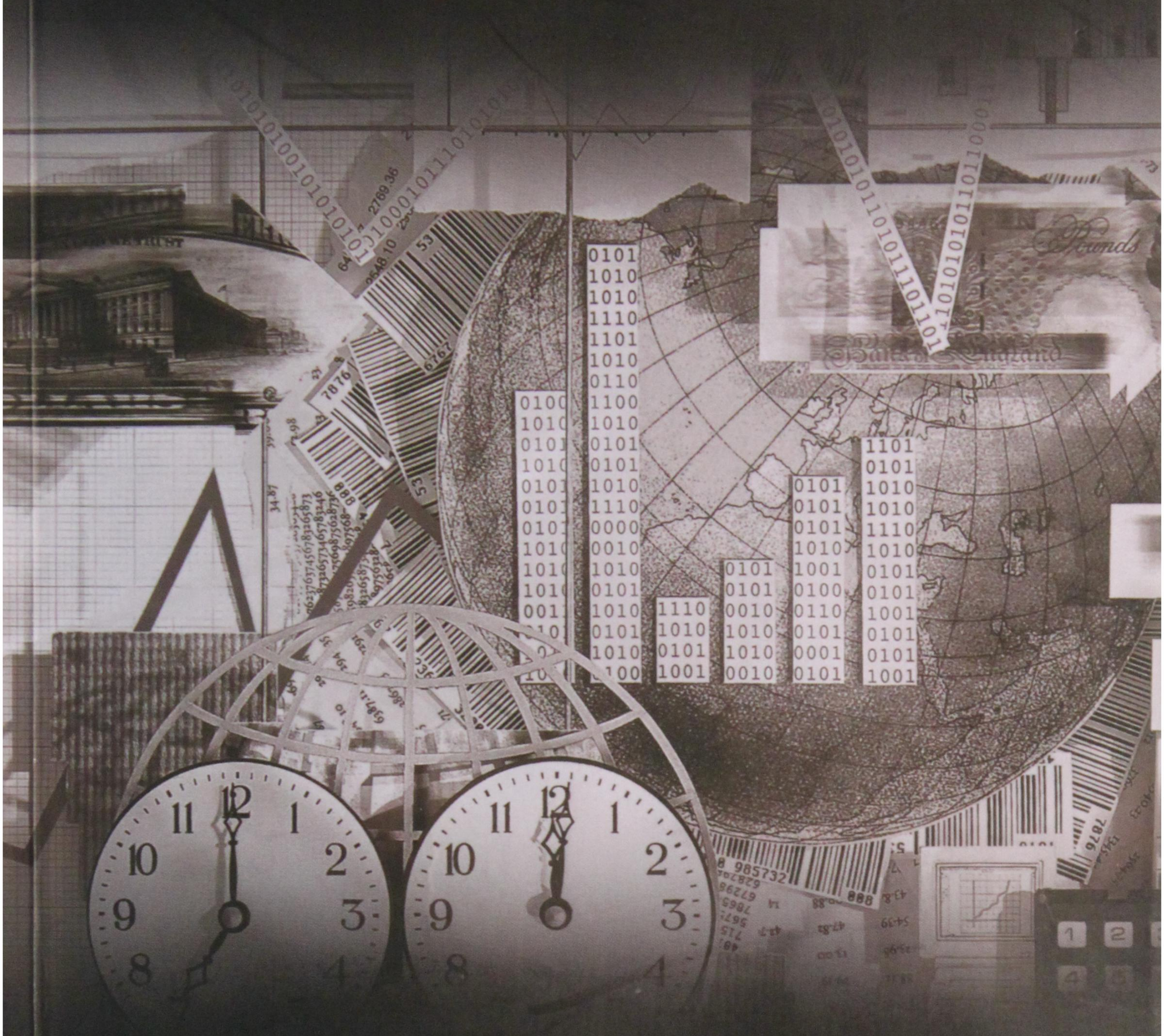
Университетская серия



Статистика

Под ред. Е. В. Улитиной

Учебное пособие

2-е издание

Market DS

Университетская серия

КК

Статистика

Под редакцией Е. В. Улитиной

Учебное пособие

2-е издание, переработанное и дополненное

Рекомендовано УМО по образованию в области антикризисного управления в качестве учебного пособия для студентов высших учебных заведений, обучающихся по специальности 351000 «Антикризисное управление» и другим экономическим специальностям.

Market DS

Москва, 2010

УДК 311.1/.2(075.8)
ББК 60.60я73-1+65в631.8я73-1
С78

Серия удостоена диплома в номинации «Лучший издательский проект» на IV Общероссийском конкурсе учебных изданий для высших учебных заведений «Университетская книга — 2008»

Печатается по решению Ученого совета Московской финансово-промышленной академии

Ответственный редактор серии
доктор экономических наук, профессор **Ю. Б. Рубин**

Улитина Е. В., Леднева О. В., Жирнова О. Л.

С78 Статистика : учеб. пособие / Е. В. Улитина, О. В. Леднева, О. Л. Жирнова. — 2-е изд., перераб. и доп. — М.: Маркет ДС, 2010. — 312 с. (Университетская серия).

ISBN 978-5-94416-107-9

Агентство CIP РГБ

Учебное пособие представляет собой классический вводный курс статистики, направленный на формирование у студентов базовых компетенций обработки и анализа информации, выраженной числовыми данными. Учебный материал охватывает основные вопросы сбора и обобщения данных, формирует системное представление о возможностях и особенностях применения богатого статистического инструментария для выявления закономерностей развития различных социальных и экономических явлений, способствует развитию навыков и компетенций применения статистических методов для решения задач на начальных этапах экономического анализа информации.

Предназначено для студентов начальных курсов высших учебных заведений экономических факультетов, аспирантов и преподавателей вузов, а также всех интересующихся вопросами статистики.

УДК 311.1/.2(075.8)
ББК 60.60я73-1+65в631.8я73-1

ISBN 978-5-94416-107-9

КРАТКОЕ СОДЕРЖАНИЕ

Раздел 1
ОПИСАТЕЛЬНАЯ СТАТИСТИКА

Глава 1.1. Статистика как наука и отрасль практической деятельности 9
Глава 1.2. Источники и методы сбора статистической информации. Статистическое наблюдение 23
Глава 1.3. Представление статистических данных с помощью таблиц и графиков . 30
Глава 1.4. Методы сбора и систематизации статистической информации. Сводка и группировка данных 47
Глава 1.5. Абсолютные и относительные статистические показатели 63

Раздел 2
ОСНОВЫ АНАЛИЗА ВАРИАЦИИ ПРИЗНАКА В РЯДАХ РАСПРЕДЕЛЕНИЯ

Глава 2.1. Статистические показатели в форме средних величин 73
Глава 2.2. Показатели вариации 90

Раздел 3
АНАЛИЗ ВЗАИМОСВЯЗИ МЕЖДУ СОЦИАЛЬНО-ЭКОНОМИЧЕСКИМИ ЯВЛЕНИЯМИ

Глава 3.1. Измерение и прогнозирование взаимосвязи с помощью корреляционно-регрессионного анализа 113
Глава 3.2. Многофакторные модели анализа взаимосвязи 153
Глава 3.3. Непараметрические методы оценки связи социально-экономических явлений 176

Раздел 4
АНАЛИЗ ДИНАМИКИ СОЦИАЛЬНО-ЭКОНОМИЧЕСКИХ ЯВЛЕНИЙ

Глава 4.1. Понятие и классификация рядов динамики 193
Глава 4.2. Аналитические показатели ряда динамики 198
Глава 4.3. Методы анализа основной тенденции в рядах динамики 207
Глава 4.4. Методы выявления сезонных колебаний в рядах динамики 220
Глава 4.5. Простейшие методы прогнозирования временных рядов 232
Глава 4.6. Индексный метод и его применение в анализе социально-экономических явлений 243

Раздел 5
ВЫБОРОЧНЫЙ МЕТОД НАБЛЮДЕНИЯ В СТАТИСТИЧЕСКИХ ИССЛЕДОВАНИЯХ

Глава 5.1. Общие понятия теории выборочного наблюдения 257
Глава 5.2. Виды выборок 269

СОДЕРЖАНИЕ

Раздел 1
ОПИСАТЕЛЬНАЯ СТАТИСТИКА

Глава 1.1. Статистика как наука и отрасль практической деятельности 9
1.1.1. Из истории развития статистики 9
1.1.2. Предмет и метод статистики 11
1.1.3. Отрасли статистики 15
1.1.4. Основные категории статистики 16

Глава 1.2. Источники и методы сбора статистической информации. Статистическое наблюдение 23
1.2.1. Сущность и виды статистического наблюдения 23
1.2.2. Точность статистического наблюдения 28

Глава 1.3. Представление статистических данных с помощью таблиц и графиков 30
1.3.1. Статистическая таблица и ее элементы 30
1.3.2. Виды статистических таблиц 31
1.3.3. Основные правила построения статистических таблиц 34
1.3.4. Основные принципы применения графического метода в статистике 36
1.3.5. Классификация основных видов статистических графиков 38

Глава 1.4. Методы сбора и систематизации статистической информации. Сводка и группировка данных 47
1.4.1. Статистическая сводка и группировка данных 47
1.4.2. Принципы построения статистических группировок 50
1.4.3. Вторичная группировка 57
1.4.4. Ряды распределения 59

Глава 1.5. Абсолютные и относительные статистические показатели . . . . .
1.5.1. Абсолютные показатели 63
1.5.2. Относительные показатели 65

Раздел 2

ОСНОВЫ АНАЛИЗА ВАРИАЦИИ ПРИЗНАКА В РЯДАХ РАСПРЕДЕЛЕНИЯ

Глава 2.1. Статистические показатели в форме средних величин 73
2.1.1. Сущность и виды средних величин 73
2.1.2. Средняя арифметическая 76
2.1.3. Свойства средней арифметической 81
2.1.4. Средняя гармоническая 83
2.1.5. Другие виды средних величин 84
2.1.6. Структурные средние 85

Глава 2.2. Показатели вариации 90
2.2.1. Для чего изучают вариацию 90
2.2.2. Показатели вариации и способы их расчета 93
2.2.3. Математические свойства показателей вариации 103
2.2.4. Правило сложения дисперсий 107

Раздел 3

АНАЛИЗ ВЗАИМОСВЯЗИ МЕЖДУ СОЦИАЛЬНО-ЭКОНОМИЧЕСКИМИ ЯВЛЕНИЯМИ

Глава 3.1. Измерение и прогнозирование взаимосвязи с помощью корреляционно-регрессионного анализа 113
3.1.1. Классификация видов взаимосвязи 113
3.1.2. Методы изучения взаимосвязи 118
3.1.3. Исследование взаимосвязи с помощью диаграммы рассеяния 119
3.1.4. Условия применения корреляционно-регрессионного анализа 123
3.1.5. Расчет линейного коэффициента корреляции 124
3.1.6. Задачи применения регрессионного анализа 132
3.1.7. Вычисление параметров линейной парной регрессии 133
3.1.8. Вычисление параметров уравнения регрессии при нелинейной зависимости . 136
3.1.9. Принятие решений на основе уравнений регрессии 143
3.1.10. Проведение корреляционно-регрессионного анализа по сгруппированным данным 146

Глава 3.2. Многофакторные модели анализа взаимосвязи 153

3.2.1. Определение множественного коэффициента корреляции и коэффициента детерминации 153

3.2.2. Построение модели множественной регрессии 158

3.2.3. Интерпретация результатов множественной регрессии 164

3.2.4. Применение корреляционно-регрессионного анализа в исследовании финансово-экономической деятельности 169

Глава 3.3. Непараметрические методы оценки связи социально-экономических явлений 176

3.3.1. Порядковая шкала и ее применение в анализе взаимосвязей 176

3.3.2. Ранговые коэффициенты корреляции Спирмена и Кендалла 178

3.3.3. Коэффициент конкордации (множественный коэффициент ранговой корреляции) 184

3.3.4. Бисериальный коэффициент корреляции 186

3.3.5. Методы изучения взаимосвязи между качественными признаками 188

Раздел 4

АНАЛИЗ ДИНАМИКИ СОЦИАЛЬНО-ЭКОНОМИЧЕСКИХ ЯВЛЕНИЙ

Глава 4.1. Понятие и классификация рядов динамики 193

4.1.1. Понятие о рядах динамики. Виды рядов динамики. 193

4.1.2. Сопоставимость уровней и смыкание рядов динамики 195

Глава 4.2. Аналитические показатели ряда динамики 198

4.2.1. Индивидуальные аналитические показатели динамики 198

4.2.2. Средние аналитические показатели в рядах динамики 202

Глава 4.3. Методы анализа основной тенденции в рядах динамики. 207

4.3.1. Тенденция ряда динамики и методы ее выявления 207

4.3.2. Метод укрупнения интервалов 208

4.3.3. Метод скользящих средних 209

4.3.4. Метод аналитического выравнивания 212

Глава 4.4. Методы выявления сезонных колебаний в рядах динамики 220

4.4.1. Базовая модель временного ряда 220

4.4.2. Методы выявления сезонной компоненты 221
4.4.3. Расчет индекса сезонности методом постоянной средней 222
4.4.4. Расчет индекса сезонности методом аналитического выравнивания 225
4.4.5. Расчет индекса сезонности методом скользящей средней 228

Глава 4.5. Простейшие методы прогнозирования временных рядов 232
4.5.1. Особенности прогнозирования при исследовании рядов динамики 232
4.5.2. Прогнозирование методом среднего уровня ряда 234
4.5.3. Прогнозирование методом среднего абсолютного прироста 235
4.5.4. Прогнозирование методом среднего темпа роста 237
4.5.5. Прогнозирование на основе экстраполяции тренда 239

Глава 4.6. Индексный метод и его применение в анализе социально-экономических явлений 243
4.6.1. Общее понятие индексов и их назначение в экономическом анализе 243
4.6.2. Индивидуальные индексы 245
4.6.3. Сводные индексы 247
4.6.4. Индексный анализ влияния структурных изменений 252
4.6.5. Средние формы сводных индексов 254

Раздел 5

ВЫБОРОЧНЫЙ МЕТОД НАБЛЮДЕНИЯ В СТАТИСТИЧЕСКИХ ИССЛЕДОВАНИЯХ

Глава 5.1. Общие понятия теории выборочного наблюдения 257
5.1.1. Значение и теоретические основы выборочного наблюдения 257
5.1.2. Основные преимущества и недостатки выборочного наблюдения 261
5.1.3. Средняя и предельная ошибки выборки. Построение доверительных границ для среднего и доли 263
5.1.4. Виды выборочного наблюдения. Способы отбора единиц в выборочную совокупность 267

Глава 5.2 Виды выборок 269
5.2.1. Собственно случайная выборка 269

5.2.2. Систематическая (механическая) выборка 280
5.2.3. Стратифицированная (типическая) выборка 282

ПРИЛОЖЕНИЯ

Приложение 1. Таблица случайных чисел 291
Приложение 2. Нормальный закон распределения 293
Приложение 3. Распределение Пирсона (χ^2-распределение) 296
Приложение 4. Распределение Стьюдента (t-распределение). 300
Приложение 5. Распределение Фишера–Снедекора (F-распределение) 302
Приложение 6. Таблица Z-преобразования Фишера 309

Раздел 1

ОПИСАТЕЛЬНАЯ СТАТИСТИКА

Глава 1.1

СТАТИСТИКА КАК НАУКА И ОТРАСЛЬ ПРАКТИЧЕСКОЙ ДЕЯТЕЛЬНОСТИ

- Из истории развития статистики
- Предмет и метод статистики
- Отрасли статистики
- Основные категории статистики

1.1.1. Из истории развития статистики

Статистика — одна из древнейших отраслей знаний. Ее возникновение связано с потребностями общества в самых разнообразных сведениях, без которых невозможно управлять государством, изучать отдельные явления и процессы, происходящие в различных областях жизни.

Первые учетные операции проводились еще в глубокой древности. Вначале они были довольно примитивны и направлены, главным образом, на получение данных о численности населения, его составе, имущественном положении. По мере развития производительных сил в обществе возрастал интерес к систематизации и расширению возможностей анализа информации о процессах, происходящих в экономической и социальной жизни, постепенно накапливался опыт сбора, обработки и обобщения сведений.

Термин «статистика» произошел от латинских слов *«stato»* (государство) и *«status»* (положение вещей, политическое состояние). До середины XVIII в. под статистикой подразумевалась совокупность сведений о государстве, о его «достопримечательностях» — системе государственного устройства, этнографической информации о быте и нравах населения, климате, финансах, армии. С 1666 г. в университетах Германии преподавался курс «Государствоведение». Термин «статистика» в научный обиход ввел немецкий ученый Готфрид Ахенваль, представитель описательной школы государствоведения. В 1746 г. Г. Ахенваль предложил заменить курс «Государствоведение» на «Статистику», положив тем самым начало развитию статистики как дисциплины и науки.

Большой вклад в статистическую науку и практику внесли русские ученые и общественные деятели. В трудах М. В. Ломоносова, И. К. Кириллова, В. Н. Татищева, а позднее и К. И. Арсеньева получили развитие идеи комплексного экономико-статистического описания страны. А. Н. Радищев сформулировал ценные предложения в области судебной статистики. В работах Д. П. Журавского показана роль группировок в статистике, предложена система статистических показателей для изучения общественной жизни. П. Л. Чебышев и его ученики сформулировали математическую базу для научно обоснованного применения выборочного метода, А. А. Чупров занимался методами установления зависимости между явлениями, разрабатывал теоретические основы математической статистики.

Во всех странах мира деятельностью по сбору и обобщению информации для внутренних и внешних потребностей экономики занимаются специально учреждаемые государствами органы — статистические комитеты и комиссии, статистические бюро. Примечательно, что развитие российской государственной статистики на принципах регулярного централизованного сбора данных о населении и экономике началось в России только в 1802 г. с создания Министерств по сбору учетных данных по подведомственным им отраслям экономики и с возникновением в 1811 г. при Министерстве полиции первого Статистического отделения. Деятельность этого учреждения привела к реализации первых глубоких статистических исследований, явившихся следствием ожидания оживления в России свободного предпринимательства и изменениями в состоянии экономики и общественной жизни. Реализация этих исследований, прежде всего в области изучения населения, во многом способствовала принятию реформы 1861 г. Пореформенный период российской статистики характеризуется как расцвет земской статистики и развитие начал современной методологии статистического учета населения. Важнейшей вехой стало проведение первой всеобщей переписи населения Российской империи 1897 г. под руководством П. П. Семенова-Тяньшанского — в тот период руководителя Центрального статистического комитета России.

В советский период статистики была разработана методология системы хозяйственных балансов, постоянно совершенствовалась система показателей планирования, углубилась теория индексного метода, получили распространение экономико-математические модели и методы.

Современная государственная статистика является неотъемлемым элементом системы государственного регулирования. Произошедшие в России изменения в общественно-экономической и социальной жизни, переход на рыночные отношения вызвали потребность в значительном совершенствовании методологии статистики, в частности в комплексном пересмотре всей системы учета и статистики в стране в соответствии с принятой международной методологией «Системы национальных счетов», сокращенно СНС.

Необходимость в получении точной, объективной и аналитичной информации о состоянии и развитии экономики для принятия решения на всех уровнях управления постоянно усиливает интегрирующую роль статистики в создании государственной информационной инфраструктуры общенационального масштаба.

1.1.2. Предмет и метод статистики

Изучаемые статистикой явления и процессы многообразны. В первую очередь, статистика изучает все, что связано с экономической деятельностью общества, — производство и реализацию промышленной и сельскохозяйственной продукции, строительство и реконструкцию основных фондов, работу транспорта и связи, формирование и движение финансовых потоков. Статистические методы широко используются в анализе социальных процессов и явлений — занятости и безработицы, доходов населения, изучении общественного мнения и др. Большую роль играет статистика в технике и производственной деятельности, например в организации контроля качества продукции. Методы статистики применяются в бухгалтерском учете, экономическом анализе, менеджменте, маркетинге, социологии, логистике, страховании, оценочной деятельности и в других научных и практических областях.

В настоящее время под термином «статистика» чаще всего понимается следующее.

Статистика — это наука о сборе и анализе данных.

Поскольку данными является любой вид зарегистрированной информации, статистика играет важную роль во всех сферах деятельности человека. Известный английский статистик У. Дж. Рейхман еще в начале прошлого века заметил: «Едва ли не в каждом своем аспекте явления природы, а также человеческая и прочая деятельности поддаются сейчас измерению при помощи статистических показателей».

Статистика — это один из видов практической деятельности человека, цель которой — сбор, обработка и анализ данных о различных социальных, общественных и экономических явлениях.

Исторически данную функцию статистики реализует государство. В нашем государстве эта работа выполняется Федеральной службой государственной статистики. В задачу таких государственных и ведомственных органов входят сбор, обобщение и публикация статистических данных в целом по стране (например, в России выходит статистический ежегодник «Россия в цифрах»), а также по регионам и отраслям экономики.

Статистикой называют также числовые (цифровые) или количественные данные, характеризующие различные объекты и явления. Но, прежде всего, статистика — это самостоятельная общественная наука, которая имеет свой предмет и специфические методы исследования.

Статистика — это наука, которая изучает количественную сторону массовых социально-экономических явлений и процессов в неразрывной связи с их качественной стороной, выявляет основные характеристики и закономерности развития этих явлений и процессов в конкретных условиях места и времени.

Из данного определения следуют основные особенности предмета статистики и статистики как науки.

Предметом статистики является количественная сторона массовых качественно определенных социально-экономических явлений и процессов, отображаемых посредством статистических показателей. Особенности статистики как науки заключаются в следующем:

1. В отличие от других общественных наук статистика ***изучает количественную сторону общественных явлений***. То есть статистика выявляет основные характеристики различных явлений общества и экономики, характеризует их, сравнивает между собой, анализирует происходящие с этими явлениями изменения, используя ***числа*** или ***цифры***.

2. Статистика ***изучает массовые явления и выявляет основные закономерности их развития***. Очевидно, что для установления закономерности развития процесса в целом необходимо проанализировать множество единичных случаев. Например, при выборе нового поставщика упаковки принять решение о том, чье качество продукции выше, невозможно, проверив лишь по одному элементу упаковки из предлагаемых поставщиками пробных партий. Для ответа на этот вопрос необходимо проанализировать качество всех элементов пробных партий обоих поставщиков, обобщить результаты и на основании их сравнения

принимать решение. Желательно при этом учесть и другие сведения, например: изучить статистику рекламаций на продукцию каждого поставщика, проверить соответствие результатов тестирования пробной партии техническим характеристикам упаковки, указанным в документации и др. В любом случае вначале потребуется обобщить, а затем сравнить результаты собранной информации. Именно поэтому статистика изучает не единичные факты, а массовые социально-экономические явления. При этом отдельные факты не игнорируются, а рассматриваются как составляющие общего явления.

Кроме того, в статистике важную роль играет ***закон больших чисел*** — общий принцип, в силу которого количественные закономерности, присущие массовым явлениям, отчетливо проявляются лишь при достаточно большом числе наблюдений. Единичные явления в большей степени подвержены действию случайных и несущественных факторов, чем масса в целом. При большом числе наблюдений случайные отклонения от общей закономерности развития взаимно погашаются. В результате этого обобщающие показатели становятся типичными, отражающими действие только постоянных и существенных факторов.

3. Для статистики ***исключительно важна качественная определенность используемых числовых данных***. Прежде чем делать обобщения и сравнивать результаты измерений, необходимо удостовериться в качественной однородности собранных данных. Например, сравнивая технические характеристики картонной упаковки различных производителей, необходимо удостовериться, что сравниваемые параметры плотности, влагостойкости относятся к картону одинаковой толщины.

Важно помнить также о соблюдении единства единиц изменения в анализируемых данных и учитывать временную определенность информации. Например, несмотря на то, что математически операция сложения размера выручки от реализации компании в июне и в декабре 2006 г. допустима, реально к какому-либо экономически значимому результату она не приведет, в отличие от суммирования размера выручки за два или более последовательно идущих месяца. Также бессмысленно из-за разницы в числовом выражении суммировать размер выручки, выраженный в миллионах рублей, с величиной выручки, выраженной в рублях или тысячах рублей.

Для того чтобы получить общее представление о статистической методологии, необходимо рассмотреть сам *процесс статистического исследования*, который включает четыре основных этапа.

1 этап. Сбор первичного статистического материала, проверка его полноты и достоверности. С этой целью применяются специальные методы *статистического наблюдения*. От качества полученных исходных данных во многом зависят окончательные результаты всего исследования.

2 этап. Предварительная обработка данных. Собранную информацию необходимо упорядочить, систематизировать, т. е. подготовить для дальнейшей работы. Основной метод, используемый на данном этапе, — *метод группировок*. В результате его реализации от больших массивов данных исследователь переходит к компактным и удобным для анализа статистическим таблицам.

3 этап. Расчет и интерпретация обобщающих статистических показателей. На данном этапе рассчитываются показатели *среднего уровня* и *вариации*, *структуры*, *взаимосвязи* и *динамики* изучаемых процессов и явлений.

4 этап. *Моделирование взаимосвязей* между социально-экономическими процессами и явлениями и выявление *закономерностей* их развития.

Полученные на каждом этапе исследования результаты *обязательно интерпретируются* — исследователь делает анализ, проверяет значимость и точность рассчитанных показателей и *формулирует выводы*. Используемые в процессе реализации всех этапов статистические приемы и методы в целом составляют методологию статистики.

Для повышения эффективности и оперативности проведения исследований с большим набором данных разработаны и успешно применяются различные автоматизированные пакеты обработки статистической информации. Наиболее распространенными среди них являются: пакет программ статистического анализа STATISTICA, пакет программ SPSS. Базовые статистические методы обработки информации являются составляющими большинства широко применяющихся программ для обработки и изображения данных, примерами таких программ являются MS Excel, Matcad и большинство математических пакетов. Применение автоматизированных специальных пакетов анализа и прогнозирования информации постоянно растет. На базе платформ, созданных компаниями SAP, SAS, SPSS, STATISTICA, другими компаниями-разработчиками, в настоящее время сформировано профессиональное отраслевое программное обеспечение, которое является комплексным инструментом, предназначенным для конкретных отраслей и видов деятельности, позволяющим не только эффективно анализировать информацию, но и оптимизировать управление бизнес-процессами на предприятиях этих отраслей в целом.

1.1.3. Отрасли статистики

Статистика изучает различные аспекты жизни общества. В зависимости от специфики объекта исследования различают отдельные отрасли статистики (рис. 1.1.1).

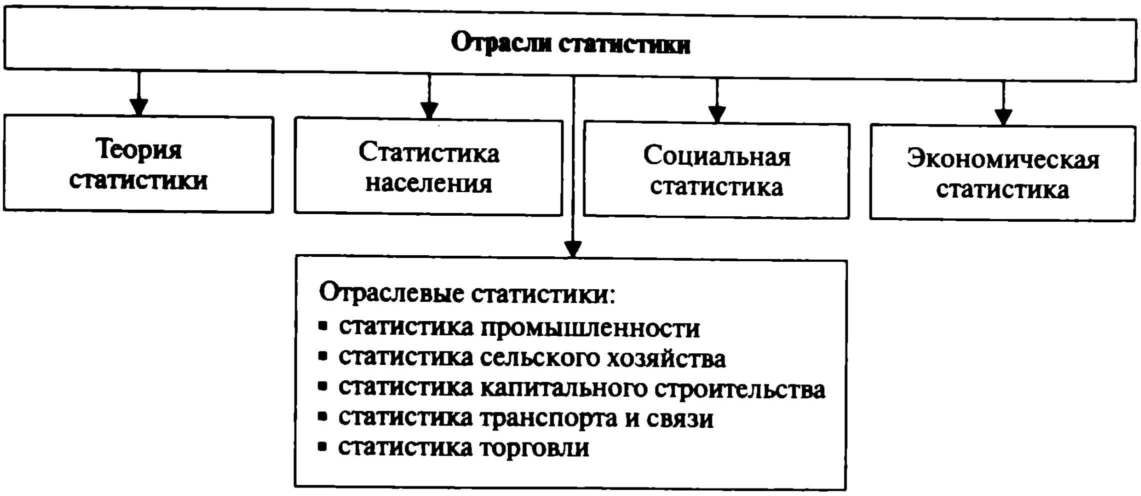


Рис. 1.1.1. Отрасли статистики

Теория статистики рассматривает общие понятия и методы сбора, обработки и анализа данных, разрабатывает общие показатели и методы изучения структуры, взаимосвязи и динамики социально-экономических процессов и явлений. Эти методы и показатели используются другими отраслевыми статистиками.

Экономическая статистика изучает количественные закономерности происходящих в экономике явлений и процессов, выявляет основные тенденции экономического развития и их воздействие на уровень жизни населения на макроуровне, т. е. в масштабе крупного региона или страны в целом. К основным показателям экономической статистики относятся валовой внутренний продукт, валовой региональный продукт, такие элементы национального богатства, как основные фонды, материальные и оборотные средства, имущество домашних хозяйств.

Отраслевые статистики изучают основные показатели, закономерности и тенденции развития отраслей экономики.

Статистика населения изучает численный и национальный состав, а также возрастно-половую структуру населения, его размещение и воспроизводство как по стране в целом, так и в разрезе территориальных единиц. Одной из основных задач статистики населения является построение краткосрочных и долгосрочных демографических прогнозов, в том числе прогнозирование ожидаемой продолжительности жизни.

Социальная статистика изучает социальную структуру населения, его уровень жизни и, в частности, доходы, уровень образования и культуры, состояние здоровья и медицинского обслуживания, использование свободного времени, общественное мнение, уровень преступности и другие аспекты социальных отношений.

1.1.4. Основные категории статистики

Как любая наука статистика, имеет свои специальные термины-категории, которые находятся в постоянном обиходе и используются при объяснении других понятий и методов. К важнейшим категориям статистики относятся: ***признак***, ***статистическая совокупность***, ***единица статистической совокупности***, ***вариация***, ***статистический показатель***.

Признак — это объективная характеристика какого-либо объекта или явления, характерная черта или свойство, которое может быть определено или измерено. Именно значения различных признаков наблюдаются и регистрируются на первом этапе исследования — наблюдении. В статистических пакетах категория «признак» часто обозначена как **«переменная»**.

Например, признаками, характеризующими промышленное предприятие, являются выручка от реализации продукции, прибыль, стоимость основных фондов, численность персонала и др. Человека можно охарактеризовать признаками: возраст, пол, место жительства, профессия, размер дохода.

Значение, которое может принимать признак, называется **вариантом**. Например, существуют всего четыре варианта значений признака «экзаменационная оценка»: «2», «3», «4», «5». Однако вариантов количества баллов, необходимого для получения определенной оценки, гораздо больше — от 0 до 100 включительно, и каждой оценке соответствует свой диапазон количества баллов: «3» — от 50 до 69, «4» — от 70 до 89, «5» — от 90 до 100. Наиболее распространенными вариантами признака «число комнат в квартире» являются 1, 2 и 3 комнаты, однако на рынке жилья представлены и четырех-, и пяти-, и шестикомнатные квартиры. А если рассматривать коттеджное жилье, то для характеристики размеров коттеджей более уместно использовать варианты общего размера площади и варианты количества этажей, среди которых, кстати, наиболее распространенными также являются варианты «1», «2» и «3».

Все признаки или переменные по содержанию принимаемых ими вариантов подразделяются на количественные и качественные (рис. 1.1.2).

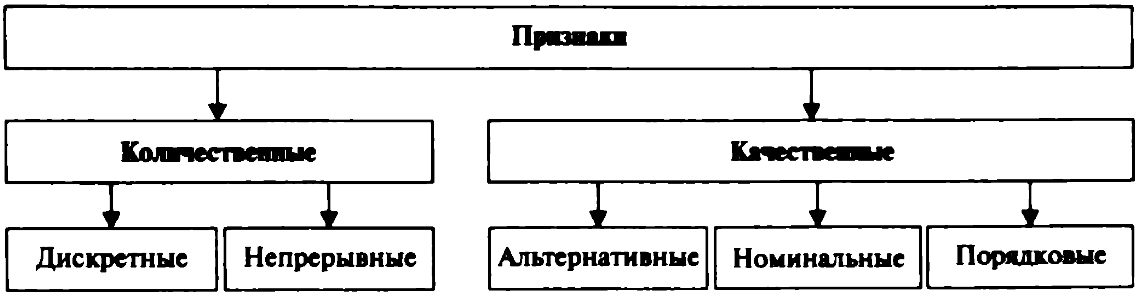


Рис. 1.1.2. Классификация видов признаков

Количественным является признак, варианты которого имеют числовое выражение и отражают размеры, масштабы некоторого объекта или явления. К количественным признакам, например, относятся доход домохозяйства, площадь жилого помещения, цена товара, стаж работы. Количественные признаки в статистике преобладают над другими видами признаков, они наиболее информативны, аналитичны, именно на их использование нацелена большая часть статистического инструментария.

Непрерывные количественные признаки могут принимать любые числовые значения, например, прибыль предприятия может быть и положительной, причем от одного рубля до триллионов рублей, и нулевой, и даже отрицательной, если предприятие терпит убытки. Единицы измерения непрерывных признаков, как правило, могут быть подвержены укрупнению или разделению. Например, время на выполнение вычислительной операции можно указать в наносекундах (миллиардная часть секунды), секундах, минутах, десятых долях часа. *Дискретные* количественные признаки принимают значения, ограниченные определенным диапазоном, широта которого зависит от особенностей признака. Как правило, варианты дискретных признаков выражаются целыми и неделимыми единицами измерения. Например, при изучении семей по числу детей типичными вариантами признака будут 0, 1, 2 и 3, хотя есть семьи, в которых растут и воспитываются и 4, и 5, и большее количество детей.

Качественные признаки, в отличие от количественных, не поддаются прямому числовому описанию. Они характеризуют различные свойства объектов и явлений, отражают их состояние или отдельные характеристики. Например, к качественным признакам относятся форма собственности предприятия, пол человека, уровень образования, квалификация специалиста.

Качественный признак называется ***альтернативным***, если он имеет только два варианта значений. Например, пол человека может быть мужским или женским, население страны или региона обычно делится на городское и сельское. Альтернативный признак может иметь и числовое выражение. Например, семья может иметь доход в размере «до 15 тыс. руб. в месяц» и «15 тыс. руб. в месяц и более».

В отличие от альтернативного, ***номинальный*** признак имеет более двух вариантов, которые выражаются в виде понятий или наименований. Например, район проживания, вид продукции, специальность, цвет автомобиля. Такие признаки имеют место в различных областях исследования, но чаще с ними работают маркетологи, социологи и психологи.

Порядковые признаки отличаются от атрибутивных тем, что имеют несколько ранжированных, т. е. упорядоченных по возрастанию или убыванию, качественных вариантов. Например, уровень образования (начальное, общее среднее, среднее полное общее, начальное профессиональное, среднее профессиональное и т. д.), уровень квалификации (врач первой категории, врач высшей категории и т. д.), воинское звание. Отдельные варианты порядкового признака трудно соизмерить количественно. Например, понятно, что образованность человека с высшим образованием выше, чем со средним специальным, но при этом нельзя утверждать, что она выше на 20% или на 30%. Водительская категория «Е» выше, чем категория «В», но количественных пропорций между ними не существует. Порядковый признак может иметь числовое выражение. Например, тарифный разряд рабочего или служащего, рейтинговые и экзаменационные оценки. Однако варианты таких признаков также не имеют количественных пропорций: рабочий 6-го разряда не обязательно в два раза больше вырабатывает продукции и в два раза больше зарабатывает, чем рабочий 3-го разряда.

Статистика изучает социально-экономические явления комплексно, поэтому, при исследовании, отдельные объекты или явления объединяются по их качественному содержанию в специальные группы.

Совокупность — это реально существующее множество объектов или явлений, объединенных хотя бы одним общим одинаковым признаком и обладающих внутренней взаимосвязью и целостностью. Именно совокупность является объектом исследования в статистике.

Статистика имеет дело с совокупностями промышленных, сельскохозяйственных, строительных и торговых предприятий, с совокупностью коммерческих банков, с совокупностью населения страны или отдельного ее региона. Так, например, всех жителей г. Москвы можно рассматривать как статистическую совокупность, так как один признак — город проживания — будет у всех одинаковым. Индивидуальный объект или явление, составляющее статистическую совокупность, называется **единицей совокупности**.

При проведении исследования именно у единиц совокупности измеряются или фиксируются важные для исследования признаки. Для отрасли единицей совокупности будет являться отдельное предприятие, для банковской системы — отдельный банк. В некоторых случаях для одной и той же совокупности можно выделить разные группы единиц. Например, при изучении половозрастной структуры населения единицей является отдельный человек, при изучении доходов, обеспеченности жильем и предметами длительного пользования (телевизоры, холодильники и т. п.) единицей будет являться семья или домохозяйство.

Общее число единиц, образующих статистическую совокупность, называется ***объемом совокупности***. Объем совокупности следует отличать от ***объема признака***, т. е. суммарного значения признака по всем единицам изучаемой совокупности. Так, число предприятий в отрасли — это объем совокупности, а общий выпуск продукции всеми предприятиями отрасли — это объем признака.

Одной из важнейших характеристик статистической совокупности является ее однородность. ***Однородной*** является совокупность, единицы которой близки между собой по значениям признаков, существенных для данного исследования. Многие методы и приемы статистического исследования применимы лишь к однородным совокупностям.

Внутри статистической совокупности отдельные объекты и явления могут сильно отличаться друг от друга по своим свойствам и характеристикам. Например, у единиц совокупности «Население г. Москвы» кроме общего города проживания есть множество признаков, по которым единицы совокупности — москвичи — отличаются друг от друга: возраст, уровень образования, размер дохода. На языке статистики этот пример звучит так: население Москвы *варьирует* по возрасту, уровню образования, доходу. Или: внутри совокупности «Население г. Москвы» *существует вариация признаков* — «возраст», «уровень образования» и «доход».

Вариация — это изменения, колебания величины признака у различных единиц совокупности, т. е. принятие единицами совокупности разных вариантов значений признака.

Именно *наличие вариации и изменчивость данных определяют необходимость статистики как науки* и делают неизбежным процесс постоянного отслеживания и анализа информации о том, как развиваются общество и экономика, как работает та или иная отрасль, насколько эффективно функционирует конкретное предприятие. Что является причиной такой изменчивости?

Во-первых, *индивидуальные особенности изучаемых объектов*. Например, коммерческие банки варьируют по величине активов в зависимости от размера уставного фонда и реализуемой финансовой политики, работники предприятия имеют разную производительность труда в зависимости от их квалификации и мотивации, но также на производительность труда влияют состояние здоровья, настроение работника и общая атмосфера в трудовом коллективе.

Во-вторых, *вариация происходит под воздействием определенных условий, в том числе и непредвиденных* (*форс-мажорных*). Например, вариация городов по численности населения зависит от географического положения города, развитости городской инфраструктуры и экономического положения в регионе. Спрос на туристические поездки в Таиланд в начале 2005 г. значительно снизился после разрушений, вызванных серией цунами, что отразилось на выручке туристических операторов, организующих отдых в этой стране.

Статистическое исследование всегда завершается расчетом и анализом различных по виду и форме выражения статистических показателей.

Статистический показатель представляет собой количественную характеристику социально-экономических явлений и процессов в конкретных условиях места и времени.

Статистический показатель непосредственно связан с внутренним содержанием изучаемого явления или процесса. Все показатели можно классифицировать по охвату единиц совокупности, по временной и пространственной определенности (рис. 1.1.3).

Рис. 1.1.3. Классификация статистических показателей

Изучаемые статистикой процессы и явления достаточно сложны и не всегда могут быть описаны одним отдельно взятым показателем. В таких случаях используется ***система статистических показателей***, позволяющих учесть различные характеристики изучаемого объекта или явления. Так, например, изучая промышленное предприятие, нельзя ограничиться только величиной выпуска продукции. Для полной экономической характеристики необходимо учесть численность персонала, производительность его труда, стоимость основных производственных фондов, финансовые показатели — затраты, прибыль, инвестиции в развитие производства.

В отличие от признака статистический показатель получается расчетным путем. Это может быть простой подсчет единиц совокупности, суммирование значений признака части или всех единиц совокупности, сравнение двух или нескольких величин, а также более сложные расчеты.

КОНТРОЛЬНЫЕ ВОПРОСЫ

1. В каких значениях может употребляться термин «статистика» в наше время?
2. Дайте характеристику основным чертам определения предмета статистики.

3. Почему статистика изучает массовые явления?
4. К каким видам (количественным или качественным) относятся следующие признаки:
   а) количество работников на предприятии;
   б) пол и возраст человека;
   в) этажность жилых помещений;
   г) количество детей в семье;
   д) прибыль торговых объединений.
5. Дайте определение статистической совокупности, приведите примеры.
6. В чем отличие статистического признака от статистического показателя?

Глава 1.2

ИСТОЧНИКИ И МЕТОДЫ СБОРА СТАТИСТИЧЕСКОЙ ИНФОРМАЦИИ. СТАТИСТИЧЕСКОЕ НАБЛЮДЕНИЕ

- Сущность и виды статистического наблюдения
- Точность статистического наблюдения

1.2.1. Сущность и виды статистического наблюдения

Статистическое наблюдение — это массовый, планомерный, научно организованный сбор данных о социально-экономических явлениях и процессах. Статистическое наблюдение заключается в регистрации отобранных признаков у каждой единицы изучаемой совокупности. Например, при переписи населения по каждому жителю страны регистрируются сведения о поле, возрасте, семейном положении, образовании и др.

Статистическое наблюдение, как правило, носит массовый характер — для получения наиболее точных данных, выявления закономерностей и взаимосвязей внутри изучаемой совокупности при проведении наблюдения необходимо получить данные от максимально возможного числа составляющих ее единиц.

Любое исследование проводится по заранее разработанному плану. Структура элементов **плана статистического наблюдения** представлена ниже.

I. Программно-методологические вопросы наблюдения

1.1. **Объект наблюдения** — совокупность, о которой должны быть собраны нужные сведения.

1.2. **Единица наблюдения** — единица изучаемой совокупности (человек, предприятие, банк), являющаяся носителем признаков, подлежащих регистрации.

1.3. **Программа наблюдения** — перечень существенных признаков, подлежащих регистрации, и перечень вопросов для их регистрации.

1.4. **Период наблюдения** — время, в течение которого осуществляется регистрация признаков единиц наблюдения по установленной программе.

1.5. **Критический момент наблюдения** — момент времени, по состоянию на который производится регистрация собираемых сведений. На практике критический момент назначается на начало периода наблюдения.

1.6. **Формуляр наблюдения** — специальный бланк для записи ответов на вопросы программы. Формуляр включает *титульную часть*, где указывается вид наблюдения; наименование организации, которая проводит наблюдение; кем и когда утвержден формуляр; *программную часть* — перечень вопросов программы, места для записи ответов на них и коды ответов; *адресную часть*, где указывается точный адрес каждой единицы или группы единиц наблюдения и др. сведения.

1.7. Инструкция по заполнению формуляра наблюдения.

II. Организационные вопросы наблюдения

2.1. Описание особенностей объекта, целей и задач наблюдения, организации (подразделения), осуществляющей(его) наблюдение.

2.2. Место и даты проведения наблюдения.

2.3. Подготовительные работы к наблюдению: подбор и обучение кадров, составление списков единиц совокупности, планирование мероприятий рекламной кампании проводимого наблюдения и т. д.

2.4. Порядок приема и сдачи материалов наблюдения и представления предварительных и окончательных итогов наблюдения.

2.5. Финансирование и материально-техническое обеспечение работ.

Основная цель статистического наблюдения — это сбор статистической информации о социально-экономических явлениях и процессах для получения обобщающих характеристик.

Различают три основные формы статистического наблюдения (рис. 1.2.1).

Рис. 1.2.1. Формы статистического наблюдения

Статистическая отчетность — это способ получения статистической информации от юридических лиц. Отчетность представляет собой специально разработанные формы, включающие в себя те признаки, которые подлежат регистрации. Формы статистической отчетности разрабатываются и утверждаются органами государственной статистики РФ. Любое юридическое лицо, являющееся субъектом экономики РФ, обязано предоставлять отчетность органам государственной статистики по месту своей регистрации по установленным отчетным формам и в установленные сроки.

Специально организованные статистические наблюдения проводятся для получения данных, не содержащихся в предоставляемой отчетности, или необходимых для проверки или уточнения данных, содержащихся в отчетах. Особым видом организованного статистического наблюдения является перепись.

Перепись — это специально проводимые широкомасштабные работы по сбору необходимой информации об изучаемых объектах в границах отрасли, региона или страны в целом.

Перепись населения — это организация сбора, обработки и публикации демографических, экономических и социальных данных обо всем населении, проживающем в определенный момент времени в стране.

Принципы переписи населения: всеобщность охвата населения переписью; непосредственное получение сведений путем опроса конкретных людей; самоопределение людей при ответах на вопросы (без предъявления подтверждающих сведения документов); конфиденциальность сообщаемых конкретным человеком сведений (результаты переписи публикуются только в сводном виде по стране, краю, области).

В 2006 г. по состоянию на 1 июля проводилась Всероссийская сельскохозяйственная перепись — массовый сбор данных от юридических и физических лиц, которые имеют земельные участки, предназначенные или используемые для производства сельскохозяйственной продукции, либо имеют сельскохозяйственных животных. Результаты этой переписи будут иметь важное значение для разработки эффективной агропромышленной политики и формирования объективной информации о состоянии продовольственного комплекса, который влияет на продовольственную и экономическую безопасность страны. Органами статистики проводятся также переписи многолетних насаждений, жилого фонда, незавершенного строительства, объектов культурного наследия.

Кроме переписей к специально организованному наблюдению относятся другие единовременные работы по сбору информации, в частности, в рамках социологических или маркетинговых исследований.

Регистровое наблюдение представляет собой постоянный мониторинг состояния и развития наблюдаемых единиц, заключающийся в размещении и своевременной актуализации информации о них в базе данных. В статистической практике ряда стран применяют регистры населения — постоянно актуализируемые списки жителей страны с указанием их основных социально-демографических признаков и регистры предприятий.

В России регистр предприятий называется *Статистический регистр хозяйствующих субъектов* — база данных об организациях, созданных на территории РФ, их местных единицах, индивидуальных предпринимателях и других типах хозяйствующих субъектов. Этот регистр ведет Федеральная служба государственной статистики. Регистры предприятий содержат данные о времени создания (регистрации) предприятия, его названии и адресе, об организационно-правовой форме, структуре, виде экономической деятельности, количестве занятых, основных экономических показателях из данных бухгалтерской отчетности и др.

Классификация видов наблюдения представлена на рис. 1.2.2.

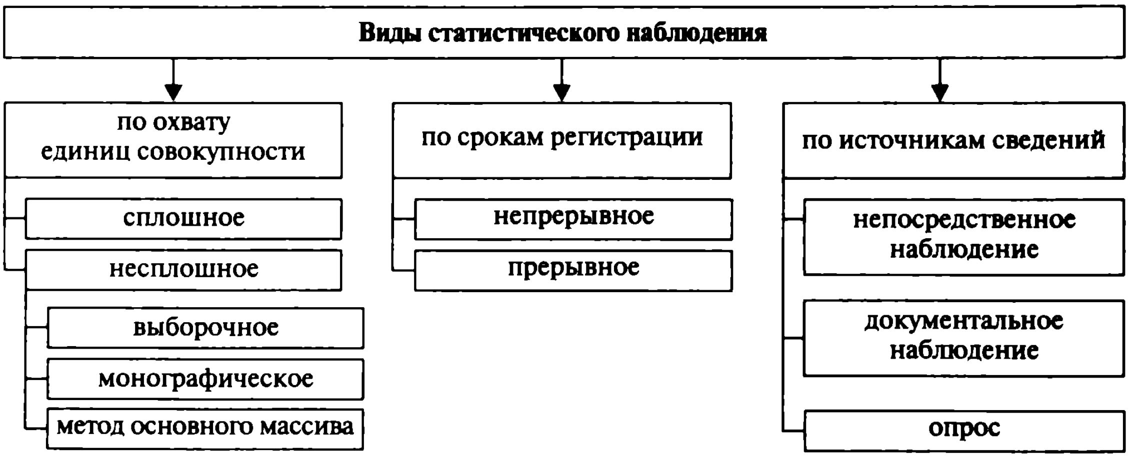


Рис. 1.2.2. Виды статистического наблюдения

По *охвату единиц совокупности* наблюдение бывает двух видов: сплошное и несплошное. При **сплошном наблюдении** обследованию подвергаются все единицы изучаемой совокупности. Примером сплошного наблюдения могут служить переписи.

При **несплошном наблюдении** обследованию подвергается только часть единиц изучаемой совокупности. Различают следующие виды несплошного наблюдения: выборочное, метод основного массива, монографическое обследование.

Выборочным называют наблюдение, основанное на принципе случайного отбора единиц изучаемой совокупности, которые должны быть подвергнуты наблюдению. Выборочное наблюдение, при правильной его организации и проведении, дает достаточно достоверные данные для характеристики изучаемой совокупности в целом. Во многих случаях им вполне можно заменить сплошной учет. При этом обеспечивается значительная экономия средств, затрачиваемых на сбор и обработку данных.

Монографическое обследование представляет собой детальное, глубокое изучение и описание отдельных единиц совокупности. Такое исследование проводится с целью выявления имеющихся резервов, оценки результатов экономических экспериментов.

Метод основного массива заключается в том, что обследованию подвергаются наиболее крупные единицы, которые имеют преобладающий удельный вес в совокупности по основному для данного исследования признаку. Например, в ряде отраслей добывающей и обрабатывающей промышленностей подавляющий объем выпуска продукции приходится на крупные и средние предприятия, поэтому результаты деятельности малых предприятий в этих отраслях практически не отражаются на обобщающих статистических показателях.

По *срокам регистрации* наблюдение может быть непрерывным (текущим) и прерывным. **Непрерывным** называют наблюдение, которое ведется постоянно, и регистрация фактов производится по мере их свершения. Так, например, осуществляется регистрация рождений, заключенных браков, разводов и др. в органах ЗАГС.

Прерывное наблюдение повторяется через определенные равные промежутки времени, например, ежегодное предоставление отчетности в органы государственной статистики или по мере необходимости, без соблюдения строгой периодичности, как, например, перепись многолетних насаждений, проведенная один раз в прошлом веке.

По *источнику сведений* различают непосредственное наблюдение, документальное наблюдение и опрос.

Непосредственным называют такое наблюдение, при котором сами регистраторы путем непосредственного замера, взвешивания или подсчета устанавливают значение признака и производят запись в формуляре наблюдения. Например, инвентаризация основных средств на предприятиях.

Документальное наблюдение предполагает запись ответов на вопросы формуляра на основании соответствующих документов. Например, сбор данных об успеваемости студентов вуза на основе зачетно-экзаменационных ведомостей, заполнение форм статистической отчетности на основании данных бухгалтерского учета и т. п.

Опрос — это наблюдение, при котором ответы на вопросы формуляра записываются со слов опрашиваемого (респондента). Этим способом проводятся переписи населения, опросы общественного мнения.

В статистике применяются следующие **способы сбора сведений**: *отчетный*, *экспедиционный*, *самоисчисление*, *анкетный*.

Отчетный способ заключается в обязательном представлении хозяйствующими субъектами статистических отчетов о своей деятельности в установленной форме и в установленные сроки.

Экспедиционный способ наблюдения заключается в том, что специально привлеченные и обученные работники посещают каждую единицу наблюдения и сами заполняют формуляр наблюдения. Этим способом собираются сведения при переписях населения.

При способе **самоисчисления** формуляры заполняют сами опрашиваемые. Обязанность специально привлеченных для получения информации сотрудников состоит в раздаче формуляров, инструктаже опрашиваемых, сборе и проверке правильности заполнения формуляров.

Анкетный способ — это сбор данных с помощью специальных вопросников, рассылаемых определенному кругу лиц или публикуемых в периодической печати. Как правило, этим способом получения информации пользуются при проведении социологических опросов. Также многие крупные производители бытовой техники, мебели и других предметов потребления вкладывают анкеты в упаковку товара с просьбой заполнить и вернуть их производителю по указанному адресу.

1.2.2. Точность статистического наблюдения

Под точностью в статистике понимают степень соответствия данных наблюдения их реальным значениям. Возникающие расхождения называются ошибками. Ошибки определяются как разность или как отношение между этими значениями. Как правило, ошибки возникают при регистрации сведений или при измерении.

Ошибки регистрации возникают вследствие неправильного установления фактов в процессе наблюдения, или ошибочной их записи, или того и другого вместе. *Случайные ошибки* возникают в результате опечаток, описок, оговорок. Например, вместо возраста человека «35 лет» указано «5 лет», у Сидоровой Елены Ивановны в графе пол отмечен «Мужской» и т. п. При большом числе наблюдений благодаря действию закона больших чисел эти ошибки более или менее взаимно погашаются.

Систематические ошибки наиболее опасны, поскольку приводят к сильному искажению данных. Наиболее показательными ошибками являются: занижение или округление населением своего возраста до цифр, оканчивающихся на 5 или 0; сокрытие экономическими субъектами реальных размеров финансовых результатов их деятельности, стремление респондентов указать заниженное значение своего дохода и т. п.

При несплошном наблюдении возникают **ошибки репрезентативности** (или ошибки представительности). Они заключаются в том, что значения признаков по отобранной выборочной совокупности не отражают реально существующей картины в целом.

С целью выявления ошибок проводится контроль полученных материалов. После проведения наблюдения весь собранный материал проверяют на полноту охвата единиц. Если выявлены упущенные единицы наблюдения, дальнейшие действия зависят от того, представляется возможным восполнение пробелов или нет. Для проверки качества заполнения формуляров и других документов наблюдения используют логический и арифметический контроль.

Логический контроль состоит в сопоставлении между собой ответов на вопросы формуляра наблюдения и выяснении их логической совместимости. При обнаружении несовместимых ответов пытаются путем дальнейших сопоставлений установить, какой из ответов является верным.

Арифметический контроль состоит в проверке различных расчетов, результаты которых проведены в формуляре наблюдения, в частности, итогов, вычисления процентов, расчетов средних величин и т. п.

КОНТРОЛЬНЫЕ ВОПРОСЫ

1. В чем состоят преимущества и недостатки сплошного и несплошного статистических наблюдений?
2. Что такое перепись населения и почему необходимо ее проводить?
3. Для чего формируются регистры предприятий?
4. Почему все зарегистрированные хозяйствующие субъекты обязаны предоставлять статистическую отчетность в органы государственной статистики?
5. К какому виду наблюдения относится регистрация факта смерти или рождения человека?
6. Что называется ошибкой регистрации? Приведите примеры таких ошибок.

Глава 1.3

ПРЕДСТАВЛЕНИЕ СТАТИСТИЧЕСКИХ ДАННЫХ С ПОМОЩЬЮ ТАБЛИЦ И ГРАФИКОВ

- Статистическая таблица и ее элементы
- Виды статистических таблиц
- Основные правила построения статистических таблиц
- Основные принципы применения графического метода в статистике
- Классификация основных видов статистических графиков

1.3.1. Статистическая таблица и ее элементы

Результаты сводки и группировки материалов статистического наблюдения, как правило, представляются в виде таблиц.

Статистическая таблица — это таблица, которая содержит сводную числовую характеристику исследуемой совокупности по одному или нескольким существенным признакам.

Табличной называется форма расположения числовой информации, при которой число располагается на пересечении четко сформулированного заголовка по вертикальному столбцу, называемому *графой*, и названия по соответствующей горизонтальной полосе — *строке*. Таблица представляет собой пересечение граф и строк, которые формируют остов (основу) таблицы (рис. 1.3.1).

Таблица «*N*»

Название таблицы (общий заголовок)*

Сказуемое / Подлежащее	Заголовки граф (колонок, столбцов)				Итоговая графа**
А	1	2	...	*N* – 1	*N*
Заголовки строк (боковые заголовки)					
Итоговая строка					

Служебная строка — строка «*А*, 1, 2, ..., *N* – 1, *N*».

* Примечания к таблице (источник данных, комментарий к отдельным значениям, разъяснения по методологии расчета данных и т. п.).

** Макет таблицы может не содержать итоговой графы.

Рис. 1.3.1. Основа статистической таблицы

Статистическая таблица содержит три вида заголовков: общий, верхние и боковые. Общий заголовок отражает содержание всей таблицы (к какому объекту, месту и времени относится числовая информация) и располагается над макетом таблицы по центру. Верхние заголовки характеризуют содержание граф (столбцов), а боковые — строк. Остов таблицы, заполненный заголовками, образует **макет** таблицы.

По логическому содержанию таблица представляет собой «статистическое предложение», состоящее из подлежащего и сказуемого.

Подлежащим статистической таблицы называется объект, который характеризуется цифрами. Обычно подлежащее таблицы дается в левой части, в наименовании строк.

Сказуемое статистической таблицы образует система показателей, которыми характеризуется подлежащее таблицы. Сказуемое формирует верхние заголовки и составляет содержание граф. Расположение подлежащего и сказуемого в отдельных случаях может меняться местами в соответствии с целями представления и анализа информации.

1.3.2. Виды статистических таблиц

По структуре подлежащего различают статистические таблицы простые и сложные (рис. 1.3.2).

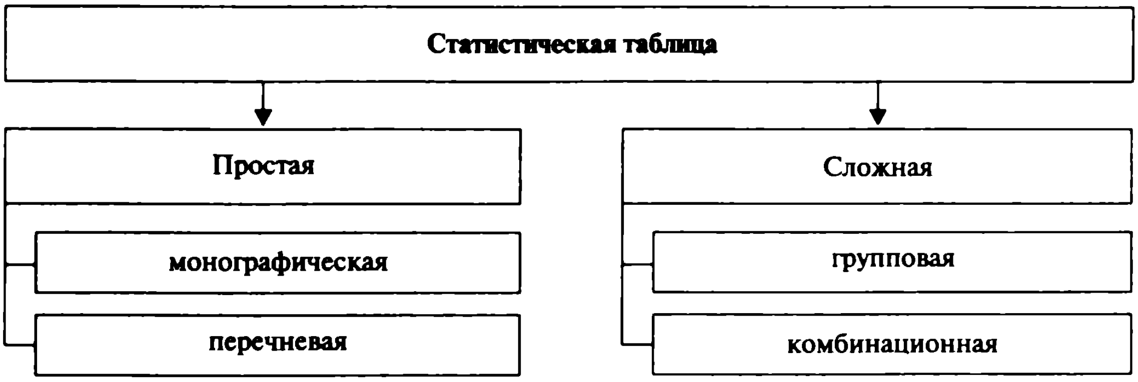


Рис. 1.3.2. Виды статистических таблиц по структуре подлежащего

Простой называется таблица, в подлежащем которой дается перечень каких-либо объектов или территориальных единиц. Простые *монографические* таблицы характеризуют не всю совокупность, а только какую-либо одну ее группу (табл. 1.3.1).

Указав в подлежащем табл. 1.3.1 наименования врачебных специальностей, получаем *простую перечневую* таблицу (табл. 1.3.2).

Таблица 1.3.1

Обеспеченность населения Российской Федерации врачами

Показатель	2003 г.	2004 г.	2005 г.
А	1	2	3
Численность врачей, тыс. чел.	686,0	688,2	690,3

Таблица 1.3.2

Обеспеченность населения врачами по отдельным специальностям

Специальности	2003 г.	2004 г.	2005 г.
А	1	2	3
Численность врачей, тыс. чел.	686,0	688,2	690,3
из них:			
терапевтов	158,3	158,5	159,7
хирургов	64,5	65,0	65,0
педиатров	69,3	68,7	68,6

Групповыми называются таблицы, подлежащее которых содержит группировку единиц совокупности по одному количественному или качественному признаку. Групповые таблицы позволяют выявить и охарактеризовать социально-экономические типы явлений, их структуру в зависимости только от одного признака (табл. 1.3.3).

Таблица 1.3.3

Распределение населения Российской Федерации по возрастным группам в 2006 г. (на начало года)

Группы населения по возрасту, лет	Численность населения	
	тыс. чел.	в % к итогу
1	2	3
0–9	13 403	9,4
10–19	20 710	14,4
20–29	23 148	16,2
30–39	19 623	13,7
40–49	23 550	16,4
50–59	18 174	12,6

Окончание табл. 1.3.3

Группы населения по возрасту, лет	Численность населения	
	тыс. чел.	в % к итогу
60–69	12 689	8,8
70 и более	12 177	8,5
Итого	**143 474**	**100,0**

Комбинационными называются таблицы, подлежащее которых содержит группировку объектов по двум и более признакам (табл. 1.3.4).

Таблица 1.3.4

Группировка предприятий пищевой промышленности по величине прибыли и численности промышленно-производственного персонала в 2004 г.

Группы предприятий по величине прибыли, млн руб.	Подгруппы предприятий по численности промышленно-производственного персонала, чел.	Число предприятий
1	2	3
50–100	250–300	7
	300–350	8
Итого по группе	–	15
100–150	250–300	3
	300–350	2
Итого по группе	–	5
Всего	–	20

Подлежащим в табл. 1.3.4 являются графы 1 и 2, т. е. группы предприятий по величине прибыли и подгруппы по численности промышленно-производственного персонала.

В сказуемом статистической таблицы указываются показатели, которые являются характеристикой изучаемого объекта. Различают статистические таблицы с простой и сложной разработкой сказуемого.

При **простой разработке сказуемого** показатель, определяющий его, не подразделяется на подгруппы и итоговые значения получаются путем простого суммирования значений по каждому признаку. Примером простой разработки сказуемого является табл. 1.3.5.

Подлежащее в данном примере расположено в графе «А», сказуемое данной статистической таблицы расположено в левой части таблицы —

в графах 1–5. Данный вид таблицы является примером простой разработки сказуемого.

Таблица 1.3.5

Численность населения Российской Федерации в 2001–2006 гг.

Годы	Все население, млн чел.	В том числе		В общей численности населения, %	
		городское	сельское	городское	сельское
А	1	2	3	4	5
2001	146,3	107,1	39,2	73,21	26,79
2002	145,2	106,4	38,8	73,28	26,72
2003	145,0	106,3	38,7	73,31	26,69
2004	144,2	105,8	38,4	73,37	26,63
2005	143,5	104,7	38,8	72,96	27,04
2006	142,8	104,1	38,7	72,90	27,10

Сложная разработка сказуемого предполагает деление признака, формирующего его, на подгруппы, например фрагмент табл. 1.3.6, где рассмотрена структура продавцов магазинов не только по полу, но внутри каждой категории еще и по стажу работы.

Таблица 1.3.6 (*фрагмент таблицы*)

Число магазинов и численность работающих продавцов в Московской области по состоянию на конец IV квартала 2006 г.

Группы магазинов по товарной специализации (наименования)	Число магазинов (всего)	Число продавцов (всего)	В том числе			
			Мужчин со стажем		Женщин со стажем	
			от 1 года до 3 лет	от 3 лет до 5 лет	от 1 года до 3 лет	от 3 лет до 5 лет

Сложная разработка сказуемого позволяет получить более полную и подробную характеристику изучаемого объекта.

1.3.3. Основные правила построения статистических таблиц

1. Таблица должна быть компактной, а цифровой материал легко восприниматься при чтении его слева направо и сверху вниз. Таблицы

нумеруются или в сквозном порядке, или через точку с указанием номера раздела или главы, в которой находится таблица. Номер таблицы указывается перед ее названием, как правило, в правой части после слова «Таблица», например «Таблица 1» или «Таблица 1.3» (последний пример указывает, что Таблица 1.3 является третьей по порядку таблицей в первом разделе или главе).

2. Заголовок таблицы и названия граф и строк должны быть четкими, краткими. В названии таблицы отражаются объект, признак, время и место совершения события. Например: «Курс евро на торгах ММВБ на 01.01.2007 г.». Названия таблицы, граф и строк пишутся без сокращений, за исключением общепринятых и сокращений единиц измерения показателей.

3. Таблица должна иметь итоговую строку «Итого» или «Всего». Она может завершать таблицу или быть первой, соединяясь с раскрывающими ее вариантами словами «В том числе».

4. Взаимосвязанные данные располагаются в соседних графах, например, число предприятий и удельный вес предприятий (в % к итогу). Графы с одинаковыми названиями, имеющие общие термины и единую смысловую нагрузку, объединяются общим заголовком.

5. Графы нумеруются в служебной строке. Графы слева, заполненные названием строк, принято обозначать буквами (А, Б, В и т. д.), а все последующие — номерами в порядке возрастания. Если таблица занимает несколько страниц, то со второй страницы таблицу можно продолжить без повторения верхних заголовков (содержания сказуемого), указав в начале страницы только содержание служебной строки.

6. В графах и строках указываются единицы измерения занесенных в них показателей. Если все данные в таблице имеют одинаковые единицы измерения, достаточно указать их один раз в названии таблицы.

7. Округление чисел в пределах одной графы следует проводить с одинаковой степенью точности (до целого числа, до десятых и т. д.). Очень малые значения данных обозначают (0,0) или (0,00).

8. Отсутствие данных отмечается в таблице многоточием «...» или «нет свед.». Если явление отсутствует полностью, то указывается прочерк (—). Если позиция вообще не подлежит заполнению, ставится знак «X».

9. Дополнения или пояснения к данным таблицы оформляются цифровой сноской и раскрываются в примечании в конце таблицы.

1.3.4. Основные принципы применения графического метода в статистике

Графиком в статистике называется изображение статистических данных в виде различных геометрических образов: точек, линий, фигур и т. п.

Визуализированные при помощи графика данные легче читаются и воспринимаются потребителем информации. Часто образное представление информации делает ее более убедительной. Например, убеждая руководство в двукратном увеличении эффективности работы компании после внедрения предложенной Вами инновационной системы контроля качества продукции, стоит построить сравнительный график эффективности работы компании до и после внедрения, поскольку двойное увеличение эффективности на графике «после» будет более показательным и, следовательно, убедительным.

При помощи графиков легче выявить закономерности развития, распределения и размещения явлений, оценить и понять соотношение между сравниваемыми объектами. Во многих случаях построение графиков помогает заменить громоздкие таблицы компактными изображениями и, кроме того, некоторые ошибки в данных выявляются именно при применении графиков.

Наиболее эффективным способом получения графического изображения данных является использование статистических или математических программ. Сегодня практически любая программа, предназначенная для ввода и обработки числовой информации, от Excel до прикладных отраслевых решений SAP или SPSS, оснащена средствами изображения данных. Однако общие принципы построения графиков неизменны.

Каждый график состоит из графического образа и вспомогательных элементов.

Графический образ (основа графика) — это геометрические знаки — точки, линии, фигуры, с помощью которых изображаются данные.

Вспомогательные элементы делают возможным чтение графика, его понимание и использование. К ним относятся: экспликация графика, масштабные ориентиры, поле графика.

Экспликация графика — словесное описание его содержания. Оно включает в себя общий заголовок графика, подписи вдоль масштабных шкал и пояснения к отдельным частям графика.

Заголовок графика должен в краткой и ясной форме отражать основное содержание (тему) изображаемых данных и объект, к которому они относятся. Если заголовок является частью текста (в книге, статье, дипломной работе), то он помещается под изображением графика. Если график представляется отдельно от текста (например, в приложении), заголовок указывается вверху графика более крупным шрифтом, чем все остальные надписи на графике.

На графике обязательно даются словесные пояснения условных знаков и смысла отдельных элементов графического образа. Сюда относятся названия и цифры масштабов, названия ломаных линий, цифры, характеризующие величины отдельных частей графика, ссылки на источники и т. д. Пояснительные надписи могут быть помещены на самом графике либо в виде **ключа или легенды**, вынесенных за его пределы.

Масштабные ориентиры статистического графика определяются *масштабом* и системой *масштабных шкал*.

Масштаб статистического графика — это мера перевода числовой величины в графическую. Например, 1 см высоты столбика равен 50 тыс. руб. уставного капитала коммерческого банка; 1 см^2 = 100 км^2 территории области. Масштабы выбирают так, чтобы на графике ясно выступало различие изображаемых величин и не терялась возможность их сравнения.

Масштабной шкалой называется линия, точки которой могут быть прочитаны как определенные числа. Шкала состоит из линии (или носителя шкалы), деления шкалы, цифрового обозначения делений. Как правило, цифровым обозначением снабжаются не все, а лишь некоторые деления. Числовое значение помещают строго против соответствующих точек, а не между ними (рис. 1.3.3).

Рис. 1.3.3. Масштабная сетка

Поле графика — то пространство, в котором размещаются образующие график геометрические знаки. Поле графика характеризуется его форматом, т. е. размером и пропорциями (соотношением сторон).

Считается, что наиболее удобной для восприятия глазом человека пропорцией является прямоугольник $1:\sqrt{2}$, т. е. 1 : 1,474 (примерно 5 : 7). Это сочетание принято в стандарте бумаги для копировально-множительной техники с форматом А4, т. е. 210 × 297 мм. Примерно такие же пропорции должны быть выдержаны и в размерах графических изображений.

1.3.5. Классификация основных видов статистических графиков

По ***форме графического образа*** различают графики объемные, линейные и плоскостные. По ***способу построения*** графики можно разделить на диаграммы и статистические карты (рис. 1.3.4).

Рис. 1.3.4. Классификация статистических графиков по способу построения и содержанию изображаемых данных

Диаграмма представляет собой чертеж, показывающий соотношение данных при помощи геометрических и изобразительных средств.

Статистические карты предназначены для графического изображения одноименных показателей, относящихся к разным территориям. Для этого в основу изображения берется географическая карта.

По ***содержанию*** можно выделить графики сравнения, графики структуры и динамики, графики взаимосвязи и размещения по территории.

Диаграммы сравнения применяются для изображения одноименных статистических данных, характеризующих разные территории или объекты.

Наиболее распространенным видом таких диаграмм являются ***столбиковые диаграммы***. Они представляют собой график, на котором размеры данных представлены расположенными в высоту прямоугольниками («столбиками») одинаковой или разной высоты. Столбиковые диаграммы применяются также для сравнения развития объектов или явлений во времени.

Высота столбика определяется по ***вертикальной масштабной шкале***. На ней записываются лишь круглые или округленные значения. На осях графика указываются единицы измерения каждого показателя.

Пример. Изобразим с помощью сравнительной столбиковой диаграммы данные о количестве предприятий торговли в Российской Федерации в 2005–2006 гг. (табл. 1.3.7).

Таблица 1.3.7

Число предприятий торговли в Российской Федерации в 2005–2006 гг. (тыс.)

Предприятия торговли	2005 г.	2006 г.
Организации, осуществляющие торговлю автотранспортными средствами и мотоциклами, их техническое обслуживание и ремонт	54,3	61,7
Организации оптовой торговли, включая торговлю через агентов, кроме торговли автотранспортными средствами и мотоциклами	645,9	605,6
Организации розничной торговли (кроме торговли автотранспортными средствами и мотоциклами) и по ремонту бытовых изделий и предметов личного пользования	206,1	198,6

Число предприятий торговли каждого вида может быть изображено на графике отдельным цветовым столбцом (рис. 1.3.5).

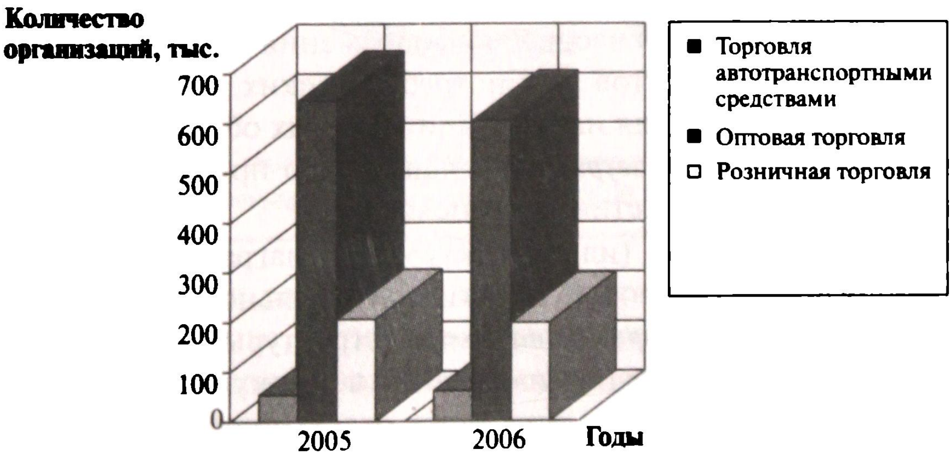


Рис. 1.3.5. Число предприятий торговли в Российской Федерации в 2005–2006 гг.

При применении сравнительных столбиковых диаграмм не рекомендуется изображать на графиках, особенно иллюстрирующих динамику в течение нескольких периодов, большое число категорий изучаемого явления (больше трех). В таких случаях желательно разбить категории на группы и построить несколько графиков или воспользоваться другим типом столбиковых диаграмм, на которых каждая категория изображается как часть общего столбца за определенный период, например рис. 1.3.6.

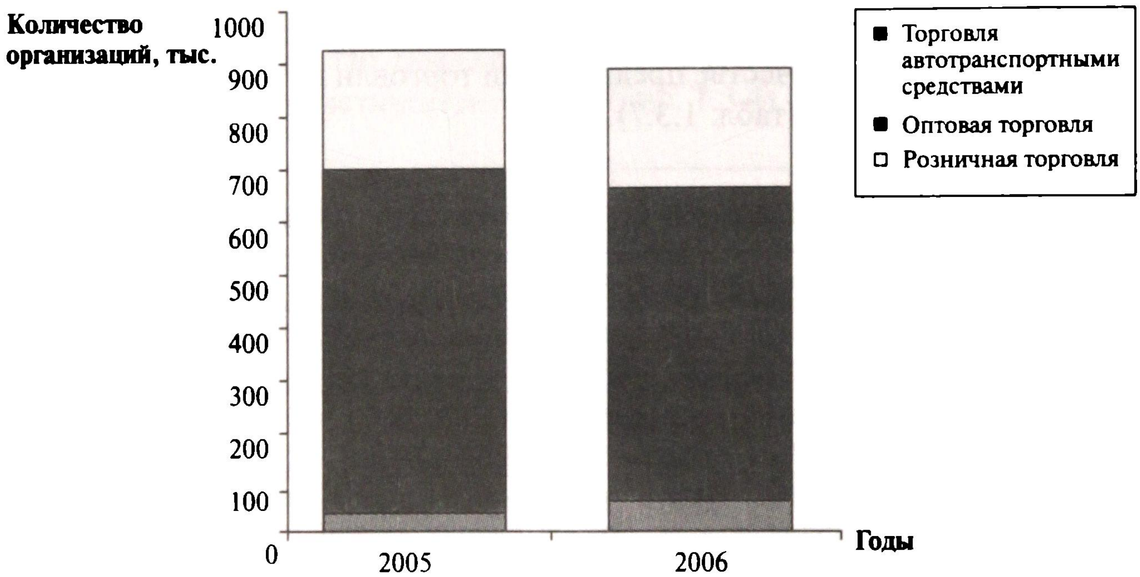


Рис. 1.3.6. Динамика числа предприятий торговли в Российской Федерации в 2005–2006 гг.

Для построения диаграмм сравнения в качестве образа может использоваться изображение объекта, к которому относятся эти данные, например: фигура (силуэт) часов для изображения данных об их продажах, изображения продуктов или потребительских товаров, схематичные изображения монет для иллюстрации данных об их эмиссии. Такие диаграммы называются ***фигурными*** и чаще всего применяются для визуализации данных в маркетинговых целях.

Диаграммы структуры (или структурные диаграммы) изображают строение статистической совокупности или отдельной ее части. Наиболее распространены ***секторные диаграммы*** структуры, которые строятся следующим образом: вся величина явления в совокупности принимается за 100%, затем рассчитываются доли отдельных составных частей. Круг разбивается на секторы пропорционально частям изображаемого целого. Таким образом, на 1% приходится 3,6 градуса.

ПРИМЕР.

Построим секторную диаграмму структуры по данным об объеме накопленных иностранных инвестиций в экономике России по основным странам-инвесторам в 2006 г. (млн долл. США) (табл. 1.3.8).

Таблица 1.3.8

Структура объема накопленных иностранных инвестиций в экономике России по основным странам-инвесторам в 2006 г. (млн долл. США)

Страна-инвестор	Накоплено на конец 2006 г.
Всего инвестиций	**142 926**
Кипр	32 276
Нидерланды	23 451
Люксембург	22 870
Германия	12 260
Великобритания	11 801
США	7698
Остальные страны	32 570

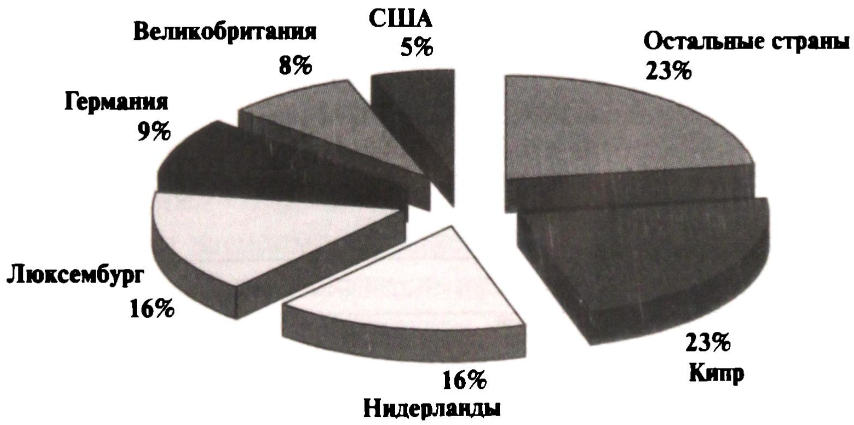


Рис. 1.3.7. Структура объема накопленных иностранных инвестиций в экономике России по основным странам-инвесторам в 2006 г. (млн долл. США)

Для изображения развития явления во времени используются **диаграммы динамики**. Чаще других в этих целях используются **линейные** диаграммы, которые воспроизводят процесс развития данных в виде непрерывной ломаной линии.

Для построения линейных диаграмм по оси X откладывается время (годы, месяцы, даты и т. д.), а по оси Y — масштабы для данных. Если масштаб шкалы на оси X очень растянут по сравнению с масштабом оси Y, то колебания в динамике явлений мало выделяются, и наоборот, преувеличение масштаба по оси Y дает на графике резкие колебания показателя.

ПРИМЕР.

Построим линейную диаграмму по данным о числе строительных организаций в Российской Федерации за период 1995–2005 гг. (табл. 1.3.9).

Таблица 1.3.9

Число действующих строительных организаций в Российской Федерации

	1995 г.	2000 г.	2001 г.	2002 г.	2003 г.	2004 г.	2005 г.
Всего строительных организаций	127 764	129 340	118 374	112 971	113 578	114 464	112 640
из них: смешанной формы собственности	12 683	7787	6936	5309	3897	2742	2004
частной формы собственности	108 639	115 331	106 191	102 471	104 954	107 357	106 628

Для изображения динамики предприятий и организаций масштабная шкала значений разрывается по оси Y недалеко от нулевой линии и на диаграмму попадает лишь информативная часть поля графика. Это позволит избежать изображения пустот, за счет чего динамика изменения показателя отобразится диаграммой более четко (рис. 1.3.8).

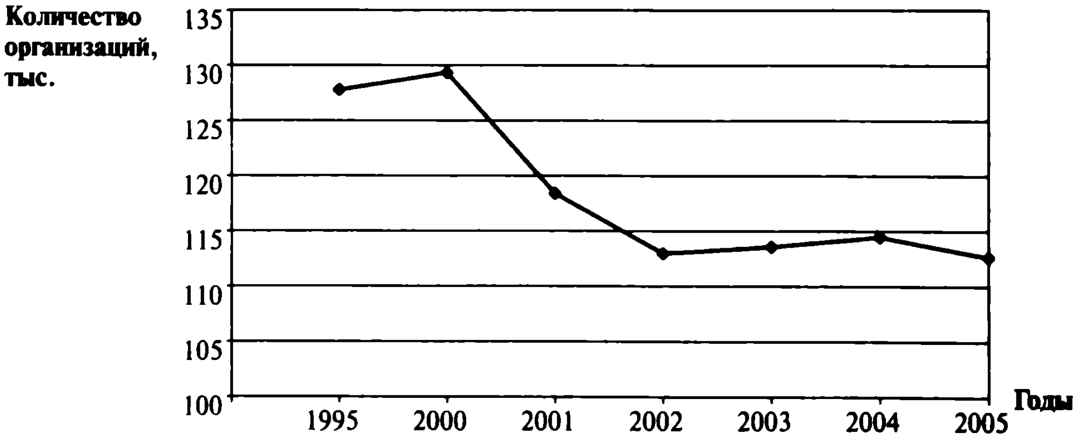


Рис. 1.3.8. Динамика общего числа действующих строительных организаций в Российской Федерации за 1995–2005 гг.

На одном графике могут изображаться несколько кривых для сравнения динамики различных показателей, имеющих одинаковые единицы измерения, или одного, но относящегося к разным объектам, например рис. 1.3.9.

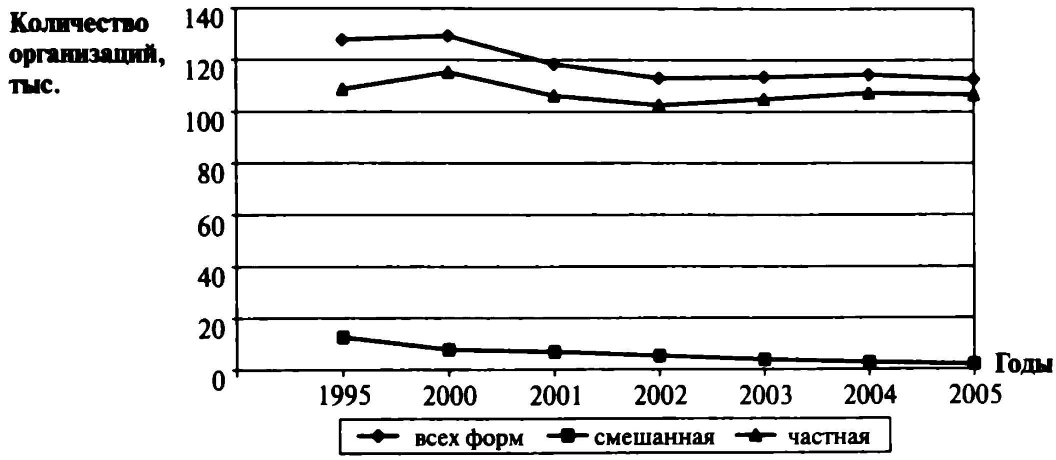


Рис. 1.3.9. Динамика числа действующих строительных организаций в Российской Федерации за 1995–2005 гг. по формам собственности

Для изображения сезонных процессов часто используют ***радиальные диаграммы динамики***. Год изображается в виде круга, радиус которого равен среднемесячному показателю. Круг делится радиусами на 12 секторов. Каждый радиус изображает месяц. На радиусах в соответствии с масштабом обозначаются размеры показателя, и отметки месяцев соединяются отрезками.

Для отображения зависимости одного показателя от другого строится **диаграмма взаимосвязи**. Один показатель принимается за X, а другой за Y (т. е. функцию от X). Строится прямоугольная система координат с масштабами для X и Y и на нее наносятся точки. При установленной взаимосвязи между X и Y точки на графике могут соединяться ломаной линией. Если взаимосвязь не установлена, соединение точек не производится.

ПРИМЕР.

Построим диаграмму взаимосвязи между временем, которое студент тратит на самоподготовку по предмету, и итоговым баллом по предмету по результатам текущих контрольных мероприятий в семестре (табл. 1.3.10). На рис. 1.3.10 показана взаимосвязь между количеством часов в неделю, отводимых студентом на подготовку по предмету, и итоговой суммой баллов по этому предмету. По графику видно, что оценка тем выше, чем больше времени уделяет студент самоподготовке.

Таблица 1.3.10

Зависимость величины итоговой суммы баллов по предмету от времени, выделенного студентом на самоподготовку

Номер студента	Время на самоподготовку, час. в неделю	Итоговая сумма баллов по предмету
1	4,0	93
2	3,5	75
3	2,5	55
4	5,0	89
5	2,9	70
6	3,5	83
7	3,0	73
8	2,0	50
9	1,8	43
10	3,7	87
11	3,1	60
12	2,0	45
13	1,5	42
14	4,5	95
15	1,0	40

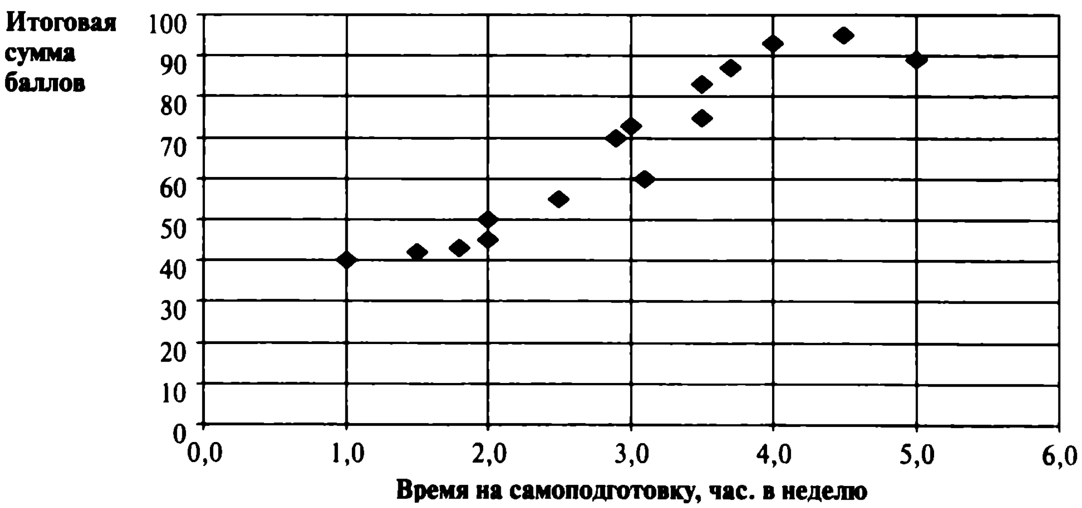


Рис. 1.3.10. Зависимость величины итоговой суммы баллов по предмету от времени, затраченного студентом на самоподготовку

Рис. 1.3.11. Картограмма величины денежного дохода в расчете на душу населения по регионам РФ (в марте 2007 г.)

* — рассчитаны по субъекту, включая АО; *1* — Белгородская область; *2* — Владимирская область; *3* — Воронежская область; *4* — Ивановская область; *5* — Калужская область; *6* — Костромская область; *7* — Курская область; *8* — Липецкая область; *9* — Московская область; *10* — Орловская область; *11* — Рязанская область; *12* — Тамбовская область; *13* — Тульская область; *14* — Ярославская область; *15* — Вологодская область; *16* — Новгородская область; *17* — Республика Адыгея; *18* — Республика Ингушетия; *19* — Кабардино-Балкарская Республика; *20* — Карачаево-Черкесская Республика; *21* — Республика Северная Осетия — Алания; *22* — Чеченская Республика; *23* — Краснодарский край; *24* — Ставропольский край; *25* — Республика Башкортостан; *26* — Республика Марий Эл; *27* — Республика Мордовия; *28* — Республика Татарстан; *29* — Удмуртская Республика; *30* — Чувашская Республика; *31* — Кировская область; *32* — Нижегородская область; *33* — Пензенская область; *34* — Пермский край; *35* — Самарская область; *36* — Саратовская область; *37* — Ульяновская область

Статистические карты представляют собой вид изображений данных на схематичной географической карте, характеризующей уровень или степень распространения изучаемого явления на определенной территории. Средствами изображения на карте являются штриховка, фоновая раскраска или геометрические фигуры. Различают картограммы и картодиаграммы.

Картограмма — это схематическая географическая карта, на которой средствами изображения показывается сравнительная интенсивность изучаемого показателя в пределах каждой единицы нанесенного на карту территориального деления.

Например, на сайте Федеральной службы государственной статистики РФ (http://www.gks.ru/gis/) в виде картограмм изображены сравнительные данные о социально-экономическом положении РФ (рис. 1.3.11). Области и регионы РФ закрашены в соответствии с принятой цветовой шкалой. При переводе курсора мыши с одного региона на другой во всплывающей строке отражается размер доходов на душу населения (например, в Москве по состоянию на март 2007 г. – 29 829,4 руб.).

Картодиаграммы представляют собой график, где в качестве изобразительных знаков используются диаграммные фигуры (столбики, квадраты, круги, фигуры, полосы), которые размещаются на контуре географической карты.

КОНТРОЛЬНЫЕ ВОПРОСЫ

1. Что такое статистические таблицы? Что такое макет таблицы и для чего он используется?
2. В какой части таблицы указывается объект исследования и система показателей для его изучения?
3. Назовите виды таблиц по характеру разработки подлежащего и сказуемого.
4. В какой части графика располагается его название и что в нем должно быть отражено?
5. Из каких элементов состоит статистический график?
6. С помощью каких графических образов можно изобразить производство текстильной продукции (в тыс. руб.) предприятиями Московской области в период с 2000 по 2005 гг.?

Глава 1.4

МЕТОДЫ СБОРА И СИСТЕМАТИЗАЦИИ СТАТИСТИЧЕСКОЙ ИНФОРМАЦИИ. СВОДКА И ГРУППИРОВКА ДАННЫХ

- Статистическая сводка и группировка данных
- Принципы построения статистических группировок
- Вторичная группировка
- Ряды распределения

1.4.1. Статистическая сводка и группировка данных

Собранные данные необходимо систематизировать и подготовить к последующей обработке. Для упорядочивания исходных данных статистика применяет два основных метода — метод сводки и метод группировки.

Сводка — это научная обработка первичных данных с целью получения обобщенных характеристик изучаемого социально-экономического явления по ряду существенных для него признаков (рис. 1.4.1).

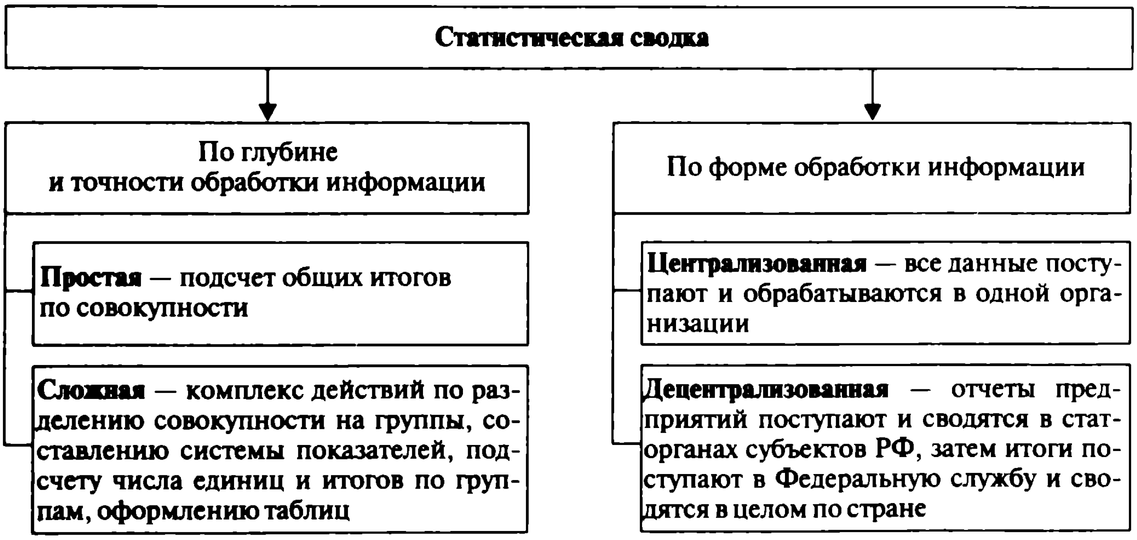


Рис. 1.4.1. Виды статистической сводки

Группировкой называется разбиение общей совокупности единиц объекта наблюдения по одному или нескольким существенным признакам на более однородные группы. Группы могут различаться между собой по числу объектов и в качественном отношении.

В соответствии с познавательными задачами различают три основных вида группировок (рис. 1.4.2).

Рис. 1.4.2. Виды группировок

Типологическая группировка — это разбиение разнородной совокупности единиц наблюдения на отдельные качественно однородные группы и выявление на их основе устойчивых социально-экономических типов явлений. Например, группировка предприятий и организаций по формам собственности, группировка торговых предприятий по принадлежности выпускаемой ими продукции к сферам промышленных и продовольственных товаров.

Структурная группировка предназначена для изучения состава совокупности по какому-либо признаку. Например, изучение возрастной структуры безработных в Российской Федерации (табл. 1.4.1).

Таблица 1.4.1

Структура безработных Росийской Федерации по возрасту (по состоянию на 1 декабря 2005 г.)

	Всего	Мужчины	Женщины
Безработные – всего	100,0	100,0	100,0
в том числе в возрасте, лет:			
до 20	10,5	10,3	10,7
20–29	30,9	32,2	29,4
30–39	20,8	20,5	21,2
40–49	22,8	22,7	22,9
50–59	12,6	12,2	13,0
60–72	2,4	2,1	2,8

Аналитическая, или **факторная группировка** выявляет взаимосвязи между изучаемыми признаками. В статистике выделяют факторные и результативные признаки. **Факторными** называются признаки, под воздействием которых изменяются другие **результативные** признаки. Взаимосвязь проявляется в том, что с возрастанием или убыванием значений факторного признака систематически возрастают или убывают значения признака результативного и наоборот.

При построении аналитической группировки, как правило, единицы совокупности группируются по факторному признаку и каждая выделенная группа характеризуется средними значениями результативного признака. Пример аналитической группировки представлен в табл. 1.4.2.

Таблица 1.4.2

Группировка зависимости суммы кредитов, выданных коммерческими банками, от размера процентной ставки

№ п/п	Группы банков по величине процентной ставки, %	Число банков	Сумма выданных кредитов, млн руб.	
			всего	в среднем на один банк
1	2	3	4	5
1	11–15	7	168,1	24,0
2	15–19	13	200,5	15,4
3	19–23	7	54,4	7,8
4	23–27	3	6,8	2,3
	Итого	30	429,8	14,3

Данные табл. 1.4.2 показывают, что с ростом процентной ставки, под которую выдается кредит, средняя сумма кредита, выдаваемая одним банком, уменьшается. Это говорит о том, что между исследуемыми признаками существует обратная связь.

По способу построения группировки бывают простые и комбинационные.

Простой называется группировка, в которой группы образованы только по одному признаку.

Комбинационной называется группировка, в которой образование групп производится по двум и более признакам, взятым в сочетании (комбинации). Рекомендуется сначала группировать единицы по качественным признакам, а затем — по количественным. Например, в группировке водителей автопарка по уровню квалификации (классу) и производительности труда вначале все водители делятся на две группы по классу, а затем внутри каждого класса производится деление по проценту выполнения плана (табл. 1.4.3).

Из табл. 1.4.3 можно заключить, что квалификация водителя оказывает большее влияние на размер заработной платы, чем производительность труда (% выполнения плана): средняя заработная плата водителей I класса, выполняющих план на 110% и больше, выше, чем у водителей II класса, работающих с той же производительностью.

Таблица 1.4.3

Группировка водителей автопарка по уровню квалификации и производительности труда

Квалификация водителя	Выполнение плана, %	Количество водителей, чел.	Средний размер заработной платы, тыс. руб.
I класс	100–110 110 и более	5 8	16,8 18,3
Итого по группе	–	13	–
II класс	100–110 110 и более	7 10	15,7 17,5
Итого по группе	–	17	–
ВСЕГО	–	30	–

1.4.2. Принципы построения статистических группировок

Построение статистических группировок включает следующие этапы.

1. Определение цели группировки и группировочного признака.
2. Определение числа групп.
3. Обозначение границ групп.
4. Выбор признаков, которые будут характеризовать выделенные группы.

Группировочным (или *основанием* группировки) называется количественный или качественный признак, по которому проводится разбиение единиц совокупности на группы.

Число выделяемых групп зависит от вида группировочного признака, степени его вариации и объема изучаемой совокупности.

Если группировочный признак качественный, то число групп будет равно числу вариантов этого признака. Например, группируя сотрудников по полу, мы получим 2 группы, которые будут обозначаться категориями «сотрудники-мужчины» и «сотрудники-женщины».

Если группировочный признак количественный, то при определении числа групп необходимо учесть **размах вариации** группировочного признака (R), который определяется по формуле:

$$R = x_{max} - x_{min}, \tag{1.4.1}$$

где x_{max} — максимальное значение группировочного признака; x_{min} — минимальное значение группировочного признака.

Чем больше размах вариации группировочного признака, тем большее число групп может быть образовано. Однако при слишком большом их числе возникает проблема «пустых» групп, т. е. не содержащих ни одного объекта.

Число групп можно определить математически или экспертным путем. Математический способ предполагает использование **формулы Стерджесса**, которая позволяет рассчитать число групп, на которое целесообразно разбить единицы изучаемой совокупности, в зависимости от объема данной совокупности:

$$m = 1 + 3{,}322 \cdot \lg n, \tag{1.4.2}$$

где m — число выделяемых групп; n — общее число единиц совокупности.

Формула (1.4.2) дает хорошие результаты при большом объеме совокупности (больше 30 единиц).

Обозначение границ групп также сильно зависит от выбранных целей группировки и группировочного признака. При группировке по количественному признаку группы могут характеризоваться конкретным вариантом группировочного признака, например группировка семей по количеству детей (семьи с одним ребенком, с двумя, тремя, четырьмя детьми), или определенным диапазоном, или интервалом значений группировочного признака.

Интервал — это значения признака, лежащие в определенных границах. **Нижней границей** называется наименьшее, а **верхней** — наибольшее значение признака в интервале. **Ширина** интервала — это разность между его верхней и нижней границами.

По ширине интервалы группировки бывают **равные** (одинаковые) (табл. 1.4.4) и **неравные** (табл. 1.4.5).

Таблица 1.4.4

Группы коммерческих банков по величине балансовой прибыли

№ группы	Балансовая прибыль, млн руб.	Число банков
1	200–400	40
2	400–600	40

4*

Окончание табл. 1.4.4

№ группы	Балансовая прибыль, млн руб.	Число банков
3	600–800	20
Итого	–	100

Ширина каждого интервала составляет 200 млн руб.:

$$400 - 200 = 600 - 400 = 800 - 600 = 200 \text{ млн руб.}$$

Таблица 1.4.5

Группы предприятий по объему инвестиций

№ группы	Объем инвестиций, млн руб.	Число фирм
1	10–20	5
2	20–40	10
3	40–80	12
Итого	–	27

Ширина интервалов неодинакова:

$$20 - 10 = 10 \text{ млн руб.}$$

$$40 - 20 = 20 \text{ млн руб.}$$

$$80 - 40 = 40 \text{ млн руб.}$$

Ширина равного интервала h определяется по следующей формуле:

$$h = \frac{R}{m} = \frac{x_{max} - x_{min}}{m}, \tag{1.4.3}$$

где x_{max}, x_{min} — максимальное и минимальное значения признака в совокупности; m — число выделяемых в совокупности групп.

Существуют следующие *правила округления ширины интервала h*:

- Если h имеет один знак до запятой (например, 0,67; 1,487; 3,82), полученные значения округляют до десятых (0,7; 1,5; 3,8).
- Если h имеет две значащие цифры до запятой (например, 14,876), это значение округляют до целого числа (15).
- В случае когда h является трех-, четырех- или более значимым числом, его величину следует округлить до ближайшего числа, кратного 100 или 50. Например, 652 следует округлить до 650 или до 700.

Интервалы группировки могут быть **закрытыми** (табл. 1.4.6) **и открытыми** (табл. 1.4.7).

Таблица 1.4.6

Группировка страховых компаний по величине прибыли

Прибыль предприятия, млн руб.	Число предприятий
20–40	5
40–60	15
60–80	20
80–100	10

Таблица 1.4.7

Группировка страховых компаний по величине прибыли

Прибыль предприятия, млн руб.	Число предприятий
до 40	5
40–60	15
60–80	20
80 и выше	10

Если максимальные или минимальные значения сильно отличаются от других значений группировочного признака, то для определения ширины интервала используют значения, несколько превышающие минимум, и несколько меньше, чем максимум. Полученную по формуле (1.4.3) величину округляют и используют в качестве ширины интервала, а первый и/или последний интервалы новой группировки открывают по верхней или нижней границе. Это делается для того, чтобы учесть в открытых интервалах единицы, имеющие аномально большие или малые значения группировочного признака.

Если значение признака у какой-то единицы совпадает с верхней границей интервала (например, в табл. 1.4.6 40 млн руб. — это верхняя граница первого интервала (20–40) и нижняя граница второго (40–60)), такая единица, как правило, относится к следующему интервалу (в данном случае ко второму интервалу 40–60).

Если в основании группировки лежит дискретный признак, то нижняя граница каждого интервала, начиная со второго, берется равной верхней границе предыдущего интервала, увеличенной на 1. Например, группируя страховые компании по числу занятого персонала, можно выделить следующие группы (чел.): 100–150, 151–200, 201–300, 301–400.

ПРИМЕР.

Известны следующие данные по основным показателям деятельности крупнейших банков одной из областей Российской Федерации (табл. 1.4.8, данные условные). Произведем группировку банков по величине собственного капитала, выделив четыре группы с равными интервалами.

Таблица 1.4.8

Показатели деятельности коммерческих банков одной из областей РФ, млн руб.

№ п/п	Сумма активов	Собственный капитал	Балансовая прибыль	Ссудная задолженность
1	645,6	12,0	8,1	30,8
2	636,9	70,4	9,5	25,7
3	629,0	41,0	38,4	26,4
4	619,6	120,8	38,4	25,3
5	616,4	49,4	13,4	20,9
6	614,4	50,3	30,1	47,3
7	608,6	70,0	37,8	43,7
8	601,1	52,4	41,1	29,1
9	600,2	42,0	9,3	56,1
10	600,0	27,3	39,3	24,9
11	592,9	72,0	8,6	39,6
12	591,7	22,4	40,5	59,6
13	585,5	39,3	45,3	44,9
14	578,6	70,0	8,4	32,2
15	577,5	22,9	12,8	45,1
16	553,7	119,3	44,7	24,5
17	543,6	49,6	8,8	31,2
18	542,0	88,6	32,2	37,1
19	517,0	43,7	20,3	23,1
20	516,7	90,5	12,2	15,8

Вначале для удобства расположим банки по убыванию величины группировочного признака – собственного капитала (табл. 1.4.9).

Определим ширину интервала:

$$h = \frac{x_{max} - x_{min}}{m} = \frac{120,8 - 12,0}{4} = 27,2.$$

Таблица 1.4.9

Показатели деятельности коммерческих банков одной из областей РФ, млн руб.

№	Собственный капитал	Сумма активов	Балансовая прибыль	Ссудная задолженность
4	120,8	619,6	38,4	25,3
16	119,3	553,7	44,7	24,5
20	90,5	516,7	12,2	15,8
18	88,6	542,0	32,2	37,1
11	72,0	592,9	8,6	39,6
2	70,4	636,9	9,5	25,7
7	70,0	608,6	37,8	43,7
14	70,0	578,6	8,4	32,2
8	52,4	601,1	41,1	29,1
6	50,3	614,4	30,1	47,3
17	49,6	543,6	8,8	31,2
5	49,4	616,4	13,4	20,9
19	43,7	517,0	20,3	23,1
9	42,0	600,2	9,3	56,1
3	41,0	629,0	38,4	26,4
13	39,3	585,5	45,3	44,9
10	27,3	600,0	39,3	24,9
15	22,9	577,5	12,8	45,1
12	22,4	591,7	40,5	59,6
1	12,0	645,6	8,1	30,8

Обозначим границы групп:

1-я группа – 12,0–39,2; 3-я группа – 66,4–93,6;

2-я группа – 39,2–66,4; 4-я группа – 93,6–120,8.

Проверим наполнение построенных групп. В первую группу вошло 4 банка (№ 1. 12, 15, 10, табл. 1.4.9). Размер собственного капитала этих банков находится в пределах определенных нами границ первой группы – от 12,0 до 39,2 млн руб. Вторая группа содержит 8 банков, третья – 6, четвертая – 2. Поскольку все группы банков по величине собственного капитала заполнены и мы не получили пустых и единичных групп, можно переходить к обобщению результатов группировки и оформлять групповую таблицу. Для характеристики построенных нами групп банков по величине собственного капитала используем все имеющиеся в табл. 1.4.9 показатели. Для этого подсчитаем суммарные размеры показателей по каждой группе (как сумму значений показателя по банкам, попавшим в соответствующую группу) и сведем всю информацию в табл. 1.4.10.

Таблица 1.4.10

Группировка коммерческих банков по величине собственного капитала

Группы банков по величине собственного капитала, млн руб.	Число банков	Сумма активов, млн руб.	Балансовая прибыль, млн руб.	Ссудная задолженность, млн руб.
А	1	2	3	4
12,0–39,2	4	2414,8	100,7	160,4
39,2–66,4	8	4707,2	206,7	279,0
66,4–93,6	6	3475,7	108,7	194,1
93,6–120,8	2	1173,3	83,1	49,8
Итого	20	11 771,0	499,2	683,3

На основе информации, представленной в табл. 1.4.10, можно построить структурную группировку. Структурная группировка коммерческих банков позволит рассчитать доли суммарных активов, балансовой прибыли, ссудной задолженности, которые приходятся на каждую выделенную нами группу банков по величине собственного капитала. Для этого необходимо разделить величину показателя соответствующей группы на суммарный размер рассматриваемого показателя по всем 20 банкам. Например, на банки с величиной собственного капитала от 12,0 до 39,2 млн руб. приходится 20,51% суммарных активов всех 20 банков ($2414{,}8/11\,771{,}0 \cdot 100 = 21{,}51\%$). Структурная группировка банков представлена в таблице 1.4.11.

Таблица 1.4.11

Структурная группировка коммерческих банков по величине собственного капитала (в %)

Группы банков по величине собственного капитала, млн руб.	Число банков	Сумма активов, млн руб.	Балансовая прибыль, млн руб.	Ссудная задолженность, млн руб.
А	1	2	3	4
12,0–39,2	4	20,51	20,17	23,47
39,2–66,4	8	39,99	41,41	40,83
66,4–93,6	6	29,53	21,77	28,41
93,6–120,8	2	9,97	16,65	7,29
Итого	20	100,00	100,00	100,00

Структура показателей, представленных в табл. 1.4.11, свидетельствует, что на банки с величиной собственного капитала от 39,2 до 66,4 млн руб. приходится около 40% суммарных размеров активов, балансовой прибыли и ссудной задолженности рассматриваемых нами 20 банков; самые крупные по величине собственного капитала банки (от 93,6 до 120,8 млн руб.) имеют самую низкую долю – 7,29% общего размера ссудной задолженности всех банков.

Аналитическая группировка (табл. 1.4.12) показывает, что величина собственного капитала банков и размер ссудной задолженности находятся в обратной зависимости.

Таблица 1.4.12

Аналитическая группировка коммерческих банков по величине собственного капитала

Группы банков по величине собственного капитала, млн руб.	Число банков	Сумма активов, млн руб.		Ссудная задолженность, млн руб.	
		всего	в среднем на 1 банк	всего	в среднем на 1 банк
А	1	2	3	4	5
12,0– 39,2	4	2414,8	603,7	160,4	40,1
39,2–66,4	8	4707,2	588,4	279,0	34,9
66,4–93,6	6	3475,7	579,3	194,1	32,4
93,6–120,8	2	1173,3	586,7	49,8	24,9
Итого	20	11 771,0	–	683,3	–
В среднем на 1 банк	–	588,6	–	34,2	–

Вывод о наличии обратной взаимосвязи следует из того, что при росте величины капитала от первой группы к последней средний размер ссудной задолженности на один банк (см. гр. 5, табл. 1.4.12) убывает. Анализ характера изменения среднего размера активов на один банк по каждой группе не позволяет предположить наличие строгой прямой или обратной взаимосвязи между этим показателем и размером собственного капитала банка: значение среднего размера активов на один банк вначале падает, а затем вновь возрастает к последней группе банков, имеющих собственный капитал в пределах от 93,6 до 120,8 млн руб.

1.4.3. Вторичная группировка

Группировки, построенные за один и тот же период времени, но для разных объектов, могут оказаться несопоставимыми из-за различий в числе выделенных групп или в значениях границ интервалов.

Вторичная группировка — операция по образованию новых групп на основе ранее осуществленной группировки. Наиболее распространенным способом является ***изменение*** (укрупнение) ***первоначальных интервалов***.

ПРИМЕР.

Требуется сравнить эффективность работы сотрудников двух филиалов компании по итогам месяца (табл. 1.4.13). Сравнивать производительность по исходным данным невозможно: выделены разные группы по проценту выполнения плана, в филиалах работает разное число сотрудников.

Таблица 1.4.13

Выполнение планового задания за месяц, %	Количество сотрудников
Филиал 1	
90–100	8
100–105	25
105–110	12
110 и более	9
Итого	54
Филиал 2	
90–95	3
95–100	7
100–110	14
110–115	9
115 и более	4
Итого	37

Для сравнения результатов необходимо перегруппировать данные. Данные по каждому филиалу путем укрупнения границ интервалов можно привести к следующим группам по проценту выполнения плана: 90–100, 100–110, 110 и более.

Для этого нам потребуется: в таблице с данными по Филиалу 1 объединить интервалы 100–105 с 25 сотрудниками и 105–110 с 12 сотрудниками и получить новый интервал 100–110, в который попадут уже 37 сотрудников (25 + 12 = 37).

В таблице с данными по Филиалу 2 объединяем первый и второй интервалы и получаем интервал 90–100 с 10 сотрудниками (3 + 7), а также интервалы 110–115 с открытым интервалом 115 и более и получаем интервал 110 и более, в который войдут 13 сотрудников (8 + 4). В результате получим новую табл. 1.4.14.

Таблица 1.4.14

Группировка сотрудников филиалов по величине выполнения планового задания

Выполнение планового задания за месяц, %	Количество сотрудников, чел.	
	Филиал 1	Филиал 2
90–100	8	10
100–110	37	14
110 и более	9	13
Итого	54	37

Чтобы можно было сравнить эффективность работы сотрудников в филиалах, вычислим удельный вес каждой группы сотрудников в общей численности занятых в каждом филиале (табл. 1.4.15).

Таблица 1.4.15

Группировка сотрудников филиалов по величине выполнения планового задания

Выполнение планового задания за месяц, %	Филиал 1		Филиал 2	
	Количество сотрудников, чел.	Удельный вес, %	Количество сотрудников, чел.	Удельный вес, %
90–100	8	14,8	10	27,0
100–110	37	68,5	14	37,8
110 и более	9	16,7	13	35,2
Итого	54	100,0	37	100,0

Теперь по данным табл. 1.4.15 понятно, что сотрудники Филиала 1 работают более продуктивно по сравнению с сотрудниками Филиала 2, так как только 14,8% из них не справляются с выполнением планового задания против 27% в Филиале 2, что составляет почти треть от общей численности его работников.

1.4.4. Ряды распределения

Ряд распределения представляет собой простейшую группировку, в которой каждая выделенная группа характеризуется только количеством входящих в нее единиц совокупности. В рассмотренном нами примере с группировкой банков по величине капитала мы также получили ряд распределения — он записан а графах А и 1 табл. 1.4.11 (с. 56).

Различают атрибутивные и вариационные ряды распределения (рис. 1.4.3).

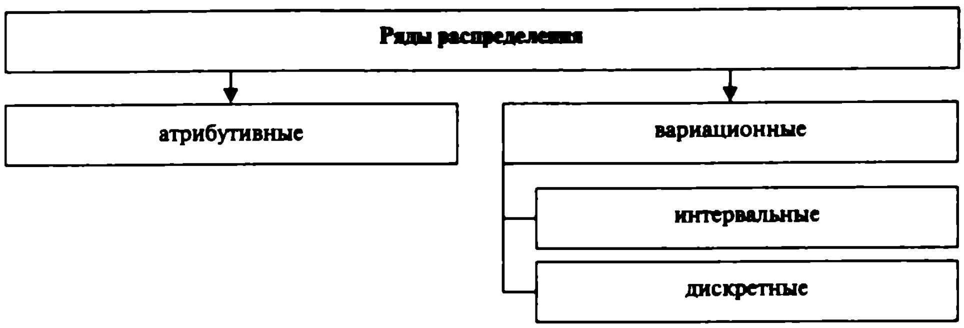


Рис. 1.4.3. Виды рядов распределения

Атрибутивными называют ряды распределения, построенные по качественным признакам. **Вариационными** называют ряды распределения, построенные по количественному признаку. Вариационный ряд состоит из двух элементов: вариантов и частот.

Вариантами называются отдельные значения признака. Варианты признака обозначаются X_i. Частотами называется количество отдельных вариант или каждой группы вариационного ряда. Частоты показывают, как часто встречаются те или иные значения признака в изучаемой совокупности и обозначаются f_i. Сумма всех частот определяет объем совокупности. **Частостями** называются частоты, выраженные в долях единицы или в процентах к итогу и обозначаются w_i. Соответственно сумма частостей равна 1 или 100%.

Дискретный вариационный ряд характеризует распределение единиц совокупности по дискретному признаку, принимающему только целые значения. Например, можно разбить студентов на группы по полученным ими баллам в сессию (табл. 1.4.16).

Таблица 1.4.16

Распределение студентов в соответствии с баллами, полученными в сессию

Балл (оценка), X_i	Количество студентов, f_i	Удельный вес студентов, w_i (%)
5	34	29,82
4	37	32,46
3	33	28,95
2	10	8,77
Итого	114	100,00

Интервальный вариационный ряд распределения — это ряд распределения, в котором группировочный признак задан интервалами значений. Например, распределение консалтинговых фирм по величине прибыли (табл. 1.4.17).

Таблица 1.4.17

Ряд распределения консалтинговых фирм по величине прибыли за 2006 г.

Прибыль, млн руб.	Количество фирм
20–40	5
40–60	15
60–80	20
80–100	10

Анализ рядов распределения наглядно можно проводить на основе их графического изображения. Для этой цели строят *полигон* и *гистограмму*.

Полигон используется для изображения дискретных вариационных рядов. По оси абсцисс (X) в одинаковом масштабе откладываются значения признака, а по оси ординат (Y) — частоты. Полученные на пересечении осей X и Y точки соединяются прямыми линиями, в результате чего получают ломаную линию, называемую полигоном частот (рис. 1.4.4).

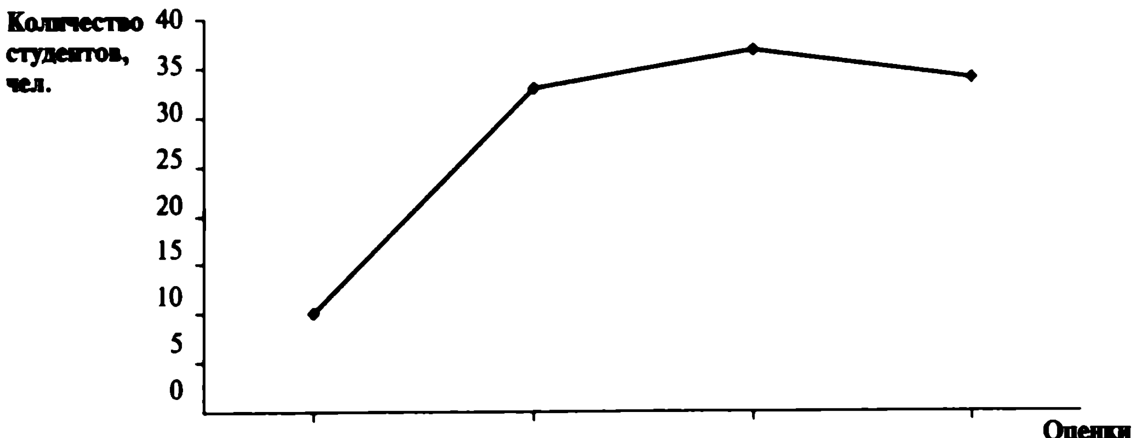


Рис. 1.4.4. Полигон распределения студентов в соответствии с баллами, полученными в сессию (табл. 1.4.16)

Гистограмма применяется для изображения интервального вариационного ряда. При ее построении на оси X откладываются величины интервалов, а на оси Y — частоты, которые изображаются прямоугольниками, построенными на соответствующих интервалах. Гистограмма может быть преобразована в полигон распределения, если середины верхних сторон прямоугольников соединить прямыми линиями (рис. 1.4.5).

При построении гистограммы распределения вариационного ряда с неравными интервалами по оси Y откладывают не частоты, а плотность распределения признака в соответствующих интервалах. **Плотность распределения** — это частота, рассчитанная на единицу ширины интервала.

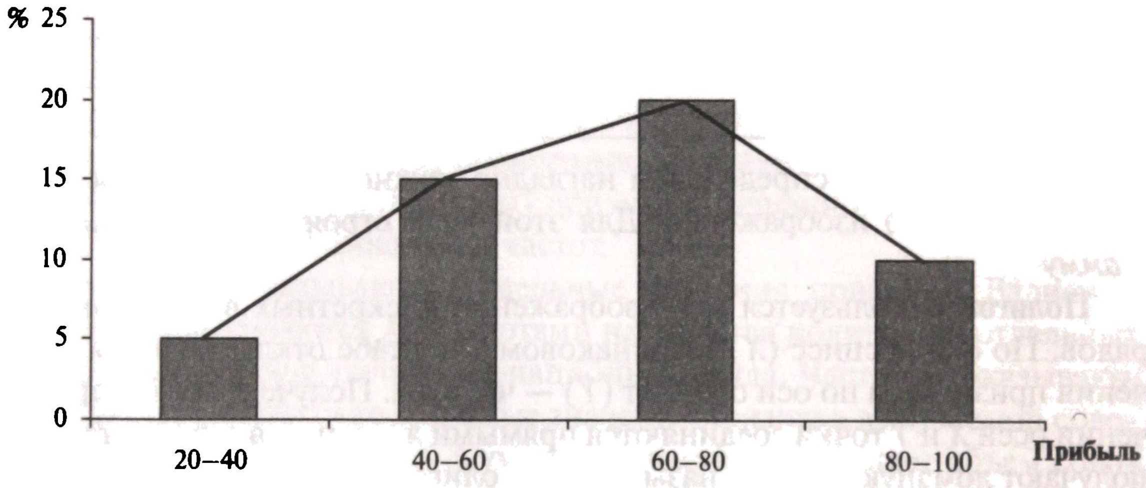


Рис. 1.4.5. Гистограмма и полигон распределения консалтинговых фирм по величине прибыли за 2006 г. (табл. 1.4.17)

КОНТРОЛЬНЫЕ ВОПРОСЫ

1. В чем заключается основная задача метода группировки? Какие виды группировок применяются?
2. Как можно определить число групп при группировке данных?
3. Может ли качественный признак являться основанием группировки?
4. Для каких целей используют аналитические группировки?
5. Что такое вторичная группировка? Каковы цели ее применения?
6. В чем отличие группировки от ряда распределения?
7. Можно ли построить гистограмму по данным о распределении семей по количеству детей?

Глава 1.5

АБСОЛЮТНЫЕ И ОТНОСИТЕЛЬНЫЕ СТАТИСТИЧЕСКИЕ ПОКАЗАТЕЛИ

- Абсолютные показатели
- Относительные показатели

1.5.1. Абсолютные показатели

Исходной формой статистических показателей являются показатели в абсолютном выражении, или абсолютные величины. Статистические показатели в форме **абсолютных величин** характеризуют размеры изучаемых статистикой процессов и явлений, а именно их массу, площадь, объем, протяженность, отражают их временные характеристики, а также могут представлять объем совокупности, т. е. число составляющих ее единиц.

Индивидуальные абсолютные показатели получают непосредственно при сборе данных как результат замера, взвешивания, подсчета и оценки интересующего количественного признака у конкретного объекта: человека, предприятия, домохозяйства. Индивидуальные абсолютные показатели могут рассчитываться как разность. Например: разность между численностью работников предприятия на конец и на начало года, разность между выручкой от реализации предприятия и суммой затрат.

Сводные абсолютные показатели характеризуют общий объем признака или объем совокупности или ее части. Их получают в результате сводки и группировки индивидуальных значений. К таким показателям относятся общая численность занятых в отрасли, совокупные активы коммерческих банков региона и т. п.

Абсолютные статистические показатели всегда имеют единицы измерения. Различают натуральные, стоимостные и трудовые единицы измерения (рис. 1.5.1).

В международной практике используются такие ***натуральные единицы измерения***, как тонны, килограммы, квадратные, кубические и простые метры, мили, километры, галлоны, литры, штуки и т. д. Например, производство электроэнергии в России за I полугодие 2006 г. составило 572 млрд кВт · ч, за этот же период добыто 277 млн т нефти и 383 млрд $м^3$ газа.

Единицы измерения абсолютных показателей

- **Натуральные**
  - единицы измерения массы, объема, расстояния, количества и т. д.
  - сочетание двух единиц измерения (тонно-километр, киловатт-час и др.)
  - условно-натуральные единицы измерения (условное топливо, баррель и т. п.)
- **Стоимостные**
  - рубли, национальные денежные единицы
  - денежные единицы других стран
- **Трудовые**

Рис. 1.5.1. Виды единиц измерения абсолютных показателей

В группу натуральных также входят *условно-натуральные измерители*, используемые в тех случаях, когда какой-либо продукт имеет несколько разновидностей и общий объем можно определить только исходя из общего для всех разновидностей потребительского свойства. Так, различные виды органического топлива переводятся в условное топливо с теплотой сгорания 29,3 МДж/кг (7000 ккал/кг), мыло разных сортов — в условное мыло с 40%-м содержанием жирных кислот, консервы различного объема — в условные консервные банки объемом 353,4 см^3 и т. д. Перевод в условные единицы измерения осуществляется на основе специальных коэффициентов, рассчитываемых как отношение потребительских свойств отдельных разновидностей продукта к эталонному значению. Так, например, 100 т торфа, теплота сгорания которого — 24 МДж/кг, будут эквивалентны 81,9 т условного топлива $\left(100 \cdot \frac{24{,}0}{29{,}3}\right)$, а 100 т нефти при теплоте сгорания 45 МДж/кг будут оцениваться в 153,6 т условного топлива $\left(100 \cdot \frac{45{,}0}{29{,}3}\right)$.

В отдельных случаях для характеристики какого-либо явления или процесса используется произведение двух единиц. Например, грузооборот и пассажирооборот, оцениваемые соответственно в тонно-километрах и пассажиро-километрах.

В условиях рыночной экономики наибольшее применение имеют стоимостные единицы измерения, позволяющие получить денежную оценку социально-экономических явлений и процессов. Так, одним из важнейших стоимостных показателей в системе национальных счетов, характеризующих общий уровень развития экономики страны, является валовой внутренний продукт, который в России за 2006 г. составил 26 781,1 млрд руб. в текущих ценах (в номинальном выражении).

При анализе и сопоставлении стоимостных показателей необходимо иметь в виду, что в условиях высоких или относительно высоких темпов инфляции они становятся несопоставимыми. Так, сравнивать ВВП России за 2006 г. с его величиной, например, за 1995 г. нецелесообразно, так как содержание рубля за этот период существенно изменилось. Для подобных сравнений осуществляют пересчет показателей в сопоставимые цены.

К ***трудовым единицам измерения***, позволяющим учитывать общие затраты труда на предприятии, трудоемкость отдельных операций технологического процесса, относятся человеко-дни и человеко-часы.

Единство единиц измерения позволяет сравнивать различные объекты, определять соотношения между ними, изучать структуру общественных явлений и процессов по различным показателям.

1.5.2. Относительные показатели

Относительный показатель представляет собой результат деления одного абсолютного показателя на другой и выражает соотношение между количественными характеристиками социально-экономических процессов и явлений.

Абсолютный показатель, находящийся в числителе, называется текущим или сравниваемым. Показатель, с которым производится сравнение и который находится в знаменателе, называется основанием, или базой сравнения. Рассчитываемая таким образом относительная величина показывает, во сколько раз сравниваемый абсолютный показатель больше базисного, какую составляет от него долю или сколько единиц первого приходится на 1, 100, 1000 и т. д. единиц второго.

Относительные показатели могут выражаться в:

- коэффициентах (долях единицы);

- процентах (%);
- промилле (‰);
- продецимилле (‱);
- быть именованными числами.

Если база сравнения принимается за 1, то относительный показатель выражается в коэффициентах, если база принимается за 100, 1000 или 10 000, то показатель соответственно выражается в (%), (‰) и (‱).

Именованный относительный показатель получается в результате соотнесения разноименных абсолютных показателей. Его наименование представляет собой сочетание наименований единиц измерения сравниваемых показателей.

Все используемые на практике относительные статистические показатели можно подразделить на следующие виды.

1. **Относительный показатель динамики (ОПД).** ОПД представляет собой отношение уровня исследуемого процесса или явления за данный период времени (по состоянию на данный момент времени) к уровню этого же процесса или явления в прошлом.

ОПД получают путем деления величины показателя, относящейся к одному временному промежутку (моменту), на величину этого же показателя за другой, как правило, прошедший промежуток или момент времени. Рассчитывая ОПД, мы сопоставляем величины или уровни одного и того же показателя, но относящиеся к различным периодам или моментам времени.

$$\text{ОПД} = \frac{\text{Текущий уровень}}{\text{Предшествующий, или базисный уровень}}. \qquad (1.5.1)$$

Рассчитанная таким образом величина показывает, во сколько раз текущий уровень превышает предшествующий (базисный) или какую долю от предшествующего (базисного) уровня он составляет. Данный показатель может быть выражен кратным отношением или переведен в проценты.

Различают относительные показатели динамики с постоянной и переменной базой сравнения. Если сравнение осуществляется с одним и тем же базисным уровнем, например первым годом периода, состоящего из нескольких лет, получают относительные показатели динамики с постоянной базой, или базисные относительные показатели. При расчете относительных показателей динамики с переменной базой сравнение осуществляется с предшествующим уровнем, т. е. знаменатель относительной величины последовательно меняется «по цепочке» от первого до предпоследнего уровня. В результате получаются цепные относительные показатели.

ПРИМЕР.

Рассчитаем ОПД по данным о производстве автомобилей в Российской Федерации за период январь–апрель 2007 г. (табл. 1.5.1).

Таблица 1.5.1

Производство автомобилей в Российской Федерации в январе–апреле 2007 г.

Наименование показателя	январь	февраль	март	апрель
Автомобили легковые, тыс. шт.	85,6	97,4	106,0	105,0

Рассчитаем относительные показатели динамики.

ОПД с переменной базой сравнения (цепные):

а) $ОПД_{февр./янв.} = \frac{97,4}{85,6}100\% = 113,785\%.$

Вывод: в феврале 2007 г. производство автомобилей в Российской Федерации составило 113,79% от уровня производства в январе (иначе: производство автомобилей в феврале 2007 г. выросло приблизительно в 1,14 раза, или на 13,785%, по сравнению с обьемом производства, достигнутым в январе 2007 г.);

б) $ОПД_{март/февр.} = \frac{106,0}{97,4}100\% = 108,83\%;$

в) $ОПД_{апр./март} \frac{105,0}{106,0}100\% = 99,06\%.$

Вывод: в апреле 2007 г. производство автомобилей в Российской Федерации составило 99,06% от уровня производства в марте (сократилось в 0,9906 раза, или на 0,94%, по сравнению с обьемом производства, достигнутым в марте).

ОПД с постоянной базой сравнения (базисные) (за базу сравнения выбран уровень производства автомобилей в январе 2007 г.):

а) $ОПД_{февр./янв.} = \frac{97,4}{85,6}100\% = 113,79\%;$

б) $ОПД_{март/янв.} = \frac{106,0}{85,6}100\% = 123,83\%;$

в) $ОПД_{апр./янв.} = \frac{105,0}{85,6}100\% = 122,66\%.$

ОПД с переменной и постоянной базами сравнения взаимосвязаны между собой следующим образом: последовательное произведение всех ОПД с переменной базой (выраженных в коэффициентах) равно ОПД с постоянной базой, рассчитанному как отношение последнего уровня рассматриваемого периода к первому уровню.

$1{,}13\,785 \cdot 1{,}0883 \cdot 0{,}9906 = 1{,}22\,668$ (небольшое увеличение значения базисного показателя произошло из-за округлений значений цепных показателей динамики).

2. Относительные показатели плана и реализации плана. Все субъекты финансово-хозяйственной деятельности, от небольших индивидуальных частных предприятий и до крупных корпораций, осуществляют как оперативное, так и стратегическое планирование, а также сравнивают реально достигнутые результаты с ранее намеченными. Для этих целей используются относительные показатели плана (**ОПП**) и относительные показатели реализации плана (**ОПРП**):

$$\text{ОПП} = \frac{\text{Уровень, планируемый на } (i+1)\text{-й период}}{\text{Уровень, достигнутый в } i\text{-м периоде}}. \qquad (1.5.2)$$

$$\text{ОПРП} = \frac{\text{Уровень, достигнутый в } (i+1)\text{-м периоде}}{\text{Уровень, планируемый на } (i+1)\text{-й период}}. \qquad (1.5.3)$$

Первый из этих показателей характеризует относительную высоту планового уровня, т. е. во сколько раз намечаемый показатель превысит достигнутый уровень или сколько процентов от этого уровня составит. Второй показатель отражает фактический объем производства или реализации в процентах или коэффициентах по сравнению с плановым уровнем.

ПРИМЕР.

Предположим, объем реализованной продукции одной из коммерческих фирм в 2005 г. составил 1 235 679 руб. В 2006 г., исходя из проведенного анализа складывающихся на рынке тенденций, руководство фирмы сочло необходимым увеличить показатель до 1 300 000 руб. Фактически же в 2006 г. было реализовано продукции на 1 289 601 руб. В этом случае ОПП, представляющий собой отношение планируемой величины к фактически достигнутой, составит $\frac{1\,300\,000}{1\,235\,679}100\% = 105{,}2\%$. Это значит, что на 2006 г. руководство фирмы установило плановое задание — повысить годовой объем реализации продукции на 5,2% (105,2 – 100,0 = 5,2) от уровня 2005 г.

Однако ОПРП, определяемый как отношение фактически достигнутой величины к ранее запланированной, по итогам 2006 г. составил лишь $\frac{1\,289\,601}{1\,300\,000}100\% = 99{,}2\%$ от запланированной величины 1 300 000 руб. То есть плановое задание по реализации продукции недовыполнено на 100–99,2 = 0,8%.

Между относительными показателями плана, реализации плана и динамики существует следующая взаимосвязь:

$$ОПП \cdot ОПРП = ОПД. \qquad (1.5.4)$$

В нашем примере: $1{,}052 \cdot 0{,}992 = 1{,}044$, или $\frac{1\,289\,601}{1\,235\,679} = 1{,}044$.

Основываясь на этой взаимосвязи, по любым двум известным величинам всегда можно определить третью неизвестную величину.

Следовательно, в 2006 г. по плану объем реализации продукции предполагалось увеличить в 1,052 раза, или на 5,2%. Фактически же произошло следующее: план был недовыполнен на 0,8% (100–99,2), но тем не менее по сравнению с 2005 г. объем реализации продукции увеличился в 1,044 раза, или на 4,4%.

3. **Относительный показатель структуры** (ОПС) представляет собой соотношение структурных частей изучаемого объекта и их целого:

$$ОПС = \frac{\text{Показатель, характеризующий часть совокупности}}{\text{Показатель по всей совокупности в целом}}. \qquad (1.5.5)$$

ОПС выражается в долях единицы или в процентах. Рассчитанные величины, соответственно называемые долями или удельными весами, показывают, какой долей обладает или какой удельный вес имеет та или иная часть в общем итоге.

ПРИМЕР.

Рассмотрим структуру произведенного ВВП РФ в 2004 г. в основных ценах (табл. 1.5.2).

Таблица 1.5.2

Структура валового внутреннего продукта Российской Федерации в 2004 г.

	Объем	
	млрд руб.	% к итогу
ВВП – всего в том числе:	14 939	100,0
производство товаров	6133,80	41,1
производство услуг	8805,70	58,9

Рассчитанные в последней графе таблицы проценты представляют собой относительные показатели структуры (в данном случае – удельные веса). Сумма удельных весов, рассчитанных по составляющим одной совокупности, всегда должна быть равна 1 или 100%.

4. Относительный показатель координации (ОПК) представляет собой отношение одной части совокупности к другой части этой же совокупности:

$$\text{ОПК} = \frac{\text{Показатель, характеризующий часть совокупности}}{\begin{array}{c}\text{Показатель, характеризующий}\\ \text{другую часть той же совокупности}\end{array}}. \quad (1.5.6)$$

При этом в качестве базы сравнения выбирается та часть, которая имеет наибольший удельный вес или является приоритетной с экономической, социальной или какой-либо другой точки зрения. В результате узнают, во сколько раз данная часть совокупности больше базисной или сколько процентов от нее составляет, или сколько единиц данной структурной части приходится на 1 единицу (иногда — на 100, 1000 и более единиц) базисной структурной части. Так, на основе данных, приведенной выше в табл. 1.5.2, мы можем вычислить, что на каждый рубль произведенных товаров в 2004 г. приходилось 1,4 руб. произведенных услуг (8805,70 : 6133,80).

5. Относительный показатель интенсивности (ОПИ) характеризует степень распространения изучаемого процесса или явления и представляет собой отношение исследуемого показателя к размеру присущей ему среды:

$$\text{ОПИ} = \frac{\text{Показатель, характеризующий явление } A}{\begin{array}{c}\text{Показатель, характеризующий}\\ \text{среду распространия явления } A\end{array}}. \quad (1.5.7)$$

Такими показателями являются демографические коэффициенты рождаемости, смертности, естественного прироста, брачности и др. Коэффициент рождаемости рассчитывается как отношение числа родившихся за год к среднегодовой численности населения:

$$K_{\text{рожд.}} = \frac{\text{Число родившихся за год}}{\text{Среднегодовая численность населения}} 1000 \text{‱}. \quad (1.5.8)$$

ОПИ получают в результате сопоставления величин двух взаимосвязанных в своем развитии явлений, поэтому наиболее часто он представляет собой именованную величину, но может быть выражен и в процентах, промилле, продецимилле.

Например, для определения уровня обеспеченности населения легковыми автомобилями рассчитывается число автомашин, приходящихся на 100 семей, для определения плотности населения рассчитывается число людей, приходящихся на 1 км^2.

Так, по данным социальной статистики, на конец 2005 г. общая численность безработных в Российской Федерации составляла 5,3 млн чел., а экономически активное население — 73,8 млн чел. Отсюда следует, что уровень безработицы составлял 7,2% = 5,3 : 73,8 · 100.

ПРИМЕР.

Рассчитаем коэффициенты рождаемости и смертности населения за 2005 г. по следующим данным: среднегодовая численность населения Российской Федерации в 2005 г. – 143,5 млн чел., число родившихся – 1457,4 тыс. чел., число умерших за 2005 г. составило 2303,9 тыс. чел.

Определим число родившихся и умерших на каждую 1000 чел. населения, используя относительный показатель интенсивности:

$$K_{\text{рожд.}} = \frac{\text{Число родившихся за год}}{\text{Среднегодовая численность населения}} 1000\,‰_{0} =$$

$$= \frac{1457{,}4}{143500{,}0} 1000\,‰_{0} = 10{,}2\,‰_{0};$$

$$K_{\text{смерт.}} = \frac{\text{Число умерших за год}}{\text{Среднегодовая численность населения}} 1000\,‰_{0} =$$

$$= \frac{2303{,}9}{143\,500{,}0} \cdot 1000\,‰_{0} = 16{,}1\,‰_{0}.$$

На каждую 1000 чел. населения в 2005 г. в Российской Федерации рождалось 10, а умирало 16 человек.

Разновидностью относительных показателей интенсивности являются **относительные показатели уровня экономического развития**, характеризующие производство продукции в расчете на душу населения и играющие важную роль в оценке развития экономики государства или региона. В расчетах используют среднюю численность населения, например, за один год.

Например, производство валового внутреннего продукта (ВВП) в Российской Федерации в 2006 г. в текущих ценах составило 26 781,1 млрд руб. Среднегодовая численность населения в 2006 г. – 142,8 млн чел. Размер ВВП на душу населения составит:

$$\frac{26\,781{,}1}{142{,}8} \cdot 1000 = 187\,543 \text{ руб.}$$

6. Относительный показатель сравнения (ОПС$_р$) представляет собой соотношение одноименных абсолютных показателей, характеризующих разные объекты (предприятия, фирмы, районы, области, страны и т. п.):

$$\text{ОПС}_\text{р} = \frac{\text{Показатель, характеризующий объект } A}{\text{Показатель, характеризующий объект } B}. \qquad (1.5.9)$$

Для выражения данного показателя могут использоваться как коэффициенты, так и проценты.

Например, среднемноголетние запасы воды в Ладожском озере составляют 911 км^3, а в озере Байкал — 23 000 км^3, относительный показатель сравнения в этом случае:

$$\text{ОПС}_\text{р} = \frac{23\,000}{911} = 25{,}2.$$

Следовательно, запас воды в озере Байкал в 25,2 раза больше, чем в Ладожском озере.

КОНТРОЛЬНЫЕ ВОПРОСЫ

1. Какие показатели относятся к абсолютным? Перечислите виды единиц измерения абсолютных статистических показателей.
2. Назовите единицы измерения относительных статистических показателей.
3. Перечислите виды относительных показателей.
4. К какому виду показателей принадлежит отношение величины прибыли предприятия за II квартал 2007 г. к величине прибыли за I квартал 2007 г.; количество рабочих-ремонтников, приходящееся на 10 водителей автобусного парка?
5. На сколько процентов реализован план по выпуску продукции, если планировалось увеличить выпуск на 14%, а по итогам месяца выпуск увеличился на 19,7%?

Раздел 2

ОСНОВЫ АНАЛИЗА ВАРИАЦИИ ПРИЗНАКА В РЯДАХ РАСПРЕДЕЛЕНИЯ

Глава 2.1

СТАТИСТИЧЕСКИЕ ПОКАЗАТЕЛИ В ФОРМЕ СРЕДНИХ ВЕЛИЧИН

- Сущность и виды средних величин
- Средняя арифметическая
- Свойства средней арифметической
- Средняя гармоническая
- Другие виды средних величин
- Структурные средние

2.1.1. Сущность и виды средних величин

При работе с числовой информацией, относящейся к объектам какой-либо совокупности или группы, практически всегда возникает потребность в обобщении, т. е. в выборе или расчете значения признака, которое в той или иной мере присуще каждому объекту, составляющему совокупность или группу, и характеризует весь набор данных одним числом. Часто такая потребность вызвана необходимостью сравнить наборы данных. Например, требуется оценить, какая из бригад скорой помощи более оперативно реагирует на вызовы по следующим наборам данных о времени: от момента получения вызова до прибытия к больному. Информация взята из журнала регистрации вызовов двух бригад за неделю:

Бригада А, в минутах:

10; 19; 15; 20; 13; 40; 11; 18; 22; 34 Общее время: 202 мин.

Бригада Б, в минутах:

31; 12; 20; 9; 42; 14; 16; 27; 11; 13; 24 Общее время: 219 мин.

Если в качестве обобщающего показателя за неделю использовать только общее время, то может показаться, что первая бригада работает оперативнее. Но для объективной оценки нужно соотнести общее время реакции на вызовы с их количеством, т. е. понять, сколько времени в среднем затрачивает каждая бригада на дорогу по одному вызову. Тогда время от момента регистрации вызова до момента начала оказания помощи больному в расчете на один вызов бригады А составит: 202 : 10 = 20,2 мин., бригады Б: 219 : 11 = 19,9 мин., т. е. в среднем на каждый вызов бригада А затрачивала 20,2 мин. движения, а бригада Б реагировала на вызов чуть быстрее — в среднем за 19,9 мин., а значит, оперативность ее работы в течение конкретно этой недели была чуть выше. Вполне вероятно, что в последующие недели картина оперативности может измениться, но если задача исследования возникнет вновь, но уже по данным за месяц, то в любом случае для вывода об оперативности необходимо будет общее время за месяц соотнести с общим количеством вызовов, т. е. рассчитать средние величины.

Средняя величина представляет собой обобщенную количественную характеристику признака в статистической совокупности в конкретных условиях места и времени.

Важнейшее свойство средней величины заключается в том, что она отражает то общее, что присуще всем единицам исследуемой совокупности или группы. Среднюю величину можно трактовать как некоторый средний размер или уровень явления (процесса) в конкретной совокупности, от которого возникают отклонения у различных единиц этой совокупности. Отклонения значений происходят под влиянием множества причин, среди которых могут быть как основные, так и случайные. Например, в случае с бригадами скорой помощи бригада Б, которая в среднем затрачивала 19,9 мин. на один вызов, могла затратить на дорогу по вызову 42 мин. из-за случайной аварии, перекрывшей движение, а за 9 мин. прибыть к больному, мимо дома которого она двигалась в момент регистрации его вызова. Или, например, компания, выставляя акции на торги на фондовой бирже по первоначальной цене, может столкнуться с ситуацией, когда котировки ее акций в отдельные дни и на отдельных биржах в силу определенных сложившихся условий могут продаваться по более высокому или заниженному курсу. Но при соотнесении общего размера признака с количеством наблюдений (т. е. при делении общего времени по всем вызовам на число вызовов, общей стоимости реализованных акций на их количество, т. е. при расчете средней величины признака на один элемент из множества данных), в особенности по большому числу данных, случайные отклонения взаимопогашаются и выявляются изменения, вызванные действием основных факторов. Это позволяет средней отражать типичный уровень признака, независимо от индивидуальных особенностей отдельных единиц.

Однако типичность средней напрямую связана с качественной однородностью используемого набора данных. Например, средняя из величин заработной платы трех сотрудников, имеющих оклады 60 000 руб., 25 000 руб., 5000 руб., будет 30 000 руб. Однако типичной ни для одного из сотрудников она не является. Поэтому определение средней величины часто производится в сочетании с применением метода группировки данных — общая средняя по совокупности дополняется расчетом групповых средних. Если исследуемая совокупность неоднородна или для анализа использовано недостаточное число фактов, на основе средней можно сделать ошибочные выводы. Кроме того, различные категории единиц совокупности могут изначально качественно отличаться по определяющим свойствам или условиям, в которых они функционируют. Поэтому, изучая массовые явления, часто средние величины рассчитывают применительно к различным категориями или территориальным единицам, например: среднюю заработную плату по профессиям или должностям сотрудников, среднедушевой доход и уровень рождаемости по субъектам РФ или крупным городам. В этом случае определяемые средние в большей степени выявляют типичный размер признака, так как соблюдается принцип однородности данных и общность условий развития изучаемых явлений.

Средняя величина, рассчитанная для однородного набора данных, определяет значение признака, вокруг которого концентрируются все другие наблюдения — числа, характеризующие размер признака у объектов изучаемой совокупности или группы.

Категорию средней можно раскрыть через понятие **определяющего свойства.** Согласно этому понятию средняя, являясь обобщающей характеристикой всей совокупности, ориентируется на определенную величину, связанную со всеми единицами изучаемой совокупности. Эту величину можно представить в виде функции:

$$f(x_1, x_2, \ldots, x_n). \qquad (2.1.1)$$

Если в функции (2.1.1) все величины $x_1, x_2, \ldots, x_n$ заменить их средней величиной $\bar{X}$ (читается как «икс среднее», или «икс с чертой»), то значение этой функции должно остаться прежним:

$$f(x_1, x_2, \ldots, x_n) = f(\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_n). \qquad (2.1.2)$$

На практике среднюю можно определить через **исходное соотношение средней** (ИСС) или ее **логическую формулу**:

$$ИСС = \bar{X} = \frac{\text{Суммарное значение или объем осредняемого признака}}{\text{Число единиц или объем совокупности}}.$$

Для каждого показателя, используемого в экономическом анализе, можно составить только одно истинное исходное соотношение для расчета средней. Например, для расчета средней заработной платы работников предприятия необходимо общий фонд заработной платы разделить на число работников:

$$ИСС = \frac{\text{Фонд заработной платы}}{\text{Списочная численность работников}}.$$

Средний размер одного вклада в банк определяется как:

$$ИСС = \frac{\text{Сумма всех вкладов}}{\text{Число вкладов}}.$$

Определить среднюю оценку удовлетворенности клиентов качеством обслуживания или сервиса, которую клиенты оценивали по 10-балльной шкале, можно рассчитать по следующему исходному соотношению:

$$ИСС = \frac{\text{Общая сумма баллов, данная всеми клиентами}}{\text{Общее количество опрошенных клиентов}}.$$

Средняя величина является наиболее распространенной формой статистических показателей, используемых в экономических исследованиях для явлений и процессов, характеристики которых могут быть выражены числами. Средние величины применяются для оценки достигнутого уровня изучаемого показателя и для сравнения величины этого уровня в различных совокупностях или территориальных единицах, при анализе и планировании производственно-хозяйственной деятельности предприятий, финансовых учреждений и других хозяйствующих единиц. Средние применяются для выявления и характеристики взаимосвязи между явлениями, при расчете нормативов отдельных показателей, моделировании и прогнозировании развития и поведения показателей и явлений при задании различных условий.

2.1.2. Средняя арифметическая

Чаще других на практике используется ***средняя арифметическая***, которая, как и все средние, может быть простой или взвешенной.

Средняя арифметическая простая используется в тех случаях, когда расчет осуществляется по несгруппированным данным.

Найдем среднюю величину по данным о размере выручки за 2005 г. 8 компаний, входящих в десятку лидеров по предоставлению услуг по оценке бизнеса и ценных бумаг на российском рынке (табл. 2.1.1).

Таблица 2.1.1

Выручка компаний, входящих в десятку лидеров в предоставлении услуг по оценке бизнеса и ценных бумаг на российском рынке в 2005 г.

Компания	Выручка за 2005 г., млн руб.
«Нексиа Пачоли Консалтинг»	47,6
«2К Аудит – Деловые консультации»	44,4
«Объединенное предприятие по оценке и экспертизе специальных объектов и инвестиций» («Спецоценка»)	43,0
Международный центр оценки	35,4
Институт оценки собственности и финансовой деятельности	33,2
Институт проблем предпринимательства	31,0
«НЭО Центр»	30,4
ФБК (PKF)	25,9

Средний размер выручки от оказания услуг по оценке за 2005 г. в расчете на одну компанию определяет следующее исходное соотношение:

$$\text{ИСС} = \frac{\text{Общая выручка всех компаний}}{\text{Число компаний}}.$$

Используя условные обозначения, запишем формулу средней арифметической простой:

$$\bar{X} = \frac{X_1 + X_2 + \ldots + X_n}{n} = \frac{\sum X_i}{n}. \qquad (2.1.3)$$

Подставив в формулу данные, рассчитаем среднюю:

$$\bar{x} = \frac{47{,}6 + 44{,}4 + 43{,}0 + 35{,}4 + 33{,}2 + 31{,}0 + 30{,}4 + 25{,}9}{8} = \frac{290{,}9}{8} = 36{,}4 \text{ млн руб.}$$

Средняя арифметическая взвешенная используется, когда все или отдельные значения осредняемого признака повторяются, т. е. для расчета средней величины по сгруппированным данным или рядам распределения.

Рассмотрим следующий условный пример.

Таблица 2.1.2

Сделки по акциям эмитента «ХХХ» за торговую сессию

Сделка	Курс продажи, X_i (руб.)	Количество проданных акций, f_i (шт.)	Удельный вес, w_i (%)
1	420	700	37,8
2	440	200	10,8
3	410	950	51,4

Определим по данному дискретному вариационному ряду средний курс продажи 1 акции, что можно сделать, используя исходное соотношение:

$$\text{ИСС} = \frac{\text{Общая сумма сделок}}{\text{Количество проданных акций}}.$$

Чтобы получить общую сумму сделок, необходимо по каждой сделке курс продажи (X_i) умножить на количество проданных акций (f_i) и полученные произведения сложить:

$$\bar{X} = \frac{420 \cdot 700 + 440 \cdot 200 + 410 \cdot 950}{700 + 200 + 950} = \frac{771\,500}{1850} = 417{,}03 \text{ руб.}$$

Мы работаем с совокупностью, единицей которой является каждая акция. Акции варьируют по величине курса продажи. Весом в данном случае является количество повторов каждого значения курса продаж акции, т. е. общее количество акций, имеющих данный размер курса продажи. Поэтому расчет среднего курса продажи произведен по формуле средней арифметической взвешенной:

$$\bar{x} = \frac{\sum x_i f_i}{\sum f_i}. \qquad (2.1.4)$$

В отдельных случаях веса могут быть представлены не абсолютными величинами, а относительными (в процентах или долях единицы). Если для расчета средней по данным табл. 2.1.2 использовать удельные веса, то, преобразовав формулу (2.1.4), получим:

$$\bar{x} = \sum\left(x_i \frac{f_i}{\sum f_i}\right). \qquad (2.1.5)$$

$$\bar{X} = \frac{420 \cdot 37{,}8 + 440 \cdot 10{,}8 + 410 \cdot 51{,}4}{37{,}8 + 10{,}8 + 51{,}4} = \frac{41\,703}{100} = 417{,}03 \text{ руб.}$$

или, выразив удельный вес в долях единицы:

$$\bar{x} = 420 \cdot 0{,}378 + 440 \cdot 0{,}108 + 410 \cdot 0{,}514 = 417{,}03 \text{ руб.}$$

При расчете средней по **интервальному вариационному ряду** от интервалов переходят к их серединам. При этом ширина открытых интервалов (первого и последнего) условно приравнивается к ширине интервалов, примыкающих к ним (второго и предпоследнего). Рассмотрим следующий пример (табл. 2.1.3).

Таблица 2.1.3

Распределение сотрудников предприятия по возрасту

Возраст (лет)	Число сотрудников (чел.)
до 25	8
25–30	32
30–40	68
40–50	49
50–60	21
60 и более	3
Итого	181

Для определения среднего возраста персонала найдем середины возрастных интервалов 22,5; 27,5; 35,0; 45,0; 55,0; 65,0.

Используя среднюю арифметическую взвешенную, определим средний возраст работников данного предприятия:

$$\bar{x} = \frac{22{,}5 \cdot 8 + 27{,}5 \cdot 32 + 35 \cdot 68 + 45 \cdot 49 + 55 \cdot 21 + 65 \cdot 3}{8 + 32 + 68 + 49 + 21 + 3} = 38{,}6 \text{ года.}$$

На практике наиболее часто встречаемая при расчете средних ошибка заключается в игнорировании весов. Например, сравним успеваемость двух студентов по итогам семестра, изучая их листы оценок (рис. 2.1.1 а, б).

Студент А	
Дисциплина	Итоговая оценка, баллов
Макроэкономика	70
Мировая экономика	80
Финансы и кредит	75
Логика	90
Теория организации	80
Теория вероятностей и математическая статистика	70
Гражданское право	75
Основы права	90

Студент Б	
Дисциплина	Итоговая оценка, баллов
Макроэкономика	90
Мировая экономика	80
Финансы и кредит	75
Логика	70
Теория организации	80
Теория вероятностей и математическая статистика	70
Гражданское право	90
Основы права	70

Рис. 2.1.1. Листы оценок студентов по итогам семестра

Как видно из листов оценок, оба студента показывают примерно одинаковые результаты обучения — их оценки по дисциплинам принимают значения 70, 75, 80 и 90 баллов. Можно ли сказать, что средний балл успеваемости студентов будет равен средней арифметической из этих четырех вариантов баллов? Для ответа на этот вопрос проанализируем распределение оценок каждого студента.

Таблица 2.1.4

Распределение оценок по предметам студента А

Оценка, баллов	Количество предметов	Частость
70	2	0,250
75	2	0,250
80	2	0,250
90	2	0,250
Итого	8	1,000

Для расчета среднего значения по ряду распределения следует использовать взвешенную формулу. Рассчитаем средний балл успеваемости каждого студента.

Таблица 2.1.5

Распределение оценок по предметам студента Б

Оценка, баллов	Количество предметов	Частость
70	3	0,375
75	1	0,125
80	2	0,250
90	2	0,250
Итого	8	1,000

Средний балл успеваемости студента А.

$$\bar{x}_{\text{Студент А}} = \frac{70\cdot 2 + 75\cdot 2 + 80\cdot 2 + 90\cdot 2}{8} = \frac{70 + 75 + 80 + 90}{4} = 78{,}75 \approx 79 \text{ баллов.}$$

Поскольку студент *А* каждую оценку получил одинаковое число раз — по два раза, для расчета балла его успеваемости действительно можно не учитывать количество повторов значений и пользовать формулу средней арифметической простой.

Однако для расчета среднего балла успеваемости студента Б необходимо использовать только взвешенную формулу, поскольку этот студент получил каждую оценку разное количество раз. Средний балл успеваемости студента Б составит:

$$\bar{x}_{\text{Студент Б}} = \frac{70\cdot 3 + 75\cdot 1 + 80\cdot 3 + 90\cdot 1}{8} = \frac{210 + 75 + 240 + 90}{4} = 76{,}875 \approx 77 \text{ баллов.}$$

Данный пример показывает, что использовать невзвешенную формулу можно только тогда, когда точно установлено отсутствие весов или их равенство (хотя отсутствие весов можно также трактовать как наличие у каждого значения X_i из набора данных веса, равного единице $f_i = 1$). Если при расчете среднего балла успеваемости студента Б мы бы воспользовались невзвешенной формулой, то его успеваемость за семестр была бы завышена на 2 балла.

2.1.3. Свойства средней арифметической

Средняя арифметическая обладает математическими свойствами, которые более полно раскрывают ее сущность и в ряде случаев используются для ее расчета. Рассмотрим эти свойства.

1. Произведение средней на сумму частот равно сумме произведений отдельных вариантов на соответствующие им частоты:

$$\bar{x}\sum f_i = \sum x_i f_i. \qquad (2.1.6)$$

Проверим свойство на примере расчета среднего курса продажи акций (табл. 2.1.2). За счет округления среднего курса правая и левая части равенства в данном случае будут несколько отличаться:

$$417{,}03 \cdot 1850 = 420 \cdot 700 + 440 \cdot 200 + 410 \cdot 950.$$

2. Сумма отклонений индивидуальных значений признака от средней арифметической равна нулю:

$$\sum (x_i - \bar{x}) f_i = 0. \qquad (2.1.7)$$

Для нашего примера:

$$(420\text{–}417{,}03) \cdot 700 + (440\text{–}417{,}03) \cdot 200 + (410\text{–}417{,}03) \cdot 950 \approx 0.$$

Математическое доказательство данного свойства:

$$\sum (x_i - \bar{x}) f_i = \sum x_i f_i - \sum \bar{x} f_i = \sum x_i f_i - \bar{x}\sum f_i = 0.$$

3. Сумма квадратов отклонений индивидуальных значений признака от средней арифметической меньше, чем сумма квадратов их отклонений от любой другой произвольной величины *C*.

Рассмотрим это свойство на упрощенном примере по ряду данных $X_i = 3,4,5$. Средняя величина для данного ряда равна 4. Рассчитаем сумму квадратов отклонений от $\bar{X} = 4$ и от величины C, приняв ее за вариант $X_i = 3$.

X_i	$X - \bar{X}_i$	$(X - \bar{X}_i)^2$
3	$3 - 4 = -1$	$(-1)^2 = 1$
4	$4 - 4 = 0$	$0^2 = 0$
5	$5 - 4 = 1$	$1^2 = 1$
Итого	0	2

X_i	$X_i - 3$	$(X_i - 3)^2$
3	$3 - 3 = 0$	$(0)^2 = 0$
4	$4 - 3 = 1$	$1^2 = 1$
5	$5 - 3 = 2$	$2^2 = 4$
Итого	3	5

$$\sum (X_i - \bar{X})^2 = 2 < \sum (X_i - 3)^2 = 5.$$

Каким бы большим или малым мы ни брали число C, сумма квадратов отклонений индивидуальных значений признака от этого числа всегда будет больше суммы квадратов отклонений от средней на величину

$$\sum(\bar{x}-C)^2 f_i \text{ или } (\bar{x}-C)^2\sum f_i. \qquad (2.1.8)$$

4. Если все осредняемые варианты уменьшить или увеличить на постоянное число A, то средняя арифметическая соответственно уменьшится или увеличится на ту же величину A:

$$\frac{\sum(x_i \pm A)f_i}{\sum f_i} = \frac{\sum x_i f_i}{\sum f_i} \pm \frac{\sum A f_i}{\sum f_i} = \bar{x} \pm A. \qquad (2.1.9)$$

Так, если все курсы продажи акций увеличить на 15 руб., то средний курс также увеличится на 15 руб.:

$$\bar{x} = \frac{435\cdot 700 + 455\cdot 200 + 425\cdot 950}{1850} = 417{,}03 + 15 = 432{,}03 \text{ руб.}$$

5. Если все варианты значений признака уменьшить или увеличить в A раз, то средняя также соответственно увеличится или уменьшится в A раз:

$$\frac{\sum \frac{x_i}{A} f_i}{\sum f_i} = \frac{\frac{1}{A}\sum x_i f_i}{\sum f_i} = \frac{1}{A}\bar{x}. \qquad (2.1.10)$$

Предположим, курс продажи в каждом случае вырос в 2 раза. Тогда и средний курс также увеличится вдвое:

$$\bar{x} = \frac{420\cdot 2\cdot 700 + 440\cdot 2\cdot 200 + 410\cdot 2\cdot 950}{1850} = 417{,}03\cdot 2 = 834{,}06 \text{ руб.}$$

6. Если все веса уменьшить или увеличить в A раз, то средняя арифметическая от этого не изменится:

$$\frac{\sum x_i \frac{f_i}{A}}{\sum \frac{f_i}{A}} = \frac{\frac{1}{A}\sum x_i f_i}{\frac{1}{A}\sum f_i} = \bar{x}. \qquad (2.1.11)$$

Так, в нашем примере удобнее было бы рассчитывать среднюю, предварительно поделив все веса на 100:

$$\bar{x} = \frac{420\cdot 7 + 440\cdot 2 + 410\cdot 9{,}5}{7 + 2 + 9{,}5} = \frac{7715}{18{,}5} = 417{,}03 \text{ руб.}$$

6*

2.1.4. Средняя гармоническая

Средняя гармоническая взвешенная используется, когда известен числитель исходного соотношения среднего, но неизвестен его знаменатель. Рассмотрим расчет средней урожайности, являющейся одним из основных показателей эффективности производства в агробизнесе.

Таблица 2.1.6

Валовой сбор и урожайность сельскохозяйственной культуры «Y» по районам области

Район	Валовой сбор, тыс. ц	Урожайность, ц/га
А	360	13
Б	530	9
В	290	15
Г	780	8
Д	200	17

Средняя урожайность любой сельскохозяйственной культуры может быть определена только на основе следующего исходного соотношения:

$$\text{ИСС} = \frac{\text{Общий валовой сбор}}{\text{Общая посевная площадь}}, \text{ ц/га}.$$

Общий валовой сбор мы получим простым суммированием валового сбора по районам, т. е. $360 + 530 + 290 + 780 + 200 = 2160$. Данные же о посевной площади отсутствуют, но их можно получить, разделив валовой сбор каждого района на его урожайность. С учетом этого определим искомое среднее:

$$\bar{x} = \frac{360+530+290+780+200}{\frac{360}{13}+\frac{530}{9}+\frac{290}{15}+\frac{780}{8}+\frac{200}{17}} = \frac{2160}{215{,}2} = 10{,}04 \text{ ц/га}.$$

Таким образом, общая посевная площадь данной культуры в целом по области составляла 215,2 тыс. га, а средняя урожайность — 10,0 ц с одного гектара. В данном случае расчет произведен по формуле средней гармонической взвешенной:

$$\bar{x} = \frac{\sum w_i}{\sum \frac{w_i}{x_i}}, \qquad (2.1.12)$$

где $w_i = x_i f_i$.

2.1.5. Другие виды средних величин

Средняя геометрическая рассчитывается по формулам:

- невзвешенная:

$$\bar{x} = \sqrt[K]{x_1 \cdot x_2 \cdot x_3 \cdot \ldots \cdot x_K} = \sqrt[K]{\prod x_i}; \qquad (2.1.13)$$

- взвешенная:

$$\bar{x} = \sqrt[\Sigma m_i]{x_1^{m_1} \cdot x_2^{m_2} \cdot x_3^{m_3} \cdot \ldots \cdot x_K^{m_K}} = \sqrt[\Sigma m_i]{\prod x_i^{m_i}}. \qquad (2.1.14)$$

Наиболее широкое применение этот вид средней получил в анализе динамики для определения среднего темпа роста, что будет рассмотрено в соответствующей теме.

Средняя квадратическая. В основе вычислений ряда сводных расчетных показателей лежит средняя квадратическая:

- невзвешенная:

$$\bar{x} = \sqrt{\frac{\sum x_i^2}{n}}; \qquad (2.1.15)$$

- взвешенная:

$$\bar{x} = \sqrt{\frac{\sum x_i^2 f_i}{\sum f_i}}. \qquad (2.1.16)$$

Данный вид средней используется при расчете показателей вариации.

2.1.6. Структурные средние

Наиболее часто используемыми в экономической практике структурными средними являются мода и медиана.

Мода (Мо) представляет собой значение признака, повторяющееся с наибольшей частотой. **Медианой** (Ме) называется значение признака, приходящееся на середину ранжированной (упорядоченной) совокупности.

Рассмотрим определение моды и медианы по **несгруппированным данным**. Предположим, что 9 торговых фирм города реализуют товар *А* по следующим оптовым ценам (тыс. руб.).

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

Так как чаще всего встречается цена 4,3 тыс. руб., то эта величина и будет модой. Для определения медианы необходимо упорядочить цены:

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

Центральной в этом ряду, т. е. пятой по счету, является цена 4,4 тыс. руб., следовательно, это медиана. Если ранжированный ряд включает в себя четное число единиц, то медиана определяется как средняя из двух центральных значений.

В отличие от моды, медиана практически выполняет функции среднего для неоднородной совокупности. Медиана используется в тех случаях, когда среднее не позволяет объективно оценить исследуемую совокупность из-за сильного влияния на нее максимальных и минимальных значений. Проиллюстрируем значение медианы следующим примером.

Допустим, нам необходимо дать характеристику среднего дохода группы людей, насчитывающей 100 человек, из которых 99 имеют доходы в интервале от 100 до 300 долл. США в месяц, а месячные доходы последнего составляют 5000 долл. США:

№ п/п	1	2	3	4 ... 50	51 ... 99	100
Доход, долл. США	100	104	104	107 ... 162	164 ... 300	5000

Средняя арифметическая покажет средний доход в 600–700 долл., который не только в несколько раз меньше дохода сотого человека, но и имеет мало общего с доходами остальной части группы. Медиана, равная 163 долл. $\left(\frac{162+164}{2}\right)$, позволит дать объективную характеристику уровня доходов 99% данной группы людей.

Рассмотрим расчет Мо и Ме по **сгруппированным данным** (табл. 2.1.7).

Таблица 2.1.7

Распределение торговых предприятий города по уровню цен на товар *А*

Цена, X_i, руб.	Число предприятий, f_i	Накопленные частоты, S_i
52	12	**12**
53	48	**60** (12 + 48)
54	56	**116** (60 + 56 = 12 + 48 + 56)
55	60	**176** (116 + 60 = 12 + 48 + 56 + 60)
56	14	**190** (176 + 14 = 12 + 48 + 56 + 60 + 14)
Всего	190	–

В дискретных рядах распределения мода — это значение, имеющее наибольшую частоту. Максимальное число предприятий (60) установили цену 55 руб., следовательно, она является модальной. Для определения медианы находят номер медианной единицы ряда:

$$N_{\text{Me}} = \frac{n+1}{2}, \qquad (2.1.17)$$

где n — объем совокупности.

В нашем случае:

$$N_{\text{Me}} = \frac{190+1}{2} = 95{,}5.$$

Полученное дробное значение, всегда имеющее место при четном числе единиц совокупности, указывает, что точная середина ряда находится между 95-м и 96-м предприятиями. Необходимо определить, в какой группе находятся предприятия с этими порядковыми номерами. Это можно сделать по накопленной частоте S_i. Очевидно, что магазинов с этими номерами нет в первой группе, где всего лишь 12 торговых предприятий, нет их и во второй группе 60 (12 + 48). Предприятия 95 и 96 находятся в третьей группе 116 (12 + 48 + 56) и, следовательно, медианой является цена 54 руб.

Определение моды и медианы по **интервальным рядам** требует проведения расчетов на основе следующих формул:

$$\text{Mo} = X_0 + i\frac{(f_{\text{Mo}} - f_{\text{Mo}-1})}{(f_{\text{Mo}} - f_{\text{Mo}-1}) + (f_{\text{Mo}} - f_{\text{Mo}+1})}, \qquad (2.1.18)$$

где X_0 — нижняя граница модального интервала (модальным называется интервал, имеющий наибольшую частоту); i — ширина модального интервала; f_{Mo} — частота модального интервала; $f_{\text{Mo}-1}$ — частота интервала, предшествующего модальному; $f_{\text{Mo}+1}$ — частота интервала, следующего за модальным.

$$\text{Me} = X_0 + i\frac{\frac{1}{2}\sum f_i - S_{\text{Me}-1}}{f_{\text{Me}}}, \qquad (2.1.19)$$

где X_0 — нижняя граница медианного интервала (медианным называется первый интервал, накопленная частота которого превышает половину общей суммы частот); i — ширина медианного интервала; $S_{\text{Me}-1}$ — накопленная частота интервала, предшествующего медианному; f_{Me} — частота медианного интервала.

Рассчитаем моду и медиану по данным табл. 2.1.8.

Таблица 2.1.8

Распределение населения региона по уровню денежного дохода

Среднедушевой денежный доход, руб.	Удельный вес населения, %
2400 и менее	2,4
2400–2500	15,4
2500–2600	20,1
2600 -2700	17,2
2700–2800	12,8
2800–2900	9,2
2900–3000	6,5
3000–3100	4,5
3100–3200	3,2
3200–3300	2,3
свыше 3300	6,4
Всего	100,0

Интервал с границами 2500–2600 в данном распределении будет модальным, так как он имеет наибольшую частоту (20,1%). По формуле (2.1.18), определим моду:

$$\text{Mo} = 2500 + 100 \cdot \frac{20{,}1 - 15{,}4}{(20{,}1 - 15{,}4) + (20{,}1 - 17{,}2)} = 2562 \text{ руб.}$$

Для определения медианного интервала необходимо определять накопленную частоту каждого последующего интервала до тех пор, пока она не превысит $^1/_2$ суммы накопленных частот (в нашем случае 50%):

Интервал	Накопленная частота, %
2400 и менее	2,4
2400–2500	17,8
2500–2600	37,9
2600–2700	55,1

Медианным является интервал с границами 2600–2700. Определим медиану по формуле (2.1.19):

$$\text{Me} = 2600 + 100 \cdot \frac{50{,}0 - 37{,}9}{17{,}2} = 2670 \text{ руб.}$$

Соотношение моды, медианы и среднего арифметического указывает на характер распределения признака в совокупности, позволяет оценить его асимметрию. Если Мо < Ме < $\bar{X}$ — имеет место правосторонняя асимметрия, при $\bar{X}$ < Ме < Мо следует сделать вывод о левосторонней асимметрии ряда.

На основе полученных в последнем примере значений структурных средних можно заключить, что наиболее распространенным, типичным является среднедушевой доход порядка 2560 руб. в месяц. В то же время более половины населения располагает доходом свыше 2670 руб. при среднем уровне 2735 руб. (среднее арифметическое взвешенное). Из соотношения этих показателей следует вывод о правосторонней асимметрии распределения населения по уровню среднедушевых денежных доходов, что позволяет предполагать достаточную емкость рынка дорогих товаров повышенного качества и товаров престижной группы.

КОНТРОЛЬНЫЕ ВОПРОСЫ

1. Что такое средняя величина? Перечислите основные виды средних.
2. В каких случаях применяются взвешенные средние величины? Почему при расчете средней величины важно учитывать веса вариантов показателя?
3. В каких случаях применяется средняя гармоническая величина?
4. Что такое мода и для чего она применяется? Можно ли рассчитать моду по интервальному ряду распределения?
5. Что характеризует медиана? Можно ли вычислить медиану по ряду распределения, не рассчитывая накопленные частоты?

Глава 2.2
ПОКАЗАТЕЛИ ВАРИАЦИИ

- Для чего изучают вариацию
- Показатели вариации и способы их расчета
- Математические свойства показателей вариации
- Правило сложения дисперсий

2.2.1. Для чего изучают вариацию

Одна из причин, по которой существует необходимость в проведении статистического анализа и постоянном сборе и обновлении информации о социальных и экономических явлениях, состоит в том, что данные меняются. Как вы уже знаете, изменения индивидуальных значений признака внутри изучаемой совокупности в статистике называются **вариацией** (от лат. *variatio* — изменение, различие).

Вариацию можно определить как степень различия между отдельными значениями признака или показателя. Почему важно изучать вариацию? В чем ценность информации о том, насколько отличаются значения признака друг от друга?

Ситуация, в которой значения признака меняются, всегда связана с долей риска и неопределенностью в будущем. Систематическое воздействие различных факторов и условий вызывает изменение отдельных вариантов признаков или показателя в целом. В большинстве случаев обнаружить такое воздействие и тем самым снизить риск можно, изучая колебания или индивидуальные различия значений, а не обобщающие величины.

Например, биржевой или финансовый аналитик для прогнозирования игры на бирже или анализа котировок конкретной ценной бумаги, извлечет больше полезной информации, изучая колебания курса акции, чем среднюю величину курса этой акции. Менеджера по персоналу при изучении производительности труда работников компании будут больше интересовать причина и размер отклонения количества обработанных за день каждым сотрудником заказов от среднего количества заказов, установленного нормой или рассчитанного для данной категории работников. А по тому, насколько мало или велико среднее отклонение от нормы в разных бригадах (отделах, сменах), менеджер сможет судить об их однородности по производительности отдельных сотрудников и принять меры для ее выравнивания и повышения в целом по предприятию.

Изменения значений одних признаков оказывают влияние на значения других взаимосвязанных с ними признаков, вызывают усиление или снижение их колеблемости. Так, например, на макроуровне изучение вариации позволяет установить, какие факторы и в какой степени влияют на смертность или рождаемость населения, финансовое положение предприятий, ритмичность работы отраслей промышленности.

Показатели изменчивости индивидуальных значений используют для организации контроля качества на производстве, оценки доходности финансовых вложений, при сравнительном анализе конкурентных возможностей предприятий, планировании затрат, маркетинговых исследованиях, для построения статистических моделей и во многих других случаях.

Каким образом можно количественно измерить, как меняется признак, и сделать вывод о степени различия между отдельными его значениями?

Очевидно, что для оценки различия между значениями необходимо сравнить их друг с другом или с некоей постоянной или объединяющей все эти значения величиной. Согласно определяющему свойству, такой общей для всех вариантов признака величиной является их средняя.

Несмотря на то что средняя является условным центром распределения вариантов признака, сама ее величина не дает представления о том, как отдельные значения группируются вокруг нее, сосредоточены ли они вблизи или значительно удалены от центра (рис. 2.2.1).

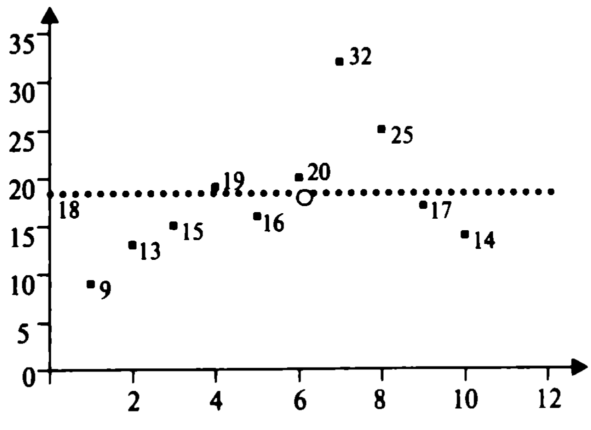


Рис. 2.2.1. Количество договоров на сервисное обслуживание, заключенных менеджерами по продажам в течение месяца

На графике по оси X отмечены № менеджеров, а по оси Y — количество заключенных ими договоров по данным табл. 2.2.1. Очевидно, что среднее количество договоров, заключенных одним менеджером, составляет 18. Средний уровень продаж отмечен на графике пунктирной линией. Степень удаленности от среднего определяется длиной перпендикуляра, опущенного из каждой точки на графике, обозначающей количество заключенных соответствующим менеджером договоров, на пунктирную линию

Чтобы оценить расположение данных, изучают разброс вариантов признака вокруг среднего значения.

Таблица 2.2.1

Итоги продаж за месяц

№ менеджера	Количество договоров
1	9
2	13
3	15
4	19
5	16
6	20
7	32
8	25
9	17
10	14
Итого	180

Чем ближе расположены значения признака к среднему значению, тем, очевидно, больше их близость по отношению друг к другу и тем выше однородность совокупности в целом.

Чтобы определить количественно, как меняется признак, необходимо оценить расстояние между каждым значением признака и общего для них среднего значения, которое определяется как разность между их величинами. Эту разность в статистике называют **отклонением от среднего значения** (рис. 2.2.2).

Вычислив среднее значение абсолютных величин индивидуальных отклонений, мы узнаем величину типичного стандартного расстояния от средней величины.

Таблица 2.2.2

Итоги продаж за месяц

№ менеджера	Количество договоров	Отклонение от 18 (от средней величины)
1	9	–9
2	13	–5
3	15	–3

Окончание табл. 2.2.2

№ менеджера	Количество договоров	Отклонение от 18 (от средней величины)
4	19	1
5	16	–2
6	20	2
7	32	14
8	25	7
9	17	–1
10	14	–4
Итого	180	0

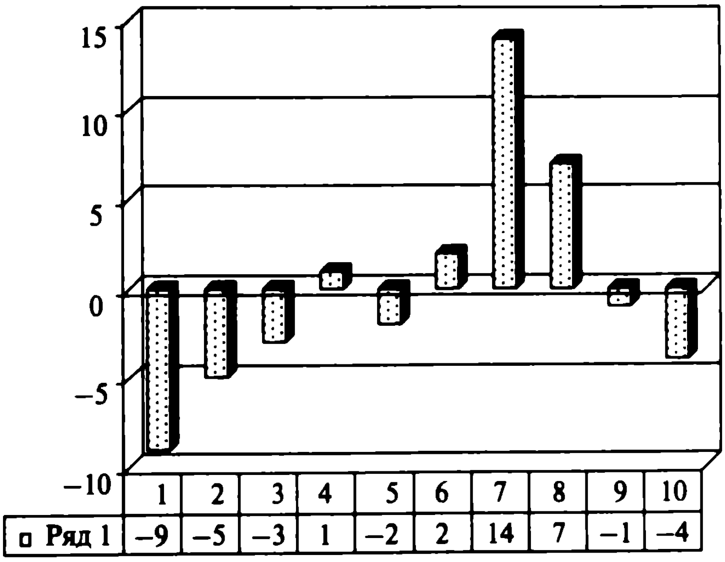


Рис. 2.2.2. Гистограмма отклонения количества заключенных договоров одним менеджером от среднего значения, равного 18

Найденное число покажет, на сколько в среднем отличается каждое значение признака в данной совокупности от средней величины, и будет служить мерой случайности отклонений отдельных значений.

2.2.2. Показатели вариации и способы их расчета

Для количественного измерения степени близости значений отдельных единиц к средней используется система абсолютных и относительных показателей. В общем виде система показателей вариации представлена на рис. 2.2.3.

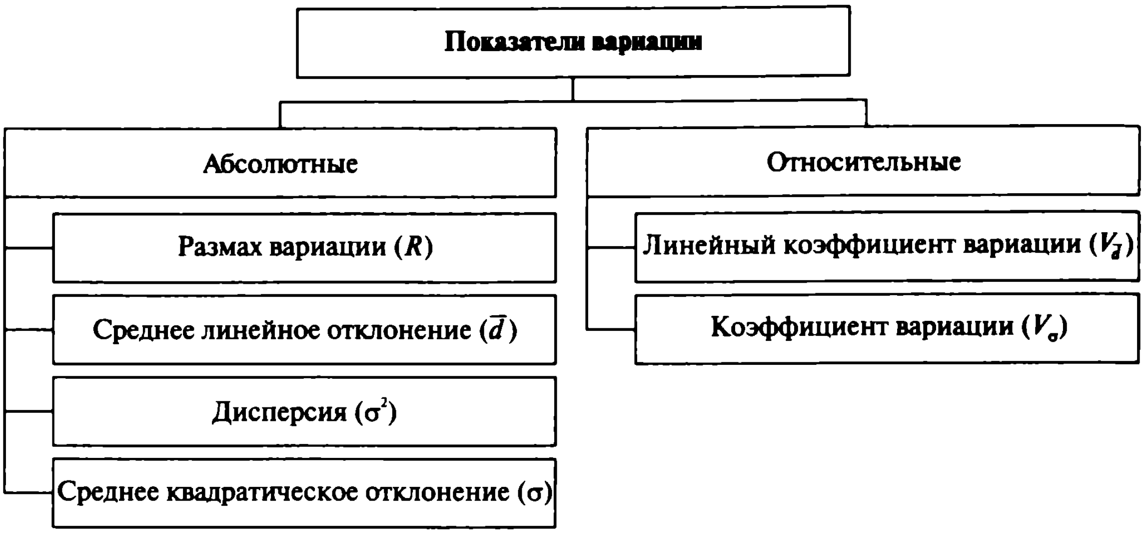


Рис. 2.2.3. Основные показатели вариации

Размах вариации — это разность между наибольшим (X_{max}) и наименьшим (X_{min}) значениями вариантов признака.

$$R = X_{max} - X_{min}. \tag{2.2.1}$$

Размах измеряется в тех же абсолютных единицах, что и значения признака. Размах показывает ширину интервала, где находятся значения данных, но не отражает величины их отклонений. Поскольку размах зависит только от крайних значений, его величина в большей мере подвержена воздействию случайности, так как в совокупности аномально большие или маленькие значения данных могут быть получены под влиянием случайных причин.

ПРИМЕР.

Определим показатель размаха вариации по данным табл. 2.2.3.

Таблица 2.2.3

Группы предприятий по объему товарооборота

Товарооборот X_i, млн руб.	Число предприятий f_i
90–100	28
100–110	48
110–120	20
120–130	4
Итого	100

Величина размаха показывает, что товарооборот предприятий изменяется (колеблется) в пределах от 90 до 130 млн руб., а общая ширина диапазона принимаемых значений (амплитуда колебаний) составляет 40 млн руб.

$$R = 130 - 90 = 40 \text{ млн руб.}$$

Размах вариации хорошо применять в случаях, когда минимальный или максимальный вариант признака имеет особое значение. Например, при определении пределов, в которых могут колебаться размеры отдельных параметров деталей, оценки пределов точности измерения приборов.

Чтобы дать обобщающую характеристику распределению отклонений, вычисляют среднее линейное отклонение $\bar{d}$, которое учитывает различие всех единиц изучаемой совокупности.

Среднее линейное отклонение показывает стандартное отличие значения каждого варианта от общей средней величины и определяется как средняя арифметическая из отклонений индивидуальных значений от среднего, без учета их знака (по модулю):

$$\bar{d} = \frac{\sum|x - \bar{x}|}{n} = \frac{|x_1 - \bar{x}| + |x_2 - \bar{x}| + \ldots + |x_n - \bar{x}|}{n}. \qquad (2.2.2)$$

Необходимость использования модуля связана со свойством средней арифметической (см. п. 2.1.3), согласно которому сумма индивидуальных отклонений от средней арифметической всегда равна нулю (см. табл. 2.2.2).

Среднее линейное отклонение измеряется в тех же абсолютных единицах, что и значения признака.

***Порядок расчета среднего линейного отклонения следующий*:**

1) по значениям признака находят среднюю арифметическую;
2) определяют отклонения каждого значения X_i от средней $|X_i - \bar{X}|$;
3) рассчитывается сумма абсолютных величин отклонений: $\sum|X_i - \bar{X}|$;
4) сумма абсолютных величин отклонений делится на число значений.

ПРИМЕР.

Рассчитаем среднее линейное отклонение стажа работников одной смены по данным табл. 2.2.4.

Средний стаж работы одного рабочего смены составляет 10 лет (50 : 5 = 10). Отклонения стажа каждого рабочего от среднего стажа, равного 10 годам, рассчитаны в графе 2, их сумма равна нулю.

Таблица 2.2.4

Стаж работников одной смены бригады рабочих на предприятии

Таб. № рабочего	Стаж работы, лет (x_i)	$x_i - \bar{x}$	$\lvert x_i - \bar{x}\rvert$
А	1	2	3
3440	2	–8	8
3537	3	–7	7
3825	12	2	2
2171	15	5	5
2845	18	8	8
Итого	50	0	30

Среднее линейное отклонение составляет:

$$\bar{d} = \frac{\sum |X_i - \bar{X}|}{n} = \frac{30}{5} = 6 \text{ лет},$$

т. е. стаж работы каждого работника данной смены в среднем отличается от среднего значения на 6 лет.

Если данные наблюдения представлены в виде дискретного или интервального ряда распределения с частотами, среднее линейное отклонение исчисляется по формуле средней арифметической взвешенной:

$$\bar{d} = \frac{\sum |X_i - \bar{X}| f_i}{\sum f_i} = \frac{|X_1 - \bar{X}| f_1 + |X_2 - \bar{X}| f_2 + \ldots + |X_n - \bar{X}| f_n}{f_1 + f_2 + \ldots + f_n}. \quad (2.2.3)$$

Порядок расчета взвешенного среднего линейного отклонения:

1) вычисляется средняя арифметическая взвешенная;

2) определяются абсолютные отклонения вариантов от среднего $|X_i - \bar{X}|$;

3) полученные отклонения умножаются на частоты $|X_i - \bar{X}| f_i$;

4) находится сумма взвешенных отклонений без учета знака: $\sum |X_i - \bar{X}| f_i$;

5) сумма взвешенных отклонений делится на сумму частот: $\dfrac{\sum |X_i - \bar{X}| f_i}{\sum f_i}$.

Для нахождения $\bar{d}$ по интервальному ряду распределения вначале находят середины каждого интервала (как среднюю арифметическую из значений границ) и используют их для расчета отклонений от общей средней аналогичным образом.

ПРИМЕР.

Рассчитаем среднее линейное отклонение по данным о производительности труда работников одной смены по данным табл. 2.2.5.

Таблица 2.2.5

Производительность труда работников одной смены бригады рабочих на предприятии

Количество деталей, обрабатываемых одним рабочим, шт.	Количество рабочих, чел. f_i	X_i	$X_i f_i$ (гр. 3 · гр. 2)	$\lvert X_i - \bar{X} \rvert$	$\lvert X_i - \bar{X} \rvert f_i$ (гр. 5 · гр. 2)
1	2	3	4	5	6
40–44	2	42	84	5	10
45–49	1	47	47	0	0
50–54	2	52	104	5	10
Итого	5	–	235	–	20

Для нахождения средней арифметической вначале были рассчитаны середины каждого интервала и записаны в графу 3. Эти значения показывают, сколько в среднем деталей обрабатывает за смену один работник по каждой группе производительности труда. Обозначим эти значения X_i и воспользуемся ими для дальнейших расчетов.

Средняя производительность труда одного рабочего данной смены составляет 235 : 5 = 47 деталей:

Индивидуальные отклонения от средней (без учета знака) указаны в графе 5. Взвешенные отклонения – в графе 6, их сумма составляет 20 единиц.

Следовательно:

$$\bar{d} = \frac{\sum \lvert X_i - \bar{X} \rvert f_i}{\sum f_i} = \frac{20}{5} = 4 \text{ детали.}$$

Это значит, что в среднем каждый работник данной смены обрабатывает за рабочее время на 4 детали меньше или больше, чем средний уровень производительности за смену, составляющий 47 деталей.

Несмотря на то что среднее линейное отклонение учитывает колебания всех значений признака, при его расчете в некотором смысле нарушается элементарное правило математики, так как отклонения от среднего значения суммируются без учета знаков.

Избежать указанного недочета позволяет возведение каждого индивидуального отклонения от средней в квадрат. Рассчитав среднюю величину из возведенных в квадрат отклонений, можно вернуться к исходному порядку данных путем извлечения квадратного корня. Найденное число будет также характеризовать типичное отклонение вариантов признака от средней величины в данной совокупности.

На практике квадратические оценки признаются более состоятельными, поскольку они не нарушают математических правил, а размер погрешности от возведения в квадрат является минимальным. Это объясняется вторым свойством средней арифметической (см. п. 2.1.3), согласно которому сумма квадратов отклонений от средней всегда меньше, чем отклонение от любого другого числа.

Основными показателями вариации, для расчета которых используются квадратические оценки, являются дисперсия и среднее квадратическое отклонение.

Дисперсия — это среднее арифметическое из квадратов отклонений каждого значения признака от общей средней. Дисперсия обычно называется средним квадратом отклонений и обозначается σ^2. В зависимости от исходных данных дисперсия может вычисляться по простой или взвешенной формулам.

Если каждый вариант признака повторяется только один раз, используют простую (невзвешенную) формулу:

$$\sigma^2 = \frac{\sum (X_i - \bar{X})^2}{n}. \tag{2.2.4}$$

Если варианты признака повторяются неодинаковое количество раз, используют взвешенную формулу:

$$\sigma^2 = \frac{\sum (X_i - \bar{X})^2 f_i}{\sum f_i}. \tag{2.2.5}$$

Среднее квадратическое отклонение (в литературе также часто используется термин ***«стандартное отклонение»***) представляет собой квадратный корень из дисперсии и обозначается σ.

Формулы для его расчета следующие:

- среднее квадратическое отклонение невзвешенное:

$$\sigma = \sqrt{\frac{\sum(X_i - \bar{X})^2}{n}}; \qquad (2.2.6)$$

- среднее квадратическое отклонение взвешенное:

$$\sigma = \sqrt{\frac{\sum(X_i - \bar{X})^2 f_i}{\sum f_i}}. \qquad (2.2.7)$$

Математические преобразования формул (2.2.6) и (2.2.7) приводят к упрощенной формуле, которая часто оказывается более удобной на практике, особенно при расчете σ по несгруппированным данным:

$$\sigma = \sqrt{\overline{X^2} - (\bar{X})^2}; \qquad (2.2.8)$$

$$\overline{x^2} = \frac{\sum x_i^2 f_i}{\sum f_i}. \qquad (2.2.9)$$

Среднее квадратическое отклонение является обобщающей характеристикой абсолютных размеров вариации признака в совокупности. Выражается оно в тех же единицах измерения, что и признак (в метрах, тоннах, процентах, гектарах и т. д.), и показывает, на сколько в среднем отличается значение каждого варианта признака от среднего значения для данной совокупности.

Среднее квадратическое отклонение всегда больше среднего линейного отклонения. Между ними имеется соотношение:

$$\sigma = \bar{d} \cdot 1{,}25. \qquad (2.2.10)$$

Чем меньше среднее квадратическое отклонение, тем лучше средняя арифметическая отражает собой всю представляемую совокупность.

Для вычисления среднего квадратического отклонения необходимо рассчитать дисперсию.

Порядок расчета дисперсии по взвешенной формуле следующий.

1. Определяют среднюю арифметическую взвешенную.
2. Рассчитывают отклонения вариантов от средней: $(X_i - \bar{X})$.
3. Возводят в квадрат отклонение каждого варианта от средней: $(X_i - \bar{X})^2$.
4. Умножают квадраты отклонений на веса (частоты): $(X_i - \bar{X})^2 f_i$.
5. Суммируют полученные произведения: $\sum(X_i - \bar{X})^2 f_i$.

6. Полученную сумму делят на сумму весов: $\dfrac{\sum (X_i - \bar{X})^2 f_i}{\sum f_i}$.

ПРИМЕР.

Найдем дисперсию и среднее квадратическое отклонение по данным о размере квартальной премии сотрудников одного из отделов фирмы (табл 2.2.6)

Таблица 2.2.6

Размер квартальной премии сотрудников отдела обслуживания фирмы

Премия x_i, тыс. руб.	Число сотрудников f_i	$x_i f_i$	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$	$(x_i - \bar{x})^2 f_i$
1	2	3	4	5	6
8	7	56	−2	4	28
9	10	90	−1	1	10
10	15	150	0	0	0
11	12	132	1	1	12
12	6	72	2	4	24
Итого	50	500	0	–	74

Вычислим среднюю арифметическую взвешенную: 500 : 50 = 10 тыс. руб.

Значения отклонений от среднего и квадраты этих величин представлены в табл. 2.2.6 в графах 4 и 5.

Определим дисперсию:

$$\sigma^2 = \frac{\sum (X_i - \bar{X})^2 f_i}{\sum f_i} = \frac{74}{50} = 1{,}48.$$

Среднее квадратическое отклонение будет равно:

$$\sigma = \sqrt{\frac{\sum (X_i - \bar{X})^2 f_i}{\sum f_i}} = \sqrt{1{,}48} = 1{,}216 \text{ тыс. руб.}$$

Это значит, что размер премии каждого сотрудника отдела обслуживания отличается от среднеквартальной величины премии по отделу, равной 10 тыс. руб., на 1 тыс. 216 руб.

Если исходные данные представлены в виде интервального ряда распределения, то сначала надо определить середины каждого интервала и далее рассчитывать показатели вариации аналогичным образом.

ПРИМЕР.

Рассчитаем дисперсию и среднее квадратическое отклонение по данным о распределении ценных бумаг по уровню доходности (табл. 2.2.7).

Таблица 2.2.7

Распределение ценных бумаг по уровню доходности

Доходность ценной бумаги, %	Количество ценных бумаг, f_i	X_i	$X_i f_i$	$(X_i - \bar{X})$	$(X_i - \bar{X})^2$	$(X_i - \bar{X})^2 f_i$
14–16	100	15	1500	–3,4	11,56	1156
16–18	300	17	5100	–1,4	1,96	588
18–20	400	19	7600	0,6	0,36	144
20–22	200	21	4200	2,6	6,76	1352
Итого	1000	–	18 400			3240

Средняя доходность одной ценной бумаги равна

$$\bar{X} = \frac{\sum X_i f_i}{\sum f_i} = \frac{18\,400}{1000} = 18{,}4\%.$$

Вычислим дисперсию:

$$\sigma^2 = \frac{\sum (X_i - \bar{X})^2 f_i}{\sum f_i} = \frac{3240}{1000} = 3{,}24.$$

Среднее квадратическое отклонение составит: $\sigma = \sqrt{3{,}24} = 1{,}8\%$.

Следовательно, с каждой ценной бумаги владелец может получить в среднем на 1,8% больше или меньше, чем величина средней доходности – 18,4%.

Для характеристики меры вариации изучаемого признака вычисляются показатели в относительных величинах. Они позволяют сравнивать характер вариации признака в разных совокупностях. Относительные показатели вариации рассчитывают как отношение типичного отклонения к средней величине, умножаемое на 100%.

Линейный коэффициент вариации характеризует долю среднего линейного отклонения от общего размера среднего значения и рассчитывается по формуле:

$$K_d = \frac{\bar{d}}{\bar{x}} 100\%. \tag{2.2.11}$$

Коэффициент вариации рассчитывается как отношение среднего квадратического отклонения к среднему значению:

$$V = \frac{\sigma}{\bar{x}}100\%. \qquad (2.2.12)$$

Учитывая, что среднеквадратическое отклонение дает обобщающую характеристику колеблемости всех вариантов совокупности, коэффициент вариации является наиболее распространенным показателем, используемым для оценки типичности средних значений. **Если *V* больше 33%**, то это говорит о больших колебаниях признака в изучаемой совокупности. Совокупность считается однородной, если коэффициент вариации **не превышает 33%**.

ПРИМЕР.

По данным табл. 2.2.7 рассчитаем коэффициент вариации:

$$V = \frac{\sigma}{\bar{x}}100\% = \frac{1,8}{18,4}100\% = 9,78\%.$$

Это означает, что ценные бумаги однородны по проценту доходности (так как $V = 9,78\% < 33\%$).

Следует отметить, что коэффициент вариации может быть более 100%, что бывает при наличии значений сильно отличающихся от средней величины, например, отрицательных значений, или аномально больших или малых значений отдельных вариантов. Такой результат означает, что в исследуемой совокупности сильна вариация признаков по отношению к средней величине и для более качественного ее исследования необходимо разбить ее на более однородные части или с помощью специальных методов исключить влияние на среднюю аномальных вариантов.

КОНТРОЛЬНЫЕ ВОПРОСЫ

1. Что такое изменчивость или вариация данных?
2. Какой показатель и почему используют для оценки изменчивости?
3. Что характеризует среднее квадратическое отклонение? В каких единицах оно измеряется?
4. В каких случаях для расчета показателей вариации используются взвешенные формулы?
5. Для чего применяется коэффициент вариации? В каких единицах он измеряется?

2.2.3. Математические свойства показателей вариации

На практике часто возникают ситуации, которые системно оказывают влияние на все значения исследуемого признака. Например, при системном сокращении или повышении расценок, переходе к использованию для расчетов условных единиц, переходе с одной валюты на другую (с доллара США к евро или фунтам стерлингов), при изменении шкалы измерений. Такие изменения отражаются в систематическом увеличении или уменьшении каждого значения показателя на некоторые фиксированные величины или в несколько раз. При таких изменениях необходимости в перерасчете всех показателей вариации не возникает, так как они обладают рядом математических свойств, зная которые можно легко определить их величину по совокупности обновленных данных. Рассмотрим основные свойства на примере данных о затратах на производство единицы продукции трех отделений корпорации, каждое из которых в среднем ежемесячно производит 2 тыс. единиц продукции (табл. 2.2.8).

Таблица 2.2.8

Затраты на производство единицы продукции по корпорации

Отделение корпорации	Затраты на единицу продукции X_i, руб.	Средний объем производства, единиц продукции в месяц
А	1	2
1 отделение	10	2000
2 отделение	12	2000
3 отделение	14	2000
Итого	–	6000

Так как каждое отделение производит одинаковое количество продукции, для расчета обобщающих показателей — среднего, дисперсии, среднего квадратического отклонения — можно воспользоваться невзвешенной формулой. Промежуточные расчеты показателей представлены в табл. 2.2.9.

Таблица 2.2.9

Затраты на производство единицы продукции по корпорации

№ отделения	X_i	$X_i - \bar{X}$	$(X_i - \bar{X})^2$
А	1	2	3
1 отделение	10	–2	4

Окончание табл. 2.2.9

№ отделения	X_i	$X_i - \bar{X}$	$(X_i - \bar{X})^2$
2 отделение	12	0	0
3 отделение	14	+2	4
Итого	36	0	8

$$\bar{X} = \frac{36}{3} = 12 \text{ руб.};$$

$$\sigma^2 = \frac{8}{3} = 2{,}6667;$$

$$\sqrt{2{,}667} = 1{,}632 \text{ руб.};$$

$$R = 14 - 10 = 4 \text{ руб.};$$

$$V = \frac{1{,}632}{12} 100\% = 13{,}6\%.$$

Средние затраты на единицу продукции корпорации составляют 12 руб. Риск превышения затрат на единицу продукции составляет 1,632 руб. — величину среднего квадратического отклонения. Общие производственные затраты в месяц составляют в среднем 12 · 6000 = 72 000 руб.

Свойство 1. Уменьшение или увеличение всех значений признака на одинаковую величину не меняет величины дисперсии, среднего квадратического отклонения и размаха вариации:

$$\sigma^2_{x \pm A} = \sigma^2_x, \ \sigma_{x \pm A} = \sigma^2_x, \ R_{x \pm A} = R_x. \tag{2.2.13}$$

Проверим, как повлияет на величину показателей вариации снижение затрат на единицу продукции на 2 руб. в результате оптимизации производственных процессов при неизменном объеме производства (табл. 2.2.10).

Как мы видим, при снижении затрат на производство единицы продукции на 2 руб. риск превышения затрат на единицу продукции остался неизменным — 1,63 руб., так же как и величины дисперсии и размаха вариации. Но это не означает, что корпорации не нужно оптимизировать производство: за счет снижения затрат на единицу общие производственные затраты корпорации составили: 10 · 6000 = 60 000 руб., что на 12 тыс. руб. (или на 16,6%) меньше, чем было до оптимизации — 72 000 руб.

Таблица 2.2.10

Затраты на производство единицы продукции по корпорации

№ отделения	X_{i-2}	$X_{i-2} - \bar{X}$	$(X_{i-2} - \bar{X})^2$
А	1	2	3
1	8	–2	4
2	10	0	0
3	12	+2	4
Итого	30	0	8

$$\bar{X} = \frac{30}{3} = 10 \text{ руб.};$$

$$\sigma^2_{(x-2)} = \frac{8}{3} = 2{,}6667;$$

$$\sigma_{(x-2)} = \sqrt{2{,}667} = 1{,}632 \text{ руб.};$$

$$R_{(x-2)} = 12 - 8 = 4 \text{ руб.}$$

Свойство 2. Увеличение всех значений признака в k раз (k — любое число) увеличивает дисперсию в k^2 раз, среднее квадратическое отклонение и размах — в k раз. Если $k < 0$, то коэффициент берется по модулю. Коэффициент вариации при этом не меняется:

$$\sigma^2_{xk} = \sigma^2_x k^2, \sigma_{xk} = \sigma_x k, R_{xk} = R_x k. \qquad (2.2.14)$$

Предположим условную ситуацию, что все затраты на единицу продукции в результате двойного удорожания цен на сырье выросли в два раза. Как это (при неизменном объеме производства) отразится на риске увеличения затрат на единицу продукции по сравнению с исходным уровнем (по данным табл. 2.2.8)? Проверим расчеты (табл. 2.2.11).

Таблица 2.2.11

Затраты на производство единицы продукции по корпорации

№ отделения	$X_i \cdot 2$	$X_i \cdot 2 - \bar{X}$	$(X_i \cdot 2 - \bar{X})^2$
А	1	2	3
1	20	–4	16
2	24	0	0
3	28	+4	16
Итого	72	0	32

$$\bar{X} = \frac{72}{3} = 24 \text{ руб.};$$

$$\sigma^2_{x \cdot 2} = \frac{32}{3} = 10{,}6667;$$

$$\sigma_{x \cdot 2} = \sqrt{10{,}667} = 3{,}266 \text{ руб.};$$

$$R_{x \cdot 2} = 28 - 20 = 8 \text{ руб.};$$

$$V = \frac{3{,}266}{24}100 = 13{,}6\%.$$

Как мы видим, при увеличении затрат на производство единицы продукции в 2 раза размах вариации вырос в два раза. Риск превышения затрат на единицу продукции составляет 3,266 руб., т. е. также в два раза больше, чем первоначальный — 1,632 руб. $\left(\frac{3{,}266}{1{,}632} = 2\right)$. Соответственно, для получения оценки риска от увеличения затрат на единицу продукции в два раза достаточно исходное среднее квадратическое отклонение 1,632 умножить на 2.

Размер дисперсии составил 10,6667. Это в 4 раза больше, т. е. в 2^2 раза, чем исходное значение $\frac{10{,}6667}{2{,}66670} = 3{,}9999 \approx 4$.

Если немного развить тему кратных преобразований исходных данных, то в результате деноминации с курсом 1 новый руб. = 10 старых руб. затраты на производство единицы продукции в среднем по корпорации составят $10 \cdot (0{,}1) = 1{,}2$ нов. руб., дисперсия — $2{,}6667 \cdot (0{,}1)^2 = 0{,}026\,667$, среднее квадратическое отклонение — $1{,}632 \cdot (0{,}1) = 0{,}1632$ нов. руб., размах — 0,4 нов. руб. $(4 \cdot 0{,}1)$. При желании перевести затраты в доллары США при курсе, например, 30 руб. за доллар, необходимо исходный размер среднеквадратического отклонения и размаха умножить на $^1/_{30} = 0{,}033$, а исходный размер дисперсии — на $0{,}033^2$.

Коэффициент вариации остался неизменным — 13,6%, как и до двукратного увеличения затрат на единицу произведенной продукции.

Свойство 3. Дисперсия отклонений значений признака от произвольного числа $A(X_i - A)(A \neq \bar{X})$ увеличивает дисперсию отклонений от средней $(X_i - \bar{X})$ на число, равное возведенной в квадрат разнице между средней и этим числом A, т. е. на $(\bar{X} - A)^2$:

$$\sigma^2_{(A)} = \sigma^2_x + (\bar{X} - A)^2 \text{ или } \sigma^2_x = \sigma^2_{(A)} - (\bar{X} - A)^2. \qquad (2.2.15)$$

Для иллюстрации свойства предположим, что руководство корпорации решило оценить уровень риска при условии, что средние затраты на единицу продукции в корпорации остались бы на уровне прошлого года — 15 руб. Проверим действие свойства, рассчитав отклонения затрат на единицу продукции от 15 (табл. 2.2.12).

Таблица 2.2.12

Затраты на производство единицы продукции по корпорации

№ отделения	X_i	$X_i - 15$	$(X_i - 15)^2$
А	1	2	3
1 отделение	10	–5	25
2 отделение	12	–3	9
3 отделение	14	–1	1
Итого	36	9	35

Получим:

$$\sigma^2_{(A)} = \frac{35}{3} = 11{,}6667 \text{ при } \sigma^2_{(X)} = 2{,}6667.$$

Таким образом, свойство подтверждено, дисперсия отклонений от числа $A = 15$ больше отклонения от среднего значения на 9:

$$11{,}6667 - 2{,}6667 = 9 = (12-15)^2 = (-3)^2.$$

2.2.4. Правило сложения дисперсий

Изучая вариацию интересующего нас признака в пределах совокупности, разделенной на группы, для более качественного анализа необходимо проследить количественные изменения признака внутри каждой группы, а также между группами. Такой анализ возможен с помощью вычисления различных видов дисперсии. Выделяют три вида дисперсии — общую, внутригрупповую и межгрупповую.

Общая дисперсия (σ^2) характеризует вариацию признака по всей совокупности как результат влияния всех факторов, определяющих индивидуальные различия единиц совокупности:

$$\sigma^2 = \frac{\sum (x_i - \bar{x})^2 f_i}{\sum f_i}. \qquad (2.2.16)$$

Внутригрупповая дисперсия (σ_i^2) отражает случайную вариацию, т. е. ту часть вариации признака, которая обусловлена действием всех прочих неучтенных факторов, кроме фактора, по которому осуществлялась группировка. Внутригрупповая дисперсия рассчитывается отдельно по каждой выделенной группе по формуле:

$$\sigma_i^2 = \frac{\sum(x_i - \overline{x_i})^2}{n_i}, \qquad (2.2.17)$$

где x_i — значения признака у единиц, входящих i-ю группу; $\bar{x}_i$ — среднее значение признака в i-й группе; n_i — число единиц в i-й группе.

Для всех групп в целом вычисляется **средняя из внутригрупповых** дисперсий $\overline{\sigma_i^2}$ по формуле:

$$\overline{\sigma_i^2} = \frac{\sum\sigma_i^2 n_i}{\sum n_i}. \qquad (2.2.18)$$

Межгрупповая дисперсия (δ_x^2) характеризует вариацию, обусловленную влиянием на значения исследуемого признака, положенного в основание группировки:

$$\delta_x^2 = \frac{\sum(\bar{x}_i - \bar{x})^2 n_i}{\sum n_i}, \qquad (2.2.19)$$

где $\bar{x}$— общая средняя величина признака; $\bar{x}_i$ — среднее значение признака в i-й группе; n_i — число единиц в i-й группе, при этом $\sum n_i = \sum f_i$.

Взаимосвязь между тремя видами дисперсий получила название **правила сложения дисперсий**. Согласно ему общая дисперсия равна сумме средней из внутригрупповой и межгрупповой дисперсий:

$$\sigma^2 = \overline{\sigma_i^2} + \delta_x^2. \qquad (2.2.20)$$

Таким образом, зная два вида дисперсий, всегда можно определить третий. Кроме того, при качественно проведенной группировке, на основании соотношения между данными видами дисперсий можно судить о степени влияния группировочного признака на изменение значений других зависящих от него признаков. Такие соотношения называются эмпирическими коэффициентами.

Эмпирический коэффициент детерминации (η^2) (от лат. *determinatio* — ограничение, определение) характеризует долю межгрупповой дисперсии в общей дисперсии и рассчитывается по формуле:

$$\eta^2 = \frac{\delta_x^2}{\sigma^2}. \qquad (2.2.21)$$

Он показывает долю общей вариации изучаемого признака, которую вызывает (определяет) вариация группировочного признака.

Если извлечь квадратный корень из коэффициента детерминации, получим **эмпирическое корреляционное отношение** $\eta_э$ (от позднелат. *correlatio* — соотношение, взаимозависимость, взаимное соответствие) — коэффициент, при помощи которого можно оценить тесноту связи между группировочным (факторным) и результативным признаками. Корреляционное отношение рассчитывается по формуле:

$$\eta_э = \sqrt{\frac{\delta_x^2}{\sigma^2}}. \qquad (2.2.22)$$

Данный коэффициент может принимать значения от 0 до 1. Чем ближе к 1 будет его величина, тем сильнее взаимосвязь между рассматриваемыми признаками.

ПРИМЕР.

Исследуем зависимость объема выполненных работ от формы собственности проектно-изыскательских организаций (табл. 2.2.13).

Таблица 2.2.13

Выполнение работ проектно-изыскательскими организациями разной формы собственности

№ группы	Форма собственности	Количество организаций	Объем выполненных работ (млн руб.)	Итого
1	Государственная	4	10, 30, 20, 40	100
2	Негосударственная	6	30, 45, 60, 20, 65, 50	270
	Итого	10	–	370

1. Рассчитаем общее среднее:

$$\bar{X} = \frac{370}{10} = 37 \text{ млн руб.}$$

2. Рассчитаем средний объем выполненных работ по каждой группе организаций:

- государственные:

$$\bar{X}_1 = \frac{10+30+20+40}{4} = \frac{100}{4} = 25 \text{ млн руб.};$$

- негосударственные:

$$\overline{X}_2 = \frac{30+45+60+20+65+50}{6} = \frac{270}{6} = 45 \text{ млн руб.}$$

3. Рассчитаем внутригрупповые дисперсии объема выполненных работ. По 1 группе — государственным организациям — σ_1^2:

$$\sigma_1^2 = \frac{\sum(x_i - \overline{x_1})^2}{n_1} = \frac{(10-25)^2+(30-25)^2+(20-25)^2+(40-25)^2}{4} =$$

$$= \frac{(-15)^2+(5)^2+(-5)^2+(15)^2}{4} = \frac{225+25+25+225}{4} = \frac{500}{4} = 125.$$

По 2 группе — негосударственным организациям — σ_2^2:

$$\sigma_2^2 = \frac{\sum(x_i - \overline{x_2})^2}{n_2} =$$

$$= \frac{(30-45)^2+(45-45)^2+(60-45)^2+(20-45)^2+(65-45)^2+(50-45)^2}{6} =$$

$$= \frac{(-15)^2+(0)^2+(15)^2+(-25)^2+(20)^2+(5)^2}{6} =$$

$$= \frac{225+0+225+625+400+25}{6} = \frac{1500}{6} = 250.$$

4. Рассчитаем среднюю из внутригрупповых дисперсий:

$$\overline{\sigma_i^2} = \frac{\sum\sigma_i^2 n_i}{\sum n_i} = \frac{\sigma_1^2 n_1 + \sigma_2^2 n_2}{n_1+n_2} = \frac{125\cdot4+250\cdot6}{4+6} = \frac{500+1500}{10} = 200.$$

5. Рассчитаем межгрупповую дисперсию:

$$\delta_x^2 = \frac{\sum(\bar{x}_i - \bar{x})^2 n_i}{\sum n_i} = \frac{(\bar{x}_1-\bar{x})^2 n_1 + (\bar{x}_2-\bar{x})^2 n_2}{n_1+n_2} =$$

$$= \frac{(25-37)^2\cdot4+(45-37)^2\cdot6}{4+6} = \frac{(-12)^2\cdot4+(8)^2\cdot6}{10} =$$

$$= \frac{144\cdot4+64\cdot6}{10} = \frac{576+384}{10} = \frac{960}{10} = 96.$$

6. Рассчитаем общую дисперсию по правилу сложения дисперсий:

$$\sigma^2 = \overline{\sigma_i^2} + \delta_x^2 = 200+96 = 296.$$

7. Для проверки правильности расчетов определим общую дисперсию по формуле:

$$\sigma^2 = \frac{\sum (x_i - \bar{x})^2}{n}.$$

Вспомогательные расчеты представим в табл. 2.2.14.

Таблица 2.2.14

Вспомогательная таблица для расчета общей дисперсии

№ группы организации	Объем работ X_i, млн руб.	$X_i - \bar{X}$	$(X_i - \bar{X})^2$
1	2	3	4
1	10	–27	729
1	30	–7	49
1	20	–17	289
1	40	3	9
2	30	–7	49
2	45	8	64
2	60	23	529
2	20	–17	289
2	65	28	784
2	50	13	169
Итого	370	0	2960

$$\sigma^2 = \frac{\sum (x_i - \bar{x})^2}{n} = \frac{2960}{10} = 296.$$

$$\sigma^2 = \overline{\sigma_i^2} + \delta_x^2 = 200 + 96 = 296.$$

Вычисление дисперсии по обычной формуле и по правилу сложения дисперсий дает одинаковый результат.

8. Проверим, в какой мере группировочный признак – форма собственности организации – оказывает влияние на вариацию объема выполненных работ, и оценим степень из взаимосвязанности.

Рассчитаем эмпирический коэффициент детерминации:

$$\eta^2 = \frac{\delta_x^2}{\sigma^2} = \frac{96}{296} = 0,3243\ (32,43\%).$$

Это значит, что объем выполненных работ на 32,43% зависит от формы собственности проектно-изыскательской организации и на 67,57% от ее внутриорганизационных возможностей.

Рассчитаем эмпирическое корреляционное отношение:

$$\eta_э = \sqrt{\frac{\delta_x^2}{\sigma^2}} = \sqrt{\frac{96}{296}} = \sqrt{0,3243} = 0,57.$$

Степень близости отношения к единице говорит о том, что форма собственности оказывает среднее по силе влияние на объем работ, выполняемый проектно-исследовательскими организациями.

Таким образом, на основе расчета эмпирических коэффициентов мы выясним, что выполнение работ проектно-изыскательскими организациями в малой степени зависит от ее формы собственности, а основной вклад в ее производство вносит внутреняя структура и организованность самого предприятия.

КОНТРОЛЬНЫЕ ВОПРОСЫ

1. Как изменится дисперсия, если все значения признака увеличить на 100?
2. Как отразится на величине коэффициента вариации сокращение каждого значения признака в 4 раза?
3. Что означает правило сложения дисперсий?
4. Что такое межгрупповая дисперсия и что она характеризует?
5. Для чего применяются эмпирический коэффициент детерминации и эмпирическое корреляционное отношение? В каких единицах они измеряются?

Раздел 3

АНАЛИЗ ВЗАИМОСВЯЗИ МЕЖДУ СОЦИАЛЬНО-ЭКОНОМИЧЕСКИМИ ЯВЛЕНИЯМИ

Глава 3.1

ИЗМЕРЕНИЕ И ПРОГНОЗИРОВАНИЕ ВЗАИМОСВЯЗИ С ПОМОЩЬЮ КОРРЕЛЯЦИОННО-РЕГРЕССИОННОГО АНАЛИЗА

- Классификация видов взаимосвязи
- Методы изучения связи
- Исследование взаимосвязи с помощью диаграммы рассеяния
- Условия применения корреляционно-регрессионного анализа
- Расчет линейного коэффициента корреляции
- Задачи применения регрессионного анализа
- Вычисление параметров линейной парной регрессии
- Вычисление параметров уравнения регрессии при нелинейной зависимости
- Принятие решений на основе уравнений регрессии
- Проведение корреляционно-регрессионного анализа по сгруппированным данным

3.1.1. Классификация видов взаимосвязи

Окружающий нас мир полон всевозможных взаимосвязей, например: между отношением к труду и производительностью, между корпоративной стратегией и долей рынка, между вмешательством государства и состоянием экономики, между объемом выпускаемой продукции и затратами на производство и т. п.

Исследование объективно существующих связей между социально-экономическими явлениями и процессами является важнейшей задачей теории статистики. В процессе статистического исследования зависимостей вскрываются причинно-следственные отношения между явлениями, что позволяет выявлять факторы (признаки), оказывающие основное влияние на вариацию изучаемых явлений и процессов. **Причинно-следственные отношения** — это такая связь явлений и процессов, при которой изменение одного из них — причины — ведет к изменению другого.

Всю статистическую информацию, с которой мы имеем дело в данном курсе, условно можно разделить на две группы: одномерные и двумерные статистические данные. **Одномерными** называются данные, состоящие из одной характеристики о каждой единице совокупности (например, численность населения страны в разрезе регионов). Когда мы исследуем социально-экономические явления, то имеем дело с **двумерными** данными (например, заработной платой и образованием сотрудников компании), так как качественные результаты возможно получить только при совместном изучении обоих явлений и выявить взаимосвязь между ними.

На финансово-экономические явления и процессы оказывает одновременное воздействие большое количество всевозможных причин. Следовательно, при изучении этих процессов необходимо выявлять главные, основные причины, абстрагируясь от второстепенных. Знание этой информации может оказать значительную помощь в долгосрочном планировании и принятии других стратегических решений. Например, вашу фирму интересуют результаты размещения рекламы в различных журналах для оценки ее маркетинговой стратегии, ориентированной на получение максимальной прибыли. По каждому рекламному объявлению вы располагаете информацией о его стоимости, объеме и количестве запросов, вызванных его появлением в журнале, планируемой целевой аудитории того или иного СМИ (характеристиках целевой аудитории выбранного СМИ: среднемесячной величине дохода клиентов журнала, уровне их образования и занимаемом месте на должностной лестнице. Для выяснения связи между количеством потенциальных клиентов, появившихся у фирмы вследствие размещения этого рекламного объявления, и затратами на его размещение, необходимо отобрать максимально существенные факторы, значимо влияющие на приток клиентов и повышение объемов продаж компании.

Признаки по их сущности и значению для изучения взаимосвязи делятся на два класса. Признаки, обуславливающие изменения других, связанных с ними признаков, называются **факторными** или просто факторами. Признаки, изменяющиеся под действием факторных признаков, называются **результативными**.

В статистике различают функциональную и стохастическую зависимости. **Функциональной** называют такую связь, при которой определенному значению факторного признака соответствует одно и только одно значение результативного признака. Например, если обозначить x независимую переменную, а y — зависимую, связь $y = x^2 + 7$ будет являться функциональной, так как каждому значению факторного признака x точно соответствует определенное значение результативного признака y (при $x = 0$ значение $y = 7$, при $x = 2$ значение $y = 11$ и т. д.).

Наиболее часто функциональные связи проявляются при изучении физических явлений, например, зависимость расстояния, пройденного объектом, от скорости его движения является функциональной зависимостью. Примером такой зависимости при изучении экономических показателей может являться размер выплат по кредиту, начисляемых на основании установленной процентной ставки; показатель доходности ценной бумаги, рассчитываемой по математической формуле, а следовательно, находящейся в функциональной зависимости от курса ценной бумаги.

Если причинная зависимость проявляется не в каждом отдельном случае, а в общем, среднем, при большом числе наблюдений, то такая зависимость называется **стохастической**. Частным случаем стохастической связи является **корреляционная связь**, при которой изменение среднего значения результативного признака обусловлено изменением факторных признаков. Например, известно, что прибыль туристической фирмы определенным образом связана с количеством клиентов. Тем не менее нельзя вычислить точный ожидаемый размер прибыли при заданном значении количества клиентов, так как помимо клиентской базы она находится в зависимости и от множества других факторов, среди которых имеются и случайные. Их действие и приводит к стохастической зависимости. Таким образом, мы можем рассчитать лишь среднее значение прибыли, которое будет получено в целом по совокупности туристических фирм со схожим объемом клиентской базы.

Связи между явлениями и их признаками классифицируются по степени тесноты, направлению и аналитическому выражению (рис. 3.1.1).

Связи различают по степени тесноты (табл. 3.1.1).

Таблица 3.1.1

Количественные критерии оценки тесноты связи

Величина показателя связи	Теснота связи
до ±0,3	практически отсутствует
±0,3 – ±0,5	слабая
±0,5 – ±0,7	умеренная
±0,7 – ±1,0	сильная

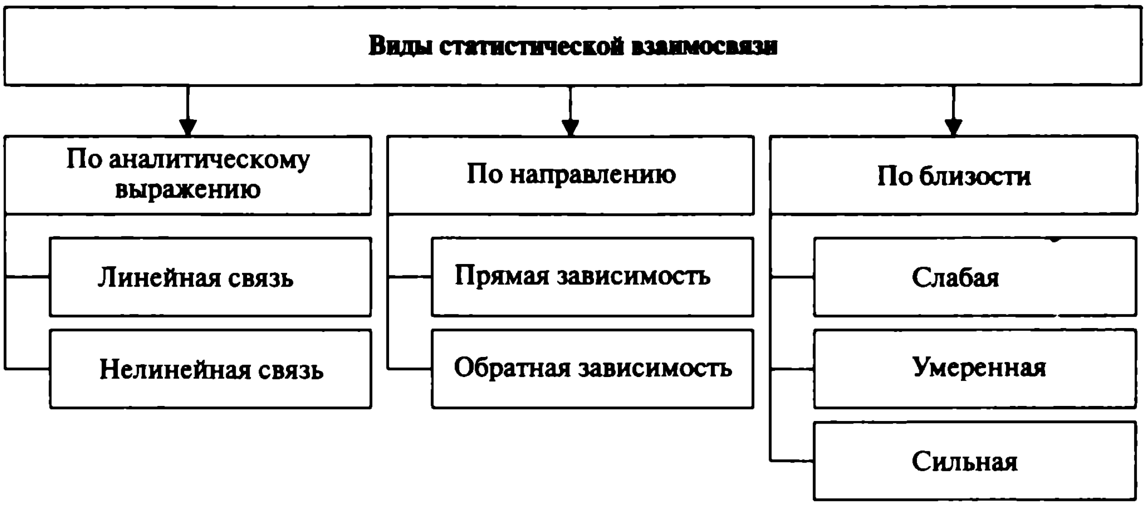


Рис. 3.1.1. Классификация видов взаимосвязи

По направлению выделяют связь прямую и обратную. Связь называется прямой, если с увеличением или уменьшением значений факторного признака происходит увеличение или уменьшение значений результативного признака. Так, рост объемов производства способствует увеличению прибыли предприятия. В случае **обратной** связи значения результативного признака изменяются под воздействием факторного, но в противоположном направлении по сравнению с изменением факторного признака. При обратной связи с увеличением или уменьшением значений факторного признака происходит уменьшение или увеличение значений признака-результата. Так, снижение себестоимости единицы производимой продукции влечет за собой рост прибыли.

По аналитическому выражению выделяют линейные и нелинейные связи. Если статистическая связь между явлениями (X и Y) может быть приблизительно описана уравнением прямой линии ($\bar{y}_x = a_0 + a_1 x$), то ее называют **линейной** связью. Если же связь можно описать уравнением какой-либо кривой, например, параболы ($\bar{y}_x = a_0 + a_1 x + a_2 x^2$), то такую связь называют **нелинейной**, или криволинейной.

Реализация корреляционно-регрессионного метода анализа включает в себя следующие этапы.

1. Выявление из совокупности наиболее информативных факторов, оказывающих существенное влияние на результативный показатель. Количество отбираемых для корреляционно-регрессионного анализа объяс-

няющих признаков X должно быть оптимальным, так как слишком большое их число приведет к «расплывчивости» модели связи, усложнит ее использование на практике, слишком малое — к неполному описанию зависимости исследуемого явления от влияния на него совокупности факторов.

2. Определение направления и количественной оценки тесноты связи между факторным и результативным признаками при парной корреляции и между результативным и множеством факторных признаков при множественной корреляции. Если исследователь работает с многофакторной моделью, то на этом этапе строится матрица парных коэффициентов корреляции, проводится проверка связи между признаками на наличие мультиколлинеарности (тесной статистической зависимости между факторными признаками, которые исследователь планировал включить в модель) и в конечном счете отбор значимых факторных признаков для построения модели регрессии.

3. Построение модели регрессии, описывающей зависимость результативного признака Y от наиболее информативных признаков X. Данная модель аналитически выражает зависимость условного среднего значения результативного признака от факторных переменных $\bar{y}_x = f(X_1, X_2, ..., X_n)$.

4. Оценка статистической значимости уравнения регрессии и коэффициентов регрессии. Определение возможной величины ошибки получаемых по этой модели прогнозных значений Y.

5. Расчет и анализ дополнительных показателей для расширения экономической интерпретации уравнения регрессии.

6. Экономическая интерпретация, формулирование выводов, построение прогнозов, разработка предложений.

В соответствии с различными целями проведения корреляционно-регрессионного анализа исследования могут различаться по степени сложности. В одних случаях необходимо лишь подтвердить связь, в других провести фундаментальное исследование путем построения многофакторной регрессионной модели. Для проведения качественного корреляционно-регрессионного анализа исследователь может воспользоваться как наиболее распространенным программным продуктом Excel, так и специализированными программами SPSS и STATISTICA, работая в которых можно построить регрессионное уравнение любой степени сложности.

3.1.2. Методы изучения взаимосвязи

Для выявления наличия связи, ее характера и направления в статистике используются методы:

- приведения параллельных данных;
- графический (диаграмма рассеяния);
- аналитических группировок;
- корреляционного анализа;
- регрессионного анализа.

Метод приведения параллельных данных основан на сопоставлении двух или нескольких рядов статистических величин. Такое сопоставление позволяет установить наличие связи и получить представление о ее характере.

ПРИМЕР.

Необходимо сравнить изменение двух величин *X* (время обращения монеты, лет) и *Y* (вес монеты, г).

Таблица 3.1.2

X, лет	5	9	14	17	23	31	35	42	46	50
Y, г	2,82	2,85	2,80	2,80	2,79	2,78	2,77	2,79	2,75	2,72

С увеличением значений факторного признака *X*, результативный признак снижается. Следовательно, связь между возрастом монет и их весом обратная, и описать ее можно с помощью линейной зависимости.

Графический метод предполагает изображение с помощью диаграммы рассеяния (поля корреляции) взаимосвязи двух признаков. Диаграмма рассеяния представляет каждую единицу совокупности в пространстве двух измерений, соответствующих двум факторам. Если одна переменная рассматривается как «причина», влияющая на другую переменную, она обозначается буквой «X» и ей соответствует горизонтальная ось. Переменная, изменяющаяся под этим влиянием, носит название «следствие», обозначается буквой «Y» и располагается по вертикальной оси.

Следовательно, при построении диаграммы рассеяния необходимо соблюдать следующие требования:

- в системе координат на оси абсцисс отложить значения факторного признака, а на оси ординат — результативного;
- каждое пересечение линий, проводимых через эти оси, обозначить соответствующими точками.

При отсутствии тесных связей имеет место беспорядочное расположение точек на графике. Чем сильнее связь между признаками, тем ближе будут группироваться точки вокруг определенной линии, выражающей форму связи.

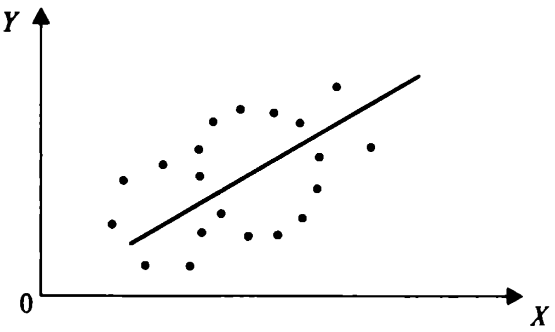


Рис. 3.1.2. Диаграмма рассеяния (поле корреляции)

Метод корреляционного анализа используется для количественного определения тесноты и направления связи между двумя признаками, при парной корреляции.

Теснота связи количественно выражается величиной коэффициента корреляции. Знаки при коэффициенте корреляции характеризуют направление связи между изучаемыми признаками. Если коэффициент корреляции является положительным числом, то говорят о прямой связи между изучаемыми показателями. При отрицательном значении коэффициента корреляции связь признается обратной.

Регрессия тесно связана с корреляцией и позволяет исследовать аналитическое выражение взаимосвязи между признаками.

Метод регрессионного анализа заключается в определении аналитического выражения связи, в котором изменение одной величины (называемой зависимой или результативным признаком) обусловлено влиянием одной или нескольких независимых величин (факторных признаков).

3.1.3. Исследование взаимосвязи с помощью диаграммы рассеяния

При анализе двумерных данных целесообразно строить диаграмму рассеяния (поле корреляции), которая позволяет увидеть структуру изучаемых социально-экономических явлений и процессов. Так же как при анализе одномерных данных, строится гистограмма, позволяющая отобразить структуру совокупности, диаграмма рассеяния иллюстрирует все процессы, происходящие с двумерными данными. Если данные содержат аномальные явления (выбросы или неожиданные особенности), то наиболее простой способ их обнаружения лежит в построении поля корреляции (диаграммы рассеяния).

Рассмотрим различные типы взаимосвязи, которые могут быть выявлены при анализе диаграммы рассеяния для двумерной совокупности данных.

I. Взаимосвязь отсутствует

Взаимосвязь в двумерной совокупности данных полностью отсутствует, если соответствующая диаграмма рассеяния точек носит совершенно случайный характер, т. е. как положительная, так и отрицательная тенденции на графике не наблюдаются. Случай полного отсутствия взаимосвязи представляет собой особый случай линейной взаимосвязи — без увеличения или уменьшения. Такая диаграмма рассеяния точек может иметь вид либо круглого, либо овального облака (причем овал может быть направлен как горизонтально, так и вертикально, но не иметь наклона).

Яркой иллюстрацией отсутствия взаимосвязи может служить рис. 3.1.3, на котором представлен совершенно произвольный разброс точек без какой-либо ярко выраженной тенденции.

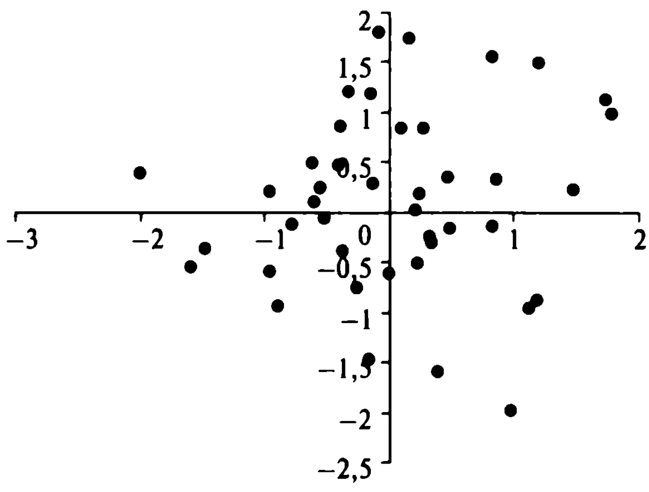


Рис. 3.1.3. Диаграмма рассеяния, свидетельствующая об отсутствии какой-либо взаимосвязи между двумя изучаемыми признаками

II. Линейная взаимосвязь

Легче всего поддается анализу и интерпретации двумерная совокупность данных, выраженная в форме линейной взаимосвязи. Данная взаимосвязь проявляется в двумерной совокупности данных, если точки на диаграмме рассеяния концентрируются вокруг прямой линии. Эти точки могут концентрироваться как довольно тесно, почти точно попадая на прямую линию, так и быть разбросаны достаточно широко, образуя некоторое облако, но в данных не должно быть сильных разбросов (резко отклоняющихся значений) (рис. 3.1.4, 3.1.5).

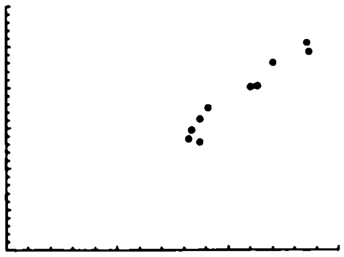


Рис. 3.1.4. Диаграмма рассеяния, выражающая линейную положительную взаимосвязь

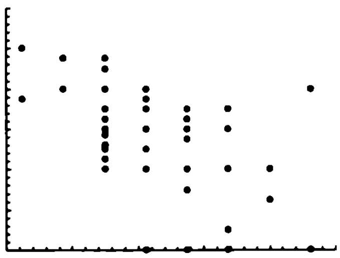


Рис. 3.1.5. Диаграмма рассеяния, выражающая линейную отрицательную взаимосвязь

III. Нелинейная взаимосвязь

Значительно сложнее поддается анализу двумерная совокупность данных в случае наличия нелинейной взаимосвязи, при которой точки на диаграмме рассеяния группируются вокруг некоторой кривой, а не вокруг прямой линии. Поскольку разновидностей кривых линий существует множество, анализ такой взаимосвязи существенно осложнен. Примеры двух наиболее распространенных видов нелинейной связи приведены на рис. 3.1.6 и 3.1.7.

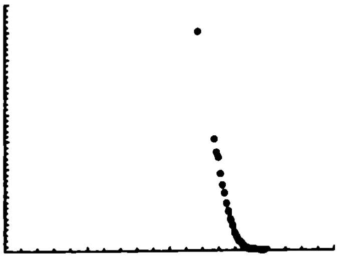


Рис. 3.1.6. Диаграмма рассеяния, характеризующая нелинейную взаимосвязь, выраженную уравнением гиперболы

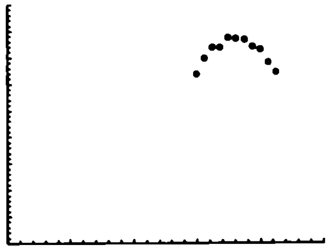


Рис. 3.1.7. Диаграмма рассеяния, характеризующая нелинейную взаимосвязь, выраженную уравнением параболы второго порядка

ПРИМЕР.

Построим диаграмму рассеяния для совокупности данных, представленных в табл. 3.1.3, и охарактеризуем тип взаимосвязи, которую иллюстрирует диаграмма рассеяния.

Таблица 3.1.3

Данные о величине списка почтовой рассылки и объеме продаж полиграфической продукции

Менеджер проекта	Величина списка контактов, тыс. чел.	Объем продаж, тыс. руб.
1	168	5178
2	21	2370
3	94	3591
4	39	2056
5	249	7325
6	43	2449
7	589	15 708
8	41	2469

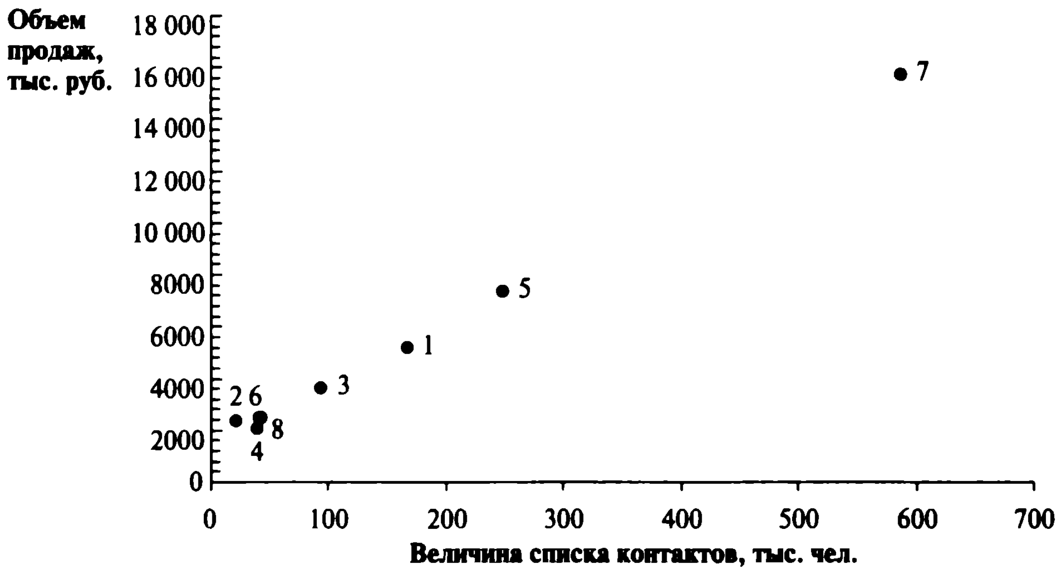


Рис. 3.1.8. Диаграмма рассеяния

Рисунок позволяет увидеть распределение количества контактов и объемов продаж, а также общую тенденцию возрастания объема продаж при увеличении количества контактов

При построении диаграммы рассеяния для данной двумерной совокупности данных необходимо определить факторный и результативный признаки. Так как затраченные усилия влияют на результат, следовательно, число контактов с клиентами признаем факторным признаком и изображаем его на горизонтальной оси, а объем продаж выступает результативным признаком и его располагаем по вертикальной оси. Далее помещаем точки на поле корреляции, иллюстрирующие, как величина объема продаж будет изменяться от количества сделанных контактов. Таким образом, нанеся все точки на график, получим диаграмму рассеяния, которая представлена на рис. 3.1.8.

На диаграмме представлена информация как о каждой отдельной единице совокупности, так и о взаимосвязи между ними. Во-первых, распределение количества контактов (ось абсцисс) находится приблизительно в диапазоне от 20 до 300, причем типичное значение составляет примерно 150. Во-вторых, распределение объемов продаж (ось ординат) находится в диапазоне приблизительно от 2000 тыс. руб. до 8000 тыс. руб., причем типичное значение равно приблизительно 5000 тыс. руб. В-третьих, взаимосвязь между количеством контактов и объемом продаж можно признать положительной: точки на диаграмме выстраиваются снизу вверх при движении слева направо. Это свидетельствует о том, что менеджеры, имеющие большее количество контактов с клиентами, обеспечили компании и большие объемы сбыта (точки, расположенные на диаграмме, имеют тенденцию к росту). В-четвертых, полученную взаимосвязь можно признать линейной, так как результативный и факторный признаки возрастают одинаково, примерно в арифметической прогрессии.

3.1.4. Условия применения корреляционно-регрессионного анализа

Методы корреляционно-регрессионного анализа могут быть применимы только к статистическим данным, удовлетворяющим следующим основным условиям:

1. **Единицы исследуемой совокупности должны иметь одинаковую размерность и методологию расчета.**

2. **Факторные и результативный признаки должны быть выражены количественно.**

3. **Используемые для исследования числовые характеристики должны являться случайно выбранными из единиц генеральной совокупности.** В противном случае исходные данные, представляющие собой определенную выборку из генеральной совокупности, не будут отражать ее характер, полученные по ним выводы о закономерности развития окажутся неверными с практической точки зрения.

4. **Единицы исследуемой совокупности должны быть независимыми друг от друга.** Зависимость единиц совокупности друг от друга в статистике называется *автокорреляцией*. В случае автокорреляции в теории корреляционно-регрессионного анализа существуют специальные методы информационной обработки.

5. **Совокупность исследуемых исходных данных должна быть однородной.** Резко выделяющиеся наблюдения могут привести к смещенным оценкам параметров уравнения регрессии и ложным выводам. Для оценки степени однородности изучаемой совокупности на практике используется вариационный анализ.

6. **Совокупность исходных данных должна подчиняться нормальному закону распределения.** Распределение данных по нормальному закону необходимо для того, чтобы при проверке значимости коэффициентов корреляции и построения для них доверительных интервалов можно было бы использовать определенные критерии. Элементарный метод проверки нормальности распределения сводится к построению диаграммы рассеяния и определению тенденции в расположении нанесенных точек. Если точки концентрируются вокруг прямой, то можно говорить об их подчинении нормальному закону распределения. В регрессионном анализе при построении уравнения регрессии требование нормальности распределения исходных данных предъявляется только к результативному признаку, а факторные признаки могут иметь произвольный закон распределения.

7. **Количество единиц совокупности, по которым проводится корреляционно-регрессионный анализ, должно превышать количество факторных признаков минимум в 3–4 раза (предпочтительнее в 8–10 раз).** Это связано с тем, что статистическая связь проявляется только при значительном числе наблюдений на основе действия закона больших чисел. Суть закона больших чисел состоит в том, что с увеличением числа наблюдений все более отчетливо проявляется статистическая закономерность.

8. **Факторные признаки не должны находиться между собой в функциональной зависимости.** Существенная связь факторных признаков в статистике называется *мультиколлинеарностью*. Ее наличие приводит к построению неустойчивых регрессионных моделей. При обнаружении мультиколлинеарности один из признаков следует исключить из дальнейшего анализа, при этом исключается тот, который оказывает минимальное влияние на результативный признак.

3.1.5. Расчет линейного коэффициента корреляции

Измерение тесноты и направления связи является важной задачей изучения и количественной оценки зависимости социально-экономических явлений. Оценка тесноты связи между признаками предполагает определение меры соответствия вариации результативного признака от одного (при парной корреляции) или нескольких (при множественной корреляции) факторов.

Линейный коэффициент корреляции, обозначаемый r, характеризует тесноту и направление связи между двумя коррелируемыми признаками в случае наличия между ними линейной зависимости.

Линейный коэффициент корреляции изменяется в пределах от -1 до 1: $-1 < r < 1$. Корреляция, равная 1, указывает на идеальную взаимосвязь в виде прямой линии, причем более высокие значения одной переменной соответствуют предсказуемым более высоким значениям другой переменной. Корреляция, равная -1, указывает на отрицательную идеальную взаимосвязь в виде прямой линии, причем одна переменная уменьшается с ростом другой. Нулевое значение коэффициента корреляции указывает на отсутствие линейной связи между признаками. Однако нельзя говорить об отсутствии связи, зависимость между x и y может быть нелинейной.

Обычная интерпретация промежуточных значений коэффициента корреляции заключается в том, что абсолютное значение корреляции указывает на «силу» взаимосвязи, а знак (положительный либо отрицательный) указывает направление (увеличение или уменьшение) связи. При значении коэффициента корреляции, близком к нулю, можно сделать вывод, что связи либо нет, либо она выражена нелинейно. В табл. 3.1.4 представлена интерпретация связи в каждом конкретном случае.

В теории разработаны и на практике используются различные модификации формулы расчета данного коэффициента:

$$r = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\sigma_x \cdot \sigma_y}, \tag{3.1.1}$$

где n — число наблюдений; $\overline{xy} = \frac{\sum_{i=1}^{n} x_i y_i}{n}$; $\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$; $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$; $\sigma_x = \sqrt{\overline{x^2} - \bar{x}^2}$; $\sigma_y = \sqrt{\overline{y^2} - \bar{y}^2}$.

Для вычисления линейного коэффициента корреляции удобно оформлять расчеты в виде вспомогательной табл. 3.1.5.

Таблица 3.1.4

Оценка линейного коэффициента корреляции

Значение линейного коэффициента связи	Характеристика связи	Интерпретация связи	Наглядное представление связи
$r = 0$	Взаимосвязь отсутствует	–	Совершенно случайное облако, не имеющее ориентации ни вверх, ни вниз при движении вправо
Близко к 0, но положительно	Незначительная положительная взаимосвязь	С увеличением факторного признака слабо возрастает результативный признак	Точки данных образуют случайное облако с незначительной ориентацией вверх и вправо
Близко к 1	Сильная положительная взаимосвязь	С увеличением факторного признака увеличивается результативный признак	Точки данных довольно плотно сгруппированы (с небольшим случайным разбросом) вокруг прямой линии, направленной вверх и вправо
Близко к 0, но отрицательно	Незначительная отрицательная взаимосвязь	С увеличением факторного признака незначительно уменьшается результативный, и наоборот	Точки данных образуют случайное облако с незначительной ориентацией вниз и вправо
Близко к –1	Сильная отрицательная взаимосвязь	С увеличением факторного признака уменьшается результативный и наоборот	Точки данных довольно плотно сгруппированы (с небольшим случайным разбросом) вокруг прямой линии, направленной вниз и вправо
$r = 1$	Функциональная	Определенному значению факторного признака соответствует строго одно значение результативного	Идеальная положительная взаимосвязь, т. е. все точки данных располагаются строго на прямой линии, направленной вверх и вправо
$r = -1$			Идеальная отрицательная взаимосвязь, т. е. все точки данных располагаются строго на прямой линии, направленной вниз и вправо

Таблица 3.1.5

Вспомогательная таблица для определения линейного коэффициента корреляции

Число наблюдений, n	Факторный признак, x	Результативный признак, y	y^2	x^2	xy
1	x_1	y_1	y_1^2	x_1^2	x_1y_1
2	x_2	y_2	y_2^2	x_2^2	x_2y_2
3	x_3	y_3	y_3^2	y_3^2	x_3y_3
Итого	$\sum_{i=1}^{n} x_i$	$\sum_{i=1}^{n} y_i$	$\sum_{i=1}^{n} y_i^2$	$\sum_{i=1}^{n} x_i^2$	$\sum_{i=1}^{n} x_i y_i$
Средняя	$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$	$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$	$\bar{y} = \frac{\sum_{i=1}^{n} y_i^2}{n}$	$\bar{x} = \frac{\sum_{i=1}^{n} x_i^2}{n}$	$\overline{xy} = \frac{\sum_{i=1}^{n} x_i y_i}{n}$

Если преобразовать формулу (3.1.1), то линейный коэффициент корреляции можно вычислить, используя итоговые суммы из вспомогательной табл. 3.1.5:

$$r = \frac{n\sum xy - \sum x \sum y}{\sqrt{\left[n\sum x^2 - \left(\sum x\right)^2\right] \cdot \left[n\sum y^2 - \left(\sum y\right)^2\right]}}. \tag{3.1.2}$$

Числитель выражает взаимодействие двух переменных и определяет знак корреляции. Если между переменными существует сильная положительная взаимосвязь, числитель примет положительное значение, если сильная обратная — числитель примет отрицательное значение. Знаменатель формул (3.1.1) и (3.1.2) всегда положителен, так как он является произведением средних квадратических отклонений по x и y, которые всегда больше нуля, поскольку рассчитываются как квадратный корень из дисперсии.

После расчета линейного коэффициента корреляции следует проверить его значимость. В связи с тем, что исходные данные, по которым устанавливается взаимосвязь признаков, являются определенной выборкой из некоторой генеральной совокупности, вычисленный коэффициент корреляции будет также выборочным. Следовательно, с его помощью можно осуществить связь только исходя из информации, которую представляют выборочные единицы совокупности. Коэффициент корреляции будет показывать реальную связь, присущую всей исследуемой совокупности, только в том случае, если отобранные единицы хорошо отражают закономерность и структуру генеральной совокупности. С увеличением числа наблюдений повышается достоверность коэффициента корреляции.

Значимость линейного коэффициента корреляции проверяется на основе *t-критерия Стьюдента*. При этом выдвигается и проверяется гипотеза H_0 о равенстве коэффициента корреляции нулю $H_0 : r_{xy} = 0$. Расчетное значение критерия определяется по формуле:

$$t_p = \sqrt{\frac{r_{xy}^2}{1 - r_{xy}^2} \cdot (n-2)} = \frac{|r_{xy}|}{\sqrt{1 - r_{xy}^2}} \cdot \sqrt{n-2}. \qquad (3.1.3)$$

Расчетное значение t-критерия сравнивается с его табличным, определяемым по таблице табулированных значений:

$$t_{\text{табл}} = \{\alpha;\ \upsilon = n - 2\}, \qquad (3.1.4)$$

где α — уровень значимости, который показывает вероятность принятия ошибочного решения; υ — число степеней свободы, характеризует количество свободно варьируемых элементов совокупности.

Возможность присутствия ошибки связана с тем, что при исследовании взаимосвязи данные взяты не из всей совокупности, а только из ее части. Обычно α принимает значения 0,05; 0,02; 0,01 и 0,001.

Если расчетное значение t-критерия по модулю превышает табличное, то коэффициент корреляции признается значимым.

Если расчетное значение t-критерия по модулю меньше критического, то гипотеза о равенстве коэффициента корреляции нулю принимается с вероятностью α и он признается незначимым, а, следовательно, не может быть использован для характеристики связи между изучаемыми признаками генеральной совокупности, так как единицы выборочной совокупности не отражают реальную структуру генеральной совокупности.

ПРИМЕР.

По данным ежегодного исследования, которое «Коммерсантъ» готовит совместно с газетой «Туринфо», в 2006 г. темпы роста (ОПД, рассмотренный в разд. 1.5.2) российской туриндустрии почти втрое замедлились и составили всего 23,5% против 66,7% в 2005 г. Компании из top10 показали 29%-й рост, а суммарный оборот 50 крупнейших турфирм России впервые превысил двухмиллиардную отметку, составив 2,472 млрд долл. США.

Пальму первенства держат столичные компании, созданные в начале 90-х годов прошлого века: «Инна Тур», TEZ Tour, «Натали Турс», Coral Travel. Их общий оборот составил почти треть от суммарного оборота компаний, участвующих в рейтинге. У восьми из первых десяти компаний финансовый оборот по итогам 2006 г. составил более 100 млн долл. Пятерку лидеров замыкает Петербургская «Фирма Нева». Ее годовой оборот составил 138 млн долл., но по количеству обслуженных туристов компания из северной столицы занимает третье место – 279 тыс. человек.

Таблица 3.1.6

Расчетная таблица для определения коэффициента корреляции

№ п/п	Название фирмы	Финансовый оборот за 2005 г. (млн долл. США)	Количество обслуженных туристов в 2005 г. (тыс. чел.)	y^2	x^2	xy
1	ООО «Инна Тур»	270,00	534,00	72 900,00	285 156	144 180,00
2	Группа компаний TEZ TOUR	235,00	505,00	55 225,00	255 025	118 675,00
3	ООО «Натали Турс»	186,00	200,00	34 596,00	40 000	37 200,00
4	Coral Travel/Odeon Tours (Группа компаний OTI)	138,60	226,00	19 209,96	51 076	31 323,60
5	ЗАО «ФИРМА НЕВА»	138,40	279,00	19 154,56	77 841	38 613,60
6	ЗАО «Спутник»	107,80	300,00	11 620,84	90 000	32 340,00
7	ОАО ВАО «Интурист»	106,88	415,00	11 423,33	172 225	44 355,20
8	ООО «Компания ПАК ГРУПП»	97,20	106,00	9447,84	11 236	10 303,20
9	ОАО «Приморское агентство авиационных компаний»	95,98	228,00	9212,16	51 984	21 883,44
10	ООО «Капитал Тур»	88,00	234,00	7744,00	54 756	20 592,00
	Итого	1463,86	3027,00	250 533,7	1 089 299	499 466,00
	Средняя	146,39	302,70	25 053,37	108 929,9	49 946,6

В табл. 3.1.6 приведена двумерная совокупность данных, включающая финансовый оборот фирмы (млн долл. США) и количество обслуженных туристов (чел.). Взаимосвязь между данными показателями должна быть положительной, поскольку более высокое число обслуженных фирмой туристов приводит к более высоким финансовым результатам.

Диаграмма рассеяния, представленная на рис. 3.1.9, отображает линейную взаимосвязь между финансовым оборотом компаний и количеством обслуженных клиентов. Данная взаимосвязь характеризуется значительным разбросом точек и тенденцией к росту.

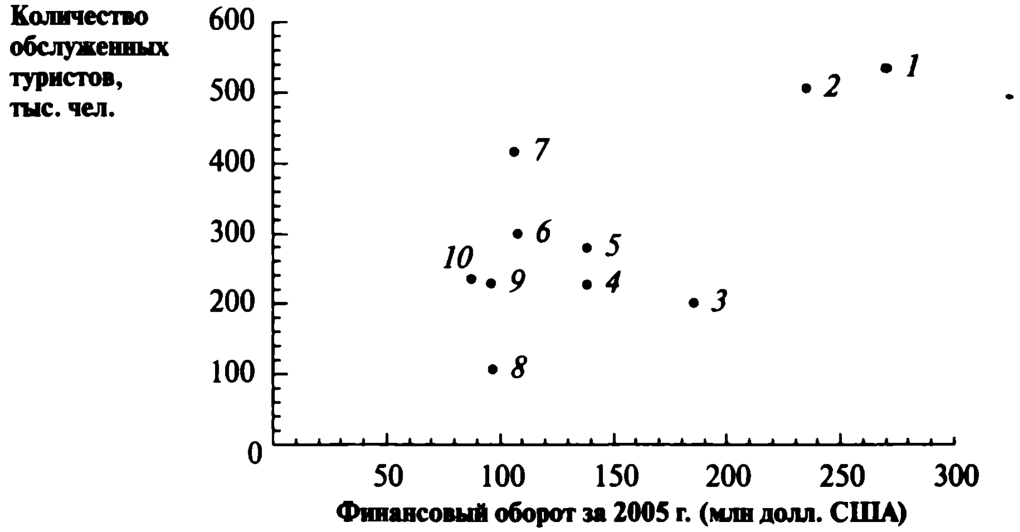


Рис. 3.1.9. Диаграмма рассеяния, характеризующая финансовый оборот наиболее крупных туристических фирм и количество обслуженных ими клиентов

Подтвердим наше предположение о наличии положительной связи между представленными показателями на основании расчета линейного коэффициента корреляции.

1. Используя формулу (3.1.1), получаем:

$$\sigma_y^2 = \overline{y^2} - (\bar{y})^2 = 25053{,}37 - (146{,}386)^2 = 3624{,}51,$$

$$\sigma_x^2 = \overline{x^2} - (\bar{x})^2 = 108\,929{,}9 - (302{,}7)^2 = 17302{,}61,$$

$$r_{xy} = \frac{49\,946{,}6 - 302{,}7 \cdot 146{,}386}{\sqrt{17\,302{,}61 \cdot 3624{,}51}} = 0{,}71.$$

2. По формуле (3.1.2) значение коэффициента корреляции составило:

$$r = \frac{10 \cdot 499466 - 3027 \cdot 1463{,}86}{\sqrt{\left[10 \cdot 1089299 - (3027)^2\right] \cdot \left[10 \cdot 250533{,}7 - (1463{,}86)^2\right]}} =$$

$$= \frac{4994660 - 4431104{,}2}{\sqrt{(10892990 - 9162729) \cdot (2505337 - 2142886)}} =$$

$$= \frac{563555{,}8}{\sqrt{1730261 \cdot 362451}} = \frac{563555{,}8}{791918{,}4489} = 0{,}71.$$

Таким образом, полученный результат 0,71 свидетельствует о сильной прямой зависимости между изучаемыми признаками.

Проверим полученный коэффициент корреляции на значимость:

$$t_p = \frac{|r_{xy}|}{\sqrt{1 - r_{xy}^2}} \cdot \sqrt{n-2} = \frac{0{,}71}{\sqrt{1 - 0{,}71^2}} \cdot \sqrt{10-2} = 4{,}05.$$

$$t_{табл} = \{\alpha = 0{,}05;\ \upsilon = 10 - 2 = 8\} = 2{,}306.$$

Так как $|t_p| > t_{табл}$, гипотеза H_0 о равенстве коэффициента корреляции нулю отвергается с вероятностью 95%, что свидетельствует о значимости данного коэффициента, а следовательно, и о подтверждении связи между финансовым оборотом и количеством обслуживаемых клиентов крупнейших российских туристических компаний.

Причинную обусловленность и корреляцию нельзя считать синонимами, так как последняя не дает объяснения, почему исследуемые переменные в двумерной совокупности связаны между собой. Следует иметь в виду, что корреляция бывает и без причинной обусловленности: корреляция представляет собой число, которое показывает на то, что большим значениям одного фактора соответствуют большие значения другого (или наоборот). Корреляция всего лишь указывает, что между этими величинами наблюдается определенное соответствие.

Наиболее распространенной причиной существования корреляции без причинной обусловленности является наличие некоторого скрытого третьего фактора, создающего впечатление, будто одна переменная выступает причиной другой переменной. Реально основной причиной для каждой из этих двух переменных является данная неизвестная третья переменная. Такая корреляция носит название ложной. Под ложной понимают высокую корреляцию, которая обеспечивается действием третьего фактора. Например, вы обнаружили высокую корреляцию между приемом на работу новых менеджеров и созданием новых проектов. Может быть, именно новые менеджеры являются причиной капиталовложений в новые проекты. Или, наоборот, создание новых проектов послужило причиной приема на работу новых менеджеров. В данном примере проявляется действие третьего фактора: высокой, рассчитанной на длительную перспективу потребности в обновлении продукции фирмы, которая и послужила причиной приема на работу новых менеджеров и создания новых проектов.

3.1.6. Задачи применения регрессионного анализа

Регрессионный анализ является заключительным этапом изучения взаимосвязи, так как основной его целью является моделирование ранее подтвержденной взаимосвязи между изучаемыми признаками. Целью моделирования является получение аналитического выражения связи в виде математической функции, которая позволяет вычислить значения одной переменой на основании другой.

Переменную, поведение которой прогнозируется, принято обозначать через y; переменную, которая используется для такого прогнозирования, принято обозначать x. Математическую функцию, описывающую зависимость результативного признака y от факторного x, называют **уравнением регрессии**, а параметры этой модели — **коэффициентами регрессии**. Зная уравнение регрессии, можно с достаточно высокой степенью надежности оценить значения зависимой переменной y при определенных заданных значениях x. Точность такой оценки будет тем выше, чем теснее будут группироваться точки вокруг определенной линии на диаграмме рассеяния, т. е. точность модели регрессии зависит от тесноты связи между результативным и факторным признаками.

Для моделирования связи между изучаемыми признаками исследователем могут быть использованы любые математические функции. При построении парной регрессии обычно используются следующие функции:

- линейная:

$$\bar{y}_x = a_0 + a_1 x; \tag{3.1.5}$$

- параболическая:

$$\bar{y}_x = a_0 + a_1 x + a_2 x^2; \tag{3.1.6}$$

- гиперболическая:

$$\bar{y}_x = a_0 + \frac{a_1}{x}; \tag{3.1.7}$$

- показательная:

$$\bar{y}_x = a_0 \cdot a_1^x; \qquad (3.1.8)$$

- степенная:

$$\bar{y}_x = a_0 \cdot x^{a_1}. \qquad (3.1.9)$$

Определить тип уравнения можно, исследуя зависимость графически, однако существуют более общие указания, позволяющие выявить уравнение связи, не прибегая к графическому изображению. Если результативный и факторный признаки возрастают одинаково, то это свидетельствует о том, что связь между ними линейная, а при обратной связи — гиперболическая. Если результативный признак увеличивается в арифметической прогрессии, а факторный значительно быстрее, то используется параболическая или степенная регрессия. Однако если при проведении исследования возможно ограничиться построением линейной модели, то следует использовать ее, так как математический аппарат линейных уравнений наиболее разработан, а сами модели достаточно просто интерпретируются. На практике большинство нелинейных форм зависимости первоначально пытаются привести к линейной форме с помощью процедур «линеаризации» функций.

3.1.7. Вычисление параметров линейной парной регрессии

Парная линейная регрессия позволяет получить аналитическое выражение связи между результативным и факторным признаками, описываемое уравнением прямой линии. Практическое ее значение состоит в том, что выделяется наиболее существенный фактор из всей совокупности показателей, максимальным образом влияющий на результативный признак.

Уравнение парной линейной регрессии имеет вид:

$$\bar{y}_x = a_0 + a_1 x, \qquad (3.1.10)$$

где $\bar{y}_x$ — среднее значение результативного признака y при определенном значении факторного признака x; a_0 — показывает усредненное влияние на результативный признак неучтенных в уравнении факторных признаков; a_1 — коэффициент регрессии, показывающий, насколько в среднем изменится значение результативного признака при изменении факторного признака на единицу собственного измерения.

Оценка параметров уравнений регрессии (a_0 и a_1) осуществляется методом наименьших квадратов, в основе которого лежит предположение о независимости наблюдений исследуемой совокупности и нахождении параметров модели, при которых минимизируется сумма квадратов отклонений эмпирических (фактических) значений результативного признака от теоретических, полученных по выбранному уравнению регрессии:

$$S = \sum (y - \bar{y}_x)^2 \to \min. \qquad (3.1.11)$$

Система нормальных уравнений для нахождения параметров линейной парной регрессии методом наименьших квадратов имеет следующий вид:

$$\begin{cases} na_0 + a_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \\ a_0 \sum_{i=1}^{n} x_i + a_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i \end{cases}, \qquad (3.1.12)$$

где n — объем исследуемой совокупности (число единиц наблюдения).

Решая систему нормальных уравнений, определяют параметры a_0 и a_1, которые подставляют в уравнение прямой (3.1.10). Рассчитанные по этому уравнению значения $\bar{y}_x$ называют **теоретическими** значениями результативного признака, полученными по уравнению связи. Для проверки правильности построенного уравнения регрессии вычисляют абсолютные отклонения (абсолютные остатки) эмпирических значений признака от теоретических, полученных по уравнению регрессии ($\varepsilon_i = y_i - \bar{y}_\xi$), сумма которых должна быть равна нулю:

$$\sum_{i=1}^{n} \varepsilon_i = 0. \qquad (3.1.13)$$

ПРИМЕР.

Быстродействие компьютеров, объединенных в сеть, при возникновении перегрузок, как правило, снижается. Естественно, чем больше загрузка компьютера X, тем большим должно быть время реакции Y. Под временем реакции понимается интервал с момента нажатия клавиши Enter до момента выдачи компьютером ответа на введенный вами запрос (табл. 3.1.7).

Предположим наличие линейной зависимости между рассматриваемыми показателями. Построим расчетную таблицу для определения параметров линейного уравнения регрессии времени реакции компьютера (табл. 3.1.8).

Таблица 3.1.7

Время реакции компьютера, с	Загрузка компьютера от полной его мощности, %
0,31	20,2
0,69	22,7
2,27	41,7
0,57	24,6
1,28	20,0
0,88	39,0
2,11	33,4
4,84	63,9
1,60	35,8
5,06	62,3

Таблица 3.1.8

Расчетная таблица для определения параметров уравнения регрессии

Время реакции компьютера y_i, с	Загрузка компьютера x_i, %	x_i^2	x_iy_i	$\bar{y}_x$	$\varepsilon_i = y_i - \bar{y}_x$	$(y_i - \bar{y}_x)^2$
0,31	20,2	408,04	6,26	0,38	–0,07	0,0049
0,69	22,7	515,29	15,66	0,62	0,07	0,0049
2,27	41,7	1738,89	94,66	2,48	–0,22	0,0484
0,57	24,6	605,16	14,02	0,81	–0,24	0,0576
1,28	20,0	400,00	25,60	0,36	0,91	0,8281
0,88	39,0	1521,00	34,32	2,22	–1,34	1,7956
2,11	33,4	1115,56	70,47	1,67	0,44	0,1936
4,84	63,9	4083,21	309,28	4,66	0,18	0,0324
1,60	35,8	1281,64	57,28	1,91	–0,31	0,0961
5,06	62,3	3881,29	315,24	4,50	0,56	0,3136
19,61	**363,6**	**15 550,08**	**942,79**	**19,61**	**0,00**	**3,3752**

Система нормальных уравнений для данного примера имеет вид:

$$\begin{cases}10a_0 + 363{,}6a_1 = 19{,}61\\ 363{,}6a_0 + 15550{,}08a_1 = 942{,}79\end{cases},$$

а ее решение есть

$$\begin{cases}a_0 = -1{,}602\\ a_1 = 0{,}098\end{cases}.$$

Подставляя полученные a_0 и a_1 в линейное уравнение $y = a_1x + a_0$, получим:

$$\bar{y}_x = -1{,}602 + 0{,}098x.$$

Значения $\bar{y}_x$ в табл. 3.1.8 получены путем пошаговой подстановки значений факторного признака x_i (загрузка компьютера) в уравнение регрессии $\bar{y}_x = 1{,}602 + 0{,}098x$.

Коэффициент регрессии $a_1 = 0{,}098$ означает, что при увеличении загрузки компьютерных мощностей на 1% время реакции компьютера повышается в среднем на 0,098 с. Данную оценку связи можно использовать для прогнозирования времени реакции компьютера при соответствующем проценте загрузки. Например, при загрузке компьютера на 50% время его реакции составит:

$$\bar{y}_{x=50} = -1{,}602 + 0{,}098 \cdot 50 = 3{,}3 \text{ с}.$$

Проведем проверку точности построенной модели регрессии, для чего вычислим абсолютные остатки и рассчитаем по ним сумму:

$$\sum_{i=1}^{n} \varepsilon_i = 0.$$

Так как сумма абсолютных отклонений фактических значений результативного признака от выровненных по уравнению связи равна нулю, а также сумма теоретических значений результативного признака полностью совпала с суммой эмпирических значений, то можно сделать вывод о том, что уравнение регрессии было построено правильно и адекватно отражает сложившуюся связь.

3.1.8. Вычисление параметров уравнения регрессии при нелинейной зависимости

Линейные связи являются основными при определении аналитического выражения связи. Однако встречаются и нелинейные связи, хорошо описываемые параболой, гиперболой и др. функциями.

Если связь между факторным и результативным признаком криволинейная и описывается уравнением параболы второго порядка вида:

$$\bar{y}_x = a_0 + a_1 x + a_2 x^2, \tag{3.1.14}$$

то система нормальных уравнений для нахождения параметров регрессии методом наименьших квадратов имеет вид:

$$\begin{cases} na_0 + a_1 \sum_{i=1}^{n} x_i + a_2 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} y_i \\ a_0 \sum_{i=1}^{n} x_i + a_1 \sum_{i=1}^{n} x_i^2 + a_2 \sum_{i=1}^{n} x_i^3 = \sum_{i=1}^{n} x_i y_i \\ a_0 \sum_{i=1}^{n} x_i^2 + a_1 \sum_{i=1}^{n} x_i^3 + a_2 \sum_{i=1}^{n} x_i^4 = \sum_{i=1}^{n} x_i^2 y_i \end{cases}. \tag{3.1.15}$$

После решения системы исследователь получит значения неизвестных коэффициентов a_0, a_1, a_2 и таким образом определит параболическое уравнение регрессии. Вычислив по уравнению регрессии теоретические значения результативного признака $\bar{y}_x$, необходимо сравнить их с фактическими данными. Таким образом расчитывается остаточная сумма квадратов отклонений. На практике считается, что минимальная сумма квадратов отклонений говорит об адекватно построенной модели регрессии и целесообразности ее использования на практике.

Оценка обратной зависимости между x и y, когда с увеличением (уменьшением) факторного признака уменьшается (увеличивается) результативный, осуществляется на основании гиперболы вида:

$$\bar{y}_x = a_0 + \frac{a_1}{x}. \tag{3.1.16}$$

Система нормальных уравнений для нахождения параметров гиперболы следующая:

$$\begin{cases} na_0 + a_1 \sum_{i=1}^{n} \frac{1}{x_i} = \sum_{i=1}^{n} y_i \\ a_0 \sum_{i=1}^{n} \frac{1}{x_i} + a_1 \sum_{i=1}^{n} \frac{1}{x_i^2} = \sum_{i=1}^{n} \frac{y_i}{x_i} \end{cases}. \tag{3.1.17}$$

ПРИМЕР.

Рассмотрим в качестве примера данные об объеме производства вишни и ее цене в фермерских хозяйствах региона (табл. 3.1.9).

Таблица 3.1.9

Данные об объеме производства и фермерской цене вишни хозяйств области

№ п/п	Объем производства x_i, тыс. т	Фермерская цена y_i, тыс. долл. за т	№ п/п	Объем производства x_i, тыс. т	Фермерская цена, y_i, тыс. долл. за т
1	0,204	0,267	9	0,276	0,163
2	0,260	0,174	10	0,404	0,262
3	0,287	0,140	11	0,331	0,126
4	0,239	0,208	12	0,373	0,165
5	0,208	0,225	13	0,200	0,299
6	0,218	0,243	14	0,198	0,325
7	0,394	0,227	15	0,356	0,150
8	0,266	0,166	16	0,384	0,188

Рис. 3.1.10. Диаграмма рассеяния, характеризующая нелинейную взаимосвязь между объемом производства вишни и ее фермерской ценой

Как правило, зависимость между производством и ценой товара нелинейна. Диаграмма рассеяния, представленная на рис. 3.1.10, иллюстрирует пример нелинейной взаимосвязи, представленной уравнением параболы второго порядка.

Построим расчетную таблицу для определения параметров нелинейного уравнения регрессии, выраженного параболической функцией (табл. 3.1.10).

Система нормальных уравнений для данного примера имеет вид (согласно формуле (3.1.15)):

$$\begin{cases} 16a_0 + 4{,}6a_1 + 1{,}41a_2 = 3{,}33 \\ 4{,}6a_0 + 1{,}41a_1 + 0{,}46a_2 = 0{,}93 \, , \\ 1{,}41a_0 + 0{,}46a_1 + 0{,}15a_2 = 0{,}28 \end{cases}$$

а ее решение:

$$\begin{cases} a_0 = 0{,}26 \\ a_1 = -0{,}04. \\ a_2 = -0{,}43 \end{cases}$$

Следовательно, по формуле (3.1.14):

$$\bar{y}_x = 0{,}26 - 0{,}04x - 0{,}43x^2.$$

Значения $\bar{y}_x$ в табл. 3.1.10 получены путем пошаговой подстановки значений факторного признака x_i (объем производства вишни) в уравнение регрессии $\bar{y}_x = 0{,}26 - 0{,}04x - 0{,}43x^2$. Коэффициент регрессии $a_1 = -0{,}04$ означает, что при увеличении объема производства вишни на 1 тыс. т фермерская цена будет в среднем снижаться на 0,04 тыс. долл.

Таблица 3.1.10

Расчетная таблица для определения параметров уравнения регрессии

Объем производства x_i, тыс. т	Фермерская цена y_i, тыс. долл. за т	x_i^2	x_i^3	x_i^4	x_iy_i	x^2y	$\bar{y}_x$
0,204	0,267	0,042	0,008	0,002	0,054	0,011	0,234
0,260	0,174	0,068	0,018	0,005	0,045	0,012	0,221
0,287	0,140	0,082	0,024	0,007	0,040	0,012	0,213
0,239	0,208	0,057	0,014	0,003	0,050	0,012	0,226
0,208	0,225	0,043	0,009	0,002	0,047	0,010	0,233
0,218	0,243	0,048	0,010	0,002	0,053	0,012	0,231
0,394	0,227	0,155	0,061	0,024	0,089	0,035	0,177
0,266	0,166	0,071	0,019	0,005	0,044	0,012	0,219
0,276	0,163	0,076	0,021	0,006	0,045	0,012	0,216
0,404	0,262	0,163	0,066	0,027	0,106	0,043	0,174
0,331	0,126	0,110	0,036	0,012	0,042	0,014	0,200
0,373	0,165	0,139	0,052	0,019	0,061	0,023	0,185
0,200	0,299	0,040	0,008	0,002	0,060	0,012	0,235
0,198	0,325	0,039	0,008	0,002	0,064	0,013	0,235
0,356	0,150	0,127	0,045	0,016	0,053	0,019	0,191
0,384	0,188	0,147	0,056	0,022	0,072	0,028	0,181
4,5979	**3,328**	**1,407**	**0,455**	**0,154**	**0,927**	**0,278**	**3,371**

Как уже ранее отмечалось, линейный коэффициент корреляции используется только для характеристики тесноты и направления связи при линейной зависимости. В случае наличия нелинейной зависимости между двумя признаками для измерения тесноты связи применяют **корреляционное отношение**. Данный показатель используется как при линейной, так и при криволинейных формах зависимости.

На практике различают теоретическое и эмпирическое корреляционное отношение. **Эмпирическое корреляционное отношение** используется, если расчет ведется по сгруппированным данным и определяется по формуле:

$$\eta_{эмп} = \sqrt{\frac{\delta^2_{\bar{y}}}{\sigma^2_y}}, \qquad (3.1.18)$$

где $\delta^2_{\bar{y}}$ — межгрупповая дисперсия, рассчитываемая по формуле:

$$\delta^2_{\bar{y}} = \frac{\sum(\bar{y}_j - \bar{y})^2 \cdot n_j}{\sum n_j}, \qquad (3.1.19)$$

где $\bar{y}_j$ — средние групповые значения результативного признака в j-й группе; $\bar{y}$ — общее среднее значение результативного признака y; n_j — число единиц в j-й группе ($n_j = \sum f_{ij}$); σ^2_y — общая дисперсия результативного признака y, рассчитываемая по формуле:

$$\sigma^2_y = \frac{\sum(y_i - \bar{y})^2}{\sum n_j}. \qquad (3.1.20)$$

В основе расчета корреляционного отношения лежит правило сложения дисперсий:

$$\sigma^2_y = \delta^2_{\bar{y}} + \overline{\sigma^2_i}. \qquad (3.1.21)$$

Подставим в приведенную формулу выражения для расчета межгрупповой и общей дисперсии. Получим следующую модификацию формулы для расчета эмпирического корреляционного отношения:

$$\eta_{эмп} = \sqrt{\frac{\sum(\bar{y}_j - \bar{y})^2 n_j}{\sum(y_i - \bar{y})^2}}. \qquad (3.1.22)$$

Эмпирическое корреляционное отношение показывает степень влияния на вариацию результативного признака y изменения показателя x, положенного в основание аналитической группировки. Квадрат эмпирического отношения называется **эмпирическим коэффициентом детерминации**, который характеризует долю вариации результативного признака, объясняемой за счет группировочного, факторного признака x.

Теоретическое корреляционное отношение рассчитывается по формуле:

$$\eta_{\text{теор}} = \sqrt{\frac{\delta^2}{\sigma_y^2}}, \tag{3.1.23}$$

где δ^2 — дисперсия теоретических значений, рассчитываемая по формуле:

$$\delta^2 = \frac{\sum(\bar{y}_x - \bar{y})^2}{n}. \tag{3.1.24}$$

Подставим вместо δ^2 выражение для ее расчета, получим следующую модификацию для расчета теоретического корреляционного отношения:

$$\eta_{\text{теор}} = \sqrt{\frac{\sum(\bar{y}_x - \bar{y})^2}{\sum(y_i - \bar{y})^2}}. \tag{3.1.25}$$

Квадрат теоретического корреляционного отношения называется **теоретическим коэффициентом детерминации**. Он показывает, какая доля вариации результативного признака может быть объяснена построенным уравнением регрессии.

Чем больше значение теоретического корреляционного отношения, тем точнее построенное уравнение регрессии описывает зависимость изучаемых признаков. При выборе наилучшего уравнения регрессии, когда форма зависимости неизвестна, необходимо учитывать данное свойство корреляционного отношения.

Корреляционное отношение изменяется в пределах от 0 до 1 $(0 < \eta < 1)$. Чем ближе значение корреляционного отношения к единице, тем связь признается теснее. Если зависимость между двумя признаками линейная, значение теоретического корреляционного отношения должно быть равно линейному коэффициенту корреляции. Интерпретация данного показателя аналогична интерпретации коэффициента корреляции.

Если рассматривается множественная линейная зависимость, то значение теоретического корреляционного отношения будет совпадать со значением множественного коэффициента корреляции, а его квадрат — теоретический коэффициент детерминации — с множественной интерпретацией коэффициента детерминации.

ПРИМЕР.

Определим, насколько хорошо полученное уравнение описывает взаимосвязь между объемом производства и фермерской ценой вишни (см. табл. 3.1.9). Все расчеты представим в виде вспомогательной табл. 3.1.11.

Таблица 3.1.11

Расчетная таблица для определения теоретического корреляционного отношения

y_i	$\bar{y}_x$	$(\bar{y}_x - \bar{y})$	$(\bar{y}_x - \bar{y})^2 \cdot 10^{-2}$	$(y_i - \bar{y})$	$(y_i - \bar{y})^2 \cdot 10^{-2}$
0,267	0,234	0,026	0,067	0,059	0,348
0,174	0,221	0,013	0,016	–0,034	0,116
0,140	0,213	0,005	0,003	–0,068	0,462
0,208	0,226	0,018	0,032	0,000	0,000
0,225	0,233	0,025	0,063	0,017	0,029
0,243	0,231	0,023	0,052	0,035	0,123
0,227	0,177	–0,030	0,093	0,019	0,036
0,166	0,219	0,011	0,012	–0,042	0,176
0,163	0,216	0,008	0,006	–0,045	0,203
0,262	0,174	–0,030	0,119	0,054	0,292
0,126	0,200	–0,010	0,010	–0,082	0,672
0,165	0,185	–0,020	0,051	–0,043	0,185
0,299	0,235	0,027	0,072	0,091	0,828
0,325	0,235	0,027	0,074	0,117	1,369
0,150	0,191	–0,020	0,028	–0,058	0,336
0,188	0,181	–0,030	0,071	–0,020	0,040
Итого		–	**0,766**	–	**5,215**

Теоретическое корреляционное отношение согласно формуле (3.1.25) $\eta^2_{теор}$ составит:

$$\eta_{теор} = \sqrt{\frac{0{,}766}{5{,}215}} = 0{,}38,\ \eta^2_{теор} = 0{,}38^2 = 0{,}14.$$

Таким образом, связь между объемом производства вишни и фермерской ценой на нее можно признать умеренной. Невысокое значение теоретического корреляционного отношения в данном случае оправданно, поскольку кроме объема производства на устанавливаемую фермерами цену оказывают сильное влияние другие факторы (уровень производственных издержек, конъюнктура спроса, цены конкурентов).

Построенное параболическое уравнение описывает вариацию результативного признака на 14% и, следовательно, не может быть использовано для дальнейшего моделирования и прогнозирования.

3.1.9. Принятие решений на основе уравнений регрессии

Полная экономическая интерпретация моделей регрессии позволяет выявить резервы развития и повышения деловой активности субъектов рыночной экономики. Любая интерпретация начинается со статистической оценки уравнения регрессии в целом и оценки значимости входящих в модель факторных признаков.

Прежде всего, необходимо рассмотреть коэффициенты регрессии. Чем больше величина коэффициента регрессии, тем значительнее влияние данного признака на моделируемый.

Знаки коэффициентов регрессии говорят о характере влияния на результативный признак. Если факторный признак имеет знак плюс, то с увеличением данного фактора результативный признак возрастает; если факторный признак имеет знак минус, то с его увеличением результативный признак уменьшается.

Если экономические процессы подсказывают, что факторный признак должен иметь положительное значение, а он имеет знак минус, то необходимо проверить расчеты параметров уравнения регрессии. Такое явление чаще всего бывает в силу допущенных в расчетах ошибок.

Проверка значимости каждого коэффициента регрессии осуществляется с помощью *t-критерия Стьюдента*, расчетное значение которого определяется по формуле:

$$t_{\text{расч}} = \frac{a_i}{\sqrt{\sigma^2_{a_i}}}, \qquad (3.1.26)$$

где $\sigma^2_{a_i}$ — дисперсия коэффициента регрессии, наиболее простая формула которого имеет вид:

$$\sigma^2_{a_i} = \frac{\sigma^2_y}{k}, \qquad (3.1.27)$$

где σ^2_y — дисперсия результативного признака y; k — число факторных признаков в уравнении связи.

При проверке значимости коэффициента регрессии на основе линейной парной зависимости дисперсия результативного признака может быть вычислена как

$$\sigma^2_{a_1} = \frac{\sigma^2_{ост}}{\sigma^2_x(n-2)}, \qquad (3.1.28)$$

где σ^2_x — дисперсия факторного признака x; $\sigma^2_{ост}$ — остаточная дисперсия, определяемая следующим образом:

$$\sigma^2_{ост} = \frac{\sum(y_i - \bar{y}_x)^2}{n}. \qquad (3.1.29)$$

Расчетное значение t-критерия сравнивается с его критическим, которое определяется по таблице табулированных значений:

$$t_{табл} = \{\alpha;\ \upsilon = n - \kappa - 1\}, \qquad (3.1.30)$$

где α — уровень значимости критерия проверки гипотезы о равенстве нулю параметров уравнения регрессии; υ — число степеней свободы, которое характеризует количество свободно варьируемых элементов совокупности; κ — количество параметров в уравнении регрессии.

Если расчетное значение t-критерия по модулю превышает табличное, то коэффициент регрессии признается значимым.

ПРИМЕР.

Определим значимость коэффициента регрессии, характеризующего изменение времени реакции компьютера от процента его загрузки (см. табл. 3.1.8), на основании следующих вычислений.

1. Определим остаточную дисперсию:

$$\sigma^2_{ост} = \frac{\sum(y_i - \bar{y}_x)^2}{n} = \frac{3{,}3752}{10} = 0{,}3375.$$

2. Рассчитаем дисперсию факторного признака:

$$\sigma^2_x = \frac{\sum x_i^2}{n} - (\bar{x})^2 = \frac{15550{,}08}{10} - (36{,}36)^2 = 232{,}96.$$

3. Подставив полученные показатели в формулу дисперсии регрессионного коэффициента a_1, получим:

$$\sigma^2_{a_1} = \frac{\sigma^2_{ост}}{\sigma^2_x(n-2)} = \frac{0{,}3375}{232{,}9584 \cdot (10-2)} = 0{,}0002.$$

4. Расчетное значение t-критерия тогда составит:

$$t_{расч} = \frac{a_i}{\sqrt{\sigma^2_{a_i}}} = \frac{0{,}098}{\sqrt{0{,}0002}} = 2{,}18.$$

5. Критическое значение t-критерия при $t_{табл}\{\alpha = 0{,}05;\ \upsilon = 8\} = 2{,}306$.

Так как $|t_{расч}| < t_{табл}$, следовательно, коэффициент регрессии признается незначимым. Таким образом, можно сделать вывод о том, что выбранная объясняющая переменная – процент загрузки компьютера – систематически не влияет на время реакции компьютера. В этом случае речь может идти о некотором скрытом факторе, который не был учтен при построении уравнения регрессии.

Проверка адекватности всей модели осуществляется с помощью *F-критерия Фишера—Снедекора*, расчетное значение которого при парной линейной регрессии определяется по формуле:

$$F_{расч} = \frac{r^2_{xy}}{1-r^2_{xy}} \cdot (n-2). \qquad (3.1.31)$$

Критическое значение F-критерия определяется по таблице F-распределения Фишера для $F_{табл} = \{\alpha;\ \upsilon_1 = к - 1;\ \upsilon_2 = n - к\}$. Если $F_{расч} > F_{табл}$, то гипотеза H_0 о несоответствии заложенных в уравнении регрессии связей реально существующим отвергается.

Для расширения возможностей экономического анализа используются **коэффициенты эластичности**, определяемые по формуле:

$$Э = a_1 \cdot \frac{\bar{x}}{\bar{y}}, \qquad (3.1.32)$$

где $\bar{x}$ — среднее значение соответствующего факторного признака; $\bar{y}$ — среднее значение результативного признака; a_1 — коэффициент регрессии при соответствующем факторном признаке.

Коэффициент эластичности показывает, на сколько процентов в среднем изменится значение результативного признака при изменении факторного признака на 1%.

ПРИМЕР.

Определим коэффициент эластичности по данным из предыдущего примера.

$$\bar{x} = \frac{363{,}6}{10} = 36{,}36, \ \bar{y} = \frac{19{,}61}{10} = 1{,}961, \ a_1 = 0{,}098;$$

$$Э = 0{,}098 \cdot \frac{36{,}36}{1{,}961} = 1{,}82.$$

Таким образом, при увеличении загрузки мощности компьютера на 1% время реакции соответственно в среднем увеличится на 1,82%.

Между линейным коэффициентом корреляции и коэффициентом регрессии существует определенная зависимость, выражаемая формулой:

$$r = a_i \frac{\sigma_x}{\sigma_y}, \qquad (3.1.33)$$

где a_i — коэффициент регрессии в уравнении связи; σ_x — среднее квадратическое отклонение факторного признака; σ_y — среднее квадратическое отклонение результативного признака.

3.1.10. Проведение корреляционно-регрессионного анализа по сгруппированным данным

На практике исследования часто проводятся по большому числу наблюдений. В этом случае исходные данные удобнее представлять в сводной *корреляционной таблице* (табл. 3.1.12).

Анализу подвергаются сгруппированные данные и по факторному x, и по результативному y признакам. Число групп по признакам может быть как равным, так и неравным. Если наибольшие значения частот каждой строки и каждого столбца располагаются на первой диагонали, связь является прямой и близкой к линейной; если наибольшие значения частот располагаются вдоль второй диагонали — связь обратная линейная. Если частоты во всех клетках таблицы примерно равны — связи нет. В табл. 3.1.12, кроме частот, приведены строки и столбцы для расчета необходимых сумм при вычислении коэффициента корреляции.

Как и средние величины признаков, так и все суммы, входящие в расчет линейного коэффициента корреляции, при группировке умножаются на соответствующие частоты. Следовательно, формула примет следующий вид:

$$r_{xy} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{m}(x'_i - \bar{x})(y'_j - \bar{y})f_j}{\sqrt{\sum_{i=1}^{k}(x'_i - \bar{x})^2 f_i \cdot \sum_{j=1}^{m}(y'_j - \bar{y})^2 f_j}}, \qquad (3.1.34)$$

где x'_i, y'_i — середины интервалов i-й категории x и j-й категории y; f_i — частота i-го значения x; f_j — частота j-го значения y; f_{ij} — частота совместного появления i-го значения x и j-го значения y (числа в клетках корреляционной таблицы); k — число интервалов группировки двумерной совокупности по факторному признаку; m — число интервалов группировки двумерной совокупности по результативному признаку.

Таблица 3.1.12

Макет сводной корреляционной таблицы

Середины интервалов, x_i	Середины интервалов, y_j						Сумма частот	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	$(x_i - \bar{x})^2 n_i$
	y_1	y_2	...	y_j	...	y_m				
x_1	n_{11}	n_{12}	...	n_{1j}	...	n_{1m}	n_1			
x_2	n_{21}	n_{22}	...	n_{2j}	...	n_{2m}	n_2			
...	...	...	...	...	..	...				
x_i	n_{i1}	n_{i2}	...	n_{ij}	...	n_{im}	n_i			
...	...	...	...	...	...	...	...			
x_k	n_{k1}	n_{k2}	...	n_{kj}	...	n_{km}	n_k			
Сумма частот	l_1	l_2	...	l_j	...	l_m	n			
$y_j - \bar{y}$										
$(y_j - \bar{y})^2$										
$(y_j - \bar{y})^2 l_j$										

Примечание: x_i – середина i-го интервала группировки по соответствующему факторному признаку; y_j – середина j-го интервала группировки по соответствующему результативному признаку; n_i – частота i-го интервала группировки по факторному признаку x, $i = 1, 2, ..., k$; l_j – частота j-го интервала группировки по результативному признаку y, $j = 1, 2, ..., m$; k – число интервалов группировки по факторному признаку; m – число интервалов группировки по результативному признаку; n_{ij} – частоты прямоугольников группировки; n – объем двумерной выборки.

ПРИМЕР.

На металлообрабатывающем заводе по 60 маркам стали провели замеры предела текучести (x, кг/мм^2) и предела прочности (y, кг/мм^2). В итоге получили 60 пар значений, представленных в табл. 3.1.13.

Таблица 3.1.13

№ п/п	Предел прочности, кг/мм^2	Предел текучести, кг/мм^2	№ п/п	Предел прочности, кг/мм^2	Предел текучести, кг/мм^2	№ п/п	Предел прочности, кг/мм^2	Предел текучести, кг/мм^2
1	178	154	21	161	145	41	101	87
2	164	133	22	107	94	42	139	88
3	75	58	23	141	113	43	98	83
4	65	51	24	97	86	44	111	106
5	114	101	25	127	121	45	104	92
6	209	169	26	138	119	46	103	85
7	140	98	27	125	112	47	118	112
8	115	97	28	97	85	48	102	98
9	101	105	29	72	41	49	108	103
10	69	44	30	113	96	50	119	99
11	116	92	31	88	45	51	128	104
12	157	141	32	109	99	52	118	107
13	93	71	33	76	73	53	193	155
14	69	39	34	85	77	54	155	136
15	147	122	35	61	47	55	81	82
16	52	33	36	85	68	56	163	136
17	117	78	37	142	137	57	79	72
18	138	114	38	123	113	58	81	66
19	149	125	39	85	42	59	61	42
20	147	133	40	179	153	60	97	85

Воспользовавшись правилами научной группировки, систематизируем полученные числовые характеристики и представим марки стали в виде группировочной табл. 3.1.14. Предположим, что большие значения предела текучести стали обусловливают большие значения прочности металла. Марки стали с низким пределом текучести имеют и низкий предел прочности. Для обоснования данного предположения о тесной положительной корреляции между пределом прочности и пределом текучести произведем расчет линейного коэффициента корреляции по сгруппированным данным.

Таблица 3.1.14

Расчетная таблица для определения коэффициента корреляции

Предел прочности, y, кг/мм^2	x_i / y_i	Предел текучести, кг/мм^2, x							f_j	$y'_i - \bar{y}$	$(y'_j - \bar{y})^2$	$(y'_j - \bar{y})^2 f_j$
		30–50	50–70	70–90	90–110	110–130	130–150	150–170				
		40	60	80	100	120	140	160				
50–70	60	5							5	–54,7	2992,1	14 960,5
70–90	80	3	3	4					10	–34,7	1204,1	12 041,0
90–110	100		1	7	6				14	–14,7	216,1	3025,4
110–130	120			1	8	4			13	5,3	28,1	365,3
130–150	140			1	1	5	2		9	25,3	640,1	5760,9
150–170	160						5		5	45,3	2052,1	10 260,5
170–190	180							2	2	65,3	4264,1	8528,2
190–210	200							2	2	85,3	7276,1	14 552,2
f_i	–	8	4	13	15	9	7	4	60	–	–	**69 494,0**
$x'_i - \bar{x}$		–56,7	–36,7	–16,7	3,3	23,3	43,3	63,3		–	–	–
$(x'_i - \bar{x})^2$		3214,9	1346,9	278,9	10,9	542,9	1874,9	4006,9		–	–	–
$(x'_i - \bar{x})^2 f_i$		25 719,2	5387,6	3625,7	163,5	4931,1	13 124,3	16 027,6	**68 979**	–	–	**64 364,7**

Для расчета линейного коэффициента корреляции по формуле (3.1.34) необходимо проделать следующие действия:

1. Определить среднюю величину прочности и среднюю величину текучести стали, воспользовавшись формулой средней арифметической взвешенной:

$$\bar{y} = \frac{\sum_{j=1}^{n} y_j f_j}{\sum_{j=1}^{n} f_j} = \frac{60 \cdot 5 + 80 \cdot 10 + 100 \cdot 14 + 120 \cdot 13 + 140 \cdot 9 + 160 \cdot 5 + 180 \cdot 2 + 200 \cdot 2}{60} = 114{,}7;$$

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i f_i}{\sum_{i=1}^{n} f_i} = \frac{40 \cdot 8 + 60 \cdot 4 + 80 \cdot 13 + 100 \cdot 15 + 120 \cdot 9 + 140 \cdot 7 + 160 \cdot 4}{60} = 96{,}7.$$

2. Вычислить числитель линейного коэффициента корреляции, для чего необходимо умножить отклонения по обоим признакам (с учетом их знаков) на частоты совместного распределения и сложить все произведения (взвешенные суммы квадратов отклонений подсчитаны и приведены в последнем столбце и последней строке табл. 3.1.14):

$$\sum_{i=1}^{k_1}\sum_{j=1}^{k_2}(x_i' - \bar{x})(y_j' - \bar{y})f_{ij} = (-56{,}7) \cdot (-54{,}7) \cdot 5 + (56{,}7) \cdot (-34{,}7) \cdot 3 + (-36{,}7) \cdot (-34{,}7) \cdot 3 +$$

$$+ (-16{,}7) \cdot (-34{,}7) \cdot 4 + (-36{,}7) \cdot (-14{,}7) \cdot 1 + (-16{,}7) \cdot (-14{,}7) \cdot 7 + 3{,}3 \cdot (-14{,}7) \cdot 6 + (-16{,}7) \cdot 5{,}3 \cdot 1 +$$

$$+ 3{,}3 \cdot 5{,}3 \cdot 8 + 23{,}3 \cdot 5{,}3 \cdot 4 + 16{,}7\ 25{,}3 \cdot 1 + 3{,}3 \cdot 25{,}3 \cdot 1 + 23{,}3 \cdot 25{,}3 \cdot 5 + 43{,}3 \cdot 25{,}3 \cdot 2 + 43{,}3 \cdot 45{,}3 \cdot 5 +$$

$$+ 63{,}3 \cdot 65{,}3 \cdot 2 + 63{,}3 \cdot 85{,}3 \cdot 2 = 64\,364{,}7.$$

Это число записано в правом нижнем углу табл. 3.1.14. Тогда линейный коэффициент корреляции согласно формуле (3.1.34) составит:

$$r_{xy} = \frac{64364{,}7}{\sqrt{68979 \cdot 69469}} = 0{,}93.$$

Таким образом, коэффициент корреляции подтверждает сделанное ранее предположение о том, что зависимость между пределом прочности и пределом текучести марок стали существует, причем по тесноте связь очень сильная, а по направлению прямая, так как коэффициент корреляции принял положительное значение.

Средние величины признаков и все суммы, входящие в расчет параметров регрессии по сгруппированным данным, также необходимо взвешивать на соответствующие частоты. Следовательно, параметры уравнения регрессии вычисляются по формулам:

$$a_1 = \frac{\sum_{i=1}^{k}\sum_{j=1}^{m}(x_i - \bar{x})(y_j - \bar{y})f_{ij}}{\sum_{i=1}^{k}(x_i - \bar{x})^2 f_i},$$

$$a_0 = \bar{y} - a_1 \cdot \bar{x}. \qquad (3.1.35)$$

ПРИМЕР.

Рассчитаем параметры линейной парной регрессии по сгруппированным данным предыдущего примера.

$$a_1 = \frac{61488}{64013{,}6} = 0{,}96,\ a_0 = 114 - 0{,}96 \cdot 95{,}8 = 22{,}03.$$

Уравнение регрессии имеет вид:

$$\bar{y}_x = 22{,}03 + 0{,}96x.$$

Следовательно, при увеличении предела текучести стали на 1 кг/мм^2 предел прочности в среднем возрастет на 0,96 кг/мм^2.

КОНТРОЛЬНЫЕ ВОПРОСЫ

1. В чем различия между стохастической и корреляционной связью?
2. В каких случаях можно говорить о прямой, а в каких об обратной зависимости между изучаемыми признаками?
3. Что такое диаграмма рассеяния? Возможно ли ее построение при: а) нелинейной связи; б) если взаимосвязь между факторным и результативным признаками отсутствует?
4. Существуют ли отличия между корреляционным и регрессионным анализом, или они преследуют одинаковые цели?
5. Что можно сказать о связи, если линейный коэффициент корреляции будет равен:
   а) нулю;
   б) единице;
   в) –0,15?

6. Если вы выявили сильную положительную корреляцию, говорит ли это о том, что большие значения факторного признака вызывают появление больших значений результативного? Если нет, какие еще возможны варианты?
7. В чем состоят основные отличия проведения корреляционно-регрессионного анализа по сгруппированным и несгруппированным данным?
8. Какими рекомендациями следует пользоваться при определении типа уравнения регрессии? Если результативный и факторный признаки имеют тенденцию к стабильному росту, то какой аналитической функцией можно описывать из взаимосвязь?
9. Важно ли при анализе коэффициентов регрессии уделять внимание знакам, стоящим перед ними? Если да, то почему?
10. Что характеризует коэффициент регрессии в уравнении связи?

Глава 3.2

МНОГОФАКТОРНЫЕ МОДЕЛИ АНАЛИЗА ВЗАИМОСВЯЗИ

- Определение множественного коэффициента корреляции и коэффициента детерминации
- Построение модели множественной регрессии
- Интерпретация результатов множественной регрессии
- Применение корреляционно-регрессионного анализа в исследовании финансово-экономической деятельности

3.2.1. Определение множественного коэффициента корреляции и коэффициента детерминации

Достаточно часто при решении реальных экономических задач необходимо оценить степень влияния на результативный показатель не одного фактора, а некоторой совокупности факторов. Таким образом, исследователь сталкивается с множественным корреляционно-регрессионным анализом, основными целями которого являются:

- описание и понимание взаимосвязи;
- прогнозирование нового наблюдения на основании имеющейся совокупности данных;
- регулирование и управление бизнес-процессами.

Предположим, ваша фирма участвует в тендере на реализацию нового проекта. Для тех проектов, конкурс на которые вам удалось выиграть, вы располагаете данными, касающимися фактических затрат (y), оценки прямых трудозатрат (x_1), оценки затрат на материалы (x_2) и затрат на управленческие функции (x_3). Допустим, что предложение цены, с которой вы выходите на конкурс, кажется вам неоправданно низким. Определив взаимосвязь между фактическими затратами и оценками, сделанными ранее, на этапе переговоров при заключении контрактов, вы сможете выяснить, какие из оценок вы систематически занижаете, или, наоборот, завышаете. На этом этапе анализа как раз и происходит описание и понимание исследуемой вззаимосвязи.

Основной причиной построения прогнозных значений исследуемой совокупности выступает необходимость принятия правильных управленческих решений, которые должны приниматься на основе тщательного анализа исходной информации. Анализ прогнозов состояния рынка необходим для оценивания возможных рисков. В противном случае вас могут опередить конкуренты, умеющие лучше оценивать и прогнозировать перспективы развития рынка. Например, для успешного планирования производства необходимо заложить дополнительные расходы на сезон повышенного спроса на продукцию вашей фирмы. Анализ структуры затрат (y) целесообразно проводить с применением модели множественной регрессии, в которой x_1 — количество выпускаемых изделий, x_2 — количество сотрудников, x_3 — объем сверхурочных работ. Результаты анализа будут полезны для принятия управленческих решений. Анализ, при котором исследователь имеет дело с одним результативным и множеством факторных признаков, в статистике называется «многомерным анализом». Многомерный анализ поможет выявить скрытые расходы, которые обнаруживают тенденцию к возрастанию с ростом объемов сверхурочных работ, и делать достаточно точные прогнозы фактических затрат на основе располагаемой вами информации.

Регулирование и управление бизнес-процессами является заключительным этапом многомерного корреляционно-регрессионного анализа. Например, страхование на рынке ценных бумаг подразумевает формирование портфеля ценных бумаг, которые в максимальной степени учитывают риски тех или иных активов. Если, предположим, вы храните определенный запас товарно-материальных ценностей, следует обеспечить страхование данного риска. Банки используют контракты на казначейские фьючерсы и опционы для страхования риска потерь в результате изменения процентных ставок по их депозитным счетам и ссудам. Процесс выбора оптимального страхового портфеля осуществляется с помощью многофакторного анализа. Взяв за основу данные прошедшего периода, можно попытаться объяснить движение цен на активы (y) изменениями курса ценных бумаг (x_1, x_2, ..., x_n). Соответствующие коэффициенты регрессии покажут, какой процент ценных бумаг того или иного вида следует включить в страховой портфель, чтобы минимизировать долю риска. Таким образом, множественная регрессия будет использоваться для регулирования и управления риском, которому подвержены активы, который и выступает в данном примере бизнес-процессом.

Основными задачами **множественного корреляционного анализа** являются определение наличия связи между отобранными признаками, установление ее направления и количественная оценка тесноты связи. Для этого сначала оценивается матрица парных коэффициентов корреляции, затем на ее основе определяются частные и множественные коэффициенты корреляции и детерминации. После нахождения значений коэффициентов корреляции следует проверить их значимость. В результате корреляционного анализа исследователь должен отобрать значимые факторные признаки для дальнейшего построения уравнения регрессии, позволяющего количественно описать и прогнозировать взаимосвязь.

На практике множественный корреляционно-регрессионный анализ проводится с использованием специализированных программных продуктов, основными из которых являются SPSS и STATISTICA. Этапы реализации многофакторного анализа в практической деятельности будут подробно рассмотрены в разд. 3.2.4.

Исходный массив данных для анализа можно записать в матричной форме как:

- вектор-столбец со значениями результативного признака размерности $n \times 1$:

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ \dots \\ y_n \end{pmatrix},$$

где i — номер наблюдения ($i = 1, 2, ..., n$); n — количество наблюдений;

- матрица со значениями факторных признаков размерности $k \times n$:

$$X = \begin{pmatrix} x_{11} & x_{21} & x_{1j} & \dots & x_{1k} \\ x_{21} & x_{22} & x_{2j} & \dots & x_{1k} \\ x_{i1} & x_{i2} & x_{ij} & \dots & x_{jk} \\ \dots & \dots & \dots & \dots & \dots \\ x_{n1} & x_{n2} & x_{nj} & \dots & x_{nk} \end{pmatrix},$$

где $j = 1, 2, ..., k$ — номер факторного признака; $i = 1, 2, ..., n$ — номер наблюдения.

Корреляционный анализ начинается с определения парных (линейных) коэффициентов корреляции, этапы расчета которых были рассмотрены в гл. 3.1. Парные коэффициенты корреляции рассчитываются как для определения связи между результативным и факторным признаками, так и между самими факторными признаками. При наличии между ними существенной связи можно сделать вывод о присутствии **мультиколлинеарности** — достаточно тесной зависимости между объясняющими x-переменными). Подтверждение мультиколлинеарности осуществляется, если парный коэффициент корреляции по абсолютной величине превышает 0,8. Явление мультиколлинеарности будет более подробно рассмотрено в п. 3.2.2. В дальнейшем при построении регрессионной модели один из взаимосвязанных факторов должен быть удален. Рекомендуется удалять тот фактор, который оказывает наименьшее влияние на зависимую переменную, т. е. имеет минимальное значение величины парного коэффициента корреляции с результативным признаком по сравнению с другими факторами.

Матрицу парных коэффициентов корреляции в общем виде можно представить так:

$$R = \begin{pmatrix} 1 & r_{yx_1} & r_{yx_2} & \dots & r_{yx_k} \\ r_{x_1 y} & 1 & r_{x_1 x_2} & \dots & r_{x_1 x_k} \\ x_{x_2 y} & r_{x_2 x_1} & 1 & \dots & r_{x_2 x_k} \\ \dots & \dots & \dots & \dots & \dots \\ x_{x_k y} & r_{x_k x_1} & r_{x_k x_2} & \dots & 1 \end{pmatrix}.$$

Каждый элемент корреляционной матрицы представляет собой парный коэффициент корреляции между соответствующими показателями. Матрица R симметрична (например, коэффициент корреляции между y и x_1 такой же, как и между x_1 и y, т. е. $r_{yx_1} = r_{x_1 y}$), на ее главной диагонали стоят единицы (связь признака с самим собой составляет единицу). Если, например, $r_{yx_1} = 0{,}80$; $r_{yx_2} = 0{,}57$; $r_{x_1 y_2} = 0{,}87$, то в регрессионное уравнение следует включить фактор x_1, а фактор x_2 не включать, так как он, во-первых, тесно связан (коллинеарен) с x_1, во-вторых, его корреляция с y слабее, чем корреляция фактора x_1. При включении в анализ факторов, имеющих функциональную связь между собой, исследователь на выходе получит вырожденную матрицу коэффициентов корреляции.

На основе корреляционной матрицы вычисляются общие показатели тесноты связи всех входящих в уравнение регрессии факторов с результативным признаком: множественный коэффициент детерминации и множественный коэффициент корреляции.

Множественный коэффициент корреляции характеризует тесноту линейной связи между одной переменной и совокупностью других переменных, рассматриваемых в корреляционном анализе. Наибольшее значение имеет расчет множественного коэффициента корреляции результативного признака с совокупностью факторных признаков, определяемый по формуле:

$$R_{y/x_1x_2,\dots,x_k} = \sqrt{1-\frac{|R|}{R_{yy}}}, \qquad (3.2.1)$$

где $|R|$ — определитель корреляционной матрицы; R_{yy} — алгебраическое дополнение r_{yy}-го элемента корреляционной матрицы R.

Если исследователь работает только с двумя факторными признаками (x_1 и x_2), то формула примет вид:

$$R_{y/x_1x_2} = \sqrt{1-\frac{|R|}{R_{yy}}} = \sqrt{\frac{r_{x_1y}^2 + r_{x_2y}^2 - 2r_{x_1x_2}r_{x_1y}r_{x_2y}}{1-r_{x_1x_2}^2}}, \qquad (3.2.2)$$

где r — парные коэффициенты корреляции.

Множественный коэффициент детерминации представляет собой квадрат множественного коэффициента корреляции и характеризует долю дисперсии результативного признака y, объясняемую за счет факторных признаков x. При увеличении числа факторных признаков коэффициент детерминации увеличивается, а при их исключении — уменьшается. Решение о включении в модель той либо иной новой переменной зависит от того, насколько она увеличивает коэффициент детерминации.

Множественный коэффициент детерминации и множественный коэффициент корреляции изменяются в пределах [0; 1]. Чем ближе значение коэффициента к единице, тем связь сильнее и тем точнее уравнение регрессии будет описывать зависимость y от x переменных. Если значение множественного коэффициента корреляции не превышает 0,3, это говорит о том, что выбранный набор факторных признаков достаточно слабо описывает вариацию результативного признака либо связь является криволинейной.

Значимость множественного коэффициента детерминации проверяется на основании F-критерия Фишера, расчетное значение которого рассчитывается по формуле:

$$F_{\text{расч}} = \frac{\frac{1}{k-1}R^2_{y/x_1x_2,\dots,x_k}}{\frac{1}{n-k}(1-R^2_{y/x_1x_2,\dots,x_k})}. \qquad (3.2.3)$$

Критическое значение F-критерия определяется по таблице F-распределения Фишера для $F_{табл} = \{\alpha;\ \upsilon_1 = к - 1;\ \upsilon_2 = n - к\}$. Если $F_{расч} > F_{табл}$, то гипотеза H_0 о равенстве множественного коэффициента детерминации нулю отвергается, следовательно, он признается значимым.

Расчет множественных коэффициентов корреляции и детерминации с использованием SPSS, а также их проверка на значимость будут подробно рассмотрены в п. 3.2.4 настоящей главы.

3.2.2. Построение модели множественной регрессии

Многофакторный регрессионный анализ является статистическим методом исследования зависимости случайной величины Y от переменных X_j, где $j = 1, 2, ..., k$. При этом зависимая переменная имеет нормальный закон распределения, а факторные переменные рассматриваются как неслучайные величины, независимо от действительного закона распределения X_j. Практически любой результативный показатель испытывает воздействие не одного, как в случае парной корреляции, а нескольких факторов. Таким образом, наиболее часто в экономическом анализе строят следующие модели множественной регрессии:

- линейная:

$$\bar{y}_x = a_0 + a_1 x + a_2 x_2 + \ldots + a_k x_k; \qquad (3.2.4)$$

- степенная:

$$\bar{y}_x = a_0 x_1^{a_1} + x_2^{a_2} + x_3^{a_3} + \ldots + x_k^{a_k}; \qquad (3.2.5)$$

- параболическая:

$$\bar{y}_x = a_0 + a_1 x_1^2 + a_2 x_2^2 + \ldots + a_k x_k^2, \qquad (3.2.6)$$

где a_0 — свободный параметр уравнения регрессии; a_j ($j = 1, 2, ..., k$) — коэффициенты регрессии, при соответствующих факторных признаках.

Для выбора оптимального регрессионного уравнения необходимо знать условный закон распределения результативной переменной y. Если графический метод не позволяет его выявить, то строят несколько регрессионных моделей, а затем выбирают наилучшую. Как уже было сказано ранее, если исследователь может ограничиться линейной моделью, то именно ее и следует использовать для анализа, поскольку ее результаты проще интерпретировать.

Если же исходные данные представлены нелинейной формой зависимости, то необходимо попытаться привести их с помощью процедуры «линеаризации» к линейной функции. Например, логарифмируя степенные уравнения или заменяя в гиперболическом уравнении величину $\frac{1}{x}$ новой переменной, переходят к линейным формам зависимости.

Линейная форма зависимости может быть выбрана в том случае, если результативный признак y_i и факторные признаки x_j в совокупности образуют $(k + 1)$-мерную случайную величину, подчиняющуюся нормальному закону распределения. При этом результативный и факторные признаки должны иметь тенденцию к возрастанию или убыванию приблизительно в арифметической прогрессии. Если форму связи изучаемых признаков установить проблематично, также прибегают к линейной форме зависимости, которая в общем виде записана в формуле (3.2.4).

Важным этапом построения уже выбранного уравнения множественной регрессии является отбор наиболее информативных переменных, оказывающих максимальное влияние на результативный признак, и включение их в модель.

Вопрос о количестве переменных в большинстве случаев решается самим исследователем. При этом желательно прийти к оптимальному решению: с одной стороны, чем больше факторных признаков включено в уравнение, тем оно лучше описывает явление, с другой — модель размерностью более 100 факторных признаков сложно реализуема. Сокращение размерности модели за счет исключения второстепенных, экономически и статистически несущественных факторов упрощает построение и повышает качество ее реализации. В то же время модель регрессии слишком малой размерности может недостаточно адекватно описывать взаимосвязь между исследуемыми явлениями. На практике считается, что число наблюдений должно минимум в 3–4 раза превышать количество факторных признаков.

Сущность метода пошаговой регрессии заключается в реализации алгоритмов последовательного «включения-исключения» факторов в уравнение регрессии и последующей проверке их статистической значимости. Алгоритм «включения» заключается в том, что факторы поочередно вводятся в уравнение так называемым «прямым методом». При проверке значимости введенного фактора определяется, насколько уменьшается сумма квадратов остатков и увеличивается величина множественного коэффициента корреляции R^2. Одновременно используется и алгоритм последовательного «исключения», сущность которого заключается в том, что исключаются факторы, ставшие незначимыми по статистическим критериям.

Фактор является незначимым, если его включение в уравнение регрессии только изменяет значения коэффициентов регрессии, не уменьшая суммы квадратов остатков и не увеличивая их значения. Если при включении в модель соответствующего факторного признака величина множественного коэффициента корреляции увеличивается, а величина коэффициента регрессии не изменяется (или меняется несущественно), то данный признак существенен и его включение в уравнение регрессии необходимо. В противном случае фактор нецелесообразно включать в модель регрессии.

При построении модели регрессии возможна, как уже было сказано ранее, проблема **мультиколлинеарности**, под которой понимается тесная зависимость между факторными признаками, включенными в модель, при этом парный коэффициент корреляции между соответствующими факторами должен превышать 0,8. Наличие мультиколлинеарности между признаками может привести к искажению величины параметров модели, которые имеют тенденцию к завышению. Таким образом, процесс определения наиболее существенных факторных признаков осложнится, а также изменится смысл экономической интерпретации коэффициентов регрессии.

Наиболее часто мультиколлинеарность между факторными признаками возникает в следующих случаях.

- Факторные признаки служат характеристикой одной и той же стороны изучаемого явления или процесса. Например, показатели объема производимой продукции и среднегодовой стоимости основных фондов одновременно включать в модель не рекомендуется, так как они оба характеризуют размер предприятия.
- Факторные признаки являются составляющими элементами друг друга. Например, показатели выработки продукции на одного работающего и численность работающих одновременно в модель включать нельзя, так как в основе расчета показателей лежит один и тот же показатель — численность работающих на предприятии.

- Факторные признаки по экономическому смыслу дублируют друг друга.

Устранение мультиколлинеарности может реализовываться через исключение из корреляционной модели одного или нескольких линейно связанных факторных признаков или преобразование исходных факторных признаков в новые, укрупненные факторы.

Вопрос о том, какой из факторов следует отбросить, решается на основании качественного, логического анализа изучаемого явления, а также на основе анализа тесноты связи между результативным и каждым из сильно коллинеарно связанных факторных признаков. Из дальнейшего анализа целесообразно исключить тот факторный признак, связь которого с результативным наименьшая.

Аналитическая форма зависимости результативного признака от нескольких факторных выражается и называется многофакторным (множественным) уравнением регрессии, или моделью связи.

Для исключения случайности нахождения регрессионных параметров после построения уравнения регрессии следует проверить его на значимость. Для этого используется следующая статистика.

1. Решение о значимости уравнения принимается на основании p-значения. Если p-значение больше 0,05, значит, соответствующая модель не является значимой (принимается нулевая гипотеза о том, что x-переменные не могут прогнозировать результативную величину). Если p-значение оказывается меньше чем 0,05, то модель является значимой. В случае $p < 0{,}01$ модель является высокозначимой.

2. Решение о значимости уравнения принимается на основании R^2. Если значения R^2 меньше, чем критическое значение R^2 в таблице, значит, соответствующая модель не является значимой и наоборот. В любом случае результат будет таким же, как и в первом случае.

3. Решение о значимости уравнения принимается на основании F-распределения. Если значение F-критерия оказывается меньшим, чем критическое значение в таблице F-распределения, значит, соответствующая модель не является значимой и наоборот.

F-критерий Фишера—Снедекора заключается в выдвижении и проверке гипотезы H_0 о равенстве нулю всех параметров в уравнении регрессии: $H_0 : a_0 = a_1 = a_2 = ... = a_k = 0$. Расчетное значение критерия определяется по формуле:

$$F_{\text{расч}} = \frac{\frac{1}{k+1}\sum \bar{y}_x^2}{\frac{1}{n-k-1}\sum (y_i - \bar{y}_x)^2}, \tag{3.2.7}$$

где y_i — наблюдаемые значения результативного признака; $\bar{y}_x$ — теоретические значения результативного признака, полученные по уравнению регрессии; n — количество наблюдений; k — количество факторных признаков в уравнении связи.

Критическое значение определяется при $\{\alpha;\ \upsilon_1 = k + 1;\ \upsilon_2 = n - k - 1\}$. Если $F_{\text{расч}} > F_{\text{кр}}$, то гипотеза о равенстве нулю всех коэффициентов в уравнении регрессии отклоняется с вероятностью α, следовательно, уравнение считается значимым. Если уравнение регрессии является незначимым, то проводить его дальнейший анализ нецелесообразно. Когда вы принимаете нулевую гипотезу, это не означает, что взаимосвязи нет: вам просто не хватает убедительных доводов в пользу наличия такой взаимосвязи. Взаимосвязь может существовать, но из-за случайных колебаний или малого размера выборки ее невозможно обнаружить.

Если уравнение регрессии признается значимым, то одна или несколько x-переменных могут быть полезны в прогнозировании результативной переменной y, и, следовательно, можно продолжить анализ с помощью t-критерия Стьюдента. Данная статистика используется для проверки значимости отдельных коэффициентов регрессии. Расчетное значение критерия для каждого коэффициента определяется по формуле:

$$t_{\text{расч}} = \frac{a_i}{\sqrt{\sigma_{a_i}^2}}, \tag{3.2.8}$$

где $\sigma_{a_i}^2$ — дисперсия соответствующего факторного признака.

Расчетное значение критерия сравнивается с его табличным при заданном уровне значимости α и числе степеней свободы $\upsilon = n - k - 1$. Если $|t_{\text{расч}}| > t_{\text{табл}}$, то соответствующий коэффициент регрессии признается значимым. Если коэффициент является незначимым, он исключается из модели, и модель строится заново: находятся новые оценки параметров регрессии, проверяется значимость уравнения и значимость коэффициентов регрессии.

Главной целью построения регрессионной модели является ее использование для дальнейшего прогнозирования значений результативного признака y. При проведении таких расчетов следует помнить об ошибках, возникающих при прогнозировании.

Первый вид ошибок связан с тем, что модель отражает лишь тенденцию развития. Так, на диаграмме рассеяния всегда видно, что точки фактических данных в действительности рассеяны вокруг линии регрессии (с небольшими отклонениями от нее). Данный факт приводит к возникновению определенного рода ошибок, показателем для измерения которых служит **остаточная дисперсия** (средняя квадратическая ошибка аппроксимации). Чем меньше величина остаточной дисперсии, тем точнее выбранная модель описывает взаимосвязь признаков и тем вернее будет сделанный прогноз. Остаточная дисперсия определяется по формуле:

$$\sigma_{\text{ост}} = \sqrt{\sigma^2_{\text{ост}}} = \sqrt{\frac{\sum (y_i - \bar{y}_x)^2}{n-k-1}}. \qquad (3.2.9)$$

Второй вид ошибки связан с тем, что найденные параметры модели являются лишь оценками своих истинных значений, поэтому для прогнозного значения всегда целесообразно определить границы *доверительного интервала*, в которых с доверительной вероятностью будут находиться значения результативной переменной y при определенных значения x-переменных. На первом этапе строят доверительные интервалы для коэффициентов регрессии. При линейной модели границы доверительного интервала для параметра a_j определяются по формуле:

$$a_j \pm t_{\text{табл}} \cdot \sigma_{a_j}, \qquad (3.2.10)$$

где $t_{\text{табл}}$ — критическое значение определяемой по таблице t-распределения Стьюдента для выбранного уровня значимости α и числа степеней свободы $\upsilon = n - k - 1$; σ_{a_j} — среднее квадратическое отклонение коэффициента регрессии a_j.

На практике множественный регрессионный анализ не всегда позволяет получить положительные результаты. Существует три основных вида проблем.

1. Проблема *мультиколлинеарности* возникает в тех случаях, когда некоторые из факторных переменных оказываются слишком схожими. Несмотря на то что эти переменные могут хорошо объяснять и про-

гнозировать y (на что указывают высокое значение R^2 и значимое F-распределение), из уравнения одну из переменных следует удалить либо перераспределить какие-то переменные, например путем деления, чтобы отличать одну переменную от другой.

2. Проблема *выбора переменных* возникает в тех случаях, когда приходится иметь дело с достаточно большим перечнем потенциально полезных объясняющих факторов и необходимо решить, какие из них необходимо включать в уравнение регрессии. При большом числе x-переменных лишние из них будут снижать качество результатов. В то же время, если отбросить нужную переменную, снизится качество прогнозов.

3. Проблема *неправильного выбора модели* связана с множеством различных потенциальных несоответствий между конкретной задачей и моделью множественной регрессии. Может сложиться ситуация, при которой конкретная задача не соответствует условиям и допущениям модели множественной регрессии.

3.2.3. Интерпретация результатов множественной регрессии

Для понимания результатов построенной модели множественной регрессии приведем краткий обзор исходных данных и основных результатов. Пусть k означает количество объясняющих переменных x, которое может быть любым разумным числом. Элементарные единицы изучаемой совокупности называют **наблюдениями**: это могут быть клиенты, фирмы, выпускаемые услуги или изделия, банки, регионы и т. д.

Параметр a_0 (сдвиг) определяет прогнозируемое значение y, когда все переменные x равны нулю. **Коэффициент регрессии** для каждой переменной x определяет влияние этой переменной на результативный признак y при условии, что остальные переменные x не меняются. Коэффициент регрессии a_j для j-й переменной x указывает, как изменится результативный признак, когда переменные x остаются неизменными, а переменная x_j увеличивается (уменьшается) на единицу своего измерения. Коэффициенты регрессии составляют **уравнение регрессии**, которое используется в целях управления и прогнозирования. **Параметры** уравнения регрессии (a_0, a_1, a_2, ..., a_k), как уже было сказано, вычисляются методом наименьших квадратов, который минимизирует сумму квадратов отклонений наблюдаемых значений результативного признака от теоретических, полученных по уравнению регрессии (ошибок прогнозирования). **Ошибки прогнозирования** (остатки) определяются как отклонения фактических значений результативного признака от теоретических, вычисленных по уравнению регрессии.

Стандартная ошибка оценки S указывает приблизительную величину ошибок прогнозирования. R^2 является коэффициентом **детерминации**, который показывает, какой процент вариации результативного признака y объясняется всеми факторными признаками, включенными в модель.

Статистический вывод начинается с проверки значимости уравнения регрессии с использованием статистики F-распределения Фишера. *Уравнение регрессии необходимо проверять на значимость, чтобы выяснить, объясняют ли переменные x значимую долю вариации y.* Дальнейший анализ проводят только в случае значимости построенного уравнения регрессии. Оценку значимости коэффициентов регрессии проводят на основании t-распределения Стьюдента, которое показывает, насколько значимым является влияние каждого факторного признака на результативный, при условии, что все остальные переменные останутся неизменными. Построение доверительных интервалов и проверка гипотез для отдельных коэффициентов регрессии основывается на стандартной ошибке. Каждый коэффициент регрессии имеет свою стандартную ошибку: S_{a_1}, S_{a_2}, ..., S_{a_k}. В табл. 3.2.1 представлены результаты множественного многофакторного корреляционно-регрессионного анализа.

Таблица 3.2.1

Результаты многофакторного корреляционно-регрессионного анализа

Название показателя	Обозначение	Описание
Постоянный коэффициент (сдвиг)	a_0	Прогнозируемое значение результативного признака, если все значения факторных признаков равны нулю
Коэффициенты регрессии	$a_1, a_2, ..., a_k$	Влияние каждой переменной x на результативную переменную при условии, что все другие переменные x останутся неизменными
Уравнение регрессии	$\bar{y}_x = a_0 + a_1x_1 + a_2x_2 + \ldots + a_kx_k$	Прогнозируемое значение результативной переменной при заданных значениях факторных переменных

Окончание табл. 3.2.1

Название показателя	Обозначение	Описание
Ошибки прогнозирования (остатки)	$y_i - \bar{y}_x$	Ошибка возникает в связи с использованием для каждого наблюдения теоретических значений, полученных по уравнению регрессии, вместо фактического значения *y* для каждого наблюдения
Стандартная ошибка оценки	σ	Приблизительная величина ошибки прогнозирования (типичная разница между фактическими значениями зависимой переменной и его прогнозом, исходя из уравнения регрессии)
Множественный коэффициент корреляции	R	Характеризует тесноту линейной связи между одной переменной и совокупностью других переменных, рассматриваемых в корреляционном анализе
Коэффициент детерминации	R^2	Процент изменчивости *y*, объясняемой всей группой *x*-переменных
F-распределение Фишера–Снедекора	$F_{расч} > F_{табл}$	Проводится проверка на значимость коэффициента детерминации и уравнения регрессии
t-распределения Стьюдента для отдельных коэффициентов регрессии	$t_{расч} > t_{табл}$	Проводится проверка значимости каждого коэффициента уравнения регрессии, т. е. влияет ли конкретная переменная *x* на результативный показатель при условии, что все другие факторы остаются неизменными
Стандартная ошибка коэффициентов регрессии	$\sigma_{a_1}, \sigma_{a_2}, \sigma_{a_k}$	Показывает выборочную оценку стандартного отклонения каждого коэффициента регрессии; используется для проверки гипотез и нахождения доверительных интервалов

ПРИМЕР.

Тарифы на размещение рекламных объявлений в журналах определяются каждым журналом самостоятельно. Какими причинами можно объяснить различия в тарифах разных печатных изданий? Предположим, здесь каким-то образом учитывается ценность рекламного объявления для рекламодателя. Журналы, располагающие большей читательской аудиторией (при прочих равных условиях), могут устанавливать большие тарифы. Кроме того, печатные издания, рассчитанные на достаточно состоятельный класс, также вправе устанавливать более высокие тарифы. Кроме вышеприведенных причин, естественно, есть и другие факторы: например, предпочтения людей разного пола, возраста, социального положения и т. д.

Для анализа влияния на тариф таких факторов, как величина читательской аудитории, структура читательской аудитории по полу и доходы читателей, воспользуемся многофакторным корреляционно-регрессионным анализом. Многомерная совокупность данных представляет собой выборку, состоящую из 50 печатных изданий. В качестве результативного признака y используем стоимость одной страницы одноразовой полноцветной рекламы. Факторными признаками будут: читательская аудитория (x_1), процент мужчин от численности всей читательской аудитории (x_2), среднемесячный доход семьи (x_3).

Результатами множественного регрессионного анализа является уравнение регрессии, полученное с помощью пакета STATISTICA:

$$\bar{y}_x = 4043 + 3{,}79x_1 - 124x_2 + 0{,}903x_3.$$

Другие пакеты программного обеспечения для статистических расчетов позволяют получить такую же базовую модель. Например, с помощью Excel также можно осуществить множественный регрессионный анализ (найдите в меню «Сервис» пункт «Анализ данных» и выберите команду «Регрессия»).

Параметр a_0 интерпретируется следующим образом: типичный тариф на размещение одностраничного цветного рекламного объявления в печатном издании, у которого нет платных подписчиков, нет мужчин среди читателей и читатели не имеют дохода, составляет 4043 руб. Данная величина рассматривается как вспомогательная для получения оптимальных прогнозов.

Коэффициент регрессии интерпретируется как влияние каждой из переменных на размер тарифа, если все другие объясняющие переменные остаются неизменными. Каждый коэффициент регрессии определяет среднее увеличение тарифа на размещение рекламы, приходящееся на единичное увеличение соответствующей ему переменной x. Коэффициент регрессии для размера читательской аудитории $a_1 = 3{,}79$ показывает, что при прочих равных условиях при увеличении численности читательской аудитории на 1 тыс. чел. стоимость рекламного объявления в журнале увеличится на 3,79 руб., т. е. каждый новый читатель увеличивает для этого журнала тариф на размещение рекламных объявлений на 0,00 379 руб.

Коэффициент регрессии для процента мужчин $a_2 = -124$ показывает, что тариф на размещение рекламных объявлений в журнале с дополнительным 1% читателей-мужчин окажется в среднем на 124 руб. ниже. Таким образом, читатели-женщины представляют большую ценность для журналов, чем мужчины.

Коэффициент регрессии для среднемесячного дохода иллюстрирует, что при увеличении среднемесячного денежного дохода читателей соответствующего журнала на 1 руб. тариф на размещение одностраничного цветного рекламного объявления увеличится на 0,903 руб. Данная величина совершенно оправданна, поскольку люди с более высоким уровнем доходов могут позволить себе тратить больше на покупку рекламируемой продукции. В таком случае более высокие уровни доходов читателей ведут к установлению более высоких тарифов на размещение рекламных объявлений.

Для нахождения прогнозируемой величины тарифа на размещение рекламных объявлений, исходя из величины читательской аудитории, процента читателей-мужчин и среднемесячного дохода читателей для конкретного журнала, подставим в уравнение прогнозирования соответствующие этому журналу значения переменных x.

Допустим, например, вы собираетесь основать новый журнал «Практическая статистика», который рассчитан на читательскую аудиторию порядка 500 000 человек, 45% которого будут составлять женщины, а среднемесячный доход его читателей составит 35 000 руб. Тогда прогнозируемый тариф на размещение рекламы в журнале составит:

$$\bar{y}_x = 4043 + 3{,}79x_1 - 124x_2 + 0{,}903x_3 =$$
$$= 4043 + 3{,}79 \cdot 500 - 124 \cdot 55 + 0{,}903 \cdot 35\,000 = 30\,723 \text{ руб.}$$

Следует учесть, что тариф на размещение рекламы в журнале ровно 30 723 руб. не составит. Это связано с тем, что даже между журналами, исследование по которым мы проводили, наблюдаются случайные колебания, а также прогноз может быть полезен только в той мере, в какой прогнозируемый журнал подобен журналам, принадлежащим исходной совокупности данных.

В некоторых случаях зависимость факторных переменных может приводить к изменению знака регрессионного коэффициента по сравнению со знаком парного коэффициента корреляции и не отражать истинное направление связи. Поэтому интерпретировать полученные коэффициенты при множественной регрессии следует с большой осторожностью. Также нельзя на основе значений регрессионных коэффициентов сравнить степень влияния на результативную переменную отдельных факторов за счет того, что переменные имеют, как правило, разные единицы измерения, учитываемые при анализе.

Поэтому в таких случаях для каждого фактора рассчитывается частный **коэффициент эластичности**, который показывает, на сколько процентов в среднем изменяется значение результативного признака при изменении фактора x_j на 1% при условии, что остальные факторы останутся неизменными. Частный коэффициент эластичности для фактора x_j можно рассчитать по формуле:

$$Э_{x_j} = a_j \cdot \frac{\bar{x}_j}{\bar{y}}, \tag{3.2.11}$$

где a_j — коэффициент регрессии при факторе x_j; $\bar{x}_j$ — среднее значение фактора x_j; $\bar{y}$ — среднее значение результативного показателя.

Определить, на сколько процентов изменится значение результативного признака при одновременном изменении всех факторов на 1%, можно, сложив частные коэффициенты эластичности, рассчитанные для каждого факторного признака:

$$Э_{общ} = \sum_{j=1}^{k} Э_{x_j}. \tag{3.2.12}$$

3.2.4. Применение корреляционно-регрессионного анализа в исследовании финансово-экономической деятельности

Имеются следующие данные о рыночной стоимости фирмы (y), стоимости активов (x_1), которыми владеет фирма, и количестве сотрудников (x_2) компаний, производящих компьютерное и офисное оборудование, представленные в табл. 3.2.2.

Таблица 3.2.2

Данные по основным показателям деятельности крупнейших американских компаний, производящих компьютерное и офисное оборудование

Название компании	Рыночная стоимость, млн долл. США	Активы, млн долл. США	Количество служащих, чел.
Intl. Business Machines	98 322	81 449	269 465
Hewlett-Packard	65 060	31 749	121 900
Compaq Computer	36 052	14 631	37 004
Xerox	31 829	27 732	91 400
Digital Equipment	7101	9693	54 900
Dell Computer	41 294	4268	16 160
Sun Microsystems	16 614	4697	21 500
Apple Computer	3499	4233	9306

Окончание табл. 3.2.2

Название компании	Рыночная стоимость, млн долл. США	Активы, млн долл. США	Количество служащих, чел.
NCR	3386	5293	38 300
Gateway 2000	6242	2039	13 369
Pitney Bowes	14 036	7893	29 901
Silicon Graphics	2636	3345	10 930
Data General	869	1135	5100
Intergraph	452	727	7653

Для многофакторного корреляционного анализа воспользуемся компьютерной программой SPSS. Корреляционная матрица, отражающая взаимосвязь каждой пары переменной исследуемой совокупности исходных данных, представлена в табл. 3.2.3.

Таблица 3.2.3

Корреляционная матрица для компаний, производящих компьютеры и офисное оборудование

	Рыночная стоимость, y	Активы, x_1	Количество служащих, x_2
Рыночная стоимость, y	1,000	0,814	0,888
Активы, x_1	0,814	1,000	0,916
Количество служащих, x_2	0,888	0,916	1,000

Анализ матрицы парных коэффициентов корреляции показывает, что в совокупности исходных данных присутствует мультиколлинеарность, поскольку каждая переменная x обозначает размер компании. Данные значения факторных признаков пересекаются по всем представленным показателям, так как крупные компании являются крупными во всех отношениях: рыночной стоимости, величине активов, количестве служащих. И наоборот, небольшие компании характеризуются относительно невысокими значениями указанных показателей.

Наглядное подтверждение коллинеарности дает коэффициент корреляции между двумя факторными переменными: количеством служащих и активами $r_{x_1x_2} = 0{,}916$. Столь высокая корреляция говорит о том, что с точки зрения количественного анализа оба факторных показателя однотипны и включают в себя идентичную информацию.

Таблица 3.2.4

Множественный коэффициент корреляции и детерминации для компаний, производящих компьютеры и офисное оборудование

Модель	R	R^2	Скорректированный R^2	Стандартная ошибка расчетов
1	0,888	0,789	0,751	14 416,25 540

Таблица 3.2.5

F-статистика для множественного коэффициента корреляции и детерминации

Модель	Сумма квадратов отклонений	υ	Средняя квадратическая ошибка	$F_{расч}$	$F_{кр}$
Регрессии	8 569 861 372,530	2	4 284 930 686,265	20,618	0,000
Остатки	2 286 112 617,184	11	207 828 419,744		
Итог	10 855 973 989,714	13			

Дисперсионный анализ системы связей используется для оценки надежности подтвержденной связи результативного признака со всеми факторами, включаемыми в модель регрессии. Необходимо сравнить дисперсии результативного признака — остаточную и объясненную: суммы соответствующих квадратов отклонений, приходящиеся на одну степень свободы вариации. Отношение дисперсии за счет факторов к остаточной дисперсии и есть F-критерий Фишера. Для нашего исследования $F_{расч} = 20{,}618$. $F_{табл}\{\alpha = 0{,}05;\ \upsilon_1 = 2;\ \upsilon_2 = 11\}$ составляет 3,98. На основании F-статистики можно сделать вывод о том, что регрессия является значимой. Используя терминологию p-значений, можно сказать, что регрессия в данном случае является очень высокозначимой ($p < 0{,}001$). На 78,9% ($R^2 = 0{,}789$) объем активов и численность сотрудников объясняет рыночную стоимость компаний, что является достаточно высоким статистическим показателем.

Уравнение множественной регрессии для прогнозирования рыночной стоимости на основании активов и количества служащих может быть записано в следующем виде: $\bar{y}_x = 4677{,}074 + 0{,}001x_1 + 0{,}360x_2$ (табл. 3.2.6).

Проверка на значимость полученных коэффициентов регрессии свидетельствует о незначимости переменной x_1 — активы, так как $t_{набл} < t_{кр}$. Таким образом, исходя из правил выполнения регрессионного анализа, переменную x_1 необходимо исключить из уравнения и построить модель регрессии заново. Но если сохранить только одну из переменных x, а именно x_2, то получим регрессию с очень высоким значимым t-тестом для этой переменной. Иными словами, как активы, так и количество служащих вносят существенный вклад в определение рыночной стоимости компьютерных фирм.

Таблица 3.2.6

t-статистика для коэффициентов множественного уравнения регрессии

Модель	Нестандартизированные коэффициенты		Стандартизированные коэффициенты	$t_{расч}$	$t_{кр}$
	a_i	Стандартная ошибка			
y	4677,074	4893,472	–	0,956	0,036
x_2	0,360	0,140	0,888	2,577	0,026
x_1	0,001	0,454	0,001	0,002	0,998

Для сохранения информации, содержащейся в обеих переменных x, одну из них можно использовать в качестве переменной, представляющей величину компании, а другую определить как некоторое отношение. Допустим, выберем в качестве переменной, представляющей величину компании, активы, поскольку они указывают на фиксированные капиталовложения, необходимые соответствующей фирме. После этого вторую переменную можно заменить на отношение количества служащих к величине активов (указывает количество служащих на миллион долларов активов). Таким образом, активы будут выступать единственной переменной, характеризующей величину компании, а другая переменная несет новую информацию об эффективности использования трудовых ресурсов. Новая совокупность данных представлена в табл. 3.2.7.

Таблица 3.2.7

Определение новых переменных x для компаний, производящих компьютерное и офисное оборудование

Название компании	Рыночная стоимость, млн долл. США	Активы, млн долл. США	Отношение количества служащих к размерам активов
Intl. Business Machines	98 322	81 449	3,308
Hewlett-Packard	65 060	31 749	3,839
Compaq Computer	36 052	14 631	2,529
Xerox	31 829	27 732	3,296

Окончание табл. 3.2.7

Название компании	Рыночная стоимость, млн долл. США	Активы, млн долл. США	Отношение количества служащих к размерам активов
Digital Equipment	7101	9693	5,664
Dell Computer	41 294	4268	3,786
Sun Microsystems	16 614	4697	4,577
Apple Computer	3499	4233	2,198
NCR	3386	5293	7,236
Gateway 2000	6242	2039	6,557
Pitney Bowes	14 036	7893	3,788
Silicon Graphics	2636	3345	3,268
Data General	869	1135	4,493
Intergraph	452	727	10,527

Проанализируем снова корреляционную матрицу, представленную в табл. 3.2.8, на наличие мультиколлинеарности между новыми переменными x. В данном случае корреляционный анализ дал достаточно хорошие результаты. Так, взаимосвязь между факторными переменными уже не является столь сильной, что иллюстрирует парный коэффициент корреляции между отношением количества служащих к размерам активов и активами $r_{x_1x_2} = -0{,}317$, и она статистически значима. Как на вариацию рыночной стоимости компьютерных компаний, так и на ее возможное изменение в динамике самое сильное влияние оказывает фактор x_1 — активы, коэффициент корреляции в этом случае составил 0,903.

Таблица 3.2.8

Корреляционная матрица для компаний, производящих компьютеры и офисное оборудование

	Рыночная стоимость, y	Активы, x_1	Отношение количества служащих к размерам активов, x_2
Рыночная стоимость, y	1,000	0,903	–0,401
Активы, x_1	0,903	1,000	–0,317
Отношение количества служащих к размерам активов, x_2	–0,401	–0,317	1,000

Для проведения дальнейшего исследования воспользуемся результатами множественной регрессии.

Таблица 3.2.9

Множественный коэффициент корреляции и детерминации для компаний, производящих компьютеры и офисное оборудование

Модель	R	R^2	Скорректированный R^2	Стандартная ошибка расчетов
1	0,911	0,830	0,799	12 962,59277

Таблица 3.2.10

F-статистика для множественного коэффициента корреляции и детерминации

Модель	Сумма квадратов отклонений	υ	Средняя квадратическая ошибка	$F_{расч}$	$F_{кр}$
Регрессии	9 007 657 065,068	2	4 503 828 532,534	26,804	0,000
Остатки	1 848 316 924,646	11	168 028 811,331		
Итог	10 855 973 989,714	13			

Произведем оценку F-критерия Фишера. Для нашего исследования $F_{расч} = 26{,}804$. $F_{табл}$ $\{\alpha = 0{,}05;\ \upsilon_1 = 2;\ \upsilon_2 = 11\}$ составляет 3,98. На основании F-статистики можно сделать вывод о том, что регрессия является значимой. Используя терминологию p-значений, можно сказать, что регрессия в данном случае является очень высокозначимой ($p < 0{,}001$). На 83,0% ($R^2 = 0{,}83$) объем активов и количество служащих на миллион долларов активов объясняют рыночную стоимость компаний, что является очень высоким статистическим индикатором.

Таблица 3.2.11

t-статистика для коэффициентов множественного уравнения регрессии

Модель	Нестандартизированные коэффициенты		Стандартизированные коэффициенты	$t_{расч}$	$t_{кр}$
	a_i	Стандартная ошибка			
y	14 691,676	9678,532		1,518	0,157
x_1	1,154	0,176	0,862	6,575	0,000
x_2	–1658,136	1706,682	–0,127	–0,972	0,352

Уравнение множественной регрессии для прогнозирования рыночной стоимости на основании активов и количества служащих (табл. 3.2.11) может быть записано в следующем виде:

$$\bar{y}_x = 14691 + 1{,}154x_1 - 1658{,}136x_2.$$

Это означает, что величина рыночной стоимости компаний в среднем по совокупности возрастает на 1,154 млн долл. при увеличении активов на 1 млн долл. и уменьшается в среднем на 1658,136 млн долл. при возрастании количества служащих на 1 млн долл. активов.

Проверка значимости полученных коэффициентов регрессии свидетельствует о значимости всех параметров по причине отсутствия «конкурирующих» переменных, характеризующих величину компании. Очевидно, для этой небольшой группы крупных компаний, производящих компьютеры и офисное оборудование, большая доля вариации рыночной стоимости может объясняться объемом активов этих фирм. Кроме того, данные о трудовых ресурсах практически не содержат новой информации о рыночной стоимости представленных наиболее успешных компаний. Возможно, при увеличении объема выборки компаний сила влияния переменной x_2 изменится.

КОНТРОЛЬНЫЕ ВОПРОСЫ

1. Какие цели преследует множественный корреляционно-регрессионный анализ?
2. Для чего используется уравнение регрессии? Как интерпретируется постоянный коэффициент уравнения регрессии и коэффициенты регрессии?
3. Какую статистику проверяют на основании F-распределения Фишера? Запишите две гипотезы F-теста. При каких условиях определяется $F_{табл}$? Если F-тест незначим, можно ли продолжать анализ и проверять на значимость отдельные коэффициенты регрессии?
4. Какую статистику проверяют на основании t-распределения Стьюдента? Запишите две гипотезы t-теста. При каких условиях определяется $t_{табл}$?
5. Что такое мультиколлинеарность? В чем заключается искажающее влияние высокой мультиколлинеарности? Каким образом решается проблема мультиколлинеарности?
6. Для получения наилучших прогнозов целесообразно ли включать в число факторных переменных все потенциально возможные показатели? Каким образом упорядоченный по приоритетам список переменных может помочь в решении проблемы выбора переменных?
7. К каким методам выбора переменных относится метод пошаговой регрессии? В чем состоит суть данного метода? Кратко опишите этапы его реализации.

Глава 3.3

НЕПАРАМЕТРИЧЕСКИЕ МЕТОДЫ ОЦЕНКИ СВЯЗИ СОЦИАЛЬНО-ЭКОНОМИЧЕСКИХ ЯВЛЕНИЙ

- Порядковая шкала и ее применение в анализе взаимосвязей
- Ранговые коэффициенты корреляции Спирмена и Кендалла
- Коэффициент конкордации (множественный коэффициент ранговой корреляции)
- Бисериальный коэффициент корреляции
- Методы изучения взаимосвязи между качественными признаками

3.3.1 Порядковая шкала и ее применение в анализе взаимосвязей

Рассмотренные в гл. 3.1 и 3.2 методы корреляционно-регрессионного анализа используются только для переменных, выраженных количественно. Когда необходимо анализировать свойства объектов, представленные в виде порядковых шкал, данные методы непригодны. Измерение по порядковой шкале означает, что исследователь получает данные, к какой категории по данному признаку принадлежит объект и в каком отношении он находится с другими объектами, принадлежащими к другим категориям по данному признаку. Важно, что при измерении по порядковой шкале получают не величину различий между объектами по данному признаку, а устанавливают порядок следования объектов.

Порядковые переменные входят в социологическую информацию, собираемую в форме опросов, интервью, заполнения вопросников. Однако все чаще такие переменные встречаются в данных экономических исследований. Это происходит в связи с необходимостью учесть человеческий фактор в бизнесе. Во всем мире, и в России в том числе, используется система бизнес-обследований, в основе которой лежит опрос руководителей организации о ее состоянии. Обычно участникам опроса рассылают новую анкету вместе с результатами предыдущего опроса, что повышает их мотивацию давать точные и оперативные ответы. Руководителям предлагается оценить фактическое состояние и ожидаемое изменение основных показателей хозяйственной деятельности в рамках альтернатив: «увеличение—уменьшение», «улучшение—ухудшение», «сохранение на прежнем уровне». Получаемые в ходе опроса данные всего лишь отражают оценки и ожидания руководителей, однако собранная информация может быть использована аналитиками для оценки и прогнозирования деятельности фирмы в целом. Причиной использования в бизнес-опросах порядковых шкал является коммерческая тайна. Вам могут не назвать конкретную сумму прибыли (убытка) за период, но скажут, как изменилось финансовое положение предприятия: улучшилось, ухудшилось, осталось прежним.

Порядковые данные становятся все более популярны в связи с построением различных рейтингов: коммерческих банков, высших учебных заведений, торговых и промышленных организаций, политических деятелей.

Еще одним фактором расширения сферы порядковых переменных служит повышение внимания к социальным проблемам (выявление удовлетворенности людей разными сторонами жизни — работой, образованием, отношениями в семье, жизненными планами и т. п.).

Для исследования взаимосвязей социально-экономических явлений и процессов в статистике разработаны **непараметрические методы** анализа, реализация которых не требует знания законов и параметров распределения показателей. При изучении взаимосвязей количественных признаков методами непараметрической статистики используются следующие показатели:

- коэффициент корреляции рангов Спирмена;
- коэффициент корреляции рангов Кендалла;
- коэффициент конкордации (множественный коэффициент ранговой корреляции).

При оценке взаимосвязи качественных признаков в статистике используются следующие показатели:

- коэффициент ассоциации;
- коэффициент контингенции;
- коэффициент взаимной сопряженности Пирсона;
- коэффициент взаимной сопряженности Чупрова.

Если необходимо оценить тесноту связи между одним количественным и одним качественным признаками, используется бисериальный коэффициент корреляции.

Широкий выбор непараметрических методов оценки взаимосвязи социально-экономических явлений и процессов предоставляют современные программные средства, например STATISTICA, которая содержит специальный модуль *«Непараметрическая статистика»*.

3.3.2. Ранговые коэффициенты корреляции Спирмена и Кендалла

В анализе социально-экономических явлений часто приходится прибегать к различным условным оценкам с помощью рангов, а взаимосвязь между отдельными признаками измерять с помощью непараметрических коэффициентов связи.

Ранжирование — это процедура упорядочения объектов изучения, которая выполняется на основе предпочтения.

Ранги — это порядковые номера значений признака, расположенных в порядке возрастания или убывания их величин. Если значения признака имеют одинаковую количественную оценку, то ранг всех этих значений принимается равным среднему арифметическому из соответствующих номеров мест, которые они определяют. Такие ранги называются *связными*.

Среди непараметрических методов оценки тесноты связи наибольшее значение имеют ранговые коэффициенты Спирмена (ρ_{xy}) и Кендалла (τ_{xy}). Эти коэффициенты могут быть использованы для определения тесноты связи как между количественными, так и между качественными признаками (рейтинги, уровни образования, квалификации и т. п.).

Коэффициент корреляции рангов Спирмена принимает значения в интервале $[-1; 1]$ и в случае несвязных рангов рассчитывается по формуле:

$$\rho_{xy} = 1 - \frac{6\sum d_i^2}{n(n^2 - 1)}, \tag{3.3.1}$$

где d_i^2 — квадраты разности рангов; n — число наблюдений, число пар рангов.

Значимость коэффициента корреляции рангов Спирмена проверяется на основе t-критерия Стьюдента. Расчетное значение критерия определяется по формуле:

$$t_{\text{расч}} = \rho_{xy} \cdot \sqrt{\frac{n-2}{1-\rho_{xy}^2}}. \tag{3.3.2}$$

Критическое значение определяется при уровне значимости α и числе степеней свободы $\upsilon = n - 2$. Значение коэффициента Спирмена признается значимым, если $t_{\text{расч}} > t_{\text{табл}}$.

При расчете коэффициента корреляции рангов Спирмена на первом этапе каждому значению показателя x и каждому значению показателя

y присваивается определенный ранг, который является порядковым номером значения в ранжированном по возрастанию (убыванию) ряду. Наименьшей (наибольшей) величине признака присваиваем ранг, равный единице. Следующему за ним во возрастанию (убыванию) значению — ранг 2 и т. д. На втором этапе определяем разность рангов и квадрат разности рангов d_i^2. Подставив полученные результаты в формулу, производят расчет коэффициента Спирмена.

ПРИМЕР.

По данным о прибыли и обьеме кредитных вложений десяти коммерческих банков одного из регионов Российской Федерации на 1 января 2006 г. определим с помощью коэффициента Спирмена зависимость между этими признаками.

Таблица 3.3.1

Расчетная таблица для определения коэффициента Спирмена

№ банка	Кредитные вложения, X млн руб.	Прибыль, Y млн руб.	Ранжирование		Сравнение рангов		Квадрат разности рангов $d_i^2 = (R_x - R_y)^2$
			X	Y	R_x	R_y	
1	2887	557	3010	628	2	4	4
2	1710	605	2887	605	10	2	64
3	3010	628	2783	589	1	1	0
4	2472	488	2535	557	5	6	1
5	2535	418	2472	501	4	8	16
6	1897	397	2003	488	7	9	4
7	2783	501	1897	437	3	5	4
8	1862	589	1862	418	8	3	25
9	1800	269	1800	397	9	10	1
10	2003	437	1710	269	6	7	1
Σ	–	–	–	–	–	–	120

Для расчета коэффициента Спирмена воспользуемся формулой (3.3.1):

$$\rho_{xy} = 1 - \frac{6 \cdot 120}{10 \cdot 99} = 1 - \frac{720}{990} = 0{,}3.$$

Таким образом, связь между кредитными вложениями и прибылью коммерческих банков можно признать слабой, а по направлению – положительной (прямой).

Проверим коэффициент корреляции рангов Спирмена на значимость:

$$t_{\text{расч}} = \rho_{xy} \cdot \sqrt{\frac{n-2}{1-\rho_{xy}^2}} = 0{,}3 \cdot \sqrt{\frac{10-2}{1-(0{,}3)^2}} = 0{,}89,$$

$$t_{\text{табл}}\{\alpha = 0{,}05;\ \upsilon = 8\} = 2{,}306.$$

Так как $t_{\text{расч}} < t_{\text{табл}}$, то сделанный ранее вывод о незначимости коэффициента корреляции Спирмена подтверждается.

Если совокупность значений по исследуемому признаку содержит связные ранги, то коэффициент корреляции Спирмена вычисляется по формуле:

$$\rho_{xy} = \frac{\frac{1}{6}\cdot\left(n^3-n\right)-\sum_{i=1}^{n} d_i^2 - T_x - T_y}{\sqrt{\left[\frac{1}{6}\cdot(n^3-n)-2T_x\right]\cdot\left[\frac{1}{6}\cdot(n^3-n)-2T_y\right]}}, \qquad (3.3.3)$$

где $T_{xy} = \frac{1}{12}\sum_{j=1}^{k}(t_j^3 - t_j)$; t_j — число одинаковых рангов в j-м ряду.

ПРИМЕР.

Производится тестирование воздействия рекламного ролика на потребителя. Цель ролика — вызвать «эффект расслабления».

Реклама проходит тестирование на эффективность создания настроения расслабления. Выборка из 15 человек была опрошена до и после просмотра рекламного ролика. Вопросник включал достаточно много пунктов, но главный вопрос был сформулирован следующим образом: «Пожалуйста, опишите свое состояние в настоящий момент, используя шкалу от 1 (очень напряженное) до 5 (полностью расслабленное)». Результаты опроса приведены в табл. 3.3.2.

Таблица 3.3.2

Расчетная таблица для определения коэффициента Спирмена

Респондент	Оценка расслабленности		Ранжирование				Сравнение рангов		Квадрат разности рангов
	до, x	после, y	x	R_x	y	R_y	R_x	R_y	$d_i^2 = (R_x - R_y)^2$
1	3	2	1	1	1	1,5	9,5	4	30,25
2	2	2	2	4,5	1	1,5	4,5	4	0,25
3	2	2	2	4,5	2	4	4,5	4	0,25

Окончание табл. 3.3.2

Респон-дент	Оценка расслабленности		Ранжирование				Сравнение рангов		Квадрат разности рангов $d_i^2=(R_x-R_y)^2$
	до, x	после, y	x	R_x	y	R_y	R_x	R_y	
4	4	5	2	4,5	2	4	13	13	0
5	2	4	2	4,5	2	4	4,5	8,5	16
6	2	1	2	4,5	3	6	4,5	1,5	9
7	1	1	2	4,5	4	8,5	1	1,5	0,25
8	3	5	3	9,5	4	8,5	9,5	13	12,25
9	3	4	3	9,5	4	8,5	9,5	8,5	1
10	2	4	3	9,5	4	8,5	4,5	8,5	16
11	5	5	3	9,5	5	13	15	13	4
12	2	3	4	13	5	13	4,5	6	2,25
13	4	5	4	13	5	13	13	13	0
14	3	5	4	13	5	13	9,5	13	12,25
15	4	4	5	15	5	13	13	8,5	20,25
Σ	–	–	–	–	–	–	–	–	124

Для расчета коэффициента Спирмена воспользуемся формулой (3.3.3):

$$\rho_{xy}=\frac{\frac{1}{6}\cdot\left(n^3-n\right)-\sum_{i=1}^{n}d_i^2-T_x-T_y}{\sqrt{\left[\frac{1}{6}\cdot(n^3-n)-2T_x\right]\cdot\left[\frac{1}{6}\cdot(n^3-n)-2T_y\right]}},$$

где $T_x=\frac{1}{12}\left[(6^3-6)+(4^3-4)+(3^3-3)\right]=24{,}5;$

$$T_y=\frac{1}{12}\left[(2^3-2)+(3^3-3)+(4^3-4)+(5^3-5)\right]=17{,}5.$$

$$\rho_{xy}=\frac{\frac{1}{6}\cdot(15^3-15)-124-24{,}5-17{,}5}{\sqrt{\left[\frac{1}{6}\cdot(15^3-15)-2\cdot24{,}5\right]\cdot\left[\frac{1}{6}\cdot(15^3-15)-2\cdot17{,}5\right]}}=0{,}76.$$

Таким образом, можно признать существенное влияние данного рекламного ролика на расслабленность респондентов после его просмотра.

Ранговый коэффициент корреляции Кендалла (τ_{xy}) также может использоваться для измерения взаимосвязи между качественными и количественными признаками, характеризующими однородные объекты и ранжированные по одному принципу. Расчет рангового коэффициента Кендалла в случае несвязных рангов осуществляется по формуле:

$$\tau_{xy} = \frac{2S}{n(n-1)}, \tag{3.3.4}$$

где n — число наблюдений; S — фактическая сумма рангов.

Коэффициент Кендалла, так же как и коэффициент Спирмена, лежит в границах $[-1;1]$, однако дает более строгую оценку связи $\tau_{x/y} < \rho^{*}_{x/y}$.

Расчет коэффициента Кендалла выполняется в следующей последовательности.

1. Значения x ранжируются в порядке возрастания или убывания.
2. Значения y располагаются в порядке, соответствующем значениям x.
3. Для каждого ранга y определяется число следующих за ним значений рангов, превышающих его величину. Суммируя таким образом числа, определяется величина P как мера соответствия последовательностей рангов по x и y и учитывается со знаком (+).
4. Для каждого ранга y определяется число следующих за ним значений рангов, меньших его величины. Суммарная величина обозначается через Q и фиксируется со знаком (–).
5. Определяется фактическая сумма рангов: $S = P + Q$.

Как правило, коэффициент Кендалла меньше коэффициента Спирмена.

Связь между признаками признается значимой, если значения коэффициентов ранговой корреляции Спирмена и Кендалла больше 0,5.

ПРИМЕР.

По данным о темпах прироста валового национального продукта (ВНП) и потребительских цен 10 развитых стран мира за 2005 г. определим ранговый коэффициент Кендалла (табл. 3.3.3).

Таким образом, коэффициент Кендалла согласно формуле (3.3.4) составит:

$$\tau_{xy} = \frac{2\cdot(13-32)}{10\cdot(10-1)} = -0{,}42.$$

Таблица 3.3.3

Темпы прироста основных макроэкономических показателей развитых стран

Страны	Прирост ВНП x, %	Индекс цен y, %	R_x	R_y	Ранжирование		P	Q
					x	y		
Япония	3,5	2,1	1	10	1	10	0	–9
США	3,2	3,9	2	4	2	4	5	–3
Германия	2,2	3,3	8	7	3	8	1	–6
Франция	2,8	2,9	4	9	4	9	0	–6
Италия	2,7	5,6	5	1	5	1	5	0
Великобритания	1,6	4,1	10	2	6	6	1	–3
Канада	3,1	3,0	3	8	7	5	1	–2
Австралия	1,8	4,0	9	3	8	7	0	–2
Бельгия	2,4	3,4	6	6	9	3	0	–1
Нидерланды	2,3	3,5	7	5	10	2	–	–
Σ	–	–	–	–	–	–	13	–32

Полученный коэффициент Кендалла свидетельствует о слабой обратной зависимости между приростом валового национального продукта и индексом цен рассмотренной совокупности развитых стран.

В случае связных рангов скорректированная формула рангового коэффициента корреляции Кендалла примет вид:

$$\tau_{xy} = \frac{S}{\sqrt{\left[\frac{n(n-1)}{2} - V_x\right] \cdot \left[\frac{n(n-1)}{2} - V_y\right]}}, \qquad (3.3.5)$$

где $V_x(V_y) = \frac{1}{2} \cdot \sum_{j=1}^{k} t_j \cdot (t_j - 1)$; t — число связных рангов соответственно в ряду x и y.

По сравнению с линейным коэффициентом корреляции коэффициенты Спирмена и Кендалла дают менее точную оценку взаимосвязи показателей. Это объясняется тем, что в процессе их расчета исследователь имеет дело не с самими значениями x и y, а с их рангами. Однако преимуществом данных показателей является простота расчета по сравнению с линейным коэффициентом корреляции, особенно в тех случаях, когда требуется лишь предварительно оценить наличие и тесноту связи между изучаемыми признаками.

При достаточно большом объеме совокупности значения данных коэффициентов имеют следующую зависимость:

$$\tau_{xy} = \frac{2}{3}\rho_{xy}. \quad (3.3.6)$$

3.3.3. Коэффициент конкордации (множественный коэффициент ранговой корреляции)

Коэффициент конкордации (множественный коэффициент ранговой корреляции) применяется для определения тесноты связи между двумя и более количественными или порядковыми качественными признаками. Формула для расчета коэффициента конкордации в случае несвязных рангов имеет вид:

$$W = \frac{12S}{m^2(n^3 - n)}, \quad (3.3.7)$$

где m — количество признаков, между которыми определяется связь; n — число единиц совокупности; S — разность между суммой квадратов рангов и средним квадратом рангов, определяемая по формуле:

$$S = \sum_{i=1}^{n}\left(\sum_{j=1}^{m} R_{ij}\right)^2 - \frac{\left(\sum_{i=1}^{n}\sum_{j=1}^{m} R_{ij}\right)^2}{n}, \quad (3.3.8)$$

где n — число наблюдений ($i = 1, 2, ..., n$); m — число признаков ($j = 1, 2, ..., m$); R_{ij} — ранг, присвоенный i-му значению j-го признака.

Коэффициент конкордации принимает значения [0; 1]. Чем ближе значение коэффициента к единице, тем связь признается более тесной.

Значимость коэффициента конкордации оценивается по χ^2-критерию Пирсона. Расчетное значение $\chi^2_{\text{расч}}$ при отсутствии связных рангов определяется по формуле:

$$\chi^2_{\text{расч}} = \frac{12S}{m \cdot n(n+1)}. \quad (3.3.9)$$

Если расчетное значение критерия $\chi^2_{\text{расч}}$ превышает табличное значение $\chi^2_{\text{табл}}$, при заданном уровне значимости α и числе степеней свободы $\upsilon = n - 1$, то коэффициент конкордации признается значимым.

ПРИМЕР.

По данным основных показателей деятельности крупнейших американских компаний, производящих компьютерное и офисное оборудование, определим коэффициент конкордации.

Таблица 3.3.4

Расчетная таблица для определения коэффициента конкордации

№ п/п	Рыночная стоимость x, млн долл. США	Активы y, млн долл. США	Количество служащих z, чел.	R_x	R_y	R_z	Сумма рангов $R_x + R_y + R_z$	Квадрат суммы рангов $(R_x + R_y + R_z)^2$
1	98 322	81 449	269 465	14	14	14	42	1764
2	65 060	31 749	121 900	13	12	13	38	1444
3	36 052	14 631	37 004	11	10	9	30	900
4	31 829	27 732	91 400	10	11	12	33	1089
5	7101	9693	54 900	7	9	11	27	729
6	41 294	4268	16 160	12	5	6	23	529
7	16 614	4697	21 500	9	6	7	22	484
8	3499	4233	9306	5	4	3	12	144
9	3386	5293	38 300	4	7	10	21	441
10	6242	2039	13 369	6	3	5	14	196
11	14 036	7893	29 901	8	8	8	24	576
12	2636	33 345	10 930	3	13	4	20	400
13	869	1135	5100	2	2	1	5	25
14	452	727	7653	1	1	2	4	16
Σ	–	–	–	–	–	–	315	8737

$$S = \sum_{i=1}^{n}\left(\sum_{j=1}^{m} R_{ij}\right)^2 - \frac{\left(\sum_{i=1}^{n}\sum_{j=1}^{m} R_{ij}\right)^2}{n} = 8737 - \frac{315^2}{14} = 1649{,}5.$$

Тогда коэффициент конкордации составит:

$$W = \frac{12S}{m^2(n^3 - n)} = \frac{12 \cdot 1649{,}5}{3^2(14^3 - 14)} = 0{,}27.$$

Значение коэффициента указывает на слабую прямую связь между рыночной стоимостью, активами и количеством служащих. Это может быть вызвано рядом причин, например мультиколлинеарностью. Если предположить, что полученные результаты исследования случайны, возникает необходимость проверки коэффициента на значимость. Расчетное значение критерия определяем по формуле:

$$\chi^2_{расч} = \frac{12S}{m \cdot n(n+1)} = \frac{12 \cdot 1649{,}5}{3 \cdot 14(14+1)} = 2{,}62.$$

Критическое значение $\chi^2_{табл}$ при $\alpha = 0{,}05$ и числе степеней свободы $\upsilon = n - 1 = 14 - 1 = 13$ равно 22,36. Так как $\chi^2_{расч} < \chi^2_{табл}$, множественный коэффициент ранговой корреляции признается незначимым.

В случае наличия связных рангов формула для расчета коэффициента конкордации примет вид:

$$W = \frac{12S}{m^2(n^3 - n) - m\sum_{j=1}^{m}(t_j^3 - t_j)}, \tag{3.3.10}$$

где t_j — число повторяющихся рангов у j-го признака.

При определении значимости коэффициента конкордации при наличии связных рангов расчетное значение $\chi^2_{расч}$ определяется по формуле:

$$\chi^2_{расч} = \frac{12S}{mn(n+1) - \frac{\sum_{j=1}^{m}(t_j - t)}{n-1}}. \tag{3.3.11}$$

Далее проверка значимости осуществляется в той же последовательности, что и в случае несвязных рангов.

3.3.4. Бисериальный коэффициент корреляции

Бисериальный коэффициент корреляции применяется для определения взаимосвязи двух признаков, один из которых является варьируемым количественным, а другой — качественным альтернативным (принимающим только два значения). Данный коэффициент вычисляется по формуле:

$$r = \frac{|\bar{y}_1 - \bar{y}_2|}{\sigma} \cdot \frac{pq}{z}, \qquad (3.3.12)$$

где $\bar{y}_1$ и $\bar{y}_2$ — средние групповые значения количественного признака, разбитого на две группы в соответствии со значениями качественного признака; p — доля единиц первой группы в общем объеме совокупности; q — доля единиц второй группы в общем объеме совокупности; z — табулированные (табличные) значения Z-распределения в зависимости от p; σ — среднее квадратическое отклонение количественного признака.

При величине бисериального коэффициента меньше 0,3 делают вывод об отсутствии связи между признаками.

ПРИМЕР.

Уровень доходов сотрудников и форма собственности предприятий, на которых они работают, представлены в табл. 3.3.5.

Таблица 3.3.5

Расчетная таблица для расчета бисериального коэффициента корреляции

Форма собственности фирмы	Размер среднемесячной заработной платы, тыс. руб.					Итого
	до 5	5–10	10–15	15–20	20 и выше	
	Середины интервалов					
	2,5	7,5	12,5	17,5	22,5	
Частная	10	17	26	24	19	96
Государственная	26	29	15	11	5	86
Итого	36	46	41	35	24	182

Поскольку количественный признак представлен в виде интервального ряда распределения, сначала необходимо определить середины интервалов. Затем рассчитать средние значения размера заработной платы по формам собственности предприятий:

$$\bar{y}_1 = \frac{2{,}5\cdot 10 + 7{,}5\cdot 17 + 12{,}5\cdot 26 + 17{,}5\cdot 24 + 22{,}5\cdot 19}{96} = 13{,}802;$$

$$\bar{y}_2 = \frac{2{,}5\cdot 26 + 7{,}5\cdot 29 + 12{,}5\cdot 15 + 17{,}5\cdot 11 + 22{,}5\cdot 5}{86} = 9{,}012.$$

Для определения величины среднего квадратического отклонения предварительно необходимо рассчитать общее среднее значение количественного признака:

$$\bar{y}_{общ} = \frac{2{,}5 \cdot 36 + 7{,}5 \cdot 46 + 12{,}5 \cdot 41 + 17{,}5 \cdot 35 + 22{,}5 \cdot 24}{182} = 11{,}538.$$

Среднее квадратическое отклонение определяем по взвешенной формуле, так как данные сгруппированы и представлены в виде интервального ряда распределения:

$$\sigma = \sqrt{\frac{(2{,}5 - 11{,}538)^2 \cdot 36 + (7{,}5 - 11{,}538)^2 \cdot 46}{182} +}$$

$$+ \sqrt{\frac{(12{,}5 - 11{,}538)^2 \cdot 41 + (17{,}5 - 11{,}538)^2 \cdot 35 + (22{,}5 - 11{,}538)^2 \cdot 24}{182}} = 6{,}57.$$

Доля единиц первой группы в общем объеме совокупности:

$$p = \frac{96}{182} = 0{,}527.$$

Доля единиц второй группы в общем объеме совокупности:

$$q = \frac{86}{182} = 0{,}473,$$

$$z_{табл} = 0{,}3977.$$

Подставив необходимые величины в формулу для расчета бисериального коэффициента корреляции, получим:

$$r = \frac{|13{,}802 - 9{,}012|}{6{,}57} \cdot \frac{0{,}527 \cdot 0{,}473}{0{,}3977} = 0{,}456.$$

Следовательно, зависимость между формой собственности предприятий и уровнем заработной платы работающих на данных предприятиях сотрудников можно признать умеренной.

3.3.5. Методы изучения взаимосвязи между качественными признаками

Коэффициенты ассоциации и контингенции. Для определения тесноты связи между качественными признаками, каждый из которых состоит только из двух вариантов, применяются коэффициенты ассоциации и контингенции.

Для их вычисления строится таблица сопряженности [2 × 2], которая показывает связь между двумя явлениями, каждое из которых состоит из двух качественно отличных друг от друга значений признака (например, изделие годное или бракованное).

Таблица 3.3.6

Таблица взаимной сопряженности для вычисления коэффициентов ассоциации и контингенции

x \ y	I	II	Всего
I	n_{11}	n_{12}	$n_{11} + n_{12}$
II	n_{21}	n_{22}	$n_{21} + n_{22}$
Всего	$n_{11} + n_{21}$	$n_{12} + n_{22}$	$n_{11} + n_{12} + n_{21} + n_{22}$

Коэффициент ассоциации вычисляется по формуле:

$$K_a = \frac{n_{11}n_{22} - n_{12}n_{21}}{n_{11}n_{22} + n_{12}n_{21}}, \qquad (3.3.13)$$

где n_{11} — число единиц, имеющих значения x_1 и y_1; n_{12} — число единиц, имеющих значения x_1 и y_2; n_{21} — число единиц, имеющих значения x_2 и y_1; n_{22} — число единиц, имеющих значения x_2 и y_2.

Коэффициент ассоциации принимает значения в интервале [–1; 1]. При равенстве коэффициента ассоциации нулю говорят об отсутствии связи между изучаемыми признаками.

Коэффициент контингенции считается более надежной мерой связи и определяется по формуле:

$$K_k = \frac{n_{11}n_{22} - n_{12}n_{21}}{\sqrt{(n_{11} + n_{12})\cdot(n_{22} + n_{21})\cdot(n_{11} + n_{21})\cdot(n_{12} + n_{22})}}. \qquad (3.3.14)$$

Коэффициент контингенции принимает значения в интервале [–1; 1]. Коэффициент контингенции принимает нулевое значение при отсутствии связи между изучаемыми показателями и равен единице — при функциональной зависимости между ними.

Коэффициент контингенции всегда меньше коэффициента ассоциации. Связь считается подтвержденной, если $K_a > 0{,}5$ или $K_k > 0{,}3$.

ПРИМЕР.

Компанию, доставляющую товары почтой, интересует, одинаков ли уровень заказов (процент отправленных по почте каталогов, которые стали основой заказа) в разных регионах. В табл. 3.3.7 приведены последние данные по регионам о количестве отправленных по почте каталогов, которые стали причиной заказа и которые не стали причиной заказа.

Таблица 3.3.7

Зависимость количества сделанных и не сделанных заказов от территориальной принадлежности клиентов

	Количество клиентов	Из них	
		Заказ сделан	Заказ не сделан
Восток	23 039	926	22 113
Запад	10 969	352	10 617
Итого	34 008	1278	32 730

$$K_a = \frac{926 \cdot 10617 - 22113 \cdot 352}{926 \cdot 10617 + 22113 \cdot 352} = 0{,}116,$$

$$K_k = \frac{926 \cdot 10617 - 22113 \cdot 352}{\sqrt{(926 + 22113) \cdot (22113 + 10617) \cdot (10617 + 352) \cdot (352 + 926)}} = 0{,}02$$

Таким образом, связь между количеством сделанных и не сделанных заказов в регионах не зависит о территориальной принадлежности клиентов фирмы.

Коэффициенты взаимной сопряженности. Когда каждый из качественных признаков состоит более чем из двух групп, то для определения тесноты связи возможно применение коэффициентов взаимной сопряженности Пирсона и Чупрова. Эти коэффициенты вычисляются по следующим формулам:

$$K_{\text{П}} = \sqrt{\frac{\varphi^2}{1+\varphi^2}}; \qquad (3.3.15)$$

$$K_{\text{Ч}} = \sqrt{\frac{\varphi^2}{\sqrt{(k_1 - 1) \cdot (k_2 - 1)}}}, \qquad (3.3.16)$$

где k_1 — число значений (групп) первого признака; k_2 — число значений (групп) второго признака; φ^2 — показатель взаимной сопряженности.

Показатель взаимной сопряженности φ^2 определяется как сумма отношений квадратов частот каждой клетки таблицы к произведению итоговых частот соответствующего столбца и строки минус единица:

$$\varphi^2 = \sum \frac{\sum \frac{n_{xy}^2}{n_x}}{n_y} - 1. \qquad (3.3.17)$$

Коэффициенты взаимной сопряженности принимают значения в интервале [0; 1]. Равенство коэффициентов нулю означает отсутствие связи, равенство единице — полную взаимосвязь.

Таблица 3.3.8

Вспомогательная таблица для расчета коэффициентов взаимной сопряженности Пирсона и Чупрова

x \ y	I	II	III	...	Всего
I	$n_{x_1y_1}$	$n_{x_1y_2}$	$n_{x_1y_3}$	...	n_{x_1}
II	$n_{x_2y_1}$	$n_{x_2y_2}$	$n_{x_2y_3}$	...	n_{x_1}
III	$n_{x_3y_1}$	$n_{x_3y_2}$	$n_{x_3y_3}$	...	n_{x_1}
...	...	...	...	...	...
Итого	n_{y_1}	n_{y_2}	n_{y_3}	...	n

ПРИМЕР.

С помощью коэффициентов взаимной сопряженности исследуем связь между себестоимостью продукции и накладными расходами на ее реализацию.

Таблица 3.3.9

Зависимость между себестоимостью продукции и накладными расходами на ее реализацию

Накладные расходы	Себестоимость			Итого
	Низкая	Средняя	Высокая	
Низкие	19	12	9	40
Средние	7	18	15	40
Высокие	4	10	26	40
Итого	30	40	50	120

Рассчитаем показатель взаимной сопряженности по формуле (3.3.17):

$$\varphi^2 = \frac{\frac{19^2}{30}+\frac{12^2}{40}+\frac{9^2}{50}}{40} + \frac{\frac{7^2}{30}+\frac{18^2}{40}+\frac{15^2}{50}}{40} + \frac{\frac{4^2}{30}+\frac{10^2}{40}+\frac{26^2}{50}}{40} - 1 =$$

$$= 0{,}431 + 0{,}356 + 0{,}414 - 1 = 1{,}183 - 1 = 0{,}183.$$

Тогда коэффициенты взаимной сопряженности согласно формулам (3.3.15) и (3.3.16) составят:

$$K_{\text{П}} = \sqrt{\frac{0{,}183}{1{,}183}} = \sqrt{0{,}155} = 0{,}39,\ K_{\text{Ч}} = \sqrt{\frac{0{,}183}{\sqrt{2 \cdot 2}}} = 0{,}21.$$

Таким образом, связь между себестоимостью продукции и расходами на ее реализацию можно признать слабой.

КОНТРОЛЬНЫЕ ВОПРОСЫ

1. Какие цели преследует исследователь при сборе нечисловой информации? Приведите практические примеры.
2. Приведите преимущества и недостатки использования непараметрических методов связи по сравнению с параметрическими.
3. Какие из непараметрических коэффициентов связи могут применяться и к количественным, и к качественным признакам?
4. Какие показатели используются для оценки тесноты связи между качественными альтернативными признаками?
5. Что представляет собой таблица сопряженности? Какие размерности таблиц сопряженности вы знаете? В каких случаях она используется?
6. Что такое показатель взаимной сопряженности? Для чего следует производить его расчет?
7. Что понимается под процедурой ранжирования в статистике и как она выполняется? Дайте понятие ранга. В чем отличие связных и несвязных рангов?
8. В каких случаях для определения тесноты связи используются ранговые коэффициенты? Какая между ними существует взаимосвязь? В каких пределах должны колебаться полученные значения?
9. Кратко опишите этапы реализации расчета коэффициента Спирмена и Кендалла.
10. В чем отличие коэффициента конкордации от ранговых коэффициентов связи Спирмена и Кендалла? Какие цели преследует проверка на значимость данного показателя? Распределение какого рода при этом необходимо использовать?

Раздел 4

АНАЛИЗ ДИНАМИКИ СОЦИАЛЬНО-ЭКОНОМИЧЕСКИХ ЯВЛЕНИЙ

Глава 4.1

ПОНЯТИЕ И КЛАССИФИКАЦИЯ РЯДОВ ДИНАМИКИ

- Понятие о рядах динамики. Виды рядов динамики
- Сопоставимость уровней и смыкание рядов динамики

4.1.1. Понятие о рядах динамики. Виды рядов динамики

Процессы и явления социально-экономической жизни общества находятся в постоянном движении и изменении, которые в статистике называют динамикой. Для отображения динамики строят особые ряды статистических показателей — **ряды динамики**.

Ряд динамики (от англ. *time series* — временной ряд) — это последовательность изменяющихся во времени значений показателя, расположенных в хронологическом порядке.

Составляющими элементами ряда динамики являются: значения показателя — **уровни ряда** и **периоды** (годы, кварталы, месяцы, сутки) **или моменты** времени (даты, часы, начало или конец года, квартала, месяца, дня). Уровни ряда обычно обозначаются через y, моменты или периоды времени, к которым относятся уровни, — через t.

Классификация рядов динамики представлена на рис. 4.1.1.

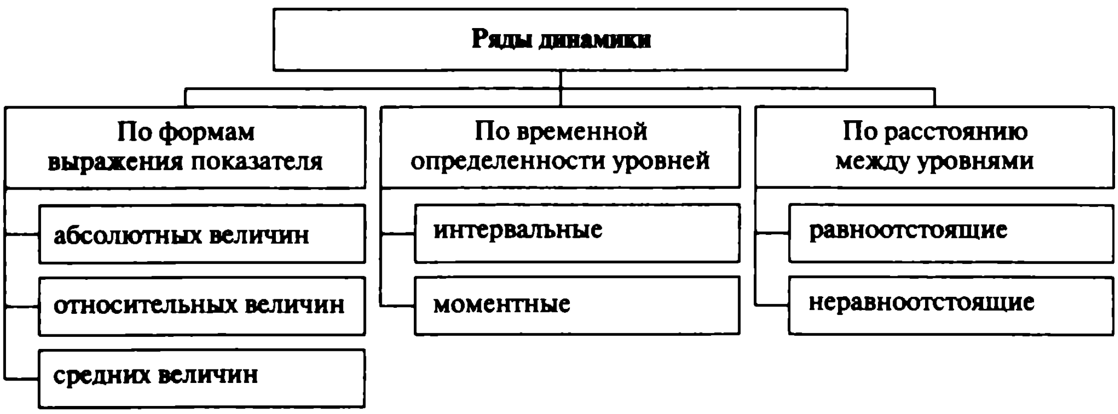


Рис. 4.1.1. Классификация рядов динамики

Ряды динамики абсолютных величин рассматриваются как исходные, а ряды относительных и средних величин — как производные. Ряд динамики **абсолютных величин** представлен в 1-й строке табл. 4.1.1, ряд **относительных величин** — во 2-й строке, ряд **средних величин** — в 3-й строке.

Таблица 4.1.1

Число построенных квартир и их средний размер в Российской Федерации

	2001 г.	2002 г.	2003 г.	2004 г.	2005 г.
Число квартир, тыс.	382,0	396,0	427,0	477,0	515,0
Число однокомнатных квартир от общего количества, %	20,0	21,0	23,0	25,0	28,0
Средний размер квартир, м^2 общей площади	83,1	85,3	85,4	86,0	84,5

Уровни **моментных** рядов динамики характеризуют явление по состоянию на определенный момент времени (табл. 4.1.2).

Таблица 4.1.2

Изменение средних фактических экспортных цен Российской Федерации на отдельные товары в 2005 г.

Вид товара	1 января	1 февраля	1 марта	1 апреля	1 мая	1 июня
Нефть сырая, млн руб.	100,2	110	123,1	132,1	130,2	137,9
Бензин автомобильный, млн руб.	96,5	97,6	105,3	107,3	107,7	110,5
Газ природный, млн руб.	105,5	102,8	100,9	115,7	113,2	111,8

Уровни моментных рядов абсолютных величин нельзя суммировать, так как в них повторяются одни и те же единицы совокупности.

Ряд, в котором уровни характеризуют результат, накопленный или вновь произведенный за определенный интервал времени, называется **интервальным** (табл. 4.1.3).

Таблица 4.1.3

Динамика браков и разводов населения Российской Федерации, тыс.

Годы	Браки	Разводы
2000	897,3	627,7
2001	1001,6	763,5
2002	1019,8	853,6

Окончание табл. 4.1.3

Годы	Браки	Разводы
2003	1091,8	798,8
2004	979,7	635,8
2005	1066,4	604,9

В отличие от моментных рядов, сумма уровней интервального ряда является реальным показателем, например динамика браков и разводов за 2000–2005 гг.

Ряды динамики следующих друг за другом периодов или повторяющихся через равные промежутки дат называются **равноотстоящими** (табл. 4.1.2 и 4.1.3). Если же в рядах даются прерывающиеся периоды или неравномерные промежутки между датами, то ряды называются **неравноотстоящими** (табл. 4.1.4).

Таблица 4.1.4

Число предприятий и организаций в Российской Федерации

	2001 г.	2002 г.	2004 г.	2005 г.	2006 г.
Число предприятий и организаций, тыс.	3347	3594	4150	4417	4767

4.1.2. Сопоставимость уровней и смыкание рядов динамики

Важнейшим условием правильного построения рядов динамики является сопоставимость входящих в него уровней. Уровни ряда динамики должны быть сопоставимы по методологии учета и расчета показателей территориальным границам и кругу охватываемых объектов, единицам измерения и другим характеристикам. Данное условие решается в процессе сбора и обработки данных либо путем их пересчета.

Основными причинами несопоставимости уровней ряда динамики являются:

- различие в единицах измерения и единицах счета;
- различие в методологии учета или расчета показателей;
- изменение круга охватываемых объектов вследствие перехода ряда объектов из одного подчинения в другое;
- изменение территориальных границ областей, районов, округов.

Следует учитывать, что в процессе развития во времени вначале происходят количественные изменения явлений, а затем на определенных ступенях совершаются качественные скачки, приводящие к изменению закономерностей явления. Поэтому научный подход к изучению динамики заключается в том, что ряды, охватывающие большие периоды времени, разделяют на части, которые включают однокачественный период развития совокупности и характеризуются одной закономерностью развития.

В тех случаях, когда уровни ряда динамики оказываются несопоставимы между собой, их необходимо привести к сопоставимому виду, применяя прием **смыкания рядов динамики**. Под смыканием понимают объединение в один ряд (более длинный) двух или нескольких рядов динамики с несопоставимыми уровнями. Для смыкания необходимо, чтобы для одного из периодов (переходного) имелись данные, исчисленные по разной методологии (или в разных границах). Оба способа рассмотрим на примере, два основных способа смыкания — абсолютный и относительный.

ПРИМЕР.

Имеются данные о числе малых предприятий одного из городов Российской Федерации за 1999–2006 гг. (табл. 4.1.5).

Таблица 4.1.5

Динамика малых предприятий одного из городов Российской Федерации

	1999 г.	2000 г.	2001 г.	2002 г.	2003 г.	2004 г.	2005 г.	2006 г.
Количество малых предприятий: По старой методологии отнесения к кругу малых предприятий	54	89	125	–	–	–	–	–
По новой методологии отнесения к кругу малых предприятий	–	–	90	138	169	194	206	220
Сомкнутый ряд динамики абсолютных величин (по новой методологии)	39	64	90	138	169	194	206	220
Сомкнутый ряд динамики относительных величин, в % к 2001 г.	43	71	100	153	188	215	229	244

Абсолютный способ заключается в том, что соотношением новых и прежних уровней момента и периода времени определяется коэффициент приведения, на который умножаются прежние уровни. Тем самым получается один приведенный ряд. В примере коэффициент приведения для 2001 г. равняется 90 : 125 = 0,72. Умножая на него предыдущие уровни, получим сопоставимый ряд динамики (третья строка табл. 4.1.5).

Относительный способ смыкания рядов заключается в том, что каждый уровень года, в котором произошли изменения, принимается за 100%, относительно которых измеряются предыдущие и новые уровни (в примере за 100% принят уровень 2001 г.). В результате получаем сомкнутый ряд динамики, который показан в последней строке табл. 4.1.5.

При сравнительном анализе развития во времени экономических показателей стран, административных и территориальных районов ряды динамики **приводятся к одному основанию** — к одному периоду или моменту времени, уровень которого принимается за базу сравнения, а все остальные уровни выражаются в виде коэффициентов или в процентах по отношению к нему.

КОНТРОЛЬНЫЕ ВОПРОСЫ

1. Дайте определение ряда динамики социально-экономических явлений.
2. Какие вы знаете виды рядов динамики?
3. В чем основное отличие моментных рядов динамики от интервальных?
4. Что означает несопоставимость уровней рядов динамики и из-за чего она возникает?
5. Что такое смыкание рядов динамики? Каковы основные способы смыкания и условия из применения?

Глава 4.2
АНАЛИТИЧЕСКИЕ ПОКАЗАТЕЛИ РЯДА ДИНАМИКИ

- Индивидуальные аналитические показатели динамики
- Средние аналитические показатели в рядах динамики

4.2.1. Индивидуальные аналитические показатели динамики

Для количественной оценки развития явлений во времени используются аналитические показатели динамики. Они характеризуют в абсолютном или относительном выражении изменение значения показателя в одном уровне рассматриваемого ряда динамики по сравнению с другим уровнем. При этом сравниваемый уровень называют **текущим** или **отчетным**, а уровень, с которым происходит сравнение, — **базисным**.

Поскольку время развивается всегда однонаправленно, каждый уровень ряда динамики можно сравнить с предшествующим или с другим, более ранним уровнем, выбранным за базу сравнения (как правило, первым в ряду). В первом случае рассчитанные аналитические показатели называют **цепными**, а при сравнении всех уровней ряда с фиксированным уровнем — **базисными**.

Базисные показатели динамики характеризуют окончательный результат всех изменений в уровнях ряда от периода, к которому относится базисный уровень, до данного (i-го) периода. Цепные показатели динамики характеризуют интенсивность изменения уровня от периода к периоду (или от даты к дате) в пределах изучаемого промежутка времени.

К основным аналитическим показателям динамики относятся: абсолютный прирост, темп роста и прироста, абсолютное значение одного процента прироста.

Абсолютный прирост (Δ) характеризует размер увеличения (или уменьшения) уровня ряда за определенный промежуток времени. Он равен разности двух сравниваемых уровней и выражает абсолютную скорость роста. Показатели абсолютного прироста имеют те же единицы измерения, что и уровни динамического ряда.

Цепные абсолютные приросты рассчитываются по формуле:

$$\Delta_{iц} = Y_i - Y_{i-1}. \qquad (4.2.1)$$

Для *базисного* абсолютного прироста, определяемого относительно начального уровня ряда:

$$\Delta_{iб} = Y_i - Y_1. \qquad (4.2.2)$$

Если в формуле (4.2.2) уровень Y_1 заменить на уровень ряда динамики, принятый за базу сравнения — Y_0, получим формулу в общем виде:

$$\Delta_{iб} = Y_i - Y_0. \qquad (4.2.3)$$

Интенсивность изменения уровня ряда оценивают при помощи относительных показателей — коэффициента роста и темпа роста.

Коэффициент роста ($K_р$) показывает, во сколько раз данный уровень ряда больше предшествующего или базисного уровня (если этот коэффициент больше единицы) или какую часть от размера предшествующего или базисного уровня он составляет (если коэффициент меньше единицы).

Темп роста ($T_р$) показывает, сколько процентов составляет величина показателя в текущем периоде по сравнению с его величиной в предшествующем или базисном периоде.

Цепной темп роста рассчитывается по формуле:

$$T_{р_{iц}} = \frac{y_i}{y_{i-1}} 100\%. \qquad (4.2.4)$$

Базисный темп роста может быть представлен в виде:

$$T_{р_{iб}} = \frac{y_i}{y_0} 100\%, \qquad (4.2.5)$$

где y_1 — уровень ряда динамики, принятый за базу сравнения.

Темп роста всегда положителен. Если темп роста равен 100%, то значение уровня не изменилось, если больше 100%, то значение уровня повысилось, а если меньше 100% — понизилось.

Темп прироста показывает, на сколько процентов изменился сравниваемый уровень по отношению к уровню, принятому за базу сравнения. Темп прироста рассчитывается как отношение абсолютного прироста к уровню, принятому за базу сравнения. Темп прироста может быть больше, меньше нуля и равным нулю.

Цепной темп прироста рассчитывается по формуле:

$$T_{пр_ц} = \frac{\Delta_{iц}}{y_{i-1}} = \frac{y_i}{y_{i-1}} 100\% - 100\% = T_{р_ц} - 100\%, \qquad (4.2.6)$$

где $T_{пр_ц}$ — соответствующий цепной темп роста; $\Delta_{iц}$ — цепной асболютный прирост.

Базисный темп прироста равен отношению базисного абсолютного прироста к уровню ряда, принятому за базу сравнения:

$$T_{\text{пр}_б} = \frac{\Delta_{iб}}{y_0} 100\% - 100\% = T_{\text{р}_б} - 100\%, \qquad (4.2.7)$$

где $T_{\text{пр}_б}$ — базисный темп роста; $\Delta_{iб}$ — базисный абсолютный прирост.

Сравнение абсолютного прироста и темпа прироста за одни и те же периоды времени показывает, что замедление темпов прироста не всегда сопровождается уменьшением абсолютных приростов. Поэтому на практике часто рассчитывают **абсолютное значение одного процента прироста**. Оно представляет собой одну сотую часть базисного уровня и в то же время — отношение абсолютного прироста к соответствующему темпу прироста:

$$|\%| = \frac{\Delta_{ц}}{T_{\text{пр}_ц} y\%} = \frac{y_i - y_{i-1}}{\dfrac{y_i - y_{i-1}}{y_{i-1}} \cdot 100} = \frac{y_{i-1}}{100} = 0{,}01 y_{i-1}. \qquad (4.2.8)$$

ПРИМЕР.

Имеются данные о динамике производства тканей в одном из регионов за 2002–2006 гг. (табл. 4.2.1).

Таблица 4.2.1

Динамика производства тканей в одном из регионов за 2000–2006 гг.

Годы	Производство тканей, млн м²	Абсолютный прирост, млн м² по сравнению		Темп роста, % по сравнению		Темп прироста, % по сравнению		Абсолютное значение 1% прироста, млн м²
		с предшествующим годом	с 2002 г.	с предшествующим годом	с 2002 г.	с предшествующим годом	с 2002 г.	
А	1	2	3	4	5	6	7	8
2002	256	–	–	–	100,0	–	–	–
2003	267	11	11	104,3	104,3	4,3	4,3	2,56
2004	279	12	23	104,5	109,0	4,5	9,0	2,67
2005	291	12	35	104,3	113,7	4,3	13,7	2,79
2006	305	14	49	104,8	119,1	4,8	19,1	2,91

1. Рассчитаем базисные показатели. Абсолютный прирост по сравнению с 2002 г. составит:

- в 2003 г.:

$$\Delta y_{i_б} = 267 - 256 = 11 \text{ млн м}^2;$$

- в 2004 г.:

$$\Delta y_{i_б} = 279 - 256 = 23 \text{ млн м}^2 \text{ и т. д.}$$

Рассчитаем цепные показатели. Абсолютный прирост составит:

- в 2003 г. по сравнению с 2002 г.:

$$\Delta y_{i_ц} = 267 - 256 = 11 \text{ млн м}^2;$$

- в 2004 г. по сравнению с 2003 г.:

$$\Delta y_{i_ц} = 279 - 267 = 12 \text{ млн м}^2 \text{ и т. д.}$$

Интенсивность изменения уровней ряда динамики оценивается отношением текущего уровня к предыдущему или базисному.

2. Базисный темп роста составит:

- в 2003 г.:

$$T_{p_{i_б}} = \frac{267}{256}100\% = 104{,}3\%;$$

- в 2004 г.:

$$T_{p_{i_б}} = \frac{279}{256}100\% = 109\% \text{ и т. д.}$$

Рассчитаем цепные показатели темпа роста:

- в 2003 г. по сравнению с 2002 г.:

$$T_{p_{i_ц}} = \frac{267}{256}100\% = 104{,}3\%;$$

- в 2004 г. по сравнению с 2003 г.:

$$T_{p_{i_ц}} = \frac{279}{267}100\% = 104{,}5\%;$$

- в 2005 г. по сравнению с 2004 г.:

$$T_{p_{i_ц}} = \frac{291}{279} 100\% = 104{,}3\% \text{ и т. д.}$$

3. Рассчитаем темпы прироста.

Для 2003 г. темп прироста по сравнению с 2002 г. составит:

$$T_{пр} = 104{,}3\% - 100\% = 4{,}3\%.$$

В 2004 г. по сравнению с 2002 г.:

$$T_{пр} = 109\% - 100\% = 9\% \text{ и т. д.}$$

4. Рассчитаем абсолютное значение одного процента прироста:

$$|\%|_{2003} = 256 \cdot 0{,}01 = 2{,}56 \text{ млн м}^2,$$

$$|\%|_{2004} = 267 \cdot 0{,}01 = 2{,}67 \text{ млн м}^2 \text{ и т. д.}$$

4.2.2. Средние аналитические показатели в рядах динамики

Каждый ряд динамики можно рассматривать как некую совокупность, состоящую из n меняющихся во времени показателей, которые можно обобщать в виде средних величин. Для обобщения данных по рядам динамики рассчитываются средний уровень ряда, средний абсолютный прирост, средний темп роста и прироста.

Определение **среднего уровня** ($\bar{y}$) зависит от вида ряда динамики.

Для *интервальных равноотстоящих рядов* средний уровень находится по формуле простой средней арифметической:

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}, \qquad (4.2.9)$$

где n — число уровней, или длина ряда.

ПРИМЕР.

В табл. 4.2.1 приведен интервальный ряд динамики с равноотстоящими уровнями. Рассчитаем по этим данным средний уровень ряда по формуле средней арифметической простой. Он равен $\bar{y} = \frac{1398}{5} = 279{,}6$ млн м^2, т. е. в среднем в период с 2003 по 2006 гг. в регионе ежегодно производилось 279,6 млн м^2 тканей.

Для *интервальных неравноотстоящих рядов* средний уровень находится по формуле взвешенной средней арифметической:

$$y = \frac{\sum_{i=1}^{n} y_i t_i}{\sum_{i=1}^{n} t_i}, \qquad (4.2.10)$$

где t_i — продолжительность интервалов времени между уровнями (число периодов времени, в течение которых значение уровня не изменялось).

Средний уровень моментного ряда динамики рассчитывается по формуле **средней хронологической**. Для *моментного равноотстоящего ряда* динамики при расчете среднего уровня используют формулу средней хронологической простой:

$$\bar{y} = \frac{\frac{y_1 + y_2}{2} + \frac{y_2 + y_3}{2} + \frac{y_3 + y_4}{2} + \ldots + \frac{y_{n-1} + y_n}{2}}{n-1} =$$

$$= \frac{\frac{y_1}{2} + y_2 + y_3 + \ldots + y_{n-1} + \frac{y_n}{2}}{n-1} \qquad (4.2.11)$$

или

$$\bar{y} = \frac{\frac{y_1 + y_n}{2} + \sum_{i=2}^{n-1} y_i}{n-1}.$$

ПРИМЕР.

Найдем средний уровень ряда динамики по данным о численности работников фирмы на 1-е число каждого месяца в I квартале 2006 г. (чел.):

1/I	1/II	1/III	1/IV
420	490	418	450

Это моментный ряд с равноотстоящими уровнями, поэтому среднемесячная численность работников фирмы за I квартал 2006 г. составила:

$$\bar{y} = \frac{\frac{420}{2} + 490 + 418 + \frac{450}{2}}{3} = \frac{1343}{3} \approx 448 \text{ чел.}$$

Средний уровень моментных рядов динамики с неравноотстоящими уровнями определяется по формуле средней хронологической взвешенной:

$$\bar{y} = \frac{(y_1 + y_2)t_1 + (y_2 + y_3)t_2 + \ldots + (y_{n-1} + y_n)t_{n-1}}{2(t_1 + t_2 + \ldots + t_{n-1})} = \frac{\sum_{i=1}^{n-1}(y_i + y_{i+1})t_i}{2\sum_{i=1}^{n-1}t_i}, \quad (4.2.12)$$

где t — продолжительность интервала времени между соседними уровнями.

ПРИМЕР.

Известна списочная численность рабочих организаций на некоторые даты 2006 г. (чел.). Ряд динамики имеет неравноотстоящие уровни во времени:

1/I	1/III	1/VI	1/IX	1/I-2007
420	418	520	550	620

Среднегодовая численность работников за 2006 г. (по формуле 4.2.12) составит:

$$\bar{y} = \frac{(420 + 418) \cdot 2 + (418 + 520) \cdot 3 + (520 + 550) \cdot 3 + (550 + 620) \cdot 4}{2 \cdot (2 + 3 + 3 + 4)} =$$

$$= \frac{12\,380}{24} = 515{,}83 \approx 516 \text{ чел.}$$

Обобщающим показателем абсолютной скорости изменения явления во времени является **средний абсолютный прирост** за весь период, ограничивающий ряд динамики. Скоростью в данном случае будем называть прирост (уменьшение) в единицу времени. Для его определения используется формула средней арифметической простой:

$$\bar{\Delta} = \frac{\sum \Delta_{i_{ц}}}{n - 1}. \quad (4.2.13)$$

Подставив в числитель выражение для цепных абсолютных приростов, получим более удобную форму записи для среднего абсолютного прироста:

$$\bar{\Delta} = \frac{y_2 - y_1 + y_3 - y_2 + \ldots + (y_n - y_{n-1})}{n - 1} = \frac{y_n - y_1}{n - 1}, \quad (4.2.14)$$

где y_n и y_1 — соответственно конечный и начальный уровни ряда динамики.

ПРИМЕР.

По данным табл. 4.2.1 определим средний абсолютный прирост производства тканей за период 2002–2006 гг. Он составит:

$$\frac{305-256}{4} = 12{,}25 \text{ млн м}^2.$$

Это значит, что в среднем ежегодно производство тканей увеличивалось на 12,25 млн $м^2$.

Сводной обобщающей характеристикой интенсивности изменения уровней ряда динамики служит **средний темп роста**. Он показывает, сколько в среднем процентов последующий уровень составляет от предыдущего в течение всего периода наблюдения.

Средний темп (коэффициент) роста рассчитывается по формуле среднего геометрического из цепных коэффициентов роста:

$$\overline{T_{\text{р}}} = \sqrt[n-1]{K_{2/1} \cdot K_{3/2} \cdot \ldots \cdot K_{n/n-1}} = \sqrt[n-1]{\prod K_{\text{ц}}}. \tag{4.2.15}$$

Выразив цепные коэффициенты (темпы) роста через соответствующие уровни ряда, получим:

$$\overline{T_{\text{р}}} = \sqrt[n-1]{\frac{y_2}{y_1} \cdot \frac{y_3}{y_2} \cdot \frac{y_4}{y_3} \ldots \frac{y_n}{y_{n-1}} 100\%} = \sqrt[n-1]{\frac{y_n}{y_1} 100\%}. \tag{4.2.16}$$

ПРИМЕР.

По данным табл. 4.2.1 рассчитаем средний темп роста производства тканей за период 2002–2006 гг. по формуле:

$$\overline{T_{\text{р}}} = \sqrt[4]{1{,}043 \cdot 1{,}045 \cdot 1{,}043 \cdot 1{,}048} = 1{,}045, \text{ или } 104{,}5\%$$

или по формуле

$$\overline{T_{\text{р}}} = \sqrt[4]{\frac{305}{256}} = 1{,}045, \text{ или } 104{,}5\%.$$

Средний темп роста равен 104,5%, следовательно, в среднем ежегодно производство тканей увеличивалось в 1,045 раза.

Когда приходится производить расчет средних темпов роста по периодам различной продолжительности (неравноотстоящие уровни), то используют среднюю геометрическую, взвешенную по продолжительности периодов. Формула средней геометрической взвешенной будет иметь вид:

$$\overline{T}_{\text{р}} = \sqrt[\sum t]{(K_{2/1})^{t_1} \cdot (K_{3/2})^{t_1} \cdot \ldots \cdot (K_{n/n-1})^{t_{n-1}}}, \qquad (4.2.17)$$

где t_i — интервал времени, в течение которого сохраняется данный темп роста.

Средний темп прироста не может быть определен непосредственно на основании цепных темпов прироста или показателей среднего абсолютного прироста. Для его вычисления необходимо сначала найти средний темп роста, а затем его уменьшить на 100%:

$$\overline{T}_{\text{пр}} = \overline{T}_{\text{р}} - 100\%. \qquad (4.2.18)$$

ПРИМЕР.

По данным табл. 4.2.1 был рассчитан средний темп роста производства тканей за 2002–2006 гг., равный 104,5%, отсюда средний темп прироста будет равен:

$$\overline{T}_{\text{р}} = 104{,}5\% - 100\% = 4{,}5\%.$$

В среднем ежегодный прирост производства тканей составил 4,5%.

КОНТРОЛЬНЫЕ ВОПРОСЫ

1. Что характеризует: а) абсолютный прирост; б) темп роста; в) темп прироста? В чем разница между этими показателями?
2. Какой показатель является обобщающим показателем абсолютной скорости изменения явления во времени?
3. Как проводится расчет среднего уровня в интервальных рядах динамики?
4. При помощи какого вида средних величин рассчитывается средний уровень в моментных рядах динамики?
5. Что характеризуют и как интерпретируются показатели среднего темпа роста и среднего темпа прироста?

Глава 4.3

МЕТОДЫ АНАЛИЗА ОСНОВНОЙ ТЕНДЕНЦИИ В РЯДАХ ДИНАМИКИ

- Тенденция ряда динамики и методы ее выявления
- Метод укрупнения интервалов
- Метод скользящих средних
- Метод аналитического выравнивания

4.3.1. Тенденция ряда динамики и методы ее выявления

Важной задачей статистики при анализе рядов динамики является определение основной тенденции развития, присущей тому или иному ряду динамики.

Под основной **тенденцией** развития ряда динамики понимают изменение, определяющее общее направление развития. Это — систематическая составляющая долговременного действия. В некоторых случаях общая тенденция ярко прослеживается в динамике рассматриваемого показателя, в других случаях она может не просматриваться из-за ощутимых случайных колебаний. Например, в отдельные моменты времени сильные колебания розничных цен могут заслонить наличие тенденции к росту или снижению этого показателя.

Для выявления основной тенденции развития ряда динамики в статистике применяются две группы методов:

1) **сглаживание или механическое выравнивание** отдельных уровней ряда динамики с использованием фактических значений соседних уровней;

2) **аналитическое выравнивание** с применением кривой, проведенной между конкретными уровнями таким образом, чтобы она отражала тенденцию, присущую ряду, и одновременно освободила его от незначительных колебаний.

Суть различных приемов сглаживания сводится к замене фактических уровней ряда расчетными уровнями, которые в меньшей степени подвержены колебаниям. Это способствует более четкому проявлению тенденции развития.

Аналитическое выравнивание временных рядов сводится к нахождению определенной модели, которая математически описывает тенденцию развития изучаемого явления во времени.

4.3.2. Метод укрупнения интервалов

Данный метод основан на укрупнении периодов времени, к которым относятся уровни. Например, ряд недельных данных можно преобразовать в ряд помесячной динамики, ряд квартальных данных заменить годовыми уровнями. Уровни нового ряда могут быть получены путем суммирования уровней исходного ряда либо могут представлять средние уровни.

ПРИМЕР.

Визуальный анализ данных о динамике продажи продукции в торговой сети, представленный в табл. 4.3.1, не позволяет сделать выводы о наличии тенденции в этом временном ряду: в отдельные месяцы (например, в марте, июне, августе и октябре) объем продаж снижался по сравнению с предшествующими периодами, в остальные месяцы – возрастал.

Таблица 4.3.1

Динамика продажи продукции в торговой сети за 2006 г.

Месяц	Продано продукции, тыс. шт.	Месяц	Продано продукции, тыс. шт.
январь	12,8	июль	15,7
февраль	13,4	август	16,4
март	15,9	сентябрь	16,9
апрель	14,7	октябрь	17,1
май	16,2	ноябрь	16,8
июнь	16,5	декабрь	17,5

Для выявления тенденции в исходном временном ряду воспользуемся методом укрупнения интервалов, образовав новый динамический ряд с более крупными временными периодами – кварталами, и рассчитаем средний объем реализации продукции по кварталам (табл. 4.3.2).

Таблица 4.3.2

Среднеквартальный объем продажи продукции в торговой сети за год

Квартал	Среднемесячный объем продаж, тыс. шт.
I	14,03
II	15,80
III	16,33
IV	17,13

По полученным более крупным интервалам четко прослеживается тенденция к возрастанию значений исследуемого признака. Таким образом, можно говорить о тенденции к росту объема продажи продукции в торговой сети.

Применение данного метода ограничивается теми ситуациями, когда исходные данные относятся к дням, неделям или месяцам года, так как более мелкие временные интервалы вносят в значения исследуемого признака большее количество случайных колебаний. Если временные промежутки представляют собой годы, то укрупнение интервалов уже становится неэффективным.

4.3.3. Метод скользящих средних

Метод простой скользящей средней. Сглаживание ряда динамики с помощью скользящей средней заключается в том, что вычисляется средний уровень из определенного числа первых по порядку уровней ряда, затем средний уровень из такого же числа уровней, начиная со второго, далее — начиная с третьего и т. д. Таким образом, при расчете средних уровней они как бы «скользят» по ряду динамики от его начала к концу, каждый раз отбрасывая один уровень вначале и добавляя один следующий. Отсюда название — скользящая средняя.

Каждое звено скользящей средней — это средний уровень за соответствующий период, который относится к середине выбранного периода, если число уровней ряда динамики нечетное.

Этапы реализации метода скользящей средней:

1. Определяют интервал сглаживания, т. е. число входящих в него уровней m ($m < n$), в соответствии со следующим правилом: если необходимо сгладить мелкие, беспорядочные колебания, то интервал сглаживания берут по возможности большим, и, наоборот, интервал сглаживания уменьшают, когда нужно сохранить более мелкие волны и освободиться от периодически повторяющихся колебаний.

2. Вычисляют среднее значение уровней, образующих интервал сглаживания по формуле:

$$\bar{y}_t = \frac{\sum_{i=t-p}^{t+p} y_i}{m}, \qquad (4.3.1)$$

где y_i — фактическое значение i-го уровня; m — число уровней, входящих в интервал сглаживания ($m = 2p + 1$); y_t — текущий уровень ряда динамики; i — порядковый номер уровня в интервале сглаживания; p — при нечетном m равно $p = \frac{m-1}{2}$.

3. Сдвигают интервал сглаживания на одну точку вправо, потом вычисляют по формуле (4.3.1) сглаженное значение для $t + 1$ члена и снова проводят сдвиг. В результате последовательного применения приведенной схемы получится $n - (m - 1)$ новых сглаженных уровней. Первые и последние p уровней ряда с помощью данного алгоритма определить нельзя, поэтому их значения теряются.

Нахождение скользящей средней по четному числу уровней рядов динамики несколько сложнее, так как средняя может быть отнесена только к середине между двумя датами, находящимся в середине интервала сглаживания. Например, средняя, найденная для четырех уровней, относится к середине между вторым и третьим, третьим и четвертым уровнями и т. д. Чтобы ликвидировать такой сдвиг, применяют так называемый способ центрирования. Центрирование заключается в нахождении средней из двух смежных скользящих средних для отнесения полученного уровня к определенной дате. При центрировании необходимо находить скользящие суммы, скользящие средние нецентрированные по этим суммам и средние из двух смежных нецентрированных скользящих средних.

Метод простой скользящей средней используется только в случае, если исходный временной ряд графически может быть представлен в виде прямой линии. В этом случае не искажается динамика исследуемого явления. Однако, когда закономерность развития исходного ряда имеет изгибы и к тому же необходимо сохранить мелкие волны, использование данного метода не даст желаемых результатов, так как простая скользящая средняя может привести к существенным искажениям тенденции ряда. В таких случаях наиболее оправданно использование метода взвешенной скользящей средней.

При построении взвешенной скользящей средней на каждом активном участке (наблюдения, которые используются для расчета среднего значения, называются активным участком сглаживания), значение центрального уровня заменяется на расчетное, определяемое по формуле средней арифметической взвешенной. Весом в данном случае будут выступать весовые коэффициенты, которые зависят от удаления данного уровня от уровня, стоящего в середине активного участка. Вследствие этого весовые коэффициенты симметричны относительного центрального уровня на активном участке. Весовые коэффициенты определяются методом наименьших квадратов.

ПРИМЕР.

Покажем расчет трехуровневой и четырехуровневой скользящих средних по данным, представленным в табл. 4.3.3.

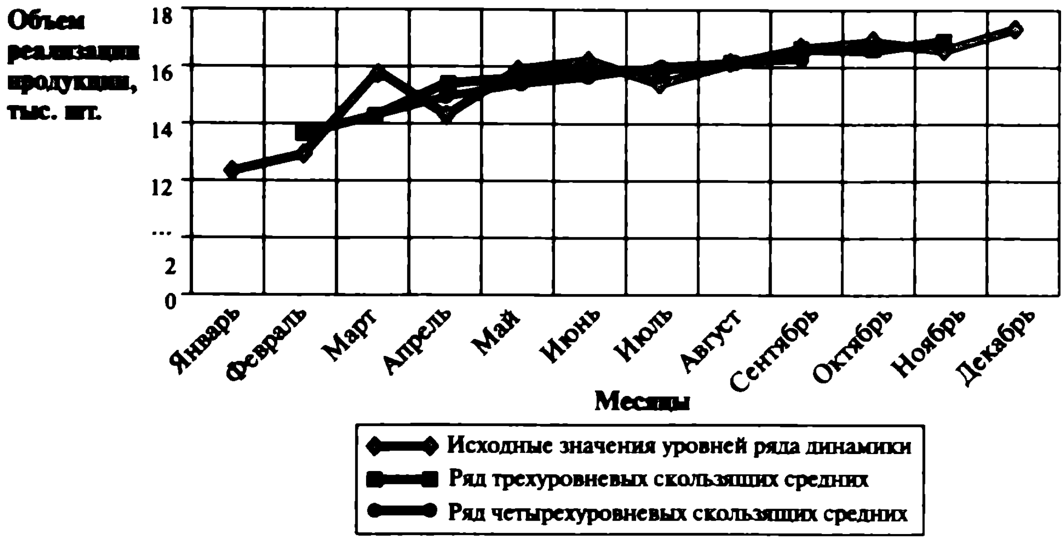


Рис. 4.3.1. Выявление основной тенденции динамики реализации продукции в торговой сети за 2006 г. методом скользящих средних

Таблица 4.3.3

Динамика продажи продукции в торговой сети за 2006 г.

Месяц	Продано продукции, тыс. шт.	Трехуровневые скользящие суммы	Трехуровневые скользящие средние	Четырехуровневые скользящие суммы	Четырехуровневые скользящие средние нецентрированные	Четырехуровневые скользящие средние центрированные
А	1	2	3	4	5	6
Январь	12,8	–	–	–	–	–
Февраль	13,4	–	14,03	–	–	–
Март	15,9	42,1	14,67	–	14,20	14,63
Апрель	14,7	44,0	15,60	56,8	15,05	15,43
Май	16,2	46,8	15,80	60,2	15,83	15,80
Июнь	16,5	47,4	16,13	63,3	15,78	15,93
Июль	15,7	48,4	16,20	63,1	16,20	16,29
Август	16,4	48,6	16,33	64,8	16,38	16,45

Окончание табл. 4.3.3

Месяц	Продано продукции, тыс. шт.	Трехуровневые скользящие суммы	Трехуровневые скользящие средние	Четырехуровневые скользящие суммы	Четырехуровневые скользящие средние нецентрированные	Четырехуровневые скользящие средние центрированные
А	1	2	3	4	5	6
Сентябрь	16,9	49,0	16,80	65,5	16,53	16,66
Октябрь	17,1	50,4	16,93	66,1	16,80	16,94
Ноябрь	16,8	50,8	17,13	67,2	17,08	–
Декабрь	17,5	51,4	–	68,3	–	–

Таким образом, по характеру изменения скользящих средних из трех и четырех уровней можем сделать вывод о том, что объем реализации продукции имеет непрерывную тенденцию к росту.

Недостаток метода простой скользящей средней состоит в том, что сглаженный ряд динамики сокращается ввиду невозможности получить сглаженные уровни для начала и конца ряда. Этот недостаток устраняется применением метода аналитического выравнивания для анализа основной тенденции.

4.3.4. Метод аналитического выравнивания

В отличие от рассмотренных выше методов, таких как укрупнение интервалов, скользящих средних, направленных в основном на то, чтобы ответить на вопрос: есть ли тенденция в динамическом ряду или нет и предварительно определить ее направление, — аналитическое выравнивание позволяет достаточно точно установить характер развития явления и описать его математически, а также, что очень важно, использовать полученную модель для прогнозирования. **Математическая модель** представляет собой систему уравнений, на основании которых исследователь может получить некий набор искусственных переменных в форме временного ряда. Аналитическое выравнивание предполагает представление уровней данного ряда динамики в виде функции времени: $y = f(t)$. Данная математическая функция, описывающая закономерность развития явления во времени, носит название **уравнение тренда**.

При таком подходе изменение явления связывают лишь с течением времени. Считается, что влияние других факторов несущественно или косвенно сказывается через фактор времени. Правильно построенная модель должна соответствовать характеру изменения тенденции исследуемого явления. Выбранная функция позволяет получить выровненные или теоретические значения уровней ряда динамики.

Для отображения основной тенденции развития явлений во времени применяются различные функции: полиномы разной степени, экспоненты, логистические кривые и другие виды. Для выбора формы кривой достаточно отобразить эмпирические значения на графике. Схема построения графика следующая: по оси абсцисс откладываются временные отрезки, по оси ординат — эмпирические значения уровней временного ряда. Чем теснее точки будут группироваться вокруг определенной линии, тем лучше функция будет описывать закономерность развития исходного ряда динамики.

Полиномы имеют следующий вид:

- полином первой степени:

$$\bar{y}_t = a_0 + a_1 t;$$

- полином второй степени:

$$\bar{y}_t = a_0 + a_1 t + a_2 t^2; \qquad (4.3.2)$$

- полином третьей степени:

$$\bar{y}_t = a_0 + a_1 t + a_2 t^2 + a_3 t^3;$$

- полином n-й степени:

$$\bar{y}_t = a_0 + a_1 t + a_2 t^2 + \ldots + a_n t^n.$$

Здесь a_0, a_1, a_2, ... a_n — параметры полиномов, t — условное обозначение времени. В статистической практике параметры полиномов невысокой степени иногда имеют конкретную интерпретацию характеристик динамического ряда. Так, например, параметр a_0 характеризует средние условия развития ряда динамики, параметр a_1 — скорость роста, параметр a_2 — ускорение роста, параметр a_3 — изменение ускорения.

Выбор в пользу выравнивания по линейной функции делают, если уровни ряда меняются в арифметической прогрессии (в этом случае рассчитанные цепные абсолютные приросты уровней приблизительно одинаковы). При ускоренном или замедленном изменении уровней динамического ряда используется уравнение параболы второго порядка. Если уровни ряда меняются в геометрической прогрессии, т. е. рассчитанные цепные коэффициенты роста относительно постоянны, для аналитического выравнивания используется показательная функция. В случае если уровни динамического ряда снижаются, постепенно замедляя скорость, но, по логике, никогда не смогут достичь нуля, для аналитического сглаживания выбирают уравнение гиперболы.

При наличии периодических колебаний (например, сезонных) выравнивание осуществляют с помощью ряда Фурье:

$$\bar{y}_t = a_0 + \sum_{k=1}^{m}(a_k \cos kt + b_k \sin kt). \qquad (4.3.3)$$

Кроме того, возможности современного программного обеспечения (например, системы SPSS и STATISTICA) позволяют использовать в качестве модели тренда математическую функцию любого произвольного вида.

Оценка параметров в моделях (4.3.2) находится методом наименьших квадратов. Суть данного метода, как вы помните из регрессионного анализа, состоит в определении таких параметров (коэффициентов), при которых сумма квадратов отклонений расчетных значений уровней от фактических значений была бы минимальной. Таким образом, эти оценки находятся в результате минимизации выражения:

$$\sum_{t=1}^{n}\left(y_t - \overline{y_t}\right)^2 \to \min, \qquad (4.3.4)$$

где y_t — фактическое значение уровня ряда динамики; $\overline{y_t}$ — расчетное значение; n — длина ряда динамики.

В результате минимизации выражения (4.3.4) получается система нормальных уравнений:

$$\begin{cases} na_0 + a_1\sum t_i + a_2\sum t_i^2 + \ldots + a_p\sum t_i^p = \sum y_i \\ a_0\sum t_i + a_1\sum t_i^2 + a_2\sum t_i^3 + \ldots + a_p\sum t_i^{p+1} = \sum y_i t_i \\ \ldots \\ \sum t_i^p + a_1\sum t_i^{p+1} + a_2\sum t_i^{p+2} + \ldots + a_p\sum t_i^{2p} = \sum y_i t_i^p \end{cases}, \qquad (4.3.5)$$

где n — число членов в ряду динамики, $t_i = 1, 2, \ldots, n$.

Система (4.3.5), состоящая из p уравнений, содержит в качестве известных величин $\sum y_i, \sum y_i t_i, ..., \sum y_i t_i^p$, т. е. суммы наблюдаемых значений уровней динамического ряда, умноженные на показатели времени в степени 1, 2, ..., p, и неизвестные величины a_j. Решение этой системы относительно a_0, a_1, ..., a_p и дает искомые значения параметров.

Системы для расчета параметров полиномов невысоких степеней намного проще. Обозначим последовательные параметры полиномов a_0, a_1, a_2. Тогда системы нормальных уравнений для оценивания параметров примут вид:

- для прямой $\bar{y}_t = a_0 + a_1 t$:

$$\begin{cases} na_0 + a_1 \sum t_i = \sum y_i \\ a_0 \sum t_i + a_1 \sum t_i^2 = \sum y_i t_i \end{cases}; \tag{4.3.6}$$

- для параболы второго порядка ($\bar{y}_t = a_0 + a_1 t + a_2 t^2$):

$$\begin{cases} na_0 + a_1 \sum t + a2 \sum t_i^2 = \sum y_i \\ a_1 \sum t_i + a_1 \sum t_i^2 + a_2 \sum t_i^3 = \sum y_i t_i \\ a_0 \sum t_i^2 + a_1 \sum t_i^3 + a_2 \sum t_i^4 = \sum y_i t_i^2 \end{cases}; \tag{4.3.7}$$

- для гиперболы $\bar{y}_t = a_0 + \frac{a_1}{t}$:

$$\begin{cases} na_0 + a_1 \sum \frac{1}{t_i} = \sum y_i \\ a_0 \sum \frac{1}{t_i} + a_1 \sum \frac{1}{t_i^2} = \sum \frac{y_i}{t_i} \end{cases}. \tag{4.3.8}$$

Для расчета параметров полиномов необходимо ввести в модель числовые значения для фактора времени t. Это можно сделать двумя способами.

1. Ввести фактор времени линейно, т. е. у первого уровня ряда $t_1 = 1$, у второго — $t_2 = 2$, у третьего — $t_3 = 3$, у предпоследнего — $t_{n-1} = n - 1$ и у последнего — $t_n = n$.

2. Другой подход заключается в обозначении фактора времени методом условного нуля. Для этого начало координат условно переносится в середину ряда динамики.

Если в ряду нечетное количество уровней, то начало координат «0» совпадает с серединным уровнем ряда. Тогда для серединного уровня ряда t_i будет равно 0, а у остальных предыдущих уровней t_i будет указываться со знаками «–» с шагом –1 от середины ряда и последующих уровней со знаком «+», с шагом 1 от середины ряда.

При четном числе уровней «0» попадает в интервал между серединными уровнями ряда. Строго следуя законам математики, серединным уровням следовало бы присвоить $t_i = -0{,}5$ и $t_i = +0{,}5$. Однако, пользуясь свойствами арифметической прогрессии и для упрощения числовых значений t_i, серединным уровням присваивают $t_i = (-0{,}5)\cdot 2 = -1$ и $t_i = 0{,}5 \cdot 2 = +1$, а для остальных уровней вводят t_i с соответствующим знаком, сохраняя шаг прогрессии 2.

Тогда, если до переноса начала координат ряд t_i выглядел как 1, 2, 3, ..., n, то после переноса:

- для нечетного числа уровней ряда $t_i = ...; -3; -2; -1; 0; 1; 2; 3; ...$
- для четного числа уровней ряда $t_i = ...; -5; -3; -1; 1; 3; 5; ...$

В любом случае $\sum t_i = 0$.

Следовательно, $\sum t_i$ и все $\sum t_i^p$, у которых p нечетное число, будут равны 0. Таким образом, все члены уравнений, содержащие $\sum t_i$ с такими степенями, могут быть исключены. Системы нормальных уравнений теперь можно упростить:

- для прямой при $\sum t_i = 0$:

$$\begin{cases} a_0 n = \sum y_i \\ a_1 \sum t_i^2 = \sum t_i y_i \end{cases};$$

$$\begin{cases} a_0 = \dfrac{\sum y_i}{n} \\ a_1 = \dfrac{\sum t_i y_i}{\sum t_i^2} \end{cases}; \qquad (4.3.9)$$

- для параболы второго порядка:

$$\begin{cases} a_0 n + a_2 \sum t_i^2 = \sum y_i \\ a_1 \sum t_i^2 = \sum t_i y_i \\ a_0 \sum t_i^2 + a_2 \sum t_i^4 = \sum t_i^2 y_i \end{cases}. \qquad (4.3.10)$$

Решая системы (4.3.9) и (4.3.10), получим величины параметров соответствующих полиномов.

Для выбора оптимального трендового уравнения, наилучшим образом описывающего основную тенденцию развития исследуемого ряда динамики, на практике используют критерий наименьшей суммы квадратов отклонения эмпирических уровней от теоретических, представленный в формуле (4.3.4).

ПРИМЕР.

Определим основную тенденцию ряда динамики фактических экспортных цен на нефть Российской Федерации за 2006 г.

Таблица 4.3.4

Таблица исходных и расчетных данных динамики фактических экспортных цен на нефть Российской Федерации за 2006 г.

Месяц	Средняя фактическая экспортная цена нефти, долл. за т, y_i	t_i	t_i^2	$t_i y_i$	$\bar{y}_t$	$(y_i - \bar{y}_t)^2$
Январь	374,3	–11	121	–4117,3	406,521	1038,1930
Февраль	383,8	–9	81	–3454,2	407,349	554,5554
Март	383,2	–7	49	–2682,4	408,177	623,8505
Апрель	418,6	–5	25	–2093	409,005	92,0640
Май	440,1	–3	9	–1320,3	409,833	916,0913
Июнь	433,2	–1	1	–433,2	410,661	508,0065
Июль	455,1	1	1	455,1	411,489	1901,9190
Август	463,4	3	9	1390,2	412,317	2609,4730
Сентябрь	429,2	5	25	2146,0	413,145	257,7630
Октябрь	392,3	7	49	2746,1	413,973	469,7189
Ноябрь	378,4	9	81	3405,6	414,801	1325,0330
Декабрь	381,3	11	121	4194,3	415,629	1178,4800
Итого	4932,9	0	572	236,9	4932,900	11 475,1500

Первые две графы – ряд динамики, подвергаемый выравниванию, – дополняются графой 3, в которой вводятся обозначения времени t. Значения выбраны таким образом, чтобы $\sum t_i = 0$. В качестве функции выравнивания выбрано уравнение прямой линии: $\bar{y}_t = a_0 + a_1 t$, параметры данного уравнения находим по упрощенным формулам (4.3.9).

Затем в графах 4 и 5 проводим необходимые расчеты и находим: $\sum y_i$ =4932,9; $\sum y_i t_i$ =336,9; $\sum t_i^2$ = 572. Отсюда:

$$\begin{cases} a_0 = \dfrac{\sum y_i}{n} = \dfrac{4932{,}9}{12} = 411{,}075 \\ a_1 = \dfrac{\sum t_i y_i}{\sum t_i^2} = \dfrac{236{,}9}{572} = 0{,}414 \end{cases}$$

Уравнение прямой будет иметь вид:

$$\bar{y}_t = 411{,}075 + 0{,}414t.$$

На основе этого уравнения находятся выровненные уровни путем подстановки в него соответствующих значений t (графа 6 табл. 4.3.4).

Полученное уравнение показывает, что средняя фактическая экспортная цена нефти растет в среднем на 0,414 долл. за тонну ежемесячно. Таким образом, величина параметра a_1 в уравнении прямой показывает среднюю величину абсолютного прироста выровненного ряда динамики.

Сумма уровней эмпирического ряда ($\sum y_i$) полностью совпала с суммой расчетных значений выровненного ряда ($\sum \bar{y}_t$).

Результаты произведенного аналитического выравнивания ряда динамики средних фактических экспортных цен на нефть за 2006 г. и фактические данные отражены на рис. 4.3.2.

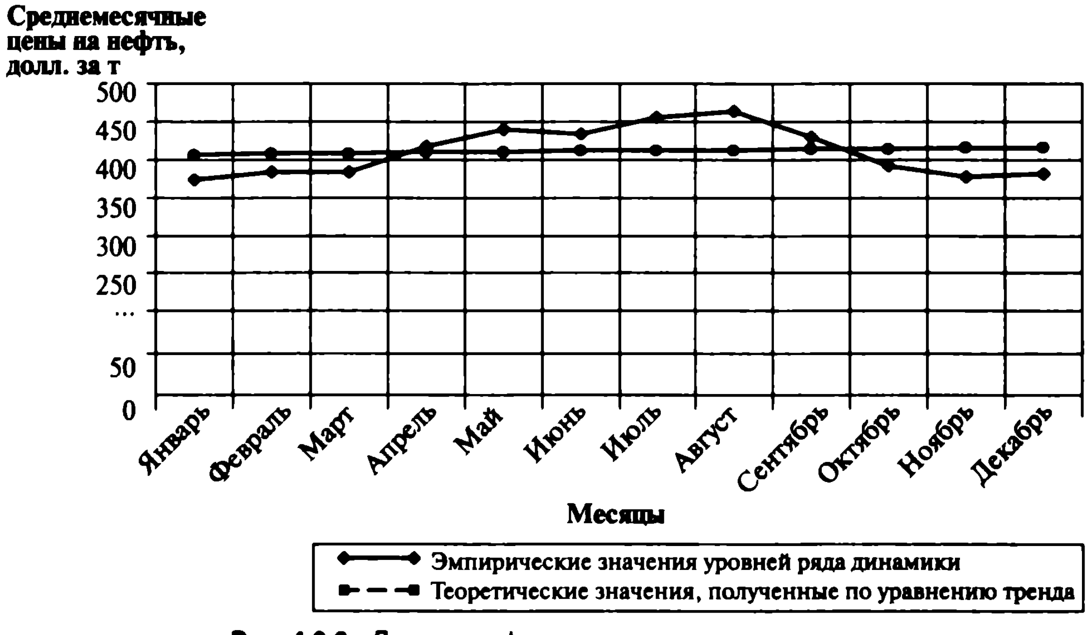


Рис. 4.3.2. Динамика фактических экспортных цен на нефть Российской Федерации за 2006 г.

Для выбора уравнения применим критерий наименьшей суммы квадратов отклонения эмпирических уровней от теоретических:

$$S = \sum_{t=1}^{n}(y_t - \bar{y}_t)^2 = 11475{,}15.$$

КОНТРОЛЬНЫЕ ВОПРОСЫ

1. Что вы понимаете под тенденцией ряда динамики?
2. Когда для выявления основной тенденции ряда динамики достаточно применить метод укрупнения первоначальных интервалов?
3. Для чего применяют метод скользящих средних? В чем основные преимущества и недостатки его использования? В каких случаях используется метод простой скользящей средней?
4. В чем состоят исходные предпосылки применения метода аналитического выравнивания? Что такое тренд и уравнение тренда?
5. Какой прием используют для упрощения расчетов параметров уравнения тренда при аналитическом выравнивании ряда динамики?
6. Что характеризует каждый из параметров уравнения тренда: a_0, a_1 и a_2?
7. В каких случаях для аналитического выравнивания используется прямая, а в каких параболическая функция?
8. Каким показателем необходимо руководствоваться при выборе оптимального уравнения тренда? Назовите этапы (последовательность действий) расчета его параметров.

Глава 4.4

МЕТОДЫ ВЫЯВЛЕНИЯ СЕЗОННЫХ КОЛЕБАНИЙ В РЯДАХ ДИНАМИКИ

- Базовая модель временного ряда
- Методы выявления сезонной компоненты
- Расчет индекса сезонности методом постоянной средней
- Расчет индекса сезонности методом аналитического выравнивания
- Расчет индекса сезонности методом скользящей средней

4.4.1. Базовая модель временного ряда

На практике в рядах динамики часто наблюдаются устойчивые периодические колебания уровней либо вокруг линии тренда, либо по отношению к среднему уровню ряда. Причем эти колебания регулярно повторяются на протяжении некоторого периода. Если значения показателя возрастают либо убывают в зависимости от момента времени, а затем по истечении некоторого временного срока, называемого **циклом**, возвращаются на предыдущие позиции (если колебания происходят вокруг среднего уровня), либо повторяется характер их изменения (если значения уровней колеблются вокруг линии тренда), то в этом случае можно говорить об устойчивых сезонных колебаниях в ряду динамики.

В зависимости от длины цикла различают периодические колебания: циклические (долгопериодические, когда длина цикла составляет несколько лет), сезонные (внутригодичные колебания по месяцам или кварталам года), недельные, дневные (регулярные колебания по дням недели), часовые (в зависимости от часа дня).

Практически любой временной ряд состоит из четырех базовых компонент: тренда, сезонности, циклической вариации и случайной компоненты. Базовая модель временного ряда представляет собой произведение данных компонент:

$$y_t = \text{Тренд} \times \text{Сезонность} \times \text{Цикличность} \times \text{Случайность}.$$

Трендовая составляющая указывает долгосрочное поведение исследуемого временного ряда в виде прямой либо кривой линии. Используется для анализа картины в целом.

Сезонная компонента определяет влияние времени года на исследуемый ряд. Каждый период времени в течение года характеризуется своим сезонным индексом, который свидетельствует о том, насколько выше или ниже соответствующий показатель в данный период времени по сравнению с другими периодами. Наиболее ярко сезонные колебания проявляются в торговле (например, продажа овощей, фруктов, ягод, носящая сугубо сезонный характер), в производстве отдельных видов продукции (например, производство косметических солнцезащитных средств), в потреблении (электроэнергия), в кредитной деятельности финансовых компаний (размер процентных ставок по отдельным видам кредитов часто носит сезонный характер).

Циклическая компонента состоит из последовательных повышений и понижений, которые не повторяются ежегодно и поэтому исключаются из сезонной компоненты. Поскольку эти понижения и повышения чередуются, то не могут быть рассмотрены как часть случайной компоненты. Так, циклические изменения уровней динамических рядов наблюдаются при анализе деятельности страховых компаний, например, в связи с экологическими катастрофами или войнами нарушается пропорция между половым составом населения определенных возрастных групп. Это вызывает циклические изменения сумм страховых возмещений и, соответственно, размеров тарифных ставок по определенным видам имущественного страхования.

Под **случайной** компонентой понимают ту часть вариации, которую невозможно объяснить. В ней проявляются действия однократных событий, происходящих случайно, а не систематически. К данным нерегулярным колебаниям можно применить только методы расчета их величин, например, стандартное отклонение. Так как даже в идеальных условиях прогноз не может быть точнее (в среднем), чем типичная величина нерегулярной вариации.

4.4.2. Методы выявления сезонной компоненты

При анализе квартальных или помесячных данных многих экономических явлений достаточно часто обнаруживаются определенные, постоянно повторяющиеся колебания, которые существенно не изменяются за длительный период времени. Данные колебания происходят в результате смены времени года, изменения различных экономических факторов. В статистике под **сезонными колебаниями** понимают периодические колебания, которые имеют определенный и постоянный период, равный годовому промежутку. Динамический ряд в этом случае называют **сезонным рядом динамики**.

Для выявления сезонных колебаний обычно берут данные за несколько лет (минимум за три года), распределенные по месяцам или кварталам. Такие данные необходимы для выявления устойчивой сезонной волны, на которой не отражаются случайные влияния одного года. Также надо учесть, что не всякие различия в месячных или квартальных уровнях являются сезонными колебаниями, а только регулярно повторяющиеся. Если колебания месячных уровней внутри одного года распределены совершенно иначе, чем в другой год, то это не сезонные, а случайные колебания, т. е. колебания, вызванные не влиянием сезона, а другими причинами. Например, такими колебаниями могут быть изменения в курсах акций, обменных курсов валют, вызванные сменой финансовой политики государства, политическими кризисами в стране и мире, слиянием и разделением компаний.

Сезонные колебания характеризуются специальными показателями, которые называются **индексами сезонности** (I_s). Совокупность этих показателей отражает сезонную волну. Индексами сезонности являются процентные отношения фактических внутригодовых уровней к постоянной или переменной средней.

Индексы сезонности можно рассчитать тремя методами:

- методом постоянной средней;
- методом аналитического выравнивания;
- методом скользящей средней.

4.4.3. Расчет индекса сезонности методом постоянной средней

Если ряд динамики не содержит ярко выраженной тенденции к развитию, то индексы сезонности вычисляются непосредственно по фактическим данным без их предварительного выравнивания.

Для каждого месяца определяется средняя величина уровня, например за три года ($\bar{y}_i$), затем из них рассчитывается среднемесячный уровень для всего ряда ($\bar{y}$) и в заключение определяется процентное отношение средних для каждого месяца к общему среднемесячному уровню ряда, т. е.:

$$I_S = \frac{\bar{y}_i}{\bar{y}} \cdot 100\%. \qquad (4.4.1)$$

ПРИМЕР.

За 2004–2006 гг. по месяцам имеются данные о депозитах и вкладах физических лиц на рублевых счетах населения *N*-го города. Требуется рассчитать индексы сезонности методом постоянной средней.

Таблица 4.4.1

Динамика депозитов и вкладов на рублевых счетах населения г. *N* за 2004–2006 гг.

Месяцы	Депозиты и вклады физических лиц на рублевых счетах, млн руб.				Индекс сезонности $(\bar{y}_i : \bar{y}) \cdot 100\%$
	2004 г. y_i	2005 г. y_i	2006 г. y_i	В среднем за три года $\bar{y}_i$	
А	1	2	3	4	5
Январь	190,0	155,0	145,0	163,3	120,7
Февраль	165,0	140,0	135,0	146,7	108,4
Март	150,0	153,0	135,0	146,0	107,9
Апрель	135,0	140,0	146,0	140,3	103,0
Май	135,0	136,0	131,0	134,0	99,0
Июнь	123,0	130,0	136,0	129,7	95,9
Июль	125,0	128,0	125,0	126,0	93,1
Август	120,0	125,0	124,0	123,0	90,9
Сентябрь	118,0	118,0	120,0	118,7	87,7
Октябрь	126,0	130,0	128,0	128,0	94,6
Ноябрь	130,0	131,0	135,0	132,0	97,6
Декабрь	138,0	131,0	139,0	136,0	100,5
Итого	1655,0	1617,0	1599,0	1623,7	–
Средний уровень ряда	137,9	134,8	133,3	$\bar{y}$ =135,3	100,0

Индекс сезонности методом постоянной средней рассчитывается по нижеприведенному алгоритму.

1. Вычисляются усредненные значения показателя по одноименным месяцам как средняя арифметическая простая:

- январь:

$$\bar{y}_1 = \frac{190 + 155 + 145}{3} = 163{,}3;$$

- февраль:

$$\bar{y}_2 = \frac{165 + 140 + 135}{3} = 146{,}7;$$

- март:

$$\bar{y}_1 = \frac{150 + 153 + 135}{3} = 146 \text{ и т. д. (графа 4 табл. 4.4.1).}$$

2. По вычисленным средним уровням определяется общий средний уровень ($\bar{y}$):

$$\bar{y} = \frac{\sum(\bar{y}_i)}{m} = \frac{137{,}9 + 134{,}8 + 133{,}3}{3} = 135{,}3$$

или

$$\bar{y} = \frac{\sum \bar{y}_i}{n} = \frac{1623{,}7}{12} = 135{,}3,$$

где m – число лет; $\sum(\bar{y}_i)$ – сумма среднегодовых уровней ряда динамики; n – количество уровней усредненного ряда $\bar{y}_i$.

3. Определяется индекс сезонности по месяцам года:

$$I_{S1} = \frac{163{,}3}{135{,}3} \cdot 100 = 120{,}7\%;$$

$$I_{S2} = \frac{146{,}7}{135{,}3} \cdot 100 = 108{,}4\%;$$

$$I_{S3} = \frac{146}{135{,}3} \cdot 100 = 107{,}9\%$$

и т. д. (графа 5 табл. 4.4.1).

Рассчитанные индексы сезонности характеризуют сезонную волну депозитов и вкладов физических лиц на рублевых счетах во внутригодовой динамике, где пик вкладов приходится на январь месяц. Изобразим полученные данные в виде линейной диаграммы динамики (рис. 4.4.1).

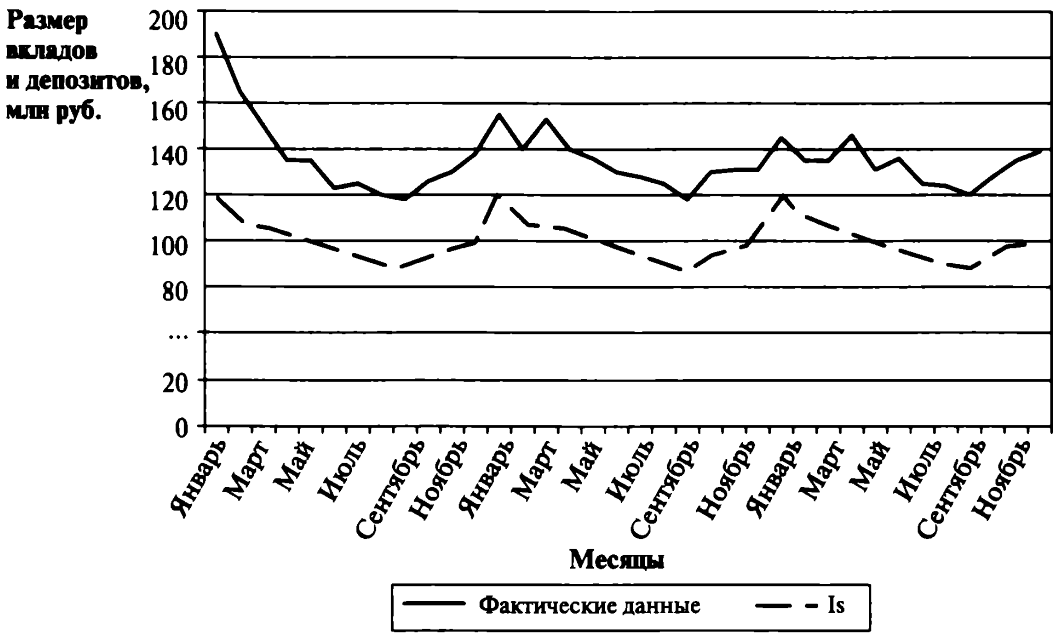


Рис. 4.4.1. Динамика размера депозитов и вкладов населения на рублевых счетах по месяцам и сезонные колебания, повторяющиеся ежегодно

4.4.4. Расчет индекса сезонности методом аналитического выравнивания

Если ряд динамики содержит определенную тенденцию к развитию, то прежде чем вычислить сезонную волну, фактические данные должны быть обработаны так, чтобы была выявлена общая тенденция. Обычно для этого прибегают к аналитическому выравниванию ряда динамики. В этом случае ход вычислений индексов сезонности следующий:

1. По эмпирическим уровням ряда динамики, применяя метод аналитического выравнивания, находят уравнение тренда. При этом вид выбранной аналитической функции должен соответствовать характеру изменения уровней ряда (прямая, парабола, гипербола и т. д.).

2. По соответствующему уравнению тренда вычисляются для каждого месяца (квартала) выровненные значения уровней на момент времени t.

3. Вычисляются индексы сезонности как отношения фактических месячных (квартальных) данных y_i к соответствующим выровненным данным $\bar{y}_t$ в процентах.

$$I_t' = \frac{y_t}{\bar{y}_t} \cdot 100\%. \qquad (4.4.2)$$

4. Поскольку для одноименных моментов времени вычисляются несколько индексов сезонности, требуется определить среднюю арифметическую величину из процентных отношений, рассчитанных по одноименным периодам в процентах:

$$\bar{I}_t' = \frac{I_1' + I_2' + \ldots + I_m'}{m}, \qquad (4.4.3)$$

где m — число одноименных периодов.

В общем виде формулу расчета индекса сезонности данным способом можно записать так:

$$I_S = \frac{\sum \frac{y_i}{\bar{y}_t}}{n} \cdot 100\%. \qquad (4.4.4)$$

Расчет заканчивается проверкой правильности вычислений индексов, так как средний индекс сезонности для всех месяцев (кварталов) должен быть 100%, то сумма полученных индексов по месячным данным равна 1200, а сумма по четырем кварталам — 400.

ПРИМЕР.

Определим индекс сезонности методом аналитического выравнивания, используя ряд поквартальных объемов продаж туристической компании, представленный в табл. 4.4.2.

1. Построим линейную модель тренда вида: $\bar{y}_t = a_0 + a_1 t$. Тогда параметры модели определим через следующую систему уравнений:

$$\begin{cases} 16a_0 + 136a_1 = 10665 \\ 136a_0 + 1496a_1 = 104635 \end{cases}, \quad \begin{cases} a_0 = 371{,}96 \\ a_1 = 36{,}13 \end{cases}.$$

Следовательно, $\bar{y}_t = 371{,}96 + 36{,}13t$.

2. Вычислим по данному уравнению тренда для каждого квартала теоретические уровни на соответствующий им момент времени. Полученные результаты запишем в графе 3 табл. 4.4.2.

3. Определим индекс сезонности по каждому кварталу года как отношение фактических квартальных данных к соответствующим выровненным данным в процентах (формула 4.4.2):

$$I_1 = \frac{438}{408,09} \cdot 100 = 107,33\%;$$

$$I_2 = \frac{432}{444,22} \cdot 100 = 97,25\%;$$

$$I_3 = \frac{591}{480,35} \cdot 100 = 123,04\%.$$

и т. д. (графа 4 табл. 4.4.2).

Таблица 4.4.2

Динамика объемов продаж туров туристической компании за период с 2003 по 2006 гг.

Год и квартал		Объем продаж y_i, тыс. руб.	t_i	Теоретические уровни $\bar{y}_t = 371,96 + 36,13t$	Индекс сезонности по каждому кварталу года $I_i^s = \frac{y_i}{\bar{y}_t} \cdot 100\%$	Индекс сезонности по одноименным кварталам I_i^s
А		1	2	3	4	5
2003	I	438	1	408,09	107,33	92,93
	II	432	2	444,22	97,25	91,06
	III	591	3	480,35	123,04	124,78
	IV	475	4	516,48	91,97	91,59
2004	I	459	5	552,61	83,06	92,93
	II	506	6	588,74	85,95	91,06
	III	736	7	624,87	117,78	124,78
	IV	542	8	661,00	81,10	91,59
2005	I	676	9	697,13	96,97	92,93
	II	645	10	733,26	87,96	91,06
	III	1084	11	769,39	140,89	124,78
	IV	819	12	805,52	101,67	91,59
2006	I	710	13	841,65	84,36	92,93
	II	817	14	877,78	93,08	91,06
	III	1073	15	913,91	117,41	124,78
	IV	862	16	950,04	90,73	91,59
Итого		10 865	136	10 865	1600,00	1600,00

Индексы сезонности вычислены верно, так как сумма всех индексов по квартальным данным составила 1600, что отвечает правилу вычислений.

4. Вычислим средние арифметические индексы (формула 4.4.3) из процентных соотношений индексов (графа 4) и представим результаты в графе 5 данной таблицы:

$$I_1^s = \frac{107{,}33 + 83{,}06 + 96{,}97 + 84{,}36}{4} = 92{,}93;$$

$$I_2^s = \frac{97{,}25 + 85{,}95 + 87{,}96 + 93{,}08}{4} = 91{,}06$$

и т. д. (графа 5 табл. 4.4.2).

В результате расчетов в табл. 4.4.2 получен ряд индексов (графа 5), характеризующих сезонную волну объемов продаж туристической компании. Таким образом, можно сделать вывод, что объем продаж в третьем квартале на 25% в среднем выше, чем во всех остальных кварталах, а во втором — на 9% в среднем ниже, чем в первом, третьем и четвертом кварталах (рис. 4.4.2).

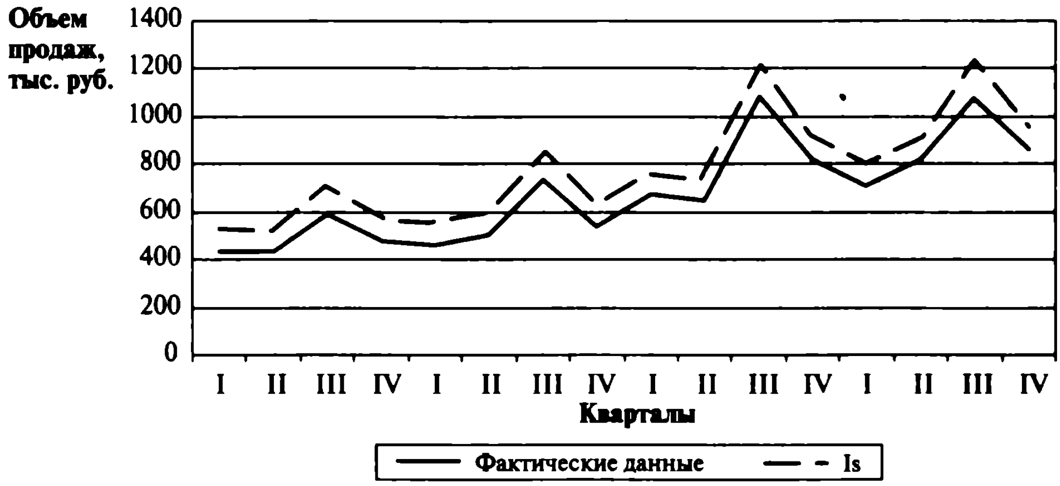


Рис. 4.4.2. Динамика поквартальных объемов продаж туристической компании и сезонные колебания продаж туров, повторяющиеся ежегодно

4.4.5. Расчет индекса сезонности методом скользящей средней

Для того чтобы выделить сезонное поведение, необходимо, прежде всего, рассчитать отношение исходных значений к скользящему среднему. Полученный результат будет включать сезонную и случайную компоненту исходного ряда динамики, поскольку скользящая средняя исключает из данных тренд и циклическую компоненту.

Для устранения случайной компоненты необходимо усреднить полученные значения для каждого сезона. Таким образом, сезонная компонента в данных будет продолжать проявляться, поскольку она присутствует ежегодно, а случайная компонента будет усреднена.

Использование данного метода базируется на делении значений ряда на гладкие скользящие средние следующим образом.

1. Скользящие средние используются для устранения сезонных эффектов путем усреднения по всему году, для уменьшения нерегулярной компоненты, получения комбинации тренда и циклической компоненты.

2. Деление исходного ряда на сглаженный ряд скользящего среднего дает отношение к скользящему среднему, которое включает как сезонные, так и нерегулярные значения. Выполняя группирование по времени года, а затем усреднение в полученных группах, находим сезонный индекс для каждого времени года. Выполняя деление каждого значения ряда на соответствующий сезонный индекс для соответствующего времени года, определяем значения с сезонной поправкой.

3. Регрессия ряда с сезонной поправкой (y) по времени (t) служит для моделирования долгосрочного тренда в виде прямой линии как функции от времени. Этот тренд не отражает сезонных колебаний и дает возможность получить прогноз с сезонной поправкой.

4. Прогнозирование выполняется с помощью сезонности тренда. Получая из уравнения регрессии прогнозируемое значение для будущих периодов времени и затем умножая их на соответствующий сезонный индекс, исследователь получает прогнозы, которые отражают как долгосрочную тенденцию, так и сезонное поведение явления.

Преимуществом метода скользящей средней является легкость вычислений и интерпретации результатов. Основной недостаток заключается в том, что соответствующая модель не является полностью определенной, как следствие — бывает достаточно сложно определить границы прогноза.

ПРИМЕР.

Определим индекс сезонности методом скользящей средней, используя данные об обьеме продаж автомобилей компании Ford Motor Company, представленные в табл. 4.4.3.

Для реализации данного метода необходимо:

1. Рассчитать четырехуровневые скользящие средние центрированные обьема продаж автомобилей (графа 2).

2. Определить отношение фактических данных к скользящей средней (графа 3).

Таблица 4.4.3

Динамика объема продаж автомобилей компании Ford Motor Company в 2001–2004 гг., млн долл.

Год и квартал		Объем продаж y_i, млн долл.	Четырех-уровневая скользящая средняя	Отношение фактических данных к скользящей средней (гр. 1 : гр. 2)	Индекс сезонности I_i^s	Объем продаж с поправкой на сезонные колебания, млн долл. (гр. 1 : гр. 4)
А		1	2	3	4	5
2001	I	26 070	–	–	1,01	25 811,88
	II	28 375	–	–	1,08	26 273,15
	III	24 926	27 100,63	0,92	0,90	27 695,56
	IV	27 766	27 602,75	1,01	1,02	27 221,57
2002	I	28 601	27 727,38	1,03	1,01	28 317,82
	II	29 861	27 645,13	1,08	1,08	27 649,07
	III	24 437	27 586,00	0,89	0,90	27 152,22
	IV	27 597	27 785,63	0,99	1,02	27 055,88
2003	I	28 297	28 276,00	1,00	1,01	28 016,83
	II	31 762	29 017,25	1,10	1,08	29 409,26
	III	26 459	29 723,25	0,89	0,90	29 398,89
	IV	31 505	30 071,13	1,05	1,02	30 887,25
2004	I	30 037	30 418,63	0,99	1,01	29 739,60
	II	32 805	30 684,75	1,07	1,08	30 375,00
	III	28 196	–	–	0,90	31 328,89
	IV	31 897	–	–	1,02	31 271,57

3. Получить сезонный индекс путем усреднения значений соответствующих отношений определенного квартала в совокупности исследуемых лет:

$$I_{S1} = \frac{1{,}03 + 1{,}00 + 0{,}99}{3} = 1{,}01;$$

$$I_{S2} = \frac{1{,}08 + 1{,}1 + 1{,}07}{3} = 1{,}08;$$

$$I_{S3} = \frac{0{,}92 + 0{,}89 + 0{,}89}{3} = 0{,}9;$$

$$I_{S4} = \frac{1{,}01 + 0{,}99 + 1{,}05}{3} = 1{,}02 \text{ (графа 4 табл. 4.4.3).}$$

4. Вычислить поквартальные значения объема продаж с поправкой на сезонные колебания как произведение исходных данных на сезонный индекс для соответствующего квартала.

5. Анализ исходного временного ряда свидетельствует о наличии трендовой компоненты: в анализируемом периоде наглядно выражена тенденция роста объема продаж автомобилей, причем характер тенденции близок к линейному развитию. В рассмотренном временном ряду также отчетливо наблюдаются сезонные колебания: устойчиво повторяющееся увеличение объема продаж автомобилей во втором квартале по сравнению с 1-м, 3-м и 4-м, в которые происходят повторяющиеся из года в год сезонные спады. Скорректированные поквартальные объемы продаж с поправкой на сезонные колебания позволяют построить достаточно точные прогнозные значения.

КОНТРОЛЬНЫЕ ВОПРОСЫ

1. Назовите базовые компоненты любого временного ряда. В чем основное отличие сезонного компонента от циклического? Подкрепите свой ответ примером.
2. Что характеризует индекс сезонности? Как он изображается графически?
3. От каких факторов зависит выбор того или иного метода для расчета индекса сезонности?
4. Почему в случае анализа трендов и сезонных колебаний используются данные как минимум за целый год?
5. Какие компоненты сохраняются в скользящем среднем? Какие уменьшаются или вообще исчезают?
6. Что нужно сделать, чтобы на основе отношения к скользящему среднему получить сезонный индекс?
7. Как внести сезонную поправку в значения временного ряда? Как интерпретируются полученные значения?
8. Как оценивается линейный тренд в анализе сезонных колебаний? Какой вид прогноза представляет линейный тренд? Какие компоненты будут представлены в этом прогнозе, а какие будут отсутствовать?

Глава 4.5

ПРОСТЕЙШИЕ МЕТОДЫ ПРОГНОЗИРОВАНИЯ ВРЕМЕННЫХ РЯДОВ

- Особенности прогнозирования при исследовании рядов динамики
- Прогнозирование методом среднего уровня ряда
- Прогнозирование методом среднего абсолютного прироста
- Прогнозирование методом среднего темпа роста
- Прогнозирование на основе экстраполяции тренда

4.5.1. Особенности прогнозирования при исследовании рядов динамики

Анализ динамики социально-экономических явлений, выявление и характеристика основной тенденции развития дают основание для прогнозирования. **Прогнозирование** — это научное, основанное на системе причинно-следственных связей и закономерностей выявление состояния и вероятностных путей развития явлений и процессов. Прогнозирование предопределяет оценку показателей и дает характеристику явлений и процессов в будущем.

В зависимости от характера воздействия на ход исследуемых процессов и явлений можно выделить три основных понятия прогнозирования: гипотеза, предсказание, прогноз.

Гипотеза — это научно обоснованное предположение либо о непосредственно ненаблюдаемом факте, либо о закономерном порядке, объясняющем известную совокупность явлений. На уровне гипотезы дается качественная характеристика объекта, выражающая общие закономерности его поведения.

Предсказание — это предвидение таких событий, количественная характеристика которых невозможна или затруднена.

Прогноз — это количественное, вероятностное утверждение в будущем о состоянии объекта или явления с достаточно высокой степенью достоверности, на основе анализа тенденций и закономерностей прошлого и настоящего.

Процесс прогнозирования предполагает, что закономерность развития, действующая в прошлом (внутри ряда динамики), сохранится и в прогнозируемом будущем, т. е. прогноз основан на экстраполяции.

Экстраполяцией называется прогнозирование явлений и процессов на основе выявленных закономерностей их развития в прошлом и настоящем периодах, представленных данным динамическим рядом. Экстраполяция всегда проводится за пределы исследуемого временного ряда: в будущее или прошлое. Экстраполяция, проводимая в будущее, называется перспективой, а в прошлое — ретроспективой. Обычно, говоря об экстраполяции рядов динамики, подразумевают чаще всего перспективную экстраполяцию. Под **интерполяцией** понимают прогнозирование неизвестных по каким-либо причинам уровней внутри самого исследуемого ряда динамики.

Точность и надежность прогнозов, получаемых при экстраполяции, зависят от следующих факторов:

- инерционности изучаемого явления, которое подвергается прогнозированию;
- точности полученного уравнения тренда, используемого для описания закономерности развития исходного ряда;
- оптимальности выбранного метода прогнозирования для конкретного временного ряда;
- периода экстраполяции (чем он короче, тем получаемые прогнозы являются наиболее точными);
- длины ряда динамики, на основании которого будет проводиться экстраполяция.

На первом этапе при построении прогнозов следует решить задачу выбора длины динамического ряда. Следовать правилу «чем больше, тем лучше» не совсем верно, так как в быстро развивающемся мире экономики и финансов длинные ряды динамики зачастую становятся несопоставимыми. Это может быть вызвано тем, что меняется методология расчета показателей, тенденции и сущность социально-экономических явлений и процессов.

На практике используется следующее правило: период упреждения прогноза не должен превышать $^1/_3$ длины базового временного ряда. Под **периодом упреждения прогноза** понимают отрезок времени от момента, для которого имеются последние статистические данные об изучаемом объекте, до момента, к которому относится прогноз. Например, если временной ряд состоит из девяти уровней, то прогнозирование проводится не дальше, чем на три уровня. Но в любом случае период упреждения прогноза тесно связан со степенью инерционности исследуемого явления. Если процесс имеет малую инерционность, то информативность уровней ряда по мере их удаления от периода прогнозирования будет соответственно снижаться: наиболее информативными будут являться последние периоды. На данном свойстве основаны методы дисконтирования информации, которые учитывают различную информационную значимость уровней ряда.

Первоначальные прогнозы, как правило, сводятся к экстраполяции тенденции. При этом могут использоваться разные методы в зависимости от исходной информации. Можно выделить следующие элементарные методы экстраполяции: на основе среднего уровня ряда динамики; среднего абсолютного прироста; среднего темпа роста и экстраполяция на основе применения метода наименьших квадратов и представления развития явлений во времени в виде уравнения тренда, т. е. математической функции зависимости уровней ряда (y) от фактора времени (t).

4.5.2. Прогнозирование методом среднего уровня ряда

Использование данного метода целесообразно только в том случае, если явление обладает высокой степенью инерционности. В этом случае точечная оценка прогнозного значения принимается равной среднему уровню ряда, а интервальная строится следующим образом:

$$\bar{y} - t_\alpha S_{\bar{y}} < \hat{y}^{*}_{t+L} < \bar{y} + t_\alpha S_{\bar{y}}, \tag{4.5.1}$$

где t_α — значение t-распределения Стьюдента, соответствующее $(n-1)$ степеням свободы и выбранному значению уровня значимости α; $S_{\bar{y}}$ — средняя квадратическая ошибка, рассчитываемая по формуле:

$$S_{\bar{y}} = \frac{S}{\sqrt{n}}, \tag{4.5.2}$$

где S — среднее квадратическое отклонение, вычисленное для членов ряда:

$$S = \sqrt{\frac{\sum (y - \bar{y})^2}{n-1}}. \tag{4.5.3}$$

При использовании данного метода прогнозирования считают, что значения исследуемого явления колеблются вокруг среднего уровня и эта тенденция сохранится в будущем.

4.5.3. Прогнозирование методом среднего абсолютного прироста

Прогнозирование на основе абсолютного прироста может быть выполнено в том случае, если есть уверенность считать общую тенденцию линейной, т. е. метод основан на предположении о равномерном изменении уровня (под равномерностью понимается стабильность абсолютных приростов).

В этом случае, когда нужно получить прогноз на L шагов вперед (L-период упреждения), достаточно воспользоваться следующей формулой:

$$\hat{y}_{t+L} = y_t + \bar{\Delta} \cdot L, \qquad (4.5.4)$$

где y_t — фактическое значение конечного уровня ряда динамики; $\hat{y}_{t+L}$ — прогнозная оценка значения $(t + L)$ уровня ряда; L — период упреждения (срок) прогноза; $\bar{\Delta}$ — значение среднего абсолютного прироста, рассчитанное по формуле:

$$\bar{\Delta} = \frac{y_t - y_1}{n-1} \text{ или } \bar{\Delta} = \frac{\sum \Delta_i}{n-1}, \qquad (4.5.5)$$

где y_t — последний уровень исходного ряда динамики; y_1 — первый уровень исходного ряда динамики.

Как видно из приведенной формулы, прогнозирование методом среднего абсолютного прироста заключается в непрерывном увеличении последнего уровня временного ряда на величину среднего абсолютного прироста на всем периоде упреждения.

Применение данного метода прогнозирования возможно при предварительной проверке следующих предпосылок.

1. Абсолютные цепные приросты должны быть приближенно одинаковыми:

$$\Delta_{iy} = y_i - y_{i-1},$$

где y_i — значение уровня i-го периода; y_{i-1} — значение уровня, предшествующего i-му периоду времени.

2. Должно выполняться неравенство вида:

$$\delta^2_{ост} < \rho^2,$$

где $\delta^2_{ост}$ — остаточная дисперсия, рассчитываемая по формуле:

$$\delta^2_{\text{ост}} = \frac{\sum_{i=1}^{n}(y_i - \bar{y}_{\bar{\Delta}})^2}{n}, \tag{4.5.6}$$

где y_i — эмпирические значения уровней ряда динамики; $\bar{y}_{\bar{\Delta}}$ — теоретические значения уровней ряда, выровненные методом среднего абсолютного прироста; n — количество уровней исходного ряда динамики.

$$\rho^2 = \frac{1}{2} \cdot \frac{\sum_{i=1}^{n}\Delta_i^2}{n}, \tag{4.5.7}$$

где Δ_i — цепные абсолютные приросты уровней исходного ряда динамики.

ПРИМЕР.

Построим прогноз на три периода упреждения средних потребительских цен кинотеатров на 2006–2008 гг.

Таблица 4.5.1

Таблица исходных и расчетных данных о динамике средних потребительских цен на услуги кинотеатров за 1999–2005 гг. (на конец года, руб. за один вид услуги)

Годы	Средние потребительские цены, руб.	Абсолютные цепные приросты $\Delta_{i\text{ц}} = y_i - y_{i-1}$	$\bar{y}_{\bar{\Delta}}$	$y_i - \bar{y}_{\bar{\Delta}}$	$(y_i - \bar{y}_{\bar{\Delta}})^2$	Δ_i^2
1999	7,36	–	7,360	0	0	–
2000	13,01	5,65	18,588	–5,578	31,118	31,923
2001	23,11	10,1	29,817	–6,707	44,979	102,010
2002	34,39	11,28	41,045	–6,655	44,289	127,238
2003	48,66	14,27	52,273	–3,613	13,056	203,633
2004	58,52	9,86	63,502	–4,982	24,817	97,220
2005	74,73	16,21	74,730	0,000	0,000	262,764
Итого	–	67,37	–	–	158,259	824,788

1. Рассчитаем абсолютные цепные приросты для проверки первой предпосылки. Можно говорить о том, что они достаточно близки между собой по значениям признака.
2. Определим средний абсолютный прирост:

$$\bar{\Delta} = \frac{\sum \Delta_i}{n-1} = \frac{67{,}37}{6} = 11{,}228 \text{ или } \bar{\Delta} = \frac{y_t - y_1}{n-1} = \frac{74{,}73 - 7{,}36}{6} = 11{,}228.$$

3. Проверим неравенство, для чего рассчитаем остаточную дисперсию по формуле (4.5.6):

$$\delta^2_{ост} = \frac{\sum_{i=1}^{n}(y_i - \bar{y}_{\Delta})^2}{n} = \frac{158{,}259}{7} = 22{,}608$$

и показатель ρ^2 по формуле (4.5.7):

$$\rho^2 = \frac{1}{2} \cdot \frac{\sum_{i=1}^{n}\Delta_i^2}{n} = \frac{1}{2} \cdot \frac{824{,}788}{7} = 58{,}913.$$

Так как на основании полученных значений можно сделать вывод о выполнении и второй предпосылки $\delta^2_{ост} < \rho^2$, следовательно, метод среднего абсолютного прироста можно использовать для нахождения прогнозных значений.

4. Построим прогноз на следующие три года средних ожидаемых потребительских цен на услуги кинотеатров (табл. 4.5.2).

Таблица 4.5.2

Прогнозные значения средних ожидаемых потребительских цен на услуги кинотеатров в 2006–2008 гг.

Годы	L	Прогнозные значения, полученные методом среднего абсолютного прироста
2006	1	$\bar{y}_{2006} = 74{,}73 + 11{,}228 \cdot 1 = 85{,}958$
2007	2	$\bar{y}_{2007} = 74{,}73 + 11{,}228 \cdot 2 = 97{,}187$
2008	3	$\bar{y}_{2008} = 74{,}73 + 11{,}228 \cdot 3 = 108{,}415$

5. Рассчитаем ошибку прогноза по формуле:

$$\sigma_{ош} = \sqrt{\frac{\sum(y_i - \bar{y}_t)^2}{n}} = \sqrt{\frac{158{,}259}{7}} = 4{,}755.$$

4.5.4. Прогнозирование методом среднего темпа роста

Прогнозирование по среднему темпу роста можно осуществлять в случае, когда есть основание считать, что общая тенденция ряда характеризуется показательной (экспоненциальной) кривой. Для нахождения прогнозного значения на L шагов вперед необходимо использовать следующую формулу:

$$\hat{y}_{t+L} = y_t \cdot \bar{T}_p^{\,L}, \tag{4.5.8}$$

где y_t — последний уровень исходного ряда динамики (для перспективного прогноза) или уровень, принятый за базу экстраполяции (во всех остальных случаях); $\overline{T}_р$ — средний темп роста, рассчитываемый по формуле вида:

$$\overline{T}_р = \sqrt[n-1]{\frac{y_n}{y_1}}, \tag{4.5.9}$$

где y_n — последний уровень исходного ряда динамики; y_1 — первый уровень исходного ряда динамики.

ПРИМЕР.

Построим прогноз на три периода упреждения средних потребительских цен кинотеатров на 2006–2008 гг.

Таблица 4.5.3

Таблица исходных и расчетных данных о динамике средних потребительских цен на услуги кинотеатров за 1999–2005 гг. (на конец года, руб. за один вид услуги)

Годы	Средние потребительские цены, руб.	$\bar{y}_{\bar{T}p}$	$(y_i - \bar{y}_{\bar{T}p})^2$
1999	7,36	7,36	0,000
2000	13,01	9,862	9,907
2001	23,11	13,216	97,899
2002	34,39	17,709	278,258
2003	48,66	23,730	621,507
2004	58,52	31,798	714,058
2005	74,73	42,610	1031,725
Итого	–	–	2753,354

1. Рассчитаем средний темп роста по формуле (4.5.9):

$$\overline{T}_p = \sqrt[n-1]{\frac{y_n}{y_1}} = \sqrt[6]{\frac{74,73}{7,36}} = 1,34.$$

2. Построим прогноз на следующие три года средних ожидаемых потребительских цен на услуги кинотеатров.

3. Рассчитаем ошибку прогноза по формуле:

$$\sigma_{ош} = \sqrt{\frac{\sum(y_i - \bar{y}_t)^2}{n}} = \sqrt{\frac{2753,354}{7}} = 19,83.$$

Таблица 4.5.4

Прогнозные значения средних ожидаемых потребительских цен на услуги кинотеатров в 2006–2008 гг.

Годы	L	Прогнозные значения, полученные методом среднего темпа роста
2006	1	$\bar{y}_{2006} = 74{,}73 \cdot 1{,}34^1 = 100{,}138$
2007	2	$\bar{y}_{2007} = 74{,}73 \cdot 1{,}34^2 = 134{,}185$
2008	3	$\bar{y}_{2008} = 74{,}73 \cdot 1{,}34^3 = 179{,}808$

Так как средняя квадратическая ошибка по методу среднего абсолютного прироста меньше, следовательно, этот метод целесообразно использовать для нахождения прогнозных значений.

К недостаткам рассмотренных методов следует отнести то, что они учитывают лишь конечный и начальный уровни ряда, исключая влияние промежуточных уровней. Тем не менее методы среднего абсолютного прироста и среднего темпа роста имеют весьма широкую область применения, что объясняется простотой их вычисления. Они могут быть использованы как приближенные, простейшие способы прогнозирования, предшествующие более глубокому количественно-качественному анализу.

4.5.5. Прогнозирование на основе экстраполяции тренда

Наиболее распространенным методом прогнозирования является аналитическое выражение тренда, целью которого является получение математической функции (уравнения тренда), которая описывает изменение уровней ряда динамики с течением времени t. При этом для выхода за границы исследуемого периода достаточно продолжить значения независимой переменной времени (t).

Экстраполяция позволяет получить точечные значения прогноза. Точечный прогноз представляет собой оценку прогнозируемого показателя в конкретный период (момент) по уравнению, описывающему тенденцию показателя. Точечная оценка рассчитывается путем подстановки номера t, на который рассчитывается прогноз, в уравнение тренда. Она является средней оценкой для прогнозируемого интервала времени.

Метод прогнозирования на основе экстраполяции тренда применяется в случае выполнения следующих условий:

1) исходный временной ряд должен описываться плавной кривой;

2) общие условия, определяющие тенденцию развития изучаемого явления в прошлом и настоящем не должны претерпевать значительных изменений в будущем;

3) исходный ряд динамики должен иметь достаточное число уровней с тем, чтобы отчетливо проявилась тенденция.

Актуальным вопросом прогнозирования по трендовым моделям является проблема точности прогноза. Точная оценка прогноза является весьма условной вследствие следующих причин.

- Выбранная для прогнозирования функция дает лишь приближенную оценку тенденции, так как она не является единственно возможной.
- Статистическое прогнозирование осуществляется на основе ограниченного объема информации, что, в свою очередь, сказывается на величине доверительных интервалов прогноза.
- Наличие в исходном временном ряду случайного компонента приводит к тому, что любой прогноз осуществляется лишь с определенной долей вероятности.

Применение трендовых моделей прогнозирования социально-экономических явлений имеет большую значимость и, несмотря на достаточно простой способ их реализации, часто применяется для прогнозирования сложных социально-экономических явлений. Если выбранная модель тренда достаточно хорошо отражает тенденцию развития ряда, то полученные на ее основе прогнозы практически всегда надежны.

Следует отметить, что прогноз по аналитическому выражению тренда имеет один существенный недостаток: при использовании данного метода прогнозируется лишь детерминированная составляющая ряда динамики и не учитывается случайный компонент. Чтобы избежать данного недочета и сделать прогноз более точным, необходимо определить закономерность изменения во времени случайного компонента. Для этого рассчитывают отклонения от тренда и определяют закономерность их изменения во времени, а затем строят прогноз случайной составляющей временного ряда. Результаты обоих прогнозов объединяются. Описанный метод дает положительные результаты при небольших случайных колебания в ряду динамики.

ПРИМЕР.

Построим прогноз на три периода упреждения средних экспортных цен на нефть Российской Федерации для января–марта 2007 г.

Таблица 4.5.5

Динамика фактических экспортных цен на нефть Российской Федерации за 2006 г.

Месяц	Средняя фактическая экспортная цена нефти, долл. за тонну	t	t^2	ty	$\bar{y}_t$	$(y_i - \bar{y}_t)^2$
Январь	374,3	1	1	374,3	406,51	1037,484
Февраль	383,8	2	4	767,6	407,34	554,1316
Март	383,2	3	9	1149,6	408,17	623,5009
Апрель	418,6	4	16	1674,4	409,00	92,1600
Май	440,1	5	25	2200,5	409,83	916,2729
Июнь	433,2	6	36	2599,2	410,66	508,0516
Июль	455,1	7	49	3185,7	411,49	1901,8320
Август	463,4	8	64	3707,2	412,32	2609,1660
Сентябрь	429,2	9	81	3862,8	413,15	257,6025
Октябрь	392,3	10	100	3923,0	413,98	470,0224
Ноябрь	378,4	11	121	4162,4	414,81	1325,6880
Декабрь	381,3	12	144	4575,6	415,64	1179,2360
Итого	4932,9	78	650	32 182,3	4932,9	11 475,1500

Построим линейную модель тренда вида: $\bar{y}_t = a_0 + a_1 t$. Тогда параметры модели определим через следующую систему уравнений:

$$\begin{cases} 12a_0 + 78a_1 = 4932{,}9 \\ 78a_0 + 650a_1 = 32182{,}3 \end{cases}; \quad \begin{cases} a_0 = 405{,}68 \\ a_1 = 0{,}83 \end{cases}.$$

Следовательно, $\bar{y}_t = 405{,}68 + 0{,}83t$.

Определим среднюю квадратическую ошибку прогноза:

$$\sigma_{\text{ош}} = \sqrt{\frac{\sum (y_i - \bar{y}_t)^2}{n - k - 1}}, \qquad (4.5.10)$$

где k – количество параметров в уравнении тренда.

$$\sigma_{\text{ош}} = \sqrt{\frac{\sum (y_i - \bar{y}_t)^2}{n - k - 1}} = \sqrt{\frac{11475{,}15}{12 - 2 - 1}} = 35{,}71.$$

Для получения прогнозных значений методом аналитического выравнивания необходимо в полученную модель подставить соответствующий период времени *t*.

Таблица 4.5.6

Прогнозные значения средней экспортной стоимости нефти Российской Федерации на январь—март 2007 г.

Месяцы	*t*	Прогнозные значения, полученные методом аналитического выравнивания
Январь	13	$\bar{y}_{01.07} = 405{,}68 + 0{,}83 \cdot 13 = 416{,}47$
Февраль	14	$\bar{y}_{02.07} = 405{,}68 + 0{,}83 \cdot 14 = 417{,}3$
Март	15	$\bar{y}_{03.07} = 405{,}68 + 0{,}83 \cdot 15 = 418{,}13$

Таким образом, прогнозирование методом экстраполяции тренда средней экспортной стоимости нефти по уравнению линейной функции дает достаточно хорошие результаты. Поэтому данный метод может быть использован для краткосрочного прогнозирования.

КОНТРОЛЬНЫЕ ВОПРОСЫ

1. Сформулируйте задачи статистического прогнозирования. Дайте понятие объекта прогнозирования.
2. Что вы понимаете под экстраполяцией? Какие два вида экстраполяции вам известны? В чем отличие экстраполяции от интерполяции уровней ряда динамики?
3. От каких условий зависит точность и надежность построения прогнозов?
4. Что вы понимаете под периодом упреждения прогноза? Какое правило исследователь должен соблюдать при выборе периода упреждения прогноза?
5. В каких случаях при прогнозировании временных рядов используется метод среднего абсолютного прироста?
6. Для каких динамических рядов при построении прогнозов целесообразно использовать метод среднего темпа роста?
7. Какие основные предпосылки должны быть выполнены для расчета прогнозных значений методом среднего абсолютного прироста?
8. Опишите алгоритм реализации метода экстраполяции тренда. Каковы необходимые условия для его применения?

Глава 4.6

ИНДЕКСНЫЙ МЕТОД И ЕГО ПРИМЕНЕНИЕ В АНАЛИЗЕ СОЦИАЛЬНО-ЭКОНОМИЧЕСКИХ ЯВЛЕНИЙ

- Общее понятие индексов и их назначение в экономическом анализе
- Индивидуальные индексы
- Сводные индексы
- Индексный анализ влияния структурных изменений
- Средние формы сводных индексов

4.6.1. Общее понятие индексов и их назначение в экономическом анализе

Существует множество экономических индексов, которые используются для оценки интенсивности изменения важнейших социально-экономических явлений и процессов, например: объема ВВП, инвестиций в основной капитал, валютных курсов и т. д. Индексы относятся к одним из самых популярных статистических показателей. Наиболее известным из них является индекс потребительских цен, который является индикатором изменения цен.

Индекс — важнейший обобщающий показатель. «Индекс» в переводе с латинского означает «указатель» или «показатель». С его помощью можно измерить динамику социально-экономического явления за два или более периода времени, динамику среднего показателя и сопоставить уровни явления в пространстве, по странам, экономическим районам, областям и т. д.

В статистике под **индексом** понимают относительный статистический показатель, который характеризует относительное изменение уровня изучаемого явления в рассматриваемом временном отрезке (*текущем* (*отчетном*) периоде) по сравнению с другим его уровнем, принятым за базу сравнения (*базисным* периодом).

В качестве такой базы сравнения может быть использован как уровень этого явления за какой-либо прошлый период времени (*динамический индекс*), так и уровень того же явления по другой территории (*территориальный индекс*).

Если оценивают изменение отдельной величины (цены конкретного товара, например, автомобиля марки Nissan или объем выпуска ноутбуков определенной модели фирмы SONY) — получают ***индивидуальный***

индекс, если анализируется изменение показателя по всей совокупности (по автомобилям всех марок или всем моделям ноутбуков) — получают ***сводный индекс***.

Сводные индексы являются незаменимым инструментом исследования в тех случаях, когда необходимо сравнить во времени или пространстве две совокупности, элементы которых непосредственно суммировать нельзя. Например, сравнить изменение количества произведенной продукции промышленности за год в абсолютном выражении нельзя, так как оно складывается из объема добычи полезных ископаемых, обрабатывающего производства, объема производства и распределения электроэнергии, объема производства промышленных машин и оборудования и др. несопоставимых по своим свойствам единиц измерения. Но если выразить объем производства через стоимость, умножив количество произведенной продукции на цену единицы каждого вида, и сопоставить суммарные величины стоимости на конец и на начало года, мы узнаем сводный индекс промышленного производства — значимую и экономически обоснованную сводную индексную величину.

Помимо оценки интенсивности изменения изучаемого явления индексы выполняют аналитическую функцию. Например, на прирост объема товарооборота влияет индивидуальное изменение цены на товары и изменение структуры продаж. На основе индексного анализа отслеживают изменения макроэкономических показателей на уровне секторов, отраслей и экономики в целом; анализируют данные государственного бюджета; оценивают влияние различных факторов на изменение размера налоговых отчислений физических и юридических лиц в счет государства; изучают финансовые результаты на микроуровне (индивидуальный анализ показателей прибыли, рентабельности, оборачиваемости активов и т. д.).

В целом индексный метод направлен на решение следующих задач:

1) характеристика общего изменения уровня сложного социально-экономического явления;

2) анализ влияния каждого фактора, составляющего сложное социально-экономическое явление, на изменение индексируемой величины;

3) анализ влияния структурных сдвигов на изменение индексируемой величины.

В практике индексного метода используются следующие общепринятые обозначения: i — индивидуальный индекс; I — сводный индекс; zq — затраты на производство; p — цена; q — количество; pq — товарооборот; z — себестоимость единицы продукции; 1 — текущий период; 0 — базисный период.

Индексный метод служит также для изучения динамики средних величин и выявления факторов, влияющих на динамику средних. С этой целью исчисляется система взаимосвязанных индексов: переменного, постоянного состава и структурных сдвигов.

4.6.2. Индивидуальные индексы

Простейшим показателем, используемым в индексном анализе, является **индивидуальный индекс**, который характеризует изменение во времени экономических величин, относящихся к одному объекту.

Изменение стоимости проданного товара в текущем периоде по сравнению с предыдущим (базисным) определяется на основе индивидуального индекса цен:

$$i_p = \frac{p_1}{p_0}, \tag{4.6.1}$$

где p_1 — цена товара в текущем периоде; p_0 — цена товара в базисном периоде.

Изменение физической массы проданного товара в натуральном выражении определяется **индивидуальным индексом физического объема реализации**:

$$i_q = \frac{q_1}{q_0}, \tag{4.6.2}$$

где q_1 — количество реализованного товара в текущем периоде; q_0 — количество реализованного товара в базисном периоде.

Изменение стоимостного объема товарооборота по данному товару отразится в значении **индивидуального индекса товарооборота (индивидуального индекса стоимости)**. Для его расчета товарооборот текущего периода (произведение цены на количество проданного товара) сравнивается с товарооборотом предшествующего периода:

$$i_{pq} = \frac{p_1 q_1}{p_0 q_0}, \tag{4.6.3}$$

где $p_1 q_1$ — общая стоимость реализованного товара (товарооборот) в текущем периоде; $p_0 q_0$ — общая стоимость реализованного товара (товарооборот) в базисном периоде.

Данный индекс также может быть получен как произведение индивидуального индекса цены и индивидуального индекса физического объема реализации:

$$i_{pq} = i_p i_q = \frac{p_1 q_1}{p_0 q_0},$$

где $i_p = \frac{p_1}{p_0}$; $i_q = \frac{q_1}{q_0}$.

Если рассматривать изменение себестоимости производства отдельного вида продукции, то система индивидуальных индексов будет следующей.

Индивидуальный индекс себестоимости единицы продукции:

$$i_z = \frac{z_1}{z_0}, \quad (4.6.4)$$

где z_1 — себестоимость единицы продукции в текущем периоде; z_0 — себестоимость единицы продукции в базисном периоде.

Индивидуальный индекс физического объема производства продукции:

$$i_q = \frac{q_1}{q_0}, \quad (4.6.5)$$

где q_1 — количество единиц продукции, произведенное в текущем периоде; q_0 — количество единиц продукции, произведенное в базисном периоде.

Индивидуальный индекс затрат на производство данного вида продукции:

$$i_{zq} = \frac{z_1 q_1}{z_0 q_0}, \quad (4.6.6)$$

где $z_1 q_1$ — общие затраты на производство всего объема продукции в текущем периоде; $z_0 q_0$ — общие затраты на производство всего объема продукции в базисном периоде.

Взаимосвязь между индексами:

$$i_z i_q = i_{zq}.$$

Индивидуальные индексы, в сущности, представляют собой относительные показатели динамики, или темпы роста, и по данным за несколько периодов времени могут рассчитываться в цепной или базисной формах.

4.6.3. Сводные индексы

В отличие от индексов индивидуальных, сводные индексы позволяют обобщить показатели по нескольким позициям (товарам, видам продукции). Исходной формой сводного индекса является агрегатная форма.

Агрегатная форма индекса позволяет найти для разнородной совокупности такой общий показатель, в котором можно объединить все ее элементы. При анализе динамики цен индивидуальные цены различных товаров складывать неправомерно, но суммировать товарооборот по этим товарам вполне допустимо. В текущем периоде такой товарооборот по *n* товарам составит:

$$p_1^1 q_1^1 + p_1^2 q_1^2 + p_1^3 q_1^3 + \ldots + p_1^n q_1^n = \sum p_1 q_1 .$$

Аналогично определяется товарооборот для базисного периода:

$$p_0^1 q_0^1 + p_0^2 q_0^2 + p_0^3 q_0^3 + \ldots + p_0^n q_0^n = \sum p_0 q_0 .$$

Если мы сравним товарооборот в текущем периоде с его величиной в базисном периоде, то получим **сводный индекс товарооборота (сводный индекс стоимости)**:

$$I_{pq} = \frac{\sum p_1 q_1}{\sum p_0 q_0}. \qquad (4.6.7)$$

Для иллюстрации этого и последующих индексов воспользуемся следующими условными данными (табл. 4.6.1).

ПРИМЕР.

Имеются следующие данные по курсу цен и количеству проданных акций нефтяных компаний на фондовом рынке (табл. 4.6.1).

Таблица 4.6.1

Цены акций нефтяных компаний, руб.

Вид акций	Сентябрь		Октябрь		Расчетные графы		
	Цена акции p_0	Стоимость проданных акций q_0p_0	Цена акции p_1	Стоимость проданных акций q_1p_1	Количество проданных акций, тыс. шт.		p_0q_1
					q_0	q_1	
А	3680	4416	3700	6300	1,200	1,703	6267
Б	2150	1505	2200	1980	0,700	0,900	1935

Окончание табл. 4.6.1

Вид акций	Сентябрь		Октябрь		Расчетные графы		
	Цена акции p_0	Стоимость проданных акций q_0p_0	Цена акции p_1	Стоимость проданных акций q_1p_1	Количество проданных акций, тыс. шт.		p_0q_1
					q_0	q_1	
В	2620	2560	2750	2780	0,977	1,011	2649
Г	2010	3120	2100	3300	1,552	1,571	3158
Итого	–	11 601	–	14 360	4,429	5,185	14 009

Рассчитаем индекс стоимости:

$$I_{pq} = \frac{14\,360}{11\,601} = 1{,}24, \text{ или } 124\%.$$

Рассчитанное значение индекса позволяет заключить, что стоимость в целом по данным видов акций в текущем периоде по сравнению с базисным увеличилась на 24% (100% – 124%). Отметим, что размер товарной группы (количество товаров), единицы измерения товаров могут быть любыми.

Величина индекса товарооборота формируется под воздействием двух факторов — изменение цен на товары и изменение объемов их реализации. Для того чтобы оценить изменение только цен (индексируемой величины), необходимо количество проданных товаров (веса индекса) зафиксировать на каком-либо постоянном уровне. При исследовании динамики таких показателей, как цена и себестоимость, физический объем реализации обычно фиксируют на уровне текущего периода. Таким способом получают **сводный индекс цен** (по методу Пааше):

$$I_p = \frac{p_1^1 q_1^1 + p_1^2 q_1^2 + \ldots + p_1^n q_1^n}{p_0^1 q_1^1 + p_0^2 q_1^2 + \ldots + p_0^n q_1^n} = \frac{\sum p_1 q_1}{\sum p_0 q_1}. \quad (4.6.8)$$

Для рассматриваемого примера получим:

$$I_p = \frac{14\,360}{14\,009} = 1{,}025, \text{ или } 102{,}5\%.$$

Таким образом, по данной товарной группе (видам акций) цены в октябре по сравнению с сентябрем в среднем увеличились на 2,5%. При построении данного индекса цена выступает в качестве индексируемой величины, а количество проданного товара — в качестве веса.

Рассмотрим сводный индекс цен более подробно. Числитель данного индекса содержит фактический товарооборот текущего периода. Знаменатель же представляет собой условную величину, которая показывает, каким был бы товарооборот в текущем периоде при условии сохранения цен на базисном уровне. Поэтому соотношение этих двух категорий и отражает имевшее место изменение цен.

Числитель и знаменатель сводного индекса цен также можно интерпретировать и по-другому. Числитель представляет собой сумму денег, фактически уплаченных покупателями за товары в текущем периоде. Знаменатель же показывает, какую сумму покупатели заплатили бы за те же товары, если бы цены не изменились. Разность числителя и знаменателя будет отражать величину экономии (если знак «–») или перерасхода («+») покупателей от изменения цен:

$$E = \sum p_1 q_1 - \sum p_0 q_1 = 14\,360 - 14\,009 = 351 \text{ тыс. руб.} \qquad (4.6.9)$$

Следовательно, в связи с увеличением стоимости акций на 2,5% наблюдается перерасход покупателями своих средств на 351 тыс. руб.

Необходимо отметить, что на практике также используется сводный индекс цен, построенный по методу Ласпейреса, когда веса или объемы продаж фиксируются на уровне базисного, а не текущего периода:

$$I_p = \frac{\sum p_1 q_0}{\sum p_0 q_0}.$$

Третьим индексом в рассматриваемой индексной системе (включающей индекс цен, рассчитанный по методу Пааше) является **сводный индекс физического объема реализации**. Он характеризует изменение количества проданных товаров не в денежных, а в физических единицах измерения. Весами в данном случае выступают цены, которые фиксируются на базисном уровне:

$$I_q = \frac{\sum q_1 p_0}{\sum q_0 p_0}. \qquad (4.6.10)$$

В нашем случае индекс составит:

$$I_q = \frac{14\,009}{11\,601} = 1{,}208, \text{ или } 102{,}8\%.$$

Физический объем проданных ценных бумаг увеличился в 1,208 раза.

Вычитая из числителя знаменатель $\sum p_1 q_0 - \sum q_0 p_0 = 14\,009 - 11\,601 = 2408$, определяем, что в абсолютном выражении за счет увеличения количества проданных акций стоимость ценных бумаг в отчетном периоде по сравнению с базисным возросла на 2408 тыс. руб.

Между рассчитанными индексами существует следующая взаимосвязь:

$$I_p \cdot I_q = I_{pq}. \tag{4.6.11}$$

Для проверки взаимосвязи переведем значения индексов из процентов в доли единицы и подставим их в формулу (4.6.10): $1{,}025 \cdot 1{,}208 = 1{,}24$, или 124%, как и по формуле (4.6.7).

На основе данной взаимосвязи по значениям двух известных индексов всегда можно определить неизвестное значение третьего индекса.

Для оценки изменения уровня производственных издержек рассчитываются сводные индексы затрат (издержек) производства, себестоимости произведенной продукции и физического объема производства.

Сводный индекс затрат на производство оценивает, во сколько раз возросли или уменьшились издержки производства продукции в текущем периоде по сравнению с базисным или сколько процентов от затрат базисного уровня они составили:

$$I_{zq} = \frac{\sum z_1 q_1}{\sum z_0 q_0}. \tag{4.6.12}$$

Разница между числителем и знаменателем сводного индекса затрат покажет, на сколько рублей увеличились или уменьшились издержки производства в текущем периоде в сравнении с базисным.

Сводный индекс себестоимости продукции показывает, как изменились издержки производства из-за изменения себестоимости в текущем периоде по сравнению с базисным:

$$I_z = \frac{\sum z_1 q_1}{\sum z_0 q_1}. \tag{4.6.13}$$

Разность числителя и знаменателя сводного индекса себестоимости характеризует, на сколько рублей изменились издержки производства в результате роста (уменьшения) себестоимости продукции.

Индекс физического объема производства оценивает, во сколько раз изменились издержки производства продукции в результате изменения объема ее производства или сколько процентов составил рост (снижение) издержек из-за изменения физического объема ее производства:

$$I_q = \frac{\sum z_0 q_1}{\sum z_0 q_0}. \tag{4.6.14}$$

Взаимосвязь между индексами выражается формулой:

$$I_z \cdot I_q = I_{zq}. \tag{4.6.15}$$

ПРИМЕР.

Процесс определения всевозможных индексов и факторного анализа сложного явления рассмотрим на примере текущих затрат на выпуск однородной продукции тремя фирмами-изготовителями. Исходные данные приведены в табл. 4.6.2.

Таблица 4.6.2

Выпуск и себестоимость продукции, произведенной тремя фирмами-изготовителями

Фирма	I квартал (базисный период)		I квартал (отчетный период)		Расчетные графы		
	Выпущено продукции q_0, шт.	Себестоимость единицы продукции z_0, тыс. руб.	Выпущено продукции, шт.	Себестоимость единицы продукции, тыс. руб.	z_0q_0	z_1q_1	z_0q_1
1	200	5,0	800	4,8	1000	3840	4000
2	500	3,5	650	3,4	1750	2210	2275
3	300	4,0	550	3,8	1200	2090	2200
Итого	1000		2000		3950	8140	8475

1. Определим свободный индекс затрат на производство по всем фирмам в целом на основании формулы (4.6.13):

$$I_{zq} = \frac{\sum z_1 q_1}{\sum z_0 q_0} = \frac{8140}{3950} = 2{,}06.$$

Таким образом, во втором квартале по сравнению с первым издержки производства продукции увеличились в 2 раза, что составило 4190 тыс. руб. ($\sum z_1 q_1 - \sum z_0 q_0 =$ $= 8140 - 3950 = 4190$ тыс. руб.).

2. Определим, за счет каких факторов произошло столь значительное увеличение. С помощью сводного индекса себестоимости продукции вычислим, каково было влияние изменения себестоимости в текущем периоде по сравнению с базисным, используя формулу (4.6.14):

$$I_z = \frac{\sum z_1 q_1}{\sum z_0 q_1} = \frac{8140}{8475} = 0{,}96.$$

За счет увеличения себестоимости продукции издержки снизились на 4%. В результате общего снижения себестоимости единицы продукции издержки производства снизились на 335 тыс. руб.

3. Используя индекс физического объема производства, определим, во сколько раз изменились издержки производства продукции в результате изменения объема ее производства согласно формуле (4.6.14):

$$I_q = \frac{\sum z_0 q_1}{\sum z_0 q_0} = \frac{8475}{3950} = 2{,}146.$$

Следовательно, за счет увеличения объема производства продукции издержки производства продукции увеличились в 2 раза.

4. Проверим взаимосвязь между рассчитанными индексами:

$$I_{zq} = I_z \cdot I_q = 0{,}96 \bullet 2{,}15 = 2{,}06.$$

4.6.4. Индексный анализ влияния структурных изменений

Индексы позволяют оценить динамику показателей, характеризующих разнородные в качественном отношении совокупности, как правило, товарные группы. Однако, даже если рассматривать данные по каждому товару отдельно или о производстве продукции одного вида, на величине результативного показателя будет отражаться влияние структурных изменений, например изменений в структуре производства или реализации данного товара по территориям. Рассмотрим случай, когда один товар или вид продукции реализуется в нескольких местах (табл. 4.6.3).

Таблица 4.6.3

Данные о ценах и объемах реализации товара X в регионах Российской Федерации

Регион	2005 г.		2006 г.	
	Цена p_0, тыс. руб.	Продано q_0, шт.	Цена p_1, тыс. руб.	Продано q_1, шт.
Воронежская обл.	50	15 000	54	27 000
Курская обл.	70	30 000	78	135 000
Итого	–	45 000	–	62 000

Проведем анализ изменения цен на данный товар. Из таблицы видно, что цена в каждом регионе возросла. Для сводной оценки этого роста воспользуемся средними показателями. Так как в данном случае реализуется один и тот же товар, вполне правомерно рассчитать его среднюю цену за 2006 и за 2005 гг. **Индекс цен переменного состава** представляет собой соотношение средних значений за два рассматриваемых периода:

$$I_{\text{р}}^{\text{пс}} = \frac{\sum p_1 q_1}{\sum q_1} : \frac{\sum p_0 q_0}{\sum q_0}. \qquad (4.6.16)$$

$$I_{\text{р}}^{\text{пс}} = \frac{54 \cdot 27000 + 78 \cdot 35000}{27000 + 35000} : \frac{50 \cdot 15000 + 70 \cdot 30000}{15000 + 30000} = \frac{67{,}5}{63{,}3} = 1{,}066, \text{ или } 106{,}6\%.$$

Рассчитанное значение индекса указывает на увеличение средней цены данного товара на 6,6%, т. е. с 63,3 тыс. руб. до 67,5 тыс. руб. На динамике средней цены данного товара отразились структурные сдвиги в рассматриваемой совокупности. Оценить воздействие этого фактора можно с помощью **индекса структурных сдвигов**:

$$I_{\text{р}}^{\text{ст}} = \frac{\sum p_0 q_1}{\sum q_1} : \frac{\sum p_0 q_0}{\sum q_0}. \qquad (4.6.17)$$

$$I_{\text{р}}^{\text{ст}} = \frac{50 \cdot 27\,000 + 70 \cdot 35\,000}{62\,000} : \frac{50 \cdot 15\,000 + 70 \cdot 30\,000}{45\,000} = \frac{61{,}3}{63{,}3} = 0{,}968, \text{ или } 96{,}8\%.$$

Первая формула в этом индексе позволяет ответить на вопрос, какой была бы средняя цена в 2006 г., если бы цены в каждом регионе сохранились на уровне 2005 г. Вторая часть формулы отражает фактическую среднюю цену 2006 г. В целом по значению индекса мы можем сделать вывод, что за счет структурных сдвигов цены снизились на 3,2% (96,8% – 100% = –3,2%).

Последним в данной системе является **индекс цен фиксированного состава**, который не учитывает влияние структуры и рассчитывается по уже известной нам формуле сводного индекса цен по методу Пааше:

$$I_{\text{р}}^{\text{фс}} = \frac{\sum p_1 q_1}{\sum p_0 q_1} = \frac{1458 + 2730}{1350 + 2450} = \frac{4188}{3800} = 1{,}102, \text{ или } 110{,}2\%.$$

Индекс цен фиксированного состава равен 110,2%, что позволяет сделать следующий вывод: если бы структура продаж не изменилась, средняя цена возросла бы на 10,2%, однако это не произошло, так как влияние структурных сдвигов оказалось сильнее.

Данное взаимодействие рассматриваемых факторов отражается в следующей взаимосвязи:

$$I_{\text{р}}^{\text{пс}} = I_{\text{р}}^{\text{фс}} \cdot I_{\text{р}}^{\text{ст}} = 1{,}102 \cdot 0{,}968 = 1{,}066. \qquad (4.6.18)$$

Аналогично строятся индексы структурных сдвигов, переменного и фиксированного состава для анализа изменения себестоимости, урожайности и других показателей.

4.6.5. Средние формы сводных индексов

В отличие от агрегатной формы индекса, средние индексы используются тогда, когда имеется информация об изменении индексируемой величины по отдельным единицам исследуемой совокупности.

Относительное изменение среднего уровня качественного показателя характеризуется с помощью системы индексов переменного, постоянного состава и структурных сдвигов, позволяющих оценить влияние каждого фактора на его динамику.

На практике при расчете индексов часть необходимой информации может отсутствовать или базироваться на результатах выборочных обследований. В подобных случаях вместо индексов в агрегатной форме удобнее использовать средние арифметические и средние гармонические индексы. Любой сводный индекс можно представить как среднюю взвешенную из индивидуальных индексов.

Предположим, мы располагаем данными о стоимости проданной продукции в текущем периоде и индивидуальными индексами цен, полученными, например, в результате выборочного наблюдения. Тогда при расчете сводного индекса цен по методу Пааше можно использовать следующую замену:

$$p_0 q_1 = \frac{1}{i_p} p_1 q_1 .$$

В целом же **сводный индекс цен** в данном случае будет выражен **в форме среднего гармонического**:

$$I_p = \frac{\sum p_1 q_1}{\sum \frac{1}{i_p} p_1 q_1}. \qquad (4.6.19)$$

Для получения значения, соответствующего индексу Ласпейреса, сводный индекс цен необходимо представить в средней арифметической форме. При этом используется следующая замена:

$$p_1 q_0 = i_p p_0 q_0 .$$

С учетом этой замены **сводный индекс цен в средней арифметической форме** можно представить следующим образом:

$$I_p = \frac{\sum i_p p_0 q_0}{\sum p_0 q_0}. \qquad (4.6.20)$$

Средняя арифметическая форма также может использоваться при расчете сводного индекса физического объема товарооборота. При этом производится замена:

$$q_1 p_0 = i_q q_0 p_0 .$$

Сводный индекс физического объема товарооборота в средней арифметической форме имеет вид:

$$I_q = \frac{\sum i_q q_0 p_0}{\sum q_0 p_0}. \qquad (4.6.21)$$

ПРИМЕР.

Рассмотрим методику расчета среднего арифметического индекса на примере. Имеются следующие данные о реализации промышленной продукции предприятия (табл. 4.6.4).

Таблица 4.6.4

Данные о реализации овощной продукции предприятия

Товар	Розничный товарооборот базисного периода p_0q_0, руб.	Изменение физического объема реализации в текущем периоде по сравнению с базисным, %
А	25 000	–6,5
Б	19 000	–8,0
В	30 000	+1,5

Последняя графа таблицы содержит информацию об изменениях индивидуальных индексов физического объема цен или их приростах. С учетом этих приростов несложно определить первоначальные значения индексов, которые по товарам А, Б и В соответственно составляют 0,935, 0,920 и 1,015. Тогда свободный индекс физического объема согласно формуле составит:

$$I_q = \frac{0{,}935 \cdot 25000 + 0{,}920 \cdot 19000 + 1{,}015 \cdot 30000}{25000 + 19000 + 30000} = 0{,}964.$$

В результате расчета мы получили, что физический объем реализации продукции рассматриваемой товарной группы в среднем снизился на 3,6%.

Рассмотрим методику расчета среднего гармонического индекса на примере.

ПРИМЕР.

Имеются данные о реализации отдельных видов товаров предприятия розничной торговли округа (табл. 4.6.5).

Таблица 4.6.5

Реализация товаров А, Б, В в натуральном и стоимостном выражении

Товар	Розничный товарооборот текущего периода, тыс. руб.	Индивидуальные индексы цен, %
А	1270	104,2
Б	920	102,3
В	1130	99,0

Рассчитаем значение сводного индекса цен в форме средней гармонической (4.6.15):

$$I_p = \frac{1270 + 920 + 1130}{\dfrac{1270}{1{,}042} + \dfrac{920}{1{,}023} + \dfrac{1130}{0{,}99}} = 1{,}019.$$

Произведенный расчет позволяет заключить, что цены по данной товарной группе в среднем возросли на 1,9%. Мы получили значение сводного индекса цен в среднегармонической форме, соответствующее сводному индексу Пааше.

КОНТРОЛЬНЫЕ ВОПРОСЫ

1. Что характеризует экономический индекс? В каких единицах измерения выражаются индексы?
2. Чем отличается индивидуальный индекс от сводного индекса?
3. Что показывает разность между числителем и знаменателем сводного индекса цен? Как интерпретируется ее отрицательное значение?
4. При помощи какого показателя можно оценить снижение издержек производства, вызванное изменением себестоимости единиц выпускаемой продукции?
5. Что представляет собой индекс в форме средней?
6. Что оценивает индекс структурных сдвигов?

Раздел 5

ВЫБОРОЧНЫЙ МЕТОД НАБЛЮДЕНИЯ В СТАТИСТИЧЕСКИХ ИССЛЕДОВАНИЯХ

Глава 5.1

ОБЩИЕ ПОНЯТИЯ ТЕОРИИ ВЫБОРОЧНОГО НАБЛЮДЕНИЯ

- Значение и теоретические основы выборочного наблюдения
- Основные преимущества и недостатки выборочного наблюдения
- Средняя и предельная ошибка выборки. Построение доверительных границ для среднего и доли
- Виды выборочного наблюдения. Способы отбора единиц в выборочную совокупность

5.1.1. Значение и теоретические основы выборочного наблюдения

Представьте такую ситуацию: вы пришли на работу в понедельник в 10 : 00 утра и прочитали служебную записку своего начальника, в которой он просит отследить реакцию клиентов вашей фирмы на несколько новых видов оказываемых услуг вашей компании. Вам необходимо предоставить доклад во вторник на девятичасовом совещании совета директоров. Каким образом возможно решение данной проблемы? Очевидно, что следует провести телефонный опрос клиентов компании, причем на одного клиента вам потребуется минимум 10–15 минут. В вашу базу данных внесено 1479 клиентов, и, конечно, за такое короткое время, располагая только тремя сотрудниками, невозможно опросить всех клиентов фирмы. В этой ситуации целесообразно построить выборку из генеральной совокупности всех клиентов компании, внесенных в базу данных. Таким образом, необходимо будет обзвонить отобранное количество клиентов. При этом следует надеяться, что выборка будет являться достаточно репрезентативной для более крупной генеральной совокупности и в отчете, подготовленном к заседанию совета директоров, основные факты будут учтены.

В статистической практике достаточно редко используют в качестве источника информации данные сплошного обследования. Гораздо чаще применяют такие несплошные методы, как выборочные, монографические и обследования основного массива.

О значении выборки можно судить по такому факту: нет ни одной отрасли экономики, где при сборе информации не прибегали бы к выборочному наблюдению. Выборочное наблюдение в настоящее время находит достаточно широкое применение в обследованиях промышленных и сельскохозяйственных предприятий, в изучении цен на потребительском рынке, в обследованиях бюджетов и занятости населения. Выборочный метод является важнейшим источником информации в контроле качества продукции, в маркетинговых и социологических исследованиях.

Выборочным наблюдением называется такое несплошное обследование, при котором признаки регистрируются у отдельных единиц изучаемой статистической совокупности, отобранных с использованием специальных методов, а полученные в процессе обследования результаты с определенным уровнем вероятности распространяются на всю исходную совокупность. Полученные статистические характеристики сопровождаются указанием на точность оценивания, которая зависит от репрезентативности (представительности) выбранной совокупности, т. е. насколько качественно она отражает закономерность генеральной совокупности.

Выборочное наблюдение нельзя отождествлять с несплошным обследованием вообще, так как оно является лишь одним из видов последнего. Выборочный метод позволяет получить достоверные результаты лишь тогда, когда соблюдается принцип равновозможности быть отобранной для каждой единицы. При этом единственным фактором, влияющим на включение в выборочную совокупность рассматриваемой единицы, является случай, за исключением механического отбора. Из всех методов несполошного наблюдения выборочный является наиболее теоретически разработанным и обоснованным в силу того, что положенный в его основу принцип случайности позволяет математически описать дальнейшее распространение выборочных характеристик на всю сововкупность.

В теории выборочного наблюдения используются специальные понятия и определения. Как вы помните, под **совокупностью** в статистике понимают группу объектов или явлений, объединенных каким-либо общим признаком или свойством качественного или количественного характера. Реализация выборочного метода базируется на понятиях генеральной и выборочной совокупностей.

Генеральной совокупностью называется вся исходная изучаемая статистическая совокупность, из которой на основе отбора единиц или групп единиц формируется совокупность выборочная. Поэтому генеральную совокупность также называют основой выборки. Количество единиц генеральной совокупности обозначается N. **Выборочная совокупность** — это небольшой набор объектов, извлеченных из генеральной совокупности по определенным правилам. Количество единиц выборочной совокупности обозначается n.

Доля выборочных единиц в объеме генеральной совокупности, выраженная в процентах, называется **долей отбора** (процентом выборки) и определяется следующим образом:

$$f = \frac{n}{N}100\%. \tag{5.1.1}$$

Например, при объеме генеральной совокупности в 2000 единиц и выборочной в 500 единиц говорят о 25%-й выборке.

Единицами отбора называют элементы отбора при формировании выборочной совокупности. **Единицей наблюдения** выступает объект, признаки которого подлежат регистрации. Единица наблюдения может не совпадать с единицей отбора.

Если исследуется количественный признак, то непосредственной задачей выборочного наблюдения выступает оценка среднего и суммарного значения. Генеральное среднее значение определяется по данным генеральной совокупности и обозначается $\bar{x}$. Форма для его расчета имеет вид:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{N}. \tag{5.1.2}$$

Выборочное среднее ($\tilde{x}$) является несмещенной, состоятельной и эффективной оценкой генерального среднего и определяется по формуле:

$$\tilde{x} = \frac{\sum_{i=1}^{n} x_i}{n}. \tag{5.1.3}$$

Генеральная дисперсия единиц количественного признака определяется по формуле:

$$\sigma^2_{ген} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{N}. \quad (5.1.4)$$

Так как генеральная дисперсия по большей части в ходе исследования остается неизвестной, то условно ее принимают равной дисперсии, рассчитываемой по выборочным данным:

$$\sigma^2_{выб} = \frac{\sum_{i=1}^{n}(x_i - \tilde{x})^2}{n}. \quad (5.1.5)$$

При малочисленных выборках в формулу необходимо внести поправку:

$$\sigma^2_{мал.\ выб.} = \frac{\sum(x_i - \tilde{x})^2}{n-1}. \quad (5.1.6)$$

Помимо определения характеристик количественных признаков можно произвести оценку характеристик альтернативных показателей. Численность единиц, обладающих изучаемыми признаками в генеральной совокупности, обозначается M, а в выборочной — m. Тогда доля единиц, обладающих исследуемыми признаками в генеральной совокупности, определяется как:

$$p = \frac{M}{N}, \quad (5.1.7)$$

в выборочной совокупности:

$$w = \frac{m}{n}. \quad (5.1.8)$$

Генеральная дисперсия доли альтернативного признака рассчитывается по формуле:

$$\sigma^2_p = pq, \quad (5.1.9)$$

где p — доля единиц, обладающих исследуемым признаком; q — доля единиц, не обладающих исследуемым признаком ($q = 1 - p$).

Выборочная дисперсия доли:

$$\sigma^2_w = w(1 - w). \quad (5.1.10)$$

При рассмотрении теории и методов выборочного наблюдения в статистике используются следующие общепринятые условные обозначения:

N — объем (число единиц) генеральной совокупности;

n — объем (число единиц) выборочной совокупности;

$\bar{x}$ — генеральная средняя, т. е. среднее значение изучаемого признака по генеральной совокупности;

$\tilde{x}$ — выборочная средняя, т. е. среднее значение изучаемого признака по выборочной совокупности;

M — численность единиц генеральной совокупности, обладающих определенным вариантом или вариантами изучаемого признака;

p — генеральная доля, т. е. доля единиц, обладающих определенным вариантом или вариантами изучаемого признака, во всей генеральной совокупности;

m — численность единиц выборочной совокупности, обладающих определенным вариантом или вариантами изучаемого признака;

w — выборочная доля, т. е. доля единиц, обладающих определенным вариантом или вариантами изучаемого признака, в выборочной совокупности;

μ — средняя ошибка выборки;

Δ — предельная ошибка выборки.

5.1.2. Основные преимущества и недостатки выборочного наблюдения

Широкое применение выборочного метода в статистической практике вызвано рядом преимуществ по сравнению со сплошным методом.

1. *Быстрота получения результатов обследования.* Значительное снижение объема наблюдаемой совокупности за счет отбора только определенной части единиц позволяет в короткие сроки собрать информацию и оперативно получить сводные результаты обследования. Кроме того, выборочный метод используется для ускоренной обработки результатов и проверки данных сплошного наблюдения.

2. *Значительное снижение стоимости наблюдения.* Все затраты на организацию наблюдения можно разделить на две группы: затраты, не зависящие от числа единиц наблюдения, и затраты, прямо пропорциональные числу единиц обследования. При использовании выборочного обследования можно существенно снизить вторую группу затрат. Наи-

большее значение данный аспект имеет при значительном объеме генеральной совокупности, так как для обеспечения одной и той же точности в больших по объему совокупностях при прочих равных условиях требуется значительно меньшая доля отбираемых единиц по сравнению с совокупностями, меньшими по численности.

3. *Возможность лучшей организации проведения обследования.* При проведении выборочного наблюдения возникает два вида ошибок: ошибки репрезентативности, обусловленные тем, что наблюдаются не все единицы совокупности, а лишь их часть, и ошибки регистрации, присущие также и сплошному наблюдению. Использование выборки позволяет минимизировать ошибки регистрации за счет лучшей организации проведения обследования. Следовательно, обеспечивается большая достоверность получаемых сведений.

4. *Возможность расширения программы наблюдения.* Снижение численности наблюдаемых единиц совокупности позволяет изучить их более детально путем расширения программы наблюдения.

5. *Возможность использования в случаях, когда проведение сплошного наблюдения методологически невозможно.* Например, при статистическом исследовании качества продукции (продолжительность горения электрических ламп, дегустация продуктов питания и т. д.) выборочное наблюдение становится единственно возможным, так как данные виды исследования связаны с порчей или уничтожением продукции.

Если генеральная совокупность объектов изучения бесконечно велика и нет возможности обследовать каждую единицу, также используется выборочный метод наблюдения (например, маркетинговые обследования покупателей, изучение пассажиропотоков и т. д.).

Наиболее существенным недостатком выборочного наблюдения является наличие ошибок репрезентативности. Ошибки получаемых выборочных характеристик можно разделить на две группы:

- ошибки регистрации;
- ошибки репрезентативности.

Ошибки регистрации являются следствием неправильного установления значения наблюдаемого признака или неправильной записи. Они свойственны не только выборочному, но и сплошному наблюдению.

Ошибки репрезентативности обусловлены тем обстоятельством, что выборочная совокупность не может по всем параметрам в точности воспроизвести совокупность генеральную. Получаемые расхождения или ошибки репрезентативности позволяют заключить, в какой степени попавшие в выборку единицы могут представлять всю генеральную совокупность. При этом следует различать систематические и случайные ошибки репрезентативности.

Систематические ошибки репрезентативности связаны с нарушением принципов формирования выборочной совокупности. Например, вследствие каких-либо причин, связанных с организацией отбора, в выборку попали единицы, характеризующиеся несколько большими или, наоборот, несколько меньшими по сравнению с другими единицами значениями наблюдаемых признаков. В этом случае и рассчитанные выборочные характеристики будут завышенными или заниженными.

Случайные ошибки репрезентативности обусловлены действием случайных факторов, не содержащих каких-либо элементов системности в направлении воздействия на рассчитываемые выборочные характеристики. Но даже при строгом соблюдении всех принципов формирования выборочной совокупности выборочные и генеральные характеристики будут несколько различаться. Получаемые случайные ошибки могут быть статистически оценены и учтены при распространении результатов выборочного наблюдения на всю генеральную совокупность.

Выборочное наблюдение. как бы грамотно с методологической точки зрения оно не было организовано, всегда связано с определенными, пусть небольшими и измеряемыми ошибками. Поэтому, когда вариация регистрируемых признаков очень сильная и процент отбора для получения выборочных значений с заданной точностью достигает 20–25%, следует правильно оценить целесообразность несплошного обследования, сопоставив достаточно большие затраты всех ресурсов на такую объемную выборку и ожидаемые погрешности статистических характеристик. Вполне вероятно, что проведение сплошного обследования в подобных случаях будет более оправданным.

5.1.3. Средняя и предельная ошибки выборки. Построение доверительных границ для среднего и доли

Из одной и той же генеральной совокупности объема N можно извлечь множество различных выборок заданного объема n. Тогда в каждом случае рассчитанные отклонения выборочных характеристик от генеральных будут различны. Если определить среднюю из ошибок всех возможных выборок заданного объема, извлеченных из одной и той же генеральной совокупности, то получим их обобщающую характеристику — **среднюю ошибку выборки** (μ), которая показывает, насколько отклоняется в среднем параметр выборочной совокупности от соответствующего параметра генеральной. В теории выборочного наблюдения выведены формулы для определения средней ошибки выборки, которые индивидуальны для различных способов отбора (повторного и бесповторного), типов используемых выборок и видов оцениваемых статистических показателей.

Как уже отмечалось, отбор единиц в выборочную совокупность может быть повторным или бесповторным. При **повторном отборе** попавшая в выборку единица подвергается обследованию, т. е. регистрации значений ее признаков, возвращается в генеральную совокупность и наравне с другими единицами участвует в дальнейшей процедуре отбора. Таким образом, некоторые единицы могут попадать в выборку дважды, трижды или даже большее число раз. И при изучении выборочной совокупности они будут рассматриваться как отдельные независимые наблюдения. Число единиц генеральной совокупности, участвующих в отборе, при таком подходе остается постоянным. Поэтому вероятность попадания в выборку для всех единиц совокупности на протяжении всего процесса отбора также не меняется.

На практике методология повторного отбора обычно используется в тех случаях, когда объем генеральной совокупности не известен и теоретически возможно повторение единиц с уже встречавшимися значениями всех регистрируемых признаков.

Например, при проведении маркетинговых исследований мы не можем сколько-нибудь точно оценить, какое число потребителей предпочитает стиральный порошок конкретной торговой марки, сколько покупателей предпочитают делать покупки именно в данном супермаркете и т. д. Поэтому возможно повторение совершенно идентичных единиц как по причине практически неограниченных объемов совокупности, так и вследствие возможной повторной регистрации. Предположим, при проведении обследования один и тот же покупатель может дважды прийти в магазин и дважды подвергнуться обследованию.

При **бесповторном отборе** попавшая в выборку единица подвергается обследованию и в дальнейшей процедуре отбора не участвует. Такой отбор целесообразен и практически возможен в тех случаях, когда объем генеральной совокупности четко определен. Получаемые при этом результаты, как правило, являются более точными по сравнению с результатами, основанными на повторной выборке.

Ошибка выборки или отклонение выборочной средней от средней генеральной находится в прямой зависимости от дисперсии изучаемого признака в генеральной совокупности, и в обратной зависимости — от объема выборки. Таким образом, среднюю ошибку выборки можно представить как

$$\mu = \sqrt{\frac{\sigma^2_{ген}}{n}}. \qquad (5.1.11)$$

При проведении выборочного наблюдения дисперсия изучаемого признака в генеральной совокупности, как правило, не известна. В то же время между генеральной дисперсией и средней из всех возможных выборочных дисперсий существует следующая зависимость:

$$\sigma^2_{ген} = \bar{\sigma}^2 \cdot \frac{n}{n-1}. \qquad (5.1.12)$$

В связи с тем, что на практике в большинстве случаев из генеральной совокупности в определенный момент времени производится только одна выборка, дисперсия изучаемого признака по этой выборке и используется при расчете ошибки. Необходимо учесть, что при достаточно большом объеме выборки отношение $\frac{n}{n-1}$ близко к 1.

Если применяется повторная собственно случайная выборка, то μ при оценивании среднего значения определяется как

$$\mu = \sqrt{\frac{\sigma^2}{n}}, \qquad (5.1.13)$$

где σ^2 — дисперсия изучаемого признака по выборочной совокупности.

Если признак альтернативный, то μ при оценивании доли определяется по формуле:

$$\mu = \sqrt{\frac{w(1-w)}{n}}. \qquad (5.1.14)$$

При бесповторной собственно случайной выборке учитывается поправка на конечность совокупности $(1-\frac{n}{N})$:

- для среднего:

$$\mu = \sqrt{\frac{\sigma^2}{n}\left(1-\frac{n}{N}\right)}; \qquad (5.1.15)$$

- для доли:

$$\mu = \sqrt{\frac{w(1-w)}{n}\left(1-\frac{n}{N}\right)}. \qquad (5.1.16)$$

При определении возможных границ значений характеристик генеральной совокупности рассчитывается **предельная ошибка выборки** (Δ), которая зависит от величины ее средней ошибки и уровня вероятности, с которым гарантируется, что генеральная средняя не выйдет за указанные границы: $\Delta = t\mu$. Уровень предельной ошибки выборки зависит от следующих факторов:

- степени вариации единиц генеральной совокупности;
- объема выборки;
- выбранных схем отбора: повторного и бесповторного;
- уровня доверительной вероятности.

Согласно теореме А. М. Ляпунова, вероятность той или иной величины предельной ошибки, при достаточно большом объеме выборочной совокупности, подчиняется нормальному закону распределения и может быть определена на основе интеграла Лапласа.

Значения интеграла Лапласа при различных величинах t табулированы и представлены в статистических справочниках. Если объем выборки больше 30, то значение t определяется по таблице нормального распределения, если меньше — по таблице распределения Стьюдента. При обобщении результатов выборочного наблюдения наиболее часто используются следующие уровни вероятности и соответствующие им значения t.

Значение доверительной вероятности P	0,6827	0,8664	0,9545	0,9973
Значение коэффициента доверия t	1,0000	1,5000	2,0000	3,0000

Например, если при расчете предельной ошибки выборки мы используем значение $t = 2$, то с вероятностью 0,954 можно утверждать, что расхождение между выборочной средней и генеральной средней не превысит двухкратной величины средней ошибки выборки.

Построение доверительных интервалов для генеральной средней и доли осуществляется следующим образом:

$$\begin{aligned} &\tilde{x} - \Delta_{\tilde{x}} < \bar{x} < \tilde{x} + \Delta_{\tilde{x}} \\ &w - \Delta_{\tilde{w}} < p < w + \Delta_{\tilde{w}} \end{aligned} . \quad (5.1.17)$$

Определение границ генеральной средней и доли состоит из следующих этапов:

- нахождение выборочного значения средней (или доли);
- определение средней ошибки выборки в соответствии с выбранной схемой отбора и видом выборки;
- задание доверительной вероятности P и определение коэффициента доверия t;
- вычисление предельной ошибки выборки Δ;
- построение доверительного интервала для среднего (доли).

5.1.4. Виды выборочного наблюдения. Способы отбора единиц в выборочную совокупность

В зависимости от состава и структуры генеральной совокупности различают виды выборок и способы отбора единиц. Выбор вида выборок определяется задачами исследования и прежде всего полнотой и особенностями информации, которой располагает исследователь об объекте наблюдения. Наиболее распространенными видами выборок, применяемыми на практике, являются:

- собственно случайная выборка;
- механическая выборка;
- типическая (стратифицированная, расслоенная) выборка;
- серийная выборка.

Отбор единиц из генеральной совокупности может быть осуществлен по одному из следующих методов:

- метод индивидуального отбора;
- метод группового отбора;
- метод комбинированного отбора.

При *индивидуальном отборе* в выборочную совокупность извлекаются отдельные единицы генеральной совокупности, например: при обследовании коммерческих банков — банки, при обследовании населения — конкретные люди и т. д. Индивидуальный отбор применяется при организации собственно случайной, механической и типической выборок.

Групповой отбор предполагает извлечение единиц из генеральной совокупности группами, например, бригады, микрорайоны и т. д. Групповой отбор применяется при серийной выборке.

При *комбинированном отборе* сочетаются индивидуальный и групповой отборы. Таким образом, отбираются сначала группы единиц (групповой отбор), а затем из них случайным образом извлекаются конкретные единицы совокупности (индивидуальный отбор). Ошибка такой выборки определяется ступенчатостью отбора.

При проведении перечисленных выше видов отбора можно использовать один из указанных ранее способов: бесповторный или повторный отбор.

КОНТРОЛЬНЫЕ ВОПРОСЫ

1. Что вы понимаете под выборочным наблюдением? Какие преимущества и недостатки по сравнению со сплошным наблюдением имеет выборочное наблюдение?
2. Какие теоремы теории вероятности послужили основой выборочного метода?
3. Дайте определения выборочной и генеральной совокупности. В чем отличие данных понятий? Приведите практические примеры.
4. В чем отличие бесповторного отбора от повторного?
5. Что показывает предельная ошибка выборки? От каких основных факторов зависит ее уровень?
6. Что показывает средняя ошибка выборки? Приведите формулы для ее расчета в случае оценивания генеральной доли при бесповторном отборе.
7. Что вы понимаете под доверительной вероятностью и коэффициентом доверия? Какая существует между ними зависимость?
8. Что такое репрезентативная выборка? Что такое смещенная выборка? Как можно извлечь репрезентативную выборку?
9. Опишите этапы определения границ генеральной средней и доли.
10. Назовите виды и способы отбора единиц из генеральной совокупности объектов.

Глава 5.2
ВИДЫ ВЫБОРОК

- Собственно случайная выборка
- Систематическая (механическая) выборка
- Стратифицированная (типическая) выборка

5.2.1. Собственно случайная выборка

Собственно случайная выборка заключается в отборе единиц из генеральной совокупности в целом, без разделения ее на группы, подгруппы или серии отдельных единиц. При этом единицы отбираются в случайном порядке, не зависящем ни от последовательности расположения единиц в совокупности, ни от значений их признаков. Простая случайная выборка строится исходя из следующих правил:

1) каждый объект генеральной совокупности имеет одинаковую вероятность быть отобранным;

2) объекты отбираются независимо друг от друга.

Таким образом, прежде чем производить собственно случайный отбор, необходимо убедиться, что все без исключения единицы генеральной совокупности имеют абсолютно равные шансы попадания в выборку, в списках или перечне отсутствуют пропуски, игнорирование отдельных единиц и т. п. Следует также установить четкие границы генеральной совокупности таким образом, чтобы включение или невключение в нее отдельных единиц не вызывало сомнений. Так, например, при обследовании торговых предприятий необходимо указать, включит ли генеральная совокупность торговые павильоны, коммерческие палатки, передвижные торговые точки и прочие подобные объекты; при обследовании студентов важно определиться, будут ли приниматься во внимание студенты-заочники, экстерны, учащиеся в магистратуре, лица, находящиеся в академическом отпуске и т. п.

Если элементы совокупности имеют равную вероятность быть извлеченными, то полученная случайная выборка будет достаточно хорошей и относительно несмещенной. Независимость отбора обеспечивает сбор максимально возможного объема независимой информации. Так как индивидуальные вкусы и человеческий фактор полностью исключены из процесса отбора, у полученной таким образом выборки будет больше шансов быть репрезентативной, чем у произвольной выборки, сделанной исследователем.

Оценим преимущества и недостатки случайной выборки перед произвольной. При извлечении случайной выборки существует гарантия того, что показатели математической статистики будут удовлетворительными. Таким образом, выборка будет являться репрезентативной, как минимум в среднем, для всех характеристик генеральной совокупности. Кроме того, случайная выборка закладывает основу для корректности статистических выводов относительно генеральной совокупности, которые могут быть сделаны из данных этой выборки. С другой стороны, например, если вы извлекаете неслучайную выборку, которая должна быть репрезентативной в отношении:

- количества мужчин и женщин;
- семейного положения;
- дохода, — то результирующая выборка может быть совершенно отличной от генеральной совокупности по важным для исследования характеристикам. Это может привести к неудачным бизнес-решениям, так не была использована случайная выборка.

Одним из способов извлечения случайной выборки является применение таблицы случайных чисел для получения номера каждого отобранного объекта генеральной совокупности. Сам объект затем находят в основе выборки. **Таблица случайных чисел** представляет собой организованную в виде таблицы последовательность чисел, в которой каждая из цифр от 0 до 9 встречается независимо друг от друга с вероятностью $^1/_{10}$. Существует множество методов составления таблицы случайных чисел, например с помощью датчика случайных чисел. Его содержат все современные статистические пакеты, а также Excel (Вставка функций — Математические — Случайное число). Для проведения отбора могут быть использованы цифры любого столбца данной таблицы, при этом необходимо учитывать объем генеральной совокупности. Ниже приведем схему извлечения случайной выборки без возврата размером n.

1. Составьте основу выборки таким образом, чтобы все элементы генеральной совокупности были пронумерованы числами от 1 до N.

2. Выберите точку начала считывания случайных чисел из таблицы. Это необходимо сделать случайным образом.

3. Начав с выбранной точки, последовательно считывайте цифры обычным способом (например, слева направо, с переходом на следующую строку).

4. Объединяем цифры в группы, размер которых равен количеству цифр в числе N. Например, при размере генеральной совокупности $N = 7643$ считываются по четыре случайные цифры за раз, так как запись числа 7643 включает четыре цифры.

5. Необходимо выполнять следующие действия до тех пор, пока не будет получена случайная выборка из n элементов:

а) если получено случайное число между 1 и N и элемент с таким номером еще не извлекался, необходимо включить его в выборку;

б) если полученное случайное число равно 0 или больше N, то его не следует учитывать, так как для него в основе выборки нет соответствующего элемента генеральной совокупности;

в) если получено такое случайное число, что элемент с соответственным номером уже был извлечен ранее, то оно также не учитывается, так как строится выборка без возврата.

При повторном отборе, если тот или иной номер случайно встретится еще один или более раз, соответствующая этому номеру единица в каждом случае повторно включается в выборочную совокупность.

ПРИМЕР.

Необходимо построить случайную выборку контрактов объемом 5 из генеральной совокупности, включающей 362 контракта без повтора. Начнем отсчет со строки 13 столбца 5 таблицы случайных чисел (Приложение 1). Поскольку число $N = 362$ состоит из трех цифр, объединим последовательность случайных чисел в группы, состоящие из трех цифр, следующим образом: 710 309 227 555 497 971 234 091 957 479 767 313 975... Отбрасываем некоторые числа из этой последовательности, которые больше 362 ($N = 362$): 710 555 497 971 957 479 767 975 578 537 519 371 820. Первым, попавшим в выборку, будет число 309. Если число встречается во второй раз, не следует включать его в выборку (так как строится выборка без возврата). Процесс продолжаем до тех пор, пока не будет отобрано 5 элементов. Поэтапная реализация метода извлечения случайной выборки показана на рис. 5.2.1.

После проведения отбора с использованием какого-либо алгоритма, реализующего принцип случайности, или на основе таблицы случайных чисел необходимо определить границы генеральных характеристик. Для этого рассчитываются средняя и предельная ошибки выборки.

Средняя ошибка собственно случайной выборки находится по формулам, представленным в табл. 5.2.1.

Таблица 5.2.1

Формулы для расчета средней ошибки собственно случайной выборки (μ)

	Схема отбора	
	повторный	бесповторный
При оценке среднего	$\mu = \sqrt{\frac{\sigma^2}{n}}$	$\mu = \sqrt{\frac{\sigma^2}{n}\left(1 - \frac{n}{N}\right)}$
При оценке доли	$\mu = \sqrt{\frac{w(1-w)}{n}}$	$\mu = \sqrt{\frac{w(1-w)}{n}\left(1 - \frac{n}{N}\right)}$
При оценке суммарного значения признака	$\mu = \sqrt{\frac{N^2\sigma^2}{n}}$	$\mu = \sqrt{\frac{N^2\sigma^2}{n}\left(1 - \frac{n}{N}\right)}$

Начинаем с использования таблицы случайных чисел:
71 030 92 275 55 497 97 123 40 919 57 479 76 731 39 755 78 537 51 937 11 680 78 820

↓

Объединяем цифры в группы по три (поскольку число 362 состоит из трех цифр):
710 309 227 555 497 971 234 091 957 479 767 313 975 578 537 519 371 168 078 820

↓

Исключаем числа, которые больше 362 или меньше 1:
309 227 234 091 313 168 078

↓

Исключаем числа, которые уже встречались ранее
(в данной выборке такие отсутствуют)

↓

Выбираем первые пять чисел:
309 227 234 091 313

Рис. 5.2.1. Извлечение без возврата случайной выборки размером $n = 5$ элементов из генеральной совокупности размером $N = 362$ элемента

Использование данных формул предполагает, что известны генеральная дисперсия и доля. Но на практике их как раз и невозможно оценить, поэтому используется выборочная дисперсия и выборочная доля.

Формулы средней ошибки выборки при оценивании доли получаются путем подстановки вместо выборочной дисперсии формулы для расчета дисперсии альтернативного признака: $w(1 - w)$.

Из формулы для расчета средней ошибки выборки следует, что ошибка выборки практически не зависит от доли отбора, так как поправка на конечную совокупность $\left(1-\frac{n}{N}\right)$ проявляется только при больших долях отбора, главным образом при небольшом числе единиц генеральной совокупности, и целиком определяется объемом выборки n. С увеличением абсолютной численности выборки ошибка уменьшается пропорционально корню квадратному из n.

Сравнение точности выборочных оценок производят с помощью коэффициента вариации оценки среднего значения V_μ, который определяется по формуле:

$$V_\mu = \frac{\mu_{\tilde{x}}}{\tilde{x}} 100\%. \tag{5.2.1}$$

Коэффициент вариации оценки среднего значения характеризует, на сколько процентов выборочная оценка отклоняется от параметров генеральной совокупности. Если она не превышает заранее установленного предельного значения (Δ), то данный способ отбора можно считать оптимальным.

ПРИМЕР.

Для изучения среднего баланса денежных средств на счете 1350 банковских вкладов было проведено выборочное обследование 810 счетов методом случайной повторной выборки, в результате которого были получены следующие данные (табл. 5.2.2).

Таблица 5.2.2

Результаты выборочного обследования среднего баланса счета совокупности банковских счетов

Баланс счета, долл.	до 250	250–500	500–750	750–1000	1000–1250	1250–1500	Σ
Количество счетов	80	105	134	231	169	91	810

В рассматриваемом примере процент отбора составляет 60%. Определим предельную ошибку выборки и границы для генеральной средней, используя следующие этапы:

1. По результатам выборочного обследования рассчитаем среднее значение и дисперсию выборочной совокупности, для чего необходимо рассчитать середины интервалов группировочного признака (табл. 5.2.3).

Таблица 5.2.3

Расчет среднего размера баланса, приходящегося на один счет, и дисперсии

Баланс счета x_i, долл.	Количество счетов f_i	Середина интервала $\bar{x}_i$	$\bar{x}_i \cdot f_i$	$\bar{x}^2{}_i \cdot f_i$
до 250	80	125	10 000	1 250 000
250–500	105	375	39 375	14 765 625
500–750	134	625	83 750	52 343 750
750–1000	231	875	202 125	176 859 375
1000–1250	169	1125	190 125	213 890 625
1250–1500	91	1375	125 125	172 046 875
Итого	810	–	650 500	631 156 250

Выборочная средняя составит:

$$\tilde{x} = \frac{550\ 500}{810} = 803{,}086.$$

Выборочная дисперсия изучаемого признака:

$$\tilde{\sigma}^2 = \left(\bar{x}_i^2\right) - \left(\tilde{x}\right)^2 = \frac{\sum (x_i)^2 f_i}{\sum f_i} - \left(\tilde{x}\right)^2 = \frac{631\,156\,250}{810} - (803{,}086)^2 = 134\ 258{,}124.$$

Тогда выборочное среднеквадратическое отклонение составит:

$$\sigma = \sqrt{\tilde{\sigma}^2} = 366{,}412.$$

2. Определим среднюю ошибку повторной случайной выборки:

$$\mu_{\tilde{x}} = \frac{\sigma}{\sqrt{n}} = \frac{366{,}412}{\sqrt{810}} = 12{,}874.$$

3. Зададим вероятность, на уровне которой будем говорить о величине предельной ошибки выборки. Чаще всего она принимается равной 0,954.

4. Определим предельную ошибку выборки с заданной вероятностью. Для этого по таблице значений вероятности функции нормального распределения находим величину коэффициента доверия t, соответствующего вероятности 0,954. При $F(t) = 0{,}954$, $t = 2$.

5. Предельная ошибка выборки с вероятностью 0,954 равна:

$$\Delta_{\tilde{x}} = t\mu_{\tilde{x}} = 2 \cdot 12{,}874 = 25{,}749.$$

6. Найдем доверительные границы для среднего значения баланса счета:

$$\tilde{x} - \Delta_{\tilde{x}} < \bar{x} < \tilde{x} + \Delta_{\tilde{x}}.$$

$$803{,}086 - 25{,}749 < \bar{x} < 803{,}086 + 25{,}749.$$

$$777{,}338 < \bar{x} < 828{,}835.$$

Таким образом, в 954 случаях из 1000 средний объем денежных средств на балансе счета у клиентов банка не будет превышать 829 долл., но не может быть ниже, чем 777 долл.

В рассмотренном примере была использована повторная схема случайной выборки. Определим, изменятся ли результаты обследования, если предположить, что отбор осуществлялся по схеме бесповторного отбора. При расчете средней ошибки собственно случайной бесповторной выборки необходимо учитывать поправку на бесповторность отбора (табл. 5.2.1).

ПРИМЕР.

Предположим, что представленные в табл. 5.2.3 данные являются результатом бесповторного отбора, средняя ошибка выборки составит:

$$\mu_{\tilde{x}} = \sqrt{\frac{\sigma^2}{n}\left(1 - \frac{n}{N}\right)} = \sqrt{\frac{134\,258{,}124}{810}\left(1 - \frac{810}{1350}\right)} = 8{,}143.$$

При вероятности 0,954 величина предельной ошибки выборки будет соответствовать:

$$\Delta_{\tilde{x}} = t\mu_{\tilde{x}} = 2 \cdot 8{,}143 = 16{,}285.$$

Доверительные границы для среднего при бесповторном случайном отборе будут иметь следующий вид:

$$\tilde{x} - \Delta_{\tilde{x}} < \bar{x} < \tilde{x} + \Delta_{\tilde{x}}.$$

$$803{,}086 - 16{,}285 < \bar{x} < 803{,}086 + 16{,}285.$$

$$786{,}801 < \bar{x} < 819{,}371.$$

Сравним точность выборочных оценок с помощью коэффициента вариации оценки среднего значения V_μ:

- для случайной повторной выборки:

$$V_\mu = \frac{\mu_{\tilde{x}}}{\tilde{x}} = \frac{12{,}874}{803{,}086} \cdot 100\% = 1{,}6\%.$$

- для случая бесповторного отбора при использовании аналогичной формулы, коэффициент вариации составит:

$$V_{\mu} = \frac{\mu_{\tilde{x}}}{\tilde{x}} = \frac{8{,}143}{803{,}086} \cdot 100\% = 1{,}01\%.$$

Коэффициент вариации оценки среднего значения иллюстрирует, на сколько процентов выборочная оценка отклоняется от параметра генеральной совокупности. Сравнив результаты бесповторного и повторного отборов, можно сделать вывод о том, что применение бесповторной случайной выборки дает более точные результаты по сравнению с применением случайного повторного отбора при одной и той же доверительной вероятности, и чем больше объем выборки, тем существеннее сужаются границы значений средней при переходе от одной схемы отбора к другой.

Помимо определения границ генеральной средней, могут быть определены границы и генеральной доли, если в качестве значений признака взять 0 (если единица не обладает данным признаком) и 1 (если обладает).

ПРИМЕР.

По данным определим, в каких границах находится генеральная доля банковских вкладов с балансом счета, не превышающим 1000 руб.

1. Рассчитаем выборочную долю. Количество банковских вкладов с балансом счета, не превышающим значение в 1000 руб., составляет 550 ед., следовательно, $m = 550$, $n = 810$:

$$w = \frac{m}{n} = \frac{550}{810} = 0{,}679.$$

2. Рассчитаем выборочную дисперсию доли:

$$\sigma_w^2 = w(1 - w) = 0{,}679(1 - 0{,}679) = 0{,}218.$$

3. Средняя ошибка выборки при использовании повторной схемы отбора составит:

$$\mu_{\tilde{w}} = \sqrt{\frac{\sigma^2}{n}} = \sqrt{\frac{w(1-w)}{n}} = \sqrt{\frac{0{,}218}{810}} = 0{,}016.$$

Предположим, что была использована бесповторная схема отбора. Тогда средняя ошибка выборки с учетом поправки на конечность совокупности составит:

$$\mu_{\tilde{w}} = \sqrt{\frac{w(1-w)}{n} \cdot \left(1 - \frac{n}{N}\right)} = \sqrt{\frac{0{,}218}{810} \cdot \left(1 - \frac{810}{1350}\right)} = 0{,}01.$$

4. Зададим доверительную вероятность и определим предельную ошибку выборки. При значении вероятности $p = 0{,}954$ по таблице нормального распределения получаем значение коэффициента доверия $t = 2$.

$$\Delta_{\tilde{w}} = t\mu_{\tilde{w}} = 2 \cdot 0{,}01 = 0{,}02.$$

5. Определим границы для генеральной доли с вероятностью 95%.

$$w - \Delta_{\tilde{w}} < p < w + \Delta_{\tilde{w}}.$$

$$0{,}679 - 0{,}02 < p < 0{,}679 + 0{,}02.$$

$$0{,}659 < p < 0{,}699.$$

Следовательно, с вероятностью 0,95 можно утверждать, что в генеральной совокупности доля банковских вкладов с балансом счета, не превышающим 1000 руб., не меньше 66% и не больше 70%.

На практике исследователь сталкивается с задачей определения границ генеральной средней и генеральной доли по результатам уже проведенного выборочного обследования, когда известен объем выборки или процент отбора единиц. На этапе же проектирования выборочного наблюдения именно объем выборочной совокупности и требует определения.

Чем больше объем выборки, тем меньше значения средней и предельной ошибок выборочного наблюдения и, следовательно, тем уже границы генеральной средней и генеральной доли. В то же время необходимо учитывать, что большой объем выборки приводит к удорожанию обследования, увеличению сроков сбора и обработки материалов, требует привлечения дополнительного персонала и соответствующего материально-технического обеспечения. Затраты всех ресурсов на 20–30%-ное выборочное наблюдение уже сопоставимы с расходами на сплошное обследование. При этом не следует забывать, что статистические характеристики, полученные по выборочной совокупности, всегда имеют вероятностную основу и всегда будут уступать результатам сплошного наблюдения по точности и надежности. Поэтому при подготовке выборочного наблюдения необходимо определить тот минимально необходимый объем выборки, который обеспечит требуемую точность полученных статистических характеристик при заданном уровне вероятности.

Предельную ошибку повторной собственно случайной выборки можно записать через среднюю ошибку:

$$\Delta_{\tilde{x}} = t \cdot \frac{\sigma}{\sqrt{n}}. \qquad (5.2.2)$$

Отсюда можно вывести формулу для определения необходимого объема собственно случайной повторной выборки:

$$n = \frac{t^2 \sigma^2}{\Delta_{\tilde{x}}^2}. \qquad (5.2.3)$$

Полученный на основе использования данной формулы результат всегда округляется в большую сторону. Например, если мы получили необходимый объем выборки, который составляет 598,1 единицы, то обследовав 598 единицы, мы не достигнем требуемой точности. Поэтому для достижения желаемого результата обследованием должны быть охвачены 599 единиц. С другой стороны, рассчитанное значение необходимого объема выборки свободно может быть увеличено в большую сторону на несколько единиц. Если мы располагаем необходимыми ресурсами, по причинам организационного порядка (компактность расположения единиц, фиксированная нагрузка на каждого регистратора и т. п.) мы вполне можем охватить больший объем, то включение в выборочную совокупность 1000 или, например, 1500 единиц только уменьшит значения полученных случайной и предельной ошибок.

В соответствии с формулой (5.2.3) необходимый объем выборки будет тем больше, чем выше заданный уровень вероятности и чем сильнее варьирует наблюдаемый признак. В то же время повышение допустимой предельной ошибки выборки приводит к снижению необходимого объема.

При определении необходимого объема выборки для определения границ генеральной доли задача оценки вариации решается значительно проще. Если дисперсия изучаемого альтернативного признака неизвестна, то можно использовать ее максимальное возможное значение:

$$\sigma^2_{w\,\max} = w(1-w) = 0{,}5(1-0{,}5) = 0{,}25. \qquad (5.2.4)$$

ПРИМЕР.

Вам необходимо провести социологическое исследование с вероятностью 95,4%, посвященное опознанию потребителями вашей торговой марки. Предельно допустимая ошибка данного исследования не должна превышать 5%. Какое количество респондентов вам необходимо опросить в порядке собственно случайной повторной выборки для решения данной проблемы?

$$n = \frac{t^2\sigma^2}{\Delta_{\tilde{x}}^2} = \frac{2^2 \cdot 0{,}25}{0{,}05^2} = 400.$$

Следовательно, обследованию необходимо подвергнуть не менее 400 респондентов на предмет узнаваемости торговой марки вашей компании.

Необходимый объем собственно случайной бесповторной выборки определяется по следующей формуле:

$$n = \frac{t^2\sigma^2 N}{t^2\sigma^2 + \Delta_{\tilde{x}}^2 N}. \qquad (5.2.5)$$

При этом важно помнить, что в процессе проведения вычислений объем генеральной совокупности должен быть выражен только в единицах, а не в тысячах или миллионах единиц. Например, подставив в данную формулу общее количество заказов, выраженное в тысячах единиц, мы не получим правильное значение необходимой численности выборки, также выраженное в тысячах единиц, как это иногда бывает в других расчетах. Результат вычислений будет неверен.

ПРИМЕР.

В табл. 5.2.2 представлены результаты 10%-го бесповторного выборочного обследования. Таким образом, объем генеральной совокупности составляет 8100 счетов. Используя значения дисперсии, полученные на основе расчетов, произведенных в табл. 5.2.3, определим необходимую численность собственно случайной бесповторной выборки для определения баланса счета с предельной ошибкой, не превышающей 16 долл. при уровне вероятности 0,954:

$$n = \frac{t^2\sigma^2 N}{t^2\sigma^2 + \Delta_{\tilde{x}}^2 N} = \frac{2^2 \cdot 134\,258{,}124 \cdot 8100}{2^2 \cdot 134\,258{,}124 + 256 \cdot 8100} = 1662{,}249 \approx 1667.$$

На основании полученного показателя делаем вывод о том, что для достижения заданной точности необходимо обследовать не менее 1667 счетов.

Востребованность собственно случайной выборки объясняется как удобством ее практического использования, так и фактом того, что она является основой многих других способов выборочного наблюдения.

5.2.2. Систематическая (механическая) выборка

Случайная выборка является не единственным способом построения выборки из генеральной совокупности. Существует много других методов, каждый из которых имеет свои преимущества и недостатки. Некоторые, как например стратифицированная случайная выборка, тщательно используют принципы случайного отбора. Другие, например систематическая выборка, основаны на совершенно других подходах и являются достаточно слабой базой для статистического анализа.

При выборе метода важно сопоставить точность получаемых сведений со стоимостью получения таких данных. Для внутреннего исследования, при котором критерий достоверности может быть не очень высоким, отсутствует необходимость в тщательном построении случайной выборки. Однако для внешнего исследования имеет смысл обратить внимание на детали и использовать тщательно спланированные методы построения случайной выборки.

Для получения **систематической выборки** необходимо выбрать в основе выборки одну случайную начальную точку (за счет этого вносится в выборку элемент случайности), и затем производить отбор элементов в основе выборки с некоторым постоянным шагом. При таком способе отбора генеральная совокупность делится на столько групп, сколько единиц наблюдения должно войти в выборку, и из каждой группы отбирается одна единица. Таким образом, все единицы генеральной совокупности нумеруются числами от 1 до N, после чего отбираются каждые N/n-е объекты для выборки, находящиеся на равном расстоянии друг от друга.

Величина $\frac{N}{n}$ называется шагом, или интервалом отбора. Если вы хотите построить систематическую выборку объемом n из генеральной совокупности объемом N, то интервал между выбираемыми элементами будет равен $\frac{N}{n}$. Например, если для 1000 единиц требуется создать 5%-ную выборочную совокупность соответственно объемом 50 единиц, то в нее попадет каждый 20-й элемент, отобранный механически через определенный интервал генеральной совокупности $\left(\frac{N}{n} = \frac{1000}{50} = 20\right)$.

При таком определении интервала возникают технические трудности в случае, если N/n не является целым числом. При выборе в качестве начальной точки отбора числа в промежутке от 1 до $\frac{N}{n}$ выборочное среднее будет являться обоснованной оценкой среднего генеральной совокупности, так как оно будет несмещенным, т. е. его величина не будет систематически завышенной или заниженной, что, в свою очередь, является положительным моментом.

Существуют два способа формирования основы механической выборки: по неранжированным (по отношению к изучаемым признакам) данным и по ранжированной генеральной совокупности.

В первом случае результаты механического отбора будут являться реализацией случайного бесповторного отбора, так как единицы наблюдения располагаются в случайном порядке. Усилить данную случайность возможно выбором начальной точки отсчета случайным образом из интервала, соответствующего первому шагу отбора.

Во втором случае единицы наблюдения определенным образом упорядочиваются по величине изучаемого признака, и отбор осуществляется в соответствии с его шагом $\frac{N}{n}$, начиная с единицы, являющейся серединой первого интервала.

Механический отбор прост в реализации и широко применялся во времена массового отсутствия средств вычислительной техники, так как вручную при большом объеме генеральной совокупности его провести значительно легче, чем случайный.

Однако метод систематического отбора имеет серьезные недостатки из-за невозможности оценить точность получаемой выборки. Недопустимо использование систематической выборки в случае повторяемости некоторого ее фрагмента, который по размеру соответствует и интервалу отбора. В этом случае выборка будет признана нерепрезентативной. Например, если во время упаковки на конвейере каждому 20-му изделию уделяют особое внимание и если по воле случая вы отбираете в свою систематическую выборку именно 20 изделий, то результаты систематической выборки будут полностью бесполезны в отношении репрезентативности качества других обычных изделий.

Для определения средней ошибки механической выборки, а также необходимой ее численности используются соответствующие формулы, применяемые при собственно случайном бесповторном отборе (5.2.2

и 5.2.5). При этом после определения необходимой численности выборки и сопоставления ее с объемом генеральной совокупности, как правило, приходится производить соответствующее округление для получения целочисленного интервала отбора.

ПРИМЕР.

Вы контролируете работу 284 супермаркетов вашего района. Определим, сколько из них необходимо отобрать в порядке механического отбора для выявления нарушений политики вашей компании с ошибкой ±2 нарушения и с вероятностью 0,95. По результатам ранее проведенного обследования известно, что среднее квадратическое отклонение численности нарушений составляет 3 единицы. Расчет проведем, используя формулу (5.2.5):

$$n = \frac{t^2\sigma^2 N}{t^2\sigma^2 + \Delta_{\tilde{x}}^2 N} = \frac{1{,}96^2 \cdot 3^2 \cdot 284}{1{,}96^2 \cdot 3^3 + 2^2 \cdot 284} = 8{,}388 \approx 9.$$

С учетом полученного необходимого объема выборки (9 супермаркетов) определим интервал отбора: $\frac{284}{9} = 31{,}56$. Определенный таким способом интервал всегда округляется в меньшую сторону, так как при округлении в большую сторону произведенная выборка не достигнет рассчитанного по формуле необходимого объема. Следовательно, в нашем примере из общего списка супермаркетов необходимо отобрать для обследования каждый 9-й магазин. При этом процент отбора составит $\frac{100\%}{9\%} = 11{,}1\%$.

5.2.3. Стратифицированная (типическая) выборка

Иногда генеральная совокупность содержит ясные, известные, легко идентифицируемые группы. Такие группы также называются слоями (стратами), в связи с чем типический отбор называют также стратифицированным или расслоенным. При обследованиях населения в качестве типических групп могут быть выбраны области, районы, социальные, возрастные или образовательные группы, при обследовании предприятий — отрасли или подотрасли, формы собственности и т. п.

Стратифицированную случайную выборку получают путем осуществления случайной выборки отдельно в каждой страте генеральной совокупности. Если генеральная совокупность однородна (гомогенна) внутри каждой страты, но отдельные страты заметно отличаются друг от друга, стратификация может увеличить точность статистического анализа. Стратификация также облегчает управление исследованием, поскольку появляется возможность поручить отбор определенным филиалам центрального офиса.

Размеры выборки для каждой из страт могут быть разными. Не обязательно отбирать одинаковое количество элементов из каждой страты или планировать размер выборки для страты в соответствии с ее процентным содержанием в генеральной совокупности. Это позволяет определять размеры выборок для страт исходя из затрат и ресурсов. Для одних страт процесс отбора может быть сложнее и дороже, чем для других, и для этих страт используются меньшие по размеру выборки. Другие страты могут иметь большую изменчивость, поэтому для них требуется использовать большие по размеру выборки.

Обозначения для размеров генеральной совокупности, размеров выборок, выборочных средних и выборочных стандартных отклонений для каждой из страт при стратифицированной выборке представим в табл. 5.2.4.

Таблица 5.2.4

Условные обозначения для стратифицированной выборки

Страта	Размер совокупности	Размер выборки	Среднее выборки	Стандартное отклонение выборки
1	N_1	n_1	$\bar{x}_1$	μ_1
2	N_2	n_2	$\bar{x}_2$	μ_2
...	...	...		
i	N_i	n_i	$\bar{x}_i$	μ_i

Рассматривать генеральную совокупность в разрезе нескольких крупных групп единиц имеет смысл только в том случае, если средние значения изучаемых признаков по группам существенно различаются. Например, с большой уверенностью можно предположить, что доходы населения крупного города будут в среднем выше доходов населения, проживающего в сельской местности; численность работников промышленного предприятия в среднем будет выше численности работников торгового или сельскохозяйственного предприятия; средний возраст студентов будет значительно меньше среднего возраста занятого населения и тем более пенсионеров. В то же время нет никакого смысла при выделении типических групп ориентироваться на признак, не связанный или очень слабо связанный с изучаемым.

Существуют следующие два вида организации отбора внутри типической группы: пропорциональный объему типических групп и пропорциональный степени колеблемости значений признака у единиц наблюдения в группах. Поскольку в выборочную совокупность в той или иной пропорции обязательно попадают представители всех групп, типизация генеральной совокупности позволяет исключить влияние межгрупповой дисперсии на среднюю ошибку выборки. В то же время в выделенных типических группах обследуются далеко не все единицы, а только включенные в выборку. Следовательно, на величине полученной ошибки будет сказываться различие между единицами внутри этих групп, т. е. внутригрупповая вариация. Поэтому ошибка типической выборки будет определяться величиной не общей дисперсии, а только ее части — средней из внутригрупповых дисперсий.

При проведении отбора, пропорционального объему типических групп, число единиц, подлежащих отбору из каждой группы, определяется следующим образом:

$$n_i = n \cdot \frac{N_i}{N}, \qquad (5.2.7)$$

где n_i — количество извлекаемых единиц для выборки из i-й типической группы; n — общий объем выборки; N_i — количество единиц генеральной совокупности, составивших i-ю типическую группу; N — общее количество единиц генеральной совокупности.

ПРИМЕР.

Общее количество счетов вашей фирмы на конец года составило 17 258 единиц, в том числе 56 крупных, 956 средних и 16 246 мелких счетов. Размер счета был определен на основании балансовой стоимости, представляющей количество денег, которое, как полагают, находится на данном счете, и контрольной стоимости, показывающей количество денег, находящихся на данном счете фактически. Вы решили подвергнуть проверке 25% всех счетов, что составило 4315 единиц, которые будут распределены пропорционально объему типических групп:

- крупные счета:

$$n_1 = 4315 \cdot \frac{56}{17\,258} = 14;$$

- средние счета:

$$n_2 = 4315 \cdot \frac{956}{17258} = 239;$$

- мелкие счета:

$$n_3 = 4315 \cdot \frac{16246}{17258} = 4062.$$

Данный способ отбора достаточно часто используется на практике, при этом извлечение единиц внутри групп происходит на случайной или механической основе, но независимо от других групп.

Формулы для расчета средней ошибки выборки для среднего и доли представлены в табл. 5.2.5.

Таблица 5.2.5

Формулы для расчета средней ошибки выборки (μ) при использовании типического отбора, пропорционального объему типических групп

	Схема отбора	
	повторный	бесповторный
При оценке среднего	$\mu = \sqrt{\frac{\overline{\sigma^2}}{n}}$	$\mu = \sqrt{\frac{\overline{\sigma^2}}{n}\left(1 - \frac{n}{N}\right)}$
При оценке доли	$\mu = \sqrt{\frac{\overline{w(1-w)}}{n}}$	$\mu = \sqrt{\frac{\overline{w(1-w)}}{n}\left(1 - \frac{n}{N}\right)}$
При оценке суммарного значения признака	$\mu = N\sqrt{\frac{\overline{\sigma^2}}{n}}$	$\mu = N\sqrt{\frac{\overline{\sigma^2}}{n}(1 - \frac{n}{N})}$
	где $w(1 - w)$ — выборочная дисперсия доли; $\overline{\sigma^2}$ — среднегрупповая дисперсия типических групп	

ПРИМЕР.

В одном из московских вузов проведено выборочное обследование студентов с целью определения показателя средней посещаемости электронной библиотеки одним студентом за семестр. Для этих целей была использована 15%-ная бесповторная типическая выборка, типические группы которой соответствуют номеру курса. При отборе пропорционального объему типических групп получены следующие данные (табл. 5.2.6).

Таблица 5.2.6

Результаты обследования студентов одного из вузов Москвы

Номер курса	Всего студентов N_i, чел.	Обследовано в результате выборочного наблюдения n_i, чел.	Среднее число посещений электронной библиотеки одним студентом за семестр $\tilde{x}_i$	Внутригрупповая выборочная дисперсия $\tilde{\sigma}_i^2$
1	650	98	11	9
2	610	92	8	15
3	580	87	5	18
4	360	54	6	24
5	350	52	10	12
Итого	2550	383	–	–

Общий объем выборочной совокупности составит:

$$n = \frac{2550}{100} \cdot 15 \approx 383 \text{ чел.}$$

Количество студентов, которое необходимо обследовать на каждом курсе, определим согласно формуле (5.2.6):

$$n_1 = n \cdot \frac{N_1}{N} = 383 \cdot \frac{650}{2550} = 98 \text{ чел.};$$

$$n_2 = 92 \text{ чел.}; \; n_3 = 87 \text{ чел.}; \; n_4 = 54 \text{ чел.}; \; n_5 = 52 \text{ чел.}$$

1. Выборочная средняя, основываясь на значениях средних типических групп, составит:

$$\tilde{x}_{\text{общ}} = \frac{\sum \tilde{x}_i n_i}{\sum n_i} = \frac{11 \cdot 98 + 8 \cdot 92 + 5 \cdot 87 + 6 \cdot 54 + 10 \cdot 52}{383} = 8{,}08.$$

2. Средняя из внутригрупповых дисперсий равна:

$$\overline{\sigma_i^2} = \frac{\sum \sigma_i^2 n_i}{\sum n_i} = \frac{9 \cdot 98 + 15 \cdot 92 + 18 \cdot 87 + 24 \cdot 54 + 12 \cdot 52}{383} = 15{,}008.$$

3. Средняя ошибка выборки:

$$\mu_{\tilde{x}} = \sqrt{\frac{\overline{\sigma^2}}{n}\left(1 - \frac{n}{N}\right)} = \sqrt{\frac{15{,}008}{383} \cdot \left(1 - \frac{383}{2550}\right)} = 0{,}18.$$

С вероятностью 0,954 находим предельную ошибку выборки:

$$\Delta_{\tilde{x}} = t \cdot \mu_{\tilde{x}} = 2 \cdot 0{,}18 = 0{,}37.$$

4. Доверительные границы для среднего находятся в пределах:

$$\tilde{x} - \Delta_{\tilde{x}} < \bar{x} < \tilde{x} + \Delta_{\tilde{x}};\ 7{,}71 < \bar{x} < 8{,}45.$$

Таким образом, с вероятностью 95,4% можно утверждать, что каждый студент за семестр посещает электронную библиотеку вуза в среднем от 8 до 9 раз.

Если вариация признака в типических группах существенно различается, то возникает необходимость прибегнуть к переменной доли отбора: чем больше колеблемость значений признаков внутри типической группы, тем большую долю отбора следует использовать для наблюдения. Таким образом, доля отбора становится прямо пропорциональна среднему квадратическому отклонению признака в этой группе σ_i. Следовательно, если внутри какой-либо типической группы наблюдаемый признак варьирует слабо, то для определения границ генеральных характеристик из данной группы достаточно обследовать относительно небольшое число единиц; при сильной же вариации признака объем выборки должен быть соответственно увеличен.

Такой отбор позволяет минимизировать величину ошибки выборки, но его использование затруднено в связи с тем, что на практике никогда не известны величины генеральных групповых дисперсий (σ_i^2).

Приблизительные величины групповых среднеквадратических отклонений рекомендуется определять до проведения основной выборки путем небольших пробных выборочных обследований. Тогда расчет количества извлекаемых единиц наблюдения из каждой группы проводится по формуле (5.2.7) при оценивании генерального среднего значения и (5.2.8) при оценивании генеральной доли:

$$n_i = \frac{n \cdot N_i \cdot \sigma_i}{\sum N_i \sigma_i}; \tag{5.2.7}$$

$$n_i = \frac{n \cdot N_i \cdot \sqrt{w_i(1-w_i)}}{\sum N_i \cdot \sqrt{w_i(1-w_i)}}. \tag{5.2.8}$$

Главное преимущество этого способа отбора заключается в том, что использование переменной доли отбора, прямо пропорциональной вариации признака внутри типических групп, позволяет уменьшить общий объем выборки при сохранении заданной точности.

Средняя и предельная ошибки типической выборки, пропорциональной степени колеблемости значений признака у единиц наблюдения, вычисляются по формулам, представленным в табл. 5.2.7.

Таблица 5.2.7

Формулы для расчета средней ошибки выборки (μ) при использовании типического отбора, пропорционального степени колеблемости значений признака внутри типических групп

	Схема отбора	
	повторный	бесповторный
При оценке среднего	$\mu = \frac{1}{N}\sqrt{\sum \frac{\sigma_i^2 N_i^2}{n_i}}$	$\mu = \frac{1}{N}\sqrt{\sum \left[\frac{\sigma_i^2 N_i^2}{n_i}\left(1 - \frac{n_i}{N_i}\right)\right]}$
При оценке доли	$\mu = \frac{1}{N}\sqrt{\sum \frac{w_i(1-w_i)N_i^2}{n_i}}$	$\mu = \frac{1}{N}\sqrt{\sum \left[\frac{w_i(1-w_i)N_i^2}{n_i}\left(1 - \frac{n_i}{N_i}\right)\right]}$
При оценке суммарного значения признака	$\mu = \sqrt{\sum \frac{\sigma_i^2 N_i^2}{n_i}}$	$\mu = \sqrt{\sum \left[\frac{\sigma_i^2 N_i^2}{n_i}\left(1 - \frac{n_i}{N_i}\right)\right]}$
	где N_i – число единиц i-й типической группы в генеральной совокупности; n_i – число единиц i-й типической группы в выборочной совокупности; $w_i(1 - w_i)$ – выборочная дисперсия доли в i-й типической группе; σ_i^2 – дисперсия i-й типической группы	

ПРИМЕР.

Для разработки маркетинговой стратегии продвижения высокотехнологичной аудио- и видеопродукции вам требуется соответствующая информация о потенциальных покупателях. В зависимости от осведомленности о данной технологии покупателей можно достаточно естественным образом разделить на две группы. Группа осведомленных покупателей желает знать технические особенности продукции; группе неподготовленных покупателей необходима лишь базовая информация общего характера. Чтобы определить, сколько денег в этом году планирует потратить на вашу продукцию типичный потенциальный покупатель, вы решили использовать в своем исследовании стратифицированную случайную выборку, так как ожидается, что группа осведомленных покупателей планирует более крупные расходы. Стратификация позволит уменьшить общую вариацию возможных объемов затрат на вашу продукцию.

Основа вашей выборки – это список имен и адресов 14 000 потенциальных покупателей, полученный из маркетинговой фирмы. Покупатели в списке уже классифицированы: 2532 из них являются осведомленными и 11 468 – неподготовленными. Вы решили подвергнуть обследованию 300 покупателей для более детального опроса, сделав акцент на сегменте подготовленных покупателей, поскольку их ожидаемая покупательская способность выше. Были получены результаты, приведенные в табл. 5.2.8.

Таблица 5.2.8

Результаты типического отбора потенциальных покупателей, пропорционального степени колеблемости значений признака внутри типических групп

Страта	Объем совокупности	Объем выборки	Среднее выборки	Стандартное отклонение выборки
Неподготовленные	$N_1 = 11\,468$	$n_1 = 136$	$\tilde{x}_1 = 7175$ руб.	$\sigma_1 = 2075$ руб.
Осведомленные	$N_2 = 2532$	$n_2 = 164$	$\tilde{x}_2 = 31\,325$ руб.	$\sigma_2 = 11\,350$ руб.

Количество единиц совокупности, извлекаемой из каждой страты, рассчитаем по формуле (5.2.8):

$$n_1 = 300 \cdot \frac{11\,468 \cdot 83}{11\,468 \cdot 83 + 2532 \cdot 454} = 136;$$

$$n_2 = 300 \cdot \frac{2532 \cdot 454}{11\,468 \cdot 83 + 2532 \cdot 454} = 164.$$

Полученные для двух слоев оценки достаточно интересны, так как подтвердили ваши предположения о том, что осведомленные потенциальные покупатели планируют потратить больше денег на приобретение вашей продукции.

Определим среднее значение расходов одного потенциального покупателя для всей генеральной совокупности по формуле средней взвешенной:

$$\bar{x} = \frac{\sum N_i \tilde{x}_i}{\sum N_i} = \frac{11\,468 \cdot 7175 + 2532 \cdot 31\,325}{14\,000} = 11\,543 \text{ руб.}$$

Значение общего среднего (11 543 руб.) намного ближе к величине расходов на покупку у неподготовленных покупателей (7175 руб.), чем к величине расходов у осведомленных – 31 325 руб. Это вызвано тем, что страта неподготовленных покупателей составляет значительно большую часть генеральной совокупности, чем страта осведомленных. Если даже увеличить объем выборки для осведомленных покупателей в два раза, это всего лишь улучшит информацию об этих покупателях, но не усилит влияние этой группы генеральной совокупности на окончательный результат.

Средняя ошибка выборки составит:

$$\mu_{\tilde{x}} = \frac{1}{N}\sqrt{\sum\left[\frac{\sigma_i^2 N_i^2}{n_i}\left(1 - \frac{n_i}{N_i}\right)\right]} =$$

$$= \frac{1}{14\,000}\sqrt{\frac{11\,468^2 \cdot 2075^2}{136} \cdot \left(1 - \frac{136}{11\,468}\right) + \frac{2532^2 \cdot 11\,350^2}{164} \cdot \left(1 - \frac{164}{2532}\right)} \approx 212 \text{ руб.}$$

С вероятностью 0,954 находим предельную ошибку выборки:

$$\Delta_{\tilde{x}} = t \cdot \mu_{\tilde{x}} = 2 \cdot 212 = 424.$$

Рассмотрим неоспоримые преимущества стратифицированной выборки по сравнению с обычной случайной выборкой размером 300 единиц из всей генеральной совокупности. Вместо объединения больших (для подготовленных покупателей) и небольших (для неосведомленных) расходов в одну выборку с большой вариацией мы разделили эту вариацию в соответствии с ее источниками, которые нам известны. Вследствие этого мы получили более точный результат. Без стратификации стандартная ошибка была бы в три раза выше (подробности здесь не приводятся), чем полученное значение 212 руб. Следует также помнить, что трехкратное сокращение стандартной ошибки достигается только при девятикратном увеличении размера выборки ($9 = 3^2$). Используя стратификацию, мы с помощью выборки обьемом 300 единиц достигли результатов, сравнимых с простой случайной выборкой размером приблизительно 900 единиц. Таким образом, стратифицированная выборка может являться отличным способом сокращения расходов на проведение исследования.

При решении различных задач стратификация может быть в большей или меньшей степени полезной, чем в рассмотренном нами примере. Стратификация более выгодна тогда, когда в пределах каждой страты элементы подобны, а сами слои отличаются один от другого. Иными словами, слои выделяют в генеральную совокупность содержательно значимые и важные части.

КОНТРОЛЬНЫЕ ВОПРОСЫ

1. Что вы понимаете под стандартной ошибкой выборки? Каким образом стандартная ошибка характеризует качество оценок, получаемых в результате оценивания?
2. Чему равна средняя ошибка выборки при использовании собственно случайной выборки и оценивании среднего?
3. В чем отличие механического отбора от случайного?
4. Перечислите основные проблемы, возникающие при использовании систематической выборки.
5. Что такое стратифицированная случайная выборка?
6. Перечислите преимущества технического отбора перед случайным. Каким образом в этом случае определяется средняя ошибка выборки?
7. Каким образом осуществляется сбор данных на основе серийного отбора? В чем состоят основные особенности гнездовой выборки?
8. Перечислите основные проблемы, возникающие при использовании систематической выборки.

ПРИЛОЖЕНИЯ

Приложение 1

ТАБЛИЦА СЛУЧАЙНЫХ ЧИСЕЛ

5489	5583	3156	0835	1988	3912	0938	7460	0869	4420
3522	0935	7877	5665	7020	9555	7379	7124	7878	5544
7555	7579	2550	2487	9477	0864	2349	1012	8250	2633
5759	3554	5080	9074	7001	6249	3224	6868	9102	2672
6303	6895	3371	3196	7231	2918	7380	0438	7547	2644
7351	5634	5323	2623	7803	8374	2191	0464	0696	9529
7068	7803	8832	5119	6350	0120	5026	3686	5657	0304
3613	1428	1796	8447	0503	5654	3254	7336	9536	1944
5143	4534	2105	0368	7890	2473	4240	8652	9435	1422
9815	5144	7649	8638	6137	8070	5345	4865	2456	5708
5780	1277	6316	1013	2867	9938	3930	3203	5696	1769
1187	0951	5991	5245	5700	5564	7352	0891	6249	6568
4184	2179	4554	9083	2254	2435	2965	5154	1209	7069
2916	2972	9885	0275	0144	8034	8122	3213	7666	0230
5524	1341	9860	6565	6981	9842	0171	2284	2707	3008
0146	5291	2354	5694	0377	5336	6460	9585	3415	2358
4920	2826	5238	5402	7937	1993	4332	2327	6875	5230
7978	1947	6380	3425	7267	7285	1130	7722	0164	8573
7453	0653	3645	7497	5969	8682	4191	2976	0361	9334
1473	6938	4899	5348	1641	3652	0852	5296	4538	4456
8162	8797	8000	4707	1880	9660	8446	1883	9768	0881
5645	4219	0807	3301	4279	4168	4305	9937	3120	5647
2042	1192	1175	8851	6432	4635	5757	6656	1660	5389

Окончание табл.

5470	7702	6958	9080	5925	8519	0127	9233	2452	7341
4504	1730	6005	1704	0345	3275	4738	4862	2556	8333
5880	1257	6163	4439	7276	6353	6912	0731	9033	5294
9083	4260	5277	4998	4298	5204	3965	4028	8936	5148
1762	8713	1189	1090	8989	7273	3213	1935	9321	4820
2023	2589	1740	0424	8924	0005	1969	1636	7237	1227
7965	3855	4765	0703	1678	0841	7543	0308	9732	1289
7690	0480	8098	9629	4819	7219	7241	5128	3853	1921
9292	0426	9573	4903	5916	6576	8368	3270	6641	0033
0867	1651	7016	4220	2533	6345	8227	1904	5138	2537
0505	2127	8255	5276	2233	3956	4118	8199	6380	6340
6295	9795	1112	5761	2575	6837	3336	9322	7403	8345
6323	2615	3410	3365	1117	2417	3176	2434	5240	5455
8672	8536	2966	5773	5412	8114	0930	4697	6919	4569
1422	5507	7596	0670	3013	1351	3886	3268	9469	2584
2653	1472	5113	5735	1469	9545	9331	5303	9914	6394
0438	4376	3328	8649	8327	0110	4549	7955	5275	2890
2851	2157	0047	7085	1129	0460	6821	8323	2572	8962
7962	2753	3077	8718	7418	8004	1425	3706	8822	1494
3837	4098	0220	1217	4732	0150	1637	1097	1040	7372
8542	4126	9274	2251	0607	4301	8730	7690	6235	3477
0139	0765	8039	9484	2577	7859	1976	0623	1418	6685
6687	1943	4307	0579	8171	8224	8641	7034	3595	3875
6242	5582	5872	3197	4919	2792	5991	4058	9769	1918
6859	9606	0522	4993	0345	8958	1289	8825	6941	7685
6590	1932	6043	3623	1973	4112	1795	8465	2110	8045

Приложение 2
НОРМАЛЬНЫЙ ЗАКОН РАСПРЕДЕЛЕНИЯ

Значения функции Лапласа $\Phi(t) = P(|T| < t_{табл})$; $\Phi(-t) = -\Phi(t)$; $\Phi(t) = \frac{2}{\sqrt{2\pi}}\int_0^t e^{-\frac{x^2}{2}}dx$

Целые и десятые доли t	Сотые доли t									
	0	1	2	3	4	5	6	7	8	9
0,0	0,0000	0,0080	0,0160	0,0239	0,0319	0,0399	0,0478	0,0558	0,0638	0,0717
0,1	0,0797	0,0876	0,0955	0,1034	0,1113	0,1192	0,1271	0,1350	0,1428	0,1507
0,2	0,1585	0,1663	0,1741	0,1819	0,1897	0,1974	0,2051	0,2128	0,2205	0,2282
0,3	0,2358	0,2434	0,2510	0,2586	0,2661	0,2737	0,2812	0,2886	0,2960	0,3035
0,4	0,3108	0,3182	0,3255	0,3328	0,3401	0,3473	0,3545	0,3616	0,3688	0,3759
0,5	0,3829	0,3899	0,3969	0,4039	0,4108	0,4177	0,4245	0,4313	0,4381	0,4448
0,6	0,4515	0,4581	0,4647	0,4713	0,4778	0,4843	0,4907	0,4971	0,5035	0,5098
0,7	0,5161	0,5223	0,5285	0,5346	0,5407	0,5467	0,5527	0,5587	0,5646	0,5705
0,8	0,5763	0,5821	0,5878	0,5935	0,5991	0,6047	0,6102	0,6157	0,6211	0,6265
0,9	0,6319	0,6372	0,6424	0,6476	0,6528	0,6579	0,6629	0,6679	0,6729	0,6778
1,0	0,6827	0,6875	0,6923	0,6970	0,7017	0,7063	0,7109	0,7154	0,7199	0,7243

Окончание табл.

Целые и десятые доли t	Сотые доли t									
	0	1	2	3	4	5	6	7	8	9
1,1	0,7287	0,7330	0,7373	0,7415	0,7457	0,7499	0,7540	0,7580	0,7620	0,7660
1,2	0,7699	0,7737	0,7775	0,7813	0,7850	0,7887	0,7923	0,7959	0,7994	0,8029
1,3	0,8064	0,8098	0,8132	0,8165	0,8198	0,8230	0,8262	0,8293	0,8324	0,8355
1,4	0,8385	0,8415	0,8444	0,8473	0,8501	0,8529	0,8557	0,8584	0,8611	0,8638
1,5	0,8664	0,8690	0,8715	0,8740	0,8764	0,8789	0,8812	0,8836	0,8859	0,8882
1,6	0,8904	0,8926	0,8948	0,8969	0,8990	0,9011	0,9031	0,9051	0,9070	0,9090
1,7	0,9109	0,9127	0,9146	0,9164	0,9181	0,9199	0,9216	0,9233	0,9249	0,9265
1,8	0,9281	0,9297	0,9312	0,9327	0,9342	0,9357	0,9371	0,9385	0,9399	0,9412
1,9	0,9426	0,9439	0,9451	0,9464	0,9476	0,9488	0,9500	0,9512	0,9523	0,9534
2,0	0,9545	0,9556	0,9566	0,9576	0,9586	0,9596	0,9606	0,9616	0,9625	0,9634
2,1	0,9643	0,9651	0,9660	0,9668	0,9676	0,9684	0,9692	0,9700	0,9707	0,9715
2,2	0,9722	0,9729	0,9736	0,9743	0,9749	0,9756	0,9762	0,9768	0,9774	0,9780
2,3	0,9786	0,9791	0,9797	0,9802	0,9807	0,9812	0,9817	0,9822	0,9827	0,9832
2,4	0,9836	0,9841	0,9845	0,9849	0,9853	0,9857	0,9861	0,9865	0,9869	0,9872
2,5	0,9876	0,9879	0,9883	0,9886	0,9889	0,9892	0,9895	0,9898	0,9901	0,9904

2,6	0,9907	0,9910	0,9912	0,9915	0,9917	0,9920	0,9922	0,9924	0,9926	0,9928
2,7	0,9931	0,9933	0,9935	0,9937	0,9939	0,9940	0,9942	0,9944	0,9946	0,9947
2,8	0,9949	0,9951	0,9952	0,9953	0,9955	0,9956	0,9958	0,9959	0,9960	0,9961
2,9	0,9963	0,9964	0,9965	0,9966	0,9967	0,9968	0,9969	0,9970	0,9971	0,9972
3,0	0,9973	0,9974	0,9975	0,9976	0,9976	0,9977	0,9978	0,9979	0,9979	0,9980
3,1	0,9981	0,9981	0,9982	0,9983	0,9983	0,9984	0,9984	0,9985	0,9985	0,9986
3,2	0,9986	0,9987	0,9987	0,9988	0,9988	0,9989	0,9989	0,9989	0,9990	0,9990
3,3	0,9990	0,9990	0,9991	0,9991	0,9992	0,9992	0,9992	0,9992	0,9993	0,9993
3,4	0,9993	0,9994	0,9994	0,9994	0,9994	0,9994	0,9995	0,9995	0,9995	0,9995
3,5	0,9995	0,9996	0,9996	0,9996	0,9996	0,9996	0,9996	0,9996	0,9997	0,9997
3,6	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9998	0,9998	0,9998
3,7	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998
3,8	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999
3,9	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999
4,0	0,999936	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999
4,5	0,999994	–	–	–	–	–	–	–	–	–
5,0	0,99999994	–	–	–	–	–	–	–	–	–

Приложение 3

РАСПРЕДЕЛЕНИЕ ПИРСОНА (χ^2-РАСПРЕДЕЛЕНИЕ)

Значения $\chi^2_{табл}$ для вероятностей $P(\chi^2 > \chi^2_{табл})$

ν_2 / ν_1	Вероятность $P(\chi^2)$										
	0,999	0,995	0,99	0,98	0,975	0,95	0,90	0,80	0,75	0,70	0,50
1	0,00000157	0,0000393	0,000157	0,000628	0,000982	0,00393	0,0158	0,0642	0,102	0,148	0,455
2	0,00200	0,0100	0,0201	0,0404	0,0506	0,103	0,211	0,446	0,575	0,713	1,386
3	0,0243	0,0717	0,115	0,185	0,216	0,352	0,584	1,005	1,213	1,424	2,366
4	0,0908	0,207	0,297	0,429	0,484	0,711	1,064	1,649	1,923	2,195	3,357
5	0,210	0,412	0,554	0,752	0,831	1,145	1,610	2,343	2,675	3,000	4,351
6	0,381	0,676	0,872	1,134	1,237	1,635	2,204	3,070	3,455	3,828	5,348
7	0,598	0,989	1,239	1,564	1,690	2,167	2,833	3,822	4,255	4,671	6,346
8	0,857	1,344	1,646	2,032	2,180	2,733	3,490	4,594	5,071	5,527	7,344
9	1,152	1,735	2,088	2,532	2,700	3,325	4,168	5,380	5,899	6,393	8,343
10	1,479	2,156	2,558	3,059	3,247	3,240	4,865	6,179	6,737	7,267	9,342
11	1,834	2,603	3,053	3,609	3,816	4,575	5,578	6,989	7,584	8,148	10,341

12	2,214	3,074	3,571	4,178	4,404	5,226	6,304	7,807	8,438	9,034	11,340
13	2,617	3,565	4,107	4,765	5,009	5,892	7,042	8,634	9,299	9,926	12,340
14	3,041	4,075	4,660	5,368	5,629	6,571	7,790	9,467	10,165	10,821	13,339
15	3,483	4,601	5,229	5,985	6,262	7,261	8,547	10,307	11,036	11,721	14,339
16	3,942	5,142	5,812	6,614	6,908	7,962	9,312	11,152	11,912	12,624	15,338
17	4,416	5,697	6,408	7,255	7,564	8,672	10,085	12,002	12,892	13,531	16,338
18	4,905	6,265	7,015	7,906	8,231	9,390	10,865	12,857	13,675	14,440	17,338
19	5,407	6,844	7,633	8,567	8,907	10,117	11,651	13,716	14,562	15,352	18,338
20	5,921	7,434	8,260	9,237	9,591	10,871	12,443	14,578	15,452	16,266	19,337
21	6,447	8,034	8,897	9,915	10,283	11,591	13,240	15,445	16,344	17,182	20,337
22	6,983	8,643	9,542	10,600	10,982	12,338	14,041	16,314	17,240	18,101	21,337
23	7,529	9,260	10,196	11,293	11,688	13,091	14,848	17,187	18,137	19,021	22,337
24	8,035	9,886	10,856	11,992	12,401	13,848	15,659	18,062	19,037	19,943	23,337
25	8,649	10,520	11,524	12,697	13,120	14,611	16,173	18,940	19,939	20,887	24,337
26	9,222	11,160	12,198	13,409	13,844	15,379	17,292	19,820	20,843	21,792	25,336
27	9,803	11,808	12,879	14,125	14,573	16,151	18,114	20,703	21,749	22,719	26,136
28	10,391	12,461	13,565	14,847	15,308	16,928	18,937	21,588	22,657	23,617	27,336
29	10,986	13,121	14,256	15,574	16,047	17,708	19,768	22,475	23,567	24,577	28,336
30	11,588	13,787	14,953	16,306	16,791	18,493	20,599	23,364	24,478	25,508	29,336

Окончание табл.

ν_2 / ν_1	Вероятность $P(\chi^2)$ 0,30	0,25	0,20	0,10	0,05	0,025	0,02	0,01	0,005	0,001
1	1,074	1,323	1,642	2,706	3,841	5,024	5,412	6,635	7,879	10,827
2	2,408	2,773	3,219	4,605	5,991	7,378	7,824	9,210	10,597	13,815
3	3,665	4,108	4,642	6,251	7,815	9,348	9,837	11,345	12,838	16,268
4	4,878	5,385	5,989	7,779	9,488	11,143	11,668	13,277	14,860	18,465
5	6,064	6,626	7,289	9,236	11,070	12,839	13,388	15,086	16,750	20,517
6	7,231	7,841	8,558	10,645	12,592	14,449	15,033	16,812	18,548	22,457
7	8,383	9,037	9,803	12,017	14,067	16,013	16,622	18,475	20,278	24,322
8	9,524	10,219	11,030	13,362	15,507	17,535	18,168	20,090	21,955	26,125
9	10,656	11,389	12,242	14,684	16,919	19,023	19,679	21,666	23,589	27,877
10	11,781	12,549	13,412	15,987	18,307	20,483	21,161	23,209	25,188	29,588
11	12,899	13,701	14,631	17,275	19,675	21,920	22,618	24,725	26,757	31,264
12	14,011	14,845	15,812	18,549	21,026	23,337	24,054	26,217	28,300	32,909
13	15,119	15,984	16,985	19,812	22,362	24,736	25,472	27,688	29,819	34,528
14	16,222	17,117	18,151	21,064	23,685	26,119	26,873	29,141	31,319	36,123
15	17,322	18,245	19,311	22,307	24,996	27,488	28,259	30,578	32,801	37,697
16	18,418	19,369	20,465	23,542	26,296	28,845	29,633	32,000	34,267	39,252

17	19,511	20,489	21,615	24,769	27,587	30,191	30,995	33,409	35,718	40,790
18	20,601	21,605	22,760	25,989	28,869	31,526	32,346	34,805	37,156	42,312
19	21,689	22,718	23,900	27,204	30,144	32,852	33,687	36,191	38,582	43,820
20	22,775	23,828	25,038	28,412	31,410	34,170	35,020	37,566	39,997	45,315
21	23,858	24,935	26,171	29,615	32,671	35,479	36,343	38,932	41,401	46,797
22	24,939	26,039	27,301	30,813	33,924	36,781	37,659	40,289	42,796	48,268
23	26,018	27,141	28,429	32,007	35,172	38,076	38,968	41,638	44,181	49,728
24	27,096	28,241	29,553	33,196	36,415	39,364	40,270	42,980	45,558	51,170
25	28,172	29,339	30,675	34,382	37,652	40,046	41,566	44,314	46,928	52,620
26	29,246	30,434	31,795	35,563	38,885	41,923	42,856	45,642	48,290	54,052
27	30,319	31,528	32,912	36,741	40,113	43,194	44,140	46,963	49,645	55,476
28	31,391	32,620	34,027	37,916	41,337	44,461	45,419	48,278	50,993	56,893
29	32,461	33,711	35,139	39,087	42,557	45,722	46,693	49,588	52,336	58,302
30	33,530	34,800	36,250	40,256	43,773	46,979	47,962	50,892	53,672	59,703

Приложение 4

РАСПРЕДЕЛЕНИЕ СТЬЮДЕНТА (t-РАСПРЕДЕЛЕНИЕ)

ν_1 \ ν_2	Вероятность $\alpha = St(t) = P(\lvert T\rvert > t_{табл})$												
	0,9	0,8	0,7	0,6	0,5	0,4	0,3	0,2	0,1	0,05	0,02	0,01	0,001
1	0,158	0,325	0,510	0,727	1,000	1,376	1,963	3,078	6,314	12,706	31,821	63,657	636,619
2	0,142	0,289	0,445	0,617	0,816	1,061	1,386	1,886	2,920	4,303	6,965	9,925	31,598
3	0,137	0,277	0,424	0,584	0,765	0,978	1,250	1,638	2,353	3,182	4,541	5,841	12,941
4	0,134	0,271	0,414	0,569	0,741	0,941	1,190	1,533	2,132	2,776	3,747	4,604	8,610
5	0,132	0,267	0,408	0,559	0,727	0,920	1,156	1,476	2,015	2,571	3,365	4,043	6,859
6	0,131	0,265	0,404	0,553	0,718	0,906	1,134	1,440	1,943	2,447	3,143	3,707	5,959
7	0,130	0,263	0,402	0,549	0,711	0,896	1,119	1,415	1,895	2,365	2,998	3,499	5,405
8	0,130	0,262	0,399	0,546	0,706	0,889	1,108	1,397	1,860	2,306	2,896	3,355	5,041
9	0,129	0,261	0,398	0,543	0,703	0,883	1,100	1,383	1,833	2,262	2,821	3,250	4,781
10	0,129	0,260	0,327	0,542	0,700	0,879	1,093	1,372	1,812	2,228	2,764	3,169	4,583
11	0,129	0,260	0,396	0,540	0,697	0,876	1,088	1,363	1,796	2,201	2,718	3,106	4,437
12	0,128	0,259	0,395	0,539	0,695	0,873	1,083	1,356	1,782	2,179	2,681	3,055	4,318
13	0,128	0,259	0,394	0,538	0,694	0,870	1,079	1,350	1,771	2,160	2,650	3,012	4,221
14	0,128	0,258	0,393	0,537	0,692	0,868	1,076	1,345	1,761	2,145	2,624	2,977	4,140

15	0,128	0,258	0,393	0,536	0,691	0,866	1,074	1,341	1,753	2,131	2,602	2,947	4,073
16	0,128	0,258	0,392	0,535	0,690	0,865	1,071	1,337	1,746	2,120	2,583	2,921	4,015
17	0,128	0,257	0,392	0,534	0,689	0,863	1,069	1,333	1,740	2,110	2,567	2,898	3,965
18	0,127	0,257	0,392	0,534	0,688	0,862	1,067	1,330	1,734	2,101	2,552	2,878	3,922
19	0,127	0,257	0,391	0,533	0,688	0,861	1,066	1,328	1,729	2,093	2,539	2,861	3,833
20	0,127	0,257	0,391	0,533	0,687	0,860	1,064	1,325	1,725	2,086	2,528	2,845	3,850
21	0,127	0,257	0,391	0,532	0,686	0,859	1,063	1,323	1,721	2,080	2,518	2,831	3,819
22	0,127	0,256	0,390	0,532	0,686	0,858	1,061	1,321	1,717	2,074	2,508	2,819	3,792
23	0,127	0,256	0,390	0,532	0,685	0,868	1,060	1,319	1,714	2,069	2,500	2,807	3,767
24	0,127	0,256	0,390	0,531	0,685	0,857	1,059	1,318	1,711	2,064	2,402	2,797	3,745
25	0,127	0,256	0,390	0,531	0,684	0,856	1,058	1,316	1,708	2,060	2,485	2,787	3,725
26	0,127	0,256	0,390	0,531	0,684	0,856	1,058	1,315	1,706	2,056	2,479	2,779	3,707
27	0,127	0,256	0,389	0,531	0,684	0,855	1,057	1,314	1,703	2,052	2,473	2,771	3,690
28	0,127	0,256	0,389	0,530	0,683	0,855	1,056	1,313	1,701	2,048	2,467	2,763	3,674
29	0,127	0,256	0,389	0,530	0,683	0,854	1,055	1,311	1,699	2,045	2,462	2,756	3,659
30	0,127	0,256	0,389	0,530	0,683	0,854	1,055	1,310	1,697	2,042	2,457	2,750	3,646
40	0,126	0,255	0,388	0,529	0,681	0,851	1,050	1,303	1,684	2,021	2,423	2,704	3,551
60	0,126	0,254	0,387	0,527	0,679	0,848	1,046	1,296	1,671	2,000	2,390	2,660	3,460
120	0,126	0,254	0,386	0,526	0,677	0,845	1,041	1,289	1,658	1,980	2,358	2,617	3,373
∞	0,126	0,253	0,385	0,524	0,674	0,842	1,036	1,282	1,645	1,960	2,326	2,576	3,291

Приложение 5

РАСПРЕДЕЛЕНИЕ ФИШЕРА–СНЕДЕКОРА (F-РАСПРЕДЕЛЕНИЕ)

Значения $F_{табл}$, удовлетворяющие условию $P(F > F_{табл})$.
Первое значение соответствует вероятности $\alpha = 0{,}05$, второе — вероятности 0,01 и третье — вероятности 0,001, где ν_1 — число степеней свободы числителя, а ν_2 — число степеней свободы знаменателя

ν_2 \ ν_1	1	2	3	4	5	6	8	12	24	∞
1	161,45	199,50	215,71	224,58	224,58	230,16	238,88	243,91	249,05	254,31
	4052,18	4999,50	5403,35	5624,58	5763,65	5858,99	5981,07	6106,32	6234,63	6365,68
	4065284,07	499999,50	540379,20	562499,58	576404,56	585937,11	598144,16	610667,82	623497,82	636604,08
2	18,51	19,00	19,16	19,25	19,30	19,33	19,37	19,41	19,45	19,50
	98,49	99,01	00,17	99,25	99,30	99,33	99,36	99,42	99,46	99,50
	998,46	999,00	999,20	999,20	999,20	999,20	999,40	999,60	999,40	999,40
3	10,13	9,55	9,28	9,12	9,01	8,94	8,84	8,74	8,64	8,53
	34,12	30,81	29,46	28,71	28,24	27,91	27,49	27,05	26,60	26,12
	67,47	148,51	141,10	137,10	134,60	132,90	130,60	128,30	125,90	123,50

4	7,71	6,94	6,59	6,39	6,26	6,16	6,04	5,91	5,77	5,63
	21,20	18,00	16,69	15,98	15,52	15,21	14,80	14,37	13,93	13,46
	74,13	61,24	56,18	53,43	51,71	50,52	49,00	47,41	45,77	44,05
5	6,61	5,79	5,41	5,19	5,05	4,95	4,82	4,68	4,53	4,36
	16,26	13,27	12,06	11,39	10,97	10,67	10,27	9,89	9,47	9,02
	47,04	36,61	33,20	31,09	20,75	28,83	27,64	26,42	25,14	23,78
6	5,99	5,14	4,76	4,53	4,39	4,28	4,15	4,00	3,84	3,67
	13,74	10,92	9,78	9,15	8,75	8,47	8,10	7,72	7,31	6,88
	35,51	26,99	23,70	21,90	20,81	20,03	19,03	17,99	16,89	15,75
7	5,59	4,74	4,35	4,12	3,97	3,87	3,73	3,57	3,41	3,23
	12,25	9,55	8,45	7,85	7,46	7,19	6,84	6,47	6,07	5,65
	29,22	21,69	18,77	17,19	16,21	15,52	14,63	13,71	12,73	11,70
8	5,32	4,46	4,07	3,84	3,69	3,58	3,44	3,28	3,12	2,99
	11,26	8,65	7,59	7,10	6,63	6,37	6,03	5,67	5,28	4,86
	25,42	18,49	15,83	14,39	13,49	12,86	12,04	11,19	10,30	9,35

Продолжение табл.

ν_2 \ ν_1	1	2	3	4	5	6	8	12	24	∞
9	5,12	4,26	3,86	3,63	3,48	3,37	3,23	3,07	2,90	2,71
	10,56	8,02	6,99	6,42	6,06	5,80	5,47	5,11	4,73	4,31
	22,86	16,39	13,90	12,56	11,71	11,13	10,37	9,57	8,72	7,81
10	4,96	4,10	3,71	3,48	3,33	3,22	3,07	2,91	2,74	2,54
	10,04	7,56	6,55	5,99	5,64	5,39	5,06	4,71	4,33	3,91
	21,04	14,91	12,55	11,28	10,48	9,92	9,20	8,45	7,64	6,77
11	4,84	3,98	3,59	3,36	3,20	3,09	2,95	2,79	2,61	2,40
	9,65	7,20	6,22	5,67	5,32	5,07	4,74	4,40	4,02	3,60
	19,69	13,81	11,56	10,35	9,58	9,05	8,35	7,62	6,85	6,00
12	4,75	3,88	3,49	3,26	3,11	3,00	2,85	2,69	2,50	2,30
	9,33	6,93	5,95	5,41	5,06	4,82	4,50	4,16	3,78	3,36
	18,64	12,98	10,81	9,63	8,89	8,38	7,71	7,00	6,25	5,42

13	4,67	3,80	3,41	3,18	3,02	2,92	2,77	2,60	2,42	2,21
	9,07	6,70	5,74	5,20	4,86	4,62	4,30	3,96	3,59	3,16
	17,81	12,31	10,21	9,07	8,35	7,86	7,21	6,52	5,78	4,97
14	4,60	3,74	3,34	3,11	2,96	2,85	2,70	2,53	2,35	2,13
	8,86	6,51	5,56	5,03	4,69	4,46	4,14	3,80	3,43	3,00
	17,14	11,78	9,73	8,62	7,92	7,44	6,80	6,13	5,41	4,60
15	4,45	3,68	3,29	3,06	2,90	2,79	2,64	2,48	2,29	2,07
	8,68	6,36	5,42	4,89	4,56	4,32	4,00	3,67	3,29	2,87
	16,59	11,34	9,34	8,25	7,57	7,09	6,47	5,81	5,10	4,31
16	4,41	3,63	3,24	3,01	2,85	2,74	2,59	2,42	2,24	2,01
	8,53	6,23	5,29	4,77	4,44	4,20	3,89	3,55	3,18	2,75
	16,12	10,97	9,01	7,94	7,27	6,80	6,20	5,55	4,85	4,06
17	4,45	3,59	3,20	2,96	2,81	2,70	2,55	2,38	2,19	1,96
	8,40	6,11	5,18	4,67	4,34	4,10	3,79	3,45	3,08	2,65
	15,72	10,66	8,73	7,68	7,02	6,56	5,96	5,32	4,63	3,85

Продолжение табл.

ν_2 \ ν_1	1	2	3	4	5	6	8	12	24	∞
18	4,41	3,55	3,16	2,93	2,77	2,66	2,51	2,34	2,15	1,92
	8,28	6,01	5,09	4,58	4,25	4,01	3,71	3,37	3,01	2,57
	15,38	10,39	8,49	7,46	6,81	6,35	5,76	5,13	4,45	3,67
19	4,38	3,52	3,13	2,90	2,74	2,63	2,48	2,31	2,11	1,88
	8,18	5,93	5,01	4,50	4,17	3,94	3,63	3,30	2,92	2,49
	15,08	10,16	8,28	7,26	6,61	6,18	5,59	4,97	4,29	3,52
20	4,35	3,49	3,10	2,87	2,71	2,60	2,45	2,28	2,08	1,84
	8,10	5,85	4,94	4,43	4,10	3,87	3,56	3,23	2,86	2,42
	14,82	9,95	8,10	7,10	6,46	6,02	5,44	4,82	4,15	3,38
21	4,32	3,47	3,07	2,84	2,68	2,57	2,42	2,25	2,05	1,82
	8,02	5,78	4,87	4,37	4,04	3,81	3,51	3,17	2,80	2,36
	14,62	9,77	7,94	6,95	6,32	5,88	5,31	4,70	4,03	3,26
22	4,30	3,44	3,05	2,82	2,66	2,55	2,40	2,23	2,03	1,78
	7,94	5,72	4,82	4,31	3,99	3,75	3,45	3,12	2,75	2,30
	14,38	9,61	7,80	6,81	6,19	5,76	5,19	4,58	3,92	3,15

23	4,28	3,42	3,03	2,80	2,64	2,53	2,38	2,20	2,00	1,76
	7,88	5,66	4,76	4,26	3,94	3,71	3,41	3,07	2,70	2,26
	14,19	9,46	7,67	6,70	6,08	5,56	5,09	4,48	3,82	3,05
24	4,26	3,40	3,01	2,78	2,62	2,51	2,36	2,18	1,98	1,73
	7,82	5,61	4,72	4,22	3,90	3,67	3,36	3,03	2,66	2,21
	14,03	9,34	7,55	6,59	5,98	5,55	4,99	4,39	3,74	2,97
25	4,24	3,38	2,99	2,76	2,60	2,49	2,34	2,16	1,96	1,71
	7,77	5,57	4,68	4,18	3,86	3,63	3,32	2,99	2,62	2,17
	13,88	9,22	7,45	6,49	5,89	5,46	4,91	4,31	3,66	2,89
26	4,22	3,37	2,98	2,74	2,59	2,47	2,32	2,15	1,95	1,69
	7,72	5,53	4,64	4,14	3,82	3,59	3,29	2,96	2,58	2,13
	13,74	9,12	7,36	6,41	5,80	5,38	4,83	4,24	3,59	2,82
27	4,21	3,35	2,96	2,73	2,57	2,46	2,30	2,13	1,93	1,67
	7,68	5,49	4,60	4,11	3,78	3,56	3,26	2,93	2,55	2,10
	13,61	9,02	7,27	6,33	5,73	5,31	4,76	4,17	3,52	2,76
28	4,19	3,34	2,95	2,71	2,56	2,44	2,29	2,12	1,91	1,65
	7,64	5,54	4,57	4,07	3,75	3,53	3,23	2,90	2,52	2,06
	13,50	8,93	7,18	6,25	5,66	5,24	4,69	4,11	3,46	2,70

Окончание табл.

ν_2 \ ν_1	1	2	3	4	5	6	8	12	24	∞
29	4,18	3,33	2,93	2,70	2,54	2,43	2,28	2,10	1,90	1,64
	7,60	5,42	4,54	4,04	3,73	3,50	3,20	2,87	2,49	2,03
	13,39	8,85	7,12	6,19	5,59	5,18	4,65	4,05	3,41	2,64
30	4,17	3,32	2,92	2,69	2,53	2,42	2,27	2,09	1,89	1,62
	7,56	5,39	4,51	4,02	3,70	3,47	3,17	2,84	2,47	2,01
	13,29	8,77	7,05	6,12	5,53	5,12	4,58	4,00	3,36	2,59
60	4,00	3,15	2,76	2,52	2,37	2,25	2,10	1,92	1,70	1,39
	7,08	4,98	4,13	3,65	3,34	3,12	2,82	2,50	2,12	1,60
	11,97	7,76	6,17	5,31	4,76	4,37	3,87	3,31	2,76	1,90
∞	3,84	2,99	2,60	2,37	2,21	2,09	1,94	1,75	1,52	1,03
	6,64	4,60	3,78	3,32	3,02	2,80	2,51	2,18	1,79	1,04
	10,83	6,91	5,42	4,62	4,10	3,74	3,27	2,74	2,13	1,05

Приложение 6

ТАБЛИЦА Z-ПРЕОБРАЗОВАНИЯ ФИШЕРА

Значения $Z = \frac{1}{2} \cdot \ln \frac{1+r}{1-r}$

Десятые доли r	Сотые доли r									
	0	1	2	3	4	5	6	7	8	9
0,0	0,0000	0,0101	0,0200	0,0300	0,0400	0,0501	0,0601	0,0701	0,0802	0,0902
0,1	0,1003	0,1104	0,1206	0,1308	0,1409	0,1511	0,1614	0,1717	0,1820	0,1923
0,2	0,2027	0,2132	0,2237	0,2342	0,2448	0,2554	0,2661	0,2769	0,2877	0,2986
0,3	0,3095	0,3205	0,3316	0,3428	0,3541	0,3654	0,3767	0,3884	0,4001	0,4118
0,4	0,4236	0,4356	0,4477	0,4599	0,4722	0,4847	0,4973	0,5101	0,5230	0,5361
0,5	0,5493	0,5627	0,5764	0,5901	0,6042	0,6184	0,6328	0,6475	0,6625	0,6777
0,6	0,6932	0,7089	0,7250	0,7414	0,7582	0,7753	0,7928	0,8107	0,8291	0,8480
0,7	0,8673	0,8872	0,9077	0,9287	0,9505	0,9730	0,9962	1,0203	1,0454	1,0714
0,8	1,0986	1,1270	1,1568	1,1881	1,2212	1,2562	1,2933	1,3331	1,3758	1,4219
0,9	1,4722	1,5275	1,5890	1,6584	1,7381	1,8318	1,9459	2,0923	2,2976	2,6467
0,99	2,6466	2,6996	2,7587	2,8257	2,9031	2,9945	3,1063	3,2504	3,4534	3,8002

Улитина Елена Валерьевна
Леднева Ольга Валерьевна
Жирнова Ольга Леонидовна

СТАТИСТИКА

Учебное пособие

Формат издания $60 \times 90^{1}/_{16}$.
Печать офсетная. Гарнитура NewtonC.
Печ. л. 19,5. Тираж 2500 экз. Заказ № 3167.

Главный редактор *Н. В. Разевиг*
Редактор *Д. А. Рогачева*
Корректор *И. Ф. Козлова*
Компьютерная верстка *В. С. Чукашев*
Дизайн обложки *Б. В. Зипунов*

ООО «Маркет ДС Корпорейшн»
125190, Москва, Ленинградский просп., д. 80, корп. Г, оф. 202(8)
Тел.: (495) 987-43-74
e-mail: book@marketds.ru
www.marketds.ru

Отпечатано в полном соответствии с качеством
предоставленных материалов в ОАО «Дом печати — ВЯТКА».
610033, г. Киров, ул. Московская, 122.

Учебное пособие представляет собой классический вводный курс статистики, направленный на формирование у студентов базовых компетенций обработки и анализа информации, выраженной числовыми данными. Учебный материал охватывает основные вопросы сбора и обобщения данных, формирует системное представление о возможностях и особенностях применения богатого статистического инструментария для выявления закономерностей развития различных социальных и экономических явлений, способствует развитию навыков и компетенций применения статистических методов для решения задач на начальных этапах экономического анализа информации.

Предназначено для студентов начальных курсов высших учебных заведений экономических факультетов, аспирантов и преподавателей вузов, а также всех интересующихся вопросами статистики.

ISBN 978-5-94416-107-9

9 785944 161079

125190, г. Москва,
Ленинградский проспект, д. 80,
корп. Г, оф. 202(8)
тел./факс: (495) 987-43-74
e-mail: book@marketds.ru
www.marketds.ru

Market DS