Министерство образования и науки Российской Федерации

**Федеральное государственное бюджетное образовательное учреждение высшего профессионального образования «Тамбовский государственный технический университет»**

**Н.П. ПУЧКОВ**

# МАТЕМАТИЧЕСКАЯ СТАТИСТИКА. ПРИМЕНЕНИЕ В ПРОФЕССИОНАЛЬНОЙ ДЕЯТЕЛЬНОСТИ

Рекомендовано Учебно-методическим объединением вузов России по университетскому политехническому образованию в качестве учебного пособия для студентов высших учебных заведений, обучающихся по направлению подготовки бакалавров «Инноватика» и направлению подготовки бакалавров «Системный анализ и управление»

Тамбов
Издательство ФГБОУ ВПО «ТГТУ»
2013

УДК 519.22(075.8)
ББК В172я73
П764

Рецензенты:

Доктор физико-математических наук, профессор,
директор Института математики, физики и информатики
ФГБОУ ВПО «ТГУ им. Г.Р. Державина»,
*Е.С. Жуковский*

Доктор технических наук, профессор,
заведующий кафедрой «Информационные процессы и управление»
ФГБОУ ВПО «ТГТУ»
*В.Г. Матвейкин*

**Пучков, Н.П.**

П764 Математическая статистика. Применение в профессиональной деятельности : учебное пособие / Н.П. Пучков. – Тамбов : Изд-во ФГБОУ ВПО «ТГТУ», 2013. – 80 с. – 100 экз.

ISBN 978-5-8265-1191-6.

Содержит базовые понятия математической статистики, изложены методы по использованию математических знаний при решении задач профессиональной деятельности, даны рекомендации по организации самостоятельной работы.

Предназначено для студентов высших учебных заведений, обучающихся по направлению подготовки бакалавров 222000 «Инноватика» и 220100 «Системный анализ и управление».

УДК 519.22(075.8)
ББК В172я73

**ISBN 978-5-8265-1191-6**

# ВВЕДЕНИЕ

Цель науки – описание, объяснение и предсказание явлений действительности на основе установленных законов, что позволяет находить решения в типичных ситуациях.

В основе научных знаний лежит наблюдение. Для обнаружения общей закономерности, которой подчиняется явление, необходимо многократно его наблюдать. Кроме того, многие явления окружающего мира взаимно связаны и влияют одно на другое. Проследить все связи и определить влияние каждой из них на явление не всегда представляется возможным. Поэтому ограничиваются изучением влияния лишь основных факторов, определяющих течение явления.

Сколько должно производиться наблюдений? Как обработать результаты наблюдений и сделать обоснованные практические выводы? Какие факторы и в какой мере учитывать при исследовании явлений? Получить ответы на эти и другие вопросы позволяет математическая статистика.

Для широкого круга явлений при сохранении постоянными основных условий испытаний отмечается неоднозначность полученных результатов. Примером таких случайных явлений служат погрешности измерений. Измеряя один и тот же параметр (предмет), получают близкие, но всё же различные результаты. Это объясняется тем, что результат каждого измерения содержит случайную погрешность. Предвидеть эту погрешность, а следовательно, и результат каждого конкретного измерения нельзя. Однако, если определённым образом систематизировать результаты измерений, то окажется, что в их изменении можно увидеть некоторую закономерность – статистическую устойчивость. Изучение этой закономерности позволяет, например, предвидеть в среднем результат серии измерений.

Математическая статистика – наука, изучающая методы обработки результатов наблюдений массовых случайных явлений, обладающих статистической устойчивостью, закономерностью, с целью выявления этой закономерности. Для вынесения более определённого заключения о закономерностях явлений математическая статистика опирается на теорию вероятностей.

Обработав результаты наблюдений, исследователь выдвигает ряд гипотез, предположений о том, что рассматриваемое явление можно описать той или иной вероятностной теоретической моделью. Далее, используя математико-статистические методы, можно дать ответ на вопрос, какую из гипотез или моделей следует принять. Именно эта модель считается закономерностью изучаемого явления. Правомерен такой вывод или нет, покажет практика использования выбранной модели. Таков типичный путь математико-статистического исследования.

Каждая математическая теория становится более понятной и доступной, если её удаётся использовать для решения практических задач. Чтобы настоящее пособие способствовало каждому обучающемуся, изучающему начала математической статистики, приобретению навыков использования теоретических знаний на практике, мы попытались провести изложение практических примеров применительно к решению следующей профессиональной задачи.

На областном уровне анализируется урожайность одной из зерновых культур, что порождает следующие вопросы: какова средняя урожайность в настоящее время, насколько она неравномерна по районам области, отдельным хозяйствам, какие факторы значимы для повышения урожайности, какие перспективы для планирования на будущее.

Известно, что во всех районах области выращиванием зерновой культуры занимается более полутысячи хозяйств (индивидуальных и коллективных). Проанализировать работу такого большого количества объектов весьма трудоёмко и затратно. Поэтому возникает первая задача (математической статистики) – в каком количестве и каким образом выбирать «экспериментальную группу» хозяйств, чтобы результаты были приемлемы для характеристики работы всех хозяйств области?

Вторая задача – как оценить точность и надёжность результатов анализа показателей работы в экспериментальной группе.

Чтобы планировать повышение урожайности, управлять этим процессом, надо знать, от чего она зависит, от каких факторов. Если эти факторы просматриваются интуитивно (качество почвы, качество её обработки, полив, удобрения и т.д.), то их значимость определяется одним из методов математической статистики – дисперсионным анализом. Если же влияние каких-то неизвестных факторов проявляется неявно (квалификация агронома, опыт работы руководителя хозяйства), то это обстоятельство исследуется методами факторного анализа.

Степень влияния факторов, её количественная оценка осуществляется методами корреляционного анализа. Регрессионный анализ делает возможным найти аналитические зависимости между значениями (неслучайными) факторных переменных (количество внесённых удобрений, объём полива, состав почвы, глубина заделки семян и др.) и средним значением анализируемой случайной величины (урожайности).

Вот в таком плане в данном учебном пособии на примерах решения перечисленных задач демонстрируются методы математической статистики.

В данном пособии теоретический материал не представлен в исчерпывающем объёме и предполагает дополнительное использование учебников и учебных пособий, например тех, которые указаны в списке литературы [1 – 3]. Кроме того, объём заданий для самостоятельного решения рассчитан на «среднего» студента, поэтому закрепление практических навыков можно продолжить, используя задачи из рекомендованных нами задачников [4 – 6].

Учитывая тот факт, что основная задача обучения в вузе – это овладение компетенциями, особыми умениями, мы рекомендуем помимо овладения навыками решения известных задач, попытаться научиться самим составлять задачи, овладевать искусством математического моделирования.

# 1. МАТЕМАТИЧЕСКАЯ СТАТИСТИКА. ОСНОВНЫЕ ПОНЯТИЯ

**I. Учебные цели.** Познакомить студентов с основными понятиями математической статистики, задачами, которые решаются в изучаемом курсе.

В результате изучения материала студенты должны иметь представление о способах сбора статистических данных, о способах их представления в удобной для статистической обработки форме (вариационный ряд, статистическое распределение выборки, полигон, гистограмма, эмпирическая функция распределения), уметь осуществлять наглядное представление статистического распределения, находить числовые характеристики вариационных рядов.

**II. Формирование компетенций.** Развитие математической культуры, совершенствование общей культуры мышления, развитие способностей применять методы математической статистики в профессиональной деятельности, умение лаконично и точно формулировать определения, давать графическую интерпретацию математических зависимостей.

**III. Введение в тему.** Математическая статистика является частью общей прикладной математической дисциплины «Теория вероятностей и математическая статистика», однако задачи, решаемые ею, носят специфический характер. Если теория вероятностей исследует явления, полностью заданные их моделью, то в математической статистике вероятностная модель определена с точностью до неизвестных параметров. Отсутствие сведений о параметрах компенсируется «пробными» испытаниями, на основе которых и восстанавливается недостающая информация. Цель математической статистики состоит в создании методов сбора и обработки статистических данных для получения научных и практических выводов.

**Вопросы для контроля усвоения излагаемого материала**

1. Что является предметом изучения математической статистики?
2. Что такое статистические данные?
3. Какие основные задачи решает математическая статистика?
4. Что такое генеральная и выборочная совокупности?
5. Какие существуют способы образования выборки?
6. Что такое вариационный ряд и статистическое распределение выборки?
7. Графики статистического распределения: полигон и гистограмма.
8. Как задаётся эмпирическая функция распределения?
9. Что такое выборочная средняя и какие у неё свойства?
10. Что такое выборочная дисперсия и какие у неё свойства?

## 1.1. ПРЕДМЕТ МАТЕМАТИЧЕСКОЙ СТАТИСТИКИ

Предметом математической статистики является изучение случайных событий и случайных величин по результатам наблюдения. В основе этой дисциплины лежит понятие статистической совокупности.

Статистической совокупностью называется совокупность предметов или явлений, объединённых каким-либо признаком. Результатом наблюдений над статистической совокупностью являются статистические данные – данные о количестве элементов в какой-либо совокупности, обладающих определённым свойством.

Например:

- количество центнеров зерна, собранного с различных полей;
- количество дождливых дней в году;
- количество жителей города в возрасте 20 лет;
- количество дубов на территории Тамбовской области.

Обработка статистических данных методами математической статистики приводит к установлению определённых закономерностей, присущих массовым явлениям.

Статистические данные, как правило, представляют собой ряд значений $\{x_1, x_2, \ldots, x_n\}$ некоторой случайной величины $X$. Её исследование начинается с обработки этого ряда значений. Затем строятся функции, характеризующие случайную величину $X$. Эти функции сопоставляют по некоторому правилу набору значений случайной величины некоторое число (своего рода характеристику) и называются статистиками.

Простейшей статистикой является, например, среднее значение одинаково распределённых случайных величин $X_1, X_2, \ldots, X_n$. Статистика $T_n = \frac{1}{n}\sum_{i=1}^{n} X_i$.

Можно выделить основные задачи математической статистики, которые решаются в изучаемом курсе:

1. Поиск способов сбора и группировки статистических данных, полученных в результате наблюдений или эксперимента.

2. Разработка методов анализа статистических данных в зависимости от целей исследования:

- оценка неизвестной вероятности события (по сути, использование статистического определения вероятности);
- оценка неизвестной функции распределения;
- оценка параметров (известного) распределения;
- оценка степени зависимости случайных величин;
- проверка статистических гипотез о виде неизвестного распределения или о величине параметров распределения, вид которого известен.

## 1.2. ГЕНЕРАЛЬНАЯ И ВЫБОРОЧНАЯ СОВОКУПНОСТИ

Генеральной совокупностью (ГС) называется совокупность объектов или наблюдений, все элементы которой подлежат изучению при статистическом анализе.

Генеральная совокупность может быть конечной или бесконечной. Число объектов в генеральной совокупности называется её объёмом.

Изучение всего набора элементов генеральной совокупности не всегда бывает возможным, в этом случае рассматривают некоторую часть генеральной совокупности, которую называют выборочной совокупностью (или выборкой).

Задача математической статистики – по результатам изучения свойств выборки «спроектировать» свойства генеральной совокупности. Для того чтобы по выборке можно было адекватно судить об изучаемой величине, она должна быть представительной (репрезентативной), т.е. представлять основные соотношения в генеральной совокупности; это условие обеспечивается случайностью её элементов: все элементы генеральной совокупности должны иметь одинаковую вероятность попадания в выборку.

Поэтому первой задачей математической статистики является поиск способов сбора и группировки статистических данных.

Различают такие способы образования выборки, как:

1) повторная выборка, когда каждый элемент, случайно отобранный и исследованный, возвращается в генеральную совокупность и может быть отобран повторно;

2) бесповторная выборка, когда отобранный элемент не возвращается в генеральную совокупность.

Повторная выборка более приемлемая, так как не нарушает исходное состояние генеральной совокупности, но не всегда возможна по той, например, причине, что может измениться сам элемент после осуществления его выборки.

Каждый из этих способов, в свою очередь, может осуществляться в виде:

- чисто случайная выборка – элемент генеральной совокупности (ГС) попадает в выборку чисто случайно (например, с помощью генератора случайных чисел);
- механическая выборка – ГС делят на столько групп, сколько объектов должно войти в выборку, из каждой берут по одному объекту;
- типическая выборка – выборка не из всей ГС, а из каждой её типической части (при заметном отличии исследуемого признака в различных типических частях);
- серийная выборка – ГС делится на серии и сплошное обследование всей серии (при отсутствии заметного отличия исследуемого признака в различных сериях).

## 1.3. ВАРИАЦИОННЫЙ РЯД И ЕГО ГРАФИЧЕСКОЕ ИЗОБРАЖЕНИЕ

Пусть из генеральной совокупности осуществлена выборка $\{x_1, x_2, \ldots, x_n\}$ объёма $n$. Элементы этой выборки (варианты) представляют собой значения случайной величины $X$ – исследуемого признака. Если они проранжированы по возрастанию, то такое представление называют рядом вариант или вариационным рядом.

Частотой варианты $x_i$ называют число $m_i$, показывающее, сколько раз эта варианта встречается в выборке. Относительной частотой (долей варианты) называют число $w_i = \dfrac{m_i}{n}$.

Количество вариант $m_x$, значения которых меньше некоторого числа $x$, называют накопленной частотой $m_x = \sum\limits_{x_i < x} m_i$.

Вариационные ряды бывают дискретными и интервальными. Вариационный ряд называется дискретным, если он представляет собой выборку значений дискретной величины, и интервальным, если представляет собой выборку значений непрерывной величины.

*Пример* 1.1. Исследуемый признак $X$ – количество хозяйств в районах области, выращивающих пшеницу и попавших в выборку. Пусть таких хозяйств 64, а число укрупнённых географических районов области – 10. Распределение количества хозяйств по номерам выделенных районов таково:

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

Если эти данные проранжировать по возрастанию, то получим вариационный ряд:

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

Для построения интервального вариационного ряда множество значений вариант, заключённых на интервале $[a_1, a_{k+1}]$, разбивают на $k$ полуинтервалов $[a_j, a_{j+1})$ $(j = \overline{1, k})$, последний из которых интервал $[a_k, a_{k+1}]$, т.е. производят их группировку (сгруппированные данные).

Если варианта находится на границе интервала, то её приравнивают к правому интервалу.

*Пример* 1.2. Исследуемый признак – случайная величина $Y$ – урожайность зерновой культуры в хозяйствах области; интересуемый диапазон значений от 18 ц/га до 48 ц/га. Рекомендуемое количество интервалов $k$ выбирают по формуле Стерджерса

$$k = 1 + 1{,}4 \ln d \ ,$$

где $d$ – диапазон изменения признака.

В рассматриваемом случае

$$k = 1 + 1{,}4\ln(48 - 18) \cong 1 + 1{,}4 \cdot 3{,}4 = 5{,}67 \approx 6.$$

Длина каждого малого интервала

$$\Delta = \frac{d}{k} = \frac{30}{6} = 5 \text{ (ц/га)}.$$

Интервальный вариационный ряд имеет вид

[18, 23); [23, 28); [28, 33); [33, 38); [38, 43); [43, 48].

Зачастую, «опорными точками» служат средние на интервалах значения вариант $c_i$, которые подсчитываются как среднеарифметические их граничных (конечных) значений: $c_i = \frac{1}{2}(a_i + a_{i+1})$, $i = \overline{1,k}$.

Статистическим распределением выборки называют ряд вариант, расположенных в порядке возрастания их значений, с соответствующими им частотами (относительными частотами).

В примере 1.1 статистическое распределение выборки имеет вид

| Варианты | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|
| Частоты | 1 | 0 | 1 | 1 | 1 | 3 | 1 | 2 |

Пусть в примере 1.2 объём выборки составил 64 хозяйства и в первый интервал попало 7 хозяйств, во второй – 11, в третий – 16, в четвёртый – 14, в пятый – 10, в шестой – 6. Соответствующее статистическое распределение выборки имеет вид

| Варианты | [18, 23) | [23, 28) | [28, 33) | [33, 38) | [38, 43) | [43, 48] |
|---|---|---|---|---|---|---|
| Средние значения | 20,5 | 25,5 | 30,5 | 35,5 | 40,5 | 45,5 |
| Частоты | 7 | 11 | 16 | 14 | 10 | 6 |

Для наглядности представления статистического распределения используются различного рода графики: полигон и гистограмма.

Полигон (частот, относительных частот) используется в случае дискретного вариационного ряда и представляет собой ломаную, соединяющую точки плоскости с координатами ($x_i$, $m_i$) или ($x_i$, $w_i$), $i = \overline{1, n}$; $n$ – количество вариант.

Так, для примера 1.1 полигон частот изображён на рис. 1.1.

Для интервального ряда также строится полигон, только его ломаная проходит через точки $(c_i, m_i)$, где $c_i$ – средние на интервалах значения.

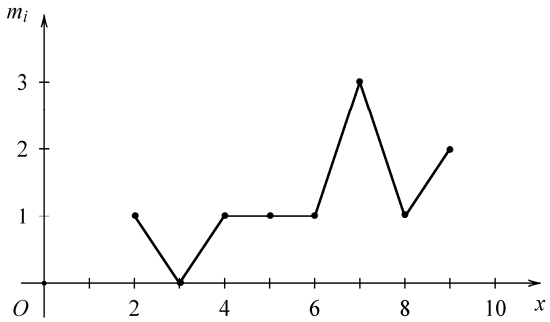


**Рис. 1.1**

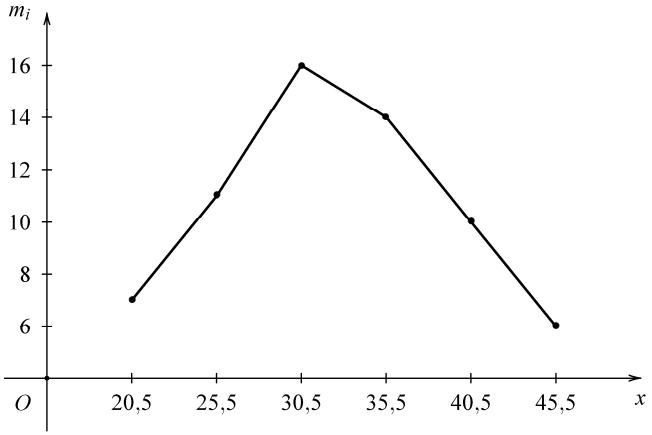


**Рис. 1.2**

Для примера 1.2 полигон частот изображён на рис. 1.2.

Гистограмма служит для представления только интервальных вариационных рядов и имеет вид ступенчатой фигуры, состоящей из прямоугольников с основаниями, равными длине интервалов $\Delta$ и высотами, равными

$$W_i = M_i/\Delta, \quad i = \overline{1, k},$$

где $M_i$ – сумма частот вариант, попавших в $i$-й интервал, $M_i/\Delta$ – плотность частоты. Таким образом, площадь каждого прямоугольника равна $\Delta \cdot M_i/\Delta = M_i$ – сумме частот.

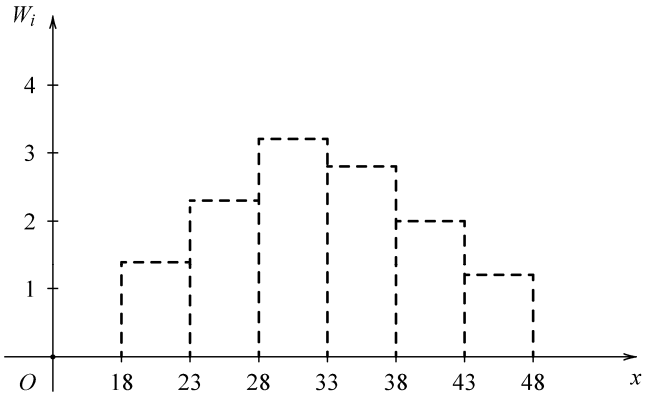


**Рис. 1.3**

Для примера 1.2 гистограмма имеет вид, представленный на рис. 1.3. Здесь $W_1 = 1{,}4$; $W_2 = 2{,}2$; $W_3 = 3{,}2$; $W_4 = 2{,}8$; $W_5 = 2$; $W_6 = 1{,}2$.

## 1.4. ЧИСЛОВЫЕ ХАРАКТЕРИСТИКИ ВАРИАЦИОННЫХ РЯДОВ

Как следует из материала прошлого параграфа, вариационные ряды (выборки) можно охарактеризовать с помощью статистического распределения. На практике бывает достаточно иметь характеристики вариационных рядов в виде отдельных чисел, а именно: выборочной средней, выборочной дисперсии, выборочного среднеквадратического отклонения.

Пусть дискретный вариационный ряд задан статистическим распределением:

| Варианты | $x_1$ | $x_2$ | … | $x_k$ |
|---|---|---|---|---|
| Частоты | $m_1$ | $m_2$ | … | $m_k$ |

$\sum_{i=1}^{k} m_i = n$ – объём выборки, $k$ – число вариант.

Выборочным средним называется величина

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{k} x_i m_i . \qquad (1.1)$$

Если статистические данные не являются сгруппированными, т.е. $m_1 = m_2 = \ldots = m_k = 1$, то выборочное среднее есть не что иное, как среднее арифметическое значений вариант $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ .

Формулой (1.1) можно пользоваться и для характеристики интервального вариационного ряда в виде

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{k} c_i\, m_i\,,$$

где $c_i$ – середина $i$-го интервала; $m_i$ – сумма частот вариант, попавших в $i$-й интервал; $k$ – число интервалов.

Свойства выборочной средней аналогичны свойствам математического ожидания случайной величины (в теории вероятностей). Укажем одно из них, необходимое для дальнейшей работы: если вариационный ряд состоит из нескольких групп, то общая выборочная средняя равна

$$\bar{x} = \sum_{i=1}^{l} \bar{x}_i \frac{n_i}{n}\,, \qquad (1.2)$$

где $\bar{x}_i$ – групповые средние; $n_i$ – объёмы групп; $l$ – число групп.

*Пример* 1.3. Дано распределение признака $X$

| $x_i$ | 2 | 3 | 4 | 8 | 9 | 12 |
|---|---|---|---|---|---|---|
| $m_i$ | 1 | 2 | 2 | 1 | 2 | 2 |

, $\sum m_i = 10$.

Общее выборочное среднее

$$\bar{x} = \frac{1}{10}\sum_{i=1}^{6} x_i m_i = 0{,}1(2+6+8+8+18+24) = 6{,}6\,.$$

Выделим две группы вариант: чётных и нечётных

I-я группа:

| $x_i$ | 2 | 4 | 8 | 12 |
|---|---|---|---|---|
| $m_i$ | 1 | 2 | 1 | 2 |

,

II-я группа:

| $x_i$ | 3 | 9 |
|---|---|---|
| $m_i$ | 2 | 2 |

,

для которых групповые выборочные средние: $\bar{x}_1 = 7$ ; $\bar{x}_2 = 6$.

По формуле (1.2):

$$\bar{x} = \frac{6}{10}\cdot 7 + \frac{4}{10}\cdot 6 = \frac{42+24}{10} = 6{,}6\,,$$

что и требовалось получить.

Выборочной дисперсией называется среднее арифметическое квадратов отклонений вариант от их выборочной средней:

$$S^2 = \frac{1}{n}\sum_{i=1}^{k} (x_i - \bar{x})^2\, m_i\,,\ n = \sum_{i=1}^{k} m_i\,.$$

Если $m_i = 1$, $i = \overline{1, n}$, то

$$S^2 = \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^2\,. \qquad (1.3)$$

Для интервального вариационного ряда

$$S^2 = \frac{1}{n}\sum_{i=1}^{k}(c_i - \bar{x})^2 ,$$

где $c_i$ – середина $i$-го интервала.

Для практических вычислений $S^2$ более удобной является формула

$$S^2 = \overline{x^2} - (\bar{x})^2 ,$$

где $\bar{x}^2$ – выборочная средняя квадратов вариационного ряда.

Выборочное среднее квадратическое отклонение определяется как квадратный корень из дисперсии: $S = \sqrt{S^2}$.

Для выборочной дисперсии справедливо свойство, которое лежит в основе раздела математики, называемого дисперсионный анализ, и состоит в том, что если вариационный ряд состоит из нескольких групп, то общая дисперсия равна сумме средней групповых дисперсий и межгрупповой дисперсии.

Пусть варианты выборки имеют обозначение $x_{ij}$, где $i$ = 1, 2, …, $l$ – номер группы, $j$ = 1, 2, …, $k_i$ – номер варианты в $i$-й группе, $m_{ij}$ – соответствующая этой варианте частота, $\bar{x}_i$ – групповые средние, $\bar{x}$ – общая выборочная средняя, $n_i$ – объём $i$-й группы, $n$ – объём выборки.

Тогда

$S_0^2 = \sum_{i=1}^{l}\sum_{j=1}^{k_i}(x_{ij} - \bar{x})^2 \frac{m_{ij}}{n}$ – общая выборочная дисперсия;

$\delta^2 = \sum_{i=1}^{l}(\bar{x}_i - \bar{x})^2 \frac{n_i}{n}$ – межгрупповая дисперсия;

$S_i^2 = \sum_{j=1}^{k_i}(x_{ij} - \bar{x}_i)^2 \frac{m_{ij}}{n_i}$ – групповые дисперсии, а

$\bar{S}_{(i)}^2 = \sum_{i=1}^{l} S_i^2 \frac{n_i}{n}$ – их средняя.

Свойство дисперсии:

$$S_0^2 = \bar{S}_{(i)}^2 + \delta^2 . \qquad (1.4)$$

Формула (1.4) может быть получена методом «разложения суммы квадратов» (см. [1, с. 286]).

Возвращаясь к примеру 1.3, найдём:

$$S_0^2 = 1/10[(2-6{,}6)^2 \cdot 1 + (3-6{,}6)^2 \cdot 2 + (4-6{,}6)^2 \cdot 2 + (8-6{,}6)^2 \cdot 1 +$$
$$+ (9-6{,}6)^2 \cdot 2 + (12-6{,}6)^2 \cdot 2] = 13{,}24.$$

$$\overline{S}^2_{(i)} = 1/10[(2-7)^2 \cdot 1 + (4-7)^2 \cdot 2 + (8-7)^2 \cdot 1 + (12-7)^2 \cdot 2 + (3-6)^2 \cdot 2 +$$

$$+ (9-6)^2 \cdot 2] = 13.$$

$$\delta^2 = (7-6{,}6)^2 \cdot \frac{6}{10} + (6-6{,}6)^2 \cdot \frac{4}{10} = 0{,}24.$$

Действительно, $S_0^2 = \overline{S}^2_{(i)} + \delta^2$.

Одновременно можно сделать вывод о том, что выделенные группы «равномощные», так как их межгрупповая дисперсия ничтожно мала по сравнению с общей.

*Пример* 1.4. По данным статистического распределения выборки в примере 1.2 (с. 8) найти среднее выборочное урожайности по 64 хозяйствам и выборочную дисперсию относительно среднего.

Имеем:

$$\bar{x} = \frac{1}{64}\left[20{,}5 \cdot 7 + 25{,}5 \cdot 11 + 30{,}5 \cdot 16 + 35{,}5 \cdot 14 + 40{,}5 \cdot 10 + 45{,}5 \cdot 6\right] =$$

$$= \frac{1}{64} \cdot 2087 \cong 32{,}6, \text{ ц/га}.$$

$$S^2 = \frac{1}{64}\left[12{,}1^2 \cdot 7 + 7{,}1^2 \cdot 11 + 2{,}1^2 \cdot 16 + 2{,}9^2 \cdot 14 + 7{,}9^2 \cdot 10 + 12{,}9^2 \cdot 6\right] =$$

$$= \frac{1}{64}\left[1024{,}87 + 554{,}51 + 70{,}59 + 117{,}74 + 624{,}1 + 998{,}46\right] = \frac{3390{,}27}{64} = 52{,}97.$$

Среднее квадратическое отклонение $S$ = 7,28 (ц/га).

## 1.5. ЗАДАЧИ ДЛЯ САМОСТОЯТЕЛЬНОГО РЕШЕНИЯ

**1.5.1.** Анализируется успеваемость по математике на учебном курсе. В выборку попали две студенческие группы (*A* и *B*) численностью 23 и 27 человек.

Экзамен был организован в форме тестирования по 100-балльной шкале. Результат – выборочные данные были сгруппированы в интервалы по 20 баллов с минимальным баллом 20 и представлены в таблице:

| Баллы | Группа *A* | Группа *B* |
|---|---|---|
| [100 – 80) | 4 | 3 |
| [80 – 60) | 8 | 10 |
| [60 – 40) | 9 | 11 |
| [40 – 20] | 2 | 3 |

Определите группу, где средний балл тестирования выше, а также группу, где разброс результатов меньше.

**1.5.2.** В течение месяца (30 дней) государственная автоинспекция зарегистрировала 77 аварий. Распределение количества аварий по числу дней представлено следующей таблицей:

| Количество аварий | 0 | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| Число дней | 6 | 4 | 5 | 3 | 6 | 6 |

Найти выборочное среднее аварийности в день и среднее квадратическое отклонение от этого среднего.

Построить гистограмму частот и относительных частот.

**1.5.3.** В течение 1 часа (60 минут) в офис компании поступило 50 звонков. Хронология звонков выглядит следующим образом:

| Временной интервал | [0, 10) | [10, 20) | [20, 30) | [30, 40) | [40, 50) | [50, 60] |
|---|---|---|---|---|---|---|
| Количество звонков | 15 | 7 | 10 | 4 | 6 | 8 |

Построить гистограмму частот по данному распределению выборки.

Найти среднее выборочное количества звонков и соответствующее среднее квадратическое отклонение относительно этого среднего.

**1.5.4.** В примере 1.3 выделены две группы

I-я группа:

| $x_i$ | 2 | 3 | 4 |
|---|---|---|---|
| $m_i$ | 1 | 2 | 2 |

II-я группа:

| $x_i$ | 8 | 9 | 12 |
|---|---|---|---|
| $m_i$ | 1 | 2 | 2 |

Найти:

- групповые средние $\bar{x}_1$ и $\bar{x}_2$;
- групповые дисперсии $S_1^2$ и $S_2^2$, а также их среднюю $\bar{S}_{(i)}^2$;
- межгрупповую дисперсию $\delta^2$;
- долю межгрупповой дисперсии $\delta^2$ от общей дисперсии $S_0^2$.

**1.5.5.** В таблице приведено распределение 50 рабочих по производительности труда $X$ (единиц за смену), разделённых на две группы: 30 и 20 человек:

| | I. Прошедшие обучение | | | | | II. Не прошедшие обучение | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $x_i$ | 34 | 85 | 96 | 102 | 103 | 63 | 69 | 83 | 89 | 106 |
| $m_i$ | 5 | 2 | 11 | 8 | 4 | 2 | 6 | 8 | 3 | 1 |

Вычислить общие, групповые средние и дисперсии и убедиться в справедливости правила сложения дисперсий.

# 2. ОЦЕНКА ПАРАМЕТРОВ ГЕНЕРАЛЬНОЙ СОВОКУПНОСТИ

**I. Учебные цели.** Познакомить студентов с постановкой одной из основных задач математической статистики – задачей оценки неизвестных параметров известного распределения, научить применять на практике метод выбора величин для оценки – метод наибольшего правдоподобия, сформировать умение осуществлять точечные и интервальные оценки неизвестных параметров.

**II. Формирование компетенций.** Формировать математическую культуру, развивать аналитическое и логическое мышление, развивать способность к обобщению.

**III. Введение в тему.** Содержание данного раздела можно сформулировать как совокупность методов, позволяющих делать научно обоснованные выводы о числовых параметрах распределения генеральной совокупности по случайной выборке из неё. Если, например, нас интересует математическое ожидание генеральной совокупности, то задача статистической оценки параметров заключается в том, чтобы найти такую выборочную характеристику, которая позволила бы получить по возможности более точное и надёжное представление об интересующем нас параметре. Так как состав выборки случаен, то выводы, сделанные в этих условиях, носят тоже случайный характер. С увеличением объёма выборки вероятность правильного вывода увеличивается. Поэтому всякому решению, принимаемому при статистической оценке параметров, стараются поставить в соответствие вероятность, характеризующую степень достоверности принятого решения.

При изучении материала обратите внимание на следующие вопросы для контроля качества его усвоения:

1. Каким требованиям должна удовлетворять точечная оценка?
2. В чём сущность метода наибольшего правдоподобия?
3. Как найти оценку параметра распределения Пуассона методом наибольшего правдоподобия?
4. Что является точечными оценками параметров нормального распределения?
5. Какова сущность интервальных оценок?
6. Как построить доверительные интервалы для параметров нормального распределения?

## 2.1. ТОЧЕЧНЫЕ ОЦЕНКИ

Различают точечные оценки (одним числом) и интервальные (парой чисел – границ интервала, с заданной вероятностью накрывающего оцениваемый параметр).

Рассмотрим первоначально точечные оценки.

Пусть $\theta$ – оцениваемый параметр, постоянное (неслучайное) число. Оценкой $\overline{\theta}_n$ параметра $\theta$ называется любая функция от значения выборки $\overline{\theta}_n = \overline{\theta}_n (x_1, x_2, \ldots, x_n)$, т.е. статистика (поэтому оценка называется статистической). Если принять утверждение, что $x_i$ является реализацией случайной величины $X_i$, то статистику $\overline{\theta}_n$ можно рассматривать как функцию от случайных величин $X_1, X_2, \ldots, X_n$.

Статистику $\overline{\theta}_n$ надо выбирать таким образом, чтобы её значения как можно точнее оценивали значение неизвестного параметра $\theta$. Различают следующие требования к оценке $\overline{\theta}_n$.

1. Состоятельность – при больших объёмах выборки $\overline{\theta}_n$ как угодно мало отличается от $\theta$ ($\overline{\theta}_n$ стремится к $\theta$ по вероятности):

$$\lim_{n\to\infty} P\left\{\left|\theta_n - \theta\right| < \varepsilon\right\} = 1.$$

2. Несмещённость – её математическое ожидание должно быть равно оцениваемому параметру $M\,\overline{\theta}_n = \theta$.

3. Эффективность – при одном и том же объёме выборки её дисперсия минимальна среди всевозможных оценок:

$$D(\overline{\theta}_n) = M(\overline{\theta}_n - \theta)^2 - \min.$$

При этом рассматриваются только несмещённые оценки.

Чтобы определить, какая величина может быть выбрана в качестве оценки $\overline{\theta}_n$, существуют различные методы. Один из них – метод наибольшего правдоподобия (МНП).

Рассмотрим сущность МНП на примере дискретной случайной величины $X$, которая в результате $n$ испытаний приняла значения $x_1, x_2, \ldots, x_n$.

Допустим, что вид закона распределения величины $X$ задан, но неизвестен параметр $\theta$, который определяет этот закон. Требуется найти его точечную оценку $\overline{\theta}_n$. Например, если $X$ имеет распределение Пуассона

$$P_m(k) = \frac{\lambda^k e^{-\lambda}}{k!},$$

где $\lambda$ – неизвестный параметр, то необходимо получить точечную оценку для $\lambda$.

Обозначим $p(x_i; \theta)$ – вероятность того, что в результате испытания величина $X$ примет значения $x_i$ ($i = 1, 2, \ldots, n$).

Функцией правдоподобия дискретной случайной величины $X$ называют функцию

$$L(x_1, x_2, \ldots, x_n; \theta) = p(x_1; \theta)\, p(x_2; \theta) \cdot \ldots \cdot p(x_n; \theta),$$

где $x_1, x_2, \ldots, x_n$ – фиксированные числа, а $\theta$ – единственный аргумент этой функции, он же – неизвестный параметр известного закона распределения.

Если рассмотреть правую часть последнего равенства, то это – произведение вероятностей, т.е. та же вероятность, которая, по сути, должна быть максимально возможной. Поэтому параметр $\theta$ имеет наиболее точную оценку $\theta^* = \theta^*(x_1, x_2, ..., x_n)$ при таком значении $\theta^*$, когда $L$ достигает максимального значения. Такое значение можно найти, используя следующий алгоритм поиска экстремума функции одной переменной:

1) найдём производную $\frac{dL}{d\theta}$;

2) из уравнения $\frac{dL}{d\theta} = 0$ находим $\theta^*$;

3) найдём $\frac{d^2L}{d\theta^2}$ и подсчитаем её при $\theta = \theta^*$,

если результат – отрицательное число, то $\theta^*$ – точка максимума, которая принимается за наиболее правдоподобную оценку неизвестного параметра.

Структурно функция $L$ представляет собой произведение $n$ функций $p(x_i; \theta)$, $i = 1, 2, \ldots, n$, поэтому целесообразно использовать логарифмическое дифференцирование, тем более, что функции $L$ и $\ln L$ достигают максимального значения при одном и том же значении аргумента $\theta$.

Метод наибольшего правдоподобия имеет ряд достоинств: полученные оценки состоятельны (хотя могут быть и смещёнными), распределены асимптотически нормально и имеют наименьшую дисперсию (по сравнению с другими асимптотическими оценками). Этот метод наиболее полно использует данные выборки об оцениваемом параметре, поэтому он ценен в случае малых выборок.

Недостаток метода – иногда требует сложных вычислений.

*Пример* 2.1. Найти методом наибольшего правдоподобия оценку параметра $\lambda$ распределения Пуассона

$$P_m(X = x_i) = \frac{\lambda^{x_i} e^{-\lambda}}{x_i!},$$

где $x_i$ – число появлений события в $i$-м опыте ($i = 1, 2, \ldots, n$), $x_i > 0$; опыт состоит из $m$ испытаний.

*Решение.* Составим функцию правдоподобия, учитывая, что $\lambda$ – неизвестный параметр:

$$L = p(x_1; \lambda)\, p(x_2; \lambda) \ldots p(x_n; \lambda) = \frac{\lambda^{x_1} e^{-\lambda}}{x_1!} \frac{\lambda^{x_2} e^{-\lambda}}{x_2!} \ldots \frac{\lambda^{x_n} e^{-\lambda}}{x_n!} = \frac{\lambda^{\sum x_i} e^{-n\lambda}}{x_1! x_2! \ldots x_n!}.$$

Логарифмическая функция правдоподобия:

$$\ln L = \sum x_i \ln \lambda - n\lambda - \ln(x_1! x_2! \ldots x_n!)\,;\ \sum x_i = \sum_{i=1}^{n} x_i\,.$$

Её первая производная по $\lambda$:

$$\frac{d \ln L}{d\lambda} = \frac{\sum x_i}{\lambda} - n = 0\,,\ \lambda = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}_B\ -$$ выборочная средняя.

Вторая производная по $\lambda$:

$$\frac{d^2 \ln L}{d\lambda^2} = -\frac{\sum x_i}{\lambda^2} < 0,$$ так как $\sum_{i=1}^{n} x_i > 0\,,\ \lambda^2 > 0\,.$

Вывод: $\lambda = \bar{x}_B$ – точка максимума, и значит в качестве оценки наибольшего правдоподобия параметра $\lambda$ распределения Пуассона надо принять выборочную среднюю $\lambda^* = \bar{x}_B$.

Аналогичным образом можно показать, что в случае нормального закона распределения, характеризующегося двумя параметрами: $a$ – математическое ожидание и $\sigma$ – среднее квадратическое отклонение, в качестве их оценок рассматривают выборочную среднюю и выборочное среднее квадратическое отклонение $S = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$ ($S^2$ – выборочная дисперсия).

Выборочная средняя $\bar{x}$ является несмещённой оценкой для генеральной средней, так как $M(\bar{x}) = M\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) = \frac{1}{n}\sum_{i=1}^{n} MX_i = \frac{1}{n}\sum_{i=1}^{n} a = \frac{1}{n} n a = a\,.$

Существуют доказательства, что эта оценка является состоятельной и эффективной.

Можно показать, что $MS^2 = \frac{n-1}{n}\sigma^2$, т.е. оценка $S^2$ является смещённой; в то же время оценка $\bar{S}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$ является несмещённой для $\sigma^2$; $\bar{S}^2 = \frac{n}{n-1}S^2$ – исправленная выборочная дисперсия. $S^2$ и $\bar{S}^2$ являются состоятельными оценками для $\sigma^2$, а $\bar{S}^2$ и эффективной. Будем далее обозначать $\bar{S}$ – несмещённую оценку для $\sigma$.

Если генеральная совокупность из $N$ элементов содержит $M$ элементов, обладающих признаком $A$, то генеральной долей признака $A$ называется величина $p = \frac{M}{N}$.

Для доли $p$ несмещённой и состоятельной оценкой будет выборочная доля

$$\omega = \frac{m}{n},$$

где $m$ – число элементов выборки, обладающих признаком $A$, $n$ – объём выборки*.

## 2.2. ИНТЕРВАЛЬНЫЕ ОЦЕНКИ

Точечная оценка $\overline{\theta}_n$ параметра θ даёт лишь его некоторое приближённое значение и не содержит информации о точности и надёжности. В тех случаях, когда это необходимо, используют интервальную оценку параметра θ, находят интервал, который с заданной вероятностью γ накрывает неизвестное значение θ. Такой интервал называется доверительным интервалом, а вероятность γ – доверительной вероятностью или уровнем надёжности.

Доверительный интервал определяется из формулы

$$p\left\{\left|\overline{\theta}_n - \theta\right| < \Delta\right\} = \gamma \text{ и имеет вид } \overline{\theta}_n - \Delta < \theta < \overline{\theta}_n + \Delta.$$

Последнее неравенство выполняется с вероятностью γ, а наибольшее отклонение $\Delta$ выборочного значения параметра $\overline{\theta}_n$ от его значения θ называется предельной ошибкой выборки: $\Delta = \Delta(\gamma)$.

### 2.2.1. Доверительные интервалы для генеральной средней и генеральной доли признака

Доверительный интервал уровня надёжности γ для генеральной средней $a$ имеет вид $\overline{x} - \Delta < a < \overline{x} + \Delta$.

Выбор формулы для Δ зависит от объёма выборки и от её вида:

– для повторной выборки $\Delta = t\dfrac{\overline{S}}{\sqrt{n}}$,

– для бесповторной $\Delta = t\dfrac{\overline{S}}{\sqrt{n}}\sqrt{1-\dfrac{n}{N}}$.

При $n \geq 30$ $t$ находят из уравнения

$$\Phi(t) = \gamma,\ \Phi(t) = \frac{2}{\sqrt{2\pi}}\int_0^t e^{-x^2/2}dx \text{ – функция Лапласа,}$$

* Если специально не оговаривается, то рассматриваются повторные выборки.

значения которой даются в табличной форме [1, с. 555]. Задавая требуемый уровень надёжности $\gamma$, по таблицам можно найти соответствующее значение параметра $t$. Например, если $\gamma = 0{,}95$, то $t = 1{,}96$; если $\gamma = 0{,}97$, то $t = 2{,}17$; если $\gamma = 0{,}99$, то $t = 2{,}58$ и т.п.

При $n \le 30$ (и только для нормальной генеральной совокупности) $t$ определяется из условия $p\{|\xi| < t\} = \gamma$, где случайная величина $\xi$ имеет распределение Стьюдента с $n - 1$ степенью свободы [1, с. 557].

Например, если $\gamma = 0{,}95$ ($p = 0{,}05$) и $\nu = 29$, то $t = 2{,}04$; если $\gamma = 0{,}99$ ($p = 0{,}01$) и $\nu = 6$, то $t = 3{,}71$.

Доверительный интервал для генеральной доли $p$: $(\omega - \Delta,\ \omega + \Delta)$.

При $n > 30$

– для повторной выборки $\Delta = t\sqrt{\dfrac{\omega(1-\omega)}{n}}$,

– для бесповторной выборки $\Delta = t\sqrt{\dfrac{\omega(1-\omega)}{n}}\sqrt{1-\dfrac{n}{N}}$, где $t$ определяется из равенства $\Phi(t) = \gamma$.

При решении статистических задач часто требуется определить необходимый объём выборки для достижения требуемой надёжности доверительного интервала (обратная задача относительно задачи нахождения предельной ошибки).

Например, при $n > 30$ и повторной выборке предельная ошибка для генеральной доли $\Delta = t\sqrt{\dfrac{\omega(1-\omega)}{n}}$, откуда $\Delta^2 = t^2\dfrac{\omega(1-\omega)}{n}$, а $n = \dfrac{t^2\omega(1-\omega)}{\Delta^2}$.

*Пример* 2.2. Используя результаты примеров 1.2; 1.4 при объёме генеральной совокупности $N = 512$ найдём доверительный интервал для оценки средней урожайности $a$ по всем хозяйствам области с надёжностью $\gamma = 0{,}95$.

*Решение*: $\bar{x} - \Delta < a < \bar{x} + \Delta$.

Имеем $\bar{x} = 32{,}6$; $\bar{S} = 7{,}34$; объём выборки $n = 64 > 30$, выборка бесповторная, поэтому

$$\Delta = t\frac{\bar{S}}{\sqrt{n}}\sqrt{1-\frac{n}{N}} = t\frac{7{,}34}{\sqrt{64}}\sqrt{1-\frac{64}{512}} \cong 0{,}858t,$$

где $t$ определяется из уравнения $\Phi(t) = 0{,}95$ и равно 1,96.

Тогда $\Delta = 1{,}68$ и $32{,}6 - 1{,}68 < a < 32{,}6 + 1{,}68$ или $30{,}92 < a < 34{,}28$.

Вывод: с вероятностью 95% можно утверждать, что средняя урожайность по области не менее, чем 30,92 ц/га и не больше, чем 34,28.

### 2.2.2. Доверительный интервал для генеральной дисперсии (среднего квадратического отклонения)

Если $\{x_1, x_2, \ldots, x_n\}$ выборка из нормальной совокупности и $a$ и $\sigma$ неизвестны, то статистика $Z = \dfrac{n\overline{S}^2}{\sigma^2}$ имеет распределение $\chi^2_{n-1}$ [4, с. 465].

Доверительный интервал для $\sigma^2$ находят из соотношения

$$p\left\{\frac{n\overline{S}^2}{z_2} < \sigma^2 < \frac{n\overline{S}^2}{z_1}\right\} = \gamma; \quad p\left\{z_1 < \frac{n\overline{S}^2}{\sigma^2} < z_2\right\} = \gamma.$$

В свою очередь $z_1$ и $z_2$ определяются из условия $p\left\{z_1 < \chi^2_{n-1} < z_2\right\} = \gamma$, а на практике таким образом, чтобы $p\left\{\chi^2_{n-1} < z_1\right\} = p\left\{\chi^2_{n-1} > z_2\right\} = \dfrac{1-\gamma}{2} = \dfrac{1}{2} - \dfrac{\gamma}{2}$.

В этом случае: $p\left\{\chi^2_{n-1} > z_1\right\} = \dfrac{1+\gamma}{2}$; $p\left\{\chi^2_{n-1} > z_2\right\} = \dfrac{1-\gamma}{2}$ – условия для выбора величин $z_1$ и $z_2$, которые можно найти из таблиц распределения $\chi^2_{n-1}$.

Для *примера* 2.2 необходимо найти доверительный интервал для $\sigma^2$ при $\gamma = 0{,}90$. Тогда $\dfrac{1+\gamma}{2} = 0{,}95$; $\dfrac{1-\gamma}{2} = 0{,}05$.

При $n = 64$ $p\left\{\chi^2_{63} > z_1\right\} = 0{,}95$; $p\left\{\chi^2_{63} > z_2\right\} = 0{,}05$. Из таблиц значений $\chi^2$ – находим, что $z_1 = 45{,}74$; $z_2 = 82{,}53$.

$$\frac{n\overline{S}^2}{z_2} = \frac{64 \cdot 53{,}81}{82{,}53} = 41{,}73; \quad \frac{n\overline{S}^2}{z_1} = \frac{64 \cdot 53{,}81}{45{,}74} = 75{,}29$$

$$41{,}73 < \sigma^2 < 75{,}29, \text{ а } 6{,}46 < \sigma < 8{,}8.$$

*Пример* 2.3. На складе хранится партия товара в количестве 625 штук с различными сроками хранения (от 2 до 10 месяцев). Осуществлена бесповторная выборка товара в количестве 25 штук; при этом оказалось, что 8 единиц товара имеют срок хранения 2 месяца, 6 единиц – 5 месяцев; 4 единицы – 6 месяцев и 67 единиц – 10 месяцев.

Найти:

а) выборочный средний срок хранения товара ($\overline{x}$);

б) выборочную (исправленную) дисперсию срока хранения ($\overline{S}^2$);

в) доверительный интервал для оценки среднего срока хранения товара $a$ с надёжностью 0,95;

г) доверительный интервал для оценки среднего квадратического отклонения $\sigma$ с надёжностью 0,95.

*Решение*. Имеем распределение выборки:

| $x_i$ | 2 | 5 | 6 | 10 |
|---|---|---|---|---|
| $m_i$ | 8 | 6 | 4 | 7 |

$n = 25$; $N = 625$.

а) $\bar{x} = \frac{1}{25}(2\cdot 8 + 5\cdot 6 + 6\cdot 4 + 10\cdot 7) = 5{,}6$;

$\overline{x^2} = \frac{1}{25}(4\cdot 8 + 25\cdot 6 + 36\cdot 4 + 100\cdot 7) = 41{,}04$;

б) $\bar{S}^2 = \frac{25}{24}\left(\overline{x^2} - (\bar{x})^2\right) = 10{,}08$; $\bar{S} = 3{,}17$;

в) $\bar{x} - \Delta < a < \bar{x} + \Delta$; $n = 25 < 30$, $\Delta = t_{n-1}\frac{\bar{S}}{\sqrt{n}}\sqrt{1 - \frac{n}{N}}$,

$t_{n-1} = t_{24}$ из таблицы распределения Стьюдента при доверительной вероятности $\gamma = 0{,}95$: $t_{24} = 2{,}06$.

Имеем: $\Delta = 2{,}06\frac{3{,}17}{\sqrt{25}}\left(\sqrt{1 - \frac{25}{625}}\right) \cong 1{,}28$;

$5{,}6 - 1{,}28 < a < 5{,}6 + 1{,}28$; $4{,}32 < a < 6{,}88$;

г) имеем: $P\left\{\frac{n\bar{S}^2}{z_2} < \sigma^2 < \frac{n\bar{S}^2}{z_1}\right\} = \gamma$;

$P\left\{\chi^2_{24} > z_1\right\} = \frac{1 + 0{,}95}{2} = 0{,}975$, по таблицам $\chi^2$ $z_1 = 12{,}4$;

$P\left\{\chi^2_{24} > z_2\right\} = \frac{1 - 0{,}95}{2} = 0{,}025 \Rightarrow z_2 = 39{,}4$;

$\frac{n\bar{S}^2}{z_2} = \frac{25\cdot 10{,}08}{39{,}4} = 6{,}4$; $\frac{n\bar{S}^2}{z_1} = \frac{25\cdot 10{,}08}{12{,}4} = 20{,}32$;

$6{,}4 < \sigma^2 < 20{,}32$; $2{,}53 < \sigma < 4{,}5$.

*Пример* 2.4. На факультете 1000 студентов; среди выбранных 50 студентов 40 получают стипендии. Найти с надёжностью 0,97 доверительный интервал для доли стипендиатов среди студентов для бесповторной выборки.

*Решение*. Имеем: $N = 1000$, $n = 50$, $m = 40$, $\gamma = 0{,}97$, $\alpha = 1 - \gamma = 0{,}03$.

$\omega - \Delta < p < \omega + \Delta$ – доверительный интервал для доли $p$. Выборочная доля $\omega = \frac{m}{n} = \frac{40}{50} = 0{,}8$. При бесповторной выборке предельная ошибка $\Delta = t\sqrt{\frac{\omega(1-\omega)}{n}}\sqrt{1 - \frac{n}{N}}$.

При $\gamma = 0{,}97$ из уравнения $\Phi(t) = \gamma$, $t = 2{,}17$.

$$\Delta = 2{,}17\sqrt{\frac{0{,}8 \cdot 0{,}2}{50}}\sqrt{1 - \frac{50}{1000}} = 0{,}123 \cdot 0{,}975 = 0{,}12.$$

Таким образом, $0{,}8 - 0{,}12 < p < 0{,}8 + 0{,}12$ или $0{,}68 < p < 0{,}92$.

*Пример* 2.5. На факультете 1000 студентов. Каков должен быть объём повторной выборки, чтобы с надёжностью 0,95 предельная ошибка выборки равна $0{,}2\sigma$?

Имеем: $\Delta = t\dfrac{\overline{S}}{\sqrt{n}}$, откуда $\sqrt{n} = \dfrac{t\overline{S}}{\Delta}$, $n = \dfrac{t^2\overline{S}^2}{\Delta^2}$.

При $\gamma = 0{,}95$ получим $t = 1{,}96$. Для повторной выборки (меняя $\overline{S}$ на $\sigma$) получим

$$n = \frac{1{,}96^2 \cdot \sigma^2}{(0{,}2\sigma)^2} = \frac{3{,}8416}{0{,}04} \cong 96.$$

## 2.3. ЗАДАЧИ ДЛЯ САМОСТОЯТЕЛЬНОГО РЕШЕНИЯ

**2.3.1.** В результате пяти измерений получены следующие результаты (верхнего) кровяного давления: 122, 124, 133, 135, 136. Найти выборочную среднюю, выборочную и исправленную дисперсии.

**2.3.2.** Некоторый товар в количестве 1000 кг решили разделить на три сорта в зависимости от качества и продавать по цене 60, 90 и 100 р. за килограмм. Для оценки возможных доходов от продажи сделали выборку в 30 кг. Оказалось, что она содержит товара низшего сорта – 5 кг, среднего – 10 кг и высшего – 15 кг. Выборка повторная.

Найти:

1. Выборочную среднюю стоимости товара.
2. Выборочную дисперсию стоимости.
3. Доверительный интервал для оценки средней стоимости товара с надёжностью 0,92.
4. Доверительный интервал для оценки среднего квадратического отклонения от средней стоимости с надёжностью 0,95.
5. Доверительный интервал для оценки доли товара высшего качества с надёжностью 0,98.

**2.3.3.** В университете работает 625 преподавателей в возрасте от 25 до 65 лет. Осуществлена бесповторная выборка в количестве 30 человек, которая показала, что в возрасте от 25 до 35 лет – 5 человек; от 35 до 45 лет – 8 человек; от 45 до 55 лет – 7 человек и от 55 до 65 лет – 10 человек.

Найти:

1. Выборочный средний возраст преподавателей.
2. Выборочную дисперсию этого возраста.
3. Доверительный интервал для оценки среднего возраста с надёжностью 0,96.
4. Доверительный интервал для оценки среднего квадратического отклонения с надёжностью 0,9.
5. Доверительный интервал для оценки доли преподавателей в возрасте до 45 лет с надёжностью 0,94.

**2.3.4.** В ящике находится 1600 монет достоинством 1, 2, 5 и 10 р. Осуществлена повторная выборка 40 монет; при этом оказалось, что монета достоинством 1 р. была извлечена 8 раз; 2 р. – 4 раза; 5 р. – 12 раз; 10 р. – 16 раз.

Найти:

1. Выборочное среднее номинала монет.
2. Выборочную дисперсию отклонения номиналов монет от среднего.
3. Доверительный интервал для оценки среднего номинала монеты с надёжностью 0,95.
4. Доверительный интервал для оценки среднего квадратического отклонения номинала с надёжностью 0,98.
5. Доверительный интервал для оценки доли монет номиналом 1 рубль с надёжностью 0,86.

**2.3.5.** В саду имеется 900 яблонь. Для оценки предполагаемого урожая выбрали 40 яблонь (выборка бесповторная). Оказалось, что урожайность составила от 100 до 260 кг для различных яблонь и подчиняется следующему распределению:

| Вес урожая (кг) | [100, 140) | [140, 180) | [180, 220) | [220, 260] |
|---|---|---|---|---|
| Количество яблонь | 20 | 6 | 8 | 6 |

Найти:

1. Выборочную среднюю урожайность одной яблони и доверительный интервал для оценки генеральной средней ($a$) с надёжностью 0,98.
2. Выборочное среднее квадратическое отклонение урожайности относительно среднего и доверительный интервал для оценки генерального среднего квадратического отклонения ($\sigma$) с надёжностью 0,95.

# 3. СТАТИСТИЧЕСКИЕ ГИПОТЕЗЫ

**I. Учебные цели.** Познакомить студентов с методикой проверки статистических гипотез.

В результате изучения материала студенты должны знать определение и виды статистических гипотез: основная, альтернативная, простая, сложная; уметь строить критерии проверки гипотез и осуществлять сами проверки.

**II. Формирование компетенций.** Формирование математической культуры, совершенствование общей культуры мышления, развитие аналитического и логического мышления, формирование способности обрабатывать и интерпретировать информацию, необходимую для разрешения научных, технических и социальных проблем.

**III. Введение в тему.** Проверка статистических гипотез тесно связана с теорией оценивания параметров. В естествознании, технике, экономике часто для выяснения того или иного случайного факта прибегают к высказыванию гипотез, которые можно проверить статистически, т.е. опираясь на результаты наблюдений в случайной выборке.

Проверка статистических гипотез используется, например, всякий раз, когда необходим обоснованный вывод о преимуществах того или иного способа инвестиций, измерений, технологического процесса, об эффективности нового метода обучения, управления, о пользе вносимого удобрения, лекарства, о доходности ценных бумаг, о значимости математической модели и т.д.

При изучении соответствующего материала обратите внимание на понимание следующих вопросов:

1. Ошибки при проверке гипотез.
2. Что такое критерий проверки гипотез?
3. Процедура построения критерия.
4. Схема проверки статистической гипотезы.
5. Что такое критерий согласия и как он строится?
6. Какие статистики используются при проверке гипотез о значениях числовых характеристик?
7. Методика проверки гипотез о равенстве числовых характеристик результатов различных экспериментов.

## 3.1. ОСНОВНЫЕ ПОНЯТИЯ

Статистической гипотезой называется любое предположение о свойствах распределения вероятностей, лежащего в основе наблюдаемых явлений.

Например:

1. Гипотезы о виде закона распределения исследуемой случайной величины (в виде функции распределения или плотности распределения);

– показательной $f(x) = 1 - e^{-\alpha x}$, $x \geq 0$ $\alpha > 0$;

– нормальной $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-a)^2}{2\sigma^2}}$ и др.

2. Гипотезы о численных значениях параметров распределения: $\alpha = \alpha_0$ (показательное распределение) или $\sigma = \sigma_0$, $a = a_0$ (нормальное распределение) и др.

Гипотезы обозначаются большими латинскими буквами $H_0$, $H_1$, …, $H_к$. Гипотеза $H_0$ называется основной в том смысле, что необходимо убедиться в справедливости именно её (например, успех в какой-либо деятельности, выздоровление больного, благополучная посадка самолёта и т.п.). Основная гипотеза одна.

Гипотезы $H_1$, $H_2$, …, $H_к$ противопоставлены $H_0$ и называются альтернативными.

Далее будем рассуждать только об одной альтернативе – $H_1$.

Принятие гипотезы $H_0$ или её альтернативы основано на исследовании выборочных данных $\{x_1, x_2, \ldots, x_n\}$ из некоторой генеральной совокупности.

Гипотеза $H$ – простая, если она полностью определяет теоретическое распределение случайной величины по имеющейся выборке её значений.

В противном случае гипотеза называется сложной.

Обычно $H_0$ – простая, а $H_1$ – простая или сложная.

Пример: выборка $\{x_1, x_2, \ldots, x_n\}$ из $N(a, \sigma)$; $\sigma$ – известна.

Надо оценить параметр $a$.

Тогда гипотезы $H_0$: $a = a_1$ и $H_1$: $a = a_2$ – простые.

Если же $H_0$: $a = a_0$, $H_1$: $a \neq a_0$, то $H_0$ – простая, $H_1$ – сложная.

Правило $K$, по которому гипотеза принимается или отвергается, называется критерием.

Гипотезу проверяют на основании выборки, полученной из генеральной совокупности. Из-за случайности выборки в результате проверки могут возникать ошибки и приниматься неправильные решения. В принципе, различают ошибки первого и второго рода. Ошибка первого рода имеет место тогда, когда отвергается правильная гипотеза $H_0$. При ошибке второго рода принимается неправильная гипотеза $H_0$.

Решение принимается по значению некоторой функции выборки, называемой статистикой или статистической характеристикой ($T_n$). Множество значений этой статистики можно разделить на два непересекающихся подмножества:

– подмножества значений статистики, при которых гипотеза $H_0$ принимается (не отвергается), называется областью принятия гипотезы (допустимой областью);

– подмножества значений статистики, при которых гипотеза $H_0$ отвергается (отклоняется) и принимается гипотеза $H_1$, называется критической областью.

При проверке гипотез разумно уменьшить вероятности принятия неправильных решений. Допустимая вероятность ошибки первого рода обозначается через $\alpha$ и называется уровнем значимости. Значение $\alpha$ обычно мало. Но уменьшение вероятности ошибки первого рода обычно вызывает увеличение вероятности ошибки второго рода ($\beta$).

Одновременно имеют место и такие понятия, как: $\gamma = 1 - \alpha$ – уровень доверия и $1 - \beta$ – мощность критерия (вероятность отвергнуть неверную гипотезу $H_0$).

Для определения критической области статистики используется уровень значимости $\alpha$ и учитывается вид альтернативной гипотезы $H_1$. Например, основная гипотеза $H_0$ о значении неизвестного параметра $\theta$ распределения выглядит так: $H_0$: $\theta = \theta_0$.

Альтернативная гипотеза может иметь следующий вид:

$H_1$: $\theta < \theta_0$ или $H_1$: $\theta > \theta_0$, или $H_1$: $\theta \neq \theta_0$.

Первому случаю соответствует левосторонняя критическая область, задаваемая условием:

$$P\left\{\overline{\theta}_n \le \theta_{\text{лев.кр}}\right\} = \alpha;$$

второму – правосторонняя:

$$P\{\overline{\theta}_n \ge \theta_{\text{пр.кр}}\} = \alpha;$$

третьему – двухсторонняя:

$$P\left\{\overline{\theta}_n \le \theta_{\text{лев.кр}}\right\} = P\left\{\overline{\theta}_n \ge \theta_{\text{пр.кр}}\right\} = \alpha/2.$$

Граничные точки $\theta_{\text{лев.кр}}$, $\theta_{\text{пр.кр}}$ критических областей определяют по таблицам распределения статистики.

Проверка статистической гипотезы состоит из следующих этапов:

1) определение гипотез $H_0$ и $H_1$;

2) выбор статистики и задание уровня значимости $\alpha$;

3) определение по таблицам, по уровню значимости $\alpha$ и по альтернативной гипотезе $H_1$ критической области;

4) вычисление по выборке значения статистики;

5) сравнение значения статистики с критической областью;

6) принятие решения: если значение статистики не входит в критическую область, то принимается гипотеза $H_0$ и отвергается гипотеза $H_1$, а если входит в критическую область, то отвергается гипотеза $H_0$ и принимается гипотеза $H_1$;

7) Результаты проверки статистической гипотезы нужно интерпретировать так: если приняли гипотезу $H_1$, то можно считать её доказанной, а если приняли гипотезу $H_0$, то признали, что гипотеза $H_0$ не противоречит результатам наблюдений. Но этим свойством могут (наряду с $H_0$) обладать другие гипотезы, поэтому в этом случае есть смысл проводить ещё дополнительные исследования.

**Основные гипотезы:**

– о виде распределения;

– о равенстве значений числовых характеристик распределения, значении доли признака в генеральной совокупности определённым числам;

– о равенстве числовых характеристик (долей признака в ГС) у распределений одного типа различных случайных величин.

## 3.2. ГИПОТЕЗА О ВИДЕ РАСПРЕДЕЛЕНИЯ

Одна из основных задач математической статистики – установление истинного закона распределения случайной величины.

На практике о нём судят по графику статистического распределения выборки, поэтому параметры закона – выборочные.

Однако, как бы мы ни выбирали вид закона распределения и его параметры, полной уверенности в том, что мы получим истинный закон распределения, к которому принадлежит имеющаяся у нас выборка, не существует. Поэтому вопрос может идти лишь о том, что на определённом уровне доверия выбранный нами закон согласуется с данными выборки.

Критерии, устанавливающие закон распределения, называются **критериями согласия –** критериями проверки гипотезы о предполагаемом законе неизвестного распределения.

Сделаем выборку из генеральной совокупности и по форме полигона частот или гистограммы составим гипотезу о её конкретном распределении, выраженном через функцию распределения $F(x)$ или плотность $f(x)$. Это распределение называется теоретическим.

По выборке можно найти эмпирическую функцию распределения $F^*(x)$. Гипотезу $H_0$ о распределении генеральной совокупности принимаем тогда, когда эмпирическое распределение хорошо согласуется с теоретическим. Для проверки этой гипотезы используют $\chi^2$ – критерий согласия Пирсона*.

Для его реализации вся область генеральной совокупности $X[a_1, a_{k+1}]$ делится на $k$ интервалов (можно различной длины). По выборке $\{x_1, x_2, \ldots, x_n\}$ строим интервальный ряд $\{\Delta_i, n_i\}$, $i = 1, 2, \ldots, k$, где $n_i$ – число элементов

* Как и любой критерий, критерий Пирсона не доказывает справедливость гипотезы, а лишь устанавливает на принятом уровне значимости её согласие или несогласие с данными наблюдений.

выборки, попавших в интервал $\Delta_i = [a_i, a_{i+1})$ (эмпирические частоты). Если в некотором интервале частота слишком мала (меньше 5), то этот интервал объединяют с соседним интервалом. (При дискретной интервальной совокупности интервал может содержать только одно значение генеральной совокупности).

По выборке вычисляем оценки параметров теоретического распределения (для нормального – выборочное среднее и среднее квадратическое отклонение). Таким образом, теоретическое распределение будет полностью определено, поэтому можно вычислить вероятность $p_i$ того, что случайная величина $X$ принимает значение из $i$-го интервала, при этом $\sum_{i=1}^{k} p_i = 1$. По формуле $m_i = p_i n$ находим теоретические частоты.

Основная гипотеза $H_0$ состоит в том, что функцией распределения случайной величины $X$ является выбранная теоретическая функция распределения $F(x)$. При таком предположении теоретические частоты $m_i$ и эмпирические частоты $n_i$ мало отличаются друг от друга.

Составим статистику

$$t = \sum_{i=1}^{k} \frac{(n_i - m_i)^2}{m_i}. \qquad (3.1)$$

Случайная величина $t$ имеет $\chi^2$ – распределение с числом степеней свободы $s = k - r - 1$, где $k$ – количество интервалов, $r$ – количество параметров теоретического распределения, оценки которых вычислялись по выборке.

Чем больше $t$, тем хуже согласованы распределения. При достаточно большом $t$ гипотезу $H_0$ надо отвергать, поэтому используется только правосторонняя критическая область.

Для заданного уровня доверия $\gamma$ по таблицам распределения $\chi^2_s$ [1, с. 558] находим критическое значение

$$\chi^2_{s,\text{кр}} : p\left(\chi^2_s < \chi^2_{s,\text{кр}}\right) = \gamma.$$

Гипотеза $H_0$ о согласии экспериментальных данных с распределением $F(x)$ принимается, если $t < \chi^2_{s,\text{кр}}$.

*Пример* 3.1. Воспользуемся данными примера 1.2, где дано статистическое распределение выборки – случайной величины $X$ – урожайности зерновых культур. Весь диапазон изменения этой величины [18, 48] (ц/га) разбит на шесть интервалов и указаны эмпирические частоты признака на каждом интервале. Полигон частот (рис. 1.2), а также гистограмма (рис. 1.3) наталкивают на мысль о том, что распределение «похоже» на нормальное

и характеризуется двумя параметрами: $a = M(X)$ – математическое ожидание и $\sigma = \sqrt{D(X)}$ – среднее квадратическое отклонение. Их оценками (по результатам выборки) являются $\bar{x}$ – выборочное среднее и $\bar{S}$ – выборочное среднее квадратическое отклонение. По данным примера 1.4. $\bar{x} = 32{,}6$; $\bar{S} = 7{,}34$. Таким образом, плотность распределения вероятностей $f(x) = \frac{1}{\bar{S}\sqrt{2\pi}} e^{\frac{-(x-\bar{x})^2}{2\bar{S}^2}}$ определена для всех $x$, что позволяет рассчитать теоретические частоты $m_i$ для каждого из шести выделенных интервалов по формуле

$$p(a_i \le X \le a_{i+1}) = \int_{a_i}^{a_{i+1}} f(x)dx, \; i = \overline{1,6}.$$

Чтобы найти этот «неберущийся интеграл» делаем замену переменной $Z = (x - \bar{x})/\bar{S}$.

Если $X$[18; 48], то $Z$[–2; 2,1], а интервалы разбиения представляются следующей последовательностью: [–2; –1,31), [–1,31; –0,63), [–0,63; 0,052), [0,052; 0,732), [0,732; 1,41), [1,41; 2,1].

Для расчёта вероятностей $p_i$ попадания случайной величины $Z$ в каждый из этих интервалов, назовём их $(b_i, b_{i+1})$, $i = \overline{1,6}$, используем функцию Лапласа в соответствии со свойством нормального распределения:

$$p_i(b_i \le Z \le b_{i+1}) = \frac{1}{2}\left[\Phi(b_{i+1}) - \Phi(b_i)\right].$$

После такого рода расчётов теоретических частот $m_i = n \cdot p_i$ получим следующую таблицу:

| Варианты | [18, 23) | [23, 28) | [28, 33) | [33, 38) | [38, 43) | [43, 48] |
|---|---|---|---|---|---|---|
| Эмпирические частоты $n_i$ | 7 | 11 | 16 | 14 | 10 | 6 |
| Теоретические частоты $m_i$ | 4, 63 | 10,87 | 16,51 | 15,44 | 10,03 | 3,90 |

По формуле (3.1) находим

$$t = \sum_{i=1}^{6} \frac{(n_i - m_i)^2}{m_i} \cong 2{,}49.$$

Случайная величина $t$ имеет $\chi^2$ – распределение с числом степеней свободы s = 6 – 2 – 1 = 3.

Пусть на уровне значимости $\alpha = 0{,}05$ требуется проверить нулевую гипотезу $H_0$: эмпирическое распределение соответствует нормальному закону распределения.

Находим, что соответствующее критическое значение $\chi^2_{0,05;3} = 7{,}82$ (по таблицам $\chi^2$, [1, с. 558]).

Так как $t < \chi^2_{0,05;3}$, то гипотеза о выбранном теоретическом нормальном законе распределения $N(32{,}6;\ 7{,}34)$ согласуется с опытными данными.

## 3.3. ГИПОТЕЗЫ О ЗНАЧЕНИЯХ ЧИСЛОВЫХ ХАРАКТЕРСТИК

Гипотезы о равенстве математического ожидания $a$ и дисперсии $\sigma^2$ определённым числам $a_0$ и $\sigma_0^2$ являются простыми гипотезами.

**3.3.1.** Пусть случайная величина $X$ имеет нормальное распределение с параметрами $a$ и $\sigma$ и имеется выборка $\{x_1, x_2, \ldots, x_n\}$ её значений. Проверим на уровне доверия $\gamma$ гипотезу $H_0$: $a = a_0$, где $a_0$ – некоторое число при условии, что дисперсия $\sigma^2$ известна, $H_1$: $a \neq a_0$.

Средняя выборочная $\bar{x} = \dfrac{1}{n}\sum_{i=1}^{n} x_i$ имеет нормальное распределение с параметрами $a$ и $\sigma/\sqrt{n}$, поэтому статистика

$$t = \frac{(\bar{x} - a)\sqrt{n}}{\sigma} \tag{3.2}$$

будет иметь стандартное нормальное распределение, если $H_0$: $a = a_0$ верна.

При заданном уровне доверия $\gamma$ по таблицам функции $\Phi(t)$ находим такое $t_{кр}$, чтобы $P\{|\xi_0| < t_{кр}\} = \gamma$, где $\xi_0$ – нормально распределённая случайная величина.

Гипотеза $H_0$ принимается, если $|t| < t_{кр}$.

Если $\sigma^2$ неизвестна, то в качестве статистики берут величину

$$t = \frac{(\bar{x} - a)\sqrt{n}}{\bar{S}}, \tag{3.3}$$

где $\bar{S}^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$ – исправленная выборочная дисперсия, а сама статистика $t$ имеет распределение Стьюдента с $n - 1$ степенями свободы: $t_{n-1}$. Для заданного уровня доверия $\gamma$ по таблицам распределения Стьюдента (например, [1, с. 557] определяется критическое значение $t_{кр}$ из условия $P\{|t_{n-1}| < t_{кр}\} = \gamma$, и гипотеза $H_0$ принимается, если $|t| < t_{кр}$.

*Пример* 3.2. Пусть $X$ – урожайность зерновой культуры имеет нормальное распределение с параметрами $a_0 = 35$ и неизвестным $\sigma^2$. Требуется на уровне значимости $\alpha = 0{,}05$ проверить нулевую гипотезу $H_0$: $a = a_0$ при $H_1$: $a \neq a_0$, если известно, что на основе выборки объёмом $n = 64$ определены оценки: $\bar{x} = 32{,}6$, и $\bar{S} = 7{,}34$.

По формуле (3.3) статистика $t = \dfrac{(32{,}6 - 35)\sqrt{64}}{7{,}34} = -2{,}62$.

Эта статистика имеет распределение Стьюдента с 63-мя степенями свободы и при $\gamma = 1 - \alpha = 0{,}95$, $t_{кр} = 2{,}0$.

Так как $|t| > t_{кр}$, то гипотеза $H_0$ отвергается: полученная по результатам выборки $\bar{x}$ на статистическом уровне не равна 35.

Нетрудно проверить, что уже при $a_0 = 34{,}4$ гипотеза $H_0$ не отвергается.

**3.3.2.** Пусть генеральная совокупность распределена нормально, генеральная дисперсия (хотя и неизвестно точно) предположительно равна $\sigma_0^2$.

Осуществляется выборка объёма $n$ и подсчитывается исправленная выборочная дисперсия $\bar{S}^2$.

Выдвигается гипотеза $H_0$: $\bar{S}^2 = \sigma_0^2$ при $H_1$: $\bar{S}^2 \neq \sigma_0$ и требуется проверить её выполнение на заданном уровне доверия $\gamma$.

Критерием проверки гипотезы $H_0$ является статистика

$$\chi_n^2 = \frac{(n-1)\bar{S}^2}{\sigma_0^2}, \qquad (3.4)$$

которая, если $H_0$ верна, имеет распределение $\chi^2$ (хи – квадрат) с числом степеней свободы $s = n - 1$.

Критическая область – двусторонняя. Левая и правая критические точки $\chi^2_{лев.кр}$ и $\chi^2_{пр.кр}$ находятся из условий:

$$P[\chi^2 < \chi^2_{лев.кр}] = \alpha/2 \quad \text{и} \quad P[\chi^2 > \chi^2_{пр.кр}] = \alpha/2,$$

где $\alpha = 1 - \gamma$.

Если $\chi^2_{лев.кр} < \chi^2_n < \chi^2_{пр.кр}$, то нет оснований отвергать нулевую гипотезу.

## 3.4. ГИПОТЕЗЫ О РАВЕНСТВЕ ЧИСЛОВЫХ ХАРАКТЕРИСТИК

### 3.4.1. Гипотеза о равенстве средних значений

На практике часто встречаются ситуации, когда среднее значение данных одного эксперимента отличается от среднего значения данных другого (проводимого при тех же условиях) эксперимента. Тогда возника-

ет вопрос, можно ли считать это расхождение незначимым, т.е. чисто случайным, или оно вызвано существенным различием двух генеральных совокупностей.

Пусть $X$ имеет нормальное распределение с параметрами $a_1$ и $\sigma_1$, $Y$ – такое же распределение с параметрами $a_2$ и $\sigma_2$ и дисперсии $\sigma_1^2$ и $\sigma_2^2$ известны. Имеются выборки $\{x_1, x_2, \ldots, x_{n_1}\}$ и $\{y_1, y_2, \ldots, y_{n_2}\}$ из генеральных совокупностей $X$ и $Y$.

Рассматриваем гипотезу $H_0$: $a_1 = a_2$ при $H_1$: $a_1 \neq a_2$.

Если выборки независимые (например, при исследовании двух различных групп испытуемых: контрольной и экспериментальной) и при $n_1$, $n_2 \geq 30$ выполняется гипотеза $H_0$, то статистика*

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \tag{3.5}$$

будет иметь стандартное нормальное распределение с нулевым математическим ожиданием и единичной дисперсией. Так как критическая область двусторонняя (и симметричная), в числителе (3.5) можно рассматривать $|\bar{x} - \bar{y}|$.

Для заданного уровня доверия $\gamma$ по таблицам интеграла Лапласа и уравнения $\Phi(t) = \gamma$ находим $t_{кр}$ и гипотеза $H_0$ принимается, если после вычисления значения $t$ удовлетворяет неравенству $|t| < t_{кр}$.

В случае зависимых выборок (например, результаты одной и той же группы испытуемых до и после воздействия независимой переменной) для определения достоверности разницы средних применяется формула

$$t = \frac{\sum d}{\sqrt{\frac{n \cdot \sum d^2 - \left(\sum d\right)^2}{n-1}}}, \tag{3.6}$$

где $d$ – разность между результатами в каждой паре; $n$ – число пар данных.

Число степеней свободы в случае зависимых выборок для определения $t_{кр}$: $s = n - 1$.

Если $|t| \geq t_{кр}$, то принимаем альтернативную гипотезу, т.е. считаем разницу средних достоверной. Если $|t| < t_{кр}$, то разница средних недостоверна.

---

* Здесь и далее мы рассматриваем только те статистики, которые используются при решении задач, сформулированных в данном пособии.

*Задача* 3.3. В результате двух серий измерений с количеством измерений $n_1 = 36$ и $n_2 = 48$ получены следующие средние значения исследуемых величин: $\bar{x} = 9{,}79$ и $\bar{y} = 9{,}60$.

Можно ли с надёжностью $\gamma = 0{,}99$ объяснить эти расхождения случайными причинами, если известно, что $\sigma_1 = \sigma_2 = 0{,}30$?

*Решение.* Вычислим нормированную разность

$$|t| = \frac{|\bar{x} - \bar{y}|}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} = \frac{|9{,}79 - 9{,}60|}{0{,}30\sqrt{1/36 + 1/48}} = \frac{0{,}19}{0{,}30\sqrt{7/144}} = \frac{0{,}19}{0{,}066} = 2{,}88.$$

По таблицам интеграла находим

$$\Phi(t) = 0{,}99 \Rightarrow t = 2{,}58 : t_{кр} = 2{,}58.$$

Имеем: 2,88 > 2,58, поэтому с надёжностью 0,99 можно считать, что расхождение средних неслучайно.

### 3.4.2. Гипотеза о равенстве дисперсий

Гипотезы о дисперсиях возникают довольно часто, поскольку дисперсии характеризуют такие важные показатели, как точность приборов, степень однородности признаков, риск, связанный с отклонением доходности от заданного уровня, разброс уровня успеваемости обучающихся относительно среднего, и т.д.

Пусть имеются выборки $\{x_1, x_2, \ldots, x_{n_1}\}$ и $\{y_1, y_2, \ldots, y_{n_1}\}$ из двух нормальных генеральных совокупностей $N(a_1, \sigma_1)$ *и* $N(a_2, \sigma_2)$. Рассмотрим гипотезу $H_0 : \sigma_1^2 = \sigma_2^2$ и альтернативную $H_1 : \sigma_1^2 \neq \sigma_2^2$.

Оценкой для $\sigma_1^2$ является исправленная дисперсия $\bar{S}_1^2$, для $\sigma_2^2$ – $\bar{S}_2^2$.

Статистика
$$t = \bar{S}_1^2 / \bar{S}_2^2 \qquad (3.7)$$
имеет распределение Фишера $F_{n_1-1,\ n_2-1}$.

По таблицам распределения Фишера [1, с. 559] находим такие $U$ и $V$, чтобы при заданном уровне доверия $\gamma = 1 - \alpha$

$$P\{F_{n_1-1,\ n_2-1} \le U\} = P\{F_{n_1-1,\ n_2-1} \ge V\} = \frac{\alpha}{2}.$$

Гипотеза $H_0$ принимается, если $U < t < V$.

*Пример* 3.4. В одном хозяйстве выборочная дисперсия урожайности зерновой культуры на 10 полях составила $\bar{S}_1^2 = 27$. В другом хозяйстве соответствующая выборочная дисперсия на 15 полях составила $\bar{S}_2^2 = 37{,}5$.

Проверить на уровне значимости $\alpha = 0{,}05$ гипотезу о том, что на статистическом уровне эти дисперсии не отличаются, т.е. генеральная дисперсия может быть оценена любой из них.

*Решение*. Имеем $H_0$: $\sigma_1^2 = \sigma_2^2$, $H_1$: $\sigma_1^2 \neq \sigma_1^2$.

Статистика $t = \dfrac{\overline{S}_2^2}{S_1^2} = \dfrac{37{,}5}{27} = 1{,}39$ имеет распределение Фишера с $s_1 = 9$ и $s_2 = 14$ степенями свободы.

По таблицам Фишера [1, с. 559] $F_{(0{,}05;\ 9;\ 14)} = 1{,}63$, $t = 1{,}39 < 1{,}63$, следовательно, предположение о равенстве дисперсий $\sigma_1^2 = \sigma_2^2$ не противоречит наблюдениям, т.е. нельзя считать, что в соседних районах значимые разбросы по урожайности.

## 3.5. ЗАДАЧИ ДЛЯ САМОСТОЯТЕЛЬНОГО РЕШЕНИЯ

**3.5.1.** По двум независимым выборкам, объёмы которых $n_1 = 16$ и $n_2 = 25$, извлечённым из нормальных генеральных совокупностей $X$ и $Y$, найдены исправленные выборочные дисперсии $\overline{S}_x^2 = 32$ и $\overline{S}_y^2 = 15$. При уровне значимости 0,05 проверить нулевую гипотезу $H_0 : D(X) = D(Y)$ о равенстве генеральных дисперсий при конкурирующей гипотезе $H_1 : D(X) > D(Y)$.

**3.5.2.** Из нормальной генеральной совокупности извлечена выборка $n = 26$ и по ней найдена исправленная выборочная дисперсия $\overline{S}^2 = 18{,}1$. Требуется при уровне значимости 0,01 проверить нулевую гипотезу $H_0 : \sigma^2 = \sigma_0^2 = 16$, приняв в качестве конкурирующей гипотезы $H_1 : \sigma^2 \neq 16$.

**3.5.3.** По двум независимым выборкам, объёмы которых $n = 35$ и $m = 42$, извлечённым из нормальных генеральных совокупностей, найдены выборочные средние $\overline{x} = 100$ и $\overline{y} = 150$. Генеральные дисперсии известны: $D(X) = 70$, $D(Y) = 90$. Требуется при уровне значимости 0,05 проверить нулевую гипотезу $H_0 : M(X) = M(Y)$ при конкурирующей гипотезе $H_1 : M(X) < M(Y)$.

**3.5.4.** Из нормальной генеральной совокупности с известным средним квадратическим отклонением $\sigma = 4{,}4$ извлечена выборка $n = 81$ и по ней найдена выборочная средняя $\overline{x} = 24{,}1$. Требуется при уровне значимости 0,02 проверить нулевую гипотезу $H_0 : a = a_0 = 22$, приняв в качестве конкурирующей гипотезы $H_1 : a \neq 22$.

**3.5.5.** Из нормальных генеральных совокупностей $X$ и $Y$ сделаны выборки $n_1 = n_2 = 11$ и найдены исправленные выборочные дисперсии $\overline{S}_x^2 = 1{,}0$ и $\overline{S}_y^2 = 2{,}7$.

1. При уровне значимости $\alpha = 0{,}1$ проверить нулевую гипотезу $H_0 : D(X) = D(Y)$ при конкурирующей $H_1 : D(X) \neq D(Y)$.

2. При уровне значимости $\alpha = 0{,}01$ проверить нулевую гипотезу $H_0 : \sigma_y^2 = \sigma_0^2 = 3$ при конкурирующей $H_1 : \sigma_y^2 > 3$.

**3.5.6.** Для проверки эффективности новых технологий отобраны две группы рабочих: $n_1 = 50$ человек; $n_2 = 70$ человек. В первой группе – новые технологии и $\overline{x} = 85$ деталей, во второй – $\overline{y} = 78$ деталей.

Известно, что $\sigma_x^2 = 100$, $\sigma_y^2 = 74$.

На уровне значимости $\alpha = 0{,}05$ выяснить влияние новой технологии на среднего производителя.

**3.5.7.** Физическая подготовка 9 спортсменов была проверена при поступлении в спортивную школу, а затем после недели тренировок. Итоги проверки в баллах оказались следующими (в первой строке указано число баллов, полученных каждым спортсменом при поступлении в школу; во второй строке – после обучения):

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| $x_i$ | 76 | 71 | 57 | 49 | 70 | 69 | 26 | 65 | 59 |
| $y_i$ | 81 | 85 | 52 | 52 | 70 | 63 | 33 | 83 | 62 |

На уровне значимости $\alpha = 0{,}05$ установить, значимо или незначимо улучшилась физическая подготовка спортсменов, в предположении, что число баллов распределено нормально.

# 4. ДИСПЕРСИОННЫЙ АНАЛИЗ

**I. Учебные цели.** Изучить основные понятия дисперсионного анализа и методику его применения.

В результате изучения материала студенты должны понимать основную идею дисперсионного анализа, знать область его применения, уметь строить факторные комплексы различной размерности, оценивать степень действия отдельных факторов.

**II. Формирование компетенций.** Формирование математической культуры, развитие способностей использовать математические знания в профессиональной деятельности, способностей анализировать результаты исследований, аргументированно и ясно строить рассуждения.

**III. Введение в тему.** Дисперсионный анализ – это статистический метод анализа результатов наблюдений, зависящих от различных, одновременно действующих факторов, выбор наиболее важных факторов и оценка их влияния.

Дисперсионный анализ находит применение в различных областях науки и техники. В данном пособии рассматриваются некоторые простейшие методы дисперсионного анализа. Идея дисперсионного анализа заключается в разложении общей дисперсии случайной величины на независимые случайные слагаемые дисперсии, каждое из которых характеризует влияние того или иного фактора или их взаимодействия.

Последующее сравнение этих дисперсий позволяет оценить существенность влияния фактора на исследуемую величину.

Пусть, например, в результате измерения величины $M$ получено значение $X$ и пусть на процесс измерения (на получение результата) влияют случайные независимые факторы $A$ и $B$. Тогда отклонение $M - X = \alpha + \beta + \gamma$, где $\alpha$ – отклонение под влиянием фактора $A$, $\beta$ – под влиянием фактора $B$, $\gamma$ – под влиянием остальных, неучтённых факторов, причём $\alpha$, $\beta$ и $\gamma$ независимы.

*Пример* 4.1. Пусть $M$ – урожайность зерновой культуры (постоянная величина), $X$ – полученное значение этой урожайности (случайная величина), $A$ – фактор влияния личности механизатора, обрабатывающего данное посевное поле, $B$ – фактор влияния качества уборочной техники.

Найдём дисперсию

$$D(M - X) = D(\alpha + \beta + \gamma).$$

По свойствам дисперсии:

$$D(M - X) = D(X) = D(\alpha) + D(\beta) + D(\gamma),$$

где $D(\alpha)$ – характеризует влияние фактора $A$; $D(\beta)$ – влияние фактора $B$; $D(\gamma)$ – влияние остальных факторов.

Дисперсия $D(\gamma)$ называется остаточной дисперсией. Для оценки влияния факторов $A$ и $B$ сравнивают соответствующие дисперсии $D(\alpha)$ и $D(\beta)$ с остаточной дисперсией $D(\gamma)$.

Если исследуется влияние одного фактора на исследуемую величину, то речь идёт об однофакторном комплексе, если изучается влияние двух факторов, то речь идёт о двухфакторном комплексе и т.д.

В данном пособии мы ограничимся знакомством с этими двумя комплексами.

При изучении материала обратите внимание на следующие вопросы для контроля качества усвоения изложенного материала:

1. В чём состоит общая идея дисперсионного анализа?
2. Какова модель однофакторного дисперсионного анализа?
3. Запишите правило разложения суммы квадратов отклонений.
4. Сформулируйте гипотезу о равенстве групповых средних.
5. Приведите пример построения однофакторного комплекса.
6. Что такое коэффициент детерминации?
7. В чём отличие схемы двухфакторного комплекса от однофакторного?

## 4.1. ОДНОФАКТОРНЫЙ АНАЛИЗ

Однофакторный дисперсионный анализ определяется как статистический метод, предназначенный для оценки влияния определённого фактора $A$ на результат эксперимента – некоторую случайную величину $X$, называемую также результативным признаком.

Рассмотрим пример 4.1 в предположении, что влияние фактора $B$ ничтожно и мы его не учитываем.

Пусть фактор $A$ имеет $m$ уровней (в хозяйстве $m$ механизаторов). Из каждого уровня (количества полей, обработанных каждым механизатором) сделаем выборку из $k$ элементов ($k$ полей). Общее количество выбранных элементов обозначим $n = m \cdot k$. Вся выборка представляет собой матрицу:

$$x = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1k} \\ x_{21} & x_{22} & \dots & x_{2k} \\ \dots & \dots & & \dots \\ x_{m1} & x_{m2} & \dots & x_{mk} \end{pmatrix} = (x_{ij}),\ i = \overline{1,\ k},\ j = \overline{1,\ m}.$$

Рассматривается задача: имеются ли существенные различия между различными уровнями фактора $A$ (различными механизаторами) по показателю качества (урожайность), т.е. проверяется влияние на урожайность одного фактора – личности механизатора.

Полагая, что выборка сделана из нормально распределённой генеральной совокупности, и задавая уровень значимости $\alpha$, нужно проверить гипотезу о равенстве средних значений на всех уровнях фактора:

$$H_0\colon\ \bar{x}_1 = \bar{x}_2 = \ldots = \bar{x}_m\,, \quad \bar{x}_i = \frac{1}{k}\sum_{j=1}^{k} x_{ij}\ .$$

При альтернативной гипотезе $H_1$: не все средние значения $\bar{x}_i$ должны быть равными.

Гипотеза $H_0$ означает, что влияние всех уровней фактора одно и то же.

Как уже отмечалось, для оценки влияния фактора сравнивают факторную и остаточную дисперсию. Поэтому исходной информацией для расчётов и принятия решений является $Q_1 = k\sum_{i=1}^{m}\left(\bar{x}_i - \bar{x}\right)^2$ сумма квадратов отклонений групповых средних $\bar{x}_i = 1/k\sum_{j=1}^{k} x_{ij}$ от общего среднего $\bar{x} = 1/m\sum_{i=1}^{m}\bar{x}_i$, характеризующая влияние фактора, и сумма квадратов внутригрупповых отклонений $Q_2 = \sum_{i=1}^{m}\sum_{j=1}^{k}\left(x_{ij} - \bar{x}_i\right)^2$, характеризующая отклонение остальных неучтённых факторов.

Основное тождество однофакторного анализа

$$Q\ = Q_1 + Q_2\,. \tag{4.1}$$

Здесь $Q = \sum_{i=1}^{m}\sum_{j=1}^{k}\left(x_{ij} - \bar{x}\right)^2$ – сумма квадратов отклонений элементов выборки относительно общего среднего арифметического.

В дальнейшем анализируются не сами $Q_1$ и $Q_2$, а так называемые средние квадраты, являющиеся несмещёнными оценками соответствующих дисперсий, которые получаются делением сумм квадратов отклонений на соответствующее число степеней свободы (равное разности общего числа наблюдений и числа связывающих их уравнений). Для межгрупповой дисперсии $Q_1$ число степеней свободы равно $m - 1$, так как наблюдается $m$ групповых средних, связанных между собой одним уравнением (общего среднего). Для внутригрупповой средней число степеней свободы равно $km - m = m(k-1)$, так как при её расчёте используется $km$ наблюдений, связанных между собой $m$ уравнениями. Если обозначить оценки как $\bar{S}_1^2$ и $\bar{S}_2^2$, то

$$\overline{S}_1^2 = Q_1/(m-1)\,;\ \overline{S}_2^2 = Q_2/(mk-m)\,. \qquad (4.2)$$

Существует доказательство, для того чтобы проверить нулевую гипотезу о равенстве групповых средних нормальных совокупностей с одинаковыми дисперсиями, достаточно проверить по критерию $F$ нулевую гипотезу о равенстве факторной и остаточной дисперсий.

В качестве статистики используют величину

$$F = \overline{S}_1^2 \big/ \overline{S}_2^2\,. \qquad (4.3)$$

Если гипотеза $H_0$ верна, то случайная величина $F$ имеет распределение Фишера со степенями свободы $s_1 = m - 1$ и $s_2 = m(k - 1)$.

При проверке гипотезы $H_0$ используют правостороннюю критическую область, определяемую уравнением

$$P(F > F_\alpha) = \alpha\,;\ F_\alpha = F_{\alpha;\, m-1;\, m(k-1)}\,.$$

Если значение статистики входит в критическую область, то гипотезу $H_0$ о равенстве средних значений на всех уровнях отвергаем, т.е. считаем влияние исследуемого фактора значимым. В противном случае принимаем гипотезу $H_0$, т.е. считаем, что значимость влияния фактора не установлена.

*Пример* 4.2. В сельскохозяйственном предприятии анализируют влияние личности механизатора на показатель урожайности зерновых культур (например, при уборке урожая). Были выбраны три механизатора (*A*, *B* и *C*), каждый из которых в равной мере участвовал в уборке четырёх культур (каждый на своём участке поля). По результатам уборки урожая были получены результаты (средние урожайности, ц/га), сведённые в следующую таблицу:

| Номер поля / Механизатор | I | II | III | IV |
|---|---|---|---|---|
| *A* | 30 | 32 | 31 | 33 |
| *B* | 33 | 33 | 34,2 | 33 |
| *C* | 28 | 31,6 | 31,2 | 28 |

На уровне значимости $\alpha$ = 0,05 выявить значимость влияния исследуемого фактора – личность механизатора.

*Решение.* Здесь три уровня фактора: $m$ = 3, в каждом по $k$ = 4 элемента, таким образом, $mk$ = 12.

• Групповые средние: $\overline{x}_A = 1/4\,(30 + 32 + 31 + 33) = 31{,}5$, аналогично $\overline{x}_B = 33{,}3;\ \overline{x}_C = 29{,}7$.

• Общая средняя: $\overline{x} = 1/3\,(\overline{x}_A + \overline{x}_B + \overline{x}_C) = 31{,}5$.

• Общая дисперсия $Q = \sum_{i=1}^{3}\sum_{j=1}^{4}(x_{ij} - \bar{x})^2 = (1{,}5^2 + 0{,}5^2 + 0{,}5^2 + 1{,}5^2) +$

$+ (1{,}5^2 + 1{,}5^2 + 2{,}7^2 + 1{,}5^2) + (3{,}5^2 + 0{,}1^2 + 0{,}3^2 + 3{,}5^2) = 5 + 14{,}04 + 25{,}6 = 44{,}64.$

• Межгрупповая дисперсия $Q_1 = 4\sum_{i=1}^{3}(\bar{x}_i - \bar{x})^2 = 4(0^2 + 1{,}8^2 + 1{,}8^2) = 25{,}92.$

• Внутригрупповая дисперсия $Q_2 = Q - Q_1 = 44{,}64 - 25{,}92 = 18{,}72.$

• Статистика $F = \dfrac{Q_1(mk-k)}{Q_2(k-1)} = \dfrac{25{,}92 \cdot 9}{18{,}72 \cdot 2} = 6{,}23.$

По таблицам распределения Фишера $F_{\text{кр}} = F_{0{,}05;\,2;\,9} = 4{,}26.$

Так как $F > F_{\text{кр}}$, то фактор (личность механизатора) значим и его влиянием нет оснований пренебрегать.

• Коэффициент детерминации $d = \dfrac{Q_1}{Q} = \dfrac{25{,}92}{44{,}64} = 0{,}58,$ следовательно, 58% общей дисперсии определяется исследуемым фактором.

## 4.2. МНОГОФАКТОРНЫЙ АНАЛИЗ

Если исследуют действие двух, трёх и более факторов, то структура дисперсионного анализа та же, что и при однофакторном анализе, усложняются лишь вычисления. Рассмотрим задачу оценки действия двух одновременно действующих факторов $A$ и $B$ в самом простейшем случае, когда для каждой пары уровней факторов имеется лишь одно наблюдение.

*Пример* 4.3. Предположим, что два преподавателя составляют тестовые задания для трёх групп студентов (по одной и той же учебной дисциплине). Требуется выяснить, значимо ли влияние личности преподавателя и состава учебной группы на результаты тестирования. Пусть фактор $A$ – влияние преподавателей, фактор $B$ – влияние состава студентов. Имеем три уровня фактора $B$: $B_1$, $B_2$, и $B_3$ и два уровня фактора $A$: $A_1$ и $A_2$. результат тестирования обозначим через $x_{ij}$* ($i$ = 1, 2; $j$ = 1, 2, 3). Результаты наблюдений запишем в виде

| $A$ \ $B$ | $B_1$ | $B_2$ | $B_3$ | $\bar{x}_{i*}$ |
|---|---|---|---|---|
| $A_1$ | $x_{11}$ | $x_{12}$ | $x_{13}$ | $x_{1*}$ |
| $A_2$ | $x_{21}$ | $x_{22}$ | $x_{23}$ | $x_{2*}$ |
| $\bar{x}_{*j}$ | $\bar{x}_{*1}$ | $\bar{x}_{*2}$ | $\bar{x}_{*3}$ | $\bar{x}$ |

* Пересечение $i$-го и $j$-го уровней образуют $ij$-ю ячейку, в которую записываются наблюдения, полученные при одновременном исследовании факторов $A$ и $B$ на $i$-м и $j$-м уровнях соответственно.

Каждому столбцу и строке соответствует среднее значение $\bar{x}_{i*}$ и $\bar{x}_{*j}$, а всей таблице общее среднее $\bar{x}$.

$$\bar{x}_{i*} = \frac{1}{3}\sum_{j=1}^{3} x_{ij}\ ; \quad \bar{x}_{*j} = \frac{1}{2}\sum_{i=1}^{2} x_{ij}\ ; \quad \bar{x} = \frac{1}{2\cdot 3}\sum_{i=1}^{2}\sum_{j=1}^{3} x_{ij}\ .$$

Основное тождество однофакторного анализа (4.1) в данном случае принимает вид

$$Q = \sum_{i=1}^{2}\sum_{j=1}^{3}(x_{ij}-\bar{x})^2 = \sum_{i=1}^{2}\sum_{j=1}^{3}(x_{ij}-\bar{x}_{i*}-\bar{x}_{*j}+\bar{x}+\bar{x}_{i*}-\bar{x}+\bar{x}_{*j}-\bar{x})^2 =$$

$$= \sum_{i=1}^{2}\sum_{j=1}^{3}(x_{ij}-\bar{x}_{i*}-\bar{x}_{*j}+\bar{x})^2 + \sum_{i=1}^{2}\sum_{j=1}^{3}(\bar{x}_{i*}-\bar{x})^2 + \sum_{i=1}^{2}\sum_{j=1}^{3}(\bar{x}_{*j}-\bar{x})^2 = Q_3 + Q_1 + Q_2.$$

Слагаемое $Q_1 = \sum_{i=1}^{2}\sum_{j=1}^{3}(\bar{x}_{i*}-\bar{x})^2 = 3\sum_{i=1}^{2}(\bar{x}_{i*}-\bar{x})^2$ представляет собой сумму квадратов разностей между средними по строкам и общим средним и характеризует изменение признака по фактору *A*.

Слагаемое $Q_2 = \sum_{i=1}^{2}\sum_{j=1}^{3}(\bar{x}_{*j}-\bar{x})^2 = 2\sum_{j=1}^{3}(\bar{x}_{*j}-\bar{x})^2$ представляет собой сумму квадратов разностей между средними по столбцам и общим средним и характеризует изменение признака по фактору *B*.

Слагаемое $Q_3$ называется остаточной суммой квадратов и характеризует влияние неучтённых факторов. Сумма $Q$ называется общей или полной суммой квадратов отклонений отдельных наблюдений от общей средней.

С учётом степеней свободы оценка дисперсий представляется в виде

$$\bar{S}^2 = \frac{1}{2\cdot 3-1}Q = Q/5;\ \ \bar{S}_1^2 = \frac{Q_1}{2-1} = Q_1;\ \ \bar{S}_2^2 = \frac{Q_2}{3-1} = Q_2/2;$$

$$\bar{S}_3^2 = \frac{Q_3}{(2-1)(3-1)} = Q_3/2.$$

В двухфакторном анализе для выяснения значимости влияния факторов *A* и *B* на исследуемый признак сравнивают оценки дисперсии по факторам с оценкой остаточной дисперсией, т.е. оценивают отношения $\bar{S}_1^2/\bar{S}_3^2$ и $\bar{S}_2^2/\bar{S}_3^2$, которые при нормальном распределении случайных величин имеют *F*-распределение (Фишера).

При выбранном уровне значимости $\alpha$ значения $F_A = \bar{S}_1^2/\bar{S}_3^2$ и $F_B = \bar{S}_2^2/\bar{S}_3^2$ сравнивают с табличными значениями $F_{\alpha(A)} = F_{\alpha,\,1,\,2}$ и $F_{\alpha(B)} = F_{\alpha,\,2,\,2}$. При $F_A < F_{\alpha(A)}$ и $F_B < F_{\alpha(B)}$ нулевая гипотеза о равенстве средних

не отвергается, т.е. влияние факторов $A$ и $B$ на исследуемый признак незначимо; при наличии противоположных неравенств, соответствующий фактор считают значимым.

Рассмотрим конкретный вид предыдущей таблицы для примера 4.3:

| $A$ \ $B$ | $B_1$ | $B_2$ | $B_3$ | |
|---|---|---|---|---|
| $A_1$ | 65 | 80 | 95 | $x_{1*} = 80$ |
| $A_2$ | 55 | 75 | 95 | $x_{2*} = 75$ |
| | $\bar{x}_{*1} = 60$ | $\bar{x}_{*2} = 77{,}5$ | $\bar{x}_{*3} = 95$ | $\bar{x} = 77{,}5$ |

На уровне значимости $\alpha = 0{,}05$ надо выяснить влияние фактора $A$ – личность преподавателя и фактора $B$ – состав студентов на результаты тестирования.

Имеем:

$Q_1 = 3[(80 - 77{,}5)^2 + (75 - 77{,}5)^2] = 37{,}5;$

$Q_2 = 2[(60 - 77{,}5)^2 + (77{,}5 - 77{,}5)^2 + (95 - 77{,}5)^2] = 1225;$

$Q = (65 - 77{,}5)^2 + (80 - 77{,}5)^2 + (95 - 77{,}5)^2 + (55 - 77{,}5)^2 +$
$+ (75 - 77{,}5)^2 + (95 - 77{,}5)^2 = 1287{,}5.$

Тогда $Q_3 = Q - Q_1 - Q_2 = 1287{,}5 - 37{,}5 - 1225 = 25.$

Находим оценки дисперсий:

$$\bar{S}_1^2 = 37{,}5/1 = 37{,}5;\ \bar{S}_2^2 = 1225/2 = 612{,}5;\ \bar{S}_3^2 = 25/2 = 12{,}5.$$

Вычисляем: $F_A = \bar{S}_1^2 / \bar{S}_3^2 = 3;\ F_B = \bar{S}_2^2 / \bar{S}_3^2 = 49.$

Для уровня значимости $\alpha = 0{,}05$ по таблицам $F$-распределения находим $F_{\alpha(A)} = F_{0{,}05;\,1;\,2} = 18{,}51$; $F_{\alpha(B)} = F_{0{,}05;\,2;\,2} = 19{,}0$.

Сравнивая табличные значения с вычисленными, имеем:

$$F_A < F_{\alpha(A)};\ F_B > F_{\alpha(B)}.$$

Полученные результаты позволяют сделать следующие выводы: нулевая гипотеза о равенстве средних по строкам подтверждается, т.е. влияние фактора $A$ – преподавателей на результаты тестирования не значимо; нулевая гипотеза о равенстве средних по столбцам не подтверждается, т.е. влияние фактора $B$ – состава учебной группы на результаты тестирования значимо.

При одном наблюдении в ячейке схема вычислений довольно проста, однако в этом случае достоверность выводов, полученных на основании проведённого анализа, недостаточна. Поэтому при решении практических

задач желательно иметь несколько наблюдений в одной ячейке. В этом случае вычисления усложняются, однако выводы получаются более достоверными. Такого рода схемы можно изучить по специальной литературе, например, [1, с. 400 – 407].

## 4.3. ЗАДАЧИ ДЛЯ САМОСТОЯТЕЛЬНОГО РЕШЕНИЯ

**4.3.1.** Анализируется вопрос значимости влияния рекламы на объём реализации рекламируемой продукции. Было выделено четыре уровня рекламирования (*А*, *В*, *С*, *Д*) и в течение трёх месяцев (помесячно) осуществлялась оценка объёма продаж (млн. р.). Результаты сведены в следующую таблицу:

| Уровень рекламы | Месяцы | | |
|---|---|---|---|
| | 1 | 2 | 3 |
| *А* | 1,4 | 1,6 | 1,5 |
| *В* | 1,8 | 1,7 | 1,3 |
| *С* | 1,55 | 1,7 | 1,85 |
| *Д* | 1,2 | 1,3 | 1,1 |

На уровне значимости $\alpha = 0{,}05$ выяснить: существенно ли влияние фактора рекламирования на объём реализации рекламируемой продукции.

**4.3.2.** В университете анализируется вопрос влияния времени начала учебных занятий на посещаемость студентов (%). Произвольным образом были выбраны 4 дня (один и тот же день недели) и выделены три уровня исследуемого фактора – время начала занятий (*А*, *В*, *С*). Количество обучающихся студентов считается статистически равным. Результаты сведены в следующую таблицу:

| Уровни исследуемого фактора | Экспериментальные дни | | | |
|---|---|---|---|---|
| | 1 | 2 | 3 | 4 |
| *А* | 80 | 78 | 82 | 80 |
| *В* | 81 | 83 | 79 | 81 |
| *С* | 77 | 79 | 75 | 73 |

На уровне значимости $\alpha = 0{,}05$ выяснить: существенно ли влияние фактора – время начала занятий на посещаемость студентами учебных занятий.

**4.3.3.** В университете анализируется влияние личности преподавателя (фактор) на результаты экзаменов (результирующий признак) по тестам, составленным этими преподавателями. Были выбраны 3 преподавателя (*А*, *В* и *С*), каждый из которых составил четыре теста (по четырём семестрам изучения одного и того же предмета, например, математики).

На основе полученных результатов была составлена таблица средних по учебной группе баллов по 100-балльной шкале.

| Семестр / Преподаватель | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| *A* | 60 | 64 | 62 | 66 |
| *B* | 67 | 66 | 69 | 62 |
| *C* | 53 | 62 | 65 | 60 |

На уровне значимости $\alpha = 0{,}05$ выявить уровень влияния исследуемого фактора.

**4.3.4.** В университете анализируется влияние на академическую успеваемость студентов – фактора *A* – соотношения часов лекционных и практических занятий в течение семестра при одинаковой общей учебной нагрузке и фактора *B* – состава учебных групп. Было выделено три уровня соотношений (*A*, *B*, *C*) и выбрано четыре группы. На основе полученных результатов была составлена таблица, где усреднённые значения академической успеваемости даны по 50-балльной шкале.

| Уровень соотношения | Группы | | | |
|---|---|---|---|---|
| | $B_1$ | $B_2$ | $B_3$ | $B_4$ |
| $A_1$ | 23 | 29 | 24 | 28 |
| $A_2$ | 32 | 38 | 32 | 42 |
| $A_3$ | 26 | 32 | 34 | 32 |

Выяснить на уровне значимости $\alpha = 0{,}05$: существенно ли влияют обозначенные факторы на академическую успеваемость?

**4.3.5.** В сельскохозяйственном предприятии анализируется влияние на величину собранного урожая (урожайность, ц/га) двух факторов: *A* – личности механизатора и *B* – качества сельскохозяйственной техники. Были выбраны четыре комбайнера (уровни фактора *A*), которые работали (попеременно) на трёх комбайнах (уровни фактора *B*) в условиях отсутствия персональной закреплённости. На основе оценки результатов уборки составлена таблица, где указаны усреднённые значения урожайности (ц/га).

| Комбайнеры | Комбайны | | |
|---|---|---|---|
| | $B_1$ | $B_2$ | $B_3$ |
| $A_1$ | 39 | 44 | 40 |
| $A_2$ | 34 | 32 | 36 |
| $A_3$ | 37 | 42 | 35 |
| $A_4$ | 34 | 44 | 39 |

На уровне значимости $\alpha = 0{,}05$ выяснить значимость влияния факторов *A* и *B* на урожайность.

# 5. ЭЛЕМЕНТЫ ТЕОРИИ КОРРЕЛЯЦИИ

**I. Учебные цели.** Изучить некоторые варианты стохастической зависимости переменных величин.

В результате изучения материала студенты должны знать определения и свойства коэффициента корреляции, понятие корреляционной зависимости, линии регрессии, уметь оценивать тесноту и характер связи случайных величин.

**II. Формирование компетенций.** Формирование математической культуры, развитие способностей использовать законы естественнонаучных дисциплин в профессиональной деятельности, изложения сути предлагаемых решений, логически верно, аргументированно и ясно строить рассуждения.

**III. Введение в тему.** В реальном мире многие явления природы происходят в обстановке действия многочисленных факторов, влияние каждого из которых ничтожно, а число их велико. В этих случаях связь теряет свою строгую функциональность и изучаемая система переходит не в определённое состояние, а в одно из возможных состояний. Поэтому речь идёт о так называемой стохастической связи, состоящей в том, что одна случайная переменная реагирует на изменение другой изменением своего закона распределения. В практике статистических исследований рассматривают частный случай стохастической связи – статистическую связь, когда условное математическое ожидание одной случайной переменной является функцией значения, принимаемого другой случайной переменной, т.е. $M_x(\overline{Y}) = f(x)$.

Знание статистической зависимости между случайными переменными имеет большое практическое значение: с её помощью можно прогнозировать значение зависимой случайной переменной в предположении, что независимая примет определённое значение. Однако такие прогнозы не могут быть безошибочными. Применяя вероятностные методы, можно вычислить вероятность того, что ошибка прогноза не выйдет за определённые границы.

### Вопросы для контроля усвоения

1. Что такое корреляционная зависимость?
2. Как найти коэффициент корреляции и что он характеризует?
3. Что такое линия регрессии?
4. Как найти уравнение линейной регрессии?
5. Что даёт статистический анализ уравнения регрессии?

## 5.1. КОРРЕЛЯЦИОННАЯ ЗАВИСИМОСТЬ И ПРЕДСТАВЛЕНИЕ ДАННЫХ В КОРРЕЛЯЦИОННОМ АНАЛИЗЕ

Пусть $X$ – независимая переменная, $Y$ – зависимая.

На практике одному значению $X$ (фактору) может соответствовать ряд значений $Y$ (определённое распределение $Y$), так как кроме фактора $X$ имеет место влияние других факторов.

Например, $x_1$ кг удобрений на $m$ различных участках поля соответствует $y_{11}, y_{12}, \ldots, y_{1m}$, кг зерна. Если найти $\bar{y}_1 = \frac{1}{m}\sum_{j=1}^{m} y_{1j}$, то $x_1$ соответствует $\bar{y}_1$. Аналогично $x_2$ соответствует $\bar{y}_2$, …, $x_n$ соответствует $\bar{y}_n$ (значениям $x_i$ соответствуют $\bar{y}_i = \sum_{j=1}^{m} y_j$, $i = 1, 2, \ldots, n$).

Корреляционной зависимостью между двумя случайными переменными $X$ и $Y$ называется функциональная зависимость между значениями одной из них и условным математическим ожиданием другой $\varphi(x) = M_x(Y)$ или $\psi(y) = M_y(X)$.

Эти уравнения носят название уравнений регрессии* («$Y$» на «$X$» и «$X$» на «$Y$»), а соответствующие графики – линии регрессии.

Если $\varphi(x) = \text{const}$, $\psi(y) = \text{const}$, то корреляционной связи нет.

Основная задача корреляционного анализа – выявление тесноты связи между переменными $X$ и $Y$ и количественная оценка этой связи.

В корреляционном анализе экспериментальные данные можно представлять в виде набора пар чисел $(x_i, y_i)$, $i = \overline{1, n}$, где $(x_1, \ldots, x_n)$ выборка значений $X$, $(y_1, y_2, \ldots, y_n)$ – выборка значений $Y$.

*Пример* 5.1. $X$ – количество дождливых дней в течение месяца: {2, 4, 6, 8, 10}, $Y$ – изменение высоты растений, м: {0,2; 0,25; 0,4; 0,45; 0,5}.

Данные для корреляционного анализа: (2; 0,2); (4; 0,25); (6; 0,4); (8; 0,45); (10; 0,5)**.

Насколько тесная связь: причины – количества дождей и следствия – изменения высоты растений – задача корреляционного анализа.

Экспериментальные данные можно задавать и таблицей (корреляционная таблица).

*Пример* 5.2. Получены статистические данные по 10 сельскохозяйственным предприятиям с целью исследования зависимости объёма основных фондов $X$ (стоимость техники и т.п.) и урожайности зерновых культур $Y$ (ц/га). В результате исследуется двумерная случайная величина $(X, Y)$, которая задана следующей таблицей.

---

* Регресс – движение назад. По изменению $Y$ мы судим о степени влияния на $Y$ переменной $X$ (или наоборот).

** Может быть и другое сочетание значений $X$ и $Y$.

| $y_i$ | $y$ | $x_i$ | | | |
|---|---|---|---|---|---|
| | | 50 | 100 | 150 | $n_j$ |
| [36, 40) | 38 | 1 | | | 1 |
| [40, 44) | 42 | | 2 | | 2 |
| [44, 48) | 46 | | 3 | 2 | 5 |
| [48, 52] | 50 | | | 2 | 2 |
| | $n_i$ | 1 | 5 | 4 | 10 |

В первой строке таблицы записаны средние значения переменной $X$ для интервалов [25, 75), [75, 125) и [125, 175], в первом столбце – интервалы изменения переменной $Y$, во втором – середины этих интервалов. Центральную часть таблицы занимают частоты $n_{ij}$ – число предприятий, соответствующих значениям переменных $X = x_i$, $Y = y_j$; $i = 1, 2, 3$; $j = 1, 2, 3, 4$. В последней строке записаны $n_i = \sum_{j=1}^{4} n_{ij}$, в последнем столбце – $n_j = \sum_{i=1}^{3} n_{ij}$. Кроме того, $\sum_{i=1}^{3} n_i = \sum_{j=1}^{4} n_j = n = 10$ (общее количество предприятий).

## 5.2. КОЭФФИЦИЕНТ КОРРЕЛЯЦИИ

Основой оценки для тесноты связи между переменными $X$ и $Y$ служит выборочный коэффициент корреляции $r$:

$$r = \frac{\overline{xy} - \bar{x}\,\bar{y}}{S_x S_y},$$

где $\bar{x}$, $\bar{y}$, $\overline{xy}$ – средние выборочные соответственно для $X$, $Y$ и $X \cdot Y$; $S_x$, $S_y$ – средние квадратические отклонения для $X$ и $Y$.

Коэффициент корреляции характеризует степень приближения зависимости между случайными величинами к линейной функциональной зависимости.

Если данные не сгруппированы, и $n$ – объём выборки, то

$$r = \frac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{j=1}^{n} y_j\right)}{\sqrt{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}\sqrt{n\sum_{j=1}^{n} y_j^2 - \left(\sum_{j=1}^{n} y_j\right)^2}}. \qquad (5.1)$$

Свойства выборочного коэффициента корреляции аналогичны свойствам коэффициента корреляции между случайными величинами $X$ и $Y$.

1. $-1 \le r \le 1$; чем ближе $|r|$ к 1, тем теснее связь.

2. Если переменные $X$ и $Y$ умножить на одно и то же число, то $r$ не изменится.

3. Если $r = \pm 1$, корреляционная связь между $X$ и $Y$ линейная.

Генеральный коэффициент корреляции $\rho$ не является случайной величиной, а выборочный коэффициент корреляции $r$ – величина случайная. Если $r \neq 0$, то возникает вопрос о значимости найденного коэффициента. Для выяснения этого вопроса на заданном уровне значимости $\alpha$ проверяется гипотеза $H_0$: $\rho = 0$ (т.е. связь между $X$ и $Y$ отсутствует) при $H_1$: $\rho \neq 0$.

При справедливости этой гипотезы статистика $t = \dfrac{r\sqrt{n-2}}{\sqrt{1-r^2}}$ имеет распределение Стьюдента с $s = n - 2$ степенями свободы. Поэтому гипотеза $H_0$ отвергается, если $|t| > t_{кр}$, где $t_{кр}$ – соответствующая критическая точка.

Если коэффициент корреляции значим, то для него можно построить доверительный интервал, который с заданной надёжностью накрывает неизвестное значение $\rho$.

В случае небольшого числа наблюдений ($n < 30$) распределение выборочного коэффициента корреляции заметно отличается от нормального. В этом случае для оценки $\rho$ используют $z$ преобразование Фишера:

$$z = \frac{1}{2}\ln\frac{1+r}{1-r}.$$

Эта статистика даже при небольших значениях $n$ имеет нормальное распределение, причём

$$M(z) = \frac{1}{2}\ln\frac{1+\rho}{1-\rho},\ \sigma_z^2 = \frac{1}{n-3}.$$

Доверительный интервал для $M(z)$:

$$z - \frac{t_{кр}}{\sqrt{n-3}} \le M(z) \le z + \frac{t_{кр}}{\sqrt{n-3}},$$

где $t_{кр}$ определяется из условия: $\Phi(t_{кр}) = \gamma = 1 - \alpha$. Так как $\rho = th(M(z))$, то

$$th\left(z - \frac{t_{кр}}{\sqrt{n-3}}\right) < \rho < th\left(z + \frac{t_{кр}}{\sqrt{n-3}}\right).$$

*Пример* 5.3. Проведено исследование 6 хозяйств для изучения зависимости урожайности зерновых культур ($Y$, ц/га) от количества внесённых минеральных удобрений на 1 га пашни ($X$, ц/га). Получены следующие данные:

| $X$ | 2,1 | 2,3 | 2,4 | 2,6 | 2,9 | 3,0 |
|---|---|---|---|---|---|---|
| $Y$ | 18,0 | 21,0 | 22,1 | 25,3 | 28 | 28,5 |

Определить тесноту связи между величиной $Y$ и величиной $X$, используя коэффициент корреляции, проверить на уровне $\alpha = 0{,}05$ его значимость.

*Решение.* Коэффициент корреляции найдём по формуле (5.1). Для её реализации составим таблицу:

| $i$ | $x_i$ | $y_i$ | $x_i y_i$ | $x_i^2$ | $y_i^2$ |
|---|---|---|---|---|---|
| 1 | 2,1 | 18,0 | 37,8 | 4,41 | 324 |
| 2 | 2,3 | 21,0 | 48,3 | 5,29 | 441 |
| 3 | 2,4 | 22,1 | 53,04 | 5,76 | 488,41 |
| 4 | 2,6 | 25,3 | 65,78 | 6,76 | 690,09 |
| 5 | 2,9 | 28 | 81,2 | 8,41 | 784 |
| 6 | 3,0 | 28,5 | 85,5 | 9,00 | 812,25 |
| $\sum_{i=1}^{n}$ | 15,3 | 142,9 | 371,62 | 39,63 | 3489,75 |

Тогда:

$$r = \frac{6 \cdot 371{,}62 - 142{,}9 \cdot 15{,}3}{\sqrt{6 \cdot 39{,}63 - (15{,}3)^2}\sqrt{6 \cdot 3489{,}75 - (142{,}9)^2}} = 0{,}991.$$

Значимость $r$ анализируем, сравнивая статистику $t = \dfrac{r\sqrt{n-2}}{\sqrt{1-r^2}} =$

$$= \frac{0{,}991 \cdot 2}{\sqrt{1-0{,}982}} = \frac{1{,}982}{0{,}134} = 14{,}79 \text{ и } t_{\text{кр}} = t_{0{,}05;\,4} = 2{,}78.$$

Так как $t > t_{\text{кр}}$, то $r$ значим.

## 5.3. СТАТИСТИЧЕСКАЯ ЗАВИСИМОСТЬ. УРАВНЕНИЕ РЕГРЕССИИ

Коэффициент корреляции – не единственная характеристика наличия статистической зависимости переменных величин. Более общие зависимости не обязательно линейные или «близкие» к ним оцениваются выборочным корреляционным моментом или выборочной ковариацией [1, с. 416].

В статистическом анализе зависимость между входными параметрами (значениями неслучайной независимой переменной $X$) и выходной переменной $Y$ рассматривается как статистическая и представляет особый интерес – установление вида зависимости $Y$ от $X_1$, $X_2$, …, $X_n$, т.е. вида уравнения регрессии. Это связано, в первую очередь, с необходимостью прогнозирования исследуемых процессов.

Установление формы зависимости, оценки функции регрессии и её параметров является задачами регрессионного анализа.

Оценкой функции регрессии $\varphi(x) = M_x(Y)$ является функция

$$y_x = \varphi(x, b_0, b_1, ..., b_n), \qquad (5.2)$$

где $x$ – значение величины $X$, $y_x = M_x(Y)$, $b_0$, $b_1$, …, $b_n$ – параметры функции регрессии.

Задача регрессионного анализа состоит в определении функции $\varphi$, её параметров и дальнейшего статистического исследования уравнения регрессии.

Ориентировочное определение вида функции $\varphi$ можно осуществить эмпирически, построив корреляционное поле.

Если функция $\varphi$ линейна по $x$, т.е. $y_x = a + bx$, то говорят, что имеет место линейная регрессия $Y$ по $X$.

Если известны значения $x_i$, $y_i$, $i = \overline{1, n}$, то используя метод наименьших квадратов, составим функцию неизвестных переменных $a$ и $b$:

$$S(a, b) = \sum_{i=1}^{n} \left(y_i - (a + bx_i)\right)^2 \to \min.$$

Из необходимого условия экстремума

$$\begin{cases} \dfrac{\partial S}{\partial a} = \sum\limits_{i=1}^{n} 2\left(y_i - (a + bx_i)\right)(-1) = 0, \\ \dfrac{\partial S}{\partial b} = \sum\limits_{i=1}^{n} 2\left(y_i - (a + bx_i)\right)(-x_i) = 0. \end{cases} \quad \text{или} \quad \begin{cases} \sum\limits_{i=1}^{n} \left(y_i - a - bx_i\right) = 0, \\ \sum\limits_{i=1}^{n} \left(x_i y_i - a x_i - b x_i^2\right) = 0. \end{cases}$$

После преобразований получим систему

$$\begin{cases} a + \bar{x}b = \bar{y}, \\ \bar{x}a + \overline{x^2}b = \overline{xy}, \end{cases}$$

где $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ ; $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$; $\overline{x^2} = \frac{1}{n}\sum_{i=1}^{n} x_i^2$; $\overline{xy} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i$ .

Решение этой системы:

$$a = \frac{\overline{x^2}\,\bar{y} - \bar{x}\,\overline{xy}}{\overline{x^2} - (\bar{x})^2} \; ; \; b = \frac{\overline{xy} - \overline{xy}}{\overline{x^2} - (\bar{x})^2} = \frac{\overline{xy} - \overline{xy}}{S_x^2} = \frac{\overline{xy} - \bar{x}\,\bar{y}}{S_x\, S_y}\,\frac{S_y}{S_x} = r\,\frac{S_y}{S_x} \quad (5.2)$$

определяет уравнение регрессии $y_x = a + bx$.

Коэффициент $b$ в уравнении регрессии называется коэффициентом регрессии $Y$ по $X$ и обозначается $b_{YX}$:

$$b_{YX} = r\,\frac{S_y}{S_x}. \quad (5.3)$$

Коэффициент регрессии $Y$ на $X$ показывает, на сколько единиц в среднем изменяется переменная $Y$ при изменении переменной $X$ на одну единицу.

Запишем (возьмём) математическое ожидание от правой и левой частей уравнения $y_x = a + bx$:

$$M(y_x) = M(a + bx) = M(a) + M(bx) = a + bM(x).$$

Тогда $\bar{y} = a + b\bar{x}$ , и уравнение регрессии можно записать в обычно принятой в математической статистике форме

$$y_x - \bar{y} = b_{YX}(x - \bar{x}), \quad (5.4)$$

где $b_{YX} = r\,\frac{S_y}{S_x}$.

Эта форма не предполагает непосредственного нахождения параметра $a$.

Таким образом, уравнение линейной регрессии можно записать, если известны выборочные средние $\bar{x}$ и $\bar{y}$ и коэффициент регрессии.

## 5.4. СТАТИСТИЧЕСКИЙ АНАЛИЗ УРАВНЕНИЯ РЕГРЕССИИ

Очевидно, что эмпирические значения $y_i$ и соответствующие (номеру $i$) расчётные (по уравнению регрессии) значения $y_{x_i}$ чаще всего будут различными. Степень их различия, выраженная, например, суммой $Q_{\text{ост}} = \sum_{i=1}^{n} (y_i - y_{x_i})^2$ характеризует факт правильности выбора регрессионной модели.

Для того чтобы установить, соответствует ли выбранная регрессионная модель экспериментальным данным используют основное уравнение дисперсионного анализа:

$$Q = Q_{ф} + Q_0,$$

где $Q_{ф}$ – сумма квадратов, обусловленная регрессией; $Q_0$ – остаточная сумма квадратов; $Q$ – общая сумма квадратов отклонений $Y$ от средней.

Для несгруппированной выборки

$$Q = \sum_{i=1}^{n}(y_i - \bar{y})^2;\ \ Q_{ф} = \sum_{i=1}^{n}\left(y_{x_i} - \bar{y}\right)^2;\ \ Q_0 = \sum_{i=1}^{n}\left(y_i - y_{x_i}\right)^2,$$

где $y_i$ – заданные (эмпирические) значения переменной $Y$, $\bar{y}$ – их среднее выборочное, $y_{x_i}$ – значения, найденные по (5.4).

Для заданного уровня $\alpha$ находим критическое значение $F_{кр}$ распределения Фишера при $s_1 = l - 1$, $s_2 = n - l$ степенях свободы, где $n$ – число наблюдений (объём выборки), $l$ – число оцениваемых параметров (при линейной функции регрессии $l = 2$).

Если статистика $t = \dfrac{Q_{ф}(n-l)}{Q_0(l-1)} > F_{кр}$, то уравнение регрессии считается значимым, т.е. соответствующим экспериментальным данным на уровне значимости $\alpha$.

Воздействие неучтённых случайных факторов в линейной модели регрессии определяется остаточной дисперсией $\sigma_0^2$. Оценкой этой дисперсии является выборочная остаточная дисперсия $S_0^2 = \dfrac{1}{n-l}Q_0$.

*Пример* 5.4. Зависимость между стоимостью эксплуатации самолёта $Y$ (млн. р.) и его возрастом $X$ (лет) выражается следующей таблицей:

| $X$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| $Y$ | 3 | 3,5 | 3,5 | 4 | 4 | 6 | 9 | 10 |

Найти линейное уравнение регрессии $Y$ по $X$, оценить его значимость на уровне $\alpha = 0{,}01$ и определить остаточную дисперсию.

*Решение.*

Уравнение регрессии $Y$ по $X$:

$$y_x - \bar{y} = r\frac{S_y}{S_x}(x - \bar{x}).$$

Объединим результаты вычислений исходных данных в следующую таблицу:

| $i$ | $x_i$ | $y_i$ | $x_i\, y_i$ | $x_i^2$ | $y_i^2$ |
|---|---|---|---|---|---|
| 1 | 1 | 3 | 3 | 1 | 9 |
| 2 | 2 | 3,5 | 7 | 4 | 12,25 |
| 3 | 3 | 3,5 | 10,5 | 9 | 12,25 |
| 4 | 4 | 4 | 16 | 16 | 16 |
| 5 | 5 | 4 | 20 | 25 | 16 |
| 6 | 6 | 6 | 36 | 36 | 36 |
| 7 | 7 | 9 | 63 | 49 | 81 |
| 8 | 8 | 10 | 80 | 64 | 100 |
| $\sum$ | 36 | 43 | 235,5 | 204 | 283 |

Имеем: $\bar{x} = \frac{36}{8} = 4{,}5$; $\bar{y} = \frac{43}{8} = 5{,}38$; $\overline{xy} = \frac{235{,}5}{8} = 29{,}4$;

$\overline{x^2} = \frac{204}{8} = 25{,}5$; $\overline{y^2} = \frac{1}{8} \cdot 283 = 35{,}4$; $S_x = 2{,}29$; $S_y = 2{,}54$.

По формуле (5.1) находим $r = 0{,}89$, тогда $r\,\frac{S_y}{S_x} = 0{,}99 \approx 1$, а уравнение регрессии:

$$y_x - 5{,}38 = x - 4{,}5 \text{ или } y_x = x + 0{,}88.$$

Для выявления значимости уравнения регрессии вычислим суммы

$$Q_{\text{ф}} = \sum_{i=1}^{8} \left(y_{x_i} - \bar{y}\right)^2,$$

где $y_{x_i} = x_i + 0{,}88$ и $Q_0 = \sum_{i=1}^{8} \left(y_i - y_{x_i}\right)^2$.

Имеем: $Q_0 = 9{,}45$, $Q_{\text{ф}} = 42$.

Для заданного уровня $\alpha = 0{,}01$ находим, что критическая точка распределения Фишера при $s_1 = 2 - 1 = 1$ и $s_2 = 8 - 2 = 6$ равна 13,74.

Статистика $t = \frac{Q_{\text{ф}}(n-l)}{Q_0(l-1)} = \frac{42 \cdot 6}{9{,}45} = 26{,}7 > F_{\text{кр}}$.

Следовательно, уравнение регрессии на уровне $\alpha = 0{,}01$ значимо.

Оценкой остаточной дисперсии (воздействия неучтённых случайных факторов в линейной модели) $\sigma_0^2$ является выборочная остаточная дисперсия

$$S_0^2 = \frac{1}{n-l} Q_0 = \frac{9{,}45}{6} = 1{,}575.$$

## 5.5. ЗАДАЧИ ДЛЯ САМОСТОЯТЕЛЬНОГО РЕШЕНИЯ

**5.5.1.** Случайные величины *X* и *Y* связаны таблицей:

| *X* | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|
| *Y* | 2 | 1,9 | 3,2 | 2,4 | 2,3 |

Определить тесноту связи *X* и *Y*, используя коэффициент корреляции и проверить на уровне $\alpha = 0{,}05$ его значимость.

**5.5.2.** В медицинском учреждении исследовали зависимость между потребляемым количеством сигарет (*X*, пачки/неделю) и вероятностью лёгочных заболеваний (*Y*, %) после 10 лет курения. Экспериментальные данные представлены таблицей:

| *X* | 8 | 12 | 16 | 20 | 24 |
|---|---|---|---|---|---|
| *Y* | 6 | 15 | 25 | 33 | 40 |

Определить тесноту связи *X* и *Y*, используя коэффициент корреляции и проверить на уровне $\alpha = 0{,}1$ его значимость.

**5.5.3.** В автохозяйстве исследовали зависимость между сроком эксплуатации автомобиля (*X* лет) и затратами на его обслуживание (ремонт, потребление топлива) (*Y*, % от первоначальной стоимости в год). Получены следующие данные:

| *X* | [1 – 3) | [3 – 5) | [5 – 9) | [9 – 13) | [13 – 19) |
|---|---|---|---|---|---|
| *Y* | 10 | 20 | 35 | 50 | 90 |

Определить тесноту связи *X* и *Y*, используя коэффициент корреляции, и проверить на уровне $\alpha = 0{,}05$ его значимость.

**5.5.4.** Определить тесноту связи общего веса *X*, г некоторого растения и веса *Y*, г его семян на основе следующих выборочных данных:

| *X* | 40 | 50 | 60 | 70 | 80 | 90 | 100 |
|---|---|---|---|---|---|---|---|
| *Y* | 20 | 25 | 28 | 30 | 35 | 40 | 45 |

Проверить значимость выборочного коэффициента корреляции при $\alpha = 0{,}05$.

**5.5.5.** Определить тесноту связи объёма полива ($X$, л/м$^2$) и урожайности ($Y$, кг/м$^2$) некоторой сельскохозяйственной культуры на основе следующих выборочных данных:

| | | | | | | |
|---|---|---|---|---|---|---|
| $X$ | 26 | 35 | 36 | 40 | 41 | 45 |
| $Y$ | 18 | 21 | 22,1 | 25,3 | 28 | 28,5 |

Проверить значимость выборочного коэффициента корреляции на уровне $\alpha = 0{,}05$.

**5.5.6.** Выборочные величины $X$ и $Y$ связаны таблицей:

| | | | | | |
|---|---|---|---|---|---|
| $X$ | 8 | 12 | 16 | 20 | 24 |
| $Y$ | 47 | 76 | 100 | 136 | 150 |

1. Определить с помощью выборочного коэффициента корреляции $r$ тесноту связи между величинами $X$ и $Y$.
2. Оценить значимость $r$ на уровне 0,05 (используя распределение Стьюдента).
3. Если коэффициент корреляции значим, то для него построить доверительный интервал, который с заданной надёжностью $\gamma = 0{,}95$ накрывает неизвестное значение $\rho$ (используя $Z$ – преобразование Фишера и таблицу значений функции Лапласа).
4. Для зависимости $Y$ от $X$, заданной корреляционной таблицей, найти оценки параметров $a$ и $b$ уравнения линейной регрессии $y_x = a + bx$, остаточную дисперсию $S_0^2$; выяснить значимость уравнения регрессии при $\alpha = 0{,}05$.

# 6. МНОГОМЕРНЫЕ СТАТИСТИЧЕСКИЕ МЕТОДЫ

**I. Учебные цели.** Познакомить обучающихся с идеями и возможностями некоторых многомерных статистических методов исследования.

В результате изучения предлагаемого материала студенты должны знать определения множественного коэффициента корреляции, выборочного коэффициента детерминации, частного коэффициента корреляции, знать модель множественной линейной регрессии и уметь применять эти знания при решении практических задач.

**II. Формирование компетенций.** Формирование общей и математической культуры, развитие способностей использовать законы естественнонаучных дисциплин в профессиональной области, способность применять аппарат математической статистики для принятия решений по выполнению работ по распределению и контролю использования производственно-технологических ресурсов.

**III. Введение в тему.** Социально-экономические процессы и явления зависят от большого числа параметров, их характеризующих, что обусловливает трудности, связанные с выявлением структуры взаимосвязей этих параметров. В подобных ситуациях, т.е. когда решения принимаются на основании анализа стохастической, неполной информации, использование методов многомерного статистического анализа является не только оправданным, но и существенно необходимым.

Многомерные статистические методы среди множества возможных вероятностно-статистических моделей позволяют обоснованно выбрать ту, которая наилучшим образом соответствует исходным статистическим данным, характеризующим реальное поведение исследуемой совокупности объектов, оценить надёжность и точность выводов, сделанных на основе ограниченного статистического материала.

Наличие множества исходных признаков, характеризующих процесс функционирования объектов, заставляет отбирать из них наиболее существенные и изучать меньший набор показателей. Чаще исходные признаки подвергаются некоторому преобразованию, которое обеспечивает минимальную потерю информации. Сжатие информации получается за счёт того, что число факторов используется значительно меньше, чем было исходных признаков.

При изучении данного материала обратите внимание на следующие вопросы:

1. Чем продиктована необходимость исследования многофакторных моделей.
2. Что такое множественный коэффициент корреляции?
3. Как проверяется значимость множественного коэффициента корреляции?

4. Что такое частные коэффициенты корреляции и как проверяется их значимость?

5. Какова область применения множественного регрессионного анализа?

6. Уравнение множественной линейной регрессии и нахождение его коэффициентов.

7. Оценка погрешности модели линейной регрессии.

8. Нелинейная регрессия и методы её линеаризации.

## 6.1. МНОЖЕСТВЕННЫЙ КОЭФФИЦИЕНТ КОРРРЕЛЯЦИИ

Задачи практики описываются чаще всего многофакторными моделями. Поэтому возникает вопрос о выяснении тесноты связи выбранного фактора со всеми остальными.

*Пример* 6.1. (пример постановки задачи). Решается задача: проведено исследование шести фермерских хозяйств для изучения зависимости урожайности зерновых культур *Z*, ц/га от качества пашни *X* (баллов) и количества внесённых удобрений *Y*, ц/га. Результаты приведены в таблице:

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 1 | *X* | 26 | 34 | 36 | 40 | 41 | 45 |
| 2 | *Y* | 2,1 | 2,3 | 2,4 | 2,6 | 2,9 | 3,0 |
| 3 | *Z* | 18 | 21 | 22,1 | 25,3 | 28 | 28,5 |

Необходимо оценить тесноту связи между величиной *Z* и величинами *X* и *Y* (одновременно).

Дальнейшие теоретические выкладки будем рассматривать относительно трёх случайных величин $X_1$, $X_2$ и $X_3$, имеющих совместное нормальное распределение. При решении конкретных задач, например 6.1, можно полагать, что $X_1 \sim X$, $X_2 \sim Y$, $X_3 \sim Z$.

Используя понятие коэффициента корреляции, по данным выборок из значений величин $X_1$, $X_2$, $X_3$ можно найти выборочные парные коэффициенты корреляции $r_{ij}$; $i, j = 1, 2, 3$ и составить матрицу, которая называется корреляционной:

$$A = \begin{pmatrix} r_{11} & r_{12} & r_{13} \\ r_{21} & r_{22} & r_{23} \\ r_{31} & r_{32} & r_{33} \end{pmatrix}, \qquad (6.1)$$

где $r_{ij} = 1$, если $i = j$; $r_{12} = r_{21}$, $r_{13} = r_{31}$, $r_{23} = r_{32}$.

Теснота линейной связи каждой из переменных $X_i$, $i = 1, 2, 3$ с совокупностью остальных переменных измеряется с помощью множественного коэффициента корреляции, который вычисляется по формуле

$$R_i = \sqrt{1 - \frac{|A|}{A_{ii}}}\,, \qquad (6.2)$$

где $|A|$ – определитель матрицы $A$; $A_{ii}$ – алгебраическое дополнение элемента $r_{ii}$. Величина $R_i$ является обобщением парного коэффициента корреляции.

Действительно, если $i$ = 1, 2, то

$$A = \begin{pmatrix} 1 & r_{12} \\ r_{21} & 1 \end{pmatrix};\ R_1 = \sqrt{1 - \frac{1 - r_{12}^2}{1}} = r_{12}\,,\ R_2 = \sqrt{1 - \frac{1 - r_{21}^2}{1}} = r_{21}\,;\ r_{12} = r_{21}.$$

Если $i$ = 1, 2, 3, то

$$R_i = \sqrt{\frac{r_{ij}^2 + r_{ik}^2 - 2 r_{ij}\, r_{ik}\, r_{jk}}{1 - r_{jk}^2}}\,, \qquad (6.3)$$

где $i \neq j$; $i \neq k$; $k \neq j$.

Величина $R_i$ удовлетворяет условию $0 \le R_i \le 1$; кроме того, для любого $j$: $R_j \ge |r_{ij}|$.

Величину $R_i^2$ называют выборочным коэффициентом детерминации. Он показывает, какой вклад в дисперсию $\sigma_i^2$ (разброс значений величины $X_i$) вносят остальные переменные. Например, $R_2^2$ показывает, какой вклад в дисперсию $\sigma_2^2$ (разброс значений $X_2$) вносят $X_1$ и $X_3$.

Множественный коэффициент корреляции значимо отличается от нуля, если статистика

$$t = \frac{R^2 (n - p)}{(1 - R^2)(p - 1)} > F_{\text{кр}}\,, \qquad (6.4)$$

где $F_{\text{кр}}$ – критическое значение распределения Фишера на уровне значимости $\alpha$ при числе степеней свободы $s_1 = p - 1$, $s_2 = n - p$; $n$ – число наблюдений величин $X_1, X_2, \ldots, X_p$, $p$ – число величин.

*Пример* 6.2. Пользуясь данными примера 6.1, на уровне значимости $\alpha$ = 0,05 найдём выборочный множественный коэффициент корреляции между величиной $Z$ – урожайность и $X$ – качеством почвы и $Y$ – количеством внесённых удобрений. Для вычисления парных коэффициентов составим таблицу:

| $i$ | $x_i$ | $y_i$ | $z_i$ | $x_i^2$ | $x_i y_i$ | $y_i^2$ | $x_i z_i$ | $z_i^2$ | $y_i z_i$ |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 26 | 2,1 | 18 | 676 | 54,6 | 4,41 | 468 | 324 | 37,8 |
| 2 | 35 | 2,3 | 21 | 1225 | 80,5 | 5,29 | 735 | 441 | 48,3 |
| 3 | 36 | 2,4 | 22,1 | 1296 | 86,4 | 5,76 | 795,6 | 488,41 | 53,04 |
| 4 | 40 | 2,6 | 25,3 | 1600 | 104 | 6,76 | 1012 | 640,09 | 65,78 |
| 5 | 41 | 2,9 | 28 | 1681 | 118,9 | 8,41 | 1148 | 784 | 81,2 |
| 6 | 45 | 3,0 | 28,5 | 2025 | 135 | 9 | 1282,5 | 812,25 | 85,5 |
| $\sum_{i=1}^{6}$ | 223 | 15,3 | 142,9 | 8503 | 579,4 | 39,63 | 5441,1 | 3489,75 | 371,62 |

Найдём выборочные парные коэффициенты корреляции, используя формулу (5.1)

$$r_{xy} = \frac{6\cdot\sum_{i=1}^{6} x_i y_i - \left(\sum_{i=1}^{6} x_i\right)\left(\sum_{i=1}^{6} y_i\right)}{\sqrt{6\cdot\sum_{i=1}^{6} x_i^2 - \left(\sum_{i=1}^{6} x_i\right)^2}\sqrt{6\cdot\sum_{i=1}^{6} y_i^2 - \left(\sum_{i=1}^{6} y_i\right)^2}} =$$

$$= \frac{6\cdot 579{,}4 - 223\cdot 15{,}3}{\sqrt{6\cdot 8503 - 223^2}\cdot\sqrt{6\cdot 39{,}63 - 15{,}3^2}} = 0{,}935;$$

$$r_{xz} = \frac{6\cdot\sum_{i=1}^{6} x_i z_i - \left(\sum_{i=1}^{6} x_i\right)\left(\sum_{i=1}^{6} z_i\right)}{\sqrt{6\cdot\sum_{i=1}^{6} x_i^2 - \left(\sum_{i=1}^{6} x_i\right)^2}\sqrt{6\cdot\sum_{i=1}^{6} z_i^2 - \left(\sum_{i=1}^{6} z_i\right)^2}} =$$

$$= \frac{6\cdot 5441{,}1 - 223\cdot 142{,}9}{\sqrt{6\cdot 8503 - 223^2}\cdot\sqrt{6\cdot 3489{,}75 - 142{,}9^2}} = 0{,}954;$$

$$r_{yz} = \frac{6\cdot\sum_{i=1}^{6} y_i z_i - \left(\sum_{i=1}^{6} y_i\right)\left(\sum_{i=1}^{6} z_i\right)}{\sqrt{6\cdot\sum_{i=1}^{6} y_i^2 - \left(\sum_{i=1}^{6} y_i\right)^2}\sqrt{6\cdot\sum_{i=1}^{6} z_i^2 - \left(\sum_{i=1}^{6} z_i\right)^2}} =$$

$$= \frac{6\cdot 371{,}62 - 15{,}3\cdot 142{,}9}{\sqrt{6\cdot 39{,}63 - 15{,}3^2}\cdot\sqrt{6\cdot 3489{,}75 - 142{,}9^2}} = 0{,}991.$$

Матрица корреляций имеет вид

$$A = \begin{pmatrix} 1 & r_{xy} & r_{xz} \\ r_{yx} & 1 & r_{yz} \\ r_{zx} & r_{zy} & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0{,}935 & 0{,}954 \\ 0{,}935 & 1 & 0{,}991 \\ 0{,}954 & 0{,}991 & 1 \end{pmatrix}.$$

По формуле (6.3) найдём

$$R_z = \sqrt{\frac{r_{xz}^2 + r_{yz}^2 - 2 r_{xz}\, r_{yz}\, r_{xy}}{1 - r_{xy}^2}} = \sqrt{\frac{0{,}954^2 + 0{,}991^2 - 2 \cdot 0{,}954 \cdot 0{,}991 \cdot 0{,}935}{1 - 0{,}935^2}} = 0{,}993.$$

Контрольный результат: $R_z > r_{yz}$; $R_z > r_{xy}$; $R_z > r_{xz}$.

Для проверки значимости $R_z$ составим статистику по формуле (6.4):

$$t = \frac{R_z^2 (n - p)}{\left(1 - R_z^2\right)(p - 1)} = \frac{0{,}993^2 (6 - 3)}{(1 - 0{,}993^2)(3 - 1)} = \frac{2{,}95}{0{,}028} = 105{,}4.$$

При $\alpha = 0{,}05$ и числе степеней свободы $s_1 = p - 1 = 2$ и $s_2 = n - p = 3$ критическое значение распределения Фишера равно $F_{кр} = 9{,}55$. Так как $t >> F_{кр}$, то выборочный коэффициент корреляции $R_z$ является заведомо значимым. Таким образом, связь между урожайностью и условиями выращивания культуры является весьма тесной.

## 6.2. ЧАСТНЫЙ КОЭФФИЦИЕНТ КОРРЕЛЯЦИИ

Выборочным частным коэффициентом корреляции между переменными $X_i$ и $X_j$ называется выражение

$$R_{ij} = \frac{-A_{ij}}{\sqrt{A_{ii} - A_{jj}}}, \qquad (6.5)$$

где $A_{ij}$, $A_{ii}$, $A_{jj}$ – алгебраические дополнения элементов $r_{ij}$, $r_{ii}$, $r_{jj}$ матрицы $A$ (6.1) соответственно.

В случае трёх переменных из (6.5) следует

$$R_{ij} = \frac{r_{ij} - r_{ik}\, r_{jk}}{\sqrt{\left(1 - r_{ik}^2\right)\left(1 - r_{jk}^2\right)}}, \qquad (6.6)$$

где $k \neq i$, $k \neq j$. Частный коэффициент корреляции по своим свойствам не отличается от парного, только при оценке его значимости число степеней свободы полагают равным $n - p + 2$ и исследуется статистика

$$t = \frac{r\sqrt{n-p+2}}{\sqrt{1-r^2}},$$

которая имеет распределение Стьюдента с $n - p + 2$ степенями свободы.

*Пример* 6.3. Пользуясь данными примера 6.2, установить тесноту связи между урожайностью $Z$ и качеством пашни $X$.

*Решение.* По формуле (6.6) находим частный коэффициент корреляции

$$R_{xz} = \frac{r_{xz} - r_{xy} \cdot r_{yz}}{\sqrt{1-r_{xy}^2}\sqrt{1-r_{yz}^2}} = \frac{0{,}954 - 0{,}935 \cdot 0{,}991}{\sqrt{1-0{,}935^2}\sqrt{1-0{,}991^2}} = 0{,}578.$$

По величине $R_{xz} = 0{,}578$ делаем вывод, что связь между $X$ и $Z$ является «средней».

В то же время

$$R_{yz} = \frac{r_{yz} - r_{xy} \cdot r_{xz}}{\sqrt{1-r_{xy}^2}\sqrt{1-r_{xz}^2}} = \frac{0{,}991 - 0{,}935 \cdot 0{,}954}{\sqrt{1-0{,}935^2}\sqrt{1-0{,}954^2}} = \frac{0{,}09901}{0{,}355 \cdot 0{,}2998} = 0{,}931.$$

Значение $R_{yz}$ говорит о том, что связь между количеством вносимых удобрений и урожайностью более тесная.

Значимость $R_{xz}$:

$$t = \frac{R_{xz}\sqrt{n-p+2}}{\sqrt{1-R_{xz}^2}} = \frac{0{,}578 \cdot \sqrt{5}}{\sqrt{1-0{,}578^2}} = \frac{1{,}292}{0{,}816} = 1{,}583.$$

$t_{кр} = t_{0{,}05;\,5}$ (распределение Стьюдента) = 2,57.

Так как $t < t_{кр}$, то $R_{xz}$ малозначим.

Значимость $R_{yz}$:

$$t = \frac{0{,}931\sqrt{5}}{\sqrt{1-0{,}931^2}} = 7{,}94.$$

Так как $t > t_{кр}$, то $R_{yz}$ достаточно значим.

## 6.3. МНОЖЕСТВЕННЫЙ РЕГРЕССИОННЫЙ АНАЛИЗ

Будем исследовать зависимость одной переменной $Z$ от переменных $X$ и $Y$ (например, в той форме, как написано в примере 6.1.), когда каждой наблюдаемой паре значений $X$ и $Y$ соответствует единственное значение величины $Z$. Однако вряд ли таким свойством будет обладать генеральная совокупность: ведь на $Z$ помимо $X$ и $Y$ влияет ряд других случайных факторов. Поэтому обратим внимание на изучение корреляционной зависимости величины $Z$ от $X$ и $Y$, т.е. зависимости условного математического

ожидания $M(Z/X = x, Y = y)$ или $M_{xy}(Z)$ (математического ожидания величины $Z$, вычисленного при условии $X = x$, $Y = y$) от значений $x$ и $y$.

Предположим, что функция регрессии линейная, т.е.

$$M_{xy}(Z) = a + bx + cy.$$

Оценочные коэффициенты (коэффициенты уравнения регрессии)

$$\hat{Z} = \hat{a} + \hat{b}x + \hat{c}y \tag{6.7}$$

найдём из требования метода наименьших квадратов:

$$F(\hat{a}, \hat{b}, \hat{c}) = \frac{1}{n}\sum_{i=1}^{n}\left(Z_i - \hat{Z}_i\right)^2 = \frac{1}{n}\sum_{i=1}^{n}\left(Z_i - \hat{a} - \hat{b}x_i - \hat{c}y_i\right)^2 \to \min .$$

Необходимое условие минимума функции $F$ образует систему равных нулю частных производных, которая в результате тождественных преобразований принимает вид

$$\begin{cases} \hat{a} + \hat{b}\overline{X} + \hat{c}\overline{Y} = \overline{Z}, \\ \hat{a}\overline{X} + \hat{b}\overline{X}^2 + \hat{c}\overline{YX} = \overline{ZX}, \\ \hat{a}\overline{Y} + \hat{b}\overline{XY} + \hat{c}\overline{Y}^2 = \overline{ZY}, \end{cases} \tag{6.8}$$

где $\overline{X}, \overline{Y}, \overline{Z}, \overline{X}^2, \overline{Y}^2, \overline{XY}, \overline{XZ}, \overline{YZ}$ – средние значения соответствующих случайных величин или их произведений, рассчитанные на основе наблюдений*.

Погрешность уравнения (6.7) определяется как

$$S_0 = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Z_i - \hat{Z}_i)^2},$$

где $Z_i - \hat{Z}_i$ – разности наблюдаемых и рассчитанных значений переменной $Z$, и сравнивается с выборочным средним квадратическим отклонением величины $Z$: $S_Z = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Z_i - \overline{Z})^2}$; отношение $\frac{S_Z}{S_0}$ показывает, во сколько раз погрешность модели $Z_i \approx \overline{Z}$ больше погрешности модели $Z_i \approx \hat{Z}_i$.

* На практике каждое из этих уравнений умножают на $n$ – число наблюдений, в результате вместо средних значений подсчитывают только суммы соответствующих величин.

Кроме того, в оценке степени линейной корреляционной зависимости величины $Z$ от $X$ и $Y$ используется величина $S_{\hat{Z}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{Z}_i - \overline{Z})^2}$ – дисперсия признака $\hat{Z}$, вычисленного по уравнению (6.7).

Выборочный множественный коэффициент корреляции $R_Z$, квадрат которого равен $R_Z^2 = S_{\hat{Z}}^2 \big/ S_Z^2$, показывает, какую долю от дисперсии $S_Z^2$ результативного признака Z составляет дисперсия $S_{\hat{Z}}^2$, «зетов», вычисленных по линейному уравнению регрессии.

Обращаем внимание на то, что с точностью до ошибок округления должно выполняться равенство

$$S_Z^2 = S_{\hat{Z}}^2 + S_0^2.$$

*Пример* 6.4. Для данных примера 6.2 найти множественное уравнение регрессии и оценить погрешность модели линейной регрессии.

*Решение.* Уравнение регрессии $\hat{Z} = a + bx + cy$ [*]. В этом случае система (6.8), где каждая строка умножена на $n$ = 6, имеет вид

$$\begin{cases} 6a + 223b + 15{,}3c = 142{,}9, \\ 223a + 8503b + 579{,}4c = 5441{,}1, \\ 15{,}3a + 579{,}4b + 39{,}63c = 371{,}62. \end{cases}$$

Её решения: $a = -5{,}12$; $b = 0{,}137$; $c = 9{,}35$.

Таким образом, $\hat{Z} = -5{,}12 + 0{,}137x + 9{,}35y$.

Значения $\hat{Z}_i = \hat{Z}(x_i, y_i)$, найденные по уравнению регрессии:

$$\hat{Z}_1 = 18{,}077;\ \ \hat{Z}_2 = 21{,}18;\ \ \hat{Z}_3 = 22{,}25;\ \ \hat{Z}_4 = 24{,}67;\ \ \hat{Z}_5 = 27{,}61;\ \ \hat{Z}_6 = 29{,}10.$$

Тогда:

$$S_0^2 = \frac{1}{6}\sum_{i=1}^{6}(Z_i - \hat{Z}_i)^2 = \frac{1}{6}(0{,}077^2 + 0{,}18^2 + 0{,}15^2 + 0{,}63^2 + 0{,}39^2 + 0{,}6^2) = 0{,}1616.$$

$S_0 = 0{,}402.$

$$S_Z^2 = \frac{1}{6}\sum_{i=1}^{6}(Z_i - \overline{Z})^2 = \frac{1}{6}(5{,}82^2 + 2{,}82^2 + 1{,}72^2 + 1{,}48^2 + 4{,}18^2 + 4{,}68^2) = 14{,}39.$$

$S_Z = 3{,}79.\quad S_Z / S_0 = 3{,}79/0{,}402 \approx 9{,}5.$

[*] Здесь и далее нет необходимости использовать надбуквенные символы для коэффициентов $a$, $b$ и $c$.

Таким образом, погрешность модели $Z_i \approx \hat{Z}_i$ в 9,5 раз меньше погрешности модели $Z_i \approx \overline{Z}$.

Находим, что $S_{\hat{Z}}^2 = S_Z^2 - S_0^2 = 14{,}23$.

Тогда $R_Z^2 = S_{\hat{Z}}^2 \big/ S_Z^2 = 14{,}23/14{,}39 = 0{,}996.$

Следовательно, более 99% дисперсии $S_Z^2$ объясняется линейной в среднем зависимостью $Z$ от $X$ и $Y$.

## 6.4. НЕЛИНЕЙНАЯ РЕГРЕССИЯ

Линейное уравнение регрессии представляет собой простейший вид регрессии. На практике, однако, часто встречаются более сложные виды зависимостей между независимыми переменными $X_1, X_2, \ldots, X_n$ и зависимой переменной $Y$, например,

$$Y = a + \frac{b}{X}, \quad Y = a + bX + aX^2, \quad Y = ae^{bX} \text{ и т.д.}$$

Выбор конкретного вида зависимостей обусловлен экономическими или техническими условиями задачи и во многом зависит от уровня квалификации исследователя, который при этом может использовать предыдущий опыт, теоретические предпосылки или иные соображения.

В последующем правильность выбора регрессионной модели оценивается остаточной дисперсией или соответствующими критериями.

Если каким-либо образом предварительно установлен вид регрессионной модели и она оказалась нелинейной, то, как правило, её можно свести к линейной путём соответствующей замены переменных.

Пусть, например, зависимость $Y$ от $X$ выражается уравнением $Y = a + b/X$. Такая зависимость может иметь место в экономике при характеристике соотношений между нормой безработицы и процентом прироста заработной платы (кривая Филипса). Обозначим $Z = 1/X$, тогда $Y = a + bZ$ – линейное уравнение регрессии, исследование которого показано в § 5.3 и 5.4.

Аналогично, если $Y = a + bX + cX^2$, то, вводя замену $X^2 = Z$, получим двухфакторное уравнение регрессии

$$Y = a + bX + cZ,$$

исследование которого показано в § 6.3.

Если $Y = ae^{bX}$, то, логарифмируя это равенство, получим

$$\ln Y = \ln a + bX.$$

Введя обозначения $\ln Y = Z$ и $\ln a = a_0$, получим линейное уравнение регрессии $Z = a_0 + bX$.

Если $Y = aX_1^b X_2^c$, $Y > 0$, $X_1 > 0$, $X_2 > 0$, то

$$\ln Y = \ln a + b \ln X_1 + c \ln X_2 .$$

Полагая $\ln Y = Z$, $\ln a = a_0$, $\ln X_1 = X$ и $\ln X_2 = Y$, получим $Z = a_0 + bX + cY$ – множественное (двухфакторное) уравнение регрессии § 6.3.

Продемонстрированные схемы линеаризации нелинейных уравнений регрессии требуют «осторожного» использования, так как реализация метода наименьших квадратов осуществляется не на исходных данных, а на преобразованных величинах. Например, оценка параметров основывается на минимизации суммы квадратов отклонений в логарифмах, а вследствие этого оценка параметров для линеаризуемых функций оказывается несколько смещённой.

Практическое применение экспоненты возможно, если результативный признак не имеет отрицательных значений. Поэтому, если исследуется, например, финансовый результат деятельности предприятия, среди которых наряду с прибыльными есть и убыточные, то данная функция не может быть использована.

*Пример* 6.5. Данные результатов наблюдений представлены в таблице:

| $X$ | –2 | –1 | 0 | 1 | 2 |
|---|---|---|---|---|---|
| $Y$ | 4,8 | 0,4 | –3,3 | –0,8 | 3,2 |

Определить методом наименьших квадратов параметры $a$, $b$, $c$ зависимости вида $y = a + bx + cx^2$.

При замене $Z = x^2$ получим таблицу двухфакторной корреляции:

| $X$ | –2 | –1 | 0 | 1 | 2 |
|---|---|---|---|---|---|
| $Z = X^2$ | 4 | 1 | 0 | 1 | 4 |
| $Y$ | 4,8 | 0,4 | –3,3 | –0,8 | 3,2 |

Система уравнений (6.8) имеет вид

$$\begin{cases} a + b\overline{X} + c\overline{Z} = \overline{Y}, \\ a\overline{X} + b\overline{X}^2 + c\overline{XZ} = \overline{YX}, \\ a\overline{X^2} + b\overline{XZ} + c\overline{Z^2} = \overline{ZY}. \end{cases} \quad \text{или} \quad \begin{cases} a + b\overline{X} + c\overline{X^2} = \overline{Y}, \\ aX + b\overline{X^2} + c\overline{X^3} = \overline{YX}, \\ a\overline{X^2} + b\overline{X^3} + c\overline{X^4} = \overline{YX^2}. \end{cases} \qquad (6.9)$$

Составим вспомогательную таблицу и произведём расчёты, необходимые для решения этой системы:

| № | $X$ | $Y$ | $X^2$ | $X^3$ | $X^4$ | $XY$ | $YX^2$ | $Y_r$ | $(Y-\bar{Y})^2$ | $(Y_r-\bar{Y})^2$ |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | –2 | 4,8 | 4 | –8 | 16 | –9,6 | 19,2 | 5,02 | 15,52 | 17,31 |
| 2 | –1 | 0,4 | 1 | –1 | 1 | –0,4 | 0,4 | –0,34 | 0,21 | 1,44 |
| 3 | 0 | –3,3 | 0 | 0 | 0 | 0 | 0 | –2,42 | 17,31 | 10,76 |
| 4 | 1 | –0,8 | 1 | 1 | 1 | –0,8 | –0,8 | –1,22 | 2,76 | 4,33 |
| 5 | 2 | 3,2 | 4 | 8 | 16 | 6,4 | 12,8 | 3,26 | 5,48 | 5,76 |
| $\sum$ | 0 | 4,3 | 10 | 0 | 34 | –4,4 | 31,6 | | $Q = 41{,}28$ | $Q_ф = 39{,}6$ |

На основании полученных результатов расчёта, система уравнений (6.9), где каждая сторона умножена на $n = 5$, примет вид

$$\begin{cases} 5a + 0b + 10c = 4{,}3, \\ 0a + 10b + 0c = -4{,}4, \\ 10a + 0b + 34c = 31{,}6. \end{cases}$$

Её решения: $a = -2{,}42$; $b = -0{,}44$; $c = 1{,}64$.

Таким образом, уравнение регрессии примет вид

$$Y_r = -2{,}42 - 0{,}44x + 1{,}64x^2.$$

Необходимо учитывать то обстоятельство, что ввиду симметричности кривой парабола второй степени далеко не всегда пригодна в конкретных исследованиях, чаще имеют дело лишь с отдельными сегментами параболы.

Чтобы решить вопрос о значимости полученного уравнения регрессии, найдём рассчитанные на его основе значения $Y_r$ – величины $Y$ и разместим их в приведённую выше таблицу.

Далее, найдём, что факторная дисперсия $Q_ф = 39{,}6$ , а общая $Q = 41{,}28$.

Тогда $Q_0 = Q - Q_ф = 1{,}68$.

Статистика $t = \dfrac{39{,}6 \cdot 3}{1{,}68} = 70{,}71.$

Уравнение регрессии значимо, так как $F_{кр} = F_{0{,}05;\,1;\,3} = 10{,}1 << t$.

Оценка остаточной дисперсии $S_0^2 = \dfrac{1{,}68}{3} = 0{,}56$ ; $S_0 = 0{,}75$.

*Пример* 6.6. Данные результатов наблюдения приведены в таблице

| $X$ | 1 | 2 | 4 | 8 | 10 |
|---|---|---|---|---|---|
| $Y$ | 7,0 | 6,5 | 5,0 | 4,0 | 3,5 |

На уровне значимости $\alpha = 0{,}05$ проверить значимость моделей регрессии:

1. Линейной: $Y = a + bX$.

2. Нелинейной вида: $Y = a_1 + b_1/X$.

3. Нелинейной вида: $Y = a_2 + b_2 X + c_2 X^2$.

Сравнить оценки соответствующих остаточных дисперсий и сделать выводы о предпочтительности (моделей).

1. Предположим, что уравнение регрессии линейное: $Y = a + bX$.

Для выполнения расчётов составим вспомогательную таблицу:

| № | $X$ | $X^2$ | $Y$ | $XY$ | $Y_x$ | $(Y-\overline{Y})^2$ | $(Y_x-\overline{Y})^2$ |
|---|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 7,0 | 1 | 6,79 | 3,24 | 2,53 |
| 2 | 2 | 4 | 6,5 | 13 | 6,41 | 1,69 | 1,46 |
| 3 | 4 | 16 | 5,0 | 20 | 5,65 | 0,04 | 0,203 |
| 4 | 8 | 64 | 4,0 | 32 | 4,13 | 1,44 | 1,145 |
| 5 | 10 | 100 | 3,5 | 35 | 3,37 | 2,89 | 3,35 |
| $1/5\sum$ | 5 | 37 | 5,2 | 21,4 | | $Q = 9{,}3$ | $Q_ф = 8{,}69$ |

Согласно формуле (5.3) имеем:

$$a = \frac{37 \cdot 5{,}2 - 5 \cdot 21{,}4}{37 - 25} = 7{,}17; \quad b = \frac{21{,}4 - 5 \cdot 5 \cdot 2}{37 - 25} = -0{,}38 .$$

Уравнение регрессии $Y = 7{,}17 - 0{,}38X$ .

Используя это уравнение, найдём $Y_x$ – расчётные значения $Y$ и разместим их в таблице. На основе данных таблицы найдём общую $Q$ и факторную $Q_ф$ дисперсии.

Далее определим, что $Q_0 = Q - Q_ф = 9{,}3 - 8{,}69 = 0{,}61$.

Статистика $t = \dfrac{8{,}69 \cdot 3}{0{,}61} = 42{,}74.$

Уравнение регрессии значимо, так как $F_{кр} = F_{0{,}05;\ 1;\ 3} = 10{,}1 < t$.

Оценка остаточной дисперсии:

$$S_0^2 = \frac{0{,}61}{3} = 0{,}203; \quad S_0 = 0{,}45.$$

2. Исследуем нелинейную модель $Y = a_1 + b_1/X$. Введём новую переменную $Z = 1/X$, тогда $y = a_1 + b_1 Z$ – уравнение линейной регрессии. В новых переменных исходная таблица имеет вид

| $Z$ | 1 | 0,5 | 0,25 | 0,125 | 0,1 |
|---|---|---|---|---|---|
| $Y$ | 7,0 | 6,5 | 5,0 | 4,0 | 3,5 |

Согласно формуле (5.3) для этого уравнения коэффициенты регрессии:

$$a_1 = \frac{\overline{Z^2}\,\overline{Y} - \overline{Z}\,\overline{ZY}}{\overline{Z^2} - (\overline{Z})^2}; \quad b_1 = \frac{\overline{ZY} - \overline{Z}\,\overline{Y}}{\overline{Z^2} - (\overline{Z})^2}.$$

Для выполнения расчётов составим вспомогательную таблицу:

| № | $Z$ | $Z^2$ | $Y$ | $ZY$ | $Y_Z$ | $(Y - \overline{Y})^2$ | $(Y_z - \overline{Y})^2$ |
|---|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 7,0 | 7,0 | 7,44 | 3,24 | 5,02 |
| 2 | 0,5 | 0,25 | 6,5 | 3,25 | 5,58 | 1,69 | 0,144 |
| 3 | 0,25 | 0,0625 | 5,0 | 1,25 | 4,65 | 0,04 | 0,303 |
| 4 | 0,125 | 0,0156 | 4,0 | 0,5 | 4,19 | 1,44 | 1,02 |
| 5 | 0,1 | 0,01 | 3,5 | 0,35 | 4,09 | 2,89 | 1,23 |
| $1/5\sum$ | 0,395 | 0,268 | 5,2 | 2,47 | | $Q$ = 9,3 | $Q_ф$ = 7,72 |

Имеем:

$$a_1 = \frac{0{,}268 \cdot 5{,}2 - 0{,}395 \cdot 2{,}47}{0{,}268 - (0{,}395)^2} = 3{,}73; \quad b_1 = \frac{2{,}47 - 0{,}395 \cdot 5 \cdot 2}{0{,}268 - (0{,}395)^2} = 3{,}71.$$

Таким образом, уравнение регрессии:

$$Y = 3{,}73 + 3{,}71Z = 3{,}73 + 3{,}71/X.$$

В таблице $Y_Z$ – рассчитанные по уравнению регрессии значения $Y$. $Q = Q_ф + Q_0$; $Q_0 = 9{,}3 - 7{,}72 = 1{,}58$.

Статистика $t = \frac{Q_ф(n-l)}{Q_0(l-1)} = \frac{7{,}72 \cdot 3}{1{,}58 \cdot 1} = 14{,}66.$

По таблице распределения Фишера при $\alpha = 0{,}05$

$$F_{кр} = F_{\alpha;\,1;\,3} = 10{,}1.$$

Так как $F_{кр} < t$, то уравнение регрессии значимо.

Оценка остаточной дисперсии

$$S_0^2 = \frac{Q_0}{n-l} = \frac{1{,}58}{3} = 0{,}53; \; S_0 = 0{,}73.$$

3. Исследуем нелинейную модель $Y = a_2 + b_2X + c_2X^2$. Введём новую переменную $Z = X^2$, тогда $Y = a_2 + b_2X + c_2Z$ – множественное уравнение регрессии, поиск нахождения коэффициентов которого изложен в 6.3.

Для выполнения расчётов составим вспомогательную таблицу.

| № | $X$ | $Y$ | $X^2$ | $X^3$ | $X^4$ | $XY$ | $X^2Y$ | $Y_{xz}$ | $(Y-\bar{Y})^2$ | $(Y_{xz}-\bar{Y})^2$ |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 7,0 | 1 | 1 | 1 | 7,0 | 7,0 | 7,04 | 5,02 | 3,39 |
| 2 | 2 | 6,5 | 4 | 8 | 16 | 13 | 26 | 6,33 | 0,144 | 1,28 |
| 3 | 4 | 5,0 | 16 | 64 | 256 | 20 | 80 | 5,20 | 0,303 | 0 |
| 4 | 8 | 4,0 | 64 | 512 | 4096 | 32 | 256 | 3,82 | 1,02 | 1,90 |
| 5 | 10 | 3,5 | 100 | 1000 | 10 000 | 35 | 350 | 3,59 | 1,23 | 2,59 |
| $\sum$ | 5 | 5,2 | 37 | 317 | 2873,8 | 21,4 | 143,8 | | $Q$ = 9,3 | $Q_ф$ = 9,16 |

Имеем систему уравнений:

$$\begin{cases} a_2 + 5b_2 + 37c_2 = 5{,}2, \\ 5a_2 + 37b_2 + 317c_2 = 21{,}4, \\ 37a_2 + 317b_2 + 2873{,}8c_2 = 143{,}8. \end{cases}$$

Её решение: $a_2 = 7{,}798$; $b_2 = -0{,}8$; $c_2 = 0{,}03788$.

Уравнение регрессии: $Y = 7{,}798 - 0{,}8X + 0{,}03788X^2$.

Используя это уравнение, найдём $Y_{xz}$ – расчётные значения $Y$ и разместим их в таблице. На основе данных таблицы найдём общую $Q$ и факторную $Q_ф$ дисперсии: $Q = 9{,}3$; $Q_ф = 9{,}16$.

Далее определяем, что остаточная дисперсия

$$Q_0 = Q - Q_ф = 9{,}3 - 9{,}16 = 0{,}14.$$

Статистика $t = \dfrac{9{,}16 \cdot 3}{0{,}14} = 196{,}3.$

Уравнение регрессии значимо, так как $F_{кр} = F_{0{,}05;\,1;\,3} = 10{,}1 << t$.

Оценка остаточной дисперсии $S_0^2 = \dfrac{0{,}14}{3} = 0{,}0467$; $S_0 = 0{,}216$.

Сравнивая оценки остаточных дисперсий, можно заключить, что наименьшая у модели квадратичной параболы. Однако и другие модели достаточно значимы.

## 6.5. ОСОБЕННОСТИ МНОЖЕСТВЕННОЙ РЕГРЕССИИ И КОРРЕЛЯЦИИ

Как было показано ранее, при исследовании зависимости результативного признака $Y$ от ряда факторов $X_1$, $X_2$, …, $X_p$ необходимо решать такие же задачи, что и при парной связи двух переменных $X$ и $Y$:

– определение вида регрессии;

– оценка параметров;

– определение тесноты связи, если $X$ и $Y$ случайные величины.

Однако наряду с этими задачами необходимо рассматривать и такие, которые характерны лишь для множественной регрессии и корреляции.

К таким задачам относится отбор факторов $X_1$, $X_2$, …, $X_k$, существенно влияющих на фактор $Y$, при наличии возможностей внутренней взаимосвязи между переменными $X_1$, $X_2$, …, $X_k$. Такой отбор требует глубокого теоретического и практического знания качественной стороны рассматриваемых явлений.

Отбор факторов осуществляется в несколько этапов. Сначала отбираются факторы, связанные с изучаемым явлением, на основе данных теоретического исследования. При этом для построения множественной регрессии и корреляции отбираются факторы, которые могут быть измерены.

Далее отобранные факторы подвергаются математико-статистической проверке существенности их влияния на изучаемый показатель. Такая проверка включает анализ матрицы парных корреляций, частных корреляций, проверку значимости коэффициентов регрессии, анализ остаточных отклонений и т.д.

Рассмотрим процедуру отбора факторов для построения множественной линейной зависимости, когда переменные $y$, $x_1$, $x_2$, …, $x_p$ являются случайными величинами*.

Наиболее простой формой зависимости и достаточно строго обоснованной для случая совместного нормального распределения является линейная, т.е. зависимость вида

$$y = a_0 + a_1 x_1 + a_2 x_2 + ... + a_p x_p . \qquad (6.10)$$

Исходная информация для построения зависимости (6.9), например, при $p = 3$, задаётся в виде некоторой таблицы вида

| № | Факторы, для которых получены данные | | | |
|---|---|---|---|---|
| | $y$ | $x_1$ | $x_2$ | $x_3$ |
| 1 | $y_1$ | $x_{11}$ | $x_{21}$ | $x_{31}$ |
| 2 | $y_2$ | $x_{12}$ | $x_{22}$ | $x_{32}$ |
| 3 | $y_3$ | $x_{13}$ | $x_{23}$ | $x_{33}$ |

Следует определить, все ли переменные необходимо включить в уравнение (6.10) или есть переменные, которые существенно не влияют на величину $y$ и их нецелесообразно включать.

* Далее используем обозначения переменных строчными буквами.

Для решения этого вопроса часто используется таблица, составленная из коэффициентов парной корреляции $r_{ij}$, $i, j = \overline{1, 3}$. Учитывая, что $r_{ij} = r_{ji}$, $i \neq j$ и $r_{ij} = 1$, если $i = j$. Эту таблицу можно записывать в упрощённой симметричной форме (треугольная форма):

| | $y$ | $x_1$ | $x_2$ | $x_3$ |
|---|---|---|---|---|
| $y$ | 1 | $r_{x_1,y}$ | $r_{x_2,y}$ | $r_{x_3,y}$ |
| $x_1$ | | 1 | $r_{x_2x_1}$ | $r_{x_3x_1}$ |
| $x_2$ | | | 1 | $r_{x_3x_2}$ |
| $x_3$ | | | | 1 |

По данной таблице можно примерно оценить, какие факторы существенно влияют на переменную $y$, а какие – несущественно, а также выявить взаимосвязь между факторами.

*Пример* 6.6. Рассмотрим конкретную таблицу:

| | $y$ | $x_1$ | $x_2$ | $x_3$ |
|---|---|---|---|---|
| $y$ | 1 | 0,65 | 0,6 | 0,03 |
| $x_1$ | | 1 | 0,5 | 0,9 |
| $x_2$ | | | 1 | 0,3 |
| $x_3$ | | | | 1 |

На основании указанных в таблице парных коэффициентов корреляции можно сделать вывод, что связь факторов $x_1$, $x_2$ с фактором $y$ существенная (коэффициенты корреляции соответственно 0,65; 0,6). В то же время величина коэффициента парной корреляции между $y$ и $x_3$ мала, в связи с этим нецелесообразно включать фактор $x_3$ в уравнение (6.9). Высок коэффициент корреляции между переменными $x_1$ и $x_3$ (0,9), что показывает их тесную корреляционную взаимосвязь. В этом случае не включают одновременно в уравнение (6.9) $x_1$ и $x_3$, а вводят один из них в зависимости от их смысла. Нецелесообразно включать в уравнение одновременно показатели, представляющие сумму некоторых факторов или их составных частей.

Кроме анализа таблицы парных коэффициентов корреляции для отбора существенных факторов вычисляют частные коэффициенты корреляции, определяют надёжность полученных коэффициентов регрессии.

Зачастую связи между изучаемыми переменными довольно сложным образом переплетаются, поэтому целесообразно рассматривать (дополнительно) вопрос о взаимосвязи между факторами при условии, что некоторые, или все остальные факторы остаются неизменными.

Для выявления такой взаимосвязи используются коэффициенты частной корреляции. Например, такой коэффициент между факторами $y$ и $x_1$ при условии, что фактор $x_2$ остаётся неизменным, определяется по формуле (6.6)

$$r_{yx_1(x_2)} = \frac{r_{yx_1} - r_{yx_2}\, r_{x_1x_2}}{\sqrt{1-r_{yx_2}^2}\sqrt{1-r_{x_1x_2}^2}}.$$

Если закреплён лишь один фактор, то такой коэффициент называется коэффициентом частной корреляции первого порядка. Если закреплены два, то – второго и т.д. При этом обычный коэффициент корреляции можно назвать частным коэффициентом корреляции нулевого порядка.

Как показывает опыт, малость коэффициентов частной корреляции низших порядков не гарантирует малость коэффициентов более высокого порядка, что надо иметь в виду при отборе существенных факторов.

После предварительного отбора факторов на основе парных и частных коэффициентов корреляции производится оценка параметров $a_0$, $a_1$, …, $a_p$, чаще всего по методу наименьших квадратов. Система уравнений в случае линейной зависимости (6.9) при $p = 3$ имеет вид

$$\begin{cases} a_0 n + a_1 \sum x_{1i} + a_2 \sum x_{2i} + a_3 \sum x_{3i} = \sum y_i, \\ a_0 \sum x_{1i} + a_1 \sum x_{1i}^2 + a_2 \sum x_{1i} \cdot x_{2i} + a_3 \sum x_{1i} \cdot x_{3i} = \sum x_{1i} y_i, \\ a_0 \sum x_{2i} + a_1 \sum x_{1i} \cdot x_{2i} + a_2 \sum x_{2i}^2 + a_3 \sum x_{2i} x_{3i} = \sum x_{2i} y_i, \\ a_0 \sum x_{3i} + a_1 \sum x_{1i} \cdot x_{3i} + a_2 \sum x_{2i} \cdot x_{3i} + a_3 \sum x_{3i}^2 = \sum x_{3i} y_i. \end{cases}$$ *

Решение такой системы может осуществляться методами Крамера, Гаусса и другими методами.

Для определения тесноты связи между фактором $y$ и совокупностью факторов $x_0$, $x_1$, …, $x_p$ в случае линейной зависимости применяется коэффициент множественной корреляции $R$ (6.2).

Если факторы – аргументы не являются случайными величинами, то коэффициенты корреляции не могут быть использованы при построении уравнения регрессии, так как не могут быть интерпретированы как показатели тесноты связи.

---

* Здесь все суммы $\sum$ от $i = 1$ до $i = n$.

Существенность вводимых факторов в случае линейной множественной регрессии может быть проверена одновременно с существенностью коэффициентов регрессии.

Для этого вычисляется отношение

$$t_i = a_i / \sigma_i \ , \ i = \overline{1, n} \ ,$$

где $a_i$ – коэффициент множественной регрессии; $\sigma_i$ – среднее квадратическое отклонение этого коэффициента.

Если $t_i > t_{табл}$, взятого по таблицам $t$-распределения Стьюдента, то с заданной вероятностью не отвергается гипотеза, что соответствующий коэффициент регрессии $a_i$ в генеральной совокупности (который не известен и который надо оценить по данным выборки) равняется нулю. В этом случае $i$-й фактор признаётся несущественным для построения уравнения регрессии.

При проведении исследования может оказаться, что вычисленные значения $t$ для нескольких факторов не превышают $t_{табл}$. В этом случае несущественные факторы из уравнения регрессии исключаются поочерёдно, начиная с наименьшего по абсолютной величине $t$. После исключения фактора, соответствующего минимальному значению $t$, из уравнения регрессии, система нормальных уравнений решается заново. Затем вновь вычисляются значения $t$ для всех оставшихся в уравнении коэффициентов, определяется минимальное значение $t$, которое сравнивается с $t_{табл}$. Если окажется, что $t_{min} < t_{табл}$, то фактор, имеющий $t_{min}$, исключается, и т.д., пока не будет выполняться соотношение $t_{min} \geq t_{табл}$. В этом случае все оставшиеся факторы существенны.

Аналогичный подход, но на последней стадии отбора существенных факторов, осуществляется и при наличии корреляционной зависимости.

Проверка значимости уравнения регрессии проводится с использованием критерия Фишера, таким образом, как, например, в § 6.4.

Практическое закрепление изложенных выше теоретических положений весьма громоздко и трудоёмко и не представляется возможным в рамках данного пособия.

## 6.6. ЗАДАЧИ ДЛЯ САМОСТОЯТЕЛЬНОГО РЕШЕНИЯ

**6.1.** При изучении финансовой деятельности компании в течение некоторого времени был собран статистический материал, содержащий данные (в условных единицах) о ежемесячной прибыли $Z$, расходах на рекламу $X$ и вложении капитала в ценные бумаги $Y$.

| $Z$ | 10 | 12 | 12 | 14 | 16 | 17 | 18 |
|---|---|---|---|---|---|---|---|
| $X$ | 0,2 | 0,5 | 0,3 | 0,5 | 0,5 | 0,6 | 0,8 |
| $Y$ | 0,8 | 0,2 | 1 | 1,2 | 0,9 | 1 | 1,1 |

1. Определить тесноту связи между переменной $Z$ и переменными $X$ и $Y$ с помощью выборочного множественного коэффициента $R$ и определить его значимость на уровне $\alpha = 0{,}05$.

2. Найти коэффициенты множественного уравнения регрессии и оценить погрешность модели линейной регрессии.

**6.2.** Для проверки факторов, влияющих на заработную плату работников, взяты данные по 10 однотипным предприятиям, содержащие сведения: $Z$ – средняя зарплата; $X$ – объём валовой продукции; $Y$ – уровень механизации труда (в баллах).

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $Z$ | 0,1 | 0,2 | 0,25 | 0,3 | 0,4 | 0,5 | 0,6 | 0,7 | 0,8 | 1 |
| $X$ | 1,0 | 0,9 | 1,0 | 1,1 | 1,2 | 0,9 | 1,6 | 1,8 | 1,9 | 2 |
| $Y$ | 2 | 2 | 4 | 5 | 6 | 8 | 7 | 9 | 9 | 10 |

1. Найти выборочный множественный $R_Z$ и частный $R_{ZX}$ коэффициенты корреляции, оценить их значимость и сделать выводы о тесноте связи между переменными $Z$ и $X$, $Y$; $Z$ и $X$.

2. Найти коэффициенты множественного уравнения регрессии и оценить погрешность модели линейной регрессии.

**6.3.** В нижеследующей таблице приведены данные для розничного товарооборота $Z$ (млрд. р.), средней численности населения $X$ (млн. человек) и среднегодового дохода $Y$ (млн. р.) для некоторого региона.

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| $Z$ | 1,2 | 1,3 | 2,5 | 1,4 | 1,2 | 0,2 | 2,4 | 4,1 | 1,1 |
| $X$ | 1,4 | 1,4 | 2,5 | 1,5 | 1,3 | 0,3 | 2,6 | 4,2 | 1,1 |
| $Y$ | 1,3 | 1,3 | 1,4 | 1,8 | 1,5 | 1,6 | 1,8 | 1,9 | 1,6 |

1. Определить тесноту связи между $Z$ и переменными $X$ и $Y$.

2. Найти коэффициенты множественного уравнения регрессии и оценить погрешность модели линейной регрессии.

**6.4.** В нижеследующей таблице указаны парные коэффициенты корреляции:

| | $y$ | $x_1$ | $x_2$ | $x_3$ | $x_4$ |
|---|---|---|---|---|---|
| $y$ | 1 | 0,71 | 0,58 | 0,08 | 0,62 |
| $x_1$ | | 1 | 0,53 | 0,2 | 0,81 |
| $x_2$ | | | 1 | 0,13 | 0,3 |
| $x_3$ | | | | 1 | 0,25 |
| $x_4$ | | | | | 1 |

Провести анализ целесообразности включения заданных факторов в уравнение регрессии.

# ВОПРОСЫ К ЭКЗАМЕНУ

1. Генеральная и выборочная совокупности. Способы образования и выборки.
2. Вариационный ряд. Статистическое распределение выборки. Полигон и гистограмма.
3. Эмпирическая функция распределения и её свойства.
4. Выборочная средняя и выборочная дисперсия, их свойства.
5. Точечные оценки. Требования к оценкам.
6. Метод наибольшего правдоподобия.
7. Точечные оценки параметров нормального распределения.
8. Интервальные оценки. Алгоритм построения доверительного интервала.
9. Статистические гипотезы. Основные понятия. Критерии проверки. Ошибки при проверке гипотез.
10. Схема проверки статистической гипотезы.
11. Критерии согласия.
12. Проверка гипотез о значениях числовых характеристик.
13. Проверка гипотез о равенстве числовых характеристик.
14. Функциональная и корреляционная зависимости. Коэффициент корреляции.
15. Коэффициент корреляции и корреляционное отношение, их свойства и оценка.
16. Уравнение регрессии. Линейная регрессия.
17. Определение параметров уравнений регрессии методом наименьших квадратов.
18. Основная идея дисперсионного анализа.
19. Дисперсионный анализ. Однофакторный комплекс.
20. Дисперсионный анализ. Двухфакторный комплекс.

# ЗАКЛЮЧЕНИЕ

Изложенный в данном учебном пособии материал по математической статистике является минимальным курсом. Поэтому мы надеемся, что изучение более содержательного курса, в большей степени соответствующего количеству часов, выделяемых на математику Государственным образовательным стандартом, позволит обосновать те тезисы, которые декларируются в минимальном курсе.

Углубление знаний при переходе от минимального курса к расширенному демонстрирует, как и в любой науке или деятельности, решая внешнюю задачу, приходится сталкиваться с внутренними техническими проблемами, что приводит к внутреннему развитию и совершенствованию этого знания, и в результате, появляются профессионалы более высокого уровня.

При изложении минимального курса мы сочли необходимым сконцентрировать усилия на формирование особых умений – компетенций. Качество формирования компетенций – весьма тонкое дело, так как балансирует на точке равновесия между теорией и практикой.

«Примеры научат лучше, нежели толкования и книги», – писал Н.И. Лобачевский. С другой стороны, безусловно, прав Гельвеций: «Знание некоторых истин избавляет от необходимости знания многих фактов». Поэтому, на самом деле, хорошая теория и хорошая практика неотделимы.

Мы желаем обучающимся руководствоваться этим условием.

# СПИСОК ЛИТЕРАТУРЫ

1. Кремер, Н.Ш. Теория вероятностей и математическая статистика: учебник для вузов / Н.Ш. Кремер. – 2-е изд., перераб. и доп. – М. : ЮНИТИ-ДАНА, 2004. – 573 с.

2. Гмурман, В.Е. Теория вероятностей и математическая статистика : учеб. пособие для вузов / В.Е. Гмурман. – 11-е изд., стер. – М. : Высш. школа, 2005. – 479 с..

3. Бородин, А.Н. Элементарный курс теории вероятностей и математической статистики : учеб. пособие / А.Н. Бородин. – 7-е изд., стер. – СПб. : Изд-во «Лань», 2008. – 256 с.

4. Гмурман, В.Е. Руководство к решению задач по теории вероятностей и математической статистике : учеб. пособие для студентов вузов / В.Е. Гмурман. – 9-е изд., стер. – М. : Высш. школа, 2004. – 404 с.

5. Емельянов, Г.В. Задачник по теории вероятностей и математической статистике : учеб. пособие / Г.В. Емельянов, В.П. Скитович. – 2-е изд., стер. – СПб. : Изд-во «Лань», 2007. – 336 с.

6. Белько, И.В. Теория вероятнгостей и математическая статистика. Примеры и задачи : учеб. пособие / И.В. Белько, Г.П. Свирид ; под ред. К.К. Кузьмича. – Минск : Новое знание, 2002. – 250 с.

# ОГЛАВЛЕНИЕ


ВВЕДЕНИЕ ..... 3
1. МАТЕМАТИЧЕСКАЯ СТАТИСТИКА. ОСНОВНЫЕ ПОНЯТИЯ ..... 5
1.1. Предмет математической статистики ..... 6
1.2. Генеральная и выборочная совокупности ..... 7
1.3. Вариационный ряд и его графическое изображение ..... 8
1.4. Числовые характеристики вариационных рядов ..... 11
1.5. Задачи для самостоятельного решения ..... 14
2. ОЦЕНКА ПАРАМЕТРОВ ГЕНЕРАЛЬНОЙ СОВОКУПНОСТИ ..... 16
2.1. Точечная оценка ..... 16
2.2. Интервальные оценки ..... 20
2.2.1. Доверительные интервалы для генеральной средней и генеральной доли признака ..... 20
2.2.2. Доверительный интервал для генеральной дисперсии (среднего квадратического отклонения) ..... 22
2.3. Задачи для самостоятельного решения ..... 24
3. СТАТИСТИЧЕСКИЕ ГИПОТЕЗЫ ..... 26
3.1. Основные понятия ..... 26
3.2. Гипотеза о виде распределения ..... 29
3.3. Гипотезы о значениях числовых характеристик ..... 32
3.4. Гипотезы о равенстве числовых характеристик ..... 33
3.4.1. Гипотеза о равенстве средних значений ..... 33
3.4.2. Гипотеза о равенстве дисперсий ..... 35
3.5. Задачи для самостоятельного решения ..... 36
4. ДИСПЕРСИОННЫЙ АНАЛИЗ ..... 38
4.1. Однофакторный анализ ..... 39
4.2. Многофакторный анализ ..... 42
4.3. Задачи для самостоятельного решения ..... 45
5. ЭЛЕМЕНТЫ ТЕОРИИ КОРРЕЛЯЦИИ ..... 47
5.1. Корреляционная зависимость и представление данных в корреляционном анализе ..... 48
5.2. Коэффициент корреляции ..... 49
5.3. Статистическая зависимость. Уравнение регрессии ..... 52
5.4. Статистический анализ уравнения регрессии ..... 53
5.5. Задачи для самостоятельного решения ..... 56
6. МНОГОМЕРНЫЕ СТАТИСТИЧЕСКИЕ МЕТОДЫ ..... 58
6.1. Множественный коэффициент корреляции ..... 59
6.2. Частный коэффициент корреляции ..... 62
6.3. Множественный регрессионный анализ ..... 63
6.4. Нелинейная регрессия ..... 66
6.5. Особенности множественной регрессии и корреляции ..... 71
6.6. Задачи для самостоятельного решения ..... 75
ВОПРОСЫ К ЭКЗАМЕНУ ..... 77
ЗАКЛЮЧЕНИЕ ..... 78
СПИСОК ЛИТЕРАТУРЫ ..... 79