 1. Исследуется связь между продолжительностью тренировки и результатом теста на скорость. Тренировка, мин: 15, 20, 25, 30, 35, 40, 45, 50. Результат: 60, 63, 66, 70, 73, 76, 80, 83.

Задача 2. Для 10 товаров определены ранги по качеству и по популярности. Ранги по качеству: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10. Ранги по популярности: 1, 2, 4, 3, 5, 7, 6, 9, 8, 10.

Задача 3. Два преподавателя ранжировали 6 рефератов по содержанию. Преподаватель 1: R1, R2, R3, R4, R5, R6. Преподаватель 2: R1, R3, R2, R5, R4, R6.

Вариант 12

Задача 1. Проверяется связь между количеством обработанных заказов и уровнем производительности. Заказы: 30, 35, 40, 45, 50, 55, 60, 65. Производительность: 48, 51, 55, 58, 62, 65, 69, 72.

Задача 2. У 8 участников соревнований сравниваются ранги по предварительному отбору и по итоговому результату. Ранги по отбору: 1, 2, 3, 4, 5, 6, 7, 8. Ранги по результату: 3, 1, 2, 5, 4, 7, 6, 8.

Задача 3. Два эксперта ранжировали 7 картин по художественной ценности. Эксперт 1: C1, C2, C3, C4, C5, C6, C7. Эксперт 2: C2, C1, C3, C5, C4, C7, C6

7. РЕГРЕССИОННЫЙ АНАЛИЗ

Это группа методов, направленных на выявление и математическое выражение тех изменений и зависимостей, которые имеют место в системе случайных величин.

Методы регрессионного анализа рассчитаны, главным образом, на случай устойчивого нормального распределения, в котором изменения от опыта к опыту проявляются лишь в виде независимых испытаний.

Выделяются различные формальные задачи регрессионного анализа. Они могут быть простыми или сложными по формулировкам, по математическим средствам и трудоемкости.

Первая цель – выявить факт изменчивости изучаемого явления при определенных, но не всегда четко фиксированных условиях.

Вторая цель – выявить тенденцию как периодическое изменение признака. Сам по себе этот признак может быть зависим или не зависим от переменной-условия (он может зависеть от неизвестных или неконтролируемых исследователем условий). Но это не важно для рассматриваемой задачи, которая ограничивается лишь выявлением тенденции и ее особенностей.

Задачи регрессионного анализа сводятся к следующему:

1. Установление формы зависимости. Относительно характера и формы зависимости между явлениями, различают положительную линейную и нелинейную и отрицательную линейную и нелинейную регрессию.
2. Определение функции регрессии в виде математического уравнения того или иного типа и установление влияния объясняющих переменных на зависимую переменную.
3. Оценка неизвестных значений зависимой переменной. С помощью функции регрессии можно воспроизвести значения зависимой переменной внутри интервала заданных значений объясняющих переменных (т. е. решить задачу интерполяции) или оценить течение процесса вне заданного интервала (т. е. решить задачу экстраполяции). Результат представляет собой оценку значения зависимой переменной.

Парная регрессия – уравнение связи двух переменных y и x: $y = f(x)$,
где y – зависимая переменная (результативный признак); x – независимая, объясняющая переменная (признак-фактор).

Различают линейные и нелинейные регрессии:

1. Линейная регрессия: $y = a + bx + \varepsilon$, где a и b – коэффициенты регрессии, ε – ошибка (погрешность).
2. Нелинейные регрессии делятся на два класса:

- регрессии, нелинейные относительно включенных в анализ объясняющих переменных, но линейные по оцениваемым параметрам, это:

✓ Полиномы разных степеней

$$y = a + b_1 * x + b_2 * x^2 + b_3 * x^3 + \cdots + \varepsilon$$

✓ Равносторонняя гипербола

$$y = a + \frac{b}{x} + \varepsilon$$

- регрессии, нелинейные по оцениваемым параметрам:

✓ Степенная $y = a * x^b * \varepsilon$

- ✓ Показательная $y = a * b^x * \varepsilon$
- ✓ Экспоненциальная $y = e^{a+b*x} * \varepsilon$

Регрессионный анализ заключается в определении аналитического выражения связи, в котором изменение одной величины (называемой зависимой или результативным признаком) обусловлено влиянием одной или нескольких независимых величин (факторов), а множество всех прочих факторов, также оказывающих влияние на зависимую величину, принимается за постоянные и средние значения.

Парная регрессия характеризует связь между двумя признаками: факторным и результативным.

Уравнение регрессии имеет вид:

$$Y = ax + b, \tag{68}$$

где Y – результативный показатель; a, b – параметры регрессии; x – факторный показатель.

Построение уравнения регрессии сводится к оценке ее параметров.

Для оценки параметров регрессий, линейных по параметрам, используют метод наименьших квадратов (МНК). МНК позволяет получить такие оценки параметров, при которых сумма квадратов отклонений фактических значений результативного признака y от теоретических y_x минимальна, то есть:

$$\sum(y - \hat{y}_x)^2 \to min, \tag{69}$$

Для парной линейной регрессии система нормальных уравнений, полученная на основе метода наименьших квадратов имеет вид:

$$\begin{cases} \sum y = na + b\sum x \\ \sum xy = a\sum x + b\sum x^2, \end{cases} \tag{70}$$

где n – количество наблюдений; a, b – параметры регрессии.

Можно воспользоваться готовыми формулами, которые вытекают из этой системы (71, 72):

$$a = \bar{y} - b * \bar{x}, \tag{71}$$

$$b = \frac{cov(x,y)}{\sigma_x^2} = \frac{\overline{xy} - \bar{y} * \bar{x}}{\overline{x^2} - \bar{x}^2}, \tag{72}$$

Тесноту связи изучаемых явлений оценивает **линейный коэффициент парной корреляции r_{xy}** (от -1 до 1) для линейной регрессии рассчитывают по формуле 73:

$$r_{xy} = b\frac{\sigma_x}{\sigma_y} = \frac{cov(x;y)}{\sigma_x * \sigma_y} = \frac{\overline{xy} - \bar{y} * \bar{x}}{\overline{x^2} - \bar{x}^2}; \tag{73}$$

Индекс корреляции ρ_{xy} для нелинейной регрессии (от 0 до 1), формула 74:

$$\rho_{xy} = \sqrt{1 - \frac{\sigma_{ост}^2}{\sigma_y^2}} = \sqrt{1 - \frac{\sum(y - \hat{y}_x)^2}{\sum(y - \bar{y})^2}}; \tag{74}$$

Оценку качества построенной модели даст коэффициент (индекс) детерминации, а также средняя ошибка аппроксимации.

Средняя ошибка аппроксимации – среднее отклонение расчетных значений от фактических (75):

$$\bar{A} = \frac{1}{n}\sum\left|\frac{y - \hat{y}}{y}\right| * 100\%, \tag{75}$$

Допустимый предел значений $\bar{A}$ не более 8–10%

Средний коэффициент эластичности $\bar{Э}$ показывает на сколько процентов в среднем по совокупности измениться результат y от своей средней величины при изменении фактора x на 1% от своего среднего значения (76):

$$\bar{Э} = f'(x) * \frac{\bar{x}}{\bar{y}}, \tag{76}$$

Проверка гипотез об отсутствии или наличии тенденции может выполняться с использованием критерия Аббе. Критерий Аббе предназначен для проверки гипотез о равенстве средних значений, установленных для $4 < n < 60$ взаимно независимых нормально распределенных выборок. Эмпирическое значение критерия Аббе вычисляется по формуле 77:

$$q = 0{,}5 * \frac{\sum_{i=1}^{n-1}(x_{i+1}-x_i)^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}, \tag{77}$$

Согласно критерию, гипотеза о равенстве средних отклоняется H_0 (принимается альтернативная гипотеза H_1), если значение статистики $q_{\text{эмп}} < q_{\text{кр}}$. Табличное (критическое) значение статистики определяется из таблицы для q-критерия Аббе.

Рассмотрим пример выявления тенденции с помощью критерия Аббе.

Пример 21. В таблице 13 представлена динамика процента студентов IV курса, на «отлично» сдававших экзамены в зимние сессии на протяжении 10 лет работы одного из факультетов университета. Требуется установить, есть ли тенденция к повышению успеваемости. В качестве H_0 проверяем гипотезу об отсутствии тенденции, т. е. о равенстве процентов.

Таблица 13

Учебный год	%
1995-96	10,8
1996-97	16,4
1997-98	17,4
1998-99	22,0
1999-00	23,0
2000-01	21,5
2001-02	26,1
2002-03	17,2
2003-04	27,5
2004-05	33,0
$x_{\text{ср}}$	21,5

Вычисляем разности между последующими и предыдущими значениями в выборке, возводим их в квадрат и суммируем:

$$\sum_{i=1}^{n-1}(x_{i+1}-x_i)^2 = (16{,}4-10{,}8)^2 + (17{,}4-16{,}4)^2 + \cdots + (27{,}5-17{,}2)^2 + (33-27{,}5)^2 = 303{,}68$$

Аналогично вычисляет знаменатель в формуле (77), суммируя квадраты разностей между каждым измерением и средним арифметическим:

$$\sum_{i=1}^{n}(x_i-\bar{x})^2=(10{,}8-21{,}5)^2+(16{,}4-21{,}5)^2+\cdots+(27{,}5-21{,}5)^2+(33-21{,}5)^2=367{,}71$$

Теперь по формуле (77) получаем:

$$q_{\text{эмп}}=0{,}5*\frac{303{,}67}{367{,}71}=0{,}41$$

В таблице критерия Аббе находим, что при $n=10$ и уровне значимости 0,05 критическое значение $q_{\text{кр}}=0{,}5311$. Это больше полученного нами 0,41, следовательно, гипотезу о равенстве процента «отличников» приходится отклонить, и можно принять альтернативную гипотезу о наличии тенденции.

Третья задача – это выявление закономерности, выраженной в виде корреляционного уравнения (регрессии).

Линейная регрессия, проверка адекватности модели (уравнения)

Пример 22. Найти выборочные числовые характеристики и выборочное уравнение линейной регрессии $y_x=ax+b$. Построить прямую регрессии и изобразить на плоскости (x и y) из таблицы. Вычислить остаточную дисперсию. Проверить адекватность регрессионной модели по коэффициенту детерминации.

x	1	2	4	6	10
y	2	6	11	16	24

Решение: здесь дана выборка объема $n=5$ из двумерной генеральной совокупности $(X;Y)$. Для решения задачи необходимо найти следующие величины:

1. Выборочные средние (или просто средние значения для выборки Xи Y:

$$\bar{x}=\frac{1}{n}\sum x_i=\frac{1}{5}23=4{,}6$$

$$\bar{y}=\frac{1}{n}\sum y_i=\frac{1}{5}59=11{,}8$$

2. Выборочные дисперсии и среднеквадратические отклонения:

$$D_x=\overline{x^2}-(\bar{x})^2=\frac{1}{n}\sum x_i^{\,2}-4{,}6^2=31{,}4-21{,}16=10{,}24$$

$$D_y=\overline{y^2}-(\bar{y})^2=\frac{1}{n}\sum y_i^{\,2}-11{,}8^2=198{,}6-139{,}24=59{,}36$$

$$\sigma_x=\sqrt{D_x}=\sqrt{10{,}24}=3{,}2$$

$$\sigma_y=\sqrt{D_y}=\sqrt{59{,}36}\approx 7{,}705$$

3. Сумму $\overline{xy}=\frac{1}{n}\sum x_iy_i=\frac{1}{5}394=78{,}8$

4. Выборочный коэффициент корреляции:

$$r_{xy} = \frac{\overline{xy} - \bar{x} * \bar{y}}{\sigma_x * \sigma_y} = \frac{78{,}8 - 4{,}6 * 11{,}8}{3{,}2 * 7{,}705} \approx 0{,}994$$

Коэффициент корреляции близок к единице, связь очень сильная (близка к функциональной), прямая.

Все расчеты представлены в таблице 14:

Таблица 14

	x_i	y_i	x_i^2	y_i^2	$x_i * y_i$
	1	2	1	4	2
	2	6	4	36	12
	4	11	16	121	44
	6	16	36	256	96
	16	24	100	576	240
Сумма	23	59	157	993	394
Среднее	4,6	11,8	31,4	198,6	78,8

5. Запишем выборочное уравнение линейной регрессии по формуле:

$$\frac{y_x - \bar{y}}{\sigma_y} = r_{xy} \frac{x - \bar{x}}{\sigma_x}, \tag{78}$$

$$\frac{y_x - \overline{11{,}8}}{7{,}705} = 0{,}994 \frac{x - \overline{4{,}6}}{3{,}2}$$

Выразим из уравнения y_x:

$$y_x = 2{,}393x + 0{,}791$$

6. По уравнению регрессии вычисляем y_x в заданных точках и заполняем таблицу 15:

Таблица 15.

	x_i	y_i	$y_x(x_i)$	$(y_x(x_i) - y_i)^2$
	1	2	3,184	1,402
	2	6	5,577	0,179
	4	11	10,363	0,406
	6	16	15,149	0,724
	16	24	24,721	0,520
Сумма	23	59	58,994	3,231

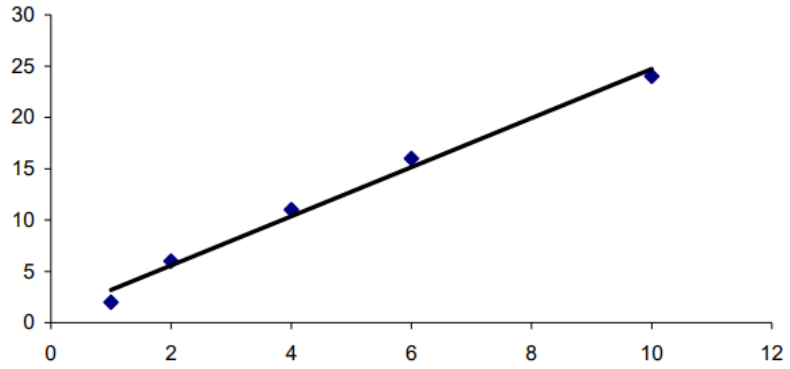
7. Рассчитываем остаточную дисперсию и коэффициент детерминации:

$$D_{\text{ост}} = \frac{1}{n}\sum(y_x(x_i) - y_i)^2 = \frac{1}{5}3{,}231 = 0{,}646, \tag{79}$$

$$R^2 = 1 - \frac{D_{\text{ост}}}{D_y} = 1 - \frac{0{,}646}{59{,}36} = 0{,}989, \tag{80}$$

Значение $R^2 = 0{,}989$ (близкое к единице) показывает, что выбранная регрессионная модель не противоречит экспериментальным данным.

По уравнению регрессии построим прямую в системе координат и на ту же систему координат нанесем точки (x, y) из таблицы.



Пример 23. Провести регрессионный анализ зависимости выручки от числа торговых точек.

Исходные данные представлены в таблице 16. В таблице 17 приведены исходные данные для расчета параметров регрессии.

Таблица 16

№	Выручка, тыс. руб.	Число торговых точек	№	Выручка, тыс. руб.	Число торговых точек
1	1598	2	16	4076	4
2	2644	5	17	1869	6
3	2197	4	18	3524	3
4	1959	5	19	3925	6
5	1052	3	20	1998	4
6	1922	3	21	2606	3
7	2385	5	22	2353	5
8	2581	5	23	2981	3
9	3105	5	24	4471	7
10	3896	4	25	2308	6
11	1510	4	26	4563	5
12	1880	2	27	4306	7
13	3620	5	28	2541	5
14	5002	6	29	6184	8
15	2819	5	30	7481	9

Таблица 17

№	x	y	x^2	$x * y$	$\hat{y}$	1	2	3
1	2	1598	4	3196	1352,1	0,154	60455,24	7,84
2	2	1880	4	3760	1352,1	0,281	278653,6	7,84
3	3	1052	9	3156	1980,6	0,883	862303,9	3,24
4	3	1922	9	5766	1980,6	0,030	3434,337	3,24
5	3	3524	9	10572	1980,6	0,438	2382074	3,24
6	3	2606	9	7818	1980,6	0,240	391121,1	3,24

№	x	y	x^2	$x*y$	$\hat{y}$	1	2	3
7	3	2981	9	8943	1980,6	0,336	1000794	3,24
8	4	2197	16	8788	2609,1	0,188	169812,3	0,64
9	4	3896	16	15584	2609,1	0,330	1656156	0,64
10	4	1510	16	6040	2609,1	0,728	1207983	0,64
11	4	4076	16	16304	2609,1	0,360	2151846	0,64
12	4	1998	16	7992	2609,1	0,306	373422,3	0,64
13	5	2644	25	13220	3237,6	0,224	352316,6	0,04
14	5	1959	25	9795	3237,6	0,653	1634722	0,04
15	5	2385	25	11925	3237,6	0,357	726863	0,04
16	5	2581	25	12905	3237,6	0,254	431074,5	0,04
17	5	3105	25	15525	3237,6	0,043	17572,84	0,04
18	5	3620	25	18100	3237,6	0,106	146258,4	0,04
19	5	2819	25	14095	3237,6	0,148	175194,7	0,04
20	5	2353	25	11765	3237,6	0,376	782451	0,04
21	5	4563	25	22815	3237,6	0,290	1756784	0,04
22	5	2541	25	12705	3237,6	0,274	485199,5	0,04
23	6	5002	36	30012	3866,0	0,227	1290400	1,44
24	6	1869	36	11214	3866,0	1,069	3988178	1,44
25	6	3925	36	23550	3866,0	0,015	3476,01	1,44
26	6	2308	36	13848	3866,0	0,675	2427496	1,44
27	7	4471	49	31297	4494,5	0,005	553,2843	4,84
28	7	4306	49	30142	4494,5	0,044	35540,54	4,84
29	8	6184	64	49472	5123,0	0,172	1125717	10,24
30	9	7481	81	67329	5751,5	0,231	2991235	17,64
Итого:	144	93356	770	497633	93356	9,437	28909087	78,8

Промежуточные расчеты таблицы 17: 1. $\frac{|\hat{y}_i - y_i|}{y_i}$; 2. $(\hat{y}_i - y_i)^2$; 3. $(x_i - \bar{x})^2$

Система уравнений для расчета параметров регрессии имеет вид:

$$\begin{Bmatrix} 93356 = 30a + 144b \\ 497633 = 144a + 770b \end{Bmatrix}$$

Для решения системы уравнений удобно их преобразовать и выразить параметры регрессии:

$$a = \bar{y} - b\bar{x}$$

$$b = \frac{\overline{xy} - \bar{x} * \bar{y}}{\sigma_x^2},$$

Проведем вспомогательные расчёты из таблицы 17:

$$\bar{x} = \frac{144}{30} = 4{,}8$$

$$\bar{y} = \frac{93356}{30} = 3111{,}9$$

$$\overline{xy} = \frac{497633}{30} = 16587{,}8$$

Дисперсия факторного признака (x):

$$\sigma_x^2 = \frac{78{,}8}{30} = 2{,}63$$

Найдем параметры регрессии:

$$b = \frac{\overline{xy} - \bar{x} * \bar{y}}{\sigma_x^2} = \frac{16587{,}8 - 4{,}8 * 3111{,}9}{2{,}63} = 628{,}5$$

$$a = \bar{y} - b\bar{x} = 3111{,}9 - 628{,}5 * 4{,}8 = 95{,}2$$

Уравнение регрессии имеет вид:

$$\hat{y} = 95{,}2 + 628{,}5 * x$$

Подставив в полученное уравнение значения факторного признака из таблицы 16, определим расчетное значение результативного показателя (табл. 17).

Для проверки значимости полученного уравнения рассчитаем **среднюю ошибку для коэффициента парной регрессии** b по данным таблицы 17:

$$\varepsilon_b = \sqrt{\frac{\frac{\sum_{i+1}^{n}(y_i - \hat{y_i})^2}{(n-2)}}{\sum_{i=1}^{n}(x_i - \bar{x})^2}} = \sqrt{\frac{\frac{28909087}{(30-2)}}{78{,}8}} = 114{,}5 \quad (81)$$

Используем критерий Стьюдента для определения регрессионной связи между количеством торговых точек и выручкой (82):

$$t_{\text{стьюд.}} = \frac{b}{\varepsilon_b} = \frac{628{,}5}{114{,}5} = 5{,}49, \quad (82)$$

Критическое значение критерия при уровне значимости ($\alpha = 0{,}01$) и числе степеней свободы ($k = n - 2 = 30 - 2 = 28$) составляет 2,763. Полученное значение критерия намного больше табличного, следовательно, гипотезу о незначительности регрессионной связи между показателями можно отклонить. Это означает, что между количеством торговых точек и выручкой существует регрессионная связь.

Для надежности полученного уравнения регрессии рассчитаем показатель средней ошибки аппроксимации (83) по таблице 17:

$$\bar{\varepsilon} = \frac{1}{n}\sum \frac{|y_i - \hat{y}_i|}{y_i} * 100\% = \frac{1}{30} * 9{,}437 * 100\% = 31{,}5\%, \quad (83)$$

Из расчета видно, что средняя ошибка аппроксимации слишком велика, следовательно полученное уравнение регрессии имеет слишком большую погрешность и не может быть использовано для прогнозирования.

Вывод: между количеством торговых точек и выручкой существует зависимость. Однако модель, построенная в результате регрессионного анализа имеет слишком высокую погрешность и для прогнозирования выручки использована быть не может.

Еще один пример расчета парной линейной регрессии, здесь уравнение регрессии имеет вид (84):

$$y_x = a_0 + a_1 x, \tag{84}$$

где a_1 – параметр, который называется коэффициент регрессии.

Оценка параметров уравнения линейной регрессии (a_0, a_1) осуществляется МНК.

Система нормальных уравнений для нахождения параметров линейной парной регрессии методом наименьших квадратов имеет следующий вид:

$$\begin{cases} na_0 + a_1 * \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \\ a_0 \sum_{i=1}^{n} x_i + a_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i \end{cases}, \tag{85}$$

где n – объем исследуемой совокупности (число единиц наблюдения).

В уравнении регрессии коэффициент a_0 показывает усредненное влияние на результативный признак неучтенных в уравнении факторных признаков. Коэффициент регрессии a_1 показывает, на сколько в среднем изменяется значение результативного признака при увеличении факторного признака на единицу собственного изменения.

Решаем систему и выражаем параметры регрессии:

$$a_1 = \frac{\overline{x*y} - \bar{x}*\bar{y}}{\overline{x^2} - \bar{x}^2} \; a_0 = \frac{\overline{x^2}*\bar{y} - \bar{x}*\overline{x*y}}{\overline{x^2} - \bar{x}^2}, \tag{86}$$

Пример 24. Составить линейное уравнение регрессии зависимости балансовой прибыли от объема реализованной продукции 10 предприятий города (исходные данные представлены в таблице 18). Определить параметры регрессии и проанализировать их.

Таблица 18

Показатели деятельности предприятий города №	Объем реализованной продукции (млн. руб.), x	Балансовая прибыль (млн. руб.), y
1	49,1	14,1
2	48,3	12,7
3	48,1	12,2
4	47,8	12,1
5	47,6	11,7
6	47,5	11,6
7	45,6	10,8
8	45,3	10,5
9	48,1	10,7
10	44,6	9,9

Решение. В качестве факторного (изменяющегося) признака выберем объем реализованной продукции, а за результативный признак (который изменяется под действием факторного) выберем балансовую прибыль.

На основе качественного анализа таблицы 18 можем предположить, что между объемом реализованной продукции и балансовой прибылью существует прямая связь, так как с уменьшением объема реализованной продукции балансовая прибыли также уменьшается

(для надежность выводов можно рассчитать коэффициент корреляции между двумя показателями x и y).

Запишем уравнении регрессии:

$$y_x = a_0 + a_1 x,$$

Параметры a_0 и a_1 определим из системы уравнений (84) по МНК. Дополнительные расчёты представим в таблице 19.

Таблица 19

Показатели деятельности предприятий города №	Объем реализованной продукции (млн. руб.), x	Балансовая прибыль (млн. руб.), y	$x * y$	x^2	y^2	$\overline{y_x}$
1	49,1	14,1	692,3	2410,8	198,8	12,0
2	48,3	12,7	613,4	2332,9	161,3	11,9
3	48,1	12,2	586,8	2313,6	148,8	11,8
4	47,8	12,1	578,4	2284,8	146,4	11,8
5	47,6	11,7	556,9	2265,8	136,9	11,7
6	47,5	11,6	551,0	2256,3	134,6	11,7
7	45,6	10,8	492,5	2079,4	116,6	11,3
8	45,3	10,5	475,7	2052,1	110,3	11,3
9	48,1	10,7	514,7	2313,6	114,5	11,8
10	44,6	9,9	441,5	1989,2	98	11,1
Итого	472	116,3	5503,2	22298,5	1366,2	116,4

Используя данные дополнительных столбов таблицы 19, составим систему уравнений:

$$\begin{cases} na_0 + a_1 * \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \\ a_0 \sum_{i=1}^{n} x_i + a_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i \end{cases} \Rightarrow \begin{cases} 10a_0 + 472a_1 = 116{,}3 \\ 472a_0 + 22298{,}5a_1 = 5503{,}2 \end{cases}$$

Решим систему уравнений, получим значения параметров уравнения регрессии:

$$a_0 = 2{,}2$$

$$a_1 = 0{,}2$$

Уравнение регрессии имеет вид:

$$y_x = 2{,}2 + 0{,}2x$$

Свободный член уравнения a_0=2,2 характеризует величину балансовой прибыли, не зависищую от объема реализованной продукции.

Коэффициент регрессии $a_1 = 0{,}2$ показывает, что при увеличении объема реализованной продукции на 1 млн. руб. балансовая прибыль увеличится на 0,2 млн. руб.

После того, как составлено уравнение регрессии на основе фактической информации по данным о 10 предприятиях города, рассчитаем теоретические значения путем подстановки

y в уравнение регрессии фактических данных об объеме реализованной продукции вместо значений x:

$$\overline{y_x} = 2{,}2 + 0{,}2 * 49{,}1 = 12{,}0$$

$\overline{y_x} = 2{,}2 + 0{,}2 * 48{,}3 = 11{,}9$ и так далее.

Принятие решений на основе уравнений регрессии

Интерпретация моделей регрессии осуществляется методами той отрасли знаний, к которой относится исследуемое явление. Всякая интерпретация начинается со статистической оценки уравнения регрессии в целом и оценки значимости входящих в модели факторных признаков. Чем больше абсолютное значение коэффициента регрессии, тем значительнее влияние факторного признака на результативный признак. Знак коэффициента регрессии говорит о характере влияния на результативный признак. Если коэффициент регрессии имеет знак +, то с увеличением фактора результативный признак возрастает. Если коэффициент регрессии имеет знак **(−)**, то с его увеличением результативный признак уменьшается.

Для рассмотренного примера коэффициент регрессии $a_1 = +0{,}2$. Это значит, что с увеличением факторного признака (объем реализованной продукции), результативный признак также возрастает (балансовая прибыль).

КОНТРОЛЬНЫЕ ВОПРОСЫ

1. Что такое регрессионный анализ и какую задачу он решает?
2. Какие основные цели регрессионного анализа выделяются в тексте?
3. Чем отличается зависимая переменная от независимой?
4. Что называют парной регрессией?
5. Как записывается уравнение парной линейной регрессии?
6. Чем линейная регрессия отличается от нелинейной?
7. Какие виды нелинейных регрессий перечислены в тексте?
8. Что означает, что регрессия линейна по параметрам?
9. В чем состоит метод наименьших квадратов?
10. Какие параметры оцениваются в уравнении линейной парной регрессии?
11. Что показывает коэффициент регрессии b в линейной модели?
12. Как определяется свободный член a в уравнении регрессии?
13. Что такое выборочный коэффициент корреляции и каков его диапазон?
14. Что показывает коэффициент корреляции r_{xy} в линейной регрессии?
15. Что такое индекс корреляции для нелинейной регрессии?
16. Что такое коэффициент детерминации и как он интерпретируется?
17. Что показывает средняя ошибка аппроксимации и каков ее допустимый предел?
18. Для чего используется средний коэффициент эластичности?
19. Что такое критерий Аббе и для чего он применяется?
20. Как интерпретируется знак и величина коэффициента регрессии при анализе модели?

ПРАКТИЧЕСКАЯ РАБОТА 6

Для каждого варианта необходимо выполнить регрессионный анализ по заданной выборке данных. По исходным значениям факторов и результативного признака построить парную линейную модель зависимости y от x, определить параметры уравнения регрессии, оценить тесноту связи между переменными и сделать содержательный вывод о характере зависимости. Для каждого задания также требуется проверить статистическую значимость полученной модели и её параметров, а при необходимости – дать экономическую или практическую интерпретацию коэффициентов регрессии.

Вариант 1

Задание 1. Парная линейная регрессия. По данным выборки найти уравнение линейной регрессии, коэффициент корреляции, остаточную дисперсию и коэффициент детерминации. x: 1; 2; 3; 5; 7; y: 3; 5; 8; 12; 15

Задание 2. Проверка значимости коэффициента регрессии. Для данных определить параметры линейной регрессии yy на xx и проверить значимость коэффициента регрессии по критерию Стьюдента. x: 2; 3; 4; 5; 6; 7; 8; 9; y: 5; 7; 8; 10; 12; 14; 15; 17.

Задание 3. Интерпретация параметров. Составить уравнение регрессии зависимости выручки y, тыс. руб., от числа торговых точек x и дать интерпретацию коэффициентов. x: 2; 3; 3; 4; 5; 5; 6; 7. y: 1480; 1820; 2050; 2280; 2710; 2950; 3180; 3890.

Вариант 2

Задание 1. Парная линейная регрессия. Для выборки найти уравнение регрессии, коэффициент детерминации и остаточную дисперсию. x: 1; 3; 4; 6; 8. y: 2; 6; 7; 11; 16.

Задание 2. Проверка качества модели. Построить линейную модель зависимости y от x и проверить адекватность модели по критерию Фишера. x: 2; 4; 5; 7; 9; 10. y: 4; 8; 9; 13; 17; 19.

Задание 3. Интерпретация параметров. Исследовать зависимость затрат на сырьё y, тыс. руб., от объема выпуска x. Найти уравнение регрессии и пояснить смысл свободного члена и коэффициента наклона. x: 5; 6; 7; 8; 9; 10. y: 18; 21; 24; 27; 29; 32.

Вариант 3

Задание 1. Регрессионный анализ. Найти уравнение линейной регрессии, коэффициент корреляции и коэффициент детерминации. x: 3; 4; 5; 6; 7. y: 6; 8; 10; 13; 15.

Задание 2. Проверка значимости модели. Оценить параметры уравнения регрессии и проверить значимость модели по критерию Фишера. x: 1; 2; 4; 6; 8; 9. y: 2; 4; 7; 11; 15; 17.

Задание 3. Интерпретация параметров. Составить уравнение регрессии зависимости себестоимости y, руб., от объема партии x и объяснить экономический смысл параметров. x: 10; 12; 14; 16; 18; 20. y: 98; 95; 93; 90; 88; 85.

Вариант 4

Задание 1. Парная линейная регрессия. Определить уравнение регрессии, остаточную дисперсию и коэффициент детерминации. x: 2; 3; 5; 7; 9. y: 4; 6; 9; 13; 18.

Задание 2. Проверка значимости коэффициента регрессии. Проверить значимость коэффициента наклона линейной регрессии по критерию Стьюдента. x: 1; 2; 3; 4; 5; 6. y: 3; 5; 6; 8; 10; 11.

Задание 3. Интерпретация параметров. Составить уравнение регрессии зависимости расхода электроэнергии y от числа станков x и интерпретировать параметры модели. x: 4; 5; 6; 7; 8; 9. y: 52; 58; 64; 70; 76; 82.

Вариант 5

Задание 1. Парная линейная регрессия. Найти уравнение линейной регрессии, коэффициент корреляции и среднюю ошибку аппроксимации. x: 10; 15; 20; 25; 30. y: 100; 140; 180; 230; 270.

Задание 2. Проверка значимости модели. Построить уравнение регрессии и проверить его значимость через F-критерий Фишера ($\alpha = 0{,}05$). x: 5; 7; 8; 10; 12; 15. y: 12; 16; 19; 24; 28; 35.

Задание 3. Интерпретация параметров. Оценить зависимость спроса (y, шт.) от цены (x, руб.). Составить уравнение и объяснить смысл свободного коэффициента и коэффициента регрессии. x: 100; 120; 140; 160; 180; 200. y: 500; 450; 410; 360; 320; 290.

Вариант 6

Задание 1. Регрессионный анализ. Для заданного ряда данных рассчитать параметры уравнения и коэффициент детерминации. x: 2; 4; 6; 8; 10. y: 5; 9; 13; 17; 21.

Задание 2. Проверка значимости коэффициента регрессии. Рассчитать t-статистику Стьюдента для коэффициента наклона и сделать вывод о его статистической значимости. x: 1; 3; 4; 6; 7; 9. y: 10; 14; 17; 22; 25; 30.

Задание 3. Интерпретация параметров. Выявить зависимость производительности труда (y, ед./час) от стажа работы (x, лет). Построить модель и дать экономический комментарий к параметрам. x: 1; 2; 3; 4; 5; 6. y: 15; 18; 22; 25; 29; 32.

Вариант 7

Задание 1. Парная линейная регрессия. Вычислить параметры регрессии, коэффициент корреляции Пирсона и остаточную дисперсию. x: 1,5; 2,5; 3,5; 4,5; 5,5. y: 4; 7; 10; 13; 16.

Задание 2. Проверка адекватности модели. Используя критерий Фишера, определить, является ли построенная линейная модель статистически надежной. x: 3; 6; 9; 12; 15; 18. y: 10; 20; 28; 40; 52; 61.

Задание 3. Интерпретация параметров. Составить уравнение зависимости веса (y, кг) от роста (x, см) для группы людей. Объяснить, на сколько в среднем меняется y при изменении x на единицу. x: 160; 165; 170; 175; 180; 185. y: 60; 64; 68; 74; 80; 86.

Вариант 8

Задание 1. Линейная регрессия и корреляция. Определить формулу связи y от x, а также оценить тесноту связи через коэффициент корреляции. x: 20; 30; 40; 50; 60. y: 45; 65; 85; 105; 125.

Задание 2. Оценка точности модели. Рассчитать параметры модели и проверить значимость коэффициента регрессии по t-критерию при уровне значимости 0,01. x: 4; 8; 12; 16; 20; 24. y: 11; 21; 33; 45; 57; 69.

Задание 3. Интерпретация параметров. Построить зависимость рекламного бюджета (y, тыс. долл.) от объема продаж (x, млн долл.). Интерпретировать значения a и b в полученном уравнении. x: 10; 12; 15; 18; 20; 25. y: 1,2; 1,5; 1,9; 2,3; 2,5; 3,2.

Вариант 9

Задание 1. Парная линейная регрессия. Рассчитать параметры линейного уравнения регрессии, определить коэффициент детерминации и коэффициент корреляции. x: 5; 10; 15; 20; 25. y: 12; 25; 37; 50; 63.

Задание 2. Проверка гипотез. Построить уравнение регрессии и проверить значимость уравнения в целом с помощью F-критерия на уровне 0,05. x: 2; 5; 8; 11; 14; 17. y: 6; 14; 22; 31; 40; 48.

Задание 3. Интерпретация параметров. Исследовать зависимость времени выполнения задачи (y, мин.) от опыта сотрудника (x, мес.). Объяснить смысл коэффициентов уравнения. x: 3; 6; 9; 12; 15; 18. y: 50; 45; 40; 36; 32; 28.

Вариант 10

Задание 1. Оценка тесноты связи. Найти линейное уравнение регрессии, оценить силу связи с помощью коэффициента корреляции и вычислить ошибку аппроксимации. x: 1; 4; 7; 10; 13. y: 10; 22; 34; 46; 58.

Задание 2. Статистический вывод. Определить коэффициент регрессии и проверить его значимость по t-критерию Стьюдента при доверительной вероятности 0,95. x: 10; 20; 30; 40; 50; 60. y: 15; 28; 42; 56; 70; 84.

Задание 3. Интерпретация параметров. Построить модель зависимости потребления ресурса (y) от температуры (x). Дать качественную оценку полученным коэффициентам регрессии. x: -10; -5; 0; 5; 10; 15. y: 80; 72; 65; 58; 50; 42.

Вариант 11

Задание 1. Регрессионный анализ. Найти коэффициенты парной линейной регрессии, построить график по исходным точкам и линии регрессии. x: 2; 6; 10; 14; 18. y: 5; 15; 25; 35; 45.

Задание 2. Проверка адекватности. Оценить адекватность модели по критерию Фишера и рассчитать доверительные интервалы для параметров уравнения. x: 5; 8; 11; 14; 17; 20. y: 18; 25; 33; 41; 49; 57.

Задание 3. Интерпретация параметров. Установить связь между расходами на обучение (x, тыс. руб.) и ростом прибыли (y, %). Прокомментировать влияние фактора x на результат. x: 10; 20; 30; 40; 50; 60. y: 2; 5; 8; 12; 15; 19.

Вариант 12

Задание 1. Основы регрессии. Рассчитать свободный член и коэффициент наклона, а также вычислить остаточную дисперсию. x: 100; 200; 300; 400; 500. y: 150; 280; 410; 540; 670.

Задание 2. Анализ значимости. Проверить гипотезу о равенстве нулю коэффициента регрессии (t-тест) для заданного набора данных. x: 3; 7; 11; 15; 19; 23. y: 9; 20; 32; 43; 55; 66.

Задание 3. Интерпретация параметров. Определить зависимость урожайности (y) от количества удобрений (x). Пояснить, как изменится y при увеличении x на одну единицу. x: 1; 2; 3; 4; 5; 6. y: 18; 23; 29; 34; 39; 45.

СПИСОК ЛИТЕРАТУРЫ

1. Гмурман В. Е. Руководство к решению задач по теории вероятностей и математической статистике. М: Высшая школа, 2002. – 405 с.

2. Вуколов Э. Я. Основы статистического анализа. Практикум по статистическим методам и исследованию операций: Учебное пособие. – М.: ФОРУМ: ИНФА-М, 2004. – 464 с.

3. Кремер Н. Ш. Теория вероятностей и математическая статистика: Учебник для вузов. – 2-е изд., перераб. и доп. – М.: ЮНИТИ-ДАНА, 2004. – 573 с.

4. Трофимова Е. А. Теория вероятностей и математическая статистика: учебное пособие / Е. А. Трофимова, Н. В. Кисляк, Д. В. Гилёв; под общей редакцией Е. А. Трофимовой; Министерство образования и науки Российской Федерации, Уральский федеральный университет имени первого Президента России Б. Н. Ельцина. – Екатеринбург: Издательство Уральского университета, 2018. – 160 с.

5. Клячкин В. Н. Статистические методы анализа данных: учеб. пособие / Ю. Е. Кувайскова, В. А. Алексеева; В. Н. Клячкин. – Электрон. изд. – Москва: Финансы и статистика, 2021. – 242 с.: ил. – 1 файл pdf: 240 с.