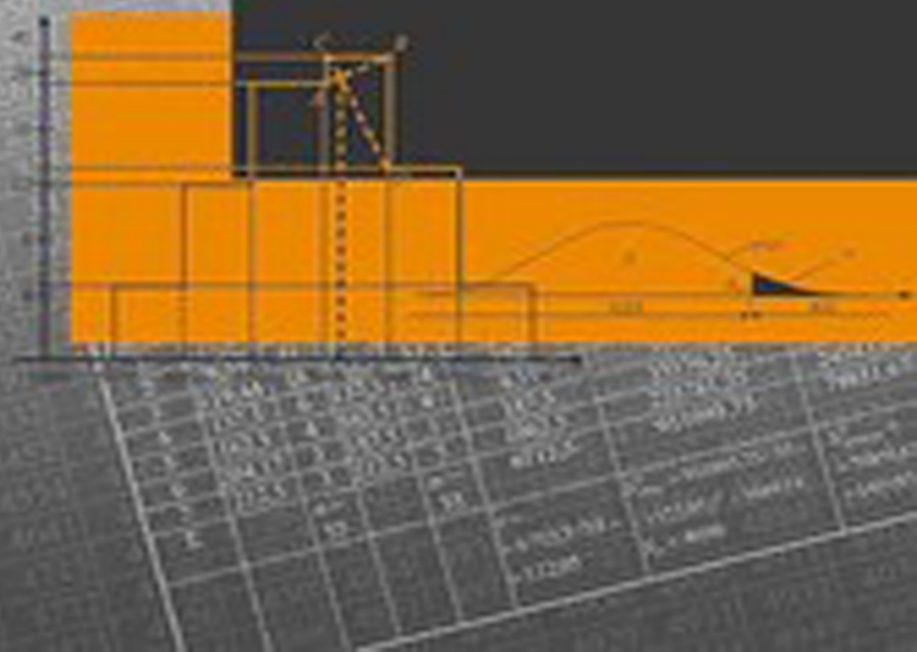


A. B. FOMINOV



ПРИКЛАДНАЯ СТАТИСТИКА



А. В. ГАНИЧЕВА

ПРИКЛАДНАЯ СТАТИСТИКА

Учебное пособие
Издание второе, стереотипное



ЛАНЬ

САНКТ-ПЕТЕРБУРГ
МОСКВА · КРАСНОДАР
2021

УДК 311.3
ББК 60.6я73

Г 19 Ганичева А. В. Прикладная статистика : учебное пособие для вузов / А. В. Ганичева. — 2-е изд., стер. — Санкт-Петербург : Лань, 2021. — 172 с. : ил. — Текст : непосредственный.

ISBN 978-5-8114-8360-0

Пособие представляет собой объединенный курс статистики и математической статистики, в котором полно и доступно, с достаточным количеством прикладных примеров показаны основные формы представления экспериментальных данных, рассмотрены статистические показатели, основные вопросы выборочного обследования, статистического изучения взаимосвязи и динамики социально-экономических явлений, проверки статистических гипотез, характеристики изменения явлений, состоящих из соизмеримых или несоизмеримых элементов.

В конце каждой главы помещены вопросы и задания для закрепления материала и приобретения практических навыков.

В конце учебного пособия имеется приложение вычислительных таблиц.

Предназначено для студентов вузов, обучающихся по направлениям: «Прикладная информатика», «Прикладная математика и информатика», «Информатика и вычислительная техника», «Бизнес-информатика», «Документоведение и архивоведение». Может быть использовано специалистами, аспирантами, научными сотрудниками, занимающимися обработкой статистического материала с целью выработки оптимальных стратегий принятия решений.

УДК 311.3
ББК 60.6я73

Рецензенты:

Е. А. АНДРЕЕВА — доктор физико-математических наук, профессор, зав. кафедрой компьютерной безопасности и математических методов управления Тверского государственного университета;

Ю. Т. ФАРИНЮК — доктор экономических наук, профессор кафедры менеджмента и предпринимательства, руководитель Центра международного сотрудничества Тверской государственной сельскохозяйственной академии.

Обложка
Е. А. ВЛАСОВА

© Издательство «Лань», 2021
© А. В. Ганичева, 2021
© Издательство «Лань»,
художественное оформление, 2021

Оглавление

Предисловие	6
Глава 1. Основные понятия статистики и формы представления статистического материала	8
§ 1. Предмет статистики	8
§ 2. Организация наблюдения и представление статистического материала	10
§ 3. Графическое изображение статистических данных в виде диаграмм	18
Вопросы и задания к главе	25
Глава 2. Статистические показатели	28
§ 1. Абсолютные и относительные величины	28
§ 2. Средние величины. Средняя арифметическая	31
§ 3. Средняя гармоническая и средняя геометрическая	35
§ 4. Структурная средняя	37
§ 5. Показатели вариации	41
Вопросы и задания к главе	49
Глава 3. Выборочное наблюдение	52
§ 1. Выборка и ее формирование	52
§ 2. Оценки характеристик генеральной совокупности	52
§ 3. Метод моментов нахождения оценок	55
§ 4. Предельная ошибка выборки. Интервальное оценивание	57
§ 5. Определение ошибки и объема репрезентативности для генеральной средней по большим выборкам	58
§ 6. Ошибка и объем репрезентативности генеральной средней для малых выборок	63
§ 7. Точность оценки доли признака при больших выборках	65
§ 8. Точность оценки доли признака при малых выборках	67
§ 9. Интервальная оценка генеральной дисперсии нормально распределенного признака	68
Вопросы и задания к главе	71
Глава 4. Статистическое изучение взаимосвязи социально-экономических явлений	73
§ 1. Классификация связей между явлениями и их признаками	73
§ 2. Оценка тесноты связи	75
§ 3. Коэффициент корреляции знаков Фехнера	76
§ 4. Эмпирическое и теоретическое корреляционные отношения	78
§ 5. Линейный коэффициент корреляции	83
§ 6. Коэффициент эластичности	89

§ 7. Коэффициент множественной корреляции	90
§ 8. Частный коэффициент корреляции	92
§ 9. Коэффициенты ассоциации K_a и контингенции K_k	93
§ 10. Коэффициенты Пирсона K_n и Чупрова K_r	95
§ 11. Коэффициент Кендалла	96
§ 12. Коэффициент Спирмена	98
§ 13. Коэффициент конкордации	101
Вопросы и задания к главе	105
Глава 5. Проверка статистических гипотез	109
§ 1. Статистические гипотезы. Критерии значимости	109
§ 2. Проверка статистической гипотезы о равенстве двух дисперсий	112
§ 3. Проверка гипотезы о равенстве дисперсии гипотетическому значению	114
§ 4. Проверка гипотезы о равенстве двух математических ожиданий при известных дисперсиях	115
§ 5. Проверка гипотезы о равенстве математических ожиданий двух совокупностей при неизвестных одинаковых дисперсиях	116
§ 6. Проверка гипотезы о равенстве математического ожидания гипотетическому значению	118
§ 7. Проверка гипотезы о равенстве вероятности появления события предполагаемому значению	120
§ 8. Проверка гипотезы о равенстве коэффициента корреляции нулю	121
§ 9. Проверка гипотезы о значении генерального коэффициента корреляции	124
§ 10. Оценка закона распределения по критерию Пирсона	126
Вопросы и задания к главе	128
Глава 6. Статистическое изучение динамики социально-экономических явлений	132
§ 1. Временные ряды, их классификация и характеристики	132
§ 2. Средние показатели динамики	135
§ 3. Методы анализа основной тенденции развития в рядах динамики	137
§ 4. Линейная однофакторная регрессия. Задача прогноза	139
Вопросы и задания к главе	144
Глава 7. Экономические индексы	147
§ 1. Понятие индекса и их классификация	147
§ 2. Средние арифметические и гармонические индексы	152

§ 3. Индексы средних уровней качественных показателей	154
§ 4. Базисные и цепные индексы	156
Вопросы и задания к главе	157
Приложения	160
Список литературы	169

Предисловие

Термин «**статистика**» был введен немецким ученым Готфридом Ахенвалем в XVII в. и означал государственное ведение. Ближе к современному пониманию статистики были представители английской школы политических арифметиков, среди которых можно отметить Дж. Граунта, К. Петти, Г. Кинга, Э. Галлея. В XIX в. статистика стала экономической наукой. Большую роль здесь сыграли ученые: А. Кетле, К. Пирсон, В. Бессет, Р. Фамер и др. В настоящее время термин «статистика» употребляется в различных значениях:

- 1) так называется практическая деятельность по сбору, накоплению, обработке и анализу массовых цифровых данных о самых различных явлениях жизни общества;
- 2) совокупность цифровых данных, характеризующих состояние массовых явлений и процессов общественной жизни;
- 3) комплекс научных дисциплин, т. е. отрасль знаний, изучающая явления в жизни общества с их количественной стороны.

Статистика тесно связана с математической статистикой, в которой изучаются математические методы систематизации, обработки и использования статистических данных для научных и практических выводов. Математическая статистика возникла тоже в XVII в. и создавалась параллельно с теорией вероятностей.

Большой вклад в развитие этих наук внесли ученые: В. Я. Буняковский (использовавший методы матстатистики к демографии и страховому делу в середине XIX в.), В. И. Романовский, А. Н. Колмогоров, Е. Е. Слуцкий, Н. В. Смирнов, Ю. В. Линник, Стиюдент, Р. Фишер, Э. Пирсон, Ю. Нейман, П. Л. Чебышев, А. А. Марков, А. М. Ляпунов, С. Н. Бернштейн, К. Пирсон, Ф. Гальтон и др.

Предлагаемое учебное пособие представляет собой объединенный курс статистики и математической статистики, состоит из семи глав.

Пособие посвящено:

- вопросам организации, представления и графического изображения статистических данных;
- точечной и интервальной оценке статистических показателей, изучению взаимосвязи социально-экономических явлений с использованием различных коэффициентов связи;
- проверке статистических гипотез;
- изучению динамики социально-экономических явлений, включая задачу прогноза и экономические индексы.

Пособие проблемно ориентировано. Содержит достаточное количество задач прикладного характера, ориентирующих обучаемых на будущую профессиональную деятельность.

К каждой главе приводится список вопросов и соответствующих заданий. Приложение содержит основные расчетные таблицы.

Авторы выражают благодарность рецензентам: Андреевой Елене Аркадьевне и Фаринюку Юрию Теодоровичу.

Глава 1. Основные понятия статистики и формы представления статистического материала

§ 1. Предмет статистики

Объектом изучения статистики является общество, протекающие в нем процессы и существующие закономерности. Предметом статистики являются размеры и количественные соотношения качественно определенных социально-экономических явлений, закономерности их связи и развития в конкретных условиях места и времени. **Характерные особенности** статистики:

- изучаются массовые общественные явления, к которым относятся численность населения, количество произведенной продукции и т. д., и их динамика: например, изменение уровня жизни населения;
- количественная сторона массовых явлений рассматривается в неразрывной связи с их качественным содержанием;
- количественная сторона общественных явлений изучается в конкретных условиях места и времени.

Метод статистики: сбор, обработка и анализ экспериментальных данных в массовых общественных явлениях. На первой стадии происходит сбор первичной статистической информации, на второй – статистическая сводка и обработка первичной информации, на третьей – обобщение и интерпретация статистической информации.

К **основным категориям** статистики относятся:

1. **Статистическая совокупность** – это множество единиц (объектов, явлений), объединенных единой закономерностью и варьируемых в пределах общего качества. Примеры: список работников согласно занимаемой должности, список депутатов по партийной принадлежности, список сотрудников с указанием наличия (отсутствия) у них больничного листа в данном месяце; список продавцов по категориям; список студентов, изучающих статистику, имеющих данный средний балл; совокупность продовольственных магазинов данного района, изучаемая с точки зрения получаемой прибыли, оклады труда, текучести кадров, качества обслуживания и т. п.

2. **Единица совокупности** – первичный элемент, выражающий качественную однородность совокупности и являющийся носителем признака. Например, магазины, банки, заводы, учебные заведения, рабочие, студенты – это единицы соответствующих объектовых совокупностей. Единицами числовых последовательностей являются составляющие их числа. Так, если в городе Энске цена 1 кв. м. муницип-

ципального жилья в Центральном, Западном, Отрадном, Пролетарском, Сахаровском и Северном районах в 1998 г. составила соответственно (в усл. ед.):

5,2; 4,9; 4,4; 4,9; 4,4; 4,7,

то эти шесть чисел являются единицами указанной совокупности.

3. **Объемом** статистической совокупности называется число единиц в ней. Единицы статсовокупности характеризуются общими свойствами – признаками.

4. **Варианта** – значение варьируемого признака статистической совокупности.

5. **Признак** – показатель, характеризующий некоторое свойство объекта совокупности, рассматриваемый как **случайная величина**. Например, единица статистической совокупности – рабочий – имеет следующие признаки: возраст, трудовой стаж, величину заработной платы и т. д. Значения каждого признака отдельной единицы совокупности могут быть различными: стаж работы может быть равен 1-му году, 2-м годам и т. д. Признаки подразделяются на **количественные**, если их варианты выражаются числовыми значениями (объем производства, издержки, численность работников и т. п.), и **атрибутивными**, не имеющими числового выражения (пол, категории работников и т. д.). Разновидностью качественных признаков являются **альтернативные**, когда единица обладает или не обладает изучаемым признаком. Количественные признаки могут быть дискретными и непрерывными, как дискретные и непрерывные случайные величины. Множество значений случайной величины **образует генеральную совокупность**. Признаки подразделяются также на **факторные** (независимые) и **результативные**. Например, объем производственных фондов – факторный признак, а производительность труда – результативный.

6. **Частотой (весом)** варианты (группы вариантов) называется численность варианты (группы вариант) в статистической совокупности. Частоты, выраженные в долях единицы или в процентах, называются **частотами (относительными частотами)**. Например, если статсовокупность, состоящая из 9 единиц:

5, 10, 6, 12, 10, 5, 6, 6, 7,

соответствует выработке 9 рабочих, то вариантами здесь являются числа 5, 6, 7, 10, 12, которые в данную совокупность входят с частотами 2, 3, 1, 2 и 1 соответственно, т. е. 5 встречается в этой последовательности 2 раза, 6 – 3 раза, 7 – 1 раз, 10 – 2 и 12 – 1 раз. Поделив

каждую частоту на их сумму, равную $2 + 3 + 1 + 2 + 1 = 9$, получим соответствующие частоты:

$$\frac{2}{9}, \frac{3}{9} = \frac{1}{3}, \frac{1}{9}, \frac{2}{9}, \frac{1}{9},$$

сумма которых равна 1. Частоты можно записать в процентах:

$$\frac{2}{9} \cdot 100\%, \frac{1}{3} \cdot 100\%, \frac{1}{9} \cdot 100\%, \frac{2}{9} \cdot 100\%, \frac{1}{9} \cdot 100\%.$$

При суммировании получим 100%.

Если рассматривается статсовокупность средних баллов студентов данного института в данной сессии, то можно сгруппировать студентов по среднему баллу и говорить о частоте данных групп.

7. **Вариация** – различия в значениях того или иного признака у отдельных единиц, входящих в данную совокупность.

8. **Статистический показатель** – количественно-качественная характеристика какого-то свойства группы единиц или совокупности в целом. Примерами статистических показателей являются: численность рабочих данного предприятия на 1 января 2014 г., средняя заработная плата, доля продукции высокого качества среди выпускаемой продукции данного предприятия. Это **количественные** показатели. Можно говорить о высокой, низкой, средней производительности труда без использования числовых данных. Такие показатели называются **качественными**.

§ 2. Организация наблюдения и представление статистического материала

Формы наблюдения:

1. **Отчетность** представляет собой предусмотренные действующим законодательством формы организации статистического наблюдения.

2. **Специально организованное статистическое наблюдение** сводится к сбору сведений посредством переписей, единовременных учетов и обследований.

Виды статистического наблюдения:

- по времени регистрации фактов – непрерывное, периодическое и единовременное;
- по степени охвата – сплошное и несплошное.

При сплошном наблюдении регистрации подлежат все без исключения единицы совокупности. При несплошном наблюдении обследованию подвергается лишь часть единиц совокупности.

Несплошное наблюдение подразделяется на: наблюдение основного массива, монографическое наблюдение (детально обследуются отдельные единицы) и выборочное наблюдение (детально обследуется ряд представителей данной совокупности).

К первичным формам представления статистического материала относятся сводка и группировка.

Статистическая сводка – это научно организованная обработка материалов наблюдения, которая заключается в систематизации и группировке данных, составлении таблиц, расчете статистических показателей. Сводка осуществляется по этапам:

1. Выбор группировочных признаков.
2. Определение порядка формирования групп.
3. Разработка системы статистических показателей.
4. Разработка статистических таблиц.

На 1-м и 2-м этапах используется **статистическая группировка**, представляющая собой процесс образования однородных групп на основе либо расчленения, либо объединения отдельных единиц в группы по существенным для них признакам. Например, группировка обучаемых согласно среднему баллу успеваемости; группировка избирателей по возрастному признаку.

Виды группировок:

1. Типологическая группировка, примером которой является группировка вкладчиков данного банка по типам вкладов. В общем случае выявляют различные типы социально-экономических явлений по качественным признакам. Число групп определяется числом значений качественного признака. Такие группировки называют также **атрибутивными рядами**.

2. Аналитическая (факторная) группировка позволяет определить зависимость между отдельными признаками. Например, группировка предприятий по производственной мощности, которая дает возможность проанализировать зависимость объема производства, производительность труда от производственной мощности.

Аналитические группировки строятся по факторному количественному признаку. Затем рассчитываются средние значения и дисперсии результативного признака в каждой группе, а также общая средняя, средняя групповых дисперсий, межгрупповая и общая дисперсии, и по ним делаются выводы о наличии или отсутствии зависимости (связи) между факторным и результативным признаками. Подробнее можно посмотреть в § 5 главы 2.

3. Структурная группировка характеризует удельный вес (долю) отдельных групп в общем объеме совокупности. Например, группировка хозяйств по реализации продукции, производительности труда и т. д.

В структурных группировках определяется количество групп и интервалы группировки. **Интервал** – это количество единиц в группе. В случае равных интервалов оптимальное количество групп k определяется по **формуле Стерджесса**:

$$k = 1 + 3,322 \cdot \lg n = 1 + 1,41 \lg n, \quad (1)$$

где n – численность единиц совокупности.

Величина (длина) интервала Δ вычисляется по формуле

$$\Delta = \frac{x_{\max} - x_{\min}}{k}, \quad (2)$$

где x_{\max} и x_{\min} – соответственно наибольшее и наименьшее значения признака, k – число групп.

Если имеется неравномерное распределение признака, то составляются группировки с неравными интервалами, причем в каждый интервал должно попасть достаточное количество единиц совокупности с учетом их возможной повторяемости. Например, для совокупности значений непрерывного признака (непрерывной случайной величины)

2, 4, 6, 2, 3, 5, 7, 3, 4, 2, 5, 6

можно рассмотреть группировку

(2, 2, 2, 3, 3) (4, 4, 5, 5, 6, 6, 7),

при которой исходная совокупность разбита на 2 группы, содержащие 5 и 7 элементов соответственно.

На основе структурных группировок строятся **вариационные ряды**, в задании которых участвуют варианты (отдельные значения изучаемого признака) и их частоты. Вариационные ряды подразделяются на а) **дискретные**, когда вариантам ставятся в соответствие их частоты, и б) **интервальные**, при которых варианты группируются в интервалы и каждому интервалу ставится в соответствие его частота, т. е. сумма частот входящих в него вариант; при этом элемент, попавший на границу двух интервалов либо относится к одному из них, либо к числу элементов каждого прибавляется по 0,5.

Интервальные ряды строятся, как правило, для непрерывных случайных величин (непрерывных признаков), когда рассматриваются большие совокупности.

Вариационные ряды можно представить в виде таблиц или изобразить графически в виде полигона, кумуляты и гистограммы. Рассмотрим это на конкретном примере.

Пример 1.

Производительность труда 20 произвольно взятых работников данного предприятия оказалась равной (в условных единицах)

14, 18, 22, 24, 18, 22, 26, 16, 16, 17,
20, 29, 22, 28, 20, 28, 14, 18, 19, 18.

Построить дискретный и интервальный ряд.

Решение.

Располагая исходные данные в порядке возрастания (убывания), получим дискретный вариационный ряд:

14, 14, 16, 16, 17, 18, 18, 18, 18, 19,
20, 20, 22, 22, 22, 24, 26, 28, 28, 29,

содержащий 11 вариантов: 14, 16, 17, 18, 19, 20, 22, 24, 26, 28, 29, которые входят в данный ряд соответственно с частотами: 2, 2, 1, 4, 1, 2, 3, 1, 1, 2, 1.

С учетом повторяемости всего будет $n = 20$ элементов.

Поделив каждую частоту на их общую сумму, равную 20, получим соответствующие частоты:

$$\frac{2}{20} = 0,1; \quad \frac{2}{20} = 0,1; \quad \frac{1}{20} = 0,05; \quad \frac{4}{20} = 0,2; \quad \frac{1}{20} = 0,05; \quad \frac{2}{20} = 0,1;$$

$$\frac{3}{20} = 0,15; \quad \frac{1}{20} = 0,05; \quad \frac{1}{20} = 0,05; \quad \frac{2}{20} = 0,1; \quad \frac{1}{20} = 0,05.$$

Сумма частостей равна 1 (проверьте!).

Полученные данные можно оформить в виде таблицы, указав в первой строке варианты x_i , во второй – их частоты n_i , а в третьей – частоты $w_i = \frac{n_i}{n}$ (табл. 1).

Таблица 1

Варианты x_i	14	16	17	18	19	20	22	24	26	28	29
Частоты n_i	2	2	1	4	1	2	3	1	1	2	1
Частоты w_i	0,1	0,1	0,05	0,2	0,05	0,1	0,15	0,05	0,05	0,1	0,05

Полученный вариационный ряд можно представить графически в виде **полигона**. Для этого по горизонтальной оси откладываются значения вариантов x_i , а по вертикальной – либо частоты n_i , либо частоты w_i . Полученные точки соединяются отрезками (рис. 1).

На всех рисунках данного параграфа, чтобы избежать громоздкости изображения, начало координат помещено несколько правее точки 0 (без соблюдения масштаба).

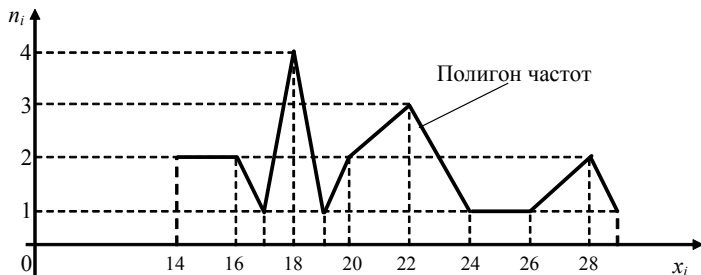


Рис. 1

Если сложить последовательно частоты, то получатся **накопленные частоты**:

$$2, 2 + 2 = 4, 4 + 1 = 5, 5 + 4 = 9, 9 + 1 = 10, 10 + 2 = 12, \\ 12 + 3 = 15, 15 + 1 = 16, 16 + 1 = 17, 17 + 2 = 19, 19 + 1 = 20,$$

которые графически представляются в виде **кумуляты** (рис. 2).

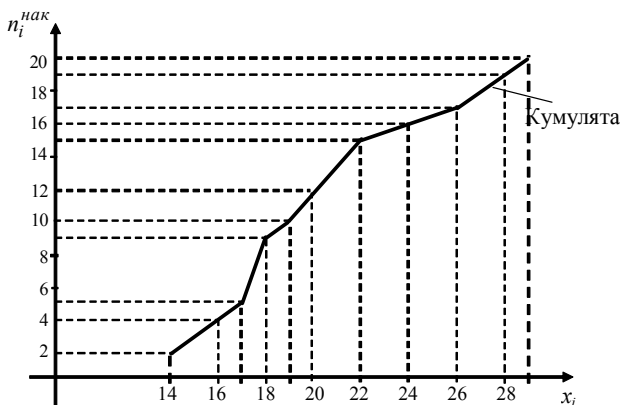


Рис. 2

Аналогичная кумулята получается при последовательном сложении частостей.

Построим для рассматриваемого примера интервальный ряд. Имеем: $n = 20$, число интервалов $k = 1 + 3,222 \lg 20 \approx 5$, $x_{\max} = 29$, $x_{\min} = 14$,

$$\Delta = \frac{29-14}{5} = 3; \text{ и интервальный ряд будет выглядеть так (табл. 2):}$$

Таблица 2

Интервалы	(14, 17)	(17, 20)	(20, 23)	(23, 26)	(26, 29)
Частоты n_i	5	7	3	2	3

Складывая почленно частоты n_i , находим накопленные частоты $n_i^{нак}$:

$$5, 5 + 7 = 12, 12 + 3 = 15, 15 + 2 = 17, 17 + 3 = 20.$$

Полигон для интервального ряда строится так же, как и для дискретного, с той лишь разницей, что соответствующие частоты (частости) откладываются (в выбранном масштабе) на перпендикулярах, проведенных из середин интервалов, отложенных на горизонтальной оси.

При построении кумуляты интервального ряда накопленные частоты (в выбранном масштабе) откладываются на перпендикулярах, проведенных из правых концов, и полученные точки соединяются.

Гистограмма определяется только для интервального ряда и строится следующим образом. На каждом интервале, как на основании, строится прямоугольник с высотой, равной (в выбранном масштабе) частоте данного интервала. Полученная ступенчатая фигура и представляет собой гистограмму. По оси ординат могут откладываться также относительные частоты (частости), а также относительные частоты, деленные на длину интервала.

Если соединить середины верхних оснований прямоугольников отрезками прямой, а середины крайних верхних оснований соединить соответственно с серединой интервала, предшествующего самому левому, и с серединой интервала, следующего за самым правым, то получится полигон частот. На рис. 5 он показан тонкой сплошной линией.

Для примера 1 полученные результаты оформлены графически в виде **полигона, кумуляты и гистограммы** (рис. 3, 4, 5).

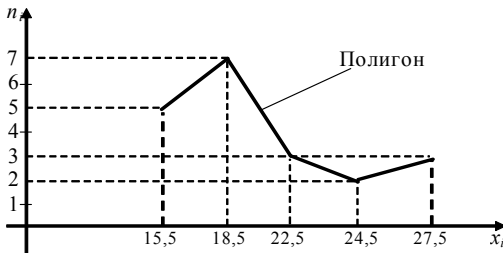


Рис. 3

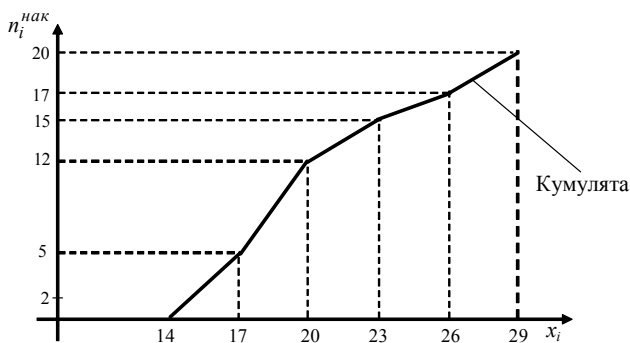


Рис. 4

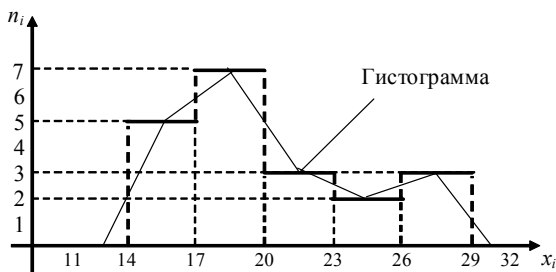


Рис. 5

Для характеристики признака используются также относительные частоты – частоты интервального ряда. Для рассмотренного примера имеем последовательность частот:

$$\frac{5}{20} = 0,25; \quad \frac{7}{20} = 0,35; \quad \frac{3}{20} = 0,15; \quad \frac{2}{20} = 0,1; \quad \frac{3}{20} = 0,15.$$

Сумма равна 1.

Аналогично строятся полигон, кумулята и гистограмма для относительных частот интервального ряда с заменой частоты относительной частотой. Читателям предлагается сделать это самостоятельно.

Отметим, что вариационный ряд является статистическим аналогом ряда распределения случайной величины (признака), но при этом имеет смысл как для дискретного, так и для непрерывного признака (случайной величины). Обратим внимание на то, что сумма площадей прямоугольников, составляющих гистограмму, равна 1, если высоты h_i прямоугольников вычисляются по формуле $h_i = w_i / \Delta$, где w_i – относительная частота i -го интервала, а Δ – его длина. Соответ-

ствующая гистограмма является статистическим аналогом плотности распределения непрерывной случайной величины. Этот факт можно использовать для приближенного вычисления вероятности того, что значение данного признака (непрерывной случайной величины) X заключается в границах от α до β . Эта вероятность приближенно равна заштрихованной площади, которая показана на рис. 6 для данных примера 1 о производительности труда при

$$h_1 = 0,25/3 \approx 0,08; \quad h_2 = 0,35/3 \approx 0,117; \quad h_3 = 0,15/3 \approx 0,05;$$

$$h_4 = 0,1/3 \approx 0,03; \quad h_5 = 0,15/3 \approx 0,05; \quad \alpha = 15,5; \quad \beta = 24,5.$$

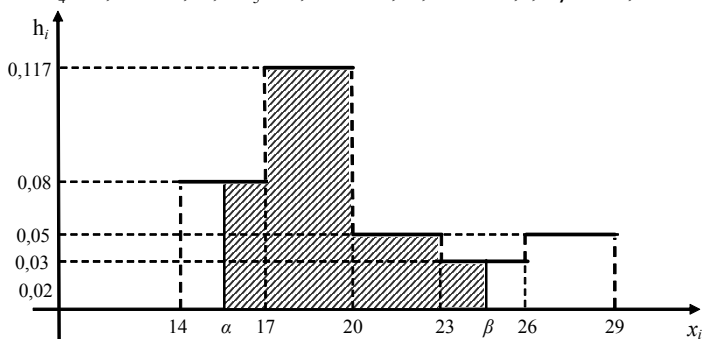


Рис. 6

Таким образом, вероятность того, что производительность труда на данном предприятии будет в границах от 15,5 до 24,5 условных единиц, численно равна (приближенно) сумме площадей заштрихованных прямоугольников:

$$1,5 \cdot 0,08 + 3 \cdot 0,117 + 1,5 \cdot 0,03 \approx 0,666.$$

Кумулята относительных частот (частостей) представляет собой статистический аналог функции распределения $F(x)$ случайной величины (признака) X .

Полигон, гистограмма и кумулята графически выражают статистическую зависимость между значениями случайной величины (признака) и соответствующими вероятностям; позволяют наглядно аппроксимировать (приближать) одно распределение другим, вычислять вероятности нахождения признаков в данных границах, выдвигать гипотезы о виде распределения случайной величины (признака), вычислять некоторые статистические показатели. Эти вопросы будут рассмотрены в следующих параграфах.

Итак, мы рассмотрели представление статистического материала в виде вариационного ряда – дискретного и интервального. Отметим,

что в ряде задач интервальный ряд может быть задан с неопределенными крайними границами, кроме того, интервалы и частоты можно, очевидно, располагать не в строках, а в столбцах, и рассматривать объединенную таблицу частот, накопленных частот, частостей и накопленных частостей (следующий пример).

Пример 2.

Распределение цены светильников в магазине представлено таблицей (1-й и 3-й столбцы табл. 3). Уточнить границы, найти частости, а также накопленные частоты и частости.

Таблица 3

Цена (руб.)	Уточненные границы ин- тервалов	Часто- та n_i	Накоплен- ная частота $n_i^{нак}$	Частости $w_i = \frac{n_i}{n}$	Накоп- ленные частости $w_i^{нак}$
до 100	от 50 до 100	5	5	$5/80=0,0625$	0,0625
от 100 до 150	от 100 до 150	10	15	$10/80=0,125$	0,1875
от 150 до 200	от 150 до 200	18	33	$18/80=0,225$	0,4125
от 200 до 250	от 200 до 250	20	53	$20/80=0,250$	0,6625
от 250 до 300	от 250 до 300	16	69	$16/80=0,2$	0,8625
от 300 до 350	от 300 до 350	8	77	$8/80=0,1$	0,9625
свыше 350	от 350 до 400	3	80	$3/80=0,0375$	1
Итого:		80		1	

Решение.

Уточненные границы указаны во 2-м столбце. Имеем интервальный ряд, причем длина каждого интервала, начиная со 2-го, и кроме последнего, равна 50 (руб.). Считаем длины 1-го и последнего интервалов также равными 50, тогда 1-й интервал будет в границах от 50 до 100, а последний – в границах от 350 до 400.

В 4-м, 5-м и 6-м столбцах записаны соответственно накопленные частоты, частости и накопленные частости.

§ 3. Графическое изображение статистических данных в виде диаграмм

Для наглядного представления результатов наблюдений удобно использовать различные **диаграммы**, на которых схематично в определенном масштабе изображается изучаемый статистический материал.

При масштабировании шкалы могут быть равномерными (обычные шкалы) и неравномерными, например логарифмическая, на кото-

рой отрезки пропорциональны логарифмам изображаемых величин. Обычно используется декартова прямоугольная система координат.

В **столбиковых** диаграммах статистические данные изображаются в виде вытянутых по вертикали прямоугольников с одинаковыми основаниями, размещенными на одинаковом расстоянии друг от друга, либо вплотную один к другому, либо наплывом, когда один столбец частично накладывается на другой.

На рис. 7 показано распределение численности студентов данного вуза по годам.

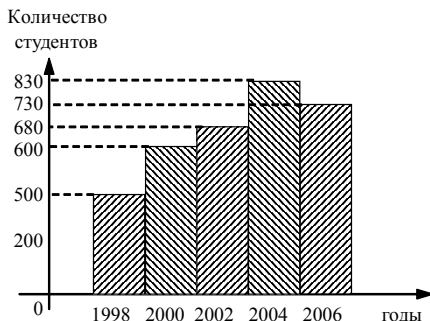


Рис. 7

При построении **квадратных** диаграмм необходимо из сравниваемых статистических величин извлечь квадратные корни, а затем построить квадраты со сторонами, пропорциональными полученным результатам.

Пример 3.

Построить квадратную диаграмму для сравнения стоимости 1 кг сахарного песка, макарон и докторской колбасы, если 1 кг сахарного песка (условно) стоит 25 руб., макарон – 14 руб., колбасы – 128 руб.

Решение.

$$\sqrt{25} = 5 \text{ (руб.)}; \sqrt{14} \approx 3,74 \text{ (руб.)}; \sqrt{128} \approx 11,31 \text{ (руб.)}.$$

Соответствующие диаграммы показаны на рис. 8.

Полосовые (ленточные) диаграммы состоят из прямоугольников, расположенных горизонтально (полосами, лентами). Масштабной шкалой является горизонтальная ось. Полосовые диаграммы находят широкое применение в статистике коммерческой деятельности.

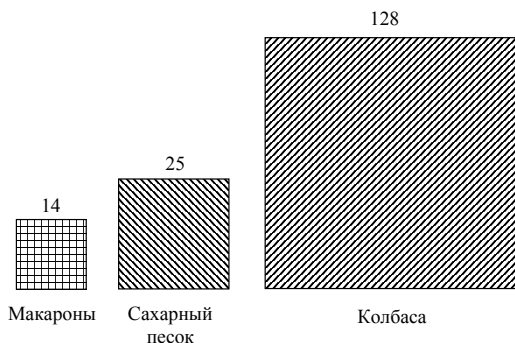


Рис. 8

Принцип построения тот же, что и столбиковых диаграмм. Так, на рис. 9 показано применение этих диаграмм для наглядного представления данных относительно уровня охвата Интернетом населения стран [9].

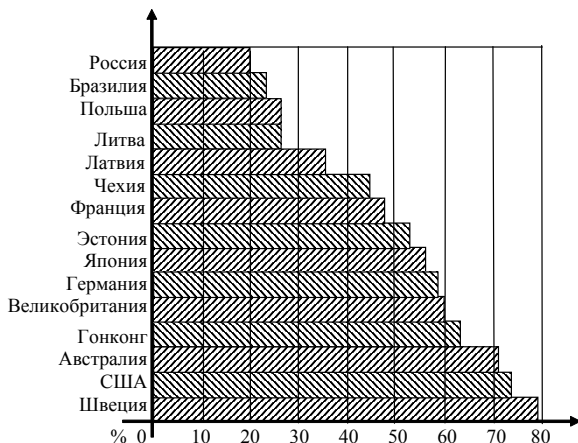


Рис. 9

Круговые диаграммы строятся аналогично, с той лишь разницей, что строятся круги, площади которых пропорциональны квадратным корням значений рассматриваемых величин.

Диаграммы **фигур-знаков** соответствуют изображению исходных данных в виде рисунков, силуэтов, фигур, количество которых пропорционально исходным данным. Например, на рис. 10 показано

возможное распределение 10 рабочих, закодированных номерами от 1 до 10, на 10 работ, изображенных кружочками с указанием номера работы.

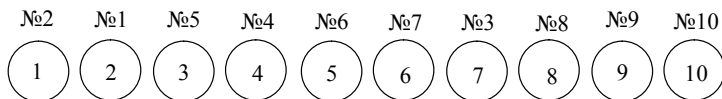


Рис. 10

Другой пример. Если речь идет о динамике производства холодильников в данном регионе и известно, что в 1995 г. их число составило 5780 штук, в 1997-м – 4950 штук, а в 1999-м – 3620, то, приняв условно один рисунок (а это может быть, например, круг, прямоугольник и т. п.) за 1000 холодильников, будем иметь следующее диаграммное представление (рис. 11).

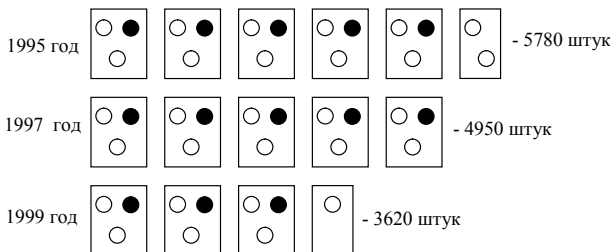


Рис. 11

Если данные о структуре какого-либо явления выражаются в абсолютных величинах, то для нахождения секторов сначала 360° делится на величину целого (значения явления), а затем полученное частное последовательно умножается на абсолютные значения частей.

В случае если сопоставляются три величины, одна из которых является произведением двух других, применяются диаграммы под названием «**знак Варзара**», представляющие собой прямоугольники, у которых один сомножитель принят за основание, другой – за высоту, а вся площадь равна произведению. Например, если x и y – произвольные блага, то $U_0 = x \cdot y$ – уровень полезности благ, геометрически изображаемый площадью прямоугольника со сторонами, численно равными x и y , представляет собой знак Варзара. Если x или y суть постоянная величина, то знак Варзара изображает **прямо пропорцио-**

нальную зависимость. При постоянной площади знак Варзара изображает **обратно пропорциональную** связь.

Линейные диаграммы находят широкое применение для характеристики изменений явлений во времени для изучения рядов распределений и выявления связи между явлениями. Эти диаграммы представляют собой ломаные линии на координатной плоскости. Например, при помощи линейных диаграмм можно показать динамику роста клиентской базы депозитария Росбанка (рис. 12) согласно данным табл. 4.

Таблица 4

2-е полуго- дие 2001 г.	1-е полуго- дие 2002 г.	2-е полуго- дие 2002 г.	1-е полуго- дие 2003 г.	2-е полуго- дие 2003 г.	1-е полуго- дие 2004 г.	2-е полуго- дие 2004 г.	1-е полуго- дие 2005 г.
622	711	889	978	1022	1156	1733	1777

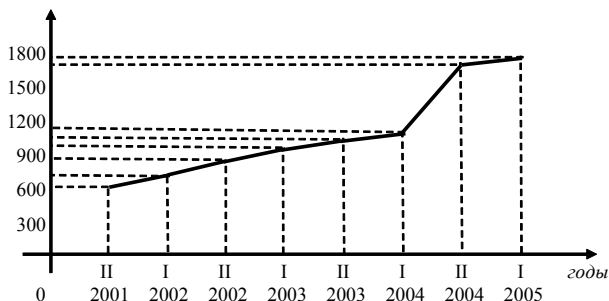


Рис. 12

На одной линейной диаграмме могут быть представлены несколько ломаных, которые дают сравнительную характеристику динамики различных явлений (показателей) или одного итого же явления (показателя) при разных условиях.

Ряды распределения чаще всего изображаются в виде **полигона**, гистограммы или кумуляты (см. предыдущий параграф).

Разновидностью линейных диаграмм являются **радиальные** диаграммы, которые применяются для изображения рядов динамики при наличии в них сезонных колебаний (глава 6).

Пример 4.

Имеются данные о численности малых предприятий Тверской области, табл. 5.

Таблица 5

Годы	1998	1999	2000	2001	2002	2003
Число малых предприятий на 1 января, тыс.	3,6	3,4	5,7	5,7	7,4	7,0

Отразить на радиальной диаграмме динамику числа малых предприятий в 1998–2003 гг.

Решение.

Сначала определяется среднее число предприятий, как их средняя арифметическая (§ 2 главы 2), т. е. все значения складываются: $3,6 + 3,4 + 5,7 + 5,7 + 7,4 + 7,0 = 32,8$ и сумма делится на число слагаемых: $32,8 : 6 = 5,47$. Далее чертим круг с радиусом R , равным полученному среднему показателю 5,47, и делим круг на число рассматриваемых промежутков – для нашего примера – на 6 (рис. 13). Получаем следующие значения показателя по годам (табл. 6).

Таблица 6

Годы	1998	1999	2000	2001	2002	2003
Число малых предприятий на 1 января, тыс.	$3,6/5,47 = 0,66$	$3,4/5,47 = 0,62$	$5,7/5,47 = 1,04$	$5,7/5,47 = 1,04$	$7,4/5,47 = 1,35$	$7/5,47 = 1,28$

Затем на радиусах последовательно в выбранном масштабе откладываются значения рассматриваемого показателя (в нашем примере – число малых предприятий).

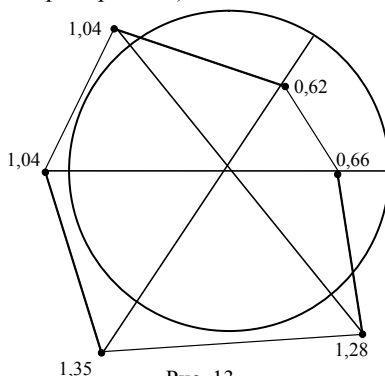


Рис. 13

При этом если значения показателя будут больше радиуса, то они отмечаются за пределами окружности на продолжении соответ-

ствующего радиуса. Полученные точки последовательно соединяются отрезками.

Для изображения циклического измерения строятся линейные графики в полярной системе координат, называемые **радиальными диаграммами**, в которых радиусы обозначают периоды времени, а окружность – величину изучаемого явления.

При **секторных** диаграммах вся величина явления принимается за 100% и изображается кругом, тогда одному проценту будет соответствовать сектор в 3,6 градусов, а любая часть явления, равная проценту a , будет изображаться сектором в $3,6a$ градусов.

Так, по данным Госкомстата, доли иностранного капитала по отраслям экономики в 2000 г. составляли:

- промышленность – 43,1%;
- строительство – 0,9%;
- сельское хозяйство – 0,5%;
- транспорт – 9,3%;
- связь – 8,5%;
- торговля и общественное питание – 17,8%;
- продукция производственно-технического назначения – 1,2%;
- коммерческая деятельность – 2,5%;
- страхование и пенсионное обеспечение – 2,5%;
- прочие инвестиции – 13,7%.

Эти данные графически представлены на рис. 14.

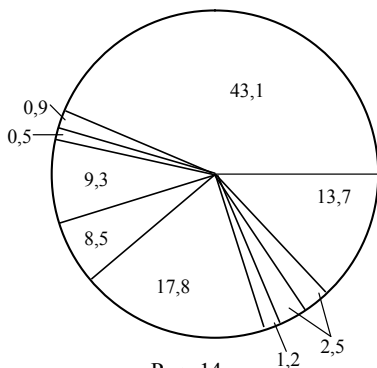


Рис. 14

Картограммы подразделяются на фоновые и точечные. Так, при помощи **фоновой** диаграммы можно изобразить количество вы-

павших осадков по регионам. При построении **точечной диаграммы** строятся точки, размещенные в пределах территориальных единиц.

При построении **картограмм** составные части диаграммы размещаются на контурной карте на площади, отведенной данному подразделению страны или стране. Например, при построении картограммы производства шифера по областям РФ за 2000 г. с использованием столбиковой диаграммы надо столбик, высота которого отражает объем производства шифера в данной области, разместить на том месте, которое отведено для нее на карте, при этом в выбранном масштабе столбики не должны выходить за пределы отведенных областей.

Вопросы и задания к главе

1. Что означает термин «статистика»?
2. Кто из ученых внес большой вклад в развитие статистики?
3. К какому времени относится становление статистики как науки?
4. Какие значения имеет термин «статистика»?
5. Что является объектом изучения статистики?
6. Что представляет собой метод статистики?
7. Дайте определение статистической совокупности и ее единицы, приведите примеры.
8. Что называется вариантой? Как связаны понятия «единица» и «варианта»?
9. Дать определение признака, привести классификацию признаков с примерами.
10. Что называется частотой и частостью (относительной частотой)?
11. Как определяется вариация?
12. Дать определение статистического показателя, привести примеры.
13. Перечислить формы наблюдения.
14. Назвать виды статистического наблюдения.
15. Охарактеризовать сводку и группировку.
16. Перечислить виды группировок, привести примеры.
17. Дать определение интервала и его длины.
18. Какие вариационные ряды используются в статистике?
19. Определить полигон, кумуляту дискретного и интервального ряда.
20. Дать определение гистограммы.
21. Статистическими аналогами каких форм закона распределения являются вариационный ряд, полигон, кумулята и гистограмма?
22. Какие виды диаграмм вы знаете? Укажите условия их применения.
23. В каких случаях знак Варзара изображает: а) прямо пропорциональную зависимость? б) обратно пропорциональную зависимость?

24. При помощи квадратной и радиальной диаграмм сопоставьте данные о жилищном фонде (в данном регионе в млн м² общей площади):

2000	2005	2010	2015
1291,6	1432,1	1533,4	1725,8

25. Имеются данные о посевной площади, валовом сборе и урожайности в целом, а также отдельных зерновых культур в данном регионе (цифры условные) (табл. 7, 8):

Таблица 7

Годы	Валовый сбор, млн т				Урожайность, ц с 1 га			
	Общий	Пшеница	Ячмень	Овес	Общая	Пшеница	Ячмень	Овес
2014	92,3	18,4	21,3	11,5	15,2	11,9	14,8	12,7
2016	93,1	17,2	23,5	10,6	15,8	10,5	16,3	12,7

Таблица 8

Годы	Посевная площадь, млн га			
	Общая	Пшеница	Ячмень	Овес
2014	100,7	14,5	13,1	9,2
2016	100,1	14,2	14,7	8,4

Изобразить эти данные при помощи диаграмм:

а) квадратных; б) круговых; в) столбиковых; г) знаков Варзара; д) секторных.

26. По данной таблице частот (табл. 9) построить: а) полигон частот; б) кумуляту частот и относительных частот (частостей); в) эмпирическую интегральную функцию распределения.

Таблица 9

x_i	4	6	11	12	15	16	17	21
n_i	2	1	6	7	3	3	4	5

27. Имеются условные данные о величине дохода на душу населения в данном регионе по годам:

2010	2011	2012	2013	2014	2015	2016
50	55	60	75	80	90	100

Изобразите данный статистический материал в виде столбиковых и линейных диаграмм.

28. Статистические данные относительно рассматриваемого экономического показателя, заданного в усл. ед., приведены в табл. 10.

Таблица 10

8	22	-36	-10	88	7	22	14	48	56
5	180	19	-17	27	43	3	40	18	52
63	70	93	110	-14	-42	95	-11	-29	13
-100	25	-42	43	49	19	-18	30	81	58
53	-80	-27	-30	3	82	75	32	-19	35

Построить: а) кумуляту частот и относительных частот; б) эмпирическую интегральную функцию распределения; в) гистограмму.

29. Допустим, еженедельная прибыль предприятия фиксировалась в течение года (табл. 11):

Таблица 11

Прибыль (в усл. ед.)	10–15	15–20	20–25	25–30	30–35	35–40
n_i (частота)	10	15	17	16	10	9

Построить: а) кумуляту частот и относительных частот; б) эмпирическую интегральную функцию распределения; в) гистограмму.

Глава 2. Статистические показатели

Статистические показатели широко применяются при характеристике социально-экономических явлений. Они подразделяются на:

1. Абсолютные величины.
2. Относительные величины.
3. Средние величины.
4. Показатели вариации.

§ 1. Абсолютные и относительные величины

Абсолютные величины характеризуют размеры, объемы, уровни изучаемых явлений и получаются, как правило, в результате измерения, которое осуществляется в натуральных, стоимостных и трудовых единицах. Натуральные единицы измерения – тонны, штуки, литры, человеко-часы и т. п. Стоимостные единицы – рубли, доллары и т. п. В трудовых единицах (человеко-днях, человеко-часах) измеряются затраты труда.

Относительные величины представляют собой показатели, получающиеся от деления одной абсолютной величины на другую и выражающие связь между ними. Эти показатели позволяют сравнивать величины друг с другом. Величина, с которой происходит сравнение (знаменатель дроби), называется **базой сравнения** или **основанием**. В зависимости от базы относительный показатель может быть представлен в долях единицы: десятых, сотых (процентное представление), тысячных (такая доля называется **промилле** и обозначается ‰), десятитысячных (доля называется продецимилле, обозначение ‰‰).

Относительные величины подразделяются на:

- величины планового задания;
- реализации плана;
- динамики (темпы роста);
- структуры;
- координации;
- интенсивности;
- сравнения.

Рассмотрим подробно каждый вид.

Величина **планового задания** характеризует плановое задание последующего периода (уровня) по сравнению с фактически достигнутым предыдущего периода (уровня). Данная величина вычисляется по формуле

$$K_{\text{план.зад.}} = x_{\text{пл}} / x_0, \quad (3)$$

где $x_{\text{пл}}$ – плановое задание на последующий период, x_0 – плановое задание предыдущего периода.

Пример 5.

Товарооборот фирмы за октябрь фактически составил 300 тыс. рублей. План на ноябрь – 450 тыс. руб. Определить величину планового задания.

Решение.

$$K_{\text{план.зад.}} = 450 / 300 = 1,5 \text{ (или 150\%)}. \quad (4)$$

Относительная величина реализации плана $K_{\text{реал.план}}$ характеризует степень выполнения планового задания. Если x_1 – фактический уровень за отчетный период, $x_{\text{пл}}$ – плановое задание, то

$$K_{\text{реал.план}} = x_1 / x_{\text{пл}}. \quad (4)$$

Пример 6.

План на декабрь был 450 тыс. рублей, фактическая реализация – 400 тыс. рублей. Определить относительную величину реализации.

Решение.

$$K_{\text{реал.план}} = 400 / 450 = 0,889 \text{ (88,9\%)}. \quad (5)$$

Относительная величина динамики характеризует изменение явлений во времени и вычисляется по формуле

$$K_{\partial} = x_1 / x_0, \quad (5)$$

где x_0 и x_1 – имеют указанный ранее смысл.

Так, для данных примеров 5 и 6 $K_{\partial} = 400 / 300 = 4/3 \approx 1,333$ (или 133,3%).

Из (3)–(5) вытекает, что рассмотренные величины связаны соотношением:

$$K_{\text{план.зад.}} \cdot K_{\text{реал.пл.}} = K_{\partial}. \quad (6)$$

Для рассмотренных примеров имеем:

$$1,5 \cdot 0,889 = 1,333.$$

Пример 7.

На предприятии планировалось повысить заработную плату рабочим в 3-м квартале по сравнению со вторым в 1,2 раза. Фактически заработная плата повысилась на 14%. Оценить выполнение плана по росту заработной платы.

Решение.

$$K_{\text{реал.пл.}} = \frac{K_{\partial}}{K_{\text{план.зад.}}} = \frac{1,14}{1,2} = 0,95 \text{ (или 95\%)}. \quad (6)$$

Относительная величина структуры является характеристикой удельного веса (доли) отдельных групп в общем объеме совокупности; вычисляется по формуле

$$K_{стр} = \frac{\text{отдельная часть совокупности}}{\text{весь объем совокупности}}. \quad (7)$$

Пример 8.

В табл. 12 приведены данные об издержках производства. В 3-м столбце указаны значения относительной величины структуры $K_{стр}$.

Таблица 12

Статьи затрат	Тыс. руб.	Структура затрат
1. Топливо и энергия	100	8,93%
2. Сырье и материалы	350	31,26%
3. Заработная плата	450	40,19%
4. Амортизация	120	10,72%
5. Прочее	100	8,93%
Итого:	1120	100%

Относительная величина координации характеризует отношение отдельных частей совокупности к одной из них, взятой за базу сравнения, т. е.

$$K_{koord} = \frac{\text{отдельная часть совокупности}}{\text{база сравнения}} \quad (8)$$

Пример 9.

Пусть в предыдущем примере за базу сравнения взято значение 120 тыс., выделенное на амортизацию. Соответствующие значения K_{koord} приведены в табл. 13.

Таблица 13

Статьи затрат	Тыс. руб.	Относительная величина координации
1	100	0,83
2	350	2,92
3	450	3,75
4	120	1
5	100	0,83

Относительная величина интенсивности используется для характеристики степени развития или распространения явления в определенной среде, выражается в промилле:

$$K_{интенс} = \frac{\text{размер изучаемого явления}}{\text{объем среды, в которой происходит развитие}} \cdot 1000. \quad (9)$$

Пример 10.

Среднегодовая численность населения исследуемого региона составляет 1 млн 600 тыс. За год родилось 8,5 тыс. человек, а умерло 10,2 тыс. человек. Определить коэффициенты рождаемости и смертности в регионе.

Решение.

$$K_{\text{рожд}} = \frac{8,5}{1600} \cdot 1000 = 5,3\text{‰}; \quad K_{\text{смертн}} = \frac{10,2}{1600} \cdot 1000 = 6,4\text{‰}.$$

Относительная величина сравнения характеризует соотношение одноименных явлений, относящихся к разным объектам:

$$K_{\text{ср}} = \frac{\text{размер явления по объекту } A}{\text{размер явления по объекту } B}. \quad (10)$$

Пример 11.

Объект A равен 5 единиц, объект B равен 4 единицы. На сколько $A > B$?

Решение.

Примем $A = 5$ за 100%, тогда $B = 4 - x\%$, т.е. $x = \frac{4 \cdot 100\%}{5} = 80\%$ и $A - B = 20\%$.

§ 2. Средние величины. Средняя арифметическая

Это обобщающие показатели, характеризующие типичный размер варьирующего признака в расчете на единицу однородной совокупности. Средняя величина отражает типичный уровень признака вне связи с индивидуальными особенностями отдельных единиц статистической совокупности, т. е. средняя величина не зависит от отклонений признака, которые обусловлены действием случайных факторов, но в ней учтены изменения, вызванные действием основных факторов.

Виды средних величин:

- * средняя арифметическая;
- * средняя гармоническая;
- * средняя геометрическая;
- * средняя хронологическая (§ 1 главы 6);
- * структурная средняя.

Средние арифметические, гармонические и геометрические объединяются под общим названием «**аналитические средние**».

Дадим определения, приведем расчетные формулы и примеры средних величин.

Средняя арифметическая простая используется в дискретных вариационных рядах, вычисляется по формуле

$$\bar{x}_{ap} = \frac{x_1 + x_2 + \dots + x_n}{n}, \quad (11)$$

где n – число единиц совокупности с учетом повторяемости, x_1, x_2, \dots, x_n – индивидуальные значения признака, т. е. все единицы совокупности складываются и сумма делится на их число.

Формулу (11) можно записать в виде:

$$\bar{x}_{ap} = \frac{\sum_{i=1}^n x_i}{n}. \quad (12)$$

Для упрощения записи в суммах подобного вида мы будем опускать индексы суммирования, т. е. вместо $\sum_{i=1}^n x_i$ будем записывать $\sum x$ или $\sum x_i$.

Замечание. При записи средней арифметической часто нижний индекс не пишут и используются обозначения \bar{x} или \bar{X} .

Пример 12.

Производство автомобилей в РФ в январе-мае 1996 г. (в тыс. штук) составило соответственно

65,0; 83,2; 79,3; 89,9; 76,6.

Определить средний объем производства за месяц в этот период.

Решение.

Воспользуемся формулой (11):

$$\bar{x}_{ap} = \frac{65,0 + 83,2 + 79,3 + 89,9 + 76,6}{5} = \frac{394,0}{5} = 78,8 \text{ (тыс. шт.)}.$$

Средняя арифметическая взвешенная определяется как для дискретного, так и для интервального ряда.

В случае когда частоты вариантов в дискретном ряде больше 1, для вычисления средней арифметической удобно использовать формулу (13) средней арифметической взвешенной:

$$\bar{x}_{ap} = \frac{x_1 \cdot n_1 + x_2 \cdot n_2 + \dots + x_k \cdot n_k}{n_1 + n_2 + \dots + n_k} = \frac{\sum x_i n_i}{n}, \quad (13)$$

где $x_i (i = \overline{1, k})$ – варианты, $n_i (i = \overline{1, k})$ – их частоты (веса) и $\sum n_i = n$.

Пример 13.

Распределение покупателей в универмаге за январь-февраль приведено в табл. 14.

Таблица 14

Количество покупателей x_i	50	65	75	80	90	110	115	125
Число дней n_i	1	4	10	14	16	9	4	2

Определить среднее число покупателей за 1 день торговли (магазин работает без выходных).

Решение.

По формуле (13):

$$\bar{x}_{ap} = \frac{50 \cdot 1 + 65 \cdot 4 + 75 \cdot 10 + 80 \cdot 14 + 90 \cdot 16 + 110 \cdot 9 + 115 \cdot 4 + 125 \cdot 2}{60} = 88,67 \text{ (покупателей)}.$$

Средняя арифметическая взвешенная для интервального ряда вычисляется по формуле

$$\bar{x}_{ap} = \frac{\sum x_i n_i}{n}, \quad (14)$$

где $x_i (i = \overline{1, k})$ – **середина** i -го интервала, $n_i (i = \overline{1, k})$ – частота i -го интервала.

В этом случае действительные значения признака заменяются средним значением интервала, которое в общем случае отличается от средней арифметической значений, включенных в данный интервал. Это отличие будет тем меньше, чем больше частота интервала и чем уже интервал.

Пример 14.

Для рассмотренного примера 1 о производительности труда найти среднюю производительность труда.

Решение.

Найдем середины интервалов:

$$x_1 = \frac{14 + 17}{2} = 15,5; \quad x_2 = \frac{17 + 20}{2} = 18,5; \quad x_3 = \frac{20 + 23}{2} = 21,5;$$

$$x_4 = \frac{23 + 26}{2} = 24,5; \quad x_5 = \frac{26 + 29}{2} = 27,5.$$

По формуле (14):

$$\bar{x}_{ap} = \frac{15,5 \cdot 5 + 18,5 \cdot 7 + 21,5 \cdot 3 + 24,5 \cdot 2 + 27,5 \cdot 3}{20} = \frac{403}{20} = 20,15 \text{ (усл. ед.)}.$$

Вычисление удобно оформить в виде (табл. 15).

Таблица 15

Производительность труда (усл. ед.)	Число работников n_i (частота)	Середина интервала x_i	$x_i \cdot n_i$
14–17	5	15,5	77,5
17–20	7	18,5	129,5
20–23	3	21,5	64,5
23–26	2	24,5	49
26–29	3	27,5	82,5
Итого:	$n = 20$		403

Для дискретного признака средняя арифметическая совпадает с его математическим ожиданием, а для непрерывного дает приближенное значение математического ожидания.

Перечислим **основные свойства средней арифметической**, вытекающие из ее определения и аналогичные свойствам математического ожидания случайной величины.

1. Средняя арифметическая постоянной равна самой постоянной.

2. Если все варианты увеличить (уменьшить) в одно и то же число раз, то средняя арифметическая увеличится (уменьшится) во столько же раз, т. е.

$$\overline{kx} = k\bar{x}.$$

3. Если все варианты увеличить (уменьшить) на одно и то же число, то средняя арифметическая увеличится (уменьшится) на то же число, т. е.

$$\overline{x + c} = \bar{x} + c$$

4. Средняя арифметическая отклонений вариантов от средней арифметической равна нулю:

$$\overline{x - \bar{x}} = 0.$$

5. Средняя арифметическая алгебраической суммы нескольких признаков равна такой же сумме средних арифметических этих признаков:

$$\overline{x + y} = \bar{x} + \bar{y}.$$

6. Если ряд состоит из нескольких групп, общая средняя равна средней арифметической групповых средних, причем весами являются объемы групп, т. е.

$$\bar{x} = \frac{1}{n} \cdot \sum_{i=1}^l \bar{x}_i \cdot n_i,$$

где \bar{x} – общая средняя (средняя арифметическая всего ряда); \bar{x}_i – групповая средняя i -й группы, объем которой равен n_i ; l – число групп.

В ряде задач может возникнуть необходимость оценки среднего значения альтернативного **качественного** признака. Например, если признак – качество продукции, то говорят о годной и бракованной продукции. Для характеристики данного признака вводится случайная величина X , принимающая два значения: 1, если данная единица совокупности годная, и 0 в противном случае.

Если среди n единиц совокупности будет n_1 годных и $n - n_1$ бракованных, то среднее значение вычислите по формуле

$$\bar{x} = \frac{1 \cdot n_1 + 0(n - n_1)}{n} = \frac{n_1}{n} = w, \quad (15)$$

где w – относительная частота.

Если отыскивается среднее значение частей некоторых величин, причем между частями нет прямо пропорциональной зависимости, то формулами средней арифметической пользоваться нельзя. В этом случае используется средняя гармоническая.

§ 3. Средняя гармоническая и средняя геометрическая

Выделяются **средняя гармоническая простая** и средняя гармоническая **взвешенная**. Первая вводится для дискретного ряда и вычисляется по формуле

$$\bar{x}_{\text{gap}} = \frac{1 + 1 + \dots + 1}{\frac{1}{x_1} + \frac{1}{x_2} + \dots + \frac{1}{x_n}} = \frac{n}{\sum \frac{1}{x}}. \quad (16)$$

Средняя гармоническая взвешенная определяется по формуле

$$\bar{x}_{\text{gap}} = \frac{\sum n_i}{\sum \frac{n_i}{x_i}} = \frac{n}{\sum \frac{n_i}{x_i}}, \quad (17)$$

причем для дискретного ряда $x_i (i = \overline{1, k})$ – варианты, $n_i (i = \overline{1, k})$ – их частоты; а для интервального ряда x_i – середины интервалов, n_i – частоты интервалов.

Для двух групп совокупности численностями n' и n'' со средними гармоническими $\bar{x}_{1\text{gap}}$ и $\bar{x}_{2\text{gap}}$ соответственно общая гармоническая вычисляется по формуле

$$\bar{x}_{\text{gap}} = \frac{n' + n''}{\frac{n'}{\bar{x}_{1\text{gap}}} + \frac{n''}{\bar{x}_{2\text{gap}}}}. \quad (18)$$

Пример 15.

В первом районе по итогам года оказалось 28 рентабельных предприятий, или 60%, во втором – 28 рентабельных предприятий, или 80%. Найти общий процент рентабельных предприятий.

Решение.

Пусть y_1 – число предприятий в первом районе, y_2 – во втором. Тогда имеем пропорции:

$$\begin{array}{l} 28 - 60\%; \quad 28 - 80\%; \\ y_1 - 100\%; \quad y_2 - 100\%; \\ y_1 = \frac{28 \cdot 100\%}{60\%} = 47; \quad y_2 = \frac{28 \cdot 100\%}{80\%} = 35. \end{array} \quad (19)$$

Теперь искомым средний процент x можно найти из пропорции:

$$47 + 35 - 100\%, \quad 28 + 28 - x\%, \text{ т. е.}$$

$$\begin{aligned} x &= \frac{2 \cdot 28}{47 + 35} \cdot 100\% = \frac{2}{\frac{47}{28 \cdot 100\%} + \frac{35}{28 \cdot 100\%}} = \\ &= \frac{2}{\frac{47}{47 \cdot 60\%} + \frac{35}{35 \cdot 80\%}} = \frac{2}{\frac{1}{60\%} + \frac{1}{80\%}}. \end{aligned}$$

Здесь сначала в выражении для x поделили числитель и знаменатель на $28 \cdot 100\%$, а затем воспользовались формулами (19). Положим $x_1 = 60\%$, $x_2 = 80\%$, тогда

$$x = \frac{2}{\frac{1}{60\%} + \frac{1}{80\%}} = \frac{2}{\frac{1}{x_1} + \frac{1}{x_2}} = \bar{x}_{\text{гар}}.$$

Таким образом, для рассмотренного примера

$$\bar{x}_{\text{гар}} = \frac{2}{\frac{1}{60\%} + \frac{1}{80\%}} = 68,5\%.$$

Обратим внимание на то, что по формуле средней арифметической получилось бы

$$\frac{60\% + 80\%}{2} = 70\%,$$

т. е. другой результат.

Средняя геометрическая простая (для дискретного ряда) вычисляется по формуле

$$\bar{x}_{\text{геом}} = \sqrt[n]{x_1 \cdot x_2 \cdot \dots \cdot x_n}, \quad (20)$$

где n – число единиц совокупности; а **средняя геометрическая взвешенная** – по формуле

$$\bar{x}_{geom} = \sqrt[n]{x_1^{n_1} \cdot x_2^{n_2} \dots x_k^{n_k}}, \quad (21)$$

причем для дискретного и интервального ряда $x_i (i = \overline{1, k})$ и $n_i (i = \overline{1, k})$ имеют тот же смысл, что и в формуле (17).

Данная величина применяется в следующих случаях:

а) для измерения степени отклонения индивидуальных значений признака относительно средней арифметической в вариационном ряду; например, если известно, что цена 1 кг апельсинов в данном городе колеблется от 24 до 30 рублей, то средняя цена в данном случае вычисляется как средняя геометрическая $\sqrt{24 \cdot 30} = 26,83$ (руб.);

б) для характеристики среднего темпа роста в рядах динамики, речь о которых пойдет в главе 6, а применение показано в следующем примере.

Имеет место **правило мажорантности средних**, т. е.

$$\bar{x}_{гарм} \leq \bar{x}_{geom} \leq \bar{x}_{арифм}.$$

Пример 16.

Годовые темпы роста продукции составили в 2008 г. 4,02; в 2009-м – 5,03; в 2010-м – 5,04; в 2011-м – 6,03. Найти средний годовой темп за четырехлетие.

$$\bar{x}_{geom} = \sqrt[4]{4,02 \cdot 5,03 \cdot 5,04 \cdot 6,03} = 4,98.$$

§ 4. Структурная средняя

Речь пойдет о моде и медиане.

Мода – это вариант с наибольшей частотой повторяемости. Так, для дискретного ряда примера 1 о производительности труда (§ 2 главы 1) мода $M_o = 18$, поскольку вариант $x_i = 18$ имеет наибольшую частоту, равную 4.

В интервальных рядах с равными интервалами мода M_o рассчитывается по формуле

$$M_o = x_{M_o} + i_{M_o} \cdot \frac{\Delta_1}{\Delta_1 + \Delta_2}, \quad (22)$$

где x_{M_o} – нижняя граница модального интервала, т. е. интервала с наибольшей частотой; i_{M_o} – величина модального интервала;

$$\Delta_1 = n_{M_o} - n_{M_o - 1}; \quad \Delta_2 = n_{M_o} - n_{M_o + 1},$$

n_{M_o} – частота модального интервала; $n_{M_o - 1}$ – частота предмодального интервала; $n_{M_o + 1}$ – частота послемодального интервала.

Пример 17.

Распределение заработной платы работников данного учреждения показано в табл. 16.

Таблица 16

Заработная плата, тыс. руб.	до 6000	6000– 6500	6500– 7000	7000– 7500	7500– 8000	свыше 8000
Число работников	5	12	19	21	13	5

Найти моду.

Решение.

Прежде всего определим модальный интервал – это интервал (7000, 7500). Найдём величину модального интервала $i_{M_o} = 7500 - 7000 = 500$;

$$\Delta_1 = n_{M_o} - n_{M_o-1} = 21 - 19 = 2;$$

$$\Delta_2 = n_{M_o} - n_{M_o+1} = 21 - 13 = 8;$$

$$M_o = 7000 + 500 \cdot \frac{2}{2+8} = 7100 \text{ (руб.)}.$$

Геометрический способ определения моды использует гистограмму, на которой ищется прямоугольник с наибольшей высотой (частотой), вершины этого прямоугольника соединяются отрезками с соответствующими вершинами двух соседних прямоугольников; абсцисса точки пересечения этих отрезков и будет модой ряда.

Для рассмотренного примера гистограмма показана на рис. 15. На этом рисунке без нарушения распределения начало координат помещено в точку (5000, 0). Абсцисса точки пересечения отрезков AB и CD является модой; из рисунка $M_o \approx 7100$.

Очевидно, имеются такие распределения, когда несколько вариантов имеют максимальную частоту. В этом случае распределение признака называют **полимодальным**.

На рис. 16 показан полигон двухмодального распределения дискретного признака.

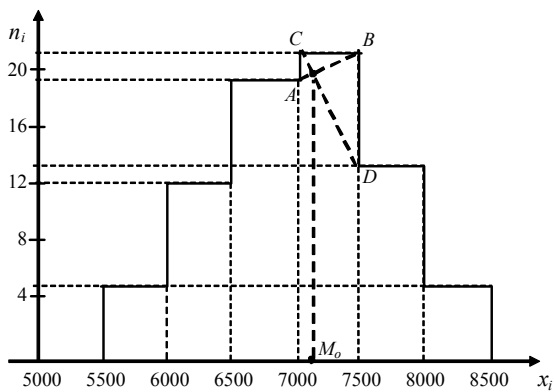


Рис. 15

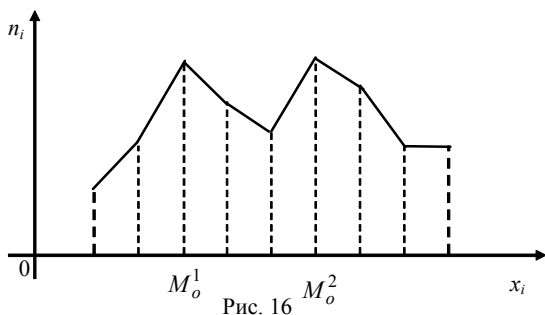


Рис. 16

На рис. 17 – пример двухмодального непрерывного признака, представленного плотностью распределения.

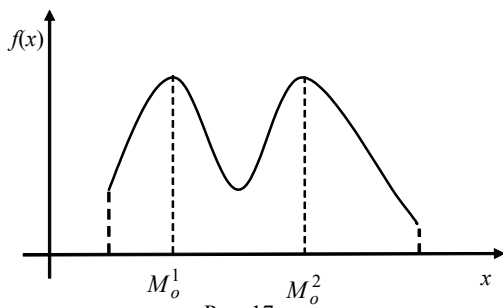


Рис. 17

Мода – это вариант, который по данному признаку является «самым-самым», например предприятие – самым прибыльным (самым убыточным), работник – высшей категории (низшей категории), ставка процента – самая высокая (самая низкая), цена – самая высокая (самая низкая) и т. п.

Еще одна структурная средняя – **медиана** M_e – представляет собой вариант в середине вариационного ряда. Для дискретного ряда номер медианы при **нечетном** числе единиц вычисляется по формуле

$$N_{M_e} = \frac{n+1}{2}, \quad (23)$$

где n – число единиц ряда. В случае **четного** числа единиц медиана равна средней из 2 вариантов, находящихся в середине ряда.

Пример 18.

Дан ряд значений дискретного признака: 1, 1, 2, 3, 4, 5, 6, 6, 6. Найти медиану.

Решение.

Число единиц $n = 9$ – нечетное, $M_e = 4$ (левее и правее находится одинаковое число единиц ряда). Добавим еще одно значение $x = 7$:

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

Тогда $n = 10$ – четное и $M_e = \frac{4+5}{2} = 4,5$.

В интервальных рядах значение медианы вычисляется по формуле

$$M_e = x_{M_e} + i_{M_e} \cdot \frac{\sum n_i - n_{M_e-1}^{нак}}{n_{M_e}}, \quad (24)$$

здесь x_{M_e} – нижняя граница медианного интервала;

i_{M_e} – величина медианного интервала;

$n_{M_e-1}^{нак}$ – сумма накопленных частот до начала медианного интервала;

n_{M_e} – частота медианного интервала.

При отыскании медианного интервала нужно помнить, что это первый интервал, накопленная частота для которого не меньше половины суммы всех частот.

Пример 19.

Найти медиану заработной платы для примера 17.

Решение.

Прежде всего определим медианный интервал. Общая сумма частот для данного примера: $n = 5 + 12 + 19 + 21 + 13 + 5 = 75$.

Поделив 75 на 2, получим 37,5. Первый интервал (считая слева направо), в котором сумма накопленных частот не меньше чем 37,5, – это интервал (7000, 7500), так как для предыдущего интервала сумма накопленных частот:

$$n_{M_e-1}^{\text{нак}} = 5 + 12 + 19 = 36 < 37,5,$$

в то время как

$$n_{M_e}^{\text{нак}} = 5 + 12 + 19 + 21 = 57 > 37,5.$$

По формуле (24):

$$M_e = 7000 + 500 \cdot \frac{75/2 - 36}{21} = 7035,7 \text{ (руб.)},$$

т. е. это значение признака такое, что в данном вариационном ряду одинаковое количество единиц, больших и меньших, чем M_e .

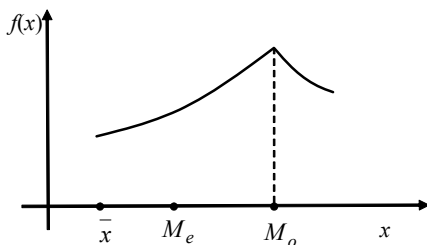


Рис. 18

Сформулируем свойство **мажорантности** средних: если $\bar{x} < M_e < M_o$, то в распределении наблюдается левосторонняя асимметрия (рис. 18). Если $\bar{x} > M_e > M_o$, то имеет место правосторонняя асимметрия.

§ 5. Показатели вариации

Средние величины не отражают **изменчивости (вариации)** значений признака, для характеристики которой вводятся **показатели вариации**:

- размах вариации;
- среднее линейное отклонение;
- дисперсия;
- среднее квадратическое отклонение;
- коэффициент вариации;
- средняя из групповых дисперсий;
- межгрупповая дисперсия;
- коэффициент детерминации.

Размах вариации рассчитывается по формуле

$$R = x_{\max} - x_{\min}. \quad (25)$$

Размах вариации показывает лишь крайние отклонения признака, но не отражает отклонения всех вариантов в ряду.

Среднее линейное отклонение подразделяется на:

а) **простое**, определяемое для дискретного ряда и вычисляемое по формуле

$$d = \frac{\sum |x - \bar{x}|}{n}, \quad (26)$$

где n – объем совокупности, \bar{x} – средняя арифметическая, x – единица совокупности;

б) **взвешенное**:

$$d = \frac{\sum |x_i - \bar{x}| \cdot n_i}{n}, \quad (27)$$

причем для дискретного ряда n_i – частота варианты x_i , \bar{x} – средняя арифметическая (взвешенная), n – объем выборки, т. е. $n = \sum_{i=1}^k n_i$, где k – число вариантов в рассматриваемой совокупности; для интервально-го ряда $x_i (i = \overline{1, k})$ – середина i -го интервала, $n_i (i = \overline{1, k})$ – его частота и $\sum_{i=1}^k n_i = n$.

Пример 20.

Найти среднее линейное отклонение производительности труда для рассмотренного в § 2 главы 1 примера 1 о производительности труда.

Решение.

Имеем интервальный ряд, воспользуемся формулой (27) и с учетом того, что $x_1 = 15,5$, $x_2 = 18,5$, $x_3 = 21,5$, $x_4 = 24,5$, $x_5 = 27,5$,

$$n_1 = 5, \quad n_2 = 7, \quad n_3 = 3, \quad n_4 = 2, \quad n_5 = 3, \quad \bar{x} = 20,5 \text{ усл. ед.}$$

получим

$$\begin{aligned} d &= \frac{1}{20} (|15,5 - 20,5| \cdot 5 + |18,5 - 20,5| \cdot 7 + |21,5 - 20,5| \cdot 3 + \\ &+ |24,5 - 20,5| \cdot 2 + |27,5 - 20,5| \cdot 3) = \frac{69,6}{20} = 3,48 \text{ (усл. ед.)}. \end{aligned}$$

Вывод: в среднем производительность труда в данной группе работников отклоняется от среднего значения на 3,48 усл. ед.

Еще один показатель вариации – **дисперсия** S^2 . Она представляет собой средний квадрат отклонений вариантов от их средней величины; в зависимости от исходных данных вычисляется как:

а) **простая**:

$$S^2 = \frac{1}{n} \sum (x - \bar{x})^2, \quad (28)$$

$$S^2 = \frac{1}{n} \sum x^2 - \bar{x}^2; \quad (29)$$

б) **взвешенная**:

$$S^2 = \frac{1}{n} \sum (x_i - \bar{x})^2 \cdot n_i, \quad (30)$$

$$S^2 = \frac{1}{n} \sum x^2 n_i - \bar{x}^2. \quad (31)$$

В формулах (28)–(31) смысл величин тот же, что и для среднего линейного отклонения. Покажем равносильность формул (28) и (29). Для формул (30) и (31) доказательство аналогичное. Итак,

$$\begin{aligned} \frac{1}{n} \sum (x - \bar{x})^2 &= \frac{1}{n} \sum (x^2 - 2x\bar{x} + \bar{x}^2) = \frac{1}{n} \sum x^2 - \frac{2\bar{x}}{n} \sum x + \frac{\bar{x}^2}{n} \sum 1 = \\ &= \frac{1}{n} \sum x^2 - 2\bar{x} + \bar{x}^2 = \frac{1}{n} \sum x^2 - \bar{x}^2. \end{aligned}$$

Дисперсия является характеристикой рассеивания и имеет размерность квадрата случайной величины.

Пример 21.

Найти дисперсию производительности труда примера 14 из главы 2.

Решение.

Воспользуемся формулой (31):

$$\begin{aligned} S^2 &= \frac{1}{20} (15,5^2 \cdot 5 + 18,5^2 \cdot 7 + 21,5^2 \cdot 3 + 24,5^2 \cdot 2 + 27,5^2 \cdot 3) - 20,15^2 = \\ &= 16,6 \text{ (усл. ед.)}^2 \end{aligned}$$

Оперировать с квадратом отклонения неудобно, поэтому вместо S^2 обычно рассматривается $S = \sqrt{S^2}$. Эта величина называется **средним квадратическим или стандартным отклонением** (СКО), которое имеет размерность случайной величины. Так, для рассмотренного примера $S = \sqrt{16,6} = 4,07$ (усл. ед.).

Чем меньше значение дисперсии (среднего квадратического отклонения), тем однороднее совокупность и тем более типичной для данной совокупности будет средняя величина.

Основные свойства дисперсии и СКО, определенные по выборке, аналогичны свойствам дисперсии и СКО случайной величины:

1. Дисперсия (СКО) постоянной равна нулю.
2. Если все варианты увеличить (уменьшить) в одно и то же число k раз, то дисперсия увеличится (уменьшится) в k^2 раз.
3. Если все варианты увеличить (уменьшить) на одно и то же число, то дисперсия не изменится.
4. Дисперсия равна разности между средней арифметической квадратов вариантов и квадратом средней арифметической:

$$S^2 = \overline{x^2} - \bar{x}^2. \quad (32)$$

Доказательство следует из (29) и (31).

5. Дисперсия суммы (разности) независимых признаков равна сумме дисперсий слагаемых. (Признаки независимы, если распределение каждого из них не зависит от значений других признаков.)

Дисперсию и СКО можно использовать и для характеристики вариации качественного признака. А именно с учетом формулы (15) имеем

$$\begin{aligned} S^2 &= \frac{(1-w)^2 \cdot n_1 + (0-w)^2 \cdot (n-n_1)}{n} = \\ &= (1-2w+w^2) \cdot w + w^2 \cdot (1-w) = w \cdot (1-w). \end{aligned} \quad (33)$$

Отсюда

$$S = \sqrt{w(1-w)}. \quad (34)$$

Пример 22.

При проверке знаний учащихся из 200 человек 30 не прошли тестирование. Пусть качественный признак соответствует хорошей подготовке обучаемых. Определить дисперсию этого признака.

Решение.

Имеем: $w = \frac{200-30}{200} = 0,85$; $1-w = 0,15$ и $S^2 = 0,85 \cdot 0,15 = 0,1275$.

В статистике часто возникает задача сравнения вариаций различных признаков. Например, вариаций стажа работы и размера заработной платы, стажа работы и производительности труда и т. п.

В этом случае рассмотренные выше показатели абсолютной колеблемости признаков непригодны: нельзя сравнивать вариацию, выраженную в годах, с вариацией, выраженной, например, в рублях. Поэтому используется относительный показатель вариации – **коэффициент вариации** V_o :

$$V_0 = \frac{S}{x} \cdot 100\%. \quad (35)$$

Если $V_0 \leq 33\%$, то совокупность считается **однородной**.

Пример 23.

Оценить вариацию рабочих по их квалификации (пример 21).

Решение.

$$\text{Найдем } V_0 = \frac{S}{x} = \frac{16,6}{20,15} \cdot 100\% = 82,4\% \text{ (т. е. } V_0 > 33\%).$$

Вывод: вариация составляет 82,4%, т. е. выборка не является однородной с точки зрения квалификации рабочих.

Используются также и другие относительные показатели вариации:

- коэффициент осцилляции

$$V_R = R / \bar{x}; \quad (36)$$

- линейный коэффициент вариации

$$V_d = d / \bar{x}. \quad (37)$$

Вариация признака обычно обусловлена различными факторами, для выявления влияния которых вся совокупность делится на группы согласно данному факторному признаку (группировочному фактору). Тогда можно определить внутригрупповую, межгрупповую и общую дисперсии (вариации). Вариация внутри каждой группы – **внутригрупповая дисперсия** S_i^2 (где i – номер группы) – это случайная вариация, т. е. часть вариации, обусловленная влиянием неучтенных факторов и **не зависящая** от факторного признака, положенного в основание группировки. Эта дисперсия может быть вычислена как простая или как взвешенная соответственно по формулам:

$$S_i^2 = \frac{\sum_j (y_{ij} - \bar{y}_i)^2}{n_i}, \quad (38)$$

$$S_i^2 = \frac{\sum_j (y_{ij} - \bar{y}_i)^2 \cdot n_{ij}}{n_i}, \quad (39)$$

где \bar{y}_i – средняя арифметическая i -й группы, n_i – число элементов в ней, n_{ij} – частота варианта x_{ij} для дискретного ряда и частота j -го интервала для интервального ряда.

Внутригрупповая дисперсия равна среднему квадрату отклонений отдельных значений признака внутри группы от средней арифметической этой группы.

Средняя внутригрупповых дисперсий:

$$\overline{S^2} = \frac{\sum S_i^2 \cdot n_i}{\sum n_i} \quad (40)$$

(здесь n_i – объем i -й группы) характеризует вариацию результативного признака во всей совокупности под влиянием случайных факторов.

Межгрупповая дисперсия $S_{\text{межгр}}^2$ характеризует систематическую вариацию результативного признака Y , обусловленную влиянием факторного признака, определяющего группировку. Она равна среднему квадрату отклонений групповых средних \bar{y}_i от общей средней (вычисленной для всей совокупности) \bar{y} :

$$S_{\text{межгр}}^2 = \frac{\sum (\bar{y}_i - \bar{y})^2 \cdot n_i}{\sum n_i}. \quad (41)$$

Общая дисперсия $S_{\text{общ}}^2$ измеряет вариацию признака во всей совокупности под влиянием всех факторов, обуславливающих эту вариацию:

$$S_{\text{общ}}^2 = \frac{\sum (y - \bar{y})^2 \cdot m_i}{\sum m_i}, \quad (42)$$

где m_i – частота варианта y или соответствующего интервала (в интервальном ряду).

Теорема (правило сложения дисперсий).

Если ряд состоит из нескольких групп наблюдений, то общая дисперсия равна сумме средней групповых дисперсий и межгрупповой дисперсии, т. е.

$$S_{\text{общ}}^2 = \overline{S}^2 + S_{\text{межгр}}^2. \quad (43)$$

Схематично это правило показано на рисунке (схема «правило сложения дисперсий»).

Докажем теорему, при этом для упрощения доказательства предположим, что вся совокупность значений количественного признака Y разбита на 2 группы, первая из которых характеризуется значениями признака y_i с частотой m_i общего объема $n_1 = \sum l_i$ с групповой средней \bar{y}_1 и групповой дисперсией S_1^2 , вторая характеризуется теми же значениями y_i с частотой n_i общего объема $n_2 = \sum k_i$ с групповой средней \bar{y}_2 и групповой дисперсией S_2^2 . Объем всей совокупности $n = n_1 + n_2$.

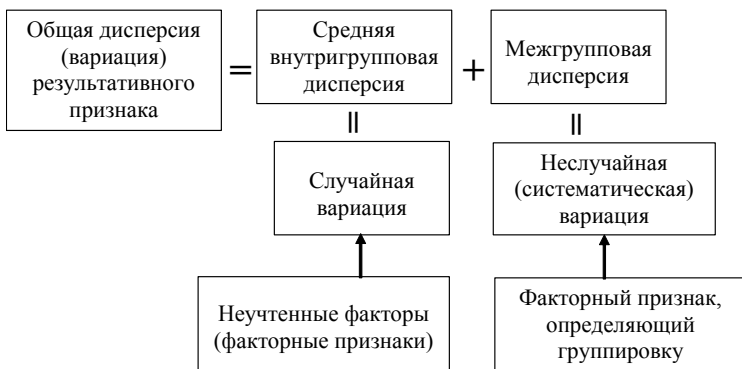


Рис. 19. Схема «правило сложения дисперсий»

Найдем общую дисперсию:

$$S_{\text{общ}}^2 = \left(\sum l_i (y_i - \bar{y})^2 + \sum k_i (\bar{y}_i - \bar{y})^2 \right) / n. \quad (44)$$

Преобразуем первое слагаемое числителя, для этого вычитаем и прибавляем y_1 :

$$\begin{aligned} \sum l_i (y_i - \bar{y})^2 &= \sum l_i [(y_i - \bar{y}_1) + (\bar{y}_1 - \bar{y})]^2 = \\ &= \sum l_i (y_i - \bar{y}_1)^2 + 2(\bar{y}_1 - \bar{y}) \sum l_i (y_i - \bar{y}_1) + \sum l_i (\bar{y}_1 - \bar{y})^2. \end{aligned}$$

Из определения S_1^2 следует, что

$$\sum l_i (y_i - \bar{y}_1)^2 = n_1 \cdot S_1^2.$$

В то же время $\bar{y}_1 = (\sum l_i y_i) / n_1$, откуда $\sum l_i y_i = \bar{y}_1 n_1$. Тогда

$$\sum l_i (y_i - \bar{y}_1) = \sum l_i y_i - \sum l_i \bar{y}_1 = \bar{y}_1 n_1 - \bar{y}_1 \sum l_i = \bar{y}_1 n_1 - \bar{y}_1 n_1 = 0.$$

Поэтому первое слагаемое числителя (44) примет вид

$$\sum l_i (y_i - \bar{y})^2 = n_1 \cdot S_1^2 + n_1 (\bar{y}_1 - \bar{y})^2.$$

Аналогично можно представить второе слагаемое числителя (44):

$$\sum k_i (\bar{y}_i - \bar{y})^2 = n_2 \cdot S_2^2 + n_2 (\bar{y}_2 - \bar{y})^2.$$

Подставив полученные выражения в (44), получим

$$\begin{aligned} S_{\text{общ}}^2 &= \left(n_1 \cdot S_1^2 + n_1 (\bar{y}_1 - \bar{y})^2 \right) / n + \left(n_2 \cdot S_2^2 + n_2 (\bar{y}_2 - \bar{y})^2 \right) / n = \\ &= (n_1 \cdot S_1^2 + n_2 \cdot S_2^2) / n + (n_1 (\bar{y}_1 - \bar{y})^2 + n_2 (\bar{y}_2 - \bar{y})^2) / n = \bar{S}^2 + S_{\text{межгр}}^2. \end{aligned}$$

Пример 24.

Для изучения влияния стажа работы на выработку рабочих произведена следующая группировка (табл. 17).

Таблица 17

Группы рабочих по стажу	Число рабочих	Выработка одного рабочего за смену (штук)
до 3 лет	5	2, 3, 3, 4, 4 – группа I
свыше 3 лет	15	2, 3, 3, 3, 3, 3, 3, 4, 4, 4, 4, 4, 4, 4 – группа II

Определить:

- 1) групповые средние и общую среднюю по выработке;
- 2) групповые дисперсии;
- 3) среднюю групповых дисперсий;
- 4) межгрупповую дисперсию;
- 5) общую дисперсию.

Решение.

В данной задаче стаж рабочих – факторный признак X , определяющий группировку, результативный признак Y – выработка рабочего за смену.

$$1) \bar{y}_1 = \frac{2+6+8}{5} = \frac{16}{5} = 3,2 \text{ – средняя выработка в группе I;}$$

$$\bar{y}_2 = \frac{2+18+32}{15} = \frac{52}{15} = 3,47 \text{ – средняя выработка в группе II;}$$

$$\bar{y}_{\text{общ}} = \frac{16+52}{5+15} = 3,4 \text{ – средняя выработка по двум группам.}$$

$$2) S_1^2 = \frac{(2-3,2)^2 + (3-3,2)^2 \cdot 2 + (4-3,2)^2 \cdot 2}{5} = 0,56 \text{ – дисперсия группы I;}$$

$$S_2^2 = \frac{(2-3,47)^2 + (3-3,47)^2 \cdot 6 + (4-3,47)^2 \cdot 8}{15} = 0,38 \text{ – дисперсия группы II.}$$

$$3) \overline{S^2} = \frac{0,56 \cdot 5 + 0,38 \cdot 15}{20} = 0,425 \text{ – средняя групповых дисперсий.}$$

$$4) S_{\text{межгр}}^2 = \frac{(3,2-3,4)^2 \cdot 5 + (3,47-3,4)^2 \cdot 15}{20} = 0,014.$$

$$5) S_{\text{общ}}^2 = 0,425 + 0,014 = 0,439.$$

Из определения межгрупповой дисперсии следует, что чем больше ее доля в общей дисперсии, тем сильнее влияние группировочного признака на изучаемый (результативный) признак. Для характеристики данной доли вводится эмпирический **коэффициент детерминации**:

$$\eta^2 = \frac{S_{\text{межгр}}^2}{S_{\text{общ}}^2}. \quad (45)$$

При отсутствии связи η^2 равен нулю, а при функциональной связи – единице.

Для рассмотренного примера $\eta^2 = \frac{0,014}{0,439} = 0,03$. Это означает, что

3% вариации выработки рабочих обусловлены различиями в их стаже, а остальные 97% связаны с другими факторами. Таким образом, связь между стажем и выработкой в рассмотренном примере фактически отсутствует. Подробно вопрос о связи (зависимости) факторов рассмотрен в главах 4–6.

Вопросы и задания к главе

1. Назовите виды статистических показателей с примерами.
2. Что характеризуют абсолютные величины? Привести примеры.
3. Как вычисляются и представляются относительные величины?
4. Что называется базой сравнения?
5. Что характеризует относительная величина: а) планового задания; б) реализации плана; в) динамики; г) структуры; д) координации; е) интенсивности; ж) сравнения?
6. По плану заработная плата во 2-м квартале должна была повыситься в 1,1 раза, а фактически повысилась на 5,2%. Как выполнен план по росту заработной платы?
7. Дайте определение средней, приведите примеры.
8. Как вычисляется средняя арифметическая простая и в каких случаях она применяется?
9. Как вычисляется средняя арифметическая взвешенная и в каких случаях она применяется?
10. Почему средняя арифметическая взвешенная для интервального ряда является приближенной? От чего зависит степень приближения?
11. Перечислите свойства средней арифметической.
12. Как оценивается среднее значение качественного признака?
13. Как определяется средняя гармоническая (простая и взвешенная)?
14. В каких случаях применяется средняя гармоническая?
15. Допустим, что пять рабочих заняты обработкой деталей. Затраты времени на обработку одной детали в мин. Следующие: 1-й – 12, 2-й – 10, 3-й – 6, 4-й – 10, 5-й – 12. Определить средние затраты времени на обработку одной детали.
16. Предприятием были выделены одинаковые денежные суммы на приобретение оборудования двух видов, при этом цена оборудования

первого вида составила 1 млн руб., второго – 1,8 млн руб. Рассчитать среднюю цену приобретенного оборудования.

17. Как определяется средняя геометрическая (простая и взвешенная)?

18. Когда используется средняя геометрическая?

19. Цена помидоров в данном районе колеблется от 34 до 100 руб. Найти среднюю цену помидоров в данном районе.

20. Темпы роста продукции составили в первом квартале 3,1, во втором – 5,1, в третьем – 6,2, в четвертом – 7,3. Найти средний квартальный темп роста продукции.

21. Дайте определение моды и приведите пример.

22. Как геометрически определяется мода?

23. Какое распределение называется полимодальным?

24. Приведите определение медианы и способ ее нахождения.

25. Сформулируйте свойство мажорантности средних.

26. Имеются данные о распределении семей относительно площади, приходящейся на одного человека (табл. 18).

Таблица 18

Площадь	5–7	7–9	9–11	11–13	13–15	15–17	17–19	19–21
Количество семей	4	7	15	25	17	16	10	6

Определить: а) среднюю площадь, приходящуюся на одного человека; б) медиану площади; в) модальную площадь.

27. Что называется вариацией признака?

28. Перечислите показатели вариации.

29. Что такое размах вариации, по какой формуле вычисляется, в чем его недостаток?

30. Дать определение дисперсии, привести расчетные формулы.

31. Какой показатель вариации называется средним квадратическим отклонением (СКО)?

32. Какую размерность имеют дисперсия и среднее квадратическое отклонение?

33. Как определяются дисперсия и СКО альтернативного признака?

34. Что характеризует коэффициент вариации?

35. На какие группы делятся факторы, вызывающие вариацию результативного признака?

36. Что характеризуют внутригрупповая и межгрупповая дисперсии?

37. Сформулировать правило сложения дисперсий.

38. В табл. 19 приведены данные по скорости автомобилей на некотором участке дорог (км/ч).

Таблица 19

35	40	45	50	38	35	45	55	60	35
62	52	48	50	52	52	38	50	48	45
65	48	47	60	60	38	55	58	52	62
65	45	57	53	54	55	35	60	48	60

Найти: а) размах вариации; б) коэффициент вариации; в) дисперсию и СКО скорости. Можно ли считать эту выборку однородной?

39. Что представляет собой коэффициент детерминации?

40. Определить коэффициент детерминации для скорости в задании 34, выделив 2, 3 или 4 группы значений скорости согласно выбранному группировочному факторному признаку. Насколько сильно влияние группировочного факторного признака на скорость?

Глава 3. Выборочное наблюдение

§ 1. Выборка и ее формирование

Выборочное наблюдение – наблюдение, при котором обследованию подвергается только часть статистической совокупности – **выборка**, а полученные результаты распространяются на всю совокупность, которая именуется **генеральной**.

Способы формирования выборки:

1. Собственно случайная выборка производится случайно или по жребию. Например, опрос людей на улице.
2. Механическая выборка: единицы генеральной совокупности упорядочиваются по определенному признаку и из полученной совокупности отбираются представители через определенный интервал. Примером является упорядочение предприятий по их доходу с последующей выборкой представителей.
3. Типическая выборка: единицы генеральной совокупности объединяются в группы по определенным признакам, после чего производится отбор представителей каждой группы случайно или механически пропорционально численности групп.
4. Серийная (гнездовая) выборка получается при отборе группами, сериями единиц.

Выборка может быть а) **повторной**, когда выбранная единица совокупности после изучения вновь возвращается в генеральную совокупность; б) **бесповторной**, при которой элементы назад не возвращаются.

§ 2. Оценки характеристик генеральной совокупности

Пусть x_1, x_2, \dots, x_N – генеральная совокупность. Элементы ее можно рассматривать как значения некоторой случайной величины (признака) X .

В качестве числовых характеристик генеральной совокупности объема N рассматриваются определенные в предыдущей главе показатели:

- 1) относительные; 2) средние; 3) вариации.

Эти характеристики называются **генеральными (истинными)**. Вычисляются они как соответствующие характеристики случайной величины X . Относительные и средние величины, а также показатели вариации, вычисленные на основе выборки, называются **выборочными** характеристиками и представляют собой **точечные оценки** соответствующих числовых характеристик генеральной совокупности (табл. 20).

Если рассматривается несколько признаков (случайных величин), то при обозначении характеристик используется соответствующая индексация, например m_x , S_x и т. д. В ряде источников вместо m_x используется запись M_x .

Таблица 20

Характеристики	Генеральные	Выборочные
1) Средняя арифметическая	m	\bar{x}
2) Дисперсия	D	S^2
3) Среднее квадратическое отклонение (СКО)	σ	S
4) Доля признака	P	w

Точечные оценки зависят от выборки (ее объема и представитель), а поэтому являются случайными величинами. К ним предъявляются следующие **требования**:

1. **Несмещенность** оценки, состоящая в том, что математическое ожидание оценки должно быть равно оцениваемой генеральной характеристике.

2. **Состоятельность** оценки, заключающаяся в том, что когда объем выборки $n \rightarrow \infty$, оценка сходится по вероятности к соответствующей характеристике генеральной совокупности, т. е. с достаточно большой вероятностью сколь угодно мало отличается от нее.

3. **Эффективность** оценки – разброс возможных значений оценки вокруг средней величины должен быть как можно меньше.

Проанализируем рассмотренные в табл. 20 оценки на несмещенность, состоятельность и эффективность.

Докажем, для примера, несмещенность средней арифметической \bar{x} . Пусть x_1, \dots, x_n – произвольная выборка, элементы которой можно трактовать как n независимых случайных величин, имеющих каждая тот же закон распределения и те же числовые характеристики, что и случайная величина (признак) X . Найдем

$$M[\bar{x}] = M\left[\frac{1}{n} \sum_{i=1}^n x_i\right] = \frac{1}{n} \sum_{i=1}^n M[x_i] = \frac{1}{n} \cdot n \cdot m = m.$$

Исследуем на несмещенность выборочную дисперсию s^2 . Во-первых,

$$\begin{aligned} \sigma_x^2 &= M[(\bar{x} - m)^2] = M\left[\left(\frac{\sum x_i}{n} - m\right)^2\right] = \frac{1}{n^2} M\left[\sum (x_i - m)^2\right] = \\ &= \frac{1}{n^2} \sum M[(x_i - m)^2] = \frac{1}{n^2} \cdot n \cdot \sigma^2 = \frac{\sigma^2}{n}, \end{aligned}$$

т. е.

$$\sigma_x^2 = \frac{\sigma^2}{n}. \quad (46)$$

Теперь найдем

$$\begin{aligned} M[S^2] &= M\left[\frac{1}{n}\sum(x_i - \bar{x})^2\right] = M\left[\frac{1}{n}\sum((x_i - m) - (\bar{x} - m))^2\right] = \\ &= M\left[\frac{1}{n}\sum(x_i - m)^2\right] - M\left[\frac{2}{n}\sum(x_i - m)(\bar{x} - m)\right] + M\left[\frac{1}{n}\sum(\bar{x} - m)^2\right] = \\ &= \frac{1}{n}M[\sum(x_i - m)^2] - 2M\left[(\bar{x} - m) \cdot \sum\frac{(x_i - m)}{n}\right] + \frac{1}{n}M[\sum(\bar{x} - m)^2] = \\ &= \frac{1}{n} \cdot n \cdot \sigma^2 - 2M\left[(\bar{x} - m) \cdot \left(\frac{1}{n}\sum x_i - \frac{1}{n} \cdot n \cdot m\right)\right] + \frac{1}{n} \cdot n \cdot M[(\bar{x} - m)^2] = \\ &= \sigma^2 - 2M[(\bar{x} - m)^2] + M[(\bar{x} - m)^2] = \sigma^2 - \sigma_x^2. \end{aligned}$$

Воспользовавшись формулой (46), получим

$$M[S^2] = \sigma^2 - \frac{\sigma^2}{n} = \frac{n-1}{n}\sigma^2 = \frac{n-1}{n}D < D.$$

Поэтому S^2 является смещенной оценкой. В то же время

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

является несмещенной оценкой. Вследствие этого часто применяется **исправленная дисперсия**:

$$\tilde{S}^2 = \sum_{i=1}^n \frac{(x_i - \bar{x})^2}{n-1}, \quad (47)$$

являющаяся несмещенной оценкой.

Однако разница между S^2 и $\frac{n}{n-1}S^2$ заметна при небольших n .

При $n > 30 - 40$ $S^2 \approx \frac{n}{n-1}S^2$, т. е. в качестве несмещенной оценки генеральной средней можно использовать рассмотренную ранее статистическую характеристику S^2 .

В качестве оценки относительного показателя доли признака p , который часто выражается в долях 1 и называется **вероятностью**, обычно рассматривается относительная частота $w = \frac{k}{n}$, где n – объем выборки, k – число единиц выборки, связанных с долей признака.

Например, это может быть доля предприятий промышленности с данной рентабельностью. Несмещенность этой оценки вытекает из

следующих рассуждений. Если каждому элементу $x_i (i = \overline{1, n})$ выборки поставить в соответствие случайную величину a_i , равную 1, если x_i входит в указанную долю признака, и 0, если нет, то, очевидно,

$$w = \frac{k}{n} = \frac{1}{n} \sum_{i=1}^n a_i, \quad (48)$$

откуда

$$M[w] = M\left[\frac{1}{n} \sum_{i=1}^n a_i\right] = \frac{1}{n} M\left[\sum_{i=1}^n a_i\right] = \frac{1}{n} \sum_{i=1}^n M[a_i]$$

Однако $x_i (i = \overline{1, n})$ входит в указанную долю признака с вероятностью p и не входит с вероятностью $q = 1 - p$. Поэтому

$$M[a_i] = 1 \cdot p + 0 \cdot (1 - p) = p,$$

$$D[a_i] = 1^2 \cdot p + 0^2 \cdot (1 - p) - p^2 = p(1 - p) = p \cdot q.$$

Таким образом,

$$M[w] = \frac{1}{n} \cdot n \cdot M[a_i] = M[a_i] = p,$$

т. е. w – несмещенная для p оценка.

Кроме того,

$$D[w] = D\left[\frac{1}{n} \sum_{i=1}^n a_i\right] = \frac{1}{n^2} \cdot n \cdot p \cdot q = \frac{pq}{n}, \text{ т. е.}$$

$$D[w] = \frac{pq}{n}. \quad (49)$$

Эта формула будет использована в § 7.

Состоятельность средней арифметической, дисперсии и среднего квадратического отклонения вытекает из теоремы Чебышева, так как \bar{x} сходится по вероятности к m . Состоятельность доли признака следует из теоремы Бернулли, поскольку w сходится по вероятности к p [1, 2].

Можно доказать, что w – эффективная оценка. Что же касается \bar{x} , S^2 , \tilde{S}^2 , то здесь эффективность зависит от вида закона распределения признака X . При нормальном распределении \bar{x} является эффективной оценкой [2–5], а S^2 , \tilde{S}^2 – асимптотически эффективными, т. е. с увеличением числа опытов n будут приближаться к эффективным.

§ 3. Метод моментов нахождения оценок

Согласно этому методу, оценки выбирают из условия, чтобы несколько важнейших генеральных характеристик были равны соответствующим выборочным аналогам (оценкам). Например, для равно-

мерного распределения $m = \frac{a+b}{2}$, $D = \frac{(b-a)^2}{12}$, где a и b – параметры.

Полагаем $m = \bar{x}$, $D = S^2$, где \bar{x} , S^2 – соответствующие оценки. Отсюда получаем оценки параметров:

$$\hat{a} = \bar{x} - \sqrt{3}S \quad \text{и} \quad \hat{b} = \bar{x} + \sqrt{3}S.$$

Приведем оценки параметров (табл. 21) для наиболее распространенных законов распределения признака (справочные данные по этим законам приведены в приложении 1).

Таблица 21

Закон распределения	Оценка параметра
Геометрический	$\hat{p} = \frac{1}{w}$
Биномиальный	$\hat{p} = w$
Пуассоновский	$\hat{a} = \bar{x}$
Показательный	$\hat{\lambda} = \frac{1}{\bar{x}}$
Равномерный	$\hat{a} = \bar{x} - \sqrt{3}S$, $\hat{b} = \bar{x} + \sqrt{3}S$
Нормальный	$m = \bar{x}$, $\sigma = S$

Пример 25.

Распределение интервала времени между поступлениями товаров в данный магазин задано следующей таблицей (табл. 22). Считая закон распределения интервала времени равномерным (свидетельством чему является приведенная таблица), оценить параметры этого закона.

Таблица 22

Интервал времени (дни)	1	2	3	4	5	6
Частота	3	2	3	3	2	3

Решение.

Сначала вычислим:

$$\bar{x} = \frac{1 \cdot 3 + 2 \cdot 2 + 3 \cdot 3 + 4 \cdot 3 + 5 \cdot 2 + 6 \cdot 3}{16} = 3,5;$$

$$S^2 = \frac{1^2 \cdot 3 + 2^2 \cdot 2 + 3^2 \cdot 3 + 4^2 \cdot 3 + 5^2 \cdot 2 + 6^2 \cdot 3}{16} - 3,5^2 = 3;$$

$$S = \sqrt{3} = 1,73.$$

Тогда $\hat{a} = 3,5 - \sqrt{3} \cdot \sqrt{3} = 0,5$; $\hat{b} = 3,5 + \sqrt{3} \cdot \sqrt{3} = 6,5$.

§ 4. Предельная ошибка выборки. Интервальное оценивание

При любом выборочном наблюдении всегда имеется ошибка, которая называется ошибкой **репрезентативности (предельной ошибкой)** или **точностью**. Она связана с тем, что рассматриваются не все элементы генеральной совокупности. Поэтому генеральные средняя, дисперсия и среднее квадратическое отклонение будут отличаться от своих статистических аналогов, вычисленных на основе выборки.

Пусть $\Delta_{\bar{x}}$ – ошибка (точность) в определении генеральной средней арифметической (генерального математического ожидания); Δ_w – ошибка в определении генеральной доли, т. е. доли единиц, обладающих данным признаком в общем числе единиц генеральной совокупности; Δ_S – ошибка в определении генеральной дисперсии. Тогда

$$|\bar{x} - m| \leq \Delta_{\bar{x}}, \quad |S^2 - \sigma^2| \leq \Delta_S, \quad |w - p| \leq \Delta_w.$$

В табл. 23 указана примерная классификация (данные приводятся из [6]) точности вычисления в зависимости от ошибки репрезентативности.

Таблица 23

Ошибка в %	Тип ошибки репрезентативности
до 3%	Повышенная точность
3–10%	Обычная точность
10–20%	Приближенная
20–40%	Ориентировочная
более 40%	Прикидочная

Как найти эти ошибки, если генеральные m , σ^2 и p неизвестны? Задачу эту решать можно, если гарантировать данную точность (ошибку) с определенной вероятностью, называемой **доверительной вероятностью** или **надежностью**. Эта вероятность обозначается через β и для m , D и p задается соответственно:

$$P(|\bar{x} - m| \leq \Delta_{\bar{x}}) = \beta, \quad (50)$$

$$P(|S^2 - D| \leq \Delta_S) = \beta, \quad (51)$$

$$P(|w - p| \leq \Delta_w) = \beta. \quad (52)$$

Обычно доверительная вероятность задается заранее, причем в качестве β берут число, близкое к 1. Наиболее часто β полагают равной 0,95; 0,99; 0,999.

Интервалы $(\bar{x} - \Delta_{\bar{x}}, \bar{x} + \Delta_{\bar{x}})$; $(S^2 - \Delta_S, S^2 + \Delta_S)$; $(w - \Delta_w, w + \Delta_w)$ называются **доверительными интервалами** соответственно для генеральных средней, дисперсии и вероятности.

В следующих параграфах будет рассмотрено построение доверительных интервалов и отыскание ошибки репрезентативности.

§ 5. Определение ошибки и объема репрезентативности для генеральной средней по большим выборкам

Для достаточно большого объема выборки n по центральной предельной теореме Ляпунова [1, 2] \bar{x} имеет нормальное распределение независимо от закона распределения X . Практически при $n > 30 - 40$ распределение \bar{x} можно считать приближенно нормальным. Поэтому

$$P\left(\bar{x} - m \leq \Delta_{\bar{x}}\right) = 2\Phi\left(\frac{\Delta_{\bar{x}}}{\sigma_{\bar{x}}}\right), \quad (53)$$

где $\sigma_{\bar{x}}$ – среднее квадратическое отклонение случайной величины \bar{X} . Таким образом,

$$\begin{aligned} \Phi\left(\frac{\Delta_{\bar{x}}}{\sigma_{\bar{x}}}\right) &= \frac{\beta}{2}, \\ \Delta_{\bar{x}} &= \sigma_{\bar{x}} \cdot \Phi^{-1}\left(\frac{\beta}{2}\right), \end{aligned} \quad (54)$$

где $\Phi^{-1}\left(\frac{\beta}{2}\right)$ – функция, обратная для функции Лапласа, она табулирована (приложение 3).

Значения функции Лапласа $\Phi(x)$ приведены в приложении 2. Это возрастающая функция, поэтому обратная для нее $\Phi^{-1}\left(\frac{\beta}{2}\right)$ тоже возрастающая. С учетом этого из формулы (54) можно сделать следующий **вывод**: чем больше доверительная вероятность β , тем больше величина ошибки $\Delta_{\bar{x}}$, т. е. меньше точность, и, наоборот, чем выше точность, тем меньше $\Delta_{\bar{x}}$ и меньше доверительная вероятность β . Следовательно, данные характеристики надо рассматривать в динамике и взаимосвязи.

Положим $\frac{\beta}{2} = \frac{\Delta_{\bar{x}}}{\sigma_{\bar{x}}}$, тогда $\Phi^{-1}\left(\frac{\beta}{2}\right) = u_{\beta}$ и формула (54) примет вид

$$\Delta_{\bar{x}} = \sigma_{\bar{x}} \cdot u_{\beta}, \quad (55)$$

откуда

$$\bar{x} - \sigma_{\bar{x}} \cdot u_{\beta} \leq m \leq \bar{x} + \sigma_{\bar{x}} \cdot u_{\beta}. \quad (56)$$

Интервал $(\bar{x} - \sigma_x^- \cdot u_\beta, \bar{x} + \sigma_x^- \cdot u_\beta)$ является **доверительным интервалом**, границы этого интервала называются **доверительными**.

Среднее квадратическое отклонение σ_x^- вычисляется по-разному, во-первых, в зависимости от того, какая выборка рассматривается: повторная или бесповторная; во-вторых, с учетом того, известно или нет генеральное среднее квадратическое отклонение σ . Это показано в табл. 24.

Таблица 24

Формулы СКО $\sigma_x^-(\sigma_x')$		
<div style="text-align: center;"> <div style="display: flex; justify-content: space-between;"> Тип выборки σ </div> </div>	Повторная выборка	Бесповторная выборка
σ известно	$\sigma_x^- = \frac{\sigma}{\sqrt{n}}$	$\sigma_x' = \sigma \sqrt{\frac{1}{n} - \frac{1}{N}}$
σ неизвестно S – оценка σ	$\sigma_x^- \approx \frac{S}{\sqrt{n}}$	$\sigma_x' \approx S \sqrt{\frac{1}{n} - \frac{1}{N}}$

Первая формула для повторной выборки получается с применением формулы (12) средней арифметической с учетом того факта, что все x_i – это независимые СВ с одной и той же дисперсией σ^2 , и согласно свойствам 2, 5 дисперсии, постоянный множитель $\frac{1}{n}$ выносится за знак дисперсии, при этом возводится в квадрат, а дисперсия суммы независимых СВ равна сумме дисперсий слагаемых. Поэтому

$$\sigma_x^2 = \sigma^2 \left[\frac{1}{n} \sum x_i \right] = \frac{1}{n^2} \sum \sigma^2 [x_i] = \frac{1}{n^2} \cdot n \cdot \sigma^2 = \frac{\sigma^2}{n},$$

откуда

$$\sigma_x^- = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}. \quad (57)$$

Вторая формула для этой выборки получается заменой σ на S . Соответствующие формулы для бесповторной выборки приведем без вывода (вывод можно посмотреть, например, в [2]).

С учетом (57) формулы (55) и (56) можно записать соответственно в виде:

$$\Delta_x^- = u_\beta \cdot \frac{\sigma}{\sqrt{n}}, \quad (58)$$

$$\bar{x} - u_\beta \cdot \frac{\sigma}{\sqrt{n}} \leq m \leq \bar{x} + u_\beta \cdot \frac{\sigma}{\sqrt{n}}. \quad (59)$$

Если σ не задана, то в формулах (58), (59) используется ее оценка S . Если $n \ll N$ (n много меньше N), то, как видно из табл. 24, расчет можно производить по формуле повторной выборки.

Минимальный объем выборки n , при котором имеет место данная ошибка репрезентативности Δ_x^- при данном уровне надежности β , например для случая повторной выборки и известном σ , получается на основе формул (55) и (57):

$$n = \frac{u_\beta^2 \cdot \sigma^2}{\Delta_x^-}. \quad (60)$$

Такой объем называется **объемом репрезентативности**. Аналогично получаются другие формулы для вычисления объема (табл. 25).

Таблица 25

Формулы объема репрезентативности		
Тип выборки σ	Повторная выборка	Бесповторная выборка
σ известно	$n = \frac{u_\beta^2 \cdot \sigma^2}{\Delta_x^-^2}$	$n' = \frac{Nu_\beta^2 \sigma^2}{u_\beta^2 \sigma^2 + N\Delta_x^-^2}$
σ неизвестно S – оценка σ	$n = \frac{u_\beta^2 \cdot S^2}{\Delta_x^-^2}$	$n' = \frac{Nu_\beta^2 S^2}{u_\beta^2 S^2 + N\Delta_x^-^2}$

Замечание 1. Поскольку данные формулы справедливы лишь при объеме, не меньшем 40, то в случае, когда $n(n')$ меньше 40, считаем необходимый объем репрезентативности равным 40.

Замечание 2. Поделив в формулах табл. 25 числитель и знаменатель на $\Delta_x^-^2$, получим связь между объемом n повторной выборки и объемом n' бесповторной выборки:

$$n' = \frac{n \cdot N}{n + N}. \quad (61)$$

Замечание 3. Из (61) следует, что при одинаковых ошибках репрезентативности и доверительных вероятностях объем репрезентативности повторной выборки будет всегда больше объема репрезентативности бесповторной выборки.

Пример 26.

Произведены 60 опытов по определению времени обслуживания клиентов в Центральном банке города Энска (в минутах). Считаем выборку повторной, т. е. один и тот же клиент может несколько раз участвовать в эксперименте. Данные приведены в табл. 26.

Определить: 1) по выборке среднее время обслуживания; 2) вероятность того, что генеральная средняя времени обслуживания отличается от среднего времени, определенного по выборке, не более чем на 0,3 минуты по абсолютной величине; 3) границы, в которых с вероятностью $\beta=0,99$ заключается генеральная средняя времени обслуживания, и тип ошибки репрезентативности; 4) объем выборки, при котором доверительные границы с ошибкой $\Delta_{\bar{x}}=4,2$ имели бы место с доверительной вероятностью 0,95.

Таблица 26

25	6	7	59	1	16	6	26	14	40	18	37
2	28	43	6	2	2	7	13	52	17	2	9
29	4	35	36	2	5	2	35	5	21	14	29
8	4	6	17	1	55	65	1	6	2	3	6
12	28	19	4	7	9	24	3	1	24	4	59

Решение.

1) В данном примере время – непрерывный количественный признак X , представленный достаточно объемной выборкой. Можно поступить двояко: либо найти \bar{x} по формуле взвешенной арифметической, либо использовать формулу (14) взвешенной арифметической интервального ряда.

В первом случае

$$\bar{x} = \frac{1}{60} \cdot \left(\begin{aligned} &1 \cdot 4 + 2 \cdot 7 + 3 \cdot 2 + 4 \cdot 4 + 5 \cdot 2 + 6 \cdot 6 + 7 \cdot 3 + 8 + \\ &+ 9 \cdot 2 + 12 + 13 + 14 \cdot 2 + 16 + 17 \cdot 2 + 18 + 19 + \\ &+ 21 + 24 \cdot 2 + 25 + 26 + 28 \cdot 2 + 29 \cdot 2 + 35 \cdot 2 + \\ &+ 36 + 37 + 40 + 43 + 52 + 55 + 59 \cdot 2 + 65 \end{aligned} \right) = \frac{1023}{60} = 17,05 \text{ (мин)}.$$

Во втором случае имеем интервальный ряд (табл. 27).

Таблица 27

Интервалы	(1, 9)	(9, 17)	(17, 25)	(25, 33)	(33, 41)	(41, 49)	(49, 57)	(57, 65)
n_i	31	7	6	5	5	1	2	3

Найдем

$$\bar{x} = \frac{1}{60} (5 \cdot 31 + 13 \cdot 7 + 21 \cdot 6 + 29 \cdot 5 + 37 \cdot 5 + 45 + 53 \cdot 2 + 61 \cdot 3) = \frac{1036}{60} = 17,27 \text{ (мин)}.$$

Вывод: при двух способах вычисления результаты получились достаточно близкими: разница составляет менее 1,3%. Так будет всегда при достаточно большом объеме выборки. Но вычисления при этом, конечно, удобнее проводить по формуле (14).

2) По формуле (31) найдем выборочную дисперсию:

$$S^2 = \frac{1}{60}(5^2 \cdot 31 + 13^2 \cdot 7 + 21^2 \cdot 6 + 29^2 \cdot 5 + 37^2 \cdot 5 + 45^2 + 53^2 \cdot 2 + 61^2 \cdot 3) - (17,27)^2 = 276,2 \quad (\text{мин}^2),$$

откуда $S = \sqrt{276,2} = 16,62$ мин.

По формуле (57) при $\sigma \approx S$ находим $\sigma_x \approx \frac{16,62}{\sqrt{60}} = 2,14$ (мин).

Используя (53), получим $P(|17,27 - m| \leq 0,3) = 2\Phi\left(\frac{0,3}{2,14}\right) = 2\Phi(0,14)$.

Из приложения 2 находим $\Phi(0,14) = 0,0557$, тогда искомая вероятность равна 0,11.

Вывод: генеральная средняя времени обслуживания m отличается от выборочной средней \bar{x} по абсолютной величине не более чем на 0,3 минуты с вероятностью 0,11.

3) Сначала по таблице 3 приложения для $\beta = 0,99$ найдем соответствующее значение $u_\beta = \Phi^{-1}\left(\frac{0,99}{2}\right) = 2,58$. Отсюда ошибка репрезентативности $\Delta_x = u_\beta \cdot \sigma_x = 2,58 \cdot 2,14 = 5,52$.

Поэтому доверительные границы будут:

$$\bar{x} - \Delta_x = 17,27 - 5,52 = 11,75 \quad (\text{мин}); \quad \bar{x} + \Delta_x = 17,27 + 5,52 = 22,79 \quad (\text{мин}).$$

Таким образом, с надежностью $\beta = 0,99$ генеральная средняя находится в границах: 11,75 мин $< m < 22,79$ мин.

Это означает, что если проанализировать несколько выборок, подобных рассмотренной и содержащих по 60 элементов, то в каждой будет свое среднее время \bar{x} . Но можно гарантировать, что примерно в 99% этих выборок их средняя \bar{x} будет отличаться от средней генеральной не более чем на 5,52 мин.

Если значение 11,75 нижней доверительной границы принять за 100%, то ошибка $\Delta_x = 5,52$ будет составлять не более 47%, что вытекает из пропорции: $11,75 - 100\% \quad 5,52 - x\%$, откуда

$$x\% = \frac{5,52 \cdot 100\%}{11,75} = 47\%,$$

что соответствует прикладной точности.

Если принять за 100% верхнюю доверительную границу, равную 22,79, то точность будет выше.

4) Расчет будем вести по формуле (60): $n = \frac{u_{\beta}^2 \cdot S^2}{\Delta_{\bar{x}}^2}$.

Параметр u_{β} найдем из приложения 3 при $\beta = 0,95$: $u_{\beta} = 1,96$. Тогда

$$n = \frac{(1,96)^2 \cdot 276,08}{(4,2)^2} \approx 61.$$

Пример 27.

Условие то же, что и в предыдущем примере, но выборка бесповторная: из 600 данных по определению времени обслуживания разных клиентов выбраны 60. Требуется найти то же самое, что и в предыдущей задаче под пунктами 3)–4).

Решение.

При решении воспользуемся данными предыдущего примера: $\bar{x} = 17,27$; $S = 16,62$.

3) Используем табл. 24 для бесповторной выборки при $n = 60$; $N = 600$ и $S = 16,62$:

$$\sigma'_x = S \sqrt{\frac{1}{n} - \frac{1}{N}} = 16,62 \sqrt{\frac{1}{60} - \frac{1}{600}} \approx 2,27 \text{ (мин)}.$$

Ошибка репрезентативности $\Delta_{\bar{x}} = u_{\beta} \cdot \sigma'_x = 2,58 \cdot 2,27 = 5,86$; доверительные границы:

$$\bar{x} - \Delta_{\bar{x}} = 17,27 - 5,86 = 11,41 \text{ (мин)}; \quad \bar{x} + \Delta_{\bar{x}} = 17,27 + 5,86 = 23,13 \text{ (мин)}.$$

$$4) \text{ Применим формулу (61): } n' = \frac{n \cdot N}{n + N} = \frac{60 \cdot 600}{60 + 600} \approx 55.$$

Замечание. Отметим, что рассмотренные в данном параграфе формулы вычисления доверительной вероятности, объема и ошибки репрезентативности являются **приближенными**, но верными для любых законов распределения признака. В следующем параграфе будут рассмотрены точные формулы, справедливые для нормального распределения признака X .

§ 6. Ошибка и объем репрезентативности генеральной средней для малых выборок

Речь пойдет о нормальном распределении признака X . В этом случае $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ имеет нормальное распределение для любого n . Поэтому, когда генеральное среднее квадратическое отклонение σ известно, при любом n можно воспользоваться формулами (54)–(60),

при этом объем и ошибка репрезентативности будут найдены **точно** для заданной доверительной вероятности β .

В случае, когда σ не дано, используется оценка S , вычисленная на основе данной выборки, поэтому ошибка и объем репрезентативности, вычисляемые по формулам (54)–(60), будут приближенными. Однако и в этом случае для данной вероятности β можно получить **точные** объем и ошибку репрезентативности, если в формуле (54) вместо $\Phi^{-1}\left(\frac{\beta}{2}\right)$ использовать **статистику** (функцию выборки)

$$t = \frac{\bar{x} - m}{S} \cdot \sqrt{n-1}, \quad (62)$$

которая имеет **распределение Стьюдента**. Число $k = n - 1$ называется числом **степеней свободы** и определяется как общее число n наблюдений признака X минус число уравнений, связывающих эти наблюдения. Число уравнений равно единице, так как наблюдения связаны уравнением для выборочной средней. Данная функция табулирована (приложение 4). Входами в таблицу являются доверительная вероятность β и число степеней свободы k . Из приложения 4 находится значение $t_{k,\beta}$. Для этого значения справедливо равенство

$$P(|\bar{x} - m| \leq \Delta_x^-) = \beta, \quad \text{где} \quad \Delta_x^- = t_{k,\beta} \cdot \sigma_x^- \quad \text{и} \quad \Delta_x^- = \frac{t_{k,\beta} \cdot S}{\sqrt{n-1}} \quad (63)$$

Из (63) при заданных S , Δ_x^- определяется объем репрезентативности

$$n = 1 + \frac{t_{k,\beta}^2 \cdot S^2}{\Delta_x^2}. \quad (64)$$

При достаточно больших n ($n > 20$) распределение Стьюдента стремится к нормальному, поэтому различие между соответствующими доверительными интервалами будет незначительным.

Пример 28.

Для проверки фасовочной установки были отобраны и взвешены 20 упаковок. Получены следующие результаты в граммах (табл. 28).

Таблица 28

Интервал	246-247,5	247,5-249	249-250,5	250,5-252	252-253,5	253,5-255
n_i	2	3	5	5	4	1

Оценить с доверительной вероятностью $\beta = 0,95$ ошибку в определении истинного (генерального) среднего веса.

Решение.

Из анализа данного статистического ряда можно сделать предположение о нормальном распределении веса упаковки. Генеральное σ не задано. Применяя (14), найдем

$$\bar{x} = \frac{1}{20} \left(246,75 \cdot 2 + 248,25 \cdot 3 + 249,75 \cdot 5 + \right. \\ \left. + 251,25 \cdot 5 + 252,75 \cdot 4 + 254,25 \cdot 1 \right) = 250,43.$$

По формуле (31) вычислим

$$S^2 = \frac{1}{20} \left[(246,75)^2 \cdot 2 + (248,25)^2 \cdot 3 + (249,75)^2 \cdot 5 + \right. \\ \left. + (251,25)^2 \cdot 5 + (252,75)^2 \cdot 4 + (254,25)^2 \cdot 1 \right] - (250,43)^2 = 4,16;$$

$$S = \sqrt{4,16} = 2,04.$$

Воспользуемся статистикой Стьюдента (приложение 4). Для $\beta = 0,95$ и $k = n - 1 = 20 - 1 = 19$ находим $t_{19; 0,95} = 2,09$ и $\Delta_x = 2,09 \cdot \frac{2,04}{\sqrt{19}} = 0,98$.

Вывод: в 95% случаев истинный средний вес будет в границах $250,43 - 0,98 < m < 250,43 + 0,98$, т. е. $249,45 < m < 251,40$.

Замечание. Если точность найдена по формуле (58) с применением обратной функции Лапласа для $\beta = 0,95$ при $u_\beta = 1,96$ и $S = 2,04$, то получим $\Delta_x \approx 1,96 \cdot \frac{2,04}{\sqrt{19}} = 0,91$.

Следовательно, в данном примере точность оценки, вычисленная с применением обратной функции Лапласа, ненамного отличается от точности, вычисленной с применением статистики Стьюдента.

§ 7. Точность оценки доли признака при больших выборках

Для генеральной доли имеет место соотношение (52):

$$P(|p - w| \leq \Delta_w) = \beta,$$

где p – генеральная доля, w – выборочная доля, Δ_w – ошибка репрезентативности, β – доверительная вероятность. При этом

$$w = \frac{k}{n} = \frac{1}{n} \sum_{i=1}^n a_i,$$

где $a_i = 1$, если элемент x_i выборки попадает в рассматриваемую долю, и $a_i = 0$ в противном случае.

Согласно центральной предельной теореме Ляпунова, при достаточно большом n ($n > 30 - 40$) w имеет нормальное распределение, поэтому соотношение (52) можно записать в виде

$$P(|p - w| \leq \Delta_w) = 2\Phi\left(\frac{\Delta_w}{\sigma_w}\right) = \beta, \quad (65)$$

отсюда

$$\Delta_w = \sigma_w \cdot u_\beta, \quad (66)$$

где $u_\beta = \Phi^{-1}\left(\frac{\beta}{2}\right)$, σ_w – СКО w , определяемое из табл. 29.

Таблица 29

Формулы средних квадратических отклонений $\sigma_w(\sigma'_w)$		
Тип выборки <i>p</i>	Повторная выборка	Бесповторная выборка
<i>p</i> известна	$\sigma_w = \sqrt{\frac{p(1-p)}{n}}$	$\sigma'_w = \sqrt{\frac{p(1-p)}{n} \left(1 - \frac{n}{N}\right)}$
<i>p</i> неизвестна <i>w</i> – оценка <i>p</i>	$\sigma_w \approx \sqrt{\frac{w(1-w)}{n}}$	$\sigma'_w \approx \sqrt{\frac{w(1-w)}{n} \left(1 - \frac{n}{N}\right)}$

Формулы для σ_w получаются на основе (49), для σ'_w – аналогично, но для зависимых x_1, x_2, \dots, x_n . Объем репрезентативности (табл. 30) получается из соответствующих формул таблицы 29 с учетом (66).

Таблица 30

Формулы объема репрезентативности		
Тип выборки <i>p</i>	Повторная выборка	Бесповторная выборка
<i>p</i> известна	$n = \frac{u_\beta^2 \cdot p \cdot q}{\Delta_w^2}$	$n' = \frac{N \cdot u_\beta^2 \cdot p \cdot q}{u_\beta^2 \cdot p \cdot q + N \cdot \Delta_w^2}$
<i>p</i> неизвестна <i>w</i> – оценка <i>p</i>	$n = \frac{u_\beta^2 w(1-w)}{\Delta_w^2}$	$n' = \frac{N \cdot u_\beta^2 \cdot w(1-w)}{u_\beta^2 \cdot w(1-w) + N \cdot \Delta_w^2}$

Пример 29.

Среди выборочно обследованных 500 семей данного района по уровню дохода на душу населения доля малообеспеченных семей составила 50. Требуется: 1) с вероятностью $\beta = 0,997$ оценить ошибку в определении доли малообеспеченных семей во всем районе; 2) границы, в которых с надежностью β заключена доля малообеспеченных семей во всем районе.

Решение.

Поскольку число N всех семей в районе достаточно велико, т. е. $N \gg n = 500$, при решении задачи можно использовать формулы повторной выборки. Итак,

$$w = \frac{50}{500} = 0,1; \quad \sigma_w = \sqrt{\frac{w(1-w)}{n}} = \sqrt{\frac{0,1 \cdot 0,9}{50}} = 0,042.$$

Из приложения 3 при $\beta = 0,997$ находим $u_\beta = 2,96$. Отсюда

$$\Delta_w = 2,96 \cdot 0,042 \approx 0,12, \text{ или } 12\%;$$

$$0,1 - 0,12 \leq w \leq 0,1 + 0,12, \text{ т. е. } 0 \leq w \leq 0,22, \text{ или } 22\%.$$

Вывод: с вероятностью 0,997 можно утверждать, что ошибка в определении процента малообеспеченных семей в данном районе не превосходит 12%, а доля таких семей не превосходит 22% во всем районе.

Пример 30.

Найти необходимое число измерений величины товарооборота, при котором для доли $p = 0,4$ с вероятностью $\beta = 0,9$ можно утверждать, что истинная (генеральная) величина товарооборота отличается от средней арифметической не более чем на 0,15 усл. ед.

Решение.

Из таблицы 30 применим формулу объема репрезентативности (количества измерений) для повторной выборки и известной доли p :

$$n = \frac{u_\beta^2 pq}{\Delta_w^2}.$$

$$\text{Имеем: } \Delta_w = 0,15; p = 0,4; q = 0,6; u_\beta = \Phi^{-1}\left(\frac{0,9}{2}\right) = 1,65 \text{ и } n = \frac{1,65^2 \cdot 0,4 \cdot 0,6}{0,15^2} \approx 30.$$

§ 8. Точность оценки доли признака при малых выборках

Пусть доля признака в генеральной совокупности равна p . Тогда вероятность того, что в выборке объема n элементов обладают этим признаком, вычисляется по формуле Бернулли [1–3]:

$$P_{k,n} = C_n^k \cdot p^k \cdot q^{n-k}.$$

Если $\tilde{k} = nw$ – фактическое число элементов выборки, обладающих данным признаком, то в качестве доверительного интервала рассматривается интервал (p_1, p_2) такой, что вероятность попадания левее p_1 и правее p_2 одна и та же и равна $(1 - \beta)/2$, т. е.

$$\sum_{k=0}^{\tilde{k}} P_{k,n} = \sum_{k=\tilde{k}}^n P_{k,n} = (1 - \beta)/2.$$

Значения p_1 и p_2 для $\beta = 0,9$ и $k = 5; 10; 15; 25; 50; 250; \infty$ приведены на рис. 20.

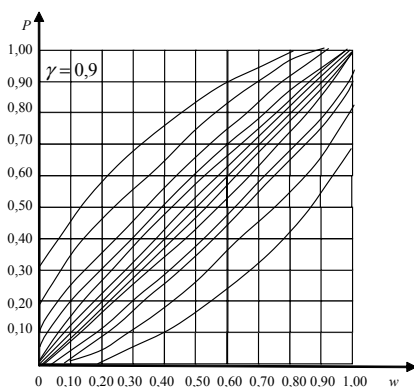


Рис. 20

На этом рисунке линии (сверху вниз) соответствуют следующим значениям n : 5, 10, 15, 25, 50, 250, ∞ , 250, 50, 25, 15, 10, 5.

Пример 31.

Случайно отобранные для опроса 25 пенсионеров высказались относительно номинации следующим образом: 6 – за, остальные – против. Указать долю пенсионеров данного города, относительно которых с уверенностью (доверительной вероятностью) $\beta = 0,9$ можно утверждать, что они поддерживают номинацию.

Решение.

Выборочная доля пенсионеров, поддерживающих номинацию, $w = k/n = 6/25 = 0,24$. Для $w = 0,24$ и $n = 25$ из рисунка находим $p_1 = 0,18$; $p_2 = 0,35$.

Вывод: с уверенностью 0,9 можно утверждать, что доля пенсионеров данного города, поддерживающих номинацию, заключена в границах от 18 до 35%.

§ 9. Интервальная оценка генеральной дисперсии нормально распределенного признака

Построение доверительного интервала для генеральной дисперсии основывается на том, что случайная величина $\chi^2 = \frac{n^2 \cdot S_0^2}{\sigma^2}$ (где $S_0^2 = \frac{1}{n} \sum (x_i - m)^2 \cdot n_i$, m – генеральная средняя, σ – генеральное отклонение) имеет χ^2 – **распределение Пирсона** с $k = n$ степенями свободы (приложение 5) и

$$P\left(\chi_1^2 < \frac{n \cdot S_0^2}{\sigma^2} < \chi_2^2\right) = \beta. \quad (67)$$

Значения χ_1^2 и χ_2^2 для данного β определяются неоднозначно. Обычно их выбирают таким образом, чтобы события $\chi^2 < \chi_1^2$ и $\chi^2 > \chi_2^2$ имели равные вероятности, т. е.

$$P(\chi^2 < \chi_1^2) = P(\chi^2 \geq \chi_2^2) = \frac{1}{2}(1 - \beta).$$

Поскольку $P(\chi^2 < \chi_1^2) = 1 - P(\chi^2 \geq \chi_1^2) = 1 - \frac{1}{2}(1 - \beta) = \frac{1}{2}(1 + \beta)$, то значения χ_1^2 и χ_2^2 определяются из приложения из равенств:

$$\begin{aligned} P(\chi^2 > \chi_1^2) &= \frac{1 + \beta}{2}, \\ P(\chi^2 > \chi_2^2) &= \frac{1 - \beta}{2}. \end{aligned} \quad (68)$$

Преобразуем неравенство $\chi_1^2 < \frac{n \cdot S_0^2}{\sigma^2} < \chi_2^2$ к виду $\frac{nS_0^2}{\chi_1^2} < \sigma^2 < \frac{nS_0^2}{\chi_2^2}$,

тогда условие (67) примет вид

$$P\left(\frac{nS_0^2}{\chi_2^2} < \sigma^2 < \frac{nS_0^2}{\chi_1^2}\right) = \beta, \quad (69)$$

которое определяет интервальную оценку генеральной дисперсии в случае известной генеральной средней.

На практике часто генеральная средняя m неизвестна, поэтому используется ее оценка \bar{x} и вместо S_0^2 рассматривается S^2 . В этом случае величина $\frac{nS^2}{\sigma^2}$ имеет распределение χ^2 с $k = n - 1$ степенями свободы. Оценка генеральной дисперсии находится из условия

$$P\left(\frac{nS^2}{\chi_2^2} < \sigma^2 < \frac{nS^2}{\chi_1^2}\right) = \beta. \quad (70)$$

Поскольку таблицы $P(\chi^2 > \chi_{k,\beta}^2)$ приложения 5 составлены при числе степеней свободы k от 1 до 30, при $k > 30$ полагают:

$$\chi_1^2 = \frac{1}{2}(\sqrt{2k-1} - u_\beta)^2, \quad \chi_2^2 = \frac{1}{2}(\sqrt{2k-1} + u_\beta)^2, \quad (71)$$

где $2\Phi(u_\beta) = \beta$.

При построении доверительного интервала для среднего квадратического отклонения извлекается корень квадратный из левой и правой границы доверительного интервала генеральной дисперсии.

Пример 32.

На основании выборочных наблюдений объема произведенной продукции 25 работников было установлено, что среднее квадратическое отклонение выработки за день составляет 10 единиц продукции. Предполагая, что объем произведенной продукции работником за один рабочий день имеет нормальное распределение, найти с доверительной вероятностью 0,9 границы, в которых заключены генеральные дисперсия и СКО выработки работников за один рабочий день.

Решение.

Имеем: $S_0 = 10$; $\beta = 0,9$; $\alpha_1 = (1 + \beta)/2 = 0,95$; $\alpha_2 = (1 - \beta)/2 = 0,05$; число степеней свободы $k = n - 1 = 25 - 1 = 24$.

Из приложения 5 для $\alpha_1 = 0,95$ находим $\chi_1^2 = 13,8$ и для $\alpha_2 = 0,05$ находим $\chi_2^2 = 36,4$. Таким образом,

$$\frac{25}{36,4} \cdot 10^2 < \sigma^2 < \frac{25}{13,8} \cdot 10^2, \quad 68,68 < \sigma^2 < 181,16, \quad 8,29 < \sigma < 13,46.$$

Вывод: с надежностью 0,9 дисперсия объема произведенной продукции работников за один рабочий день заключена в границах от 68,68 до 181,16 усл. ед., а среднее квадратическое отклонение – от 8,29 до 13,46 усл. ед.

Пример 33.

Решить предыдущую задачу для выборочных наблюдений объема произведенной продукции за один рабочий день 40 работников.

Решение.

При $\beta 2\Phi(u_\beta) = 0,9$, $\Phi(u_\beta) = 0,9/2 = 0,45$ и из приложения 3 находим $u_\beta = 1,645$. Число степеней свободы $k = n - 1 = 39$. Вычисляем:

$$\chi_1^2 = \frac{1}{2}(\sqrt{2 \cdot 39 - 1} - 1,645)^2 = 25,42; \quad \chi_2^2 = \frac{1}{2}(\sqrt{2 \cdot 39 - 1} + 1,645)^2 = 54,29.$$

Доверительные интервалы:

$$\frac{40 \cdot 10^2}{54,29} < \sigma^2 < \frac{40 \cdot 10^2}{25,42}, \quad 73,68 < \sigma^2 < 157,36, \quad 8,58 < \sigma < 12,54.$$

Вопросы и задания к главе

1. Дайте определение выборочного наблюдения.
2. Определите генеральную совокупность и выборку, приведите примеры.
3. Перечислите способы формирования выборки с указанием их особенностей.
4. Какая выборка называется повторной? Приведите примеры.
5. Что представляет собой бесповторная выборка? Приведите примеры.
6. Назовите числовые характеристики генеральной совокупности.
7. Дайте определение выборочных характеристик и приведите примеры.
8. Определите понятие точечной оценки и приведите примеры.
9. В чем заключается требование несмещенности точечной оценки?
10. Что означает состоятельность оценки?
11. Какие оценки называются эффективными?
12. Что можно сказать о несмещенности, состоятельности и эффективности точечных оценок генеральных средних, дисперсии и вероятности?
13. В чем сущность метода моментов?
14. Дайте определение предельной ошибки выборки. Приведите соответствующие формулы.
15. Что называется точностью (точечной) оценки?
16. Как определяется доверительная вероятность (надежность)?
17. Дайте определение доверительного интервала.
18. Как определяются доверительные интервалы для генеральных средних, дисперсии, среднего квадратического отклонения и вероятности для выборок большого и малого объема?
19. Имеются данные показателя качества X (табл. 31):

Таблица 31

x_i	15	17	20	24	25	29	30	32
n_i	12	15	20	25	14	17	13	10

Найти точечные и с доверительной вероятностью, равной 0,9, интервальные оценки генеральных среднего значения и разброса показателя качества.

20. Найти доверительный интервал для математического ожидания нормально распределенной случайной величины X при:

а) $\sigma_x = 6$; $\bar{x} = 16$; $n = 25$; $\beta = 0,95$ для повторной выборки;

б) для бесповторной выборки при тех же условиях и $N = 100$.

21. Найти доверительную вероятность для оценки математического ожидания при:

а) $\bar{x} = 20$; $n = 8$; $\sigma = 4$; $\Delta_{\bar{x}} = 2,5$;

б) $\bar{x} = 20$; $n = 8$; $S_x = 6$; $\Delta_{\bar{x}} = 2,5$;

в) $\bar{x} = 20$; $n = 8$; $\sigma = 4$; $\Delta_{\bar{x}} = 2,5$; $N = 300$; выборка бесповторная.

22. Построить доверительный интервал для дисперсии и СКО нормально распределенной случайной величины при:

а) $n = 20$; $S_x^2 = 3,1$; $\beta = 0,98$;

б) $n = 41$; $S_x^2 = 3,2$; $\beta = 0,95$.

23. Вероятность p стандартного изделия равна 0,85. Построить 95% доверительный интервал для этой вероятности: а) для выборки из 60 изделий, в которой 53 изделия оказались стандартными; б) для такой же выборки, взятой из партии, содержащей 2200 изделий.

Глава 4. Статистическое изучение взаимосвязи социально-экономических явлений

§ 1. Классификация связей между явлениями и их признаками

Выделяются следующие виды связи (зависимости).

а) **По направлению** – прямая и обратная. При прямой связи с увеличением (уменьшением) значений факторного признака результирующий признак имеет тенденцию к увеличению (уменьшению). Например, с увеличением роста вес имеет тенденцию к увеличению (рис. 21).

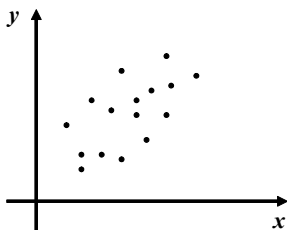


Рис. 21

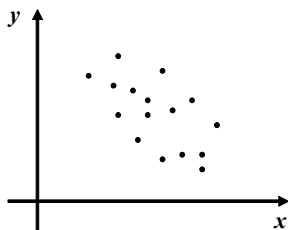


Рис. 22

При обратной связи с увеличением (уменьшением) факторного признака результирующий имеет тенденцию к уменьшению (увеличению) – рис. 22. Так, с повышением цены на товар понижается покупательная способность.

В общем случае связь может иметь колебательный характер (рис. 23, 24). В этом случае можно говорить о прямой (обратной) связи на отдельном участке изменения X . Если же рассматривается весь диапазон изменения X , то целесообразно говорить о преобладании тенденции к прямой или обратной связи.

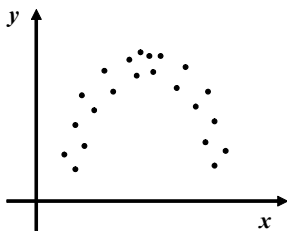


Рис. 23

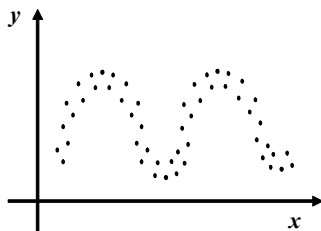


Рис. 24

б) По степени тесноты выделяется функциональная и стохастическая взаимосвязь. При **функциональной** связи определенному значению факторного признака соответствует одно-единственное значение результативного. **Стохастической** (статистической или вероятностной) называется зависимость, при которой каждому значению факторного признака соответствует определенное (условное) распределение результативного признака. Например, между урожайностью и количеством выпавших осадков имеет место прямая стохастическая зависимость.

Частным случаем стохастической зависимости является **корреляционная** зависимость. Это функциональная зависимость между значениями факторного признака X и условным математическим ожиданием $M_x(y)$ результативного признака Y .

в) По **аналитическому выражению** связи подразделяются на прямолинейные (линейные) и нелинейные (криволинейные) связи (зависимости). Если статистическую связь приближенно можно выразить уравнением прямой линии, то это линейная связь или более полно, линейная вероятностная связь, и в этом случае говорят о сглаживании экспериментальных данных по прямой; если это уравнение параболы, гиперболы, степенной, показательной и т. п. функции, то такую связь называют нелинейной (криволинейной) и говорят о сглаживании по кривой.

Уравнения аппроксимирующих линий называются **уравнениями регрессии**. Подробно об этом будет рассказано в главе 6.

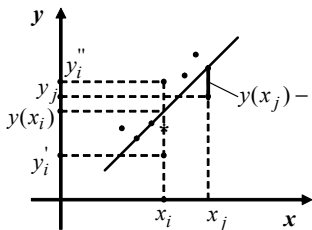


Рис. 25

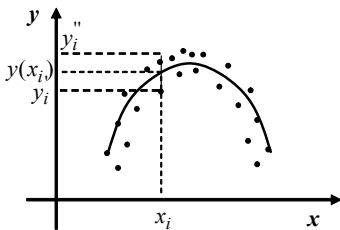


Рис. 26

На рис. 25 показана зависимость между X и Y , аппроксимируемая прямой линией, на рис. 26 зависимость аппроксимируется параболой. Причем данному x_i в выборке соответствуют 2 разных значения y_i' и y_i'' . В этом случае для удобства и упрощения каждому такому x_i ставится в соответствие средняя арифметическая всех соответствующих y_i . Для

рассматриваемого рисунка значению x_i поставится в соответствие значение $\frac{1}{2}(y_i' + y_i'')$.

На рис. 25 соответствующая точка отмечена звездочкой. Поэтому можно считать, что каждому x_i соответствует единственное y_i , которое заменяется на $y(x_i)$ согласно аппроксимирующей функциональной зависимости. Такое приближенное описание связи называется **моделированием (моделью)** данной зависимости.

На рис. 25 жирной чертой показана разность $y(x_i) - y_j$ для данного x_j . Сумма абсолютных величин таких разностей используется для оценки погрешности приближения (аппроксимации) исходных данных функциональной зависимостью (следующий параграф). Сумма квадратов указанных отклонений применяется для отыскания наилучшей аппроксимирующей функциональной зависимости, т. е. зависимости с наименьшей погрешностью приближения. Речь об этом пойдет в главе 6.

§ 2. Оценка тесноты связи

При исследовании зависимостей между признаками решаются следующие **основные задачи**:

- 1) предварительный анализ свойств моделируемой совокупности единиц;
- 2) установление факта наличия связи, определение ее направления и формы;
- 3) измерение степени тесноты между признаками;
- 4) построение регрессионной модели, т. е. нахождение аналитического выражения связи;
- 5) оценка адекватности модели, т. е. насколько она согласуется с опытными данными;
- 6) экономическая интерпретация полученной модели и ее практическое использование.

В данном пособии решение перечисленных задач рассмотрено в данной, а также в шестой главе.

Для оценки **тесноты** связи используется погрешность аппроксимации стохастической зависимости функциональной или применяются коэффициенты связи.

Погрешность вычисляется по одной из формул:

$$\Delta_1 = \frac{1}{n} \sum_{i=1}^n |y_i - y(x_i)|, \quad (72)$$

$$\Delta = \frac{1}{n} \sum_{i=1}^n \left| \frac{y_i - y(x_i)}{y_i} \right| \cdot 100\%, \quad (73)$$

где n – объем выборки; $y_i (i = \overline{1, n})$ – элементы исходной выборки резуль-
тативного признака Y , соответствующие значениям x_i фактора X ;
 $y(x_i)$ – приближенное значение для y_i , вычисленное по x_i согласно
аппроксимирующей (приближенной) связи, т. е. регрессии (рис. 25,
26). Если Δ находится в пределах 5–7%, то это свидетельствует о хо-
рошем подборе модели к исходным данным, Δ_1 при этом не превос-
ходит величины $\Delta \cdot y_{\max}$ (здесь $y_{\max} = \max y_i$). Если изначально имеет-
ся не стохастическая, а функциональная зависимость (типа прямоли-
нейной, гиперболической, параболической, степенной, показатель-
ной), Δ_1 и Δ равны нулю.

Для изучения связи между **качественными** признаками расчи-
тываются коэффициенты:

- ассоциации и контингенции;
- взаимной сопряженности Пирсона – Чупрова;
- ранговой корреляции Спирмена;
- ранговой корреляции Кенделла;
- множественной конкордации.

Для характеристики связи между **количественными** показате-
лями используются следующие коэффициенты:

- корреляции знаков Фехнера;
- эмпирическое корреляционное отношение;
- коэффициент эластичности;
- линейный коэффициент корреляции;
- коэффициент множественной корреляции;
- частные коэффициенты корреляции.

Перейдем к рассмотрению этих коэффициентов.

§ 3. Коэффициент корреляции знаков Фехнера

Применяется для описания связи между количественными пока-
зателями. Данный коэффициент связан с согласованностью **направ-
лений отклонений** индивидуальных значений факторного X и ре-
зультативного Y признаков от соответствующих средних. Суть за-
ключается в следующем. Исходные значения x_i и y_i заносятся в
столбцы таблицы. Вычисляются \bar{x} и \bar{y} . Каждое x_i сравнивается с \bar{x} .

Если $x_i \geq \bar{x}$, то в выделенном столбце таблицы ставится «+», если $x_i < \bar{x}$, то ставится «-». Аналогично все y_i сравниваются с \bar{y} , результат заносится в отдельный столбец. Затем полученные два столбца знаков «+» и «-» для X и Y сравниваются для соответствующих x_i и y_i . Для удобства результаты сравнения записываются в столбце, где ставится буква «а», если знаки столбцов для данного i совпадают, в противном случае записывается буква «b».

Коэффициент Фехнера K_ϕ вычисляется по формуле

$$K_\phi = \frac{r_a - r_b}{r_a + r_b}, \quad (74)$$

где r_a – число совпадений знаков (число букв «а» в последнем столбце); r_b – число несовпадений знаков (число букв «b» в последнем столбце).

Из (74) следует, что $-1 \leq K_\phi \leq 1$.

При совпадении знаков всех отклонений K_ϕ будет равен 1, что свидетельствует о возможном наличии прямой связи. Если же знаки всех отклонений будут разыми, то $K_\phi = -1$, что дает основание предположить наличие обратной связи.

Приведем классификацию прямой (обратной) связи (табл. 32), подобную классификации Чэддока для эмпирического корреляционного отношения, речь о котором пойдет в следующем параграфе.

Таблица 32

Величина коэффициента Фехнера	Прямая (обратная) связь
до 0,3 (-0,3)	практически отсутствует
0,3–0,5 (от -0,5 до -0,3)	
0,5–0,7 (от -0,7 до -0,5)	
0,7–1,0 (от -1,0 до -0,7)	
	слабая
	умеренная
	сильная

В связи со сказанным коэффициент Фехнера можно рассматривать как меру близости связи к прямой (обратной).

Пример 34.

Для изучения динамики реализации сельскохозяйственных продуктов на рынках города за данный период времени были получены следующие статистические данные (табл. 33).

Таблица 33

№ п/п	Наименование товара	Количество проданных товаров X	Оборот тыс. руб. Y	Знаки отклонений от \bar{x}	Знаки отклонений от \bar{y}	Совпадение a несовпадение b
1	Картофель, ц	299,8	405	+	+	a
2	Капуста, ц	26,3	106	-	-	a
3	Лук, ц	75,4	302	+	-	b
4	Свекла, ц	31,9	80	-	-	a
5	Морковь, ц	22,1	148	-	-	a
6	Огурцы, ц	26,9	123	-	-	a
7	Помидоры, ц	13,0	72	-	-	a
8	Яблоки, ц	85,1	144	+	-	b
9	Говядина, ц	106,8	1015	+	+	a
10	Свинина, ц	59,9	631	-	+	b
		$\bar{x}=74,72$	$\bar{y}=302,6$			

Что можно сказать о зависимости оборота от количества проданного товара на рынках города?

Решение.

По формуле (74) $K_\phi = \frac{7-3}{7+3} = 0,4$. Следовательно, в выборке между факторами имеет место слабая прямая связь.

§ 4. Эмпирическое и теоретическое корреляционные отношения

Эмпирическое корреляционное отношение (ЭКО) определяется выражением

$$\eta = \sqrt{\frac{S_{\text{межгр}}^2}{S_{\text{общ}}^2}}, \quad (75)$$

здесь обе дисперсии определяются для результативного признака \bar{Y} . Этот показатель используется для характеристики тесноты связи между факторами, т. е. оценки того, насколько данная связь близка к функциональной (табл. 34 классификации Чеддока).

В самом деле, из формулы (75) определения ЭКО следует, что чем больше влияние факторного признака X , тем меньше $S_{\text{межгр}}^2$ отличается от $S_{\text{общ}}^2$, и η будет приближаться к 1. И, наоборот, чем ближе η к 1, тем меньше отличие $S_{\text{межгр}}^2$ от $S_{\text{общ}}^2$, а поэтому влияние X на Y возрастает, т. е. зависимость Y от X становится сильнее.

Таблица 34

Величина ЭКО	Характер связи
до $ \pm 3 $	практически отсутствует
$ \pm 0,3 - \pm 0,5 $	слабая
$ \pm 0,5 - \pm 0,7 $	умеренная
$ \pm 0,7 - \pm 1,0 $	сильная
± 1	функциональная

Если $\eta \neq 0$ и это объясняется действительно существующей корреляционной связью между X и Y , а не следствием случайного отбора элементов в выборке (при другом отборе, например, $\eta = 0$ или имеет другой знак), то говорят, что η **значим**.

Проверка значимости ЭКО основана на том, что статистика

$$F = \frac{\eta^2 \cdot (n - m)}{(1 - \eta^2)(m - 1)}, \quad (76)$$

где m – число интервалов по группировочному признаку, имеет F -распределение **Фишера-Снедекора** с $k_1 = m - 1$ и $k_2 = n - m$ степенями свободы на уровне значимости α (приложение 6). Поэтому η **значим** (значимо отличается от нуля, т.е. с надежностью $\beta = 1 - \alpha$ можно утверждать, что между признаками имеет место связь), если $F > F_{k_1, k_2, \alpha}$, где $F_{k_1, k_2, \alpha}$ – табличное значение F -критерия на уровне значимости α при числе степеней свободы $k_1 = m - 1$ и $k_2 = n - m$.

Пример 35.

Зависимость стоимости фондов Y (тыс. руб.) от затрат на производство работ X (%) по $n = 55$ фирмам показана в табл. 35.

Таблица 35

Затраты X	Стоимость Y							Всего фирм n_i
	50– 75	75– 100	100– 125	125– 150	150– 175	175– 200	200– 225	
2–6	4	6	-	-	-	-	-	10
6–10	2	3	6	-	-	-	-	11
10–14	-	3	7	8	-	-	-	18
14–18	-	-	-	6	-	-	-	6
18–22	-	-	-	-	4	-	-	4
22–26	-	-	-	-	-	1	2	3
26–30	-	-	-	-	-	-	3	3
Всего фирм n_j	6	12	13	14	4	1	5	55

Вычислить ЭКО и проверить его значимость.

Решение.

Сначала вычислим $\bar{y}_i (i=\overline{1,7})$ – средние стоимости активной части основных фондов для i -го интервала:

$$\bar{y}_5 = \frac{1}{4}(4 \cdot 162,5) = 162,5; \quad \bar{y}_6 = \frac{1}{3}(1 \cdot 187,5 + 2 \cdot 212,5) = 204,17;$$

$$\bar{y}_7 = \frac{1}{3}(3 \cdot 212,5) = 212,5.$$

$$\bar{y}_1 = \frac{1}{10}(4 \cdot 62,5 + 6 \cdot 87,5) = 77,5; \quad \bar{y}_2 = \frac{1}{11}(2 \cdot 62,5 + 3 \cdot 87,5 + 6 \cdot 112,5) = 96,59;$$

$$\bar{y}_3 = \frac{1}{18}(3 \cdot 87,5 + 7 \cdot 112,5 + 8 \cdot 137,5) = 119,44; \quad \bar{y}_4 = \frac{1}{6}(6 \cdot 137,5) = 137,5;$$

Остальное решение оформим в виде табл. 36, в которой n_i – число фирм, отнесенных к i -му интервалу затрат, \bar{y} – средняя по всем группам (интервалам), y_j – средняя j -го интервала стоимости Y ; n_j – число фирм, попавших в j -й интервал стоимости.

Таблица 36

Интер-валы	\bar{y}_i	n_i	y_j	n_j	$y_j \cdot n_j$	$y_j^2 \cdot n_j$	$(\bar{y}_i - \bar{y})^2 n_i$
1	77,5	10	62,5	6	375	23437,5	19847
2	96,59	11	87,5	12	1050	91875	7130,33
3	119,44	18	112,5	13	1462,5	164531,25	122,62
4	137,5	6	137,5	14	1925	264687,5	1432,22
5	162,5	4	162,5	4	650	105625	6544,81
6	204,17	3	187,5	1	187,5	35156,25	20231,08
7	212,5	3	212,5	5	1062,5	225781,25	24543,61
Σ		$n = 55$		$n = 55$	6712,5	911093,75	79851,67
				$\bar{y} =$ $= 6712,5 / 55 =$ $= 122,05$	$S_{\text{общ}}^2 = 911093,75 / 55 -$ $-(122,05)^2 = 1669,14;$ $S_y = 40,86$	$S_{\text{межгр}}^2 =$ $= 79851,67 / 55$ $= 1451,85$	

$$\text{Тогда } \eta = \sqrt{\frac{S_{\text{межгр}}^2}{S_{\text{общ}}^2}} = \sqrt{\frac{1451,85}{1669,14}} = 0,933.$$

Вывод: имеется тесная связь между стоимостью активной части основных фондов и затратами на производство работ в данной выборке 55 строительных фирм.

Возникает **вопрос:** справедлив ли данный вывод для генеральной совокупности, т. е. можно ли говорить о тесной связи между данными факторами во всех строительных фирмах, подобных рассмотренным?

Ответ таков: вывод о **наличии** связи (но не о **степени тесноты**) можно обобщить на всю генеральную совокупность, если коэффициент η значим (значимо отличается от нуля). Для проверки значимости ЭКО по формуле (76) вычисляется статистика F для $\alpha = 0,05$; $\eta = 0,95$; $n = 55$ и $m = 7$.

Для рассматриваемого примера

$$F = \frac{0,95^2 \cdot (55 - 7)}{(1 - 0,95^2)(7 - 1)} = 51,26.$$

Табличное значение (приложение 6) для $\alpha = 0,05$; $k_1 = 7 - 1 = 6$; $k_2 = 55 - 7 = 48$:

$$F_{k_1, k_2, \alpha} = F_{6, 48, 0,05} \approx 2,25.$$

Так как $F > F_{6, 48, 0,05}$, то η значимо отличается от нуля и можно сделать **вывод** о наличии связи между затратами на производство строительно-монтажных работ и стоимостью активной части основных фондов во всей генеральной совокупности.

Теоретическое корреляционное отношение (ТКО) задается выражением

$$\eta_{теор} = \sqrt{\frac{S_{y(x)}^2}{S_{общ}^2}}, \quad (77)$$

где $S_{y(x)}^2$ – дисперсия, вычисляемая аналогично межгрупповой, но групповые средние \bar{y}_i заменены на $y(x_i)$, которые вычисляются согласно уравнению регрессии (§ 1 данной главы и § 4 главы 6), т. е.

$$S_{y(x)}^2 = \frac{\sum (y(x_i) - \bar{y}) \cdot n_i}{n}. \quad (78)$$

ТКО является показателем тесноты связи (табл. 34). Теоретическое корреляционное отношение по-другому называется **индексом корреляции Y по X**.

Достоинством ЭКО и ТКО является то, что они могут быть вычислены при любой форме связи между признаками. При этом ЭКО несколько занижает тесноту связи по сравнению с ТКО, но для его вычисления не нужно знать уравнение аппроксимирующей линии (уравнение регрессии).

Индекс корреляции $\eta_{теор}$ значим, если значение статистики

$$F = \frac{\eta_{теор}^2 (n - 2)}{1 - \eta_{теор}^2} \quad (79)$$

больше табличного $F_{k_1, k_2, \alpha}$ для $k = 1$ и $k = n - 2$.

Пример 36.

В главе 6 в примере 80 для данных примера 35 настоящего параграфа построено уравнение регрессии: $y(x_i) = 5,77x + 51,14$. По данному уравнению найти $y_i (i = \overline{1,7})$, вычислить $\eta_{теор}$ и для $\alpha = 0,05$ проверить его значимость.

Решение.

В примере 35 было получено $\bar{y} = 122,05$; $S_{общ}^2 = 1669,14$. Значения n_i возьмем из табл. 36, а x_i представляют собой середины интервалов изменения X . Вычисления оформим в виде табл. 37.

Таблица 37

№ интервала	n_i	x_i	$y(x_i)$	$(y(x_i) - \bar{y})^2 \cdot n_i$
1	10	4	74,22	22877,09
2	11	8	97,3	6738,19
3	18	12	120,38	50,2
4	6	16	143,46	2750,33
5	4	20	166,54	7917,44
6	3	24	189,62	13697,11
7	3	28	212,7	24652,27
Σ	55			78682,63
				$S_{y(x)}^2 = \frac{1}{n} \sum (y(x_i) - \bar{y})^2 n_i =$ $= (78682,63) / 55 = 1430,59$

Таким образом,

$$\eta_{теор}^2 = \frac{1430,59}{1669,14} = 0,857; \quad \eta_{теор} = \sqrt{0,857} = 0,926.$$

Величина коэффициента детерминации $\eta_{теор}^2 = 0,857$ показывает, что в данной выборке вариация зависимой переменной Y (стоимости активной части основных фондов) на 85,7% объясняется вариацией независимой переменной X (затратами на производство строительно-монтажных работ).

Читателям предлагается для данного примера самостоятельно найти коэффициент корреляции и убедиться, что он будет равен 0,926, т. е. коэффициенту $\eta_{теор}$, что говорит о тесной линейной связи между рассматриваемыми признаками.

Для проверки значимости индекса корреляции, согласно формуле (79), найдем

$$F = \frac{0,857 \cdot (55 - 2)}{1 - 0,857} = 318,$$

это больше, чем $F_{1; 53; 0,05} \approx 4$. Поэтому индекс корреляции значим.

§ 5. Линейный коэффициент корреляции

Для генеральной совокупности данный коэффициент определяется по формуле

$$r_{xy} = \frac{K_{xy}}{\sigma_x \cdot \sigma_y} = \frac{M[(X - m_x)(Y - m_y)]}{\sigma_x \cdot \sigma_y}, \quad (80)$$

где $K_{xy} = M[(X - m_x)(Y - m_y)]$ – ковариация X и Y ,

m_x, m_y – генеральные средние признаков X и Y соответственно,

σ_x, σ_y – генеральные средние квадратические отклонения.

Поскольку

$$M[(X - m_x)(Y - m_y)] = M[(XY - m_x Y - m_y X + m_x m_y)] = M(XY) - m_x m_y,$$

формулу (80) можно записать в виде:

$$r_{xy} = \frac{M(XY) - m_x m_y}{\sigma_x \cdot \sigma_y}, \quad (81)$$

которая при выборках x_1, x_2, \dots, x_n и y_1, y_2, \dots, y_n для выборочного коэффициента корреляции \hat{r}_{xy} примет вид

$$\hat{r}_{xy} = \frac{\overline{x \cdot y} - \bar{x} \cdot \bar{y}}{S_x \cdot S_y}, \quad (82)$$

где \bar{x} – выборочная средняя для X , \bar{y} – выборочная средняя для Y , $\overline{x \cdot y}$ – выборочная средняя произведений соответствующих значений X и Y , S_x – оценка σ_x , S_y – оценка σ_y .

Квадрат коэффициента корреляции \hat{r}_{xy}^2 равен коэффициенту детерминации η^2 , введенному в главе 2, § 5 [2, 4].

Пример 37.

В табл. 38 приведены коэффициенты эластичности отраслей промышленности Тверской области, которые показывают, на сколько процентов изменяются объем промышленного производства (k_1) и налоговые поступления (k_2) отрасли при изменении доли инвестиций на 1% (данные относятся к 1997–2000 гг.).

Таблица 38

Отрасли промышленности	k_1	k_2
Электроэнергетика	1,137	3,08
Машиностроение	0,086	1,00
Пищевая	0,092	0,31
Химическая	0,016	1,48

Требуется найти связь между коэффициентом эластичности объема производства k_1 и коэффициентом эластичности налоговых поступлений k_2 указанных отраслей промышленности Тверской области.

Решение.

В данной задаче $X = k_1$, $Y = k_2$. Находим:

$$\begin{aligned}\bar{k}_1 &= \frac{1,137 + 0,086 + 0,092 + 0,016}{4} = 0,333; \quad \bar{k}_2 = \frac{3,08 + 1,00 + 0,31 + 1,48}{4} = 1,468; \\ \overline{k_1 k_2} &= \frac{1,137 \cdot 3,08 + 0,086 \cdot 1 + 0,092 \cdot 0,31 + 0,016 \cdot 1,48}{4}; \\ S_x^2 &= \frac{1,137^2 + 0,086^2 + 0,092^2 + 0,016^2}{4} - 0,333^2 = 1,1 \quad S_x = \sqrt{1,197} = 1,094; \\ S_y^2 &= \frac{3,08^2 + 1^2 + 0,31^2 + 1,48^2}{4} - 1,468^2 = 10,617; \quad S_y = \sqrt{10,617} = 3,258; \\ \hat{r}_{xy} &= \frac{0,91 - 0,33 \cdot 1,468}{1,094 \cdot 3,258} = 0,12.\end{aligned}$$

Вывод: связь между коэффициентами в данной выборке слабая.

Квадрат коэффициента корреляции – коэффициент детерминации – в данном случае будет равен $0,12^2 = 0,0144$. Это означает, что 1,41% вариации коэффициента эластичности K_2 связано с вариацией коэффициента эластичности K_1 , остальная вариация приходится на случайные факторы, т. е. связь между факторами K_1 и K_2 в данной выборке больше случайная, чем закономерная.

Вопрос о том, возможно ли обобщение данного вывода на всю генеральную совокупность, имея, например, в виду все области РФ, будет рассмотрен в примере 39. Формулу (82) можно записать в развернутом виде:

$$\hat{r}_{xy} = \frac{n \cdot \sum xy - \sum x \sum y}{\sqrt{[n \sum y^2 - (\sum y)^2] \cdot [n \sum x^2 - (\sum x)^2]}}. \quad (83)$$

Пример 38.

Имеются выборочные данные по показателям X и Y (табл. 39).

Таблица 39

$Y \backslash X$	20–40	40–60	60–80	80–100	100–120	120–140	140–160	160–180
30–60	2							
60–90	1	4	2		4		7	
90–120		6	1	5		2		
120–150			3	8	3	5	8	4
150–180				9		3	1	2

Числа, стоящие во внутренних клетках таблицы, означают частоту n_{ij} попадания X в i -й интервал, а Y – в j -й интервал, n_i – частота попадания X в i -й интервал, n_j – частота попадания Y в j -й интервал. Найти коэффициент корреляции.

Решение.

Имеем $n = \sum_{i,j} n_{ij} = 80$.

Частоты для X и Y определяются сложением чисел n_{ij} соответственно по строкам и столбцам. Для X : $n_1 = 2, n_2 = 18, n_3 = 14, n_4 = 31, n_5 = 15$; для Y : $n_1 = 3, n_2 = 10, n_3 = 6, n_4 = 22, n_5 = 7, n_6 = 10, n_7 = 16, n_8 = 6$.

Найдем x_i и y_j – середины указанных в таблице интервалов:

$$x_1 = 45, x_2 = 75, x_3 = 105, x_4 = 135, x_5 = 165;$$

$$y_1 = 30, y_2 = 50, y_3 = 70, y_4 = 90, y_5 = 110, y_6 = 130, y_7 = 150, y_8 = 170.$$

Коэффициент \hat{r}_{xy} можно вычислить по формуле:

$$\hat{r}_{xy} = \frac{n \sum_i \sum_j n_{ij} \cdot x_i \cdot y_j - \left(\sum_i n_i x_i \right) \cdot \left(\sum_j n_j y_j \right)}{\sqrt{n \sum_i n_i (x_i)^2 - \left(\sum_i n_i x_i \right)^2} \cdot \sqrt{n \sum_j n_j (y_j)^2 - \left(\sum_j n_j y_j \right)^2}}. \quad (84)$$

Имеем:

$$\begin{aligned} n \sum_{i,j} n_{ij} \cdot x_i \cdot y_j &= 80 \cdot (2 \cdot 45 \cdot 30 + 75 \cdot 30 + 4 \cdot 75 \cdot 50 + 6 \cdot 105 \cdot 50 + 2 \cdot 75 \cdot 70 + \\ &+ 105 \cdot 70 + 3 \cdot 135 \cdot 70 + 5 \cdot 105 \cdot 90 + 8 \cdot 135 \cdot 90 + 9 \cdot 165 \cdot 90 + 4 \cdot 75 \cdot 110 + \\ &+ 3 \cdot 135 \cdot 110 + 2 \cdot 105 \cdot 130 + 5 \cdot 135 \cdot 130 + 3 \cdot 165 \cdot 130 + 7 \cdot 75 \cdot 150 + 8 \cdot 135 \cdot 150 + \\ &+ 165 \cdot 150 + 4 \cdot 135 \cdot 170 + 2 \cdot 165 \cdot 170) = 80 \cdot 1\,046\,100 = 83\,688\,000; \\ \sum_i n_i \cdot x_i &= 2 \cdot 45 + 18 \cdot 75 + 14 \cdot 105 + 31 \cdot 135 + 15 \cdot 165 = 9570; \end{aligned}$$

$$\begin{aligned}
\sum_j n_j y_j &= 3 \cdot 30 + 10 \cdot 50 + 6 \cdot 70 + 22 \cdot 90 + 7 \cdot 110 + 10 \cdot 130 + \\
&\quad + 16 \cdot 150 + 6 \cdot 170 = 8480; \\
n \sum_i n_i x_i^2 &= 80 \cdot (2 \cdot 45^2 + 18 \cdot 75^2 + 14 \cdot 105^2 + 31 \cdot 135^2 + 15 \cdot 165^2) = \\
&= 80 \cdot 1\,233\,000 = 98\,640\,000; \\
n \sum_j n_j \cdot y_j^2 &= 80 \cdot (3 \cdot 30^2 + 10 \cdot 50^2 + 6 \cdot 70^2 + 22 \cdot 90^2 + 7 \cdot 110^2 + \\
&\quad + 10 \cdot 130^2 + 16 \cdot 150^2 + 6 \cdot 170^2) = 80 \cdot 1\,022\,400 = 81\,792\,000; \\
\hat{r}_{xy} &= \frac{83\,688\,000 - 9570 \cdot 8480}{\sqrt{98\,640\,000 - (9570)^2} \cdot \sqrt{81\,792\,000 - (8480)^2}} = 0,3,
\end{aligned}$$

т. е. связь между данными показателями прямая и слабая.

Факт совпадения и несовпадения значений теоретического корреляционного отношения η и линейного коэффициента корреляции \hat{r}_{xy} используется для оценки формы связи: установлено, что если $|\eta^2 - \hat{r}_{xy}^2| = 0,1$, то гипотезу о линейной вероятностной зависимости можно считать подтвержденной.

Рассмотрим свойства коэффициента корреляции.

1. Абсолютная величина коэффициента корреляции генеральных совокупностей X и Y равна 1 тогда и только тогда, когда имеет место линейная функциональная зависимость $Y = b_1 X + b_0$, где b_0 и b_1 – постоянные величины.

Доказательство.

Пусть $Y = b_1 X + b_0$. Имеем:

$$\begin{aligned}
K_{xy} &= M[(X - m_x)(Y - m_y)] = M[(X - m_x)(b_1 X + b_0 - b_1 m_x - b_0)] = \\
&= b_1 M[(X - m_x)^2] = b_1 D_x = b_1 \sigma_x^2; \quad D_y = D[b_1 X + b_0] = D[b_1 X] + D[b_0] = b_1^2 D_x,
\end{aligned}$$

т. е. $\sigma_y = |b_1| \sigma_x$. Отсюда

$$r_{xy} = \frac{K_{xy}}{\sigma_x \sigma_y} = \frac{b_1 \sigma_x^2}{\sigma_x \cdot |b_1| \cdot \sigma_x} = \frac{b_1}{|b_1|} = \pm 1.$$

Доказательство того факта, что при $r_{xy} = \pm 1$ связь представляет собой линейную функциональную зависимость, можно посмотреть в [2].

2. Чем меньше $|\hat{r}_{xy}|$, тем больше погрешность Δ_1 .

Доказательство.

Преобразуем (72) для случая линейной аппроксимации:

$$\Delta_1 = \frac{1}{n} \sum |y_i - y(x_i)| = \frac{1}{n} \sum |y_i - b_1 x_i - b_0|,$$

при этом, как будет показано в § 4 главы 6,

$$b_1 = \hat{r}_{xy} \frac{S_y}{S_x} \text{ и } b_0 = \bar{y} - b_1 \bar{x}. \quad (85)$$

Таким образом,

$$\Delta_1 = \frac{1}{n} \sum \left| y_i - \hat{r}_{xy} \frac{S_y}{S_x} x_i - \bar{y} + \hat{r}_{xy} \frac{S_y}{S_x} \bar{x} \right| = \frac{1}{n} \sum \left| (y_i - \bar{y}) - \hat{r}_{xy} \frac{S_y}{S_x} (x_i - \bar{x}) \right|. \quad (86)$$

При $\hat{r}_{xy} = \pm 1$ имеет место линейная функциональная зависимость, и правая часть формул (72) и (86) равна нулю. А поэтому $(y_i - \bar{y}) - \hat{r}_{xy} \frac{S_y}{S_x} (x_i - \bar{x}) = 0$, т.е. $y_i - \bar{y} = \hat{r}_{xy} \frac{S_y}{S_x} (x_i - \bar{x})$.

Следовательно, коэффициент \hat{r}_{xy} , а значит, и r_{xy} можно использовать для характеристики тесноты линейной связи между признаками. Чем теснее (ближе к линейной функциональной) эта связь, тем ближе значение коэффициента корреляции к ± 1 («+» в случае возрастания, «-» в случае убывания, рис. 27, 28). Близость коэффициента к 0 означает либо отсутствие связи, либо существенное отклонение зависимости от линейной (рис. 29, 30).

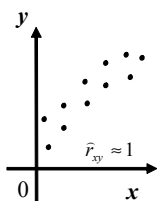


Рис. 27

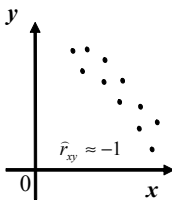


Рис. 28

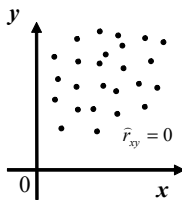


Рис. 29

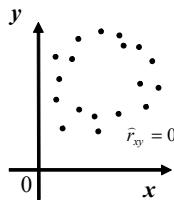


Рис. 30

При этом можно руководствоваться, например, таблицей оценок Чеддока.

3. Если все значения переменных увеличить (уменьшить) на одно и то же число или в одно и то же число раз, то величина коэффициента корреляции не изменится.

Доказательство.

Пусть X увеличилась на C_1 , Y – на C_2 , тогда

$$\begin{aligned} & M[(X + C_1 - m_{x+C_1})(Y + C_2 - m_{y+C_2})] = \\ & = M[(X + C_1 - m_x - C_1)(Y + C_2 - m_y - C_2)] = M[(X - m_x)(Y - m_y)]; \\ & \sigma(X + C_1) = \sigma(X); \quad \sigma(Y + C_2) = \sigma(Y); \end{aligned}$$

$$r_{x+c_1, y+c_2} = \frac{M[(X - m_x) \cdot (Y - m_y)]}{\sigma(X) \cdot \sigma(Y)} = r_{x,y}. \quad (87)$$

Аналогично рассматривается случай увеличения (уменьшения) переменных в данное число раз. Этот результат верен и для \hat{r}_{xy} .

4. ЭКО η и индекс корреляции $\eta_{теор}$ связаны с r_{xy} следующим образом:

$$0 \leq r_{xy} \leq \eta_{теор} \leq \eta \leq 1. \quad (88)$$

5. В случае линейной аппроксимации (линейной модели связи)

$$\eta_{теор} = |\hat{r}_{xy}|. \quad (89)$$

Для оценки **значимости** выборочного коэффициента корреляции \hat{r}_{xy} используется t -критерий Стьюдента для статистики

$$t = \hat{r}_{xy} \sqrt{\frac{n-2}{1-\hat{r}_{xy}^2}}, \quad (90)$$

где $k = n - 2$ – число степеней свободы при данном уровне значимости α и объеме выборки n . Из приложения 4 находится $t_{k;1-\alpha}$.

Выдвигается гипотеза $H_0 : r_{xy} = 0$ (r_{xy} – генеральный коэффициент корреляции равен нулю). Если H_0 отвергается, то выборочный коэффициент корреляции \hat{r}_{xy} значимо отличается от нуля и с надежностью $1 - \alpha$ можно утверждать о наличии связи между признаками во всей генеральной совокупности. В противном случае с вероятностью $1 - \alpha$ можно утверждать, что связь между признаками в генеральной совокупности отсутствует. Итак,

$|t| > t_{k;1-\alpha} - H_0$ отвергается (выборочный коэффициент корреляции значим);

$|t| \leq t_{k;1-\alpha} - H_0$ принимается (выборочный коэффициент корреляции незначим).

Пример 39.

Проверить значимость выборочного коэффициента корреляции примера 37: а) при $\alpha = 0,05$; б) $\alpha = 0,01$.

Решение.

Дано: $\hat{r}_{xy} = 0,12$; $n = 4$; $k = n - 2 = 2$; а) $\alpha = 0,05$; б) $\alpha = 0,01$.

Найдем: $t = 0,12 \sqrt{\frac{4-2}{1-0,12^2}} = 0,17$;

а) $t_{2;0,95} = 4,3$; $t < t_{2;0,95}$. б) $t_{2;0,99} = 9,92$; $t < t_{2;0,99}$.

Вывод: в обоих случаях гипотеза H_0 принимается, т. е. $r_{xy} = 0$ и с вероятностью 0,95 (0,99) можно утверждать, что и в генеральной совокупности отсутствует связь между признаками.

Пример 40.

Для условия предыдущего примера найти, каков должен быть объем выборки n , чтобы гипотеза о значимости коэффициента \hat{r}_{xy} могла быть принятой при $\alpha = 0,1$.

Решение.

Имеем

$$t = 0,12 \sqrt{\frac{n-2}{1-0,12^2}} > t_{n-2; 0,95}.$$

Возведем обе части данного неравенства в квадрат, получим:

$$0,0144 \cdot \frac{(n-2)}{1-0,12^2} > t_{n-2; 0,95}^2, \text{ т. е. } 0,0146n > 0,02922 + t_{n-2; 0,95}^2.$$

Из анализа таблицы приложения 4 видно, что это неравенство не выполняется ни при каком n .

Пример 41.

Для условия примера 39 найти значение \hat{r}_{xy} , при котором он будет значим ($\alpha = 0,05$).

Решение.

$$t = \hat{r}_{xy} \sqrt{\frac{4-2}{1-\hat{r}_{xy}^2}} > 4,3 = t_{2; 0,95}, \text{ т. е. } 2\hat{r}_{xy}^2 + (4,3\hat{r}_{xy})^2 - (4,3)^2 > 0 \text{ и в результате } \hat{r}_{xy} > 0,95.$$

§ 6. Коэффициент эластичности

Данный коэффициент показывает, на сколько процентов в среднем изменяется результативный признак при изменении факторного признака на 1%. Определяется по формуле

$$\varepsilon = b_1 \frac{x}{y}, \quad (91)$$

где $b_1 = \frac{\overline{x \cdot y} - \bar{x} \cdot \bar{y}}{S_x^2}$, \bar{x} , \bar{y} , $\overline{x \cdot y}$, S_x имеют тот же смысл, что и в определении выборочного коэффициента корреляции (82).

Коэффициент линейной корреляции и коэффициент эластичности связаны соотношением

$$\hat{r}_{xy} = \frac{S_x}{S_y} \cdot \varepsilon. \quad (92)$$

Пример 42.

На основе статистических данных экспертами оценивались вкусовые качества разных вин для контроля цены в магазинах. Данные приводятся в первых трех столбцах табл. 40. Оценить обоснованность изменения цен в зависимости от балла, поставленного экспертами данной марке вина.

Таблица 40

Марка вина	Оценка в баллах X	Цена в усл. ед. Y	$x \cdot y$	x^2
1	11	1,57	17,27	121
2	14	1,60	22,4	196
3	17	2,00	34,0	289
4	15	2,10	31,5	225
5	13	1,70	22,1	169
6	13	1,85	24,05	169
7	18	1,80	32,4	324
8	10	1,15	11,5	100
9	19	2,30	43,7	361
10	25	2,40	60,0	625
	$\bar{x} = \frac{\sum x}{10} = 15,5$	$\bar{y} = \frac{\sum y}{10} = 1,85$	$\overline{xy} = \frac{\sum xy}{10} = 29,89$	$\overline{x^2} = \frac{\sum x^2}{10} = 257,9$

Решение.

Поскольку в определении коэффициента эластичности участвуют $\overline{x \cdot y}$ и S_x^2 , в последних двух столбцах таблицы записаны значения произведений $x \cdot y$ и x^2 . Итак, находим:

$$b_1 = \frac{29,89 - 15,5 \cdot 1,85}{257,9 - 15,5^2} = 0,069;$$

$$\Theta = 0,069 \cdot \frac{15,5}{1,85} = 0,58.$$

Вывод: при увеличении оценки на 1% цена вин в среднем увеличивается на 0,58%.

§ 7. Коэффициент множественной корреляции

Характеризует тесноту линейной связи одного признака X_i с совокупностью других; в частности в случае трех признаков он вычисляется по формуле

$$R_{i,j,k} = \sqrt{\frac{r_{ij}^2 + r_{ik}^2 - 2r_{ij} \cdot r_{ik} \cdot r_{jk}}{1 - r_{jk}^2}}. \quad (93)$$

здесь r_{ij} – коэффициент парной корреляции X_i и X_j ;

r_{ik} – коэффициент парной корреляции X_i и X_k ;

r_{jk} – коэффициент парной корреляции X_j и X_k .

Для оценки значимости множественного коэффициента корреляции R используется F -распределение Фишера–Снедекора с $k_1 = m - 1$ и $k_2 = n - m$ степенями свободы (приложение 6), где n – объем выборки, m – число факторных признаков.

Вычисляется статистика

$$F = \frac{R^2(n-m)}{(1-R^2)(m-1)}, \quad (94)$$

которая сравнивается с табличным значением $F_{k_1, k_2, \alpha}$ F -критерия по правилу:

если $F > F_{k_1, k_2, \alpha}$, то R значимо отличается от нуля;

если $F \leq F_{k_1, k_2, \alpha}$, то $R = 0$.

Пример 43.

Для исследования зависимости между рентабельностью труда X_1 , объемом продукции X_2 и авансированным капиталом X_3 была произведена выборка из 20 однотипных хозяйств. Вычисленные парные коэффициенты оказались значимыми и равными: $r_{12} = 0,95$; $r_{13} = 0,3$; $r_{23} = 0,15$. Вычислить множественный коэффициент корреляции $R_{1,23}$ и оценить его значимость при $\alpha = 0,01$.

Решение.

Из (93) и (94) находим:

$$R_{1,23} = \sqrt{\frac{0,95^2 + 0,3^2 - 2 \cdot 0,95 \cdot 0,3 \cdot 0,15}{1 - 0,15^2}} = 0,96;$$

$$F = \frac{0,96^2(20-3)^2}{(1-0,96^2)(3-1)} = 100 > F_{2;17;0,01} = 3,59.$$

Вывод: между рентабельностью, с одной стороны, и объемом продукции и авансированным капиталом – с другой, существует тесная связь. В данном примере множественный коэффициент детерминации $R_{1,23}^2 = 0,96^2 = 0,92$ показывает, что вариация рентабельности на 92% объясняется вариацией объема продукции и авансированным капиталом.

§ 8. Частный коэффициент корреляции

Характеризует степень тесноты связи между двумя признаками при фиксированном значении других факторных признаков. Например, если число факторных признаков $m = 3$ и анализируется зависимость X_i от X_j при фиксированном X_k , то соответствующий коэффициент корреляции вычисляется по формуле

$$r_{ij,k} = \frac{r_{ij} - r_{ik} \cdot r_{jk}}{\sqrt{(1 - r_{ik}^2)(1 - r_{jk}^2)}}; \quad (95)$$

Доказано, что $-1 \leq r_{ij,k} \leq 1$. Формулу (95) можно использовать и для результативного признака Y_i .

Значимость частного коэффициента корреляции оценивается так же, как и коэффициента парной корреляции (§ 5), но при этом вместо n рассматривают $n' = n - m + 2$.

Пример 44.

При условии предыдущего примера найти частные коэффициенты корреляции и оценить их значимость при $\alpha = 0,05$.

Решение.

По формуле (95) находим

$$r_{12,3} = \frac{0,95^2 - 0,3 \cdot 0,15}{\sqrt{(1 - 0,3^2)(1 - 0,15^2)}} = 0,91; \quad r_{13,2} = \frac{0,3^2 - 0,95 \cdot 0,15}{\sqrt{(1 - 0,95^2)(1 - 0,15^2)}} = -0,17;$$

$$r_{23,1} = \frac{0,15^2 - 0,95 \cdot 0,3}{\sqrt{(1 - 0,95^2)(1 - 0,3^2)}} = -0,89.$$

Оценим значимость коэффициента $r_{13,2}$. Полагаем $n' = n - m + 2 = 20 - 3 + 2 = 19$. По формуле (90) находим

$$t = \frac{-0,17\sqrt{19-2}}{\sqrt{1-0,17^2}} = -0,71.$$

Из приложения 4 $t_{17;0,05} = 2,11$. Поскольку $|t| < t_{17;0,05}$, то этот коэффициент незначим. Оценим значимость $r_{12,3}$:

$$t = \frac{0,91\sqrt{19-2}}{\sqrt{1-0,91^2}} = 9,04;$$

$|t| = 9,04 > t_{17;0,05} = 2,11$, поэтому $r_{12,3}$ значим.

Для оценки $r_{23,1}$ находим

$$t = \frac{-0,89\sqrt{19-2}}{\sqrt{1-0,89^2}} = -7,98;$$

$|t| = 7,98 > t_{17;0,05} = 2,11$, следовательно, $r_{23,1}$ значим.

Сравним исходные коэффициенты парной корреляции с соответствующими частными коэффициентами (табл. 41).

Таблица 41

$r_{12} = 0,95$	$r_{12,3} = 0,91$
$r_{13} = 0,3$	$r_{13,2} = -0,17$
$r_{23} = 0,15$	$r_{23,1} = -0,89$

Наибольшему изменению подвергся коэффициент r_{23} , который изменил знак и абсолютное значение почти в 6 раз.

Вывод: между рентабельностью и объемом продукции хозяйств данного типа существует тесная линейная корреляционная зависимость. Между рентабельностью и авансированным капиталом имеет место прямая корреляционная связь (слабая), которая при устранении влияния переменной X_2 (авансированного капитала) в чистом виде находится в обратной по направлению (и очень слабой по тесноте) связи с авансированным капиталом ($r_{13,2} = -0,17$). Между объемом продукции и авансированным капиталом связь почти отсутствует, при постоянной рентабельности X_1 эта связь обратная и сильная.

§ 9. Коэффициенты ассоциации K_a и контингенции K_k

Характеризуют связь двух качественных признаков A и B . Используется табл. 42.

Таблица 42

	B	\bar{B}	
A	a	b	$a+b$
\bar{A}	c	d	$c+d$
	$a+c$	$b+d$	$a+b+c+d$

Здесь a – количество обследованных единиц, удовлетворяющих признакам A и B ; b – то же самое для A и \bar{B} ; c – то же самое для \bar{A} и B ; d – то же самое для \bar{A} и \bar{B} .

Расчеты ведутся по формулам:

$$K_a = \frac{a \cdot d - b \cdot c}{a \cdot d + b \cdot c}; \quad (96)$$

$$K_k = \frac{a \cdot d - b \cdot c}{\sqrt{(a+b)(b+d)(a+c)(c+d)}}. \quad (97)$$

Покажем, что $K_k < K_a$. Предположим противное:

$$\frac{a \cdot d - b \cdot c}{a \cdot d + b \cdot c} < \frac{a \cdot d - b \cdot c}{\sqrt{(a+b)(b+d)(a+c)(c+d)}}.$$

Числители одинаковые, сравниваем знаменатели:

$$ad + bc < \sqrt{(a+b)(b+d)(a+c)(c+d)},$$

поскольку при возведении обеих частей данного неравенства в квадрат и раскрытия скобок в правой части будет сумма 15 слагаемых, 3 из которых совпадают с левой частью.

Можно показать, что

$$-1 \leq K_a \leq 1 \quad \text{и} \quad -1 \leq K_k \leq 1,$$

причем чем они меньше отличаются от ± 1 , тем сильнее связаны между собой изучаемые признаки. Связь считается подтвержденной, если $|K_a| > 0,5$ или $|K_k| > 0,3$.

Пример 45.

Исследовалась социально-демографическая картина случайного употребления спиртных напитков в зависимости от профессии (табл. 43).

Таблица 43

Группы потребителей спиртного	Профессия		Всего
	рабочие	служащие	
Не употребляют	8	10	18
Употребляют	5	5	10
Итого:	13	15	28

На основе данного статистического материала выяснить, можно ли говорить о влиянии профессии на употребление спиртных напитков.

Решение.

По формулам (96) и (97) соответственно находим:

$$K_a = \frac{8 \cdot 5 - 10 \cdot 5}{8 \cdot 5 + 10 \cdot 5} = \frac{40 - 50}{40 + 50} \approx -0,11;$$

$$K_k = \frac{8 \cdot 5 - 10 \cdot 5}{\sqrt{(8+10)(10+5)(8+5)(5+5)}} = \frac{-10}{\sqrt{18 \cdot 15 \cdot 13 \cdot 10}} = \frac{-10}{187,35} \approx -0,05.$$

Вывод: так как $|K_a| < 0,5$ и $|K_k| < 0,3$, употребление спиртных напитков не зависит от профессии.

§ 10. Коэффициенты Пирсона K_n и Чупрова K_r

Используются для изучения связи между двумя качественными признаками, каждый из которых состоит более чем из двух групп. Вычисляются соответственно по формулам:

$$K_n = \sqrt{\frac{\varphi^2}{1 + \varphi^2}}; \quad K_r = \sqrt{\frac{\varphi^2}{\sqrt{(k_1 - 1)(k_2 - 1)}}}, \quad (98)$$

где φ^2 – показатель взаимной сопряженности;

$$\varphi^2 = \sum \frac{n_{xy}^2}{n_x \cdot n_y} - 1,$$

n_x – объемы признака X по группам,

n_y – объемы признака Y по группам,

n_{xy} – объемы выборок, относящихся к X и Y одновременно;

k_1 – число значений (групп) первого признака,

k_2 – число значений (групп) второго признака.

Можно показать, что $0 < K_n < 1$ и $0 < K_r < 1$, и использовать качественную оценку связи, как для ЭКО.

Пример 46.

Известны следующие данные о распределении рабочих предприятия по стажу и разряду работы (табл. 44).

Таблица 44

Разряд рабочего X	Стаж работы Y			Итого:
	до 5 лет	от 5 до 10	более 10	
1–2	80	30	0	110
2–4	20	60	50	130
4–6	5	40	50	95
Итого	105	130	100	335

Исследовать зависимость между стажем рабочих и их разрядом.

Решение.

$$\varphi^2 = \frac{80^2}{110} + \frac{30^2}{130} + \frac{20^2}{105} + \frac{60^2}{130} + \frac{50^2}{100} + \frac{5^2}{105} + \frac{40^2}{130} + \frac{50^2}{100} - 1 = 1,01.$$

$$K_n = \sqrt{\frac{(1,01)^2}{1 + (1,01)^2}} = 0,71; \quad K_r = \sqrt{\frac{(1,01)^2}{\sqrt{2 \cdot 2}}} = 0,72.$$

Вывод: связь между стажем рабочих и их разрядом сильная.

§ 11. Коэффициент Кендалла

В определении данного коэффициента используется понятие **ранга**. Это номер в ранжированной (расположенной по возрастанию или убыванию значимости значений признака) последовательности. Если значения признака имеют одинаковую количественную оценку, то ранг этих значений принимается равным средней арифметической номеров. Такие ранги называются **связанными**.

Прежде чем приводить формулу коэффициента Кендалла τ , укажем шаги ее построения.

1. Значения факторного признака X ранжируют по убыванию.
2. Значения результативного признака Y располагают в порядке, соответствующем значениям X .

3. Значения X нумеруют: самому большому числу присваивается номер (ранг) 1, следующему номер – 2 и т. д. (получается последовательность R_1). Так же нумеруются значения Y (получается последовательность R_2).

4. Для каждого ранга в R_2 определяется число следующих за ним больших рангов. Сумма таких чисел R^+ рассматривается как мера соответствия последовательностей рангов по X и Y .

5. Для каждого ранга в R_2 определяется число следующих за ним меньших рангов. Сумма этих чисел обозначается R^- и характеризует несоответствие последовательностей рангов по X и Y .

6. Вычисляется разность $S = R^+ - R^-$.

Ранговый коэффициент корреляции Кендалла τ определяется по формуле

$$\tau = \frac{2S}{n(n-1)}, \quad (99)$$

где n – число наблюдений.

Приведенный алгоритм удобно оформить в виде табл. 45.

Пример 47.

Имеются данные опроса работников фирмы (первые 4 столбца таблицы) по двум вопросам:

- 1) Довольны ли местом работы; 2) Устраивает ли оплата труда?

Таблица 45

№ п/п	Возраст (лет)	Процент положительных ответов		$X\%$	$Y\%$	R_1	R_2	R^+	R^-
		1-й вопрос $X\%$	2-й вопрос $Y\%$						
1	от 20 до 25	80,2	60	90,5	62	1	2	3	1
2	от 25 до 30	90	74	90	74	2	1	4	0
3	от 30 до 40	82	58	82	58	3	4	2	1
4	от 40 до 50	90,5	62	80,2	60	4	3	2	0
5	от 50 до 60	50	40	70	50	5	5	1	0
6	от 60 до 70	70	50	50	40	6	6	0	0

Связаны ли ответы на вопросы с возрастом отвечающих?

Решение.

В таблице в 4-м столбце указана ранжированная по убыванию последовательность значений факторного признака X , в 5-м столбце перечислены соответствующие значения результативного признака Y . В столбце R_1 приводятся ранги значений факторного признака (большему значению присвоен меньший ранг), в столбце R_2 – ранги результативного признака (большему значению соответствует меньший ранг).

В столбце R^+ для каждого значения y записывается число следующих за ним значений y с большими рангами.

В последнем столбце R^- для каждого y записывается число следующих значений y с меньшими рангами.

Для рассматриваемого примера $S = 13 - 2 = 11$.

$$\tau = \frac{2 \cdot 11}{6(6-1)} = \frac{22}{30} = 0,73.$$

Вывод: для данной выборки ответы на вопросы в достаточной степени характеризуют возрастные группы опрошенных.

Можно ли полученный вывод обобщить на генеральную совокупность? Для этого надо проверить значимость τ . При этом исходят из статистики

$$u = \tau \sqrt{\frac{9n(n-1)}{2(2n+5)}} \quad (100)$$

гипотезы $H_0: \tau = 0$ и критического значения $u_{1-\alpha} = \Phi^{-1}((1-\alpha)/2)$.

Используется правило:

$|u| > u_{1-\alpha} - H_0$ отвергается (τ значим); $|u| \leq u_{1-\alpha} - H_0$ принимается.

Так, для примера 47 при $\alpha = 0,05$; $n = 6$ имеем

$$u = 0,73 \sqrt{\frac{9 \cdot 6 \cdot 5}{2 \cdot 17}} = 2,06 > 1,96 = u_{1-\alpha}.$$

Вывод: на уровне значимости 0,05 (с уверенностью 0,95) можно утверждать, что в генеральной совокупности ответы на вопросы зависят от группы опрошенных.

Если в изучаемой совокупности есть **связанные ранги**, т. е. одинаковые по величине значения x (или y), то расчеты проводятся по формуле

$$\tau = \frac{S}{\sqrt{[n(n-1)/2 - V_x][n(n-1)/2 - V_y]}}, \quad (101)$$

$V_x = V_y = \frac{1}{2} \sum_i (b_i^2 - b_i)$ – число баллов, корректирующих (уменьшающих) максимальную сумму баллов за счет повторений (объединений) b_i рангов в каждом ряду;

$b_x(b_y)$ – число равных значений признака X (Y);

S – фактическая сумма баллов, в которой суммируются +1 при одинаковом порядке изменения и -1 при обратном порядке изменения.

Случаи следования одинаковых повторяющихся баллов оцениваются баллом 0 (они не учитываются при расчете).

§ 12. Коэффициент Спирмена

Используется понятия ранга: значения признаков X и Y ранжируются, вводится величина d_i^2 – квадрат разности соответствующих рангов X и Y . Коэффициент Спирмена вычисляется по формуле

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}, \quad (102)$$

где n – число наблюдений (число пар рангов). Коэффициент Спирмена удовлетворяет неравенству

$$-1 < \rho < 1,$$

кроме случаев, когда ранги X и Y расположены в обратном порядке и $\rho = -1$ либо $\rho = 1$ при равенстве соответствующих рангов X и Y .

Пример 48.

Оцените тесноту связи между интересами юношей и девушек по данным выборочного обследования первокурсников данного вуза (табл. 46, в которой R_1 и R_2 обозначают ранги признаков X и Y соответственно).

Таблица 46

№ п/п	Интересы	Процент к числу опрошенных		R_1	R_2	$d_i =$ $= R_2 - R_1$	d_i^2
		юноши $X\%$	девушки $Y\%$				
1	Музыка	97,7	92,4	1	2	1	1
2	Телепередачи	90,8	94,5	2	1	-1	1
3	Спорт	80,4	70,3	3	4	1	1
4	Чтение художественной литературы	70,2	72,4	5	3	-2	4
5	Экскурсии	72,4	62,4	4	5	1	1
6	Иностранные языки	60,3	58,2	6	6	0	0
7	Приготовление пищи	20,4	40,8	7	7	0	0
8	Живопись	18,9	20,5	8	8	0	0
9	Садоводство	10,5	12,3	9	9	0	0
10	Туризм	10	10,2	10	10	0	0
11	Отсутствие интересов	5	7	11	11	0	0

Решение.

Найдем $\rho = 1 - \frac{6 \cdot 8}{11 \cdot 120} = 0,96$.

Вывод: существует тесная связь между интересами данных девушек и юношей.

При наличии **связных** рангов ранговый коэффициент Спирмена вычисляется по формуле

$$\rho = 1 - \frac{\sum d_i^2}{\frac{1}{6}(n^3 - n) - (B_r + B_s)}, \quad (103)$$

где $B_r = \frac{1}{12} \sum (b_r^3 - b_r)$; $B_s = \frac{1}{12} \sum (b_s^3 - b_s)$; суммирование происходит по всем группам неразличимых (одинаковых) рангов признаков X и Y соответственно; b_r, b_s – число рангов, входящих в группу неразличимых рангов признаков X и Y соответственно.

Пример 49.

По результатам двух сессий были выборочно ранжированы средние баллы студентов-первокурсников экономического факультета. В табл. 47 представлены выборочные данные.

Оцените по результатам двух сессий стабильность в усвоении учебного материала данными учащимися, характеризующую отсутствием заметных колебаний по итогам успеваемости каждого студента в данных сессиях.

Таблица 47

Студенты п/п	Ранги		$d_i = s_i - r_i$	d_i^2
	1-я сессия r_i	2-я сессия s_i		
1	4,5	4	-0,5	0,25
2	4,5	5	0,5	0,25
3	3	2,5	-0,5	0,25
4	1	2,5	1,5	2,25
5	10,5	10	-0,5	0,25
6	13	12	-10	1
7	13	13,5	0,5	0,25
8	13	13,5	0,5	0,25
9	17	18	-1	1
10	16	15	-1	1
11	10,5	11	0,5	0,25
12	8	7	-1	1
13	8	8,5	0,5	0,25
14	8	8,5	0,5	0,25
15	19,5	20	0,5	0,25
16	19,5	19	-0,5	0,25
17	15	16	1	1
18	18	17	-1	1
19	2	1	1	1
20	6	6	0	0

Решение.

В первую сессию имеется 3 группы неразличимых рангов с числом рангов $b_r = 2$ (3 раза) и $b_r = 3$ (2 раза); во вторую сессию 2 группы с $b_s = 2$ (3 раза). Найдем:

$$B_r = \frac{1}{12}[(2^3 - 2) + (3^2 - 3)] = 1; \quad B_s = \frac{1}{12}(2^3 - 2) = 0,5;$$

$$\rho = 1 - \frac{11}{\frac{1}{6}(8000 - 20) - (1 + 0,5)} = 0,99.$$

Вывод: по результатам двух сессий наблюдается фактически неизменный результат усвоения учебного материала данными двадцатью студентами.

При проверке значимости ρ исходят из того, что в случае справедливости $H_0: \rho = 0$ при $n > 10$ статистика

$$t = \frac{\rho\sqrt{n-2}}{\sqrt{1-\rho^2}} \quad (104)$$

имеет t -распределение Стьюдента с $k = n - 2$ степенями свободы, поэтому: $|t| > t_{k;1-\alpha} - H_0$ отвергается (ρ значим); $|t| \leq t_{k;1-\alpha} - H_0$ принимается.

Пример 50.

Для примера 49 при $\alpha = 0,05$ проверить, можно ли обобщить полученный результат на генеральную совокупность.

Решение.

Имеем: $\rho = 0,99$; $k = 18$; $t_{18;0,95} = 2,1$; $t = \frac{0,99 \cdot \sqrt{8}}{\sqrt{1 - 0,99^2}} = 20$; $t > t_{18;0,95}$, сле-

довательно, ρ значим, и на уровне значимости 0,05 можно утверждать о неизменном результате усвоения учебного материала в данных двух сессиях всеми студентами-первокурсниками экономического факультета данного вуза.

При достаточно большом числе наблюдений между коэффициентами Кендалла и Спирмена существует следующее соотношение:

$$\rho \approx \frac{3}{2} \tau. \quad (105)$$

§ 13. Коэффициент конкордации

Коэффициент конкордации (согласованности) рангов Кендалла вычисляется по формуле

$$W = \frac{12 \sum_{i=1}^n D_i^2}{m^2 (n^3 - n)}, \quad (106)$$

где $D_i = \sum_j R_{ij} - \frac{1}{n} \sum_j \sum_i R_{ij}$,

m – количество порядковых переменных,

n – число наблюдений (объектов),

R_{ij} – ранги ($i = \overline{1, n}$; $j = \overline{1, m}$),

причем $0 \leq W \leq 1$ и $W = 1$ при совпадении всех ранжировок.

Формула (106) применяется в случае отсутствия связных рангов.

Пример 51.

Группа из 6 экспертов оценивает ущерб, нанесенный талыми водами жителям верхних этажей (в усл. ед.) в 50 однотипных домах поселка. Выборочные данные по 8 домам помещены в табл. 48.

Таблица 48

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

Установить, имеет ли место согласованность во мнении экспертов относительно нанесенного ущерба данным 8 домам.

Решение.

Решение оформим в расчетной табл. 49, в которой в столбцах с номерами домов указаны соответствующие ранги причиненного ущерба, причем R_{ij} – ранг ущерба j -го дома по мнению i -го эксперта, ранги расположены по возрастанию степени причиненного ущерба.

В данном примере дома – это объекты, эксперты – это факторы (признаки).

Таблица 49

Дома, i	Эксперты, j						Сумма рангов		
	1	2	3	4	5	6	$\sum_{j=1}^6 R_{ij}$	D_i	D_i^2
1	3	2	6	5	3	1	20	20-27=-7	49
2	2	1	4	2	1	2	12	12-27=-15	225
3	6	8	1	4	8	5	32	32-27=5	25
4	8	6	7	7	6	6	40	40-27=13	169
5	1	4	2	6	4	3	20	20-27=-7	49
6	4	3	3	1	2	4	17	17-27=-10	100
7	5	7	5	3	7	8	35	35-27=8	64
8	7	5	8	8	5	7	40	40-27=13	169
Итого:	36	36	36	36	36	36	216		850

Среднее значение сумм рангов:

$$\frac{1}{n} \sum_{i=1}^8 \sum_{j=1}^6 R_{ij} = 216/8 = 27;$$

в предпоследнем столбце таблицы указаны разности D_i , а в последнем – их квадраты, сумма которых равна 850. Найдем

$$W = \frac{12 \cdot 850}{6^2(8^3 - 8)} = 0,56.$$

Вывод: мнения данных экспертов умеренно согласованны в оценке ущерба для указанных 8 домов.

Чтобы обобщить этот результат для произвольного количества домов данного типа (генеральная совокупность имеет 50 таких домов), используется χ^2 -критерий Пирсона, согласно которому вычисляется статистика ($n > 7$):

$$\chi^2 = \frac{12 \cdot \sum_{i=1}^n D_i^2}{m \cdot n \cdot (n+1)}, \quad (107)$$

для рассмотренного примера принимающая значение

$$\chi^2 = \frac{12 \cdot 850}{6 \cdot 8 \cdot 9} = 23,6.$$

При $\alpha = 0,05$ и $k = n - 1 = 7 - 1 = 6$ из приложения 5 находим $\chi_{6;0,05} = 12,6$, что подтверждает значимость коэффициента конкордации и свидетельствует с 95-процентной гарантией о наличии согласованности в оценках данных экспертов по ущербу, нанесенному всем пострадавшим 50 однотипным домам привокзального квартала.

В случае **связных** рангов коэффициент конкордации вычисляется по формуле

$$W = \frac{12 \sum_{i=1}^n D_i^2}{m^2(n^3 - n) - m \sum_{i=1}^m B_i}, \quad (108)$$

где $B_i = 1/12 \sum_{l=1}^{k_i} (t_l^3 - t_l)$ – характеристика связанности рангов по i -й переменной, t_l – количество рангов в l -й связке, k_i – число связей i -й переменной.

При оценке значимости W , вычисленного по формуле (108), используется статистика:

$$\chi^2 = \frac{12 \sum_{i=1}^n D_i^2}{m \cdot n(n+1) - \frac{1}{n-1} \sum_{i=1}^m B_i}, \quad (109)$$

представляющая собой распределение χ^2 с $k = n - 1$ степенями свободы. Можно показать, что в случае связанных рангов $-1 \leq W \leq 1$.

Пример 52.

Среди участвовавших в конкурсе красоты девушек произвольно были выбраны 8. Их оценки в баллах, выставленные шестью экспертами, представлены в табл. 50. При $\alpha = 0,01$ оценить согласованность мнений экспертов.

Таблица 50

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

Решение.

Ранги оценок, расположенные по возрастанию значений, представлены в табл. 51.

Таблица 51

Девушки, i	Эксперты, j						Сумма рангов		
	1	2	3	4	5	6	$\sum_{j=1}^6 R_{ij}$	D_i	D_i^2
1	4	7	1	2	4	7	25	-2	4
2	5,5	7	3,5	6,5	6	5,5	35	7	49
3	2	1	7	5	7	2	28	1	1
4	5,5	3	3,5	2	2	1	17	-10	100
5	8	5	6	2	3	5,5	29,5	2,5	6,25
6	3	6	5	8	5	3	30	5	25
7	1	3	2	4	1	8	19	-6	36
8	7	3	7	6,5	8	4	35,5	8,5	72,25
Итого:	36	36	36	36	36	36	216		293,5

Среднее значение сумм рангов равно $216/8=27$; $\sum_{i=1}^8 D_i^2 = 293,5$, и поскольку имеются связанные ранги,

$$W = \frac{12 \cdot 293,5}{6^2 \cdot (8^3 - 8) - 6 \sum_i B_i},$$

где $\sum_i B_i = \frac{1}{12} \cdot (6 \cdot (2^3 - 2) + 2 \cdot (3^3 - 3)) = 7$.

Таким образом, $W \approx 0,2$, что говорит о слабой связи (слабой согласованности) между мнениями экспертов.

$$\text{По формуле (109)} \quad \chi^2 = \frac{12 \cdot 293,5}{6 \cdot 8 \cdot 9 - 7 \cdot 6 / 7} = 8,27 < \chi_{7,0,01}^2 = 18,5.$$

Вывод: на уровне значимости $\alpha=0,01$ можно утверждать, что коэффициент конкордации генеральной совокупности равен нулю, а это означает, что в случае репрезентативной выборки с вероятностью 0,99 можно утверждать об отсутствии согласованности мнений экспертов.

Вопросы и задания к главе

1. В чем состоит отличие между стохастической и функциональной связью? Приведите примеры.
2. Как определяется корреляционная связь? Приведите примеры такой связи.
3. Как классифицируется связь между явлениями и их признаками:
а) по направлению; б) по степени тесноты; в) по аналитическому выражению?
4. Какие основные задачи решаются при изучении зависимостей между признаками?
5. Дайте определение коэффициента корреляции знаков Фехнера. В каком диапазоне он изменяется и что характеризует?
6. Как определяется эмпирическое и теоретическое корреляционные отношения? Что они характеризуют?
7. Какую роль играет индекс корреляции?
8. Как измерить долю общей вариации результативного признака, которая объясняется влиянием вариации факторного признака?
9. Как определяется и для чего используется коэффициент эластичности?
10. Приведите определение линейного коэффициента корреляции. Что он характеризует?
11. Чему будет равен коэффициент корреляции, если в случае линейной зависимости 60% вариации результативного признака объясняется влиянием факторного признака?
12. Перечислите его свойства.
13. Что характеризуют коэффициенты множественной корреляции и частные коэффициенты корреляции?
14. Дайте определение коэффициентам ассоциации и контингенции. Когда они используются?

15. В каких случаях используются коэффициенты Пирсона и Чупрова?
16. Дайте определение ранга. Какие ранги называются связными?
17. Как определяется ранговый коэффициент Кендалла? В каких случаях он применяется?
18. Приведите определение коэффициента Спирмена. Когда он используется?
19. Как связаны при достаточно большом числе наблюдений коэффициенты Кендалла и Спирмена?
20. Как определяется и в каких случаях используется коэффициент множественной корреляции?
21. В таблице 52 приводятся выборочные данные по стажу работы X и производительности труда Y для 20 работников данного предприятия.

Таблица 52

x_i лет	2	2,5	3	3	3,5	4	4,5	4,5	5	7
y_i усл. ед.	12	12	17	10	10	20	15	15	10	25
x_i лет	10	10	15	16	20	20	20	25	27	29
y_i усл. ед.	20	22	20	21	25	15	17	10	20	22

Используя разные коэффициенты связи, оцените данную зависимость.

22. Для задания № 21 найти коэффициент эластичности. Какой вывод можно сделать?
23. В табл. 53 представлены выборочные данные результатов тестирования (в баллах) первокурсников данного факультета по двум предметам A и B .

Таблица 53

Тест	Студенты									
	1	2	3	4	5	6	7	8	9	10
A	22	10	40	8	30	29	14	35	40	27
B	32	24	52	25	69	57	28	47	30	22

Вычислить ранговые коэффициенты корреляции Спирмена и Кендалла между результатами тестирования и для $\alpha = 0,05$ оценить их значимость.

24. Выборочные результаты опроса 100 жителей данного квартала приведены в табл. 54.

Таблица 54

Возрастные категории	Удовлетворенность уровнем жизни		Итого:
	Вполне удовлетворен	Не удовлетворен	
от 20 до 35 лет	30	40	70
от 35 до 60 лет	10	20	30
Итого:	40	60	100

Рассчитайте коэффициенты ассоциации и контингенции. Какие выводы можно сделать?

25. Для выявления неиспользованных возможностей и повышения качества работы на предприятии «Лангуста» были приглашены 4 независимых эксперта, которые оценили деятельность данного предприятия по 8 показателям (производительность труда, рентабельность, средняя заработная плата и т. д.) следующим образом (табл. 55, в которой оценки даны в условных единицах).

Таблица 55

Эксперты	Показатели							
	1	2	3	4	5	6	7	8
1	20	10	30	4	2	41	22	15
2	18	12	32	5	2	36	23	10
3	20	9	30	3	3	51	22	8
4	16	10	30	4	3	58	24	14

Определить с помощью коэффициента конкордации Кендалла, насколько согласованы данные оценки экспертов, а также будут ли существенные различия, если продолжить список оцениваемых показателей, характеризующих деятельность данного предприятия и подобных ему. Положить $\alpha = 0,05$.

26. В табл. 56 указаны в усл. ед. стоимости перевозки единицы товара из пунктов отправления $A_i (i = \overline{1,3})$ в пункты назначения $B_j (j = \overline{1,8})$.

Таблица 56

$B_j \backslash A_i$	B_1	B_2	B_3	B_4	B_5	B_6	B_7	B_8
A_1	7	8	5	6	9	7	7	6
A_2	8	6	6	7	8	5	6	7
A_3	8	9	6	5	6	9	10	7

Имеется ли существенная разница между стоимостью перевозки: а) из разных пунктов отправления; б) в разные пункты назначения? Положить $\alpha = 0,01$; $\alpha = 0,02$; $\alpha = 0,05$.

27. Структурная матрица торговли восьми стран S_1, S_2, \dots, S_8 имеет вид табл. 57.

Таблица 57

0,1	0,2	0,1	0,15	0,1	0	0,7	0,2
0,1	0,1	0,3	0,15	0,1	0,2	0,1	0,08
0,05	0,1	0,1	0,1	0	0,2	0,03	0
0,3	0,15	0,05	0	0	0	0,3	0,02
0,2	0,15	0,05	0,1	0,3	0,1	0,1	0,5
0,05	0,05	0,2	0	0,2	0,3	0,2	0,1
0,1	0,05	0,2	0,3	0,15	0,2	0	0,05
0,1	0,2	0	0,2	0,15	0	0,2	0,05

Где a_{ij} – доля национального дохода, которую страна S_j тратит на покупку товаров у страны S_i . Выяснить, есть ли существенное различие в структуре торговли этих стран, если продолжить список стран, у которых данные 8 стран закупают товары. Уровень значимости α взять, как в задании № 26.

Глава 5. Проверка статистических гипотез

§ 1. Статистические гипотезы. Критерии значимости

Основная задача: как обобщить результаты, полученные для вариационного ряда, на всю генеральную совокупность? Для этого выдвигаются статистические гипотезы, которые затем проверяются по определенным правилам.

Статистической называется гипотеза, выдвигаемая на основе статистического материала:

1) о виде неизвестного распределения; 2) о параметрах распределения (гипотезы второго типа называются **параметрическими**). Например, гипотеза о равенстве математического ожидания случайной величины определенному значению.

Выдвинутую гипотезу называют нулевой (основной) и обозначают H_0 . Гипотезу, противоречащую нулевой, называют конкурирующей (альтернативной) и обозначают H_1 . Конкурирующие гипотезы могут быть трех типов. Например, для $H_0: m_x = c$ могут применяться следующие альтернативные гипотезы:

1) $H_1: m_x > c$ (« m_x больше, чем c » – положительное отклонение оцениваемого параметра m_x от гипотетического значения c);

2) $H_1: m_x < c$ (« m_x меньше, чем c » – отрицательное отклонение оцениваемого параметра от гипотетического значения);

3) $H_1: m_x \neq c$ (« m_x не равно c » – отклонение оцениваемого параметра от гипотетического значения).

Правила, с помощью которых проверяются статистические гипотезы, называются **критериями значимости**, а критерии для проверки гипотез о виде закона распределения называют обычно **критериями согласия**. Критерии значимости и критерии согласия представляют неравенства вида:

$$K_{расч} \leq K_{кр}, \quad (110)$$

где $K_{расч}$ и $K_{кр}$ – соответственно расчетное (фактическое) и критическое (табличное) значения статистики критерия. **Статистика** – некоторая функция K исходных данных x_1, \dots, x_n , представляющая собой случайную величину, точное или приближенное распределение которой известно в предположении справедливости нулевой гипотезы H_0 .

При выполнении неравенства (110) принимается H_0 , т. е. она **согласуется** с опытными данными, но при этом может быть как истинной, так и ложной. В случае ее ложности имеет место ошибка II рода. При невыполнении (110) H_0 отвергается в пользу H_1 , т. е. H_0 не согла-

судется с опытными данными, но при этом она может быть как ложной, так и истинной. В случае ее истинности совершается ошибка I рода.

Так же, как и построение доверительных интервалов, построение критериев значимости основано на знании законов распределения оценок (статистик). Для построения критерия значимости весь диапазон изменения величины K , относительно которой проверяется H_0 , разбивается на две области: **область допустимых значений** (ОДЗ) и **критическую область** (КО). При попадании проверяемой величины в область допустимых значений принимается H_0 , а при ее попадании в критическую область H_0 отвергается в пользу альтернативной гипотезы H_1 . Области допустимых значений соответствует вероятность β , а критической области – вероятность $\alpha = 1 - \beta$. Вероятность α называется **уровнем значимости критерия**; это есть вероятность попадания в критическую область в случае, когда справедлива H_0 , т. е. вероятность совершения ошибки первого рода $p(I)$: $\alpha = p(I)$. Эти вероятности на графике плотности распределения случайной величины K численно равны площади под кривой распределения в соответствующих границах (рис. 31–33).

Принятие (отклонение) основной гипотезы H_0 зависит от альтернативной гипотезы H_1 , уровня значимости α и рассматриваемой выборки (в частности, от ее объема).

Уровень значимости выражают в процентах или долях единицы. Обычно используют значения $\alpha = 5\%$ (вероятность попадания в критическую область равна пяти процентам), 2% или 1% . Чем меньше α , тем меньше вероятность ошибки I рода. В то же время при $\alpha \rightarrow 0$ область допустимых значений стремится к бесконечности, в результате чего увеличивается вероятность ошибки II рода $\gamma = p(II)$. На практике вместо $p(II)$ часто используют вероятность $1 - p(II)$, называемую **мощностью критерия**. Мощность критерия есть вероятность отвергнуть H_0 , когда она не верна, а верна H_1 .

Вероятность ошибки второго рода зависит от вида критической области, которую выбирают так, чтобы вероятность ошибки второго рода была минимальной, а мощность критерия – максимальной.

Таким образом, с нулевой и альтернативной гипотезами связаны четыре вероятности:

- 1) вероятность принять H_0 , когда она верна (доверительная вероятность β);
- 2) вероятность отвергнуть H_0 , когда она верна, – уровень значимости критерия (вероятность ошибки I рода α);

- 3) вероятность принять H_0 , когда верна H_1 , – вероятность ошибки II рода γ ;
 4) вероятность отвергнуть H_0 , когда она неверна, – мощность критерия $1 - \gamma$.

Схематично это показано в табл. 58

Таблица 58

$P(H_0/H_0)=\beta$ – доверительная вероятность	$P(H_1/H_0)=\alpha$ – уровень значимости
$P(H_0/H_1)=\gamma$ – вероятность ошибки II рода	$P(H_1/H_1)=1-\gamma$ – мощность критерия

При этом $\beta + \alpha = 1$ и $\alpha + \gamma \leq 1$.

Экономическая трактовка этих вероятностей: α – риск поставщика, связанный с браковкой по результатам выборочного контроля изделий всей партии; γ – риск потребителя, связанный с принятием по анализу выборки партии, не удовлетворяющей стандарту; β – вероятность принять по выборке партию, удовлетворяющую стандарту; $1 - \gamma$ – вероятность забраковать по выборке всю партию, не удовлетворяющую стандарту.

Каждому виду альтернативной гипотезы соответствует определенная критическая область, обеспечивающая минимум вероятности ошибки II рода (максимум мощности критерия). Альтернативной гипотезе H_1 – положительное отклонение проверяемого параметра относительно гипотетического значения – соответствует правосторонняя критическая область (рис. 31). На рисунках 31–33 статистика имеет нормальное распределение, другие статистики представлены в приложениях 4–6.

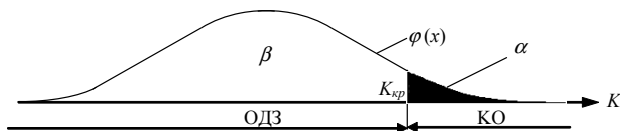


Рис. 31

Гипотезе H_1 – отрицательное отклонение проверяемого параметра – соответствует левосторонняя критическая область (рис. 32).

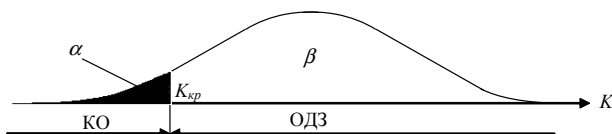


Рис. 32

Наконец, гипотезе H_1 – отклонение оцениваемого параметра по абсолютной величине – соответствует двусторонняя критическая область (рис. 33).

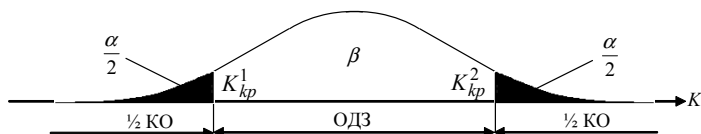


Рис. 33

В дальнейшем случайную величину K будем обозначать через u , если она имеет нормальное распределение, f , если она распределена по закону Фишера – Снедекора, t – по закону Стьюдента, χ^2 – по закону «хи квадрат» и т. д.

Замечание. Если для данной выборки гипотеза H_0 отклоняется, а вы уверены, что эта гипотеза верна, рассмотрите другую выборку, измените α или увеличьте объем выборки. Если результат будет тот же, попробуйте воспользоваться другими критериями.

Внимание! При проверке параметрических гипотез в данном учебном пособии рассматриваются нормально распределенные СВ (признаки).

§ 2. Проверка статистической гипотезы о равенстве двух дисперсий

Пусть \tilde{S}_1^2 и \tilde{S}_2^2 – оценки дисперсий (исправленные) двух случайных величин, найденные по независимым выборкам объемом n_1 , n_2 , и $\tilde{S}_1^2 > \tilde{S}_2^2$. Требуется проверить гипотезу $H_0: \sigma_1^2 = \sigma_2^2$ (о равенстве генеральных дисперсий) при альтернативной гипотезе $H_1: \sigma_1^2 > \sigma_2^2$ (генеральная дисперсия первой случайной величины больше генеральной дисперсии второй случайной величины).

Для принятия H_0 используется **F-критерий Фишера–Снедекора**:

$$f = \frac{\tilde{S}_1^2}{\tilde{S}_2^2} \leq f_{kp} = f_{n_1-1, n_2-1, \alpha}, \quad (111)$$

где $K = f = \frac{\tilde{S}_1^2}{\tilde{S}_2^2}$ – статистика, $K_{kp} = f_{kp} = f_{n_1-1, n_2-1, \alpha}$, f_{kp} – критическое значение критерия, равное табличному значению $f_{n_1-1, n_2-1, \alpha}$ распределения Фишера со степенями свободы $n_1 - 1$ и $n_2 - 1$ для уровня значимости α (приложение 6). Имеем правостороннюю критическую область.

Для $H_1: \sigma_1^2 < \sigma_2^2$ H_0 отвергается, если $f < f_{n_1-1, n_2-1, 1-\alpha}$ (левосторонняя критическая область). Для $H_1: \sigma_1^2 \neq \sigma_2^2$ H_0 отклоняется, если $f < f_{n_1-1, n_2-1, 1-\alpha/2}$ либо $f > f_{n_1-1, n_2-1, \alpha/2}$ (двусторонняя критическая область).

Для F -критерия справедливо соотношение

$$f_{n_1-1, n_2-1, 1-\alpha/2} = \frac{1}{f_{n_1-1, n_2-1, \alpha/2}}, \quad (112)$$

здесь $f_{n_1-1, n_2-1, 1-\alpha/2} = f_{kp_1}$ – левая (меньшая единицы) граница, $f_{n_1-1, n_2-1, \alpha/2} = f_{kp_2}$ – правая (большая единицы) граница.

По таблицам F -критерия можно найти лишь правую границу.

Пример 53.

Для проверки качества выпускаемой продукции двумя контейнерами были произведены соответственно две выборки со следующими данными: $n_1=12$; $n_2=15$. Для $\tilde{S}_1^2 = 11,51$; $\tilde{S}_2^2 = 6,7$; $\alpha = 0,05$ проверить $H_0: \sigma_1^2 = \sigma_2^2$ при $H_1: \sigma_1^2 > \sigma_2^2$. Указать критическую область.

Решение.

Воспользуемся критерием (111):

$$f = \frac{\tilde{S}_1^2}{\tilde{S}_2^2} = \frac{11,51}{6,7} = 1,72; \quad f_{kp} = f_{11;14;0,05} = 2,56; \quad f = 1,72 < f_{kp} = 2,56.$$

Критерий (111) выполняется, следовательно, гипотеза H_0 принимается с уровнем значимости 0,05 (с доверительной вероятностью 0,95). Критическая область: $f > 2,56$.

Пример 54.

При оценке прибыли двух предприятий, получаемой на протяжении нескольких лет, были произведены соответствующие выборки со следующими данными: $n_1 = 16$; $n_2 = 14$; $\tilde{S}_1^2 = 12,52$; $\tilde{S}_2^2 = 10,35$. Для $\alpha = 0,1$ проверить $H_0: D_1 = D_2$ (равенство генеральных дисперсий) при $H_1: D_1 \neq D_2$. (неравенство генеральных дисперсий).

Решение.

$$\text{Найдем } f = \frac{\tilde{S}_1^2}{\tilde{S}_2^2} = \frac{12,52}{10,35} = 1,21; \quad f_{кр_2} = f_{n_1-1; n_2-1; \alpha/2} = f_{15; 13; 0,05} = 2,53;$$

с учетом (112):

$$f_{кр_1} = f_{n_1-1; n_2-1; 1-\alpha/2} = 1/2,53 = 0,4.$$

В случае отклонения H_0 должно быть либо $f < f_{кр_1}$, либо $f > f_{кр_2}$. Критическая область: $f < 0,4$ либо $f > 2,53$. Однако $0,4 < 1,21 < 2,53$. Поэтому на уровне значимости 0,1 принимается H_0 .

Пример 55.

Исходя из данных примера 54 для выполнения $H_1 : D_1 \neq D_2$ найти

- а) минимальное \tilde{S}_1^2 ; б) максимальное \tilde{S}_2^2 ;
в) максимальное n_1 ; г) максимальное n_2 .

Решение.

а) $f = \frac{\tilde{S}_1^2}{12,25} > 2,53$, т. е. $\tilde{S}_1^2 > 30,9925$; если точность до 4-го знака,

можно считать минимальным значением $\tilde{S}_1^2 = 30,9926$;

б) $f = \frac{10,25}{\tilde{S}_2^2} > 2,53$, т. е. $\tilde{S}_2^2 < 0,2468$; аналогично можно считать максимальным значением $\tilde{S}_2^2 = 0,2467$.

в) Из приложения 6 следует, что даже при $n_1 = \infty$ $f_{кр_2} = 2,21$; и это существенно больше, чем $f = 1,21$, а $f_{кр_1} = 1/2,21 = 0,45 < f = 1,21$, т. е. H_1 не проходит ни при одном n_1 .

г) Аналогично п. в) даже при $n_2 = \infty$ $f_{кр_2} = 1,67$, а $f_{кр_1} = 1/1,67 = 0,6$, поэтому при любом n_2 H_1 не проходит.

§ 3. Проверка гипотезы о равенстве дисперсии гипотетическому значению

Проверяется гипотеза $H_0 : \sigma_x^2 = \sigma_0^2$, при альтернативной гипотезе $H_1 : \sigma_x^2 > \sigma_0^2$. **Критерий Пирсона** для принятия H_0 :

$$\chi^2 = \frac{\tilde{S}_x^2(n-1)}{\sigma_0^2} \leq \chi_{кр}^2 = \chi_{n-1, \alpha}^2, \quad (113)$$

где $\chi_{кр}^2$ – критическое значение критерия, равное табличному значению $\chi_{n-1, \alpha}^2$ распределения χ^2 Пирсона для числа степеней свободы $k = n - 1$ и уровня значимости α .

Замечание. В левой части (113) может использоваться $n \cdot S_x^2 / \sigma_0^2$.

Если неравенство (113) выполняется с противоположным знаком, H_0 отвергается в пользу H_1 . При $H_1: \sigma_x^2 < \sigma_0^2$ H_0 отклоняется, если $\chi^2 > \chi_{n-1,1-\alpha}^2$. В случае $H_1: \sigma_x^2 \neq \sigma_0^2$ гипотеза H_0 отклоняется, если $\chi^2 > \chi_{n-1,\alpha/2}^2$ либо $\chi^2 < \chi_{n-1,1-\alpha/2}^2$.

Пример 56.

Проверить $H_0: \sigma_x^2 = \sigma_0^2$ при $n = 20$; $\tilde{S}_x^2 = 16,8$; $\sigma_0^2 = 14$; $\alpha = 0,01$; $H_1: \sigma_x^2 > \sigma_0^2$.

Решение.

$$\frac{\tilde{S}_x^2(n-1)}{\sigma_0^2} = \frac{16,8 \cdot 19}{14} = 22,8; \quad \chi_{19;0,01}^2 = 36,2,$$

гипотеза H_0 принимается с уровнем значимости 1% (надежность 99%).

Пример 57.

Можно ли по данной выборке утверждать, что детали, изготавливаемые на данном заводе, имеют стандартное отклонение, отличное от 1,15 (усл. ед.), если $n = 20$; $\tilde{S}_x = 1,2$; $\alpha = 0,02$?

Решение.

$$\chi^2 = \frac{1,2^2 \cdot (20-1)}{1,15^2} = 20,69; \quad \chi_{19;0,99}^2 = 7,63 < \chi^2 < 36,2 = \chi_{19;0,01}^2 = \chi_{n-1,\alpha/2}^2.$$

Следовательно, гипотеза $H_0: \sigma_x = 1,15$ усл. ед. принимается на уровне значимости 0,02 (с надежностью 0,98).

§ 4. Проверка гипотезы о равенстве двух математических ожиданий при известных дисперсиях

Пусть n_1 и n_2 – объемы двух выборок. Найдены средние значения \bar{x} , \bar{y} . Известны генеральные дисперсии D_x и D_y . Проверяется гипотеза $H_0: m_x = m_y$ (равенство генеральных средних) при $H_1: m_x \neq m_y$ (неравенство генеральных средних).

Используется статистика:

$$u = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{D_x}{n_1} + \frac{D_y}{n_2}}}. \quad (114)$$

Критерий для принятия H_0 :

$$|u| \leq u_{(1-\alpha)/2}, \quad (115)$$

где $u_{(1-\alpha)/2}$ – аргумент функции Лапласа, для которого:

$$\Phi(u_{(1-\alpha)/2}) = (1-\alpha)/2.$$

При $H_1: m_x > m_y$ или $H_1: m_y > m_x$ в правой части неравенства (115) вместо $u_{(1-\alpha)/2}$ следует записывать $u_{0,5-\alpha}$ – аргумент функции Лапласа, для которого $\Phi(u_{0,5-\alpha}) = 0,5 - \alpha$.

При этом если $u > u_{0,5-\alpha}$ то принимается $H_1: m_x > m_y$, если $u < -u_{0,5-\alpha}$ то принимается $H_1: m_y > m_x$.

Замечание. Критерий применяется для нормальных распределений X и Y при любых объемах выборки; для других законов – при $n_1, n_2 > 30 - 40$.

Пример 58.

Для проверки эффективности использования новой технологии отобраны 2 группы рабочих. В первой группе численностью 60 человек, где применялась новая технология, среднее значение выработки составило 95 изделий, во второй группе численностью 70 человек средняя выработка оказалась равной 88 изделий. Известно, что дисперсии первой и второй выборок соответственно равны 90 и 74. На уровне значимости $\alpha = 0,05$ выяснить влияние новой технологии на среднюю производительность.

Решение.

Имеем: $n_1 = 60$, $n_2 = 70$, $\bar{x} = 95$, $\bar{y} = 88$, $D_x = 90$, $D_y = 74$, $\alpha = 0,05$.

Найдем: $u = \frac{95 - 88}{\sqrt{90/60 + 74/70}} = 4,38$;

$\Phi(u_{(1-\alpha)/2}) = \frac{1 - 0,01}{2} = 0,495$, откуда $u_{0,495} = 2,58$.

Поскольку $|u| > u_{0,495}$, H_0 отвергается с уровнем значимости 1%, с надежностью 99% можно принять гипотезу $H_1: m_1 \neq m_2$, т. е. новая технология оказывает влияние на среднюю производительность труда. Увеличивает или уменьшает? Из приложения 2 найдем $u_{0,5-\alpha} = u_{0,5-0,01} = u_{0,49} = 2,33 < 4,38 = u$. Поэтому $m_x > m_y$, а это означает, что новая технология увеличивает производительность труда.

§ 5. Проверка гипотезы о равенстве математических ожиданий двух совокупностей при неизвестных одинаковых дисперсиях

Известны \bar{x} , \bar{y} , n_1 , n_2 , \tilde{S}_x^2 , \tilde{S}_y^2 нормально распределенных случайных величин X и Y . Проверяется гипотеза $H_0: m_x = m_y$ при $H_1: m_x \neq m_y$.

Используется статистика

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{(n_1 - 1)\tilde{S}_x^2 + (n_2 - 1)\tilde{S}_y^2}} \cdot \sqrt{\frac{n_1 n_2 (n_1 + n_2 - 2)}{n_1 + n_2}}, \quad (116)$$

и t -критерий Стьюдента для принятия H_0 :

$$|t| \leq t_{n_1+n_2-2, 1-\alpha}, \quad (117)$$

где $t_{n_1+n_2-2, 1-\alpha}$ – табличное значение распределения Стьюдента для числа степеней свободы $k = n_1 + n_2 - 2$ и уровня значимости α (приложение 4). Если $t > t_{n_1+n_2-2, 1-2\alpha}$, принимается $H_1: m_x > m_y$, если $t < -t_{n_1+n_2-2, 1-2\alpha}$, принимается $H_1: m_y > m_x$.

Пример 59.

Для $\bar{x} = 3,3$; $\bar{y} = 2,48$, $n_1 = 5$, $n_2 = 6$, $\tilde{s}_x^2 = 0,25$, $\tilde{s}_y^2 = 0,108$ проверить $H_0: m_y = m_x$ при $\alpha = 0,05$ и $H_1: m_x \neq m_y$.

Решение.

Вначале проверим гипотезу о равенстве дисперсий по критерию (111):

$$\frac{\tilde{s}_x^2}{\tilde{s}_y^2} = \frac{0,25}{0,108} = 2,31; \quad f_{4;5;0,95} = 5,19.$$

Гипотеза о равенстве дисперсий принимается. Теперь проверяем условие (116). Расчетное значение критерия:

$$\frac{3,3 - 2,48}{\sqrt{4 \cdot 0,25 + 5 \cdot 0,108}} \cdot \sqrt{\frac{5 \cdot 6 \cdot (5 + 6 - 2)}{5 + 6}} = 3,27;$$
$$t_{9;0,95} = 2,26.$$

Следовательно, H_0 отвергается с уровнем значимости 0,05.

Пример 60.

Плотность потока обслуженных клиентов (число обслуженных клиентов за час) в банках № 1 и № 2 зависит от времени и является случайной величиной. Для ответа на вопрос: какой банк работает более эффективно, – были проведены следующие расчеты. Для произвольно фиксированных моментов времени были составлены выборки числа обслуженных клиентов за час с $n_1 = 12$; $\bar{x} = 15$; $\tilde{s}_1^2 = 1,5$; $n_2 = 16$; $\bar{y} = 20$; $\tilde{s}_2^2 = 1$. При уровне значимости $\alpha = 0,05$ проверить гипотезу H_0 о более эффективной работе второго банка.

Решение.

Проверим гипотезу о равенстве дисперсий:

$$\frac{\tilde{s}_1^2}{\tilde{s}_2^2} = \frac{1,5}{1} = 1,5; \quad f_{11;15;0,95} = 2,72.$$

Гипотеза о равенстве дисперсий принимается. Вычислим

$$t = \frac{15 - 20}{\sqrt{11 \cdot 1,5 + 15 \cdot 1}} \cdot \sqrt{\frac{12 \cdot 16 \cdot (12 + 16 - 2)}{12 + 16}} = -11,9;$$

$$t_{n_1+n_2-2; 1-2\alpha} = t_{26; 0,9} = 1,71.$$

Поскольку $-t_{26; 0,9} = -1,71 > -11,9 = t$, принимается гипотеза $H_1: m_y > m_x$, т. е. с доверительной вероятностью 0,95 можно утверждать о том, что работа банка № 2 осуществляется более эффективно, чем банка № 1.

§ 6. Проверка гипотезы о равенстве математического ожидания гипотетическому значению

Рассмотрим два варианта решения этой задачи: при известной дисперсии случайной величины σ_x^2 и известной ее оценке S_x^2 нормального распределения случайной величины.

В первом случае известны \bar{x}, n, σ_x^2 . Проверяется $H_0: m_x = m_0$ при $H_1: m_x \neq m_0$.

Критерий для принятия H_0 :

$$\left| \frac{\bar{x} - m_0}{\sigma_x / \sqrt{n}} \right| \leq u_{кр}, \quad (118)$$

где $u_{кр}$ – аргумент функции Лапласа для вероятности $(1 - \alpha)/2$:

$$\Phi(u_{кр}) = (1 - \alpha)/2.$$

Данный критерий можно использовать для произвольного распределения при $n > 30 - 40$.

Пример 61.

Один из показателей эффективности работы автозаправочной станции – среднее число обслуженных машин за единицу времени, т. е. средняя плотность обслуженных машин, например, за час. Считая среднюю плотность случайной величиной, по выборке объема $n = 10$ с $\bar{x} = 22$, $\sigma_x = 2$ проверить гипотезу H_0 о равенстве генеральной средней значению $m_0 = 19$. Уровень значимости положить равным 0,05.

Решение.

$$\text{Найдем } \left(\frac{\bar{X} - m_0}{\sigma_x / \sqrt{n}} \right) = \frac{22 - 19}{2 / \sqrt{10}} = 0,47.$$

$$\Phi(u_{кр}) = 0,475, \text{ откуда } u_{кр} = 1,96.$$

Условие (118) выполнено, поэтому гипотеза о равенстве генеральной средней плотности потока обслуженных машин значению $m_0 = 19$ принимается на уровне значимости $\alpha = 0,05$.

Во втором случае известны \bar{x} , n , S_x^2 . Проверяется $H_0: m_x = m_0$ при $H_1: m_x \neq m_0$.

Применяется t -критерий Стьюдента для принятия H_0 :

$$\left| \frac{\bar{x} - m_0}{S_x / \sqrt{n-1}} \right| \leq t_{n-1, 1-\alpha}, \quad (119)$$

где $t_{n-1, 1-\alpha}$ – критическое значение t -распределения Стьюдента для числа степеней свободы $n-1$ и уровня значимости α .

И в первом, и во втором случае для $H_1: m_x > m_0$ и $H_1: m_x < m_0$ гипотеза H_0 принимается (отклоняется) аналогично тому, как это было показано в § 4-5.

Пример 62.

Для $\bar{x} = 17$, $n = 17$, $S_x = 4,5$, $m_0 = 15$ проверить H_0 при $H_1: m_x \neq m_0$ и $\alpha = 0,05$.

Решение.

Вычислим
$$\left| \frac{\bar{x} - m_0}{S_x / \sqrt{n-1}} \right| = \frac{17-15}{4,5 / \sqrt{16}} = 1,98; \quad t_{16, 0,95} = 2,12.$$

H_0 принимается с уровнем значимости 0,05.

Пример 63.

При условии примера 61 оценить: а) минимальный объем выборки, необходимый для принятия H_0 ; б) максимальное значение σ_x для принятия H_1 .

Решение.

а) $\frac{22-19}{2} \cdot \sqrt{n} \leq 1,96$, откуда $n = 2$;
 б) $\frac{22-19}{\sigma_x} \cdot \sqrt{10} > 1,96$, откуда $\sigma_x < 4,8402$;

при точности до четвертого знака можно считать максимальное значение σ_x равным 4,8401.

Пример 64.

При условии примера 62 найти мощность критерия, если в действительности $m_x = 14$.

Решение.

Поскольку $m_x = 14 < 15 = m_0$, критическая область левосторонняя. Применяя (119), получаем

$$\bar{x}_{\text{кр}} = m_0 - t_{n-1, 1-2\alpha} \cdot S_x / \sqrt{n-1} = 15 - 1,75 \cdot 4,5 / 4 = 13,03,$$

так как критическая область значений для \bar{x} есть интервал $(-\infty; 13,03)$. Мощность критерия равна вероятности отвергнуть H_0 , когда верна H_1 , т. е.

$$P = P(-\infty < \bar{x} < 13,03) = \frac{1}{2} (1 + P(|\bar{x}| \leq 13,03)).$$

Вероятность $P(|\bar{x}| < 13,03)$ находится из приложения 4 для распределения Стьюдента при $k = n - 1 = 17 - 1 = 16$ и

$$t = \frac{\bar{x}_{\text{кр}} - m_x}{S_x / \sqrt{n-1}} = \frac{13,03 - 14}{4,5 / 4} = -0,8.$$

Поскольку распределение Стьюдента симметрично относительно нуля, поэтому можно взять $t = 0,8$. В приложении 4 для $k = 16$ нет значения $t = 0,8$, поэтому применим линейную интерполяцию для $t_1 = 0,69$; $\beta_1 = 0,5$; $t_2 = 0,86$; $\beta_2 = 0,6$. Имеем:

$$\frac{t - 0,69}{0,86 - 0,69} = \frac{\beta - 0,5}{0,6 - 0,5}; \quad \beta = 0,59t + 0,09.$$

При $t = 0,8$ $\beta = 0,59 \cdot 0,8 + 0,09 = 0,562$; $P = \frac{1}{2} (1 + 0,562) = 0,781$. Итак, мощность критерия равна 0,781.

§ 7. Проверка гипотезы о равенстве вероятности появления события предполагаемому значению

Известна наблюдаемая частота появления события в n испытаниях k/n . Проверяется $H_0: p = p_0$ при $H_1: p \neq p_0$ (p – вероятность появления события в n испытаниях, другое название – **доля признака**).

Критерий для принятия H_0 :

$$|u| = \left| \frac{k/n - p_0}{\sqrt{p_0 \cdot q_0 / n}} \right| \leq u_{\text{кр}}; \quad (120)$$

где $q_0 = 1 - p_0$, $u_{\text{кр}}$ – значение аргумента функции Лапласа для вероятности $(1 - \alpha)/2$: $\Phi(u_{\text{кр}}) = (1 - \alpha)/2$. В противном случае принимается гипотеза $H_1: p \neq p_0$.

Случаи $H_1: p > p_0$ и $H_1: p < p_0$ рассматриваются аналогично тому, как показано в § 4–5.

Пример 65.

По 100 независимым испытаниям найдена относительная частота 0,08. Проверить $H_0: p = p_0$ для $p_0 = 0,12$ и $\alpha = 0,05$ при $H_1: p \neq p_0$.

Решение.

Найдем $|u| = \frac{|0,08 - 0,12|}{\sqrt{0,12 \cdot 0,88} / \sqrt{100}} = 1,25$; $\Phi(u_{кр}) = 0,475$, откуда $u_{кр} = 1,96$.

Значит, H_0 принимается с уровнем значимости 0,05.

Пример 66.

Известно, что доля нестандартных деталей среди 50 отобранных составила 1 деталь. Проверить гипотезу H_1 о том, что генеральная доля нестандартных деталей составляет менее 1,8%. Уровень значимости взять равным 0,05.

Решение.

Имеем: $k/n = 1/50 = 0,02$; $p_0 = 0,018$; $q_0 = 0,982$;

$$u = \frac{0,02 - 0,018}{\sqrt{0,018 \cdot 0,982} / \sqrt{50}} = 0,11.$$

Выдвигается гипотеза $H_1: p < 1,8\%$. Находится аргумент $u_{0,5-\alpha}$ функции Лапласа, для которого $\Phi(u_{0,5-\alpha}) = 0,5 - \alpha$ (приложение 2). Для нашего случая $\Phi(u_{0,5-0,05}) = \Phi(u_{0,45}) = 0,45$ и $u_{0,45} = 1,65$.

В силу того, что $u = 0,11 < 1,65 = u_{0,45}$, нельзя сделать вывод о том, что генеральная доля нестандартных изделий больше или меньше 1,8% (§ 4). Это же следует из неравенства (115), которое для рассматриваемого случая примет вид

$$u = 0,11 \leq u_{(1-\alpha)/2} = u_{(1-0,05)/2} = u_{0,475} = 1,96,$$

что означает принятие гипотезы H_0 о равенстве генеральной доли значению 1,8% при уровне значимости 5% ($\alpha = 0,05$).

§ 8. Проверка гипотезы о равенстве коэффициента корреляции нулю

Если генеральный коэффициент корреляции r_{xy} не равен нулю, это говорит о наличии линейной корреляционной связи между факторами X и Y во всей генеральной совокупности, и связь между этими факторами в рассматриваемой выборке не случайна, а имеет закономерный характер. В этом случае говорят, что выборочный коэффициент корреляции \hat{r}_{xy} **значимо** (существенно) отличается от нуля на уровне значимости α (или с надежностью $\beta = 1 - \alpha$), т. е. для $\beta\%$ выборок данной генеральной совокупности будет иметь место линейная корреляционная связь между факторами X и Y .

Пусть известна оценка \hat{r}_{xy} коэффициента корреляции r_{xy} . Требуется проверить нулевую гипотезу о равенстве генерального коэффи-

иента корреляции нулю, т. е. $H_0: r_{xy} = 0$ при альтернативной гипотезе $H_1: r_{xy} \neq 0$.

Используется **критерий Стьюдента** для принятия H_0 :

$$|t| = \frac{|\hat{r}_{xy}| \cdot \sqrt{n-2}}{\sqrt{1-\hat{r}_{xy}^2}} \leq t_{n-2;1-\alpha}, \quad (121)$$

где $t_{n-2;1-\alpha}$ – табличное значение распределения Стьюдента для числа степеней свободы $k = n - 2$ и уровня значимости α , \hat{r}_{xy} – выборочный коэффициент корреляции.

Для $H_1: r_{xy} > 0$ и $H_1: r_{xy} < 0$ гипотеза H_0 отклоняется (принимается), как показано в § 5.

Пример 67.

Проверить $H_0: r_{xy} = 0$ при $n = 100$; $\hat{r}_{xy} = 0,23$; $\alpha = 0,05$ и а) $H_1: r_{xy} \neq 0$; б) $H_1: r_{xy} < 0$; в) $H_1: r_{xy} > 0$.

Решение.

Воспользуемся статистикой t из (121) и найдем ее значение для исходных данных: $t = \frac{0,23 \cdot \sqrt{100-2}}{\sqrt{1-0,23^2}} = 2,34$;

а) $t_{98;0,95} = 1,98$ и H_0 отвергается в пользу H_1 с уровнем значимости 5%; б), в) $t_{98;1-2\alpha} = t_{98;0,9} = 1,66$; $2,34 = t > 1,66 = t_{98;0,9}$, поэтому принимается $H_1: r_{xy} > 0$.

Пример 68.

В табл. 59 приведены данные по капитализации X и обороту Y некоторых мировых компаний нефтегазового сектора на 31.03.05 в млн долларов (данные взяты из [9]).

Таблица 59

Капитализация X	380567,2	221365,3	123536,3	75036,1	30332,2	24542,6	23815,0
Оборот Y	291252,0	285059,0	150865,0	135076,0	14515,0	15749,3	22872,0

На уровне значимости $\alpha = 0,01$ проверить значимость выборочного коэффициента корреляции \hat{r}_{xy} .

Решение.

Вычислим выборочный коэффициент корреляции:

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

$$\begin{aligned}\text{где } \bar{x} &= \frac{1}{7}(380567,2 + 221365,3 + 123536,2 + 75036,1 + 30332,2 + 24542,6 + \\ &+ 23815) = \frac{879194,7}{7} = 125599,24;\end{aligned}$$

$$\begin{aligned}\bar{y} &= \frac{1}{7}(291252 + 285059 + 150865 + 135076 + 14515 + 15749,3 + 22872) = \\ &= \frac{915388,3}{7} = 130769,76;\end{aligned}$$

$$\begin{aligned}\overline{xy} &= \frac{1}{7}(380567,2 \cdot 291252 + 221365,3 \cdot 285059 + \dots + 23815 \cdot 22872) = \\ &= \frac{10^4}{7}(11084095 + 6310217,1 + 1863730,3 + 1013557,6 + 44027,188 + \\ &+ 38652,877 + 54469,688) = \frac{10^4}{7} \cdot 20408747 = 2915535,2 \cdot 10^4;\end{aligned}$$

$$\begin{aligned}S_x^2 &= \frac{1}{7}(380567,2^2 + 221365,3^2 + \dots + 23815^2) - 125599,22^2 = \\ &= \frac{10^4}{7} \cdot 21681510 - 1577516,4 \cdot 10^4 = 1519842,4 \cdot 10^4;\end{aligned}$$

$$\begin{aligned}S_y^2 &= \frac{1}{7}(291252^2 + 285059^2 + \dots + 22872^2) - 130769,75^2 = \\ &= \frac{10^4}{7} \cdot 20807398 - 1710072,7 \cdot 10^4 = 1262412,6 \cdot 10^4;\end{aligned}$$

$$S_x = 123282; \quad S_y = 112357;$$

$$\hat{r}_{xy} = \frac{2915535,2 \cdot 10^4 - 125599,22 \cdot 130769,75}{123282 \cdot 112357} \approx 0,92;$$

$$t = \frac{0,92\sqrt{7-2}}{\sqrt{1-0,92^2}} = 5,25; \quad t_{n-2;1-\alpha} = t_{5;0,99} = 4,03.$$

Вывод: поскольку $t > t_{5;0,99}$, гипотеза $H_0: r_{xy} = 0$ отвергается, т. е. с уверенностью 99% можно утверждать, что \hat{r}_{xy} значим.

Пример 69.

При условии предыдущего примера проверить гипотезу $H_1: r_{xy} > 0$.

Решение.

$$t_{n-2;1-2\alpha} = t_{5;0,98} = 3,36 < 5,25,$$

поэтому гипотеза H_1 принимается.

§ 9. Проверка гипотезы о значении генерального коэффициента корреляции

Используется z -преобразование Фишера [2, 4]:

$$z = \frac{1}{2} \ln \frac{1 + \hat{r}_{xy}}{1 - \hat{r}_{xy}},$$

имеющее приближенно уже при небольших n нормальное распределение (практическая рекомендация: $n \geq 10$) с

$$M_z = \frac{1}{2} \ln \frac{1 + r_{xy}}{1 - r_{xy}} + \frac{r_{xy}}{2(n-1)}, \quad \sigma_z^2 = \frac{1}{n-3}. \quad (122)$$

Проверяется $H_0: r_{xy} = r_0$ против $H_1: r_{xy} > r_0$ при уровне значимости α .

По условию, $P(z_{кр} < z < \infty) = \alpha$ – вероятность того, что z принимает значение, большее, чем критическое $z_{кр}$. Вычисляется M_z подстановкой вместо r_{xy} данного значения r_0 :

$$P(z_{кр} < z < \infty) = \Phi(\infty) - \Phi\left(\frac{z_{кр} - M_z}{\sigma_z}\right) = 0,5 - \Phi\left(\frac{z_{кр} - M_z}{\sigma_z}\right) = \alpha.$$

Отсюда

$$\Phi\left(\frac{z_{кр} - M_z}{\sigma_z}\right) = 0,5 - \alpha, \quad \frac{z_{кр} - M_z}{\sigma_z} = \Phi^{-1}(0,5 - \alpha),$$

здесь $\Phi^{-1}(0,5 - \alpha) = u_{0,5-\alpha}$ и $\Phi(u_{0,5-\alpha}) = 0,5 - \alpha$;

$$z_{кр} = M_z + \sigma_z \cdot u_{0,5-\alpha}. \quad (123)$$

Критерий принятия:

$$z_{кр} > z. \quad (124)$$

Аналогично, если $z < -z_{кр}$, принимается $H_1: r_{xy} < r_0$ и $z_{кр}$ определяется из (123); если $|z| > z_{кр}$, то принимается $H_1: r_{xy} \neq r_0$ и $z_{кр}$ определяется как в (123), но с $t_{(1-\alpha)/2}$.

Пример 70.

Проверить $H_0: r_{xy} = 0,2$ при $n=50$; $\hat{r}_{xy}=0,3$; $\alpha=0,05$; $H_1: r_{xy} > r_0$.

Решение.

$$M_z = \frac{1}{2} \ln \frac{1 + 0,2}{1 - 0,2} + \frac{0,2}{2(50-1)} = 0,21; \quad \sigma_z^2 = \frac{1}{50-3} = 0,02; \quad \sigma_z = 0,14;$$

$$u_{0,5-\alpha} = 1,65 \text{ (приложение 2); } z_{кр} = 0,21 + 0,14 \cdot 1,65 = 0,44;$$

$$z = \frac{1}{2} \ln \frac{1 + 0,3}{1 - 0,3} = 0,31; \quad 0,44 = z_{кр} > z = 0,31.$$

Вывод: гипотеза H_0 принимается, т. е. она согласуется с опытными данными, и с вероятностью 0,95 можно утверждать, что генеральный коэффициент корреляции равен 0,2. Если рассматривать множество таких совокупностей, то 95% из них будут иметь коэффициент корреляции 0,2.

Пример 71.

Как известно, эффективность любого мероприятия во многом определяется категорией эгоизма исполнителей. Эгоизм можно рассматривать как вектор с координатами, характеризующими, например, несвоевременное выполнение поручений, халатное отношение к труду и т. п. Тогда степень эгоизма можно охарактеризовать длиной вектора, соотнесенной с длиной вектора-нормы. В табл. 60 приведены условные данные относительно степени эгоизма (в усл. ед.) и соответствующей производительности труда (в усл. ед.) на примере 8 представителей данного объединения.

Таблица 60

Степень эгоизма X	2,3	3,4	4,1	6,2	7,5	7,8	8,1	9,0
Производительность труда Y	39,0	38,2	37,1	35,4	33,2	32,4	26,1	25,7

Проверить гипотезу H_0 о равенстве коэффициента корреляции значению $r_{xy} = -0,87$. Уровень значимости положить равным 0,05.

Решение.

$$\bar{x} = \frac{1}{8}(2,3 + 3,4 + 4,1 + 6,2 + 7,5 + 7,8 + 8,1 + 9) = 6,05;$$

$$\bar{y} = \frac{1}{8}(39 + 38,2 + 37,1 + 35,4 + 33,2 + 32,4 + 26,1 + 25,7) = 33,39;$$

$$S_x^2 = \frac{1}{8}(2,3^2 + 3,4^2 + \dots + 9^2) - 6,05^2 = \frac{335,8}{8} - 36,6 = 5,375;$$

$$S_y^2 = \frac{1}{8}(39^2 + 38,2^2 + \dots + 25,7^2) - 33,39^2 = \frac{9103,51}{8} - 1114,89 = 23,05;$$

$$S_x = \sqrt{5,375} = 2,32; \quad S_y = \sqrt{23,05} = 4,8;$$

$$\overline{xy} = \frac{1}{8}(2,3 \cdot 39 + 3,4 \cdot 38,2 + 4,1 \cdot 37,1 + 6,2 \cdot 35,4 + 7,5 \cdot 33,2 + 7,8 \cdot 32,4 +$$

$$+ 8,1 \cdot 26,1 + 9 \cdot 25,7) = 191,95; \quad \hat{r}_{xy} = \frac{191,95 - 6,05 \cdot 33,39}{2,32 \cdot 4,8} = -0,9;$$

$$M_z = \frac{1}{2} \ln \frac{1 - 0,87}{1 + 0,87} - \frac{0,87}{2(8-1)} = -1,31; \quad \sigma_z^2 \frac{1}{8-3} = 0,2; \quad \sigma_z = 0,45;$$

$u_{0,5-\alpha}=1,65$ (приложение 2); $z_{кр} = -1,31 + 1,65 \cdot 0,45 = -0,57$;

$$z = \frac{1}{2} \ln \frac{1-0,9}{1+0,9} = -1,5; \quad z_{кр} > z.$$

Вывод: гипотеза H_0 о равенстве генерального коэффициента корреляции значению $-0,87$ согласуется с опытными данными.

§ 10. Оценка закона распределения по критерию Пирсона

Распределение случайной величины задано таблицей частот 61.

Таблица 61

$x_1 \div x_2$	$x_2 \div x_3$	$x_m \div x_{m+1}$
n_1	n_2		n_m

Здесь m – число интервалов разбиения, x_i – левая, x_{i+1} – правая граница i -го интервала, $(x_{i+1} - x_i)$ – его ширина, $i = \overline{1, m}$, n_i – частота попадания в i -й интервал. Проверить гипотезу о том, что случайная величина имеет данное распределение.

Применяется **критерий χ^2 Пирсона** для принятия гипотезы о данном законе распределения случайной величины:

$$\chi^2 = \sum_{i=1}^m \frac{(n_i - n_i^T)^2}{n_i^T} < \chi_{m-r-1; \alpha}^2, \quad (125)$$

где $n_i^T = n \cdot p_i$ – теоретические частоты, p_i – вероятность попадания в i -й интервал; $\chi_{m-r-1; 1-\alpha}^2$ – табличное значение распределения χ^2 для числа степеней свободы $k=m-r-1$ и уровня значимости α ; r – число параметров гипотетического закона распределения. При применении данного критерия должно выполняться условие $np_i \geq 5$ [3, 7]. В [7] указывается на достаточное выполнение неравенств: $np_i \geq 1$ и $n \geq 50$, в [2] используется условие $n_i \geq 5$. В противном случае соответствующий интервал присоединяется к соседнему и число степеней свободы вычисляется для нового числа интервалов.

Количество параметров основных законов распределения указано в табл. 62.

Таблица 62

Закон распределения	Количество параметров
1. Равномерный	2
2. Показательный	1
3. Нормальный	2
4. Биномиальный	1
5. Пуассоновский	1

Для нормального закона распределения $r = 2$, т. е. два параметра: m_x и D_x , при этом

$$n_i^T = n \cdot p_i = n \cdot [\Phi(u_{i+1}) - \Phi(u_i)],$$

где $u_i = (x_i - \bar{x}) / S_x$; $u_{i+1} = (x_{i+1} - \bar{x}) / S_x$. Наименьшее u_1 полагают равным $-\infty$, наибольшее u_{k+1} – равным $+\infty$; $\Phi(z)$ – функция Лапласа.

Пример 72.

Проверить гипотезу о нормальном распределении при $\alpha = 0,05$ для случайной величины – ошибки в оценке прибыли, – заданной таблицей, в которой i – номер интервала, общее количество которых равно 7, x_i – левая, x_{i+1} – правая граница i -го интервала, n_i – частота i -го интервала (табл. 63).

Таблица 63

i	1	2	3	4	5	6	7
x_i	4	9	14	19	24	29	34
x_{i+1}	9	14	19	24	29	34	39
n_i	6	8	15	40	16	8	7

Решение.

$$n = \sum_{i=1}^7 n_i = 6 + 8 + 15 + 40 + 16 + 8 + 7 = 100.$$

По данным таблицы находим:

$$\bar{x} = \frac{1}{100} (6,5 \cdot 6 + 11,5 \cdot 8 + 16,5 \cdot 15 + 21,5 \cdot 40 + 26,5 \cdot 16 + 31,5 \cdot 8 + 36,5 \cdot 7) = 21,7;$$

$$S_x^2 = \frac{1}{100} (6,5^2 \cdot 6 + 11,5^2 \cdot 8 + 16,5^2 \cdot 15 + 21,5^2 \cdot 40 + 26,5^2 \cdot 16 + 31,5^2 \cdot 8 + 36,5^2 \cdot 7) - 21,7^2 = 52,998; S_x = 7,28.$$

Перейдем к нормированным границам интервалов u_i , u_{i+1} (табл. 64, при этом левый конец первого интервала принимается за $-\infty$, а правый конец последнего интервала за $+\infty$).

Таблица 64

i	1	2	3	4	5	6	7
u_i	$-\infty$	-1,74	-1,06	-0,37	0,32	1,0	1,69
u_{i+1}	-1,74	-1,06	-0,37	0,32	1,0	1,69	∞

Находим теоретические n_i^T частоты попадания в интервалы и слагаемые $(n_i - n_i^T)^2 / n_i^T$ левой части критерия (125) Пирсона (табл. 65).

Таблица 65

i	1	2	3	4	5	6	7
$\Phi(u_i)$	-0,5	0,459	0,355	0,144	0,125	0,341	0,454
$\Phi(u_{i+1})$	0,459	0,355	0,144	0,125	0,341	0,454	0,5
p_i	0,0409	0,1037	0,2111	0,2698	0,2158	0,1132	0,0455
$n_i^T = np_i$	4,09	10,37	21,11	26,98	21,58	11,32	4,55
$\frac{(n_i - n_i^T)^2}{n_i^T}$	0,892	0,542	1,768	6,283	1,443	0,974	1,32

Из последней строки этой таблицы получаем

$$\chi^2 = 13,22.$$

Поскольку $\chi^2_{4;0,05} = 9,5 < 13,22 = \chi^2$, гипотеза о нормальном распределении случайной величины отвергается.

Аналогично оцениваются другие законы распределения. В случае дискретной случайной величины при подсчете числа степеней свободы k вместо числа интервалов берут число разрядов с учетом их объединения. Подробнее об этом можно посмотреть, например, в [3, 7].

Замечание. Когда первоначально трудно определить с гипотезой относительно закона распределения, рекомендуется построить гистограмму по имеющемуся статистическому материалу.

Вопросы и задания к главе

1. Дать определение статистической гипотезы, привести примеры.
2. Какие гипотезы называются параметрическими?
3. Какая гипотеза называется нулевой, конкурирующей?
4. Перечислить типы альтернативной гипотезы.
5. Определить критерии значимости и согласия.
6. Что называется статистикой?
7. Что означает утверждение: данная гипотеза согласуется с опытными данными?
8. Определить ошибки I и II рода. Дать их экономическую интерпретацию.
9. Дать определение уровня значимости, в каком диапазоне он обычно изменяется?
10. Что называется мощностью критерия?
11. Как связаны уровень значимости и доверительная вероятность? Уровень значимости и вероятность ошибки II рода?
12. Как связаны мощность критерия и вероятность ошибки II рода?

13. Как определяются область допустимых значений и критическая область?
14. Как определяются односторонние и двусторонние критические области?
15. Что предпринять, если проверяемая гипотеза H_0 отклоняется, а вы уверены, что эта гипотеза верна?
16. Перечислите основные статистики, используемые при задании критериев значимости (согласия).
17. Для $\tilde{S}_1^2 = 12,56$; $\tilde{S}_2^2 = 8,9$; $n_1 = 14$; $n_2 = 16$; $\alpha = 0,05$ проверить $H_0: \sigma_1^2 = \sigma_2^2$ при: а) $H_1: \sigma_1^2 > \sigma_2^2$; б) $H_1: \sigma_1^2 < \sigma_2^2$; в) $H_1: \sigma_1^2 \neq \sigma_2^2$.
18. Был проведен выборочный опрос жителей двух регионов относительно их заработной платы (табл. 66, 67).

Таблица 66

Зарплата (ден. ед.)	120–160	160–200	200–240	240–280	280–320	320–360
Количество человек	5	9	14	11	7	4

Таблица 67

Зарплата (ден. ед.)	150–180	180–210	210–240	240–270	280–300	300–330
Количество человек	4	7	13	15	6	5

Считая распределение средней заработной платы нормальным, оценить, в каком регионе заработная плата имеет: а) больший; б) меньший разброс.

19. Проверить $H_0: \sigma_x^2 = \sigma_0^2$ при $n = 20$; $S_x^2 = 16$; $\sigma_0^2 = 15,6$; $\alpha = 0,01$ для: а) $H_1: \sigma_x^2 > \sigma_0^2$; б) $\sigma_x^2 < \sigma_0^2$; в) $\sigma_x^2 \neq \sigma_0^2$. Распределение признака считать нормальным.
20. При условии задания 18 проверить гипотезу $H_0: D_x = 14$ для различных альтернативных гипотез.
21. Считая распределения признаков x и y нормальными, проверить $H_0: m_x = m_y$ при $n_1 = 52$; $n_2 = 60$; $\bar{x} = 130$; $\bar{y} = 140$; $D_x = 40$; $D_y = 42$; $\alpha = 0,01$ и разных альтернативных гипотезах H_1 .
22. Для данных задания 18 проверить гипотезу о равенстве средних при разных вариантах H_1 .
23. Условие, как в примере 21, но значения генеральных дисперсий не известны, известно только, что они равны.

24. Для каждого из двух регионов (задание 18) сравнить генеральную среднюю со значением 23.

25. Проверить $H_0: p = p_0$ при $n=150$; $m=20$; $p_0=0,15$; $\alpha=0,05$ и разных H_1 .

26. При выборочном контроле в данном регионе из 100 машин в течение данных суток 30 нарушили правила парковки. Проверить гипотезу о том, что в данном регионе в целом в течение этих суток доля машин, нарушающих правила парковки, составит 20% (при разных альтернативных гипотезах и $\alpha=0,01$; $\alpha=0,05$).

27. Проверить гипотезу о равенстве коэффициента корреляции нулю, если $n=80$; $\hat{r}_{xy}=0,1$; $\alpha=0,05$ при разных альтернативных гипотезах.

28. Показатели X и Y представлены следующими выборками (табл. 68–70).

Таблица 68

x_i	7	8	11	15	17	18	20	24
y_i	5	9	10	8	8	15	14	16

Таблица 69

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

Таблица 70

$X \backslash Y$	20-30	30-40	40-50	50-60	60-70	70-80	80-90	90-100
50-70	2							
70-90	1	3	2		4	2	7	
90-110		7	1	5		6		4
110-130			4	8	3	7	8	3

На уровне значимости $\alpha=0,05$ проверить гипотезу о значимости выборочного коэффициента корреляции.

29. Для задания 28 выдвинуть гипотезу о значении генерального коэффициента корреляции и проверить ее при уровне значимости $\alpha=0,02$; $\alpha=0,05$.

30. Для активизации борьбы со стихийными бедствиями были проанализированы выборочные данные относительно проведенных мероприятий. В табл. 71 указаны плотности потоков (в усл. ед.) соответствующих мероприятий.

Таблица 71

2	6	14	5	6	9	7	9	8	12	12	16
14	4	8	15	18	16	18	20	15	22	21	4

На уровне значимости $\alpha = 0,05$ проверить гипотезу о равном распределении времени между проводимыми мероприятиями. Оценить среднее время наступления очередного мероприятия.

31. С целью повышения эффективности труда выборочно была проконтролирована норма выработки 50 рабочих (табл. 72).

Таблица 72

Норма выработки (усл. ед.)	240–260	260–280	280–300	300–320	320–340	340–360
Число работников	2	8	16	14	7	3

На уровне значимости $\alpha = 0,05$ проверить гипотезу о нормальном распределении нормы выработки.

32. При анализе демографического состояния работающего населения данного региона выборочно были взяты данные относительно возраста 1000 работающих респондентов. Были выделены возрастные категории: нулевая – младше 20 лет; первая – 20–40 лет; вторая – 40–60 лет; третья – 60–80 лет; четвертая – свыше 80 лет. Коды данных категорий соответственно: 0, 1, 2, 3, 4. Картина получилась следующая (табл. 73).

Таблица 73

Возрастная категория x_i	0	1	2	3	4
n_i (число работающих данной категории)	30	240	460	268	2

На уровне значимости $\alpha = 0,05$ проверить гипотезу о биномиальном распределении указанных возрастных категорий. В случае ее принятия найти вероятность того, что данный работник относится либо к 2-й, либо к 4-й категории; б) из трех работников хотя бы один будет из 3-й категории.

Глава 6. Статистическое изучение динамики социально-экономических явлений

§ 1. Временные ряды, их классификация и характеристики

Динамика – это изменение процессов (явлений) во времени, которое моделируется на основе **рядов динамики**. Ряды динамики состоят из двух последовательностей:

- 1) показателей времени (это годы, месяцы, дни и т. д.);
- 2) уровней динамики (значений экономического показателя для каждого показателя времени).

Обычно i -й уровень динамики обозначается J_i .

Существуют два вида рядов динамики:

1. **Моментальные ряды**, в которых уровни представлены в определенные моменты времени (например, численность служащих на момент выдачи заработной платы).
2. **Интервальные ряды**, в которых уровни представлены в определенные периоды (интервалы) времени (например, число новорожденных за 1 год).

Для характеристики рядов динамики используются **средние величины (средние уровни)**:

- 1) для интервальных рядов динамики с равными интервалами:

$$\bar{J} = \frac{\sum J}{n}, \quad (126)$$

где n – число интервалов;

- 2) для интервальных рядов с неравными интервалами:

$$\bar{J} = \frac{\sum J_i T_i}{\sum T_i}, \quad (127)$$

где J_i – уровень i -го периода,

T_i – длина i -го периода;

- 3) для моментальных рядов с равноотстоящими датами:

$$\bar{J} = \frac{1/2 J_1 + J_2 + \dots + J_{n-1} + 1/2 J_n}{n-1}; \quad (128)$$

эта средняя называется **средней хронологической**;

- 4) для моментальных рядов с неравноотстоящими датами:

$$\bar{J} = \frac{\sum \bar{J}_i T_i}{\sum T_i}, \quad (129)$$

где \bar{J}_i – средний уровень i -го периода.

Пример 73.

В табл. 74 представлены данные нарушений трудовой дисциплины в апреле месяце в ООО «Медуза». Найти среднее число нарушений.

Таблица 74

Календарные периоды апреля	Число работников J_i	Длины периодов в днях T_i	$J_i T_i$
1–4	3	4	12
5–11	4	7	28
12–20	1	9	9
21–27	0	7	0
28–30	2	3	6
Итого:		30	55

Решение.

Среднее число нарушений найдем по формуле (126): $\bar{J} = \frac{55}{30} = 1,83$.

Пример 74.

Определить среднегодовое число обучающихся первокурсников-информатиков в ТФ МЭСИ по следующим данным:

- 01.09.2001 – 24 студента,
- 01.09.2002 – 30 студентов,
- 01.09.2003 – 36 студентов,
- 01.09.2004 – 28 студентов.

Решение.

Воспользуемся формулой (128):

$$\bar{J} = \frac{24/2 + 30 + 36 + 28/2}{4 - 1} = \frac{92}{3} \approx 3 \text{ студента.}$$

Пример 75.

Определить среднюю численность работников данного предприятия за один год по следующим данным:

- на 01.01 – 100 человек,
- на 01.06 – 98 человек,
- на 01.08 – 94 человека,
- на 01.09 – 102 человека,
- на 01.01 (следующего года) – 106 человек.

Решение.

Введем обозначения: \bar{J}_i – средняя арифметическая двух соседних уровней, T_i – соответствующий период. Воспользуемся формулой (129), вычисления поместим в табл. 75.

Таблица 75

\bar{J}_i	T_i	$\bar{J}_i T_i$
$\bar{J}_1 = \frac{100 + 98}{2} = \frac{198}{2} = 99$	5	495
$\bar{J}_2 = \frac{98 + 94}{2} = \frac{192}{2} = 96$	2	192
$\bar{J}_3 = \frac{94 + 102}{2} = \frac{196}{2} = 98$	1	98
$\bar{J}_4 = \frac{102 + 106}{2} = \frac{208}{2} = 104$	4	416
Всего:	12	1201

$$\bar{J} = \frac{1201}{12} = 100,08.$$

Перейдем к рассмотрению **показателей динамики**. К ним относятся:

1) абсолютные приросты; 2) темпы роста; 3) темпы прироста; 4) абсолютное содержание одного процента прироста.

Эти показатели можно рассчитывать как: а) **цепные**, когда каждый уровень ряда сравнивается с предыдущим; б) **базисные**, когда каждый уровень ряда сравнивается с одним и тем же уровнем, взятым за базу сравнения.

Так, **абсолютные приросты** определяются как цепные $\Delta J_y = J_i - J_{i-1}$ и как базисные $\Delta J_o = J_i - J_0$, где J_i – i -й уровень, J_{i-1} – $i-1$ -й уровень, J_0 – базисный (начальный) уровень.

Пример 76.

В табл. 76 представлены изменения продукции по годам:

Таблица 76

2013	2014	2015
11	10	12

Определить абсолютные приросты.

Решение.

Цепные приросты: $\Delta J_y(14) = 10 - 11 = -1$; $\Delta J_y(15) = 12 - 10 = 2$. Базисные приросты: $\Delta J_o(14) = 10 - 11 = -1$; $\Delta J_o(15) = 12 - 11 = 1$ ($y_0 = 2013$).

Темпы роста показывают, во сколько раз изменяется изучаемый уровень, и вычисляются по формулам:

$$T_y = \frac{J_i}{J_{i-1}} \text{ — цепной показатель, } T_o = \frac{J_i}{J_0} \text{ — базисный показатель.}$$

Для рассмотренного примера в предположении, что 2013 – базисный период, имеем:

$$T_u = \left(\frac{14}{13}\right) = \frac{10}{11} = 0,909; \quad T_u = \left(\frac{15}{14}\right) = \frac{12}{10} = 1,2;$$

$$T_{\delta} = \left(\frac{14}{13}\right) = \frac{10}{11} = 0,909; \quad T_{\delta} = \left(\frac{15}{13}\right) = \frac{12}{11} = 1,091.$$

Произведение n цепных темпов роста равно n -базисному темпу. Для нашего примера

$$T_u \left(\frac{14}{13}\right) \cdot T_u \left(\frac{15}{14}\right) = T_{\delta} \left(\frac{15}{13}\right) = 0,909 \cdot 1,2 = 1,091.$$

Темпы прироста дают относительную оценку абсолютного прироста, а именно

$$\Delta T_u = \frac{\Delta J_u}{J_{i-1}} = \frac{J_i - J_{i-1}}{J_{i-1}} = \frac{J_i}{J_{i-1}} - 1 = T_u - 1,$$

т. е. для цепных темпов прироста

$$\Delta T_u = T_u - 1. \quad (130)$$

Для базисных темпов прироста

$$\Delta T_{\delta} = \frac{\Delta J_{\delta}}{J_0} = \frac{J_i - J_0}{J_0} = \frac{J_i}{J_0} - 1 = T_{\delta} - 1, \text{ т. е.}$$

$$\Delta T_{\delta} = T_{\delta} - 1. \quad (131)$$

Для нашего примера:

$$\Delta T_u \left(\frac{14}{13}\right) = 0,91 - 1 = -0,09; \quad \Delta T_u \left(\frac{15}{14}\right) = 1,2 - 1 = 0,2;$$

$$\Delta T_{\delta} \left(\frac{14}{13}\right) = 0,91 - 1 = -0,09; \quad \Delta T_{\delta} \left(\frac{15}{13}\right) = 1,09 - 1 = 0,9.$$

Абсолютное содержание одного процента прироста $A\%$ показывает, сколько абсолютных единиц содержится в одном проценте прироста и вычисляется по формуле

$$A\% = \frac{\Delta J_u}{\Delta T_u \cdot 100} = \frac{\Delta J_u}{\frac{\Delta J_u \cdot 100}{J_{i-1}}} = \frac{J_{i-1}}{100} = 0,01 \cdot J_{i-1}.$$

Для примера 76: $J_{i-1} = 10$ (2014 год), $A\% = 0,01 \cdot J_{i-1} = 0,1\%$.

§ 2. Средние показатели динамики

К ним относятся:

1. **Средние абсолютные приросты**, которые подразделяются на:

$$\text{а) цепные: } \Delta \bar{J}_y = \frac{\sum \Delta J_y}{n}; \quad \text{б) базисные: } \Delta \bar{J}_\delta = \frac{\Delta J_\delta}{m-1},$$

где n – число интервалов прироста, m – число уровней ряда динамики.

Для рассмотренного примера 76:

$$\Delta \bar{J}_y = \frac{-1+2}{2} = 0,5; \quad \Delta \bar{J}_\delta = \frac{12-11}{2} = \frac{1}{2} = 0,5.$$

2. Средние темпы роста рассчитываются по формуле средней геометрической двумя способами:

а) по цепным темпам роста:

$$\bar{T}_y = \sqrt[n]{T_{y1} \cdot T_{y2} \cdot \dots \cdot T_{yn}},$$

б) по базисным темпам роста:

$$\bar{T}_\delta = \sqrt[m]{T_\delta}.$$

Для примера 76 имеем:

$\bar{T}_y = \sqrt[2]{0,909 \cdot 1,2} = 1,045 = \bar{T}_\delta = \sqrt[3]{1,091}$, т.е. средний за год выпуск продукции увеличился на 4,5%.

3. Средние темпы прироста определяются так:

$$\Delta \bar{T}_y = \bar{T}_y - 1, \quad \Delta \bar{T}_\delta = \bar{T}_\delta - 1.$$

Для рассмотренного примера $\Delta \bar{T}_y = 1,045 - 1 = 0,045 = \Delta \bar{T}_\delta$, т.е. в среднем производство увеличилось на 4,5%.

Средние показатели динамики можно использовать для составления прогнозов, если исходить из предположения, что выявленная тенденция развития явления сохранится в будущем. Один из способов связан с использованием средних абсолютных приростов согласно формуле

$$J_{n+t} = J_n + (\Delta \bar{J})t, \quad (132)$$

где J_n – уровень на начало планового (базисного) периода, t – длина периода.

Прогнозирование методом среднего абсолютного прироста используется, если тенденция развития явления наилучшим образом аппроксимируется линейной зависимостью, а абсолютные цепные приросты примерно одинаковы.

Другой способ основывается на среднем темпе роста:

$$J_{n+t} = J_n + (\bar{T})^t. \quad (133)$$

Он применим, если темпы роста за данный период времени примерно одинаковы, а тенденция развития явления подчиняется геометрической прогрессии и может быть описана показательной (экспоненциальной) кривой.

Пример 77.

Используя данные примера 76, сделать прогноз относительного объема продукции на 2017 год.

Решение.

По первому способу: $J_n=12$, $\Delta\bar{J}_6=0,5$, $t=2$, $J_{2017}=12+(0,5)\cdot 2=12+2=13$.

По второму способу: $J_{2017}=12+1,045^2=13,09$.

§ 3. Методы анализа основной тенденции развития в рядах динамики

Одна из основных задач статистики заключается в определении тенденции развития изучаемых явлений. Для этого применяются **действия** над рядами динамики, заключающиеся в **выравнивании динамического ряда**, что может быть осуществлено либо укрупнением интервалов ряда, либо методом скользящей средней, либо методом аналитического выравнивания. Рассмотрим эти методы.

Укрупнение интервалов осуществляется путем суммирования уровней в рядах динамики, что приводит к новым уровням на более длительные периоды. Так, для рассмотренного примера 76 можно три периода (интервала) объединить в один. Тогда с 2013 по 2015 г. выпуск продукции составит $11 + 10 + 12 = 33$ (усл. ед.).

Метод **скользящей средней** основан на замене исходных уровней средними арифметическими за определенные периоды времени. В результате случайные колебания погашаются, а основная тенденция развития выражается в виде некоторой плавной линии скользящей средней. Если при сглаживании рядов динамики звенья скользящей средней составляются из нечетного числа уровней $2k + 1$, то для каждого уровня (там, где это возможно) выделяются подряд k предыдущих и k последующих уровней и ищется средняя арифметическая этих $2k + 1$ уровней. Если звенья скользящей средней определяются для четного числа уровней, то они попадают в промежутки между соответствующими уровнями, поэтому для определения сглаженных уровней находится среднее значение каждой соседней пары полученных средних.

Пример 78.

В табл. 77 приведен фактический уровень урожайности данной культуры по годам (первые два столбца таблицы). Выровнять данные с использованием скользящей средней (трехлетней и пятилетней) и показать скользящие средние на графике.

Таблица 77

Годы t	Фактический уровень урожайности y (усл. ед.)	Скользящая средняя	
		Трехлетняя	Пятилетняя
2006	15,4	-	-
2007	14	$\frac{15,4+14+17,6}{3}=15,7$	-
2008	17,6	$\frac{14+17,6+15,4}{3}=15,7$	$\frac{15,4+14+17,6+15,4+10,9}{5}=14,7$
2009	15,4	$\frac{17,6+15,4+10,9}{3}=14,6$	15,1
2010	10,9	14,6	15,2
2011	17,5	14,5	17,1
2012	15	17	16,8
2013	18,5	15,9	17,6
2014	14,2	15,9	-
2015	14,9	-	-

Решение.

В третьем и четвертом столбцах таблицы приведены данные значений урожайности для сглаженных уровней. На рис. 34 сплошной линией показана зависимость урожайности по годам, пунктиром показана зависимость для случая, когда средние значения определяются за каждые 3 и 5 лет. При этом 2005 г. и минимальное значение урожайности 10,9 (усл. ед.) совпадают с началом координат.

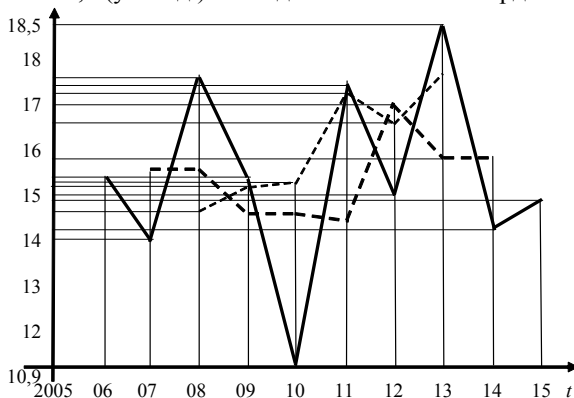


Рис. 34

Из рисунка видно, что исходные статистические данные, представленные графически сплошной линией, дают плохое представление об основной тенденции развития, поскольку имеются резкие колебания значений y . Совсем по-другому выглядит картина динамики урожайности, представленная трехлетней и пятилетней скользящими средними – отчетливо видна основная тенденция изменения урожайности.

Метод **аналитического выравнивания** ряда (сглаживание экспериментальных данных) сводится к аппроксимации исходных данных **наилучшим** образом подобранными с точки зрения приближений функциональными зависимостями, геометрически представленными линиями (однофакторная регрессия) или поверхностями (многофакторная) регрессия. Уравнения аппроксимирующих линий (поверхностей) называются **уравнениями регрессии**, а сами линии (поверхности) – **линиями (поверхностями) регрессии**.

Регрессионный анализ является разделом эконометрики. В данной книге мы ограничимся рассмотрением линейной однофакторной регрессии, описываемой уравнением прямой: $y = b_0 + b_1x$.

§ 4. Линейная однофакторная регрессия. Задача прогноза

Пусть имеются две выборки: x_1, x_2, \dots, x_n и y_1, y_2, \dots, y_n . Изобразим точки (x_i, y_i) на плоскости (рис. 35).

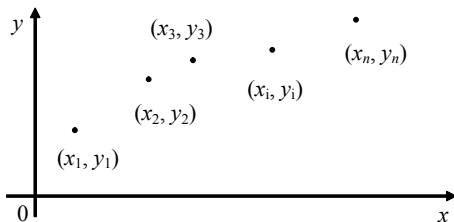


Рис. 35

Задача заключается в определении функциональной зависимости (приближенной) между признаками X и Y , описываемой в общем виде уравнением

$$\hat{y}_x = \varphi(x), \quad (134)$$

где \hat{y}_x – регрессия Y по X (подробнее, например, [1–3, 10]). Здесь X – факторный признак (независимая переменная), Y – результативный признак (зависимая переменная).

Наиболее распространенными типами уравнения (134) являются: $\hat{y}_x = b_1x + b_0$ – линейная связь (рис. 21, 22), $\hat{y}_x = b_2x^2 + b_1x + b_0$ – параболическая связь (рис. 23), $\hat{y}_x = ba^x$ – показательная связь, $\hat{y}_x = bx^n$ – степенная связь, $\hat{y}_x = \frac{1}{ax + b}$ – гиперболическая связь.

Подбор уравнения регрессии (аппроксимация) осуществляется по **методу наименьших квадратов**, согласно которому сумма квадратов отклонений статистических y_i от полученных приближенных значений \hat{y}_x должна быть минимальной. Ограничимся рассмотрением линейной связи. При линейной однофакторной регрессии аппроксимация осуществляется по прямой линии

$$\hat{y}_x = b_1x + b_0, \quad (135)$$

т. е. \hat{y}_x зависит от одного фактора x и эта связь – линейная. Коэффициенты b_1 и b_0 вычисляются соответственно по формулам:

$$b_1 = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{S_x^2}, \quad b_0 = \bar{y} - b_1\bar{x}. \quad (136)$$

По формуле (135) для любого значения x можно найти соответствующее значение \hat{y}_x , т. е. решить **задачу прогноза**.

Докажем формулы (136). Имеем задачу исследования функции на безусловный экстремум [11, 12]:

$$\sum (y_i - \hat{y}_x)^2 = \sum (y_i - b_0 - b_1x_i)^2 \rightarrow \min.$$

Найдем частные производные от минимизируемой функции, приравняем их к нулю, получим систему двух линейных уравнений с двумя неизвестными (систему так называемых нормальных уравнений), откуда найдем коэффициенты b_0 и b_1 .

Итак,

$$\begin{cases} nb_0 + b_1 \sum x_i = \sum y_i, \\ b_0 \sum x_i + b_1 \sum x_i^2 = \sum x_i y_i \end{cases} \quad (137)$$

и

$$\begin{aligned} b_1 &= \frac{n \sum x_i y_i - \sum x_i \sum y_i}{n \sum x_i^2 - (\sum x_i)^2}, \\ b_0 &= \frac{\sum y_i}{n} - b_1 \frac{\sum x_i}{n}. \end{aligned} \quad (138)$$

Равенства (138) преобразуются к виду (136). Тогда (135) можно записать в виде

$$\hat{y}_x - \bar{y} = b_1(x - \bar{x}).$$

Коэффициент b_1 показывает отклонение величины Y от ее среднего значения, приходящееся на единицу отклонения величины X от ее среднего значения. Для другого коэффициента регрессии анализируется его знак. Если $b_0 > 0$, то относительное изменение результата Y происходит медленнее, чем изменение фактора X , в противном случае – наоборот [10].

Вспомнив формулу (82) для вычисления коэффициента корреляции, получим связь между \hat{r}_{xy} и параметром b_1 :

$$\hat{r}_{xy} = b_1 \frac{S_x}{S_y}. \quad (139)$$

Пример 79.

Имеются данные о числе зарегистрированных браков на 1000 жителей России за 1977–1989 гг. (табл. 78).

Таблица 78

Годы	1977	1978	1979	1980	1981	1982	1983
Число браков, %	11,2	10,9	10,7	10,6	10,6	10,4	10,4
Годы	1984	1985	1986	1987	1988	1989	
Число браков, %	9,6	9,7	9,8	9,9	9,5	9,4	

Определить, какое количество браков, предположительно, могло быть зарегистрировано в 1990 г.

Решение.

Для упрощения выкладок закодируем годы следующим образом:

1977 – 0, 1978 – 1, 1979 – 2, ..., 1989 – 12.

Тогда 1990 г. будет иметь номер 13. Найдем

$$\bar{x} = \frac{1}{13}(0+1+2+3+4+5+6+7+8+9+10+11+12) = 6;$$

$$\bar{y} = \frac{1}{13}(11,2+10,9+10,7+10,6+10,6+10,4+10,4+9,6+9,7+9,8+9,9+9,5+9,4) = 10,21;$$

$$\overline{xy} = \frac{1}{13}(10,9+2 \cdot 10,7+3 \cdot 10,6+4 \cdot 10,6+5 \cdot 10,4+6 \cdot 10,4+7 \cdot 9,6+8 \cdot 9,7+9 \cdot 9,8+10 \cdot 9,9+11 \cdot 9,5+12 \cdot 9,4) = 59,25;$$

$$S_x^2 = \frac{1}{13}(1^2+2^2+3^2+4^2+5^2+6^2+7^2+8^2+9^2+10^2+11^2+12^2)-6^2 = 14;$$

$$b_1 = \frac{58,78-6 \cdot 10,21}{14} = -0,14; \quad b_0 = 10,21+0,14 \cdot 6 = 11,06.$$

Уравнение регрессии: $\hat{y}_x = -0,14x + 11,06$.

Найдем $\hat{y}_x(13) = -0,14 \cdot 13 + 11,06 \approx 9$.

Вывод: в 1990 г. число браков приблизительно было равно 9.

Пример 80.

Построить уравнение регрессии для примера 35 из главы 4, отражающее связь между стоимостью активной части Y основных фондов и затратами на производство работ X . Оценить значение стоимости y при затратах $x = 30$ (усл. ед.). Сделать анализ уравнения регрессии.

Решение.

Решение представим в виде табл. 79, в которой x_i – середины интервалов для фактора X , y_j – середины интервалов для Y , n_{ij} – частоты пар (x_i, y_j) , n_i – число фирм с данным значением x_i затрат, n_j – число фирм с данной стоимостью активной части y_j .

Таблица 79

x_i y_j	4	8	12	16	20	24	28	n_j	$y_j n_j$
62,5	4	2						6	375
87,5	6	3	3					12	1050
112,5		6	7					13	1462,5
137,5			8	6				14	192,5
162,5					4			4	650
187,5						1		1	187,5
212,5						2	3	5	1062,5
n_i	10	11	18	6	4	3	3	$\Sigma=55$	$\Sigma=6712,5$
$x_i n_i$	40	88	216	96	80	72	84	$\Sigma=676$	
$x_i^2 n_i$	160	704	2592	1536	1600	1728	2352	$\Sigma=10672$	
$\sum_{j=1}^7 y_j n_{ij}$	775	1062,5	2150	825	650	612,5	637,5		
$x_i \sum_{j=1}^7 y_j n_{ij}$	3100	8500	25800	13200	13000	14700	17850	$\Sigma=96150$	

$$\text{Отсюда } \bar{x} = \frac{676}{55} = 12,29; \bar{y} = \frac{6712,5}{55} = 122,05; \overline{xy} = \frac{96150}{55} = 1748,18;$$

$$\overline{x^2} = \frac{10672}{55} = 194,04; b_1 = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\overline{x^2} - \bar{x}^2} = \frac{1748,18 - 12,29 \cdot 122,05}{194,04 - 12,29^2} = 5,77;$$

$$b_0 = \bar{y} - b_1 \bar{x} = 122,05 - 5,77 \cdot 12,29 = 51,14;$$

уравнение регрессии:

$$\hat{y}_x = 5,77x + 51,14;$$

найдем $\hat{y}_x(30) = 5,77 \cdot 30 + 51,14 = 224,24$ (усл. ед.).

Анализ уравнения регрессии для исходных данных:

- 1) при затратах $x = 30$ (усл. ед.) стоимость y будет приблизительно равна 224,24 (усл. ед.);
- 2) при отклонении затрат от среднего значения на 1 единицу стоимость изменяется по сравнению со средним значением на 5,77 усл. ед.;
- 3) $\varepsilon = b_1 \frac{\bar{x}}{y} = 5,77 \cdot \frac{12,29}{122,05} = 0,58$, т. е. с возрастанием затрат x на 1% следует ожидать увеличения стоимости y на 0,45%;
- 4) составим табл. 80 отклонений $\varepsilon_i = y_i - \hat{y}_x$.

Таблица 80

y_i	62,5	87,5	112,5	137,5	162,5	187,5	212,5
x_i	4	8	12	16	20	24	28
$y(x_i)$	74,22	97,3	120,38	143,46	166,54	189,62	212,7
ε_i	-11,72	-9,8	-7,88	-5,96	-4,04	-2,12	-0,2

Все отклонения отрицательные, т. е. при данных затратах стоимость будет ниже ожидаемого результата.

Итак, в примере 79 был получен прогноз числа браков, в примере 80 – прогноз стоимости при данных затратах.

Насколько можно доверять полученным результатам? Чтобы ответить на этот вопрос, надо сделать **статистическую оценку** полученного уравнения регрессии, включающую:

- 1) оценку точности аппроксимации;
- 2) интервальную оценку коэффициентов b_0 и b_1 , которые являются случайными величинами, поскольку зависят от выборки;
- 3) интервальную оценку прогноза (индивидуального и среднего) значения y , также являющегося случайной величиной;
- 4) оценку значимости коэффициентов b_0 и b_1 ;
- 5) оценку значимости уравнения регрессии.

Мы рассмотрим первый пункт, остальные рассматриваются в курсе эконометрики.

Как уже отмечалось в § 2 главы 4, точность (погрешность) аппроксимации данного статистического материала можно оценить по формуле

$$\Delta = \frac{1}{n} \sum_{i=1}^n \left| \frac{y_i - y(x_i)}{y_i} \right| \cdot 100\%.$$

Если Δ находится в пределах 5–7%, это свидетельствует о хорошем подборе модели к исходным данным.

Оценим Δ для рассмотренного примера. Для этого сначала по уравнению регрессии вычислим все $\hat{y}_x(x_i)$ для $x_i = 0, 1, 2, \dots, 12$. Соответствующие значения помещены в табл. 81

Таблица 81

Годы	0	1	2	3	4	5
Приближенное число браков в %	11,29	11,11	10,93	10,75	10,57	10,39
Годы	7	8	9	10	11	12
Приближенное число браков в %	10,03	9,85	9,67	9,49	9,31	9,13

Затем найдем

$$\Delta = \frac{1}{13} \left(\left| \frac{11,2 - 11,29}{11,2} \right| + \left| \frac{10,9 - 11,11}{10,9} \right| + \left| \frac{10,7 - 10,93}{10,7} \right| + \left| \frac{10,6 - 10,75}{10,6} \right| + \right. \\ \left. + \left| \frac{10,6 - 10,57}{10,6} \right| + \left| \frac{10,4 - 10,39}{10,4} \right| + \left| \frac{10,4 - 10,21}{10,4} \right| + \left| \frac{9,6 - 10,03}{9,6} \right| + \left| \frac{9,7 - 9,85}{9,8} \right| + \right. \\ \left. + \left| \frac{9,8 - 9,67}{9,9} \right| + \left| \frac{9,9 - 9,49}{9,9} \right| + \left| \frac{9,5 - 9,31}{9,5} \right| + \left| \frac{9,4 - 9,13}{9,9} \right| \cdot 100\% \right) = 3\%.$$

Вывод: модель хорошо аппроксимирует данный статистический материал.

Вопросы и задания к главе

1. Как понимается динамика явлений?
2. Как определяются ряды динамики? Из каких последовательностей они состоят? Какие показатели времени обычно рассматриваются? Что представляют собой уровни динамики?
3. Какие виды рядов динамики существуют?
4. Как определяются: а) моментальные ряды; б) интервальные ряды? Привести примеры моментальных и интервальных рядов.
5. Как определяются средние уровни для интервальных и моментальных рядов? Как определяется хронологическая средняя?
6. Перечислить и охарактеризовать показатели динамики: а) цепные и б) базисные.
7. Дать определение абсолютных приростов. Привести примеры.
8. Что показывают темпы роста?
9. Какую оценку дают темпы прироста?
10. Что показывает абсолютное содержание одного процента?
11. Как классифицируются средние абсолютные приросты? Дать характеристику каждого вида.
12. Как вычисляются средние темпы роста?
13. Определить средние темпы прироста.
14. Указать два основных способа использования средних показателей динамики для составления прогнозов.
15. В чем заключается сущность выравнивания динамического ряда: а) укрупнением интервального ряда; б) методом скользящей средней; в) методом аналитического выравнивания?

16. Что называется: а) однофакторной; б) многофакторной регрессией; в) уравнением регрессии; г) линией (поверхностью) регрессии? Назовите метод, на основании которого определяется наилучшая аппроксимация (регрессия).

17. Что представляет собой линейная однофакторная регрессия? Привести формулу вычисления коэффициентов b_1 и b_0 .

18. Имеются данные (табл. 82) относительно объема продукции в данном цехе в указанные периоды (в усл. ед.)

Таблица 82

Рассматриваемые периоды в усл. ед.	Объем продукции J_i	Длины периодов в усл. ед. T_i	$J_i T_i$
1–5	150	5	750
5–11	280	7	1960
11–17	270	7	1890
17–25	450	9	4050
Итого:		28	8650

Найти средний объем произведенной продукции.

19. Для условия предыдущего примера найти средний объем, если объем (в усл. един.) составил:

на 1.09 – 150;

на 1.10 – 280;

на 1.11 – 270;

на 2.02 – 450.

20. В табл. 83 показаны изменения численности работников на данном предприятии по годам:

Таблица 83

2013	2014	2015
119	102	124

Найти: 1) абсолютные приросты; 2) темпы роста численности; 3) темпы прироста; 4) абсолютное содержание одного процента прироста.

21. Для условия предыдущего задания найти следующие показатели рядов динамики: 1) средние абсолютные приросты (цепные и базисные); 2) средние темпы роста (цепные и базисные); 3) средние темпы прироста.

22. Имеются данные (табл. 84) изменения некоторого экономического показателя по годам. Выводить данные с использованием скользящей средней (трехлетней и пятилетней). Показать скользящие средние на графике.

Таблица 84

Годы t	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016
Значения показателя y (усл. ед.)	17,8	19,2	20,1	18,0	17,4	24,6	27,8	22,2	30,2	34,1

23. Изменение средней прибыли данного предприятия по годам показано в табл. 85.

Таблица 85

Годы	2003	2006	2007	2010	2011	2013	2014	2016
прибыль усл. ед.	280	270	275	295	290	310	325	320

Построить линейное уравнение регрессии. Оценить его точность. Чему была равна прибыль в 2000 г.? Какую прибыль можно ожидать в 2018 г.? Исходя из тенденции роста прибыли, определить, через сколько лет она составит 360 усл. ед.

Глава 7. Экономические индексы

§ 1. Понятие индекса и их классификация

Индексом в статистике называется относительный показатель, который характеризует изменение какого-либо явления, состоящего из соизмеримых или несоизмеримых элементов, во времени, пространстве или по сравнению с любым эталоном.

Индексы **подразделяются**:

1. По охвату элементов совокупности:
 - а) индивидуальные;
 - б) сводные (общие).
2. По способу построения:
 - а) агрегатные (количественные и качественные);
 - б) средние (арифметические и гармонические).

Рассмотрим виды индексов подробнее.

Индивидуальные индексы дают относительную оценку изменения явления по отдельным элементам совокупности. К ним относятся:
- индивидуальные индексы физического **объема** продукции i_q , характеризующие количество (объем) произведенного одноименного товара в текущем (отчетном) q_1 и базисном q_0 периодах соответственно:

$$i_q = \frac{q_1}{q_0}. \quad (140)$$

- индивидуальные индексы **цен** i_p , представляющие собой отношение цены единицы одноименного товара в текущем периоде к соответствующей цене в базисном периоде:

$$i_p = \frac{p_1}{p_0}; \quad (141)$$

- индивидуальные индексы **товарооборота** (стоимости продукции), определяемые как отношение выручки $q_1 p_1$ текущего периода к выручке $q_0 p_0$ базисного периода:

$$i_{qp} = \frac{q_1 p_1}{q_0 p_0}. \quad (142)$$

Из (140)–(142) следует, что указанные индексы связаны соотношением:

$$i_{qp} = i_q \cdot i_p. \quad (143)$$

Значения индексов выражают в коэффициентах или процентах. При этом если из значения индекса, выраженного в процентах, вычесть 100%, то полученная разность покажет, на сколько процентов возросла

(уменьшилась) индексируемая величина. Так, если во II квартале 2003 г. цена 1 кг мяса на рынке равнялась 60 руб., а в III квартале – 65 рублей, то $i_p = 65 : 60 = 1,08$ или 108%, т.е. цена на мясо повысилась на 8%.

Сводные (общие) индексы дают относительную оценку изменения явления по всей разнородной совокупности, т. е. совокупности, состоящей из элементов разных наименований, которые непосредственно не складываются, например объем и цена. В этом случае удобно использовать **общий индекс** стоимости продукции или товарооборота:

$$I_{pq} = \frac{\sum p_1 q_1}{\sum p_0 q_0}, \quad (144)$$

где $\sum p_1 q_1$ – стоимость продукции (товарооборот) отчетного периода; $\sum p_0 q_0$ – то же самое для базисного периода.

Данный индекс показывает, во сколько раз возросла (уменьшилась) стоимость продукции отчетного периода по сравнению с базисным, или сколько процентов составляет рост (снижение) стоимости продукции.

К **агрегатным количественным** индексам относятся индексы физического объема продукции, стоимости продукции (товарооборота) национального дохода и др. Это количественные показатели. При их расчете количества оцениваются в одинаковых, сопоставимых ценах.

Качественные индексы: индексы цен, себестоимости, трудоемкости, производительности труда и т. п. Качественные показатели измеряют не общий объем, а интенсивность, эффективность явления или процесса. Как правило, они выражаются средними или относительными величинами и рассчитываются на базе одинаковых, неизменных количеств продукции.

Агрегатные индексы зависят либо от двух факторов, либо от одного фактора, когда другой фактор рассматривается либо только в базисном (первая группа индексов), либо только в отчетном периоде (вторая группа). Основные агрегатные индексы одного фактора приведены в табл. 86.

Таблица 86

Индексы первой группы	Индексы второй группы
Физического объема	
$I_q = \frac{\sum q_1 p_1}{\sum q_0 p_1}$	$I_q = \frac{\sum q_1 p_0}{\sum q_0 p_0}$
Цен	
$I_p = \frac{\sum p_1 q_1}{\sum p_0 q_1}$ – Пааше	$I_p = \frac{\sum p_1 q_0}{\sum p_0 q_0}$ – Ласпейреса

Индексы первой группы	Индексы второй группы
Себестоимости	
$I_z = \frac{\sum z_1 q_1}{\sum z_0 q_1}$	$I_z = \frac{\sum z_1 q_0}{\sum z_0 q_0}$
Производительности труда	
$I_t = \frac{\sum t_1 q_1}{\sum t_0 q_1}$	$I_t = \frac{\sum t_1 q_0}{\sum t_0 q_0}$

Индексы первой группы имеют следующей смысл:

- физического объема – характеризуют, во сколько раз увеличился (уменьшился) физический объем продукции в отчетном периоде по сравнению с базисным при ценах отчетного периода;
- индекс цен Пааше показывает, во сколько раз (на сколько в процентном соотношении) изменились цены отчетного периода по сравнению с базисным по товарам, реализованным в отчетном периоде;
- индекс себестоимости – увеличение (уменьшение) себестоимости в отчетном периоде по товарам, реализованным в отчетном периоде;
- индекс производительности труда – увеличение (уменьшение) этого фактора в отчетном периоде относительно производства товаров в отчетном периоде.

Индексы второй группы характеризуют то же самое, но в базисном периоде. Так, индекс цен Ласпейреса показывает, во сколько раз (на сколько процентов) подорожали (подешевели) товары базисного периода из-за изменения цен на них в отчетном периоде, т. е. на сколько изменились цены в отчетном периоде по сравнению с базисным, по той продукции, которая была реализована в базисном периоде.

К агрегатным индексам, зависящим от двух факторов, относятся:

- индекс стоимости продукции (**товарооборота**):

$$I_{pq} = I_q \cdot I_p = \frac{\sum q_1 p_0}{\sum q_0 p_0} \cdot \frac{\sum p_1 q_1}{\sum p_0 q_1} = \frac{\sum p_1 q_1}{\sum p_0 q_1} \cdot \frac{\sum p_1 q_0}{\sum p_0 q_0} = \frac{\sum p_1 q_1}{\sum p_0 q_0};$$

- индекс издержек производства

$$I_{zq} = I_z \cdot I_q = \frac{\sum z_1 q_1}{\sum z_0 q_1} \cdot \frac{\sum q_1 z_0}{\sum q_0 z_0} = \frac{\sum z_1 q_1}{\sum z_0 q_0}.$$

Индекс стоимости продукции или товарооборота зависит от двух факторов: изменения количества продукции (объема) и цен; показывает, во сколько раз возросла (уменьшилась) стоимость продукции (товарооборот) отчетного периода по сравнению с базисным, или сколько процентов составляет прирост (снижение) стоимости про-

дукции. Если из значения индекса стоимости вычесть 100%, то разность покажет, на сколько процентов возросла (уменьшилась) стоимость продукции в отчетном периоде по сравнению с базисным.

Пример 81.

Имеются данные о продаже товаров на рынке (табл. 87, в которой указаны товары двух видов).

Таблица 87

Товар	Цена 1 усл. ед. (ден. ед.)		Физический объем (усл. ед.)	
	базисный п-д	отчетный	базисный	отчетный
	p_0	p_1	q_0	q_1
I	8	6	143,5	167,1
II	11	10	38,9	45

Рассчитать индексы:

а) цен Ласпейреса и Пааше; б) физического объема; в) индексы товарооборота.

Решение.

а) Индекс цен Пааше:

$$\sum p_1 q_1 = 6 \cdot 167,1 + 10 \cdot 45 = 1452,6; \quad \sum p_0 q_1 = 8 \cdot 167,1 + 11 \cdot 45 = 1786,3;$$

$$I_p = \frac{1452,6}{1786,3} = 0,81, \text{ или } 81\%.$$

Вывод: индекс цен Пааше показывает, что по данному ассортименту товаров в отчетном периоде цены снизились на 19%.

Индекс цен Ласпейреса:

$$\sum p_1 q_0 = 6 \cdot 143,5 + 10 \cdot 38,9 = 1250; \quad \sum p_0 q_0 = 8 \cdot 143,5 + 11 \cdot 38,9 = 1575,9;$$

$$I_p = \frac{1250}{1575,9} = 0,79, \text{ или } 79\%.$$

Вывод: индекс цен Ласпейреса показывает, что товары базисного периода (данного ассортимента) подорожали на 21%, т. е. в отчетном периоде, по сравнению с базисным, цены уменьшились на 21% по той продукции, которая была реализована в базисном периоде.

б) Индекс физического объема в ценах отчетного периода:

$$I_q = \frac{\sum q_1 p_1}{\sum q_0 p_1} = \frac{1452,6}{1250} = 1,16.$$

Вывод: по данному ассортименту товаров прирост физического объема в текущем периоде составил 16% в ценах отчетного периода.

Индекс физического объема в ценах базисного периода:

$$I_q = \frac{\sum q_1 p_0}{\sum q_0 p_0} = \frac{1786,8}{1575,9} = 1,13.$$

Вывод: физический объем реализованных товаров в отчетном периоде по сравнению с базисным увеличился на 13% в ценах базисного периода.

в) Индекс товарооборота:

$$I_{qp} = I_q \cdot I_p = 1,13 \cdot 0,81 = 0,92.$$

Вывод: товарооборот уменьшился на 8% в отчетном периоде по сравнению с базисным.

Для качественной характеристики изменения явления используется также **индекс цен Фишера**:

$$I_{p\text{Фишера}} = \sqrt{I_{p\text{Пааше}} \cdot I_{p\text{Ласпейреса}}} \quad (145)$$

$$\text{Для рассмотренного примера } I_{p\text{Фишера}} = \sqrt{0,81 \cdot 0,79} = 0,8.$$

Рассмотренные агрегатные индексы позволяют сделать качественную оценку, для количественной характеристики применяются **абсолютные изменения**, которые представляют собой разность между числителем и знаменателем в формулах агрегатных индексов.

Знак «-» в этой разности означает абсолютную экономию денежных средств покупателей в результате изменения цен на эти товары. Знак «+» соответствует перерасходу денежных средств.

Здесь используются следующие показатели:

1) изменение физического объема:

$$\Delta_{qp_1} = \sum q_1 p_1 - \sum q_0 p_1 \quad (\text{в ценах отчетного периода}); \quad (146)$$

$$\Delta_{qp_0} = \sum q_1 p_0 - \sum q_0 p_0 \quad (\text{в ценах базисного периода}); \quad (147)$$

2) изменение цен:

$$\Delta_{pq_1} = \sum p_1 q_1 - \sum p_0 q_1 \quad (\text{в отчетном периоде}); \quad (148)$$

$$\Delta_{pq_0} = \sum p_1 q_0 - \sum p_0 q_0 \quad (\text{в базисном периоде}); \quad (149)$$

3) изменение себестоимости:

$$\Delta_{zq_1} = \sum z_1 q_1 - \sum z_0 q_1 \quad (\text{в отчетном периоде}); \quad (150)$$

$$\Delta_{zq_0} = \sum z_1 q_0 - \sum z_0 q_0 \quad (\text{в базисном периоде}); \quad (151)$$

4) изменение производительности труда:

$$\Delta_{tq_1} = \sum t_1 q_1 - \sum t_0 q_1 \quad (\text{в отчетном периоде}); \quad (152)$$

$$\Delta_{tq_0} = \sum t_1 q_0 - \sum t_0 q_0 \quad (\text{в базисном периоде}); \quad (153)$$

5) изменение стоимости продукции (товарооборота):

$$\Delta_{pq} = \sum q_1 p_1 - \sum q_0 p_0; \quad (154)$$

6) изменение издержек производства:

$$\Delta_{zq} = \sum z_1 q_1 - \sum z_0 q_0; \quad (155)$$

Из формул (143), (146), (151) и (145), (146), (151) соответственно получаем:

$$\Delta_{pq} = \Delta_{qp_1} + \Delta_{pq_0}, \quad (156)$$

$$\Delta_{pq} = \Delta_{pq_1} + \Delta_{qp_0}. \quad (157)$$

Пример 82. Найти абсолютное изменение физического объема, цен и товарооборота для примера 81.

Решение.

1) абсолютное изменение физического объема:

$$\Delta_{qp_1} = 1452,6 - 1250 = 202,6 \text{ (тыс. руб.)};$$

$$\Delta_{qp_0} = 1786,3 - 1575,9 = 210,4 \text{ (тыс. руб.)}.$$

Вывод: физический объем проданных товаров в текущем периоде при ценах этого периода увеличился по сравнению с базисным периодом на 202,6 тыс. руб., а при ценах базисного периода – на 210,4 тыс. руб.

2) абсолютное изменение цен:

$$\Delta_{pq_1} = 1452,6 - 1786,3 = -333,7 \text{ (тыс. руб.)};$$

$$\Delta_{pq_0} = 1250 - 1575,9 = -325,9 \text{ (тыс. руб.)}.$$

Вывод: цены товаров в текущем периоде снизились по сравнению с базисным периодом на:

а) 333,7 тыс. руб. по товарам текущего периода;

б) 325,9 тыс. руб. по товарам базисного периода,

что представляет собой соответствующую условную экономию средств населения (покупателей) от повышения цен.

3) абсолютное изменение товарооборота:

$$\Delta_{pq} = 1452,6 - 1575,9 = -123,3 \text{ (тыс. руб.)}.$$

Вывод: товарооборот уменьшился на 123,3 тыс. руб. за счет изменения цен и физического объема.

§ 2. Средние арифметические и гармонические индексы

Данные индексы применяются, когда не хватает данных для вычисления агрегатных индексов, но в то же время известны индивидуальные индексы. Пусть, например, $I_q = \frac{\sum q_1 p_0}{\sum q_0 p_0}$ и $i_q = \frac{q_1}{q_0}$ – индивидуальный индекс физического объема, тогда $q_1 = i_q \cdot q_0$ и **средний арифметический** индекс:

$$I_q = \frac{\sum i_q q_0 p_0}{\sum q_0 p_0}. \quad (158)$$

Пример 83.

В табл. 88 приведены данные о продаже одежды в магазине.

Таблица 88

Отдел	Товарооборот за 1-й квартал в тыс. руб. $q_0 p_0$	Изменение товарооборота во 2-м квартале по сравнению с 1-м i_q (в %)
Женская одежда	620	+15%
Мужская одежда	510	-8%
Детская одежда	300	Без изменения

Определить, на сколько процентов в среднем изменится объем реализации в целом по магазину.

Решение.

$$I_q = \frac{620 \cdot 1,15 + 510 \cdot 0,92 + 300}{620 + 510 + 300} = \frac{713 + 469,2 + 300}{620 + 510 + 300} = 1,033,$$

или 103,3%, т. е. объем реализации увеличился на 3,3%.

Для расчета сводных индексов качественных показателей используются средние гармонические индексы. Пусть $I_q = \frac{\sum p_1 q_1}{\sum p_0 q_1}$ и

$i_q = \frac{p_1}{p_0}$. Тогда $p_0 = \frac{p_1}{i_p}$ и **средний гармонический индекс**:

$$I_q = \frac{\sum p_1 q_1}{\sum \frac{p_1 q_1}{i_p}}. \quad (159)$$

Пример 84.

Данные по промтоварному магазину приведены в табл. 89:

Таблица 89

Секции	Товарооборот в тыс.руб.		Изменение цен на товары в % во II полугодии по сравнению с I i_q
	I полугодие $q_0 p_0$	II полугодие $q_1 p_1$	
Одежда	710	600	+6
Обувь	1050	1240	-2
Бытовая химия	200	250	+10

Определить: 1) на сколько процентов изменились цены и объем реализованных товаров в магазине во II полугодии по сравнению с I;

2) абсолютное изменение товарооборота (общее), а также его изменение за счет: а) цены; б) объема реализации.

Решение.

$$1) I_p = \frac{\sum q_1 p_1}{\sum q_1 p_0} = \frac{\sum q_1 p_0}{\sum q_1 \frac{p_0}{i_p}} = \frac{600 + 1240 + 250}{\frac{600}{1,6} + \frac{1240}{0,98} + \frac{250}{1,1}} = \frac{2090}{1867,6} = 1,11, \text{ или } 111\%;$$

$$I_{qp} = \frac{\sum q_1 p_1}{\sum q_0 p_0} = \frac{600 + 1240 + 250}{710 + 1050 + 200} = \frac{2090}{1960} = 1,07 \approx 1,1, \text{ или } 110\%;$$

$$I_q = \frac{I_{qp}}{I_p} = \frac{1,1}{1,11} = 0,99, \text{ или } 99\%.$$

$$2) \Delta_{pq}^{(\text{общее})} = \sum q_1 p_1 - \sum q_0 p_0 = (600 + 1240 + 250) - (710 + 1050 + 200) = 130 \text{ (тыс. руб.)}.$$

$$a) \Delta_{pq}^{(\text{за счет цены})} = \sum q_1 p_1 - \sum q_1 p_0 = 2090 - 1867,6 = 213,4 \text{ (тыс. руб.)}.$$

$$б) \Delta_{pq}^{(\text{за счет объема})} = \Delta_{pq}^{(\text{общее})} - \Delta_{pq}^{(\text{за счет цены})} = 130 - 213,4 = -83,4 \text{ (тыс. руб.)}.$$

§ 3. Индексы средних уровней качественных показателей

Рассматриваются три разновидности:

- 1) индексы переменного состава;
- 2) индексы постоянного состава;
- 3) индексы структурных сдвигов.

Индекс переменного состава зависит от изменения уровня качественного показателя для разных объектов (например, цены) и изменения структуры совокупности (например, объема объектов с разным уровнем качественного показателя). Для определения влияния 1-го фактора вводится индекс постоянного состава, для определения влияния 2-го фактора вводится индекс структурных сдвигов. Перечисленные индексы вычисляются соответственно по формулам:

$$I_{\text{перем.сост.}} = \frac{\sum q_1 p_1}{\sum q_1} \cdot \frac{\sum q_0 p_0}{\sum q_0} = \frac{\overline{p_1}}{\overline{p_0}}, \quad (160)$$

$$\text{здесь } \overline{p_1} = \frac{\sum q_1 p_1}{\sum q_1}, \quad \overline{p_0} = \frac{\sum q_0 p_0}{\sum q_0};$$

$$I_{\text{норм.сост.}} = \frac{\sum p_1 q_1}{\sum q_1} \cdot \frac{\sum p_0 q_1}{\sum q_1} = \frac{\overline{p_1}}{\overline{p_q}}, \quad (161)$$

здесь $\overline{p}_ц$ – условная цена;

$$I_{\text{стр.сдв.}} = \frac{\sum p_0 q_1}{\sum q_1} : \frac{\sum p_0 q_0}{\sum q_0} = \frac{\overline{p}_ц}{p_0}. \quad (162)$$

Как следует из (160)–(162), указанные индексы связаны соотношением

$$I_{\text{пост.сост.}} \cdot I_{\text{стр.сдв.}} = I_{\text{перем.сост.}}. \quad (163)$$

Пример 85.

В табл. 90 приведены данные по поставкам обуви в данный магазин с трех складов.

Таблица 90

Склады	Цена (усл. ед.)		Объем поставок (пар обуви)		Структура поставок (в %)	
	2013 г.	2014 г.	2013 г.	2014 г.	2013 г.	2014 г.
1	33	35	400	500	50	10
2	30	32	300	400	30	67
3	25	30	200	300	20	33
Итого:	p_0	p_1	q_0	q_1		
	88	97	900	1200	100	100

Определить:

- 1) индивидуальные индексы объема поставок и цены;
- 2) общие индексы средней цены:
 - а) индекс переменного состава; б) индекс постоянного состава;
 - в) индекс структурных сдвигов.

Решение.

$$1) i_{q_1} = \frac{500}{400} = 1,25; \quad i_{p_1} = \frac{35}{33} = 1,06; \quad i_{q_2} = \frac{400}{300} = 1,33;$$

$$i_{p_2} = \frac{32}{30} = 1,06; \quad i_{q_3} = \frac{300}{200} = 1,5; \quad i_{p_3} = \frac{30}{25} = 1,2.$$

2) По формулам (160)–(162):

$$a) I_{\text{перем.сост.}} = \frac{35 \cdot 500 + 32 \cdot 400 + 30 \cdot 300}{500 + 400 + 300} : \frac{33 \cdot 400 + 30 \cdot 300 + 25 \cdot 200}{400 + 300 + 200} =$$

$$= \frac{17500 + 12800 + 9000}{1200} : \frac{13200 + 9000 + 5000}{900} =$$

$$= \frac{39300}{1200} : \frac{27200}{900} = 1,08, \text{ или } 108\%;$$

$$б) I_{\text{пост.сост.}} = \frac{17500 + 12800 + 9000}{1200} : \frac{33 \cdot 500 + 30 \cdot 400 + 25 \cdot 300}{1200} =$$

$$= \frac{39300}{1200} : \frac{3600}{1200} = 1,09, \text{ или } 109\%;$$

$$в) I_{стр.сдв.} = \frac{1,08}{1,09} = 0,99, \text{ или } 99\%.$$

Вывод: средняя цена увеличилась на 8%, в том числе за счет повышения цены на каждом складе увеличилась на 9%, а за счет неблагоприятных структурных сдвигов в объеме поставок уменьшилась на 1%.

§ 4. Базисные и цепные индексы

Базисные и цепные индексы применяются, когда изменение индексируемых величин изучают не за два, а за ряд последовательных периодов.

- Базисные индивидуальные индексы:

$$i_{p_{1/0}} = \frac{p_1}{p_0}; \quad i_{p_{2/0}} = \frac{p_2}{p_0}; \quad i_{p_{3/0}} = \frac{p_3}{p_0}. \quad (164)$$

- Цепные индивидуальные индексы:

$$i_{p_{1/0}} = \frac{p_1}{p_0}; \quad i_{p_{2/1}} = \frac{p_2}{p_1}; \quad i_{p_{3/2}} = \frac{p_3}{p_2}. \quad (165)$$

Верны соотношения:

$$i_{p_{3/0}} = i_{p_{1/0}} \cdot i_{p_{2/1}} \cdot i_{p_{3/2}} = \frac{p_1}{p_0} \cdot \frac{p_2}{p_1} \cdot \frac{p_3}{p_2} = \frac{p_3}{p_0}; \quad (166)$$

т. е. произведение последовательных цепных индивидуальных индексов дает базисный индекс последнего периода;

$$i_{p_{3/2}} = i_{p_{3/0}} : i_{p_{2/0}}, \quad (167)$$

т. е. отношение базисного индекса отчетного периода к базисному индексу предшествующего периода дает цепной индекс отчетного периода.

Перечислим базисные и цепные индексы цен и физического объема продукции.

- Базисные индексы:

- цен Пааше:

$$I_{p_{1/0}} = \frac{\sum p_1 q_1}{\sum p_0 q_1}; \quad I_{p_{2/0}} = \frac{\sum p_2 q_2}{\sum p_0 q_2}; \dots; \quad I_{p_{n/0}} = \frac{\sum p_n q_n}{\sum p_0 q_n};$$

- цен Ласпейреса:

$$I_{p_{1/0}} = \frac{\sum p_1 q_0}{\sum p_0 q_0}; \quad I_{p_{2/0}} = \frac{\sum p_2 q_0}{\sum p_0 q_0}; \dots; \quad I_{p_{n/0}} = \frac{\sum p_n q_0}{\sum p_0 q_0};$$

- физического объема продукции:

$$I_{q_{1/0}} = \frac{\sum p_1 q_0}{\sum q_0 p_0}; \quad I_{q_{2/0}} = \frac{\sum p_2 q_0}{\sum q_0 p_0}; \dots; \quad I_{p_{n/0}} = \frac{\sum q_n p_0}{\sum q_0 p_0}.$$

2) Цепные индексы:

- цен Пааше:

$$I_{p_{1/0}} = \frac{\sum p_1 q_1}{\sum p_0 q_1}; \quad I_{p_{2/1}} = \frac{\sum p_2 q_2}{\sum p_1 q_2}; \dots; \quad I_{p_{n/n-1}} = \frac{\sum p_n q_n}{\sum p_{n-1} q_n};$$

- цен Ласпейреса:

$$I_{p_{1/0}} = \frac{\sum p_1 q_0}{\sum p_0 q_0}; \quad I_{p_{2/1}} = \frac{\sum p_2 q_0}{\sum p_1 q_0}; \dots; \quad I_{p_{n/n-1}} = \frac{\sum p_n q_0}{\sum p_{n-1} q_0};$$

- физического объема продукции:

$$I_{q_{1/0}} = \frac{\sum p_1 q_0}{\sum q_0 p_0}; \quad I_{q_{2/1}} = \frac{\sum p_2 q_0}{\sum q_1 p_0}; \dots; \quad I_{q_{n/n-1}} = \frac{\sum q_n p_0}{\sum q_{n-1} p_0}.$$

Вопросы и задания к главе

1. Дать определение индекса и привести классификацию индексов.
2. Охарактеризовать индивидуальные, сводные (общие), агрегатные (количественные и качественные), средние (арифметические и гармонические) индексы.
3. Привести формулы вычисления индексов.
4. Как связаны индивидуальные индексы физического объема, цен и товарооборота?
5. Как подразделяются агрегатные индексы одного фактора?
6. Назовите агрегатные индексы, зависящие от двух факторов.
7. Как связаны индексы цен Пааше, Ласпейреса и Фишера?
8. Для чего вводятся абсолютные изменения и как они определяются?
9. Как трактуются: а) отрицательные; б) положительные абсолютные изменения?
10. Приведите примеры показателей абсолютных изменений.
11. В каких случаях применяются арифметические и гармонические индексы?
12. Какие разновидности средних уровней качественных показателей рассматриваются? От каких показателей они зависят? Как связаны друг с другом?
13. В каких случаях применяются базисные и цепные индексы? Как они определяются и подразделяются?
14. В табл. 91 приведены данные о продаже товаров двух видов данного ассортимента.

Таблица 91

Товар	Цена 1 усл. ед. (ден. ед.)		Физический объем (усл. ед.)	
	базисный п-д	отчетный	базисный	отчетный
	p_0	p_1	q_0	q_1
I	10	7	148,6	170,2
II	12	10	195,8	203,4

Рассчитать индексы: а) цен Ласпейреса; б) физического объема; в) товарооборота. Сделать выводы.

15. Найти абсолютное изменение физического объема, цен и товарооборота для предыдущей задачи. Сделать выводы.

16. По приведенным в табл. 92 данным о продаже обуви в магазине определить, насколько процентов в среднем изменится объем реализации в целом по магазину.

Таблица 92

Отдел	Товарооборот за 1-й квартал в тыс. руб. $q_0 p_0$	Изменение товарооборота во 2-м квартале по сравнению с 1-м i_q (в %)
Женская обувь	830	+12%
Мужская обувь	670	без изменения
Детская обувь	570	-8%

17. Данные по хозяйственным товарам приводятся в табл. 93.

Определить: 1) на сколько процентов изменились цены и объем реализованных товаров во II полугодии по сравнению с I; 2) абсолютное изменение товарооборота (общее), а также его изменение за счет: а) цены; б) объема реализации.

Таблица 93

Товары	Товарооборот в ден. ед.		Изменение цен на товары в % во II полугодии по сравнению с I-м i_q
	I полугодие $q_0 p_0$	II полугодие $q_1 p_1$	
Светильники	640	600	-2
Бытовая химия	980	1050	+4
Электронагревательные приборы	440	480	+6

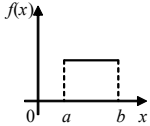
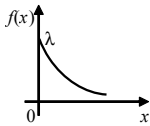
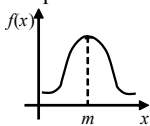
18. Имеются данные (табл. 94) относительно поставок мужских костюмов в магазин от трех поставщиков.

Таблица 94

Поставщики	Цена (ден. ед)		Объем поставок (число костюмов)		Структура поставок (в %)	
	2005 г.	2006 г.	2005 г.	2006 г.	2005 г.	2006 г.
1	35	33	70	80	30	25
2	30	30	90	85	40	30
3	25	28	100	105	25	35
Итого:	p_0	p_1	q_0	q_1		
	90	91	260	270	95	90

Определить:

- 1) индивидуальные индексы объема поставок и цены;
- 2) общие индексы средней цены: а) индекс переменного состава;
- б) индекс постоянного состава; в) индекс структурных сдвигов.

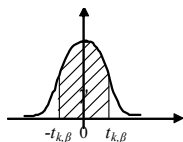
Основные законы распределения случайных величин		
Название закона	Формульное описание и числовые характеристики	Вероятность попадания на участок $[\alpha, \beta]$
1. Биномиальный	$P_{k,n} = P(X = k) =$ $= C_n^k \cdot p^k \cdot (1-p)^{n-k},$ $m_x = np, \quad D_x = np(1-p)$	$P[\alpha \leq X < \beta] = \sum_{k=\alpha}^{\beta-1} P_{k,n}$
2. Пуассоновский	$P(X = k) = \frac{a^k}{k!} e^{-a},$ где $a = \lambda \cdot l, l$ – мера области, $m_x = D_x = a$	$P[\alpha \leq X < \beta] = \sum_{k=\alpha}^{\beta-1} \frac{a^k}{k!} e^{-a}$
3. Равномерное распределение 	$f(x) = \begin{cases} \frac{1}{b-a}, & a \leq x \leq b, \\ 0, & x \notin [a, b], \end{cases}$ $m_x = \frac{a+b}{2}, \quad D_x = \frac{(b-a)^2}{12}$	$P[\alpha \leq X < \beta] = \frac{\beta' - \alpha'}{b-a},$ где $[\alpha, \beta] \cap [a, b] = [\alpha', \beta']$
4. Показательный 	$f(x) = \begin{cases} \lambda e^{-\lambda x}, & x \geq 0, \\ 0, & x < 0, \end{cases}$ $m_x = \frac{1}{\lambda}, \quad D_x = \frac{1}{\lambda^2}$	$P[\alpha \leq X < \beta] = e^{-\alpha\lambda} - e^{-\beta\lambda}$
5. Нормальный 	$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-m)^2}{2\sigma^2}}$ $m_x = m, \quad \sigma_x = \sigma$	$P[\alpha \leq X < \beta] =$ $= \Phi\left(\frac{\beta-m}{\sigma}\right) - \Phi\left(\frac{\alpha-m}{\sigma}\right)$

$$\text{Значения функции } \Phi(x) = \frac{1}{\sqrt{2\pi}} \cdot \int_0^x e^{-\frac{t^2}{2}} dt$$

x	0	1	2	3	4	5	6	7	8	9
0,0	0,000	0040	0080	0120	0160	0199	0239	0279	0319	0359
0,1	0398	0438	0478	0517	0557	0596	0636	0675	0714	0753
0,2	0793	0832	0871	0910	0948	0987	1026	1064	1103	1141
0,3	1179	1217	1255	1293	1331	1368	1406	1443	1480	1517
0,4	1554	1591	1628	1664	1700	1736	1772	1808	1844	1879
0,5	1915	1950	1985	2019	2054	2088	2123	2157	2190	2224
0,6	2257	2291	2324	2357	2389	2422	2454	2486	2517	2549
0,7	2580	2611	2642	2673	2708	2734	2764	2794	2823	2852
0,8	2881	2910	2939	2967	2995	3023	3051	3078	3106	3133
0,9	3159	3186	3212	3238	3264	3289	3315	3340	3365	3389
1,0	3413	3438	3461	3485	3508	3531	3554	3577	3599	3621
1,1	3643	3665	3696	3708	3729	3749	3770	3790	3810	3830
1,2	3849	3869	3883	3907	3925	3944	3962	3980	3997	4015
1,3	4032	4049	4066	4082	4099	4115	4131	4147	4162	4177
1,4	4192	4207	4222	4236	4251	4265	4279	4292	4306	4319
1,5	4332	4345	4357	4370	4382	4394	4406	4418	4429	4441
1,6	4452	4463	4474	4484	4495	4505	4515	4525	4535	4545
1,7	4554	4564	4573	4582	4591	4599	4608	4616	4625	4633
1,8	4641	4649	4656	4664	4671	4678	4686	4693	4699	4706
1,9	4713	4719	4726	4732	4738	4744	4750	4756	4761	4767
2,0	4772	4778	4783	4788	4793	4798	4803	4808	4812	4817
2,1	4821	4826	4830	4834	4838	4842	4846	4850	4854	4857
2,2	4861	4864	4868	4871	4875	4878	4881	4884	4887	4890
2,3	4893	4896	4898	4901	4904	4906	4909	4911	4913	4916
2,4	4918	4920	4922	4925	4927	4929	4931	4932	4934	4936
2,5	4938	4940	4941	4943	4945	4946	4948	4949	4951	4951
2,6	4953	4955	4956	4967	4959	4960	4961	4962	4963	4964
2,7	4965	4966	4967	4968	4969	4970	4971	4972	4973	4974
2,8	4974	4975	4976	4977	4977	4978	4979	4979	4980	4981
2,9	4981	4982	4982	4983	4984	4984	4985	4985	4986	4986

x		x		x		x		x	
3,0	0,49865	3,4	0,49966	3,8	0,49993	4,2	0,499987	4,6	0,499998
3,1	0,49903	3,5	0,49977	3,9	0,49995	4,3	0,499991	4,7	0,499999
3,2	0,49931	3,6	0,49984	4,0	0,499968	4,4	0,499995	4,8	0,499999
3,3	0,49952	3,7	0,49989	4,1	0,499979	4,5	0,4999966	4,9	0,499999

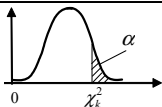
Значения $\Phi^{-1}\left(\frac{\beta}{2}\right)$		
β – доверительная вероятность	$\alpha = 1 - \beta$, α – уровень значи- мости	$u_{\beta} = \Phi^{-1}\left(\frac{\beta}{2}\right)$
0,9	0,1	1,65
0,95	0,05	1,96
0,98	0,02	2,3
0,99	0,01	2,58
0,9975	0,0025	3,02



Приложение 4

Значения $t_{k,\beta}$ -критерия Стьюдента

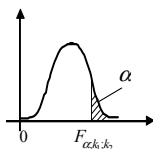
Число степеней свободы k	Доверительная вероятность $\beta = 1 - \alpha$, α – уровень значимости												
	сти												
	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	0,95	0,98	0,99	
1	0,16	0,32	0,51	0,73	1,00	1,38	1,96	3,08	6,31	12,71	31,82	63,66	
2	14	29	44	62	0,82	06	34	1,89	2,92	4,30	6,96	9,92	
3	14	28	42	58	76	0,98	25	64	35	3,18	4,54	5,84	
4	13	27	41	57	74	94	19	53	13	2,78	3,75	4,60	
5	13	27	41	56	73	92	16	48	01	57	36	03	
6	0,13	0,26	0,40	0,55	1,72	1,91	1,13	1,44	1,94	2,45	3,14	3,71	
7	13	26	40	55	71	90	12	41	89	36	00	50	
8	13	26	40	55	70	89	11	40	86	31	2,90	35	
9	13	26	40	54	70	88	10	38	83	26	82	25	
10	13	26	40	54	70	88	09	37	81	23	76	17	
11	0,13	0,26	0,40	0,54	0,70	0,88	1,09	1,36	1,80	2,20	2,72	3,11	
12	13	26	39	54	69	87	08	36	78	18	68	05	
13	13	26	39	54	69	87	08	35	77	16	65	01	
14	13	26	39	54	69	87	08	34	76	14	62	2,98	
15	13	26	39	54	69	87	07	34	75	13	60	95	
16	0,13	0,26	0,39	0,53	0,69	0,86	1,07	1,34	1,75	2,12	2,58	2,92	
17	13	26	39	53	69	86	07	33	74	11	57	90	
18	13	26	39	53	69	86	07	33	73	10	55	88	
19	13	26	39	53	69	86	07	33	73	09	54	86	
20	13	26	39	53	69	86	06	32	72	09	53	84	
21	0,13	0,26	0,39	0,53	0,69	0,86	1,06	1,32	1,72	2,08	2,52	2,83	
22	13	26	39	53	69	86	06	32	72	07	51	82	
23	13	26	39	53	68	86	06	32	71	07	50	81	
24	13	26	39	53	68	86	06	32	71	06	49	80	
25	13	26	39	53	68	86	06	32	71	06	48	79	
26	0,13	0,26	0,39	0,53	0,68	0,86	1,06	1,31	1,71	2,06	2,48	2,78	
27	13	26	39	53	68	85	06	31	70	05	47	77	
28	13	26	39	53	68	85	06	31	70	05	47	76	
29	13	26	39	53	68	85	05	31	70	04	46	76	
30	13	26	39	53	68	85	05	31	70	04	46	75	
40	0,13	0,25	0,39	0,53	0,68	0,85	1,05	1,30	1,68	2,02	2,42	2,70	
60	13	25	39	53	68	85	05	30	67	00	39	66	
120	0,13	0,25	0,39	0,53	0,68	0,84	1,04	1,29	1,66	1,98	2,36	2,62	
∞	13	25	38	52	67	84	04	28	64	96	33	58	



Приложение 5

Значения $\chi_{k,\alpha}^2$ -критерия Пирсона

Число степеней свободы k	Уровень значимости α												
	0,99	0,98	0,95	0,90	0,80	0,70	0,50	0,30	0,20	0,10	0,05	0,02	0,01
1	0,00	0,00	0,00	0,02	0,06	0,15	0,45	1,07	1,64	2,71	3,84	5,41	6,64
2	0,02	0,04	0,10	0,21	0,45	0,71	1,39	2,41	3,22	4,60	5,99	7,82	9,21
3	0,11	0,18	0,35	0,58	1,00	1,42	2,37	3,66	4,64	6,25	7,82	9,84	11,3
4	0,30	0,43	0,71	1,06	1,65	2,20	3,36	4,88	5,99	7,78	9,49	11,7	13,3
5	0,55	0,75	1,14	1,61	2,34	3,00	4,35	6,06	7,29	9,24	11,1	13,4	15,1
6	0,87	1,13	1,63	2,20	3,07	3,83	5,35	7,23	8,56	10,6	12,6	15,0	16,8
7	1,24	1,56	2,17	2,83	3,82	4,67	6,35	8,38	9,80	12,0	14,1	16,6	18,5
8	1,65	2,03	2,73	3,49	4,59	5,53	7,34	9,52	11,0	13,4	15,5	18,2	20,1
9	2,09	2,53	3,32	4,17	5,38	6,39	8,34	10,7	12,2	14,7	16,9	19,7	21,7
10	2,56	3,06	3,94	4,86	6,18	7,27	9,34	11,8	13,4	16,0	18,3	21,2	23,2
11	3,05	3,61	4,58	5,58	6,99	8,15	10,3	12,9	14,6	17,3	19,7	22,6	24,7
12	3,57	4,18	5,23	6,30	7,81	9,03	11,3	14,0	15,8	18,5	21,0	24,1	26,2
13	4,11	4,76	5,89	7,04	8,63	9,93	12,3	15,1	17,0	19,8	22,4	25,5	27,7
14	4,66	5,37	6,57	7,79	9,47	10,8	13,3	16,2	18,1	21,1	23,7	26,9	29,1
15	5,23	5,98	7,26	8,55	10,3	11,7	14,3	17,3	19,3	22,3	25,0	28,3	30,6
16	5,81	6,61	7,96	9,31	11,1	12,6	15,3	18,4	20,5	23,5	26,3	29,6	32,0
17	6,41	7,26	8,67	10,1	12,0	13,5	16,3	19,5	21,6	24,8	27,6	31,0	33,4
18	7,02	7,91	9,39	10,9	12,9	14,4	17,3	20,6	22,8	26,0	28,9	32,3	34,8
19	7,63	8,57	10,1	11,6	13,7	15,3	18,3	21,7	23,9	27,2	30,1	33,7	36,2
20	8,26	9,24	10,8	12,4	14,6	16,3	19,3	22,8	25,0	28,4	31,4	35,0	37,6
21	8,90	9,92	11,6	13,2	15,4	17,2	20,3	23,9	26,2	29,6	32,7	36,3	38,9
22	9,54	10,6	12,3	14,0	16,3	18,1	21,3	24,9	27,3	30,8	33,9	37,7	40,3
23	10,2	11,3	13,1	14,8	17,2	19,0	22,3	26,0	28,4	32,0	35,2	39,0	41,6
24	10,9	12,0	13,8	15,7	18,1	19,9	23,3	27,1	29,6	33,2	36,4	40,3	43,0
25	11,5	12,7	14,6	16,5	18,9	20,9	24,3	28,2	30,7	34,4	37,7	41,7	44,3
26	12,2	13,4	15,4	17,3	19,8	21,8	25,3	29,2	31,8	35,6	38,9	42,9	45,6
27	12,9	14,1	16,1	18,1	20,7	22,7	26,3	30,3	32,9	36,7	40,1	44,1	47,0
28	13,6	14,8	16,9	18,9	21,6	23,6	27,3	31,4	34,0	37,9	41,3	45,4	48,3
29	14,3	15,6	17,7	19,8	22,5	24,6	28,3	32,5	35,1	39,1	42,6	46,7	49,6
30	14,9	16,3	18,5	20,6	23,4	25,5	29,3	33,5	36,2	40,3	43,8	48,0	50,9



Приложение 6

Значения $F_{k_1, k_2, \alpha}$ -критерия Фишера–Снедекора

(k_1 – число степеней свободы большей дисперсии,
 k_2 – число степеней свободы меньшей дисперсии)

$k_2 \backslash k_1$	Уровень значимости $\alpha = 0,01$									
	1	2	3	4	5	6	7	8	9	10
1	4052	4999,5	5403	5625	5764	5859	5928	5982	6022	6056
2	98,50	99,00	99,17	99,25	99,30	99,33	99,36	99,37	99,39	99,40
3	34,12	30,82	29,46	28,71	28,24	27,91	27,67	27,49	27,35	27,23
4	21,20	18,00	16,69	15,98	15,52	15,21	14,98	14,80	14,66	14,55
5	16,26	13,27	12,06	11,39	10,97	10,67	10,46	10,29	10,16	10,05
6	13,75	10,92	9,78	9,15	8,75	8,47	8,26	8,10	7,98	7,87
7	12,25	9,55	8,45	7,85	7,46	7,19	6,99	6,84	6,72	6,62
8	11,26	8,65	7,59	7,01	6,63	6,37	6,18	6,03	5,91	5,81
9	10,56	8,02	6,99	6,42	6,06	5,80	5,61	5,47	5,35	5,26
10	10,04	7,56	6,55	5,99	5,64	5,39	5,20	5,06	4,94	4,85
11	9,65	7,21	6,22	5,67	5,32	5,07	4,89	4,74	4,63	4,54
12	9,33	6,93	5,95	5,41	5,06	4,82	4,64	4,50	4,39	4,30
13	9,07	6,70	5,74	5,21	4,86	4,62	4,44	4,30	4,19	4,10
14	8,86	6,51	5,56	5,04	4,69	4,46	4,28	4,14	4,03	3,94
15	8,68	6,36	5,42	4,89	4,56	4,32	4,14	4,00	3,89	3,80
16	8,53	6,23	5,29	4,77	4,44	4,20	4,03	3,89	3,78	3,69
17	8,40	6,11	5,18	4,67	4,34	4,10	3,93	3,79	3,68	3,59
18	8,29	6,01	5,09	4,58	4,25	4,01	3,84	3,71	3,60	3,51
19	8,18	5,93	5,01	4,50	4,17	3,94	3,77	3,63	3,52	3,43
20	8,10	5,85	4,94	4,43	4,10	3,87	3,70	3,56	3,46	3,37
21	8,02	5,78	4,87	4,37	4,04	3,81	3,64	3,51	3,40	3,31
22	7,95	5,72	4,82	4,31	3,99	3,76	3,59	3,45	3,35	3,26
23	7,88	5,66	4,76	4,26	3,94	3,71	3,54	3,41	3,30	3,21
24	7,82	5,61	4,72	4,22	3,90	3,67	3,50	3,36	3,26	3,17
25	7,77	5,57	4,68	4,18	3,85	3,63	3,46	3,32	3,22	3,13
26	7,72	5,53	4,64	4,14	3,82	3,59	3,42	3,29	3,18	3,09
27	7,68	5,49	4,60	4,11	3,78	3,56	3,39	3,26	3,15	3,06
28	7,64	5,45	4,57	4,07	3,75	3,53	3,36	3,23	3,12	3,03
29	7,60	5,42	4,54	4,04	3,73	3,50	3,33	3,20	3,09	3,00
30	7,56	5,39	4,51	4,02	3,70	3,47	3,30	3,17	3,07	2,98
40	7,31	5,18	4,31	3,83	3,51	3,29	3,12	2,99	2,89	2,80
60	7,08	4,98	4,13	3,65	3,34	3,12	2,95	2,82	2,72	2,63
120	6,85	4,79	3,95	3,48	3,17	2,96	2,79	2,66	2,56	2,47
∞	6,63	4,61	3,78	3,32	3,02	2,80	2,64	2,51	2,41	2,32

Продолжение приложения 6

$k_2 \backslash k_1$	Уровень значимости $\alpha = 0,01$							
	12	15	20	24	30	40	60	∞
1	6106	6157	6209	6235	6261	6287	6313	6366
2	99,42	99,43	99,45	99,46	99,47	99,47	99,48	99,50
3	27,05	26,87	26,69	26,60	26,50	26,41	26,32	26,13
4	14,37	14,20	14,02	13,93	13,84	13,75	13,65	13,46
5	9,89	9,72	9,55	9,47	9,38	9,29	9,20	9,02
6	7,72	7,56	7,40	7,31	7,23	7,14	7,06	6,88
7	6,47	6,31	6,16	6,07	5,99	5,91	5,82	5,65
8	5,67	5,52	5,36	5,28	5,20	5,12	5,03	4,86
9	5,11	4,96	4,81	4,73	4,65	4,57	4,48	4,31
10	4,71	4,56	4,41	4,33	4,25	4,17	4,08	3,91
11	4,40	4,25	4,10	4,02	3,94	3,86	3,78	3,60
12	4,16	4,01	3,86	3,78	3,70	3,62	3,54	3,36
13	3,96	3,82	3,66	3,59	3,51	3,43	3,34	3,17
14	3,80	3,66	3,51	3,43	3,35	3,27	3,18	3,00
15	3,67	3,52	3,37	3,29	3,21	3,13	3,05	2,87
16	3,55	3,41	3,26	3,18	3,10	3,02	2,93	2,75
17	3,46	3,31	3,16	3,08	3,00	2,92	2,83	2,65
18	3,37	3,23	3,08	3,00	2,92	2,84	2,75	2,57
19	3,30	3,15	3,00	2,92	2,84	2,76	2,67	2,49
20	3,23	3,09	2,94	2,86	2,78	2,69	2,61	2,42
21	3,17	3,03	2,88	2,80	2,72	2,64	2,55	2,36
22	3,12	2,98	2,83	2,75	2,67	2,58	2,50	2,31
23	3,07	2,93	2,78	2,70	2,62	2,54	2,45	2,26
24	3,03	2,89	2,74	2,66	2,58	2,49	2,40	2,21
25	2,99	2,85	2,70	2,62	2,54	2,45	2,36	2,17
26	2,96	2,81	2,66	2,58	2,50	2,42	2,33	2,13
27	2,93	2,78	2,63	2,55	2,47	2,38	2,29	2,10
28	2,90	2,75	2,60	2,52	2,44	2,35	2,26	2,06
29	2,87	2,73	2,57	2,49	2,41	2,33	2,23	2,03
30	2,84	2,70	2,55	2,47	2,39	2,30	2,21	2,01
40	2,66	2,52	2,37	2,29	2,20	2,11	2,02	1,80
60	2,50	2,35	2,20	2,12	2,03	1,94	1,84	1,60
120	2,34	2,19	2,03	1,95	1,86	1,76	1,66	1,38
∞	2,18	2,04	1,88	1,79	1,70	1,59	1,47	1,00

Продолжение приложения 6

$k_2 \backslash k_1$	Уровень значимости $\alpha = 0,05$									
	1	2	3	4	5	6	7	8	9	10
1	161	200	216	225	230	234	237	239	240	242
2	18,5	19,0	19,2	19,2	19,3	19,3	19,3	19,4	19,4	19,4
3	10,1	9,55	9,28	9,12	9,01	8,94	8,89	8,85	8,81	8,79
4	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6,00	5,96
5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,77	4,74
6	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10	4,06
7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,64
8	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39	3,35
9	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,14
10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,98
11	4,84	3,98	3,59	3,36	3,20	3,09	3,01	2,95	2,90	2,85
12	4,75	3,89	3,49	3,26	3,11	3,00	2,91	2,85	2,80	2,75
13	4,67	3,81	3,41	3,18	3,03	2,92	2,83	2,77	2,71	2,67
14	4,60	3,74	3,34	3,11	2,96	2,85	2,76	2,70	2,65	2,60
15	4,54	3,68	3,29	3,06	2,90	2,79	2,71	2,64	2,59	2,54
16	4,49	3,63	3,24	3,01	2,85	2,74	2,66	2,59	2,54	2,49
17	4,45	3,59	3,20	2,96	2,81	2,70	2,61	2,55	2,49	2,45
18	4,41	3,55	3,16	2,93	2,77	2,66	2,58	2,51	2,46	2,41
19	4,38	3,52	3,13	2,90	2,74	2,63	2,54	2,48	2,42	2,38
20	4,35	3,49	3,10	2,87	2,71	2,60	2,51	2,45	2,39	2,35
21	4,32	3,47	3,07	2,84	2,68	2,57	2,49	2,42	2,37	2,32
22	4,30	3,44	3,05	2,82	2,66	2,55	2,46	2,40	2,34	2,30
23	4,28	3,42	3,03	2,80	2,64	2,53	2,44	2,37	2,32	2,27
24	4,26	3,40	3,01	2,78	2,62	2,51	2,42	2,36	2,30	2,25
25	4,24	3,39	2,99	2,76	2,60	2,49	2,40	2,34	2,28	2,24
26	4,23	3,37	2,98	2,74	2,59	2,47	2,39	2,32	2,27	2,22
27	4,21	3,35	2,96	2,73	2,57	2,46	2,37	2,31	2,25	2,20
28	4,20	3,34	2,95	2,71	2,56	2,45	2,36	2,29	2,24	2,19
29	4,18	3,33	2,93	2,70	2,55	2,43	2,35	2,28	2,22	2,18
30	4,17	3,32	2,92	2,69	2,53	2,42	2,33	2,27	2,21	2,16
40	4,08	3,23	2,84	2,61	2,45	2,34	2,25	2,18	2,12	2,08
60	4,00	3,15	2,76	2,53	2,37	2,25	2,17	2,10	2,04	1,99
120	3,92	3,07	2,68	2,45	2,29	2,17	2,09	2,02	1,96	1,91
∞	3,84	3,00	3,60	2,37	2,21	2,10	2,01	1,94	1,83	1,83

Продолжение приложения 6

$k_2 \backslash k_1$	Уровень значимости $\alpha = 0,05$								
	12	15	20	24	30	40	60	120	∞
1	244	246	248	249	250	251	252	253	254
2	19,4	19,4	19,4	19,4	19,5	19,5	19,5	19,5	19,5
3	8,74	8,70	8,66	8,64	8,62	8,59	8,57	8,55	8,53
4	5,91	5,86	5,80	5,77	5,75	5,72	5,69	5,66	5,63
5	4,68	5,62	4,56	4,53	4,50	4,46	4,43	4,40	4,36
6	4,00	3,94	3,87	3,84	3,81	3,77	3,74	3,70	3,67
7	3,57	3,51	3,44	3,41	3,38	3,34	3,30	3,27	3,23
8	3,28	3,22	3,15	3,12	3,08	3,04	3,01	2,97	2,93
9	3,07	3,01	2,94	2,90	2,86	2,83	2,79	2,75	2,71
10	2,91	2,85	2,77	2,74	2,70	2,66	2,62	2,58	2,54
11	2,79	2,72	2,65	2,61	2,57	2,53	2,49	2,45	2,40
12	2,69	2,62	2,54	2,51	2,47	2,43	2,38	2,34	2,30
13	2,60	2,53	2,46	2,42	2,38	2,34	2,30	2,25	2,21
14	2,53	2,46	2,39	2,35	2,31	2,27	2,22	2,18	2,13
15	2,48	2,40	2,33	2,29	2,25	2,20	2,16	2,11	2,07
16	2,42	2,35	2,28	2,24	2,19	2,15	2,11	2,06	2,01
17	2,38	2,31	2,23	2,19	2,15	2,10	2,06	2,01	1,96
18	2,34	2,27	2,19	2,15	2,11	2,06	2,02	1,97	1,92
19	2,31	2,23	2,16	2,11	2,07	2,03	1,98	1,93	1,88
20	2,28	2,20	2,12	2,08	2,04	1,99	1,95	1,90	1,84
21	2,25	2,18	2,10	2,05	2,01	1,96	1,92	1,87	1,81
22	2,23	2,15	2,07	2,03	1,98	1,94	1,89	1,84	1,78
23	2,20	2,13	2,05	2,01	1,96	1,91	1,86	1,81	1,76
24	2,18	2,11	2,03	1,98	1,94	1,89	1,84	1,79	1,73
25	2,16	2,09	2,01	1,96	1,92	1,87	1,82	1,77	1,71
26	2,15	2,07	1,99	1,95	1,90	1,85	1,80	1,75	1,69
27	2,13	2,06	1,97	1,93	1,88	1,84	1,79	1,73	1,67
28	2,12	2,04	1,96	1,91	1,87	1,82	1,77	1,71	1,65
29	2,10	2,03	1,94	1,90	1,85	1,81	1,75	1,70	1,64
30	2,09	2,01	1,93	1,89	1,84	1,79	1,74	1,68	1,62
40	2,00	1,92	1,84	1,79	1,74	1,69	1,64	1,58	1,51
60	1,92	1,84	1,75	1,70	1,65	1,59	1,53	1,47	1,39
120	1,83	1,75	0,66	1,61	1,55	1,50	1,43	1,35	1,25
∞	1,75	1,67	1,57	1,52	1,46	1,39	1,32	1,22	1,00

Список литературы

1. *Ганичева, А. В.* Теория вероятностей : учеб. пособие для вузов. – Тверь : ТФ МЭСИ, 2005. – 156 с.
2. *Кремер, Н. Ш.* Теория вероятностей и математическая статистика: учебник для вузов. – М. : ЮНИТИ-ДАНА, 2009. – 551 с.
3. *Ганичева, А. В.* Теория вероятностей и математическая статистика : учеб. пособие для вузов / А. В. Ганичев, А. В. Ганичева. – Тверь : ТФ РГСУ, 2008. – 254 с.
4. *Ганичев, А. В.* Практикум по математической статистике с примерами в Excel / А. В. Ганичев, А. В. Ганичева. – Тверь : Тв. ГТУ, 2016. – 104 с.
5. *Ганичева, А. В.* Математика для психологов / А. В. Ганичева, В. П. Козлов. – М. : Аспект-Пресс, 2005. – 239 с.
6. *Савюк, Л. К.* Правовая статистика. – М. : Юрист, 2004. – 588 с.
7. *Палий, И. А.* Прикладная статистика. – М. : Высш. шк., 2004. – 176 с.
8. Инновационные процессы в регионах: экономика, образование, право : сб. науч. тр. – Тверь : ТФ МЭСИ, 2003. – 180 с.
9. Бизнес и социально-экономическое развитие региона: инновационный и инвестиционный аспект : сб. науч. тр. – Тверь : ТФ МЭСИ, 2006. – 168 с.
10. *Елисеева, И. И.* Эконометрика. – М. : Финансы и статистика, 2007. – 576 с.
11. Высшая математика для экономистов: учебник для студентов вузов, обучающихся по экономическим специальностям / Н. Ш. Кремер [и др.] ; под ред. Н. Ш. Кремера. – М. : ЮНИТИ-ДАНА, 2007. – 479 с.
12. *Ганичева, А. В.* Высшая математика для экономистов. – Тверь : Технопак, 2001. – 319 с.

Антонина Валериановна ГАНИЧЕВА
ПРИКЛАДНАЯ СТАТИСТИКА
Учебное пособие
Издание второе, стереотипное

Редакция
естественнонаучной литературы

ЛР № 065466 от 21.10.97
Гигиенический сертификат 78.01.10.953.П.1028
от 14.04.2016 г., выдан ЦГСЭН в СПб
Издательство «ЛАНЬ»
lan@lanbook.ru; www.lanbook.com
196105, Санкт-Петербург, пр. Юрия Гагарина, д. 1, лит. А
Тел./факс: (812) 336-25-09, 412-92-72
Бесплатный звонок по России: 8-800-700-40-71

Подписано в печать 11.06.21.
Бумага офсетная. Гарнитура Школьная. Формат 84×108¹/₃₂.
Печать офсетная. Усл. п. л. 9,03. Тираж 30 экз.

Заказ № 731-21.

Отпечатано в полном соответствии
с качеством предоставленного оригинал-макета
в АО «Т8 Издательские Технологии».
109316, г. Москва, Волгоградский пр., д. 42, к. 5.