ВНИМАНИЮ ЧИТАТЕЛЕЙ!

В издательстве "Финансы и статистика" имеется в продаже учебное пособие

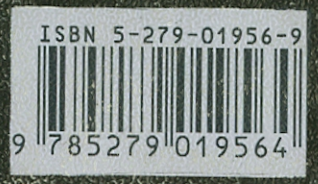
ПРАКТИКУМ ПО ТЕОРИИ СТАТИСТИКИ / Под ред. проф. Р.А. Шмойловой.– 2001. – 23 л.

Составлен в соответствии с типовой учебной программой курса "Теория статистики". Содержит краткий обзор основных понятий общей теории статистики: группировка статистических данных и ее роль в анализе информации, абсолютные, относительные и средние величины, статистические распределения, выборочное наблюдение, корреляционная связь, ряды динамики, индексы и их использование в экономико-статистических исследованиях и т.д. Представлены типовые примеры с решениями и задачи (с ответами) по изучаемому материалу. В приложениях даны необходимые для решения задач математико-статистические таблицы, а также основные формулы теории статистики.

Для преподавателей, аспирантов, студентов экономических вузов, экономистов, менеджеров.

Книгу Вы можете заказать по почте
или приобрести в издательстве по адресу:
101000, Москва, ул. Покровка, 7
(метро "Китай-город", выход на ул. Маросейка)
Тел.: (095) 925-35-02, 925-47-08, факс (095) 925-09-57
E-mail: mail@finstat.ru *http://www.finstat.ru*

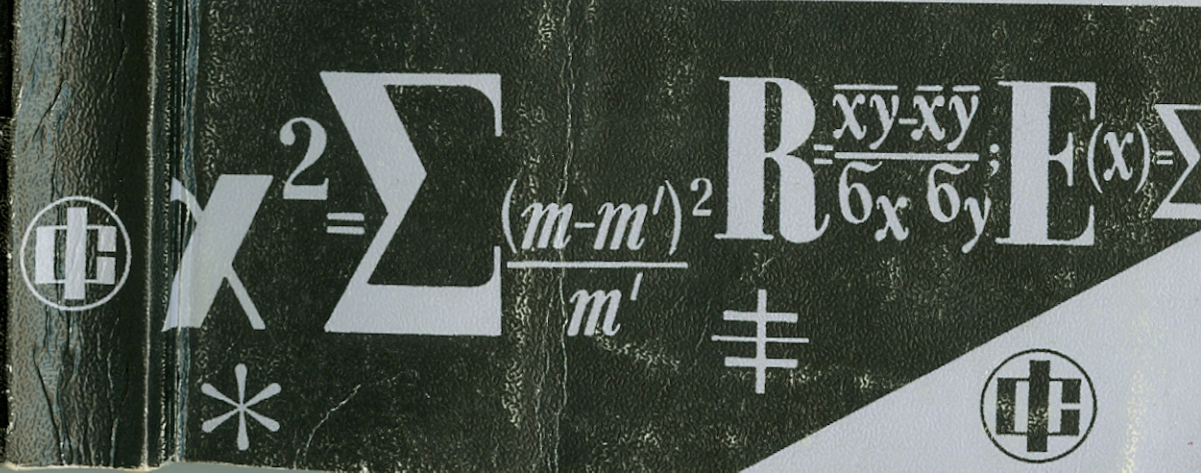
ISBN 5-279-01956-9

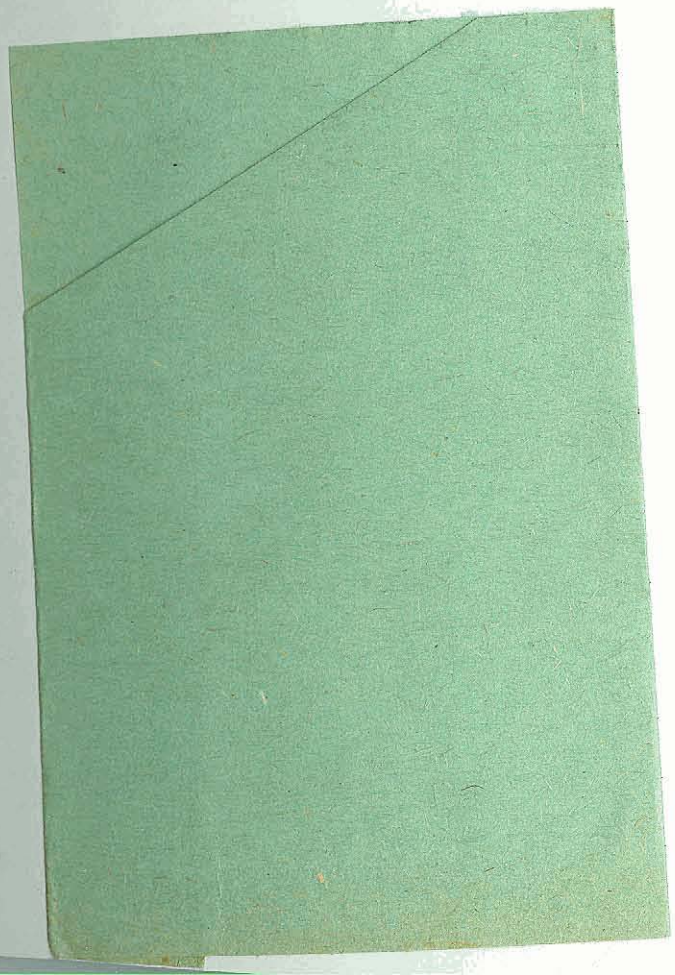


ОБЩАЯ ТЕОРИЯ СТАТИСТИКИ

И.И.ЕЛИСЕЕВА, М.М.ЮЗБАШЕВ

ОБЩАЯ ТЕОРИЯ СТАТИСТИКИ





И.И.ЕЛИСЕЕВА, М.М.ЮЗБАШЕВ

ОБЩАЯ ТЕОРИЯ СТАТИСТИКИ

1610

Под редакцией члена-корреспондента
Российской Академии наук
И.И. Елисеевой

ЧЕТВЕРТОЕ ИЗДАНИЕ

Рекомендовано
Министерством образования
Российской Федерации
в качестве учебника
для студентов высших учебных заведений,
обучающихся по направлению и специальности
"Статистика"

Москва
"Финансы и статистика"
2002

УДК 311(075.8)
ББК 60.6я73
Е51

АВТОРЫ:

И.И. Елисеева,
д-р экон. наук, проф., чл.-корр. РАН — предисловие, главы 1, 2, 4, 6, 7, 10, приложение;

М.М. Юзбашев,
д-р экон. наук, проф. — главы 3, 5, 8, 9, 11.
Параграфы 1.3, 5.9, 6.3, 7.9 написаны авторами совместно

РЕЦЕНЗЕНТЫ:

кафедра общей теории статистики
Московского Государственного университета экономики, статистики и информатики (МЭСИ);

Г.Л. Громыко,
д-р экон. наук, профессор МГУ им. М.В. Ломоносова

Елисеева И.И., Юзбашев М.М.

Е51 Общая теория статистики: Учебник/ Под ред. чл.-корр. РАН И.И. Елисеевой. – 4-е изд., перераб. и доп. – М.: Финансы и статистика, 2002. – 480 с.: ил.

ISBN 5-279-01956-9

Излагаются статистические методы: группировки, выборочный, индексный, корреляционный, анализ динамики. Показаны их взаимосвязи и возможности применения с использованием ПЭВМ в рыночной экономике: в сборе информации в связи с увеличением числа хозяйственных единиц и их типов, аудите, финансовом менеджменте, прогнозировании. Четвертое издание (3-е изд. – 1997 г.) полностью переработано, расширено изложение методов многомерной классификации данных, подробнее рассмотрены применение выборочного метода, методы совмещения индексов и регрессий; введен анализ соотношения индексов экономических показателей. Включена глава, посвященная статистическому изучению структуры данных и ее изменений.

Для студентов, обучающихся по специальности «Статистика», а также для экономических вузов и факультетов.

Е $\frac{0702000000 - 012}{010(01) - 2002}$ 183–2001

УДК 311(075.8)
ББК 60.6я73

ISBN 5-279-01956-9

© И.И. Елисеева, М.М. Юзбашев, 1995
© И.И. Елисеева, М.М. Юзбашев, 1999

ПРЕДИСЛОВИЕ

«Общая теория статистики» – одна из основных дисциплин в системе экономического образования и важнейшая для тех, кто избрал статистику своей профессией.

Термин «статистика» возник во второй половине XVIII в. в связи с познанием государств, описанием их особенностей, достопримечательностей, как тогда предпочитали говорить. К тому же времени относится начало преподавания статистики в университетах Германии.

История развития человечества показала, что без статистических данных невозможно управление государством, развитие отдельных отраслей и секторов экономики, обеспечение оптимальных пропорций между ними. Необходимость сбора и обобщения множества данных о населении страны, предприятиях, банках, фермерских хозяйствах и т. д. приводит к существованию специальных статистических служб – учреждений государственной статистики. В зависимости от того, по какой отрасли организуются сбор, обработка и анализ статистических данных, различают статистику населения, промышленности, сельского хозяйства, капитального строительства, финансов и т. д. Все эти разделы статистики призваны вырабатывать методы сбора и обобщения данных, построения сводных показателей для отражения процессов в соответствующей отрасли. Статистика рассчитывает и общеэкономические показатели – валовой национальный продукт, валовой внутренний продукт, совокупный общественный продукт, национальный доход и т. д.

Статистик нужен и для предприятия, и для страны. Статистические методы позволяют разрабатывать стратегию развития фирмы на основе прогнозирования динамики основных показателей и соотношений между ними. Важное значение для успешной работы фирмы имеют статистические методы контроля и анализа качества продукции. Динамика макроэкономических показателей дает основания для разработки перспективных планов развития экономики в целом, измерения эффективности общественного производства и т. д.

Несмотря на разнообразие сфер применения статистики, имеются общие методы статистической работы, которыми нужно руководствоваться всегда и

везде. Курс «Общая теория статистики» знакомит с общими правилами сбора, обработки и анализа статистических данных.

Статистик имеет дело с числовой и нечисловой информацией, с большими и малыми выборками, с вычислениями, таблицами и графиками. Имеется множество отечественных и зарубежных пакетов прикладных программ статистической обработки данных на персональных компьютерах и больших ЭВМ. Есть специальные программы, предназначенные для обучения студентов, которые содержат подробные объяснения всех процедур и тесты для проверки их усвоения.

С развитием рыночной экономики – увеличением числа хозяйственных единиц, их типов, развитием аудита, финансового менеджмента, статистического прогнозирования и моделирования – задачи отечественной статистики значительно расширились. В практику государственной статистики России внедряются методики, принятые в международной статистике. В учебнике рассмотрены основные процедуры сбора, обработки и анализа массовых данных; возможности их реализации на персональных компьютерах. Особое внимание уделяется обоснованию вероятностного характера статистического вывода, выборочному методу, проверке статистических гипотез.

Этот учебник дает представление об основных статистических методах, их возможностях и границах применения. Если вы захотите более глубоко изучить соответствующий раздел статистики, то в конце каждой главы вы найдете список рекомендуемой литературы.

При изложении материала авторы стремились показать, что статистика не является скучной и трудной наукой, как иногда думают, она может доставить удовольствие и принести пользу.

Авторы считают своим долгом поблагодарить рецензентов книги – коллектив кафедры общей теории статистики МЭСИ и особенно профессора кафедры статистики МГУ им. М. В. Ломоносова д-ра экон. наук Г. Л. Громыко за советы и замечания, а также доцента кафедры статистики Санкт-Петербургского университета экономики и финансов, канд. экон. наук Т. С. Кадибур за поддержку в работе.

В 4-м издании учебника полностью переработана глава 3; внесены дополнения в главы 2, 4, 5, 8, 9; расширено изложение методов многомерной классификации данных (глава 6); подробнее рассмотрены вопросы применения выборочного метода (глава 7); изложены методы совмещения индексов и регрессий; введен анализ соотношения индексов экономических показателей (глава 10); заново написана глава 11, посвященная статистическому изучению структуры данных и ее изменений.

Авторы старались учесть советы и замечания, которые были высказаны при обсуждении предыдущего издания учебника на заседании секции социально-экономических проблем и статистики Санкт-Петербургского Дома ученых РАН[1].

[1] Вопросы статистики. – 1995. – № 12. – С. 30–32.

1 Глава ПОНЯТИЕ О СТАТИСТИКЕ

1.1. ЧТО ТАКОЕ СТАТИСТИКА

Слово «статистика» используется в нескольких значениях: прежде всего как синоним слова *«данные»*. Именно в этом смысле можно сказать: «статистика рождаемости и смертности в России» или «статистика преступлений». Статистикой называется *отрасль знаний,* объединяющая принципы и методы работы с числовыми данными, характеризующими массовые явления. Статистикой называют также *отрасль практической деятельности,* направленной на сбор, обработку, анализ статистических данных.

Слово «статистика» происходит от латинского слова *status* – состояние, положение вещей. Первоначально оно употреблялось в значении «политическое состояние». Отсюда итальянское слово *stato* – государство и *statista* – знаток государства. В научный обиход слово «статистика» вошло в XVIII в. и первоначально употреблялось в значении «государствоведение». В настоящее время *статистика может быть определена как собирание, представление, анализ и интерпретация числовых данных.* Это особый метод, который используется в различных сферах деятельности, в решении разнообразных задач.

Исторически развитие статистики было связано с развитием государств, с потребностями государственного управления. Хозяйственные и военные нужды уже в древний период истории человечества требовали наличия данных о населении, его составе, имущественном положении. С целью налогообложения организовывались переписи населения, проводился учет земель и т. д. Первые работы такого рода отмечены даже в священных книгах разных народов. В античном мире был организован учет родившихся; молодые люди, достигшие 18 лет, вносились в списки военнообязанных, а по достижении 20 лет – в списки полноправных граждан. Составлялись

земельные кадастры, в которые вносились сведения о строениях, рабах, скоте, инвентаре, получаемых доходах. Появились описания государств. Большая заслуга в этом принадлежит греческому философу Аристотелю (384–322 г. до н. э.); он составил описание 157 городов и государств своего времени.

Средневековье оставило уникальный памятник – «Книгу страшного суда» (1061 г.) – это свод материалов всеобщей переписи населения Англии и его имущества (включает данные о 240 тыс. дворов). Со временем собирание данных о массовых общественных явлениях приобрело регулярный характер; с середины XIX в. благодаря усилиям великого бельгийца – математика, астронома и статистика Адольфа Кетле (1796–1874) были выработаны правила переписей населения и регулярность их проведения в развитых странах. Для координации развития статистики по инициативе А. Кетле проводились международные статистические конгрессы, а в 1885 г. был основан Международный статистический институт, существующий и сейчас. Международной статистикой занимаются международные организации – ООН, ФАО, ЮНЕСКО, МОТ, ЕС, Мировой банк и др. Международные организации и государственная статистика каждой отдельной страны занимаются сбором, представлением, сравнением, интерпретацией социально-экономических данных. Сложились методы работы, продолжающие традиции государствоведения.

Другие разделы статистики были развиты при анализе азартных игр (подсчет игровых шансов), изучении процессов воспроизводства населения. Эти достаточно сложные методы, основанные на теории вероятностей, нашли применение прежде всего в страховании и биологии, затем в других естественных науках, психологии и, наконец, с начала XX в. – в социально-экономических исследованиях, в изучении уровня жизни населения, покупательского спроса, качества продукции и т. д. Сегодня трудно найти сферу, где бы не применялась статистика. Статистика нужна для расчета страховых тарифов, оценки финансовых и предпринимательских рисков; она используется в работе аудитора, при постановке управленческого учета в фирме, в контроле и анализе качества продукции, в медицине, спорте и маркетинге. Может быть, только в области искусства статистика не нашла пока широкого применения.

При изучении разных объектов в разных задачах, конечно же, используются различные методы. Тем не менее существуют некоторые общие принципы и методы статистической работы. В учебнике «Теория статистики» английских статистиков Дж. Э. Юла и М. Дж. Кендэла говорится: «Независимо от того, в какой отрасли

знания получены числовые данные, они обладают определенными свойствами, для выявления которых может потребоваться особого рода научный метод обработки. Последний известен как статистический метод или, короче, статистика».

Статистические методы включают как простые методы, которые могут быть понятны любому человеку, так и сложные математические процедуры, доступные специалистам высокого класса. В этом учебнике излагаются простые и часто используемые методы при изучении социальных и экономических явлений и процессов.

1.2. СТАТИСТИЧЕСКАЯ ЗАКОНОМЕРНОСТЬ. СТАТИСТИЧЕСКИЕ СОВОКУПНОСТИ

Статистика позволяет выявить и измерить закономерности развития социально-экономических явлений и процессов, взаимосвязей между ними. Познание закономерностей возможно лишь в том случае, если изучаются не отдельные явления, а *совокупности* явлений – ведь закономерности общественной жизни проявляются в полной мере лишь в массе явлений. В каждом отдельном явлении *необходимое* – то, что присуще всем явлениям данного вида, проявляется в единстве со *случайным*, индивидуальным, присущим лишь этому конкретному явлению. Так, например, реклама какого-либо товара может не оказать влияния на рост объема продажи этого товара, однако обобщение данных о затратах на рекламу товаров и объеме их реализации показывает наличие прямой связи между этими показателями. Поэтому рекламу и называют «двигателем торговли».

Закономерности, в которых необходимость неразрывно связана в каждом отдельном явлении со случайностью и лишь во множестве явлений проявляет себя как закон, называются ***статистическими.***

Кто дольше живет – мужчины или женщины? Можно привести примеры долгожительства мужчин, например существует свидетельство, что англичанину Фоме Карне, родившемуся в 1588 г., удалось прожить 207 лет. Абсолютно точно известно, что азербайджанец Ширали Мислимов прожил 168 лет (1805–1973). Однако это лишь частные примеры. Только при обобщении данных по всему населению выявляются закономерные соотношения. Изучая ожидаемую продолжительность жизни при рождении, видим, что при всем различии в уровне развития стран, их культуры, времени расчета показателя общим для всех является большая ожидаемая (и фактическая) продолжительность жизни женщин (табл. 1.1).

Таблица 1.1

Ожидаемая продолжительность жизни при рождении[1]

Страна	Все население	Мужчины	Женщины
Россия	65	58	72
Австрия	76	73	79
Венгрия	69	65	74
Китай	69	68	71
США	76	73	79
Япония	80	76	83

Источник. Российский статистический ежегодник. 1996. – М.: Логос, 1996. – С. 636.

[1] По России – 1995 г., по остальным странам – 1993 г.

Понятию статистической закономерности противостоит понятие динамической закономерности, проявляющейся в отдельном явлении. Так, площадь круга изменяется с изменением его радиуса, и эта связь выражается формулой $s = 2\pi r^2$, которая справедлива для любого круга. Свойство статистических закономерностей проявляется лишь в массе явлений при обобщении данных по достаточно большому числу единиц; оно получило название *закона больших чисел.*

Соответственно предметом статистического изучения всегда выступают *совокупности* тех или иных явлений, *включающие все множество проявлений исследуемой закономерности.*

Статистические совокупности часто называют *массовыми явлениями.* Они обладают свойством устойчивости – в течение более или менее длительного промежутка времени их характеристики остаются примерно постоянными. Так, доля мальчиков и девочек среди новорожденных, доля лиц разных возрастов среди вступающих в брак и т. д. обнаруживает от года к году не очень значительные колебания. Этот факт представляет громадный интерес. Устойчивость определяет возможность существования и развития общества, на этом свойстве базируются прогнозы, скажем, прогноз пропорций между отраслями и секторами экономики и т. д.

Каждое единичное явление рассматривается статистикой как особый, частный случай изучаемой закономерности. Статистика дает количественную характеристику исследуемой закономерности, а это возможно лишь при обобщении всего множества ее проявлений,

взятых в целом, т.е. на основе совокупности явлений. Количественная характеристика каждого отдельного явления отражает его сущность. Но эта частная характеристика ограничена в своем значении для познания закономерности, так как сложилась в конкретных условиях и в силу этого соединяет в себе как типичные черты, присущие всем явлениям данного вида, так и случайные, присущие именно этой конкретной единице. Так, в Санкт-Петербурге имеются крупные семьи, включающие родителей и 5–6 детей; бездетные семьи, состоящие лишь из мужа и жены, но и те, и другие не являются типичными, закономерными для этого города. Лишь обобщив данные по всем петербургским семьям, можно говорить о том, какой размер семьи закономерен для этого города на современном этапе (3,1 человека), а также о том, что типичными являются так называемые простые семьи, состоящие из родителей и одного-двух детей, без прочих родственников. Поскольку статистика призвана выявлять закономерное, она, опираясь на данные о каждом отдельном проявлении изучаемой закономерности, обобщает их и таким путем получает *количественное выражение этой закономерности.*

Статистика не связана с каким-либо конкретным измерителем. Она использует как стоимостные, так и натуральные показатели. Для анализа динамики стоимостные показатели выражаются не только в текущих ценах, но и в так называемых неизменных ценах, т. е. ценах, установленных за определенный период или на определенную дату, применяемых в течение ряда лет для оценки продукции в отдельных отраслях материального производства.

Стоимостное выражение позволяет агрегировать данные: например, рассчитывать валовую продукцию предприятия, объединения, отрасли. Универсальным измерителем являются затраты труда в человеко-часах, человеко-днях и т.д. При обобщении натуральных показателей могут возникнуть трудности из-за несопоставимости данных. Преодолеть их позволяют *условно-натуральные измерители.* Например, рыбные консервы выпускаются в больших и маленьких банках, высоких и низких, причем в разные годы соотношение между ними меняется. Для того чтобы подсчитать, сколько всего произведено консервов, сравнить эту цифру с прошлым периодом, используют так называемые условные банки. Чтобы обобщить мощность двигателей по совокупности предприятий, их выражают в лошадиных силах, а затем суммируют. Топливо разной теплотворной способности пересчитывают в условное топливо; скот (коров, быков и т.д.) пересчитывают условно в крупный рогатый скот и т. д.

Несмотря на материальные различия изучаемых статистикой совокупностей, все они имеют общие черты.

достаточно лишь изучить одну единицу, чтобы знать все о всех явлениях этого вида.

Вариация – основа существования мира и источник его развития. Если бы люди не делились на мужчин и женщин, человечество прекратило бы существование; если бы не было различных мнений – истина была бы недостижимой, а жизнь без вариаций – невыносимо скучной!

1.3. ПРИЗНАКИ И ИХ КЛАССИФИКАЦИЯ

Единицы совокупности обладают определенными свойствами, качествами. Эти свойства принято называть *признаками.* Например, признаки человека: возраст, образование, занятие, рост, вес, семейное положение и т.д.; признаки предприятия: форма собственности, специализация (отрасль), численность работников, величина уставного фонда, экономическая эффективность его деятельности и т. д.

Статистика изучает явления через их признаки: чем более однородна совокупность, тем больше общих признаков имеют ее единицы и меньше варьируют их значения.

Таблица 1.2

Классификация признаков в статистике

Основная классификация				
по характеру их выражения	по способу измерения	по отношению к характеризуемому объекту	по характеру вариации	по отношению ко времени
1. Описательные	1. Первичные или учитываемые	1. Прямые (непосредственные)	1. Альтернативные	1. Моментные
2. Количественные	2. Вторичные или расчетные	2. Косвенные	2. Дискретные	2. Интервальные
			3. Непрерывные	

Признаки различаются способами их измерения и другими особенностями, влияющими на приемы статистического изучения. Это дает основание для классификации признаков (табл. 1.2).

Описательные признаки выражаются словесно: национальность человека, разновидность почв, материал стен здания. Описательные признаки подразделяются на *номинальные и порядковые.* Эти термины взяты из теории измерений. Отличия между ними в том, что *номинальные* – это описательные признаки, по которым нельзя ранжировать данные, а *порядковые* – это признаки, по которым можно ранжировать, упорядочивать данные. Например, пользуясь оценками экспертов, ранжируют фигуристов по технике и артистичности исполнения программы или работников по мастерству и т. д.

Количественные признаки выражены числами. Они играют преобладающую роль в статистике. Таковы возраст человека, площадь пашни, заработная плата рабочих, население города, доход кооператива и т. д.

Первичные признаки характеризуют единицу совокупности в целом. Это абсолютные величины. Они могут быть измерены, сосчитаны, взвешены и существуют сами по себе, независимо от их статистического изучения. Например, площадь пашни, мощность двигателей на предприятии, численность населения города, число автомобилей, произведенных в стране.

Вторичные, или *расчетные,* признаки не измеряются непосредственно, а рассчитываются. Они являются продуктами человеческого сознания, результатом познания изучаемого объекта. Например, себестоимость единицы продукции, производительность труда, рентабельность, урожайность и т. п. Вторичные признаки представляют собой соотношения первичных признаков: деление объема выпущенной продукции на численность работников дает показатель производительности труда; деление суммы затрат на произведенную продукцию на число единиц данной продукции дает себестоимость и т. д. Несмотря на расчетный характер признаков, они тоже имеют объективный характер. Процесс познания есть отражение объективных свойств явлений и процессов, и расчеты, статистические методы познания являются таким же необходимым средством отражения объективных свойств совокупности, как измерение, взвешивание. Вторичный – не означает второстепенный. Термин определяет только путь познания: сначала надо измерить значения первичных признаков, а уже потом, во вторую очередь, на основе первичных признаков рассчитать значения вторичных.

Прямые (непосредственные) признаки – это свойства, непосредственно присущие тому объекту, который ими характеризуется.

Таковы возраст человека, поголовье коров на ферме, объем продукции завода, численность его рабочих.

Косвенные признаки являются свойствами, присущими не самому объекту, а другим совокупностям, относящимся к объекту, входящим в него. Например, продуктивность коров как косвенный признак фермы. Хотя продуктивность не фермы, а коров – это их прямой признак, но ведь продуктивность характеризует и ферму, которой принадлежат эти коровы (или даже целую область). Такова и оплата труда рабочих по отношению к заводу. Это косвенный признак завода, но очень важный для того, кто собирается поступать на работу и выбирает предприятие.

Практически деление признаков на прямые и косвенные совпадает с их делением на первичные и вторичные. Признаки различаются в статистике и по характеру их вариации, т.е. по различиям их значений у разных единиц совокупности. Выделяются *альтернативные* признаки, которые могут принимать только два значения. Таковыми являются признаки обладания или необладания чем-то. Например, все садовые участки по признаку наличия посадок вишни можно разделить на имеющие посадки вишни и не имеющие их. Альтернативным признаком являются пол человека, место проживания (город, село), двигатель трактора (гусеничный или колесный).

К дискретным относятся количественные признаки, которые могут принимать только отдельные значения, без промежуточных значений между ними. Дискретные признаки, как правило, целочисленные. Это число членов семьи, количество этажей здания, комнат в квартире.

Непрерывные, точнее, непрерывно варьирующие признаки способны принимать любые значения, конечно, в определенных границах. К непрерывным относятся расчетные вторичные признаки. Ведь их значения – результат деления, а оно может приводить к любым числам – целым, дробным, иррациональным. На практике значения непрерывных признаков округляют с конечной степенью точности, так что они становятся квазидискретными. С другой стороны, дискретные по существу признаки, например число работников предприятия на 1 января, поголовье коров на ту же дату, имеют такое громадное число возможных значений, что на практике статистика вынуждена обращаться с ними, как с квазинепрерывными. Об этом будет сказано в главах 5 и 6 при обсуждении метода группировок.

Моментные признаки характеризуют изучаемый объект в какой-то момент времени, установленный планом статистического исследования. Они существуют на любой момент времени и характеризуют наличие чего-либо: численность населения, стоимость фондов, количество скота, размеры жилой площади.

К интервальным относятся признаки, характеризующие результаты процессов. Поэтому их значения могут возникать только за интервал времени: год, месяц, сутки, но не на момент времени. Таковы число родившихся, умерших, объем промышленной продукции, надой молока, сумма полученной прибыли. Различие между моментными и интервальными признаками существенно при изучении динамики (см. гл. 9).

Единицы измерения моментных признаков относятся только к характеризуемым ими свойствам объектов, а единицы измерения интервальных признаков содержат еще и указание того отрезка времени, за который определено значение признака. Так, стоимость основных производственных фондов предприятия на 1 января выражается в миллионах рублей, а объем продукции за январь – в тысячах или миллионах рублей за месяц.

1.4. ОПРЕДЕЛЕНИЕ ПРЕДМЕТА СТАТИСТИКИ – ОСНОВА СТАТИСТИЧЕСКОЙ МЕТОДОЛОГИИ

Как уже отмечалось, предметом статистического изучения всегда выступает совокупность явлений. Как правило, она включает в себя несколько *частных совокупностей,* представляющих особые *типы* явлений, иначе говоря, особые модификации изучаемой закономерности. Единицы разных частных совокупностей в рамках общего качества отличаются кругом признаков и их значений.

В большинстве случаев правильным будет представление частной совокупности (однородной группы), состоящей из *ядра* и окружающих его явлений – *слоя.* Ядро – концентрированное выражение всех специфических свойств типа (группы), определяющих качественное отличие данного типа от всех иных. Кроме единиц, составляющих ядро, тип включает явления переходного качества, принадлежность которых к данному типу может быть установлена с определенной вероятностью. Такие явления образуют, так сказать, «полосу размыва» между типами.

Среди студентов можно встретить такой тип, как «идеальный студент»: прекрасно учится, много читает, хороший товарищ. Есть студенты не такие разносторонние, для которых важны только специальные знания; есть и другие типы. «Качество» одних студентов, их принадлежность к тому или иному типу можно определить практически безошибочно, тогда как в отношении других вопрос реша-

ется в определенной степени условно. Они-то и представляют собой явления переходного качества.

Соотношение между ядром и его окружением в разных типах будет, конечно, различным: это зависит от устойчивости типа, длительности его существования, взаимодействия с другими типами той же совокупности, с другими совокупностями. Однако ядро должно составлять большинство единиц того или иного типа, так как именно ядро определяет «лицо» типа, его характерные свойства.

Социально-экономическая статистика изучает совокупности однокачественных явлений в конкретных условиях места и времени. Таким образом, статистика располагает всегда *ограниченным числом данных.* Каждое явление возникает как результат множества факторов. В естественных науках можно проследить интересующие взаимосвязи с помощью специально проведенных лабораторных экспериментов, которые называют активными экспериментами, так как исследователь практически полностью контролирует ход эксперимента и может выделить в более или менее чистом виде влияние каждого из выбранных факторов, элиминируя влияние остальных. Иная ситуация в социально-экономических исследованиях. «При анализе экономических форм нельзя пользоваться ни микроскопом, ни химическими реактивами. То и другое должна заменить сила абстракции», – писал К.Маркс (Маркс К., Энгельс Ф. Соч.– Изд. 2-е. – Т. 23.– С. 4). Применяя различные методы анализа, мы проводим «пассивный» эксперимент, причем ни один метод не позволяет определить «чистый» вклад каждого из факторов по отдельности в совокупный результат.

Важно то, что в центре социально-экономических явлений и процессов находится человек со своими субъективными установками, активным воздействием на окружающий мир; это делает достоверность данных особой проблемой статистики.

Обобщая сказанное, можно указать следующие *особенности социально-экономических явлений:* 1) сложность их материальной природы, многообразие количественных и качественных определений; 2) ограниченность численности; 3) динамичность; 4) многообразие видов и форм, в которых проявляются единые по своей сущности процессы, отсюда – разделение на частные совокупности, на группы особого качества; 5) взаимосвязанность явлений и признаков; невозможность элиминирования действия факторов и раздельной оценки их действия.

Специфика предмета статистики обусловливает специфику *статистического метода.* Он включает *сбор данных* (статистическое

наблюдение), их *обобщение, представление, анализ* и *интерпретацию.* Статистические данные могут быть взяты из публикаций, а можно собрать новую информацию по каждой единице совокупности (фирме, человеку, виду продукции, товару). Получение исходных данных является одной из наиболее трудных и важных задач, которые встают перед статистикой. Главное – использовать те данные, которым можно доверять. Проблемы статистического наблюдения рассматриваются в гл. 2 и 7.

Обобщение данных наблюдения включает *группировку* – разграничение общей совокупности на группы однородных единиц и *сводку* – обобщение значений признаков в сводные статистические показатели для характеристики каждой частной совокупности, группы и совокупности в целом (см. гл. 3, 5, 6).

Чтобы пользоваться результатами обобщения или непосредственно исходной информацией, данные должны быть представлены в подходящей форме, компактно и наглядно. С этой целью строятся таблицы и графики (см. гл. 4).

Процесс анализа охватывает все стадии статистического исследования. Каждый следующий этап статистической работы зависит от предыдущего. Этап обобщения данных оказывает влияние на статистическое наблюдение — ведь именно тем, что мы хотим получить в результате исследования, определяются границы объекта наблюдения, программа наблюдения (какие признаки мы будем регистрировать у единиц совокупности).

Выделение типов в результате классификации или группировки данных обеспечивает их однородность. Тем самым создается основа для расчета сводных показателей, анализа вариации и связей. Однородность обобщаемых данных определяет устойчивость всех статистических показателей. Так, по-разному рассчитывается средний надой молока в целом по России, объединяя центральные районы, Северный Кавказ, Северо-Запад и т. д., и этот же показатель по отдельным территориям страны с достаточно однородными природно-климатическими условиями.

При изучении связей статистика помогает установить круг важнейших факторов, измерить хотя бы и условно силу их влияния (см. гл. 8, 10). В решении этой задачи всегда существует опасность установления ложных связей – принять за причину просто сопутствующие явления. Например, считать черного кота или разбитое зеркало предвестием неудач.

Важным направлением анализа является изучение динамики. Чтобы предсказать развитие в будущем (сколько автомобилей будет произведено и продано на внутреннем рынке, какова будет числен-

1620

ность населения в 2000 г. и т.д.), нужно знать фактическую динамику в прошлом: как изменялись показатели, имелась ли тенденция в их изменении, каков характер колеблемости данных.

Каждый шаг исследования завершается интерпретацией полученных результатов: какое заключение можно сделать, исходя из проведенного анализа, что говорят нам цифры – подтверждают ли они исходные предположения или открывают что-то новое? Интерпретация данных ограничена исходным материалом. Если заключения основаны на данных выборки, то она должна быть репрезентативной, чтобы выводы были отнесены к совокупности в целом (см. гл. 7). Статистика позволяет выяснить все то полезное, что содержится в исходных данных, и определить, что и как можно использовать в принятии решений.

Рекомендуемая литература к главе 1

1. *Елисеева И. И.* Моя профессия – статистик.– М.: Финансы и статистика, 1992.
2. *Плошко Б. Г.* Группировка и системы статистических показателей.– М.: Статистика, 1978.

Глава 2

СТАТИСТИЧЕСКОЕ НАБЛЮДЕНИЕ

2.1. ОРГАНИЗАЦИЯ ГОСУДАРСТВЕННОЙ СТАТИСТИКИ В РОССИЙСКОЙ ФЕДЕРАЦИИ И МЕЖДУНАРОДНОЙ СТАТИСТИКИ

Как уже отмечалось, статистическая работа состоит в том, чтобы собрать числовые данные о массовых явлениях, обработать их, представить в форме, удобной для анализа, проанализировать и интерпретировать полученные результаты.

Собирание данных лежит в основе всего исследования. От качества используемых данных, от их достоверности и точности зависит достоверность результатов анализа. Люди по-разному относятся к статистической информации: одни не воспринимают ее, другие безоговорочно верят, третьи согласны с мнением английского политика Б. Дизраэли (1804–1881): «Есть ложь, есть наглая ложь, а есть статистика». Однако ему же принадлежит следующее утверждение: «В жизни, как правило, преуспевает больше тот, кто располагает лучшей информацией». На основе статистической информации правительство разрабатывает свою экономическую и социальную политику, оценивает ее результаты, составляет экономические прогнозы. Статистическая информация обеспечивает подготовку двухсторонних и многосторонних экономических соглашений между государствами. Статистика дает информацию для решения региональных задач, для предпринимательской деятельности – об уровне цен на товары в разных регионах, объемах реализации товаров, условиях кредитования, уровне и темпах инфляции, занятости и т.д.; наконец, в той или иной степени статистика нужна каждому из нас для принятия решений по выбору стратегии поведения.

На любом уровне и в любой сфере эффективность использования статистики во многом определяется качеством исходных данных.

Где можно получить статистические данные?

Статистические данные могут быть прежде всего получены из различных публикаций, например такие статистические данные, как валютные курсы, объемы продаж валют, колебания цен, темпы и формы приватизации и т. д. приводятся в журнале «Эксперт», газетах «Экономика и жизнь», «Финансовая газета», «Коммерсант daily» и др.

Но главными источниками опубликованной статистической информации являются издания органов государственной статистики. Наиболее полную информацию о Российской Федерации содержит статистический ежегодник «Российская Федерация в 19... году», издаваемый Государственным комитетом по статистике Российской Федерации (Госкомстатом РФ) – высшим органом государственной статистики нашей страны.

Этот сборник содержит макроэкономические показатели, такие, как валовой внутренний продукт (ВВП) – его формирование и использование; произведенный национальный доход и валовой общественный продукт, государственный бюджет РФ, показатели различных отраслей экономики; данные о развитии системы образования и здравоохранения, о заболеваемости, численности и естественном движении населения, т. е. рождаемости, смертности, брачности, разводимости, ожидаемой продолжительности жизни, миграции населения – перераспределении населения между территориями РФ, миграции в пределах СНГ, эмиграции в страны дальнего зарубежья. Специальный раздел статистического ежегодника составляют данные о потреблении населения, его совокупных и денежных доходах. Сборник содержит некоторые статистические данные и по другим странам. Госкомстат РФ издает и специализированные статистические сборники по отраслям народного хозяйства, социальной статистике. Много данных государственной статистики публикуется в журнале «Вестник статистики» (с 1994 г. – «Вопросы статистики»). Издаются экспресс-информация, экономические доклады, бюллетени, журнал «Статистическое обозрение».

Предоставление статистической информации – главная задача органов государственной статистики и продукция их деятельности. Как любая продукция, она имеет стоимость. Особенно дорогой является та информация, получение которой выходит за пределы программы работы государственной статистики.

Структура органов государственной статистики соответствует административно-территориальному делению страны. В двух городах – Москве и Санкт-Петербурге – имеются местные комитеты по статистике, то же – в автономных республиках. В краях и областях также работают комитеты статистики. Низовым звеном являются районные инспектуры государственной статистики, которые имеются в административных районах краев и областей, крупных городов.

Местные статистические органы издают региональные статистические сборники. Например, Петербургкомстат издает ежегодник «Народное хозяйство Санкт-Петербурга и Ленинградской области в 19... году»; Комитет по статистике г. Москвы – сборник «Москва в цифрах» и т. д. Издаются и специализированные статистические сборники по отдельным отраслям экономики, социальной статистике, статистике населения и т. д. Однако тираж их, как правило, крайне ограничен, и потому сборники малодоступны неспециалистам.

Основные функции всех статистических органов состоят в сборе, обработке, анализе и представлении данных в удобном пользователю виде. Статистические службы должны оперативно предоставлять информацию органам управления, осуществлять обмен информацией с Центральным банком РФ и его конторами на местах, Министерством финансов РФ и его местными органами, Госкомимуществом РФ и его службами, Комитетом по труду и занятости РФ и т. д.

Все статистические органы, кроме райинспектур, имеют внутреннюю структуру: отделы (а в Госкомстате РФ – управления) статистики предприятий, сельского хозяйства, капитального строительства и т. д. Государственная статистика призвана отражать комплексность развития народного хозяйства страны и отдельных регионов, взаимосвязи между территориями. Эти функции выполняют отделы (управления) балансовых работ и системы национальных счетов, статистики финансов, сводный отдел.

Оперативность и качество статистических работ зависят от развития технологии сбора, передачи, обработки и хранения информации. Все областные, краевые и республиканские управления и комитеты по статистике имеют вычислительные центры. Мощный вычислительный центр имеет Госкомстат РФ (ГВЦ РФ). Все большее значение приобретают локальные вычислительные сети, связывающие банки данных статистических служб, других держателей региональной и федеральной информации. Государственный комитет по статистике РФ входит в структуру федеральных органов исполнительной власти.

Госкомстат РФ является методологическим и организационным центром работы всех служб государственной статистики. В его

структуре имеется специальное Управление методологии статистических работ. Здесь разрабатываются федеральный план статистических работ на год и перспективу, методология расчета статистических показателей, сбора и разработки статистических данных.

Большую роль в методологической работе играет Научно-исследовательский институт статистики Госкомстата РФ. В этой работе принимает участие и Научно-методологический совет Госкомстата РФ, который объединяет ведущих работников государственной статистики и представителей экономической и статистической науки.

В последние годы методологическая работа Госкомстата РФ направлена на внедрение интегрированной системы учета и статистики, соответствующей международным стандартам, прежде всего на разработку системы национальных счетов РФ, позволяющей исследовать формирование основных пропорций экономики и рассчитывать важнейшие макроэкономические показатели, используемые в международной практике, а также на измерение инфляции и уровня жизни. Эта работа ведется при участии международных статистических организаций и национальных статистических служб развитых стран. Широко распространились международные связи между национальными статистическими службами и на региональном уровне.

Организация *международной статистики* осуществляется статистическими службами Организации Объединенных Наций (ООН), специализированных учреждений (МОТ, ФАО, ВОЗ и др.) и других международных организаций – Организации экономического сотрудничества и развития (ОЕСД), Европейского сообщества (ЕС), Международного валютного фонда (МВФ), Мирового банка и т.д. Деятельность статистических служб этих организаций включает разработку международных стандартов, обеспечивающих сравнимость статистических показателей разных стран, осуществление международных сопоставлений, публикацию данных по группам стран, регионам и миру в целом. Зарубежные национальные и международные статистические публикации, так же как и отечественные, можно подразделить на сводные, включающие данные по всем разделам статистики, и специализированные, включающие данные по какой-либо одной отрасли, например по финансовой, демографической, сельскохозяйственной и другой статистике. Из сводных изданий важнейшим является ежегодник ООН – Statistical Yearbook. Научным международным центром в области статистики является Международный статистический институт (ISI).

Координация деятельности статистических служб стран – членов СНГ осуществляется созданным в 1992 г. Статистическим комитетом Содружества Независимых Государств. Публикуются статистические сборники по странам СНГ и другим государствам ближнего зарубежья.

Статистические публикации – это один из возможных источников статистической информации. Используя его, следует критически относиться к статистическим данным, прикидывая, насколько та или иная цифра реальна. Полезно иметь данные из разных источников. Если они различаются значительно в несколько раз, то данные не вызывают доверия. Пользоваться лучше теми данными, способ получения которых понятен. Достоверность данных государственной статистики определяется тем, что это результат профессиональной деятельности специально подготовленных работников, использующих единую методологию, соответствующую в большинстве случаев международным стандартам, дающую возможность проследить динамику какого-либо показателя за ряд лет.

Если нет соответствующих данных в статистических сборниках, то можно получить их самим, т. е. провести *статистическое наблюдение* – научно организованный сбор данных. В системе государственной статистики не менее трети всего объема работ связано с получением данных.

Кем бы и когда бы ни проводилось статистическое наблюдение, оно должно быть организовано по определенным правилам, соблюдение которых позволяет обеспечить надежную основу статистического исследования.

2.2. ТРЕБОВАНИЯ, ПРЕДЪЯВЛЯЕМЫЕ К СОБИРАЕМЫМ ДАННЫМ. ФОРМЫ ОРГАНИЗАЦИИ И ВИДЫ СТАТИСТИЧЕСКОГО НАБЛЮДЕНИЯ

Собираемые данные должны отвечать двум требованиям: достоверности и сопоставимости. *Достоверность* – это соответствие данных тому, что есть на самом деле. Вся методика, организация и техника проведения статистического наблюдения должны быть нацелены на обеспечение достоверных данных. Чтобы понять характер задач, возникающих при этом, представим статистическое наблюдение в виде взаимодействующих компонентов (схема 2.1).

Очевидно, что достоверность данных зависит как от характеристик самого статистика – его профессиональной подготовки, ком-

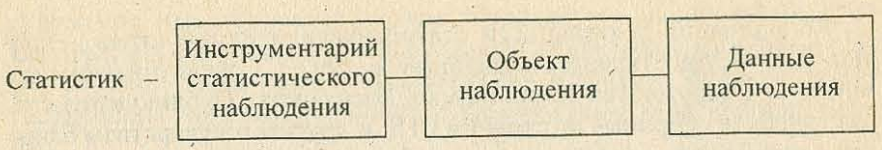


Схема 2.1. Составляющие статистического наблюдения

муникабельности, организационных навыков и т. д., так и от качества используемого инструментария – программы наблюдения, бланков, анкет, инструкций по их заполнению. Они в конечном счете тоже зависят от статистика. На достоверность данных влияет и подготовленность объекта к статистическому обследованию. Это может быть сделано в форме предварительного извещения населения о предстоящем обследовании — в газетах, по радио, телевидению. Влияет на достоверность и упорядочение названия улиц и нумерации домов, квартир и т. д.

На достоверность данных влияет социальная функция показателя. Известно, например, о фактах недостоверности данных о младенческой смертности (смертности детей до одного года). Основной недоучет составляют случаи, когда факт рождения ребенка, умершего вскоре после рождения, умышленно регистрируется как мертворождение, а часть случаев мертворождений записывается как поздние выкидыши и не регистрируются в органах загса. Недостоверными могут быть данные о характере и числе преступлений, профессиональной заболеваемости и т. п., т. е. те данные, которые свидетельствуют о «здоровье» общества.

Общими условиями обеспечения достоверности являются полнота охвата наблюдаемого объекта; полнота и точность регистрации данных по каждой единице наблюдения.

Чтобы данные об отдельных явлениях можно было обобщать, они должны быть сопоставимы друг с другом: собираться в одно и то же время, по единой методике. Кроме того, должна быть обеспечена сравнимость с прошлыми исследованиями, чтобы можно было понять, как изменяется явление. Для этого должна быть полная ясность организации и методологии статистического наблюдения, чтобы были понятны характер и причины различий в данных наблюдений, если таковые были вызваны именно организационно-методологическими факторами.

Сравнимость данных разных наблюдений выполняется, если использовались одно и то же определение единицы наблюдения, одна и та же методика регистрации первичных признаков и методика расчета вторичных признаков, таких, как себестоимость, производительность труда, рентабельность, ликвидность и т. д.

Важным условием сравнимости является сохранение времени проведения наблюдения и периода или момента, к которому относятся регистрируемые данные. Например, численность студентов университета определяется на начало учебного года, стипендиальный фонд – на полгода (или год) и т. д. Обычно рекомендуется, чтобы данные соответствовали хотя бы одному полному циклу изучаемого процесса, например учебному, хозяйственному или финансовому году и т.д. Если сильно влияет сезонность, данные должны собираться по месяцам или по кварталам. Время наблюдения выбирается таким образом, чтобы наблюдаемый объект находился в наиболее стабильном состоянии.

Статистическое наблюдение подразделяется на виды – по *времени* наблюдения и по *охвату* единиц наблюдения.

По времени регистрации фактов различают непрерывное (текущее), периодическое и единовременное наблюдение. *Непрерывное (текущее) наблюдение* ведется систематически, постоянно, непрерывно, по мере возникновения явлений. Например, регистрируются в загсе рождения и смерти, браки и разводы, на предприятиях учитываются выпуск продукции, явки и неявки работников, расчеты с дебиторами и кредиторами, поступление денег в кассу и денежные выплаты и т. п. При *периодическом наблюдении* регистрация проводится через определенные, обычно одинаковые промежутки времени, например учет успеваемости студентов по данным экзаменационных сессий. *Единовременное наблюдение* проводится один раз для решения какой-либо задачи или повторяется через неопределенные промежутки времени по мере надобности, например перепись жилого фонда, школьная перепись и т. д.

Применение на практике того или иного вида наблюдения зависит от специфики исследуемого объекта. Так, функционирование общественного производства носит непрерывный характер: ежедневно производится и потребляется множество различных видов продукции, изменяются их запасы и т. д. Обеспечение бесперебойного производства требует непрерывного систематического учета затрат на производство и его результатов. Иной характер носят изменения в составе населения по социальному или национальному признаку, образованию и пр. В обычных условиях для больших групп населения эти признаки несущественно изменяются в короткие промежутки времени, поэтому нет надобности в непрерывной их регистрации.

Бывает, что для изучения одного и того же процесса используется как текущее, так и единовременное наблюдение. Например, потребление населением изучается государственной статистикой по

данным текущего наблюдения (бюджетные обследования). В то же время многими исследовательскими коллективами потребление изучается по данным единовременных наблюдений: фиксируются «обычные» дневные покупки продовольствия, иногда эти данные дополняются данными фактических покупок за последние 2–3 дня; фиксируются наличие предметов длительного пользования, покупки непродовольственных товаров за последний месяц, квартал или полгода и т. д.

По охвату единиц совокупности различают сплошное и несплошное наблюдение.

При *сплошном наблюдении* регистрации подлежат все без исключения единицы совокупности. Оно применяется, например, при переписи населения, сборе данных в форме отчетности, охватывающей предприятия разных форм собственности, учреждения и организации и т. д.

Развитие многоукладной экономики увеличило число объектов экономической деятельности. Это способствовало расширению практики *несплошного наблюдения*, которое, в свою очередь, подразделяется на способ *основного массива*, *выборочное* и *монографическое*.

При способе *основного массива* обследованию подвергается основной массив – та часть единиц, которая вносит наибольший вклад в изучаемое явление. Часть совокупности, о которой заведомо известно, что она не играет большой роли в характеристике совокупности, исключается из наблюдения, т. е. при этом методе отбираются и обследуются наиболее крупные единицы. Логика метода состоит в том, что крупные единицы могут практически определять интересующие нас статистические показатели. Например, вследствие концентрации производства в отрасли несколько наиболее крупных предприятий могут давать основной объем продукции, в то время как большая масса мелких предприятий выпускает ее незначительную часть. Это бывает при высоком уровне монополизма в отрасли экономики, особенно в условиях региона. Так, в Санкт-Петербурге в 1991 г. всего лишь на 7 предприятиях машиностроения и металлообработки, которые составляли 1,3% от числа промышленных предприятий города, работало около 20% работников. На каждом из этих предприятий было занято свыше 10 тыс. человек, в эту группу входили такие гиганты, как «Кировский завод» – 25 тыс. человек, «Ленинец» – 22,9 тыс. человек и т. д. В подобных условиях логично наблюдать только наиболее крупные предприятия, а мелкие либо вообще игнорировать, либо провести досчет приходящейся на них доли продукции. Поскольку их доля невелика, то ошибка

при распространении данных «основного массива» на всю совокупность будет незначительной. Точность досчета зависит от того, какими сведениями о не охваченной наблюдением части совокупности мы располагаем.

Часто применение метода основного массива требует установления *ценза – значения признака, которое ограничивает объект* наблюдения. Например, обследуются предприятия с числом работников 500 человек и более, или устанавливается, что обследованию подлежат малые предприятия с численностью работников до 100 человек (или до 200 человек). Такой метод называется *цензовым*.

Следует иметь в виду, что термин «ценз» применяется в статистике не только в смысле пограничного значения признака, но и для обозначения переписей. В США, Англии цензами называют переписи населения, промышленности и т.д.

При *выборочном наблюдении* обследованию подвергается отобранная в определенном порядке часть единиц совокупности, а получаемые результаты распространяются на всю совокупность.

В выборке полностью реализуется основная идея несплошного наблюдения. При этом получают информацию о всей совокупности, изучив лишь ее часть. Чтобы понять, хорошее пиво или нет, не обязательно выпить целую бочку, то же можно сказать в отношении проверки качества любой продукции. В решении такого рода задач, да и во многих других случаях, может помочь только выборка.

Выборочный метод играет все большую роль в отечественной статистике, поэтому планирование выборки, методы отбора, оценки ее репрезентативности специально рассматриваются в гл. 7.

Обследования основного массива и выборки – это массовые наблюдения, охватывающие множество единиц. При *монографическом наблюдении* подробно описываются отдельные единицы совокупности в целях их углубленного изучения, которое не может быть столь же детальным при массовом наблюдении. Главное внимание обращается на качественные стороны явления, его поведение, ориентацию, перспективы развития и т. д. Примерами монографических обследований являются этнографические обследования, когда изучается образ жизни семьи или нескольких семей, и др.

В любом обследовании источником получения первичных данных могут быть непосредственное наблюдение, документы и опрос.

Непосредственное наблюдение осуществляется путем регистрации изучаемых единиц и их признаков на основе непосредственного осмотра, подсчета, взвешивания, показаний приборов и т. д. Например, во время переписи вагонов проводится осмотр каждого

вагона. Примером непосредственного наблюдения являются регистрация цен и объема реализации товаров на рынках; метеорологические наблюдения – регистрация температуры воздуха, снежного покрова, суммы осадков; инвентаризация остатков товарно-материальных ценностей на складе.

Документальный способ наблюдения основан на использовании в качестве источника статистических сведений различных документов первичного учета предприятий, учреждений и организаций, поэтому этот способ наблюдения часто называют отчетным. Он применяется, например, при переоценках основных фондов (средств) предприятий и организаций, которые дают основу начисления амортизации, анализа использования фондов и их структуры, особенно в условиях инфляции. При заполнении государственной статистической отчетности по переоценке каждым самостоятельным предприятием любой отрасли и формы собственности используются следующие данные первичной учетной информации: инвентаризационные описи, инвентарные карточки основных фондов, технические паспорта или другая соответствующая документация и данные бухгалтерского учета.

Непосредственное наблюдение и документальный способ обеспечивают наибольшую достоверность статистических данных.

При *опросе* источником данных являются сведения, которые дают опрашиваемые лица. При этом могут быть использованы разные способы собирания данных: экспедиционный, корреспондентский и саморегистрация.

Экспедиционный способ заключается в том, что специально подготовленные регистраторы на основе опроса заполняют переписные формуляры, одновременно контролируя правильность получаемых ответов. Этот способ обеспечивает достаточно точные результаты, но он дорогостоящий. В отечественной статистике экспедиционный способ используется при переписях населения.

Корреспондентский способ заключается в том, что статистические или другие организации рассылают специально разработанные бланки и инструкции к их заполнению отдельным организациям или специально подобранным лицам, давшим согласие периодически заполнять бланки и присылать статистическому органу в установленные сроки. Например, научно-исследовательский институт по изучению спроса населения на товары народного потребления и конъюнктуры торговли создал сеть корреспондентов в каждом регионе, которые периодически сообщают в центр сведения о покупательском спросе населения, товарном обеспечении в данной местности и другую информацию. Преимуществом этого способа являет-

ся его дешевизна, однако он не всегда обеспечивает хорошее качество сведений, так как зависит от уровня восприятия вопросов опрашиваемым, от его ответственности – отправит он заполненную анкету или нет.

При *саморегистрации* или самоисчислении работники той организации, которая проводит опрос, раздают опросные листы или анкеты опрашиваемым лицам, инструктируют их, а затем собирают заполненные формуляры, контролируя полноту и правильность полученных сведений. Этот способ используется в государственной статистике при бюджетных обследованиях семей, проведении некоторых переписей и т. д.

Заметим, что при любом методе проведения статистическое наблюдение пассивно: статистика хочет как можно точнее зарегистрировать данные без какого-либо влияния на наблюдаемый процесс. Принципиально иным способом собирания данных является *эксперимент.* В этом случае статистику принадлежит активная роль: он должен не только наблюдать, а полностью контролировать ситуацию, планировать эксперимент и реализовать свой план. Эксперимент позволяет выявить влияние каких-либо установленных ограничений или нагрузок на поведение людей. Например, влияние на скорость реакций человека пребывания без сна в течение одних, двух, трех суток. Эксперимент традиционно входил в круг методов биологической, медицинской статистики, приложений статистического метода в естественных науках. В настоящее время все большее распространение получают идеи «социального эксперимента».

2.3. ПОДГОТОВКА СТАТИСТИЧЕСКОГО НАБЛЮДЕНИЯ

Чтобы провести статистическое наблюдение, нужно сформулировать его цель и основные гипотезы, которые должны быть проверены по данным наблюдения. Эта стадия работы определяет все последующие, поэтому обычно все решения вырабатываются коллективно в ходе обсуждения проблем предстоящего наблюдения. На этой стадии работы дается определение объекта и единицы наблюдения, разрабатывается и утверждается программа наблюдения.

Определение *объекта наблюдения* включает определение единицы наблюдения, территории и времени наблюдения.

Единица наблюдения – это то явление, признаки которого подлежат регистрации. Совокупность единиц наблюдения составляет *объект наблюдения.* Как уже отмечалось, для определения границ объекта наблюдения нередко устанавливается *ценз* – значение при-

знака (или нескольких признаков), позволяющее отделить единицы наблюдения от других явлений. В самом деле, трудно установить границы даже, казалось бы, очевидного объекта – совокупности промышленных предприятий: что входит в понятие промышленного предприятия, а что – нет. Входят ли в круг промышленных предприятий предприятия по ремонту и мойке автомобилей, закупке и переработке фруктов и т. д.? Устанавливать ли цензовые значения только по численности работников или по стоимости производственного оборудования? При проведении переписи населения возникают вопросы, учитывать ли тех граждан, которые длительное время работают за границей, как учитывать тех, кто находится в заключении, на службе в армии и т. д. Все эти вопросы требуют всестороннего обсуждения. Их решение основано на том, что является конечным результатом, что должно быть получено в результате всего исследования. Если не предусмотреть чего-то на начальной стадии, это скажется на качестве всего исследования.

Территория проведения наблюдения охватывает все места нахождения единиц наблюдения; ее границы зависят от определения единицы наблюдения.

Время наблюдения – это то время, к которому относятся собираемые данные. Время регистрации данных для всех единиц устанавливается единое – для предупреждения неполного учета или повторного счета, а также для обеспечения сопоставимости данных.

При изучении объектов наблюдения, численность и характеристика которых непрерывно изменяются, устанавливается *критическая дата,* по состоянию на которую собираются сведения. При переписях обычно устанавливают время начала и время окончания регистрации данных. Так, последняя Всесоюзная перепись населения проводилась в течение 8 дней – с 12 по 19 января 1989 г.; 5%-ная микроперепись населения РФ проводилась в течение 10 дней – с 14 по 23 февраля 1994 г. И в том, и в другом случае время наблюдения приходилось на зимний период и те даты, когда у работающих меньше отпусков, нет государственных праздников или каникул у школьников и студентов.

При изучении такого подвижного объекта, как население, недостаточно установить время наблюдения – ведь состав населения России и его характеристики постоянно меняются: в среднем в каждую минуту в нашей стране рождается 3 человека и умирает 3–4 человека. Поэтому данные регистрируются по состоянию на определенный момент времени, называемый *критическим моментом наблюдения.* В качестве критического момента в микропереписи населения 1994 г. было принято 0 часов ночи с 13 на 14 февраля.

Соответственно в бланки микропереписи заносились все живущие на данный момент времени и не вносились родившиеся после 0 часов ночи с 13 на 14 февраля и умершие до этого времени.

При переоценке основных фондов устанавливается *критическая дата,* по состоянию на которую учитываются основные фонды (здания, сооружения, оборудование, транспорт и т. д.). Одна из последних переоценок проводилась по состоянию на 1 января 1994 г. Все предприятия, владевшие основными фондами на эту дату, должны были показать сведения о них в своем отчете; если в период между 1 января 1994 г. и моментом заполнения бланка отчетности какие-либо фонды были проданы, переданы другому владельцу, то новый владелец не включал их в свой отчет во избежание двойного учета.

Определение объекта наблюдения, его территориального размещения важно для определения объема работ, который нужно выполнить в период наблюдения. Если наблюдение планируется провести в форме отчетности, то составляется список подотчетных предприятий и организаций. При специально организованном наблюдении определение объема работ необходимо для расчета *численности работников,* требуемых для выполнения обследования в установленные сроки. Рассчитывается дневная норма работы 1 регистратора (счетчика) с учетом сложности программы наблюдения, трудоемкости заполнения формуляра наблюдения и размещения объекта. В сельской местности, например, где плотность застройки намного ниже городской, дневная норма устанавливается меньше, чем в городах. В целях лучшей организации наблюдения и контроля качества материала вся территория разбивается на отдельные счетные участки; 20–30 счетных участков при переписях населения образуют инструкторский участок, руководимый инструктором.

Проведение массовых работ требует участия множества исполнителей (в переписях населения участвуют тысячи счетчиков). Все они должны пройти специальное обучение – *инструктаж* и провести пробное заполнение тех формуляров, которые предполагается использовать в статистическом наблюдении. Должна быть составлена смета на проведение специального обследования, в которой предусматриваются размножение материалов наблюдения (бланков, инструкций), оплата услуг средств связи, транспорта, работы инструкторов, счетчиков и др. Статистическое обследование – дорогостоящая и трудоемкая процедура. Проведение обследований должно быть обосновано и подкреплено финансовыми, материальными и трудовыми ресурсами.

Программа наблюдения включает признаки, подлежащие регистрации по каждой единице наблюдения. Ее содержание зависит от

целей и задач обследования. В какой-то мере программа наблюдения зависит и от выделенных средств: мало средств, программа может быть короче или число наблюдаемых единиц меньше. Поэтому первый принцип составления программы наблюдения: никаких сведений, не относящихся к данному обследованию (на всякий случай). Второй принцип немаловажный для получения достоверных данных при опросах: не включать в программу наблюдения те вопросы, которые могут показаться людям подозрительными и на которые можно заведомо ожидать неточных ответов. Например, при изучении потенциальной эмиграции не стоит включать в анкету прямой вопрос типа: «Собираетесь ли вы уехать за границу на длительное время или навсегда?» Более эффективно использовать систему вопросов, составленных таким образом, чтобы их сочетание позволяло сделать те заключения, которые бы вы хотели получить через ответы на прямой вопрос. Или, понимая, что точную сумму доходов и сбережений состоятельные люди скорее всего не укажут, имеет смысл задать косвенные вопросы, например: «Есть ли среди Ваших знакомых люди с месячным доходом 100 тыс. долл и выше?» и т. д.

Программа наблюдения всегда включает опознавательные признаки: вопросы, прямо связанные с целью исследования; контрольные вопросы. Выделение последних весьма условно, так как один и тот же вопрос может выполнять как содержательную, так и контрольную функцию. Так, программа переписи населения содержит вопросы о возрасте, образовании, семейном положении, наличии детей, их возрасте, образовании и т. д. Все они логически связаны, что позволяет контролировать правильность ответов. То же в бюджетных обследованиях – вопросы о доходах и расходах выполняют и познавательную функцию, и функцию взаимного контроля.

Опознавательные признаки позволяют идентифицировать единицу совокупности, к которой относятся регистрируемые данные. В социологических обследованиях опрос обычно анонимный. Однако чтобы избежать недоучета и повторного счета, каждой единице наблюдения (опрашиваемому) присваивается какой-либо номер (шифр), а также фиксируется место проживания (населенный пункт). При сборе данных в форме отчетности опознавательными признаками являются название предприятия (организации), его шифр в регистре государственной статистики, отраслевая принадлежность, адрес, номер телефона, факса и т. п.

Все вопросы программы наблюдения ориентированы на определенную форму ответа: цифровую, альтернативную («да» или «нет»), многовариантную, когда ответ состоит в выборе одного или

нескольких вариантов из множества предлагаемых вариантов ответа. Так, на вопрос о возрасте ответ дается в количественной форме – указывается число исполнившихся лет, то же на вопрос о стаже работы, на вопрос о наличии автомобиля или дачи ответ будет в альтернативной форме – «да» или «нет», на вопрос о степени удовлетворенности работой или учебой ответ выбирается из предлагаемого меню. Обычно такое меню строится по принципу симметрии: абсолютно негативное (или, наоборот, абсолютно позитивное) отношение, затем – более мягкая оценка, затем – выражение полной индифферентности, после чего оценки переходят в противоположную область: если были негативные, то теперь – позитивные и наоборот.

Предлагаемые варианты ответов называются *подсказом*. Наличие подсказа обеспечивает единообразное понимание вопросов программы и облегчает последующую обработку данных, так как каждый предлагаемый вариант ответа имеет свой код или шифр и работа по подготовке данных к обработке ведется лишь по тем вариантам ответов, которые не были предусмотрены в подсказе и вписывались самими опрашиваемыми *(респондентами)*.

Приведем в качестве примера фрагмент из анкеты читателей молодежной газеты «Смена».

Как к вам попал этот номер «Смены»?

001 – подписчиком газеты являюсь лично я;
002 – взял у знакомых;
003 – купил в газетном киоске;
004 – газету выписывают у меня дома;
005 – другой ответ.

Наличие кодов облегчает обработку собранного материала, которая начинается сразу же, как только статистик убедился в том, что получены данные от всех единиц и что даны ответы на все вопросы.

В переписях населения и других специальных обследованиях, проводимых государственной статистикой, подсказы обычно включают все варианты ответов (без дописывания). Например, вопрос о типе жилого помещения в программе микропереписи 1994 г. включал варианты ответов: индивидуальный дом, отдельная квартира, общая (коммунальная) квартира, общежитие, другое жилое помещение, снимает жилое помещение.

Составление программы наблюдения – сложная и ответственная задача. В государственной статистике разработкой программы специальных обследований занимаются специалисты Госкомстата РФ и НИИ при участии представителей Научно-методологического совета и заинтересованных организаций. Программы таких важных и

массовых работ, как перепись населения, переоценка основных фондов и другие, обсуждаются на специальных совещаниях, в печати, что обеспечивает их высокое качество.

Инструментарий статистического наблюдения включает формуляры и инструкции по их заполнению. Формуляры наблюдения – это бланки, опросные листы, анкеты и т. д., на которых напечатаны вопросы программы наблюдения и куда затем заносятся собираемые сведения. Соответственно в формуляре должно быть предусмотрено место для вопроса и ответа. Обычно в верхней части формуляра или на первой странице располагаются опознавательные признаки, слева – вопросы программы наблюдения, справа – место для ответов. Формуляр наблюдения может быть *карточным* (индивидуальным) или *списочным.* В первом случае он предназначен для записи данных только по одной единице наблюдения, во втором – по нескольким. В микропереписи населения РФ 1994 г. была принята списочная форма – формуляр заполнялся на домохозяйство. При этом если число членов домохозяйства превышало 5 человек, то использовался дополнительный бланк, а в опознавательной части проставлялись буквенные обозначения бланка (а, б и т. д.).

Качество данных статистического наблюдения зависит не только от перечисленных факторов, но и от подготовленности счетчиков (регистраторов, интервьюеров). Для них организуется инструктаж по разъяснению вопросов анкеты (или другого формуляра наблюдения) и пользованию инструкцией. Объясняется, например, что при наличии подсказов счетчик обязан ознакомить респондента со всеми вариантами ответов, не выделяя те из них, каторые он сам считает наиболее вероятными. Затем проводится пробное заполнение анкет, итоги которых коллективно обсуждаются.

Доброжелательность счетчика, его умение вступать в контакт с людьми влияют на атмосферу опроса, а значит, и на его результаты. Важной этической проблемой является анонимность данных опроса. Уверенность в анонимности снимает напряженность при регистрации мнений, суждений, пожеланий, а также характеристик благосостояния (чем владеет респондент, имеет ли сбережения, что из «крупных» вещей приобрел за последний год и т.д.). Иногда в интересах планирования наблюдения, контроля данных полной анонимности респондентов нет – можно говорить лишь о доверительности. Так, если для проведения опроса с целью изучения уровня бедности в России в качестве основы выборки использовались списки избирателей, то соответствующий код респондента позволяет идентифицировать его. В таких случаях респондент должен быть убежден, что его ответы как персональные никогда не будут использованы. Они войдут в общую совокупность ответов и послужат основой расчета обобщающих показателей.

Как бы тщательно ни была составлена программа и разработан формуляр, для обеспечения единообразия его заполнения, толкования вопросов программы наблюдения все же необходима инструкция. Этот документ содержит объяснения вопросов программы с конкретными примерами, указания по взаимосвязи вопросов. Инструкция издается либо в виде отдельной брошюры, либо дается в подсказах, либо на самом формуляре наблюдения (обычно на оборотной стороне).

2.4. СТАТИСТИЧЕСКАЯ ОТЧЕТНОСТЬ

Статистическая отчетность – особая форма организации сбора данных, присущая только государственной статистике. Она проводится в соответствии с федеральной программой статистических работ. Государственная статистика включает все виды статистических наблюдений (регулярные и периодические отчеты, единовременные учеты, различного рода переписи, выборочные, анкетные, социологические, монографические обследования и т. д.), формы и программы которых утверждены Государственным комитетом РФ по статистике или по согласованию с ним органами государственной статистики республик в составе Российской Федерации, краев, областей, автономной области и автономных округов, городов Москвы и Санкт-Петербурга. Сведения о деятельности предприятий, организаций поступают в статистические органы в установленные сроки в виде определенных документов (отчетов). Бланки таких отчетов называют *формами статистической отчетности.* Каждая из них имеет свой шифр и название. Например, форма № 5-нт (образцы) «Отчет о созданных впервые в России образцах нового типа машин, оборудования, аппаратов, приборов в 199... году» или форма № 2 – инновация «Сведения о технологических инновациях промышленного предприятия (объединения) за 199... год» и т. д. Программа отчетности, т.е. перечень собираемых сведений, методика их определения и форма бланка отчетности, разрабатывается и утверждается Госкомстатом РФ. Формы отчетности, включающие финансовые результаты, утверждаются, кроме того, и Минфином РФ.

Отчетность различается по периодичности: срочная – содержит данные за месяц и менее; квартальная; полугодовая; годовая. Наиболее подробной является программа годовой отчетности. Срочная отчетность часто передается по телеграфу (или электронной почте)

и называется телеграфной; отчетность за более длительное время является почтовой.

Статистическое наблюдение в форме отчетности использует только один источник данных – документы. Прежде всего это документы бухгалтерского учета предприятий, организаций. Например, форма отчетности 10-ф (квартальная) практически полностью основана на данных бухгалтерского баланса и приложения к нему «Отчет о прибылях и убытках». Госкомстат России проводит политику **унификации** отчетности предприятий разных отраслей экономики.

Предприятия и организации любых форм собственности обязаны представлять отчетность в установленные сроки по утвержденной форме. С мая 1992 г. введен в действие Закон РФ «Об ответственности за нарушение порядка представления государственной статистической отчетности», который определяет правовую ответственность предприятий, учреждений, организаций и объединений за нарушение порядка представления государственной статистической отчетности и других данных, необходимых для проведения государственных статистических наблюдений. Закон предоставляет органам государственной статистики РФ право применять к руководителям и другим должностным лицам предприятий, учреждений, организаций и объединений независимо от их форм собственности административные взыскания в виде предупреждения или штрафа в размере от трехкратного до восьмикратного установленного законом размера минимальной месячной оплаты труда за нарушение порядка представления государственной статистической отчетности, выразившееся в непредставлении отчетов и других данных, необходимых для проведения государственных статистических наблюдений, искажении отчетных данных или нарушении сроков представления отчетов, а за те же действия, совершенные повторно в течение года после наложения административного взыскания, – в виде штрафа в размере от восьмикратного до десятикратного установленного законом размера минимальной месячной оплаты труда.

Нарушением сроков представления государственной статистической отчетности является опоздание на одни сутки, а опоздание более одних суток рассматривается как ее непредставление. Искажением отчетных данных считается неправильное их отражение в государственной статистической отчетности, допущенное как в результате умышленных действий должностных лиц с целью сокрытия доходов и в других корыстных целях, так и вследствие нарушения действующих инструкций и методологических указаний по составлению статистической отчетности, а также арифметических ошибок.

Важной функцией государственной статистики является определение круга подотчетных единиц. С этой целью все предприятия, организации, объединения независимо от формы собственности, а также граждане, занимающиеся предпринимательской деятельностью, представляют в органы государственной статистики учредительные документы для присвоения идентификационных кодов, определения классификационных признаков на основании общероссийских классификаторов технико-экономической информации для включения в единый государственный регистр предприятий и организаций (ЕГРПО) и отражения в государственной статистической отчетности.

При реорганизации или ликвидации предприятия, учреждения, организации, объединения представляют органам статистики государственную статистическую отчетность за период своей деятельности в отчетном году до момента ликвидации на бланках форм годовой отчетности, а также нормативные акты о своей реорганизации или ликвидации для внесения изменений в ЕГРПО.

Отчетность дает необходимую информацию для государственных органов управления. Данные отчетности позволяют следить за динамикой объема промышленного производства и продукции других отраслей народного хозяйства, оценивать комплексность развития страны и регионов, изучать соотношение разных форм собственности по отраслям и регионам и сравнивать эффективность деятельности государственных и негосударственных предприятий и организаций.

Большое значение имеют стабильность отчетности, содержания ее форм. Только при этом условии обеспечивается возможность построения протяженных рядов динамики, а значит, выявления тенденций, анализа колеблемости, разработки прогнозов. Конечно, содержание отчетности – перечень форм, показателей – меняется со временем, но прежде чем внести какое-либо изменение, нужно решить, является ли оно действительно необходимым. Ведь отчетность подготавливают десятки тысяч работников бухгалтерий и финансовых отделов предприятий и организаций. Очевидно, что такая массовая форма сбора данных может давать надежные данные, если она достаточно стабильна.

Формирование содержательной части форм отчетности осуществляется с учетом требований Государственной программы перехода Российской Федерации на принятую в международной практике систему учета и статистики, а также систему статистических показателей для оценки экономических преобразований в стране. Развитие малого бизнеса потребовало специального решения воп-

роса об отчетности малых предприятий. Они отчитываются по минимальному количеству форм статотчетности и показателей. Утверждена унифицированная интегрированная форма статистической отчетности квартальной периодичности для малых предприятий всех организационно-правовых форм и форм собственности и годовая форма, содержащая показатели производственной деятельности этих предприятий.

Данные статистической отчетности поступают от предприятий и организаций в органы государственной статистики – либо в райинспектуры, либо прямо в областные (краевые) управления статистики. После проверки поступивших данных они разрабатываются: составляются сводные таблицы по формам, установленным Госкомстатом РФ. Данные обобщаются по отраслям, организационно-правовым формам, формам собственности, территориям и т. д. Сводные таблицы из местных статистических органов отправляются в Госкомстат РФ, где составляются сводные таблицы по стране в целом, рассчитываются сводные показатели с учетом тех же группировок данных (по отраслям, территориям, формам собственности и т. д.). Местные органы государственной статистики, как правило, проводят дополнительные разработки данных отчетности по заказам местной администрации и научных организаций.

2.5. ОШИБКИ СТАТИСТИЧЕСКОГО НАБЛЮДЕНИЯ. МЕТОДЫ КОНТРОЛЯ ДАННЫХ НАБЛЮДЕНИЯ

Как бы тщательно ни был составлен инструментарий наблюдения, проведен инструктаж исполнителей, всегда материалы наблюдения нуждаются в контроле. Это объясняется массовым характером статистических работ и сложностью их содержания.

Прежде всего проверяется *полнота охвата единиц наблюдением.* С этой целью производится сверка данных по спискам предприятий и организаций, ЕГРПО; пересчитываются заполненные анкеты. При проведении массовых социологических обследований часто кроме основного списка составляется дополнительный список респондентов на тот случай, если респонденты из основного списка почему-либо не могли быть опрошены. Дополнительный список формируется так, чтобы при замене респондентов общая структура опрашиваемых сохранялась. Поэтому при проверке устанавливается соответствие фактически опрошенных основному и дополнительному спискам. Проверка полноты охвата единиц не означает применения только сплошного наблюдения. Она связана с выявлением

недоучета или повторного счета и обеспечения проектируемых пропорций собранных данных.

Одновременно на этой стадии проверяется *полнота заполнения каждого формуляра наблюдения* – формы отчетности, анкеты и т. д.

После такой общей проверки проводится детальная проверка каждого полностью заполненного формуляра.

Чтобы хорошо организовать проверку, нужно представлять характер возможных ошибок. Все ошибки наблюдения можно назвать *ошибками регистрации.* Но они имеют разный характер и по-разному сказываются на результатах статистического исследования. Ошибки могут быть *случайными и систематическими.* Те и другие чаще всего возникают при опросе, но могут быть допущены и при непосредственном или документальном наблюдении.

Во всех случаях источниками ошибок может быть как информация, поступившая от объекта наблюдения (ошибки в ответах опрашиваемого, искажения в показаниях приборов, регистрирующих какие-либо свойства объекта, ошибки в учетных документах), так и ошибки регистратора или экономиста предприятия, представляющего данные (неправильная запись ответов опрашиваемого, ошибки при переносе на формуляры наблюдения показателей приборов, данных учетных документов).

Случайные ошибки не имеют какой-либо направленности. Это описки, оговорки, перестановки цифр при записи цифровых данных и т.д. При обобщении массового материала они взаимопогашаются и не могут исказить значения сводных показателей и результаты анализа.

Другое дело *систематические ошибки* – они являются неслучайными и имеют определенную направленность. Такие ошибки очень опасны, так как приводят к искажению результатов статистического исследования. Эти ошибки, как правило, являются преднамеренными. Известно, например, что люди предпочитают преуменьшать свои доходы, округлять возраст, стараются показать большую осведомленность в области культуры, науки, чем это есть на самом деле. Предприятия также могут внести элементы недостоверности в свою информацию, особенно в те характеристики, от которых зависят величина налоговых платежей, расчеты с кредиторами и т. п. Все ошибки такого рода необходимо выявить и исправить. Поэтому после проверки полноты данных проводится их контроль – счетный и логический.

Счетный контроль основан на жесткой связи между признаками, которая может быть проверена арифметическими действиями: сложением, вычитанием, умножением, делением. Связь такого рода часто отражается в заголовках граф отчетности и в подсказах тако-

го рода: графа X равна графе Y плюс графа Z или графа X равна графе Y, деленной на графу Z, и т. д. Счетный контроль используется для проверки итоговых сумм. Если представленное число слагаемых не является полным, то сумма слагаемых должна быть меньше либо равна общему итогу, но не может превышать его.

Счетный контроль совершенно определенно устанавливает наличие ошибки, тогда как логический контроль может лишь поставить под сомнение правильность данных. *Логический контроль* основан на логической взаимосвязи между признаками. Классическим примером является пример взаимосвязи данных при переписи населения: вопросы о возрасте, образовании, семейном положении взаимоконтролируются. Если, например, окажется, что гражданин десяти лет женат или пятилетний имеет общее среднее образование, то ясно, что при заполнении формуляра допущены ошибки либо при записи возраста, либо другой характеристики.

Логический контроль основан и на сравнении с данными прошлого периода. Например, достоверность данных о выпуске продукции по видам может быть проведена сравнением с данными прошлого периода для того же предприятия. Кроме того, логический контроль опирается на представления о пределах возможных значений признака: минимуме и максимуме. Скажем, при проверке отчетности по форме 10-ф можно прикинуть, каким будет срок погашения дебиторской задолженности.

$$\text{Срок погашения дебиторской задолженности} = \frac{\text{Количество дней в отчетном периоде}}{\text{Оборачиваемость дебиторской задолженности}},$$

$$\text{Оборачиваемость дебиторской задолженности} = \frac{\text{Чистая выручка от реализации}}{\text{Средняя сумма дебиторской задолженности}}$$

Величина оборачиваемости дебиторской задолженности выражается в разах. Маловероятно, чтобы этот показатель был меньше 5 или больше 12 за год. При проверке срока погашения дебиторской задолженности мы можем использовать и нормативное значение этой величины (обычно 30 дней). Если реальный срок погашения намного (на несколько недель) отличается от нормативного в ту или иную сторону, необходимо поставить под сомнение резко отличающиеся данные и сделать запрос на предприятие.

Обычно для проверки поступающего материала наблюдения составляется *схема контроля,* в которую включаются все увязки между вопросами программы наблюдения: как арифметические, так и логические.

Никогда не следует произвольно вносить исправления в формуляр. Необходимо либо самому статистику провести повторное наблюдение (повторный опрос и т. д.), либо обратиться к лицам, отвечающим за представленную информацию (директору, главному бухгалтеру предприятия).

Данные наблюдения считаются принятыми, если они прошли контроль и если потребовалось, в них внесены исправления. Проверкой собранных данных завершается начальный этап статистического исследования.

Рекомендуемая литература к главе 2

1. *Воронов Ю.П.* Методы сбора информации в социологическом исследовании. – М.: Статистика, 1974.
2. *Деев Г., Крутова Т.* Метод основного массива в статистических наблюдениях // Вестник статистики. – 1992. – № 5. – С. 39–43.
3. *Деев Г., Мухин П.* Несплошное статистическое наблюдение: исторический опыт, практика, перспективы // Вопросы статистики.– 1996. – № 3. – С. 21–27.
4. *Елисеева И.И.* Моя профессия – статистик. – М.: Финансы и статистика, 1992.
5. *Моргенштерн О.* О точности экономико-статистических наблюдений. – М.: Статистика, 1968.
6. Об ответственности за нарушение порядка представления государственной статистической отчетности // Вестник статистики. – 1992. – № 1. – С. 3–7.
7. *Рябушкин Т.В., Симчера В.М.* Очерки международной статистики. – М.: Наука, 1981.

3 Глава СТАТИСТИЧЕСКИЕ ПОКАЗАТЕЛИ

3.1. СУЩНОСТЬ И ЗНАЧЕНИЕ СТАТИСТИЧЕСКИХ ПОКАЗАТЕЛЕЙ. ПОКАЗАТЕЛЬ И ЕГО АТРИБУТЫ

В первой главе сказано, что статистика выражает массы явлений, процессы количественно в числовой форме. Но «числа», применяемые в статистике, это не абстрактные числа математики, характеризуемые только величиной, знаком, формой (целые – дробные; мнимые – действительные; рациональные – иррациональные и т. п.). Статистика применяет, собственно говоря, не числа, а *показатели*, точнее – *статистические показатели*.

Что же такое статистический показатель; каково его содержание и построение; какие виды показателей применяются в статистике; какое значение имеют статистические показатели в познании массовых явлений и процессов, в управлении производством, в жизни общества в целом? Ответам на эти вопросы посвящена данная глава.

Не умея правильно понять содержание, форму, свойства того или иного статистического показателя, нельзя корректно применить его в анализе социально-экономических явлений и процессов и понять смысл статистической информации и жизни страны и мира.

С философской точки зрения статистический показатель – это *мера*, т. е. единство качественного и количественного отражения свойств объективных явлений и процессов в научном сознании. Поскольку статистика, как уже сказано, изучает массовые явления, *статистический показатель* – это *обобщающая характеристика* какого-то свойства совокупности, группы. Этим он отличается от индивидуальных значений, которые, как отмечалось, называются *признаками*. Например, средняя продолжительность ожидаемой жизни родившегося поколения людей в стране – статистический показатель. Продолжительность жизни конкретного человека – признак.

Рассмотрим содержание и форму статистического показателя на примере ввода в действие жилых домов в Российской Федерации в 1993 г., составившего 41880,2 тыс. м^2. Показателем является не одно только число 41880,2, а весь текст, поясняющий его содержание. Качественная сторона этого показателя – ввод в действие жилых домов. Далее статистический показатель имеет количественную сторону, которая выражается числом и единицей измерения: 41880,2 тыс. м^2.

Не всегда статистический показатель является именованным числом. Он может быть абстрактным и отвлеченным числом без наименования, может быть выражен в долях единицы: в процентах, промилле и т. п. Именованными числами являются абсолютные статистические показатели.

Статистический показатель имеет указание на территориальные границы объекта (жилье на определенной территории – Российской Федерации) и границы во времени – 1993 год. Без указания территориальных, отраслевых или ведомственных границ объекта и без привязки к определенному интервалу времени или моменту статистический показатель не существует. Итак, структура статистического показателя представлена на схеме 3.1.

Качественная сторона: объект, его свойство	Количественная сторона: число и единицы измерения	Территориальные, отраслевые и иные границы объекта	Интервал или момент времени

Схема 3.1. Атрибуты статистического показателя

Являясь отображением каких-либо свойств изучаемых явлений и процессов, статистический показатель служит орудием их познания. Но всякое знание всегда ограничено, неполно соответствует изучаемому объекту. Ни один статистический показатель, ни целая их *система* не могут отразить все свойства, все особенности объекта и даже часть этих свойств с абсолютной точностью. Статистический показатель – приближенное, неточное и неполное отображение свойств изучаемого объекта, доступное при имеющемся уровне знаний и возможностях учета, измерения, сбора и передачи информации. Каждому ясно, что невозможно точно измерить вес собранного картофеля без примеси песка, глины, частиц почв и камней, невозможно в масштабах целой республики избежать ошибок во взвешивании, записи, передаче сведений об урожае. Это один из

возможных примеров. Известно, что бывают сознательные искажения данных – приписки. Если же речь идет о сложных свойствах жизни общества, как-то: уровень материального благосостояния, эффективность производственного процесса, культурный уровень населения, то главной причиной неточности, неполноты отображения этих сторон общественной жизни статистическими показателями является недостаточное развитие тех наук, которые формируют указанные категории, и трудности перехода от их качественного описания к количественному измерению.

Статистические показатели поэтому не есть нечто раз навсегда застывшее. Одни развиваются, улучшаются, от иных отказываются за ненадобностью, создаются новые.

Признак и показатель

Остановимся на соотношении между признаком и статистическим показателем.

Признак – это свойство, присущее единице совокупности. Признак входит в качественное содержание показателя, он существует объективно независимо от того, отражает ли его наука с помощью тех или иных показателей. Например, возраст человека – это его признак, который можно измерять с разной степенью точности – в годах, месяцах, в сутках или охарактеризовать датой рождения.

Показатель – характеристика группы единиц или совокупности в целом. Его построение зависит от цели исследования и изобретательности статистика. Средний возраст работников фирмы или жителей города – это статистические показатели, дающие возрастную характеристику определенных групп, совокупностей людей. Другим видом возрастных показателей может служить ряд распределения людей по возрасту и вычисленные на основе этого ряда системы показателей для характеристики структуры такого ряда и размеров вариации (см. гл. 5).

3.2. КЛАССИФИКАЦИЯ СТАТИСТИЧЕСКИХ ПОКАЗАТЕЛЕЙ

Объектами статистического исследования могут быть самые разнообразные явления и процессы. Поэтому чрезвычайно велико и разнообразие статистических показателей. В данном параграфе рассматривается только наиболее общая классификация статистических показателей. Конкретные их виды и формы представлены в

последующих главах учебника, в курсах математической, социально-экономической и отраслевых статистических дисциплин. На схеме 3.2 представлена классификация видов статистических показателей.

По качественной стороне показателей	По количественной стороне показателей	По отношению к характеризуемому свойству
1. Показатели свойств конкретных объектов	1. Абсолютные	1. Прямые
2. Показатели статистических свойств любых массовых явлений и процессов	2. Относительные	2. Обратные

Схема 3.2. Классификация видов статистических показателей

В последующих главах рассматриваются и другие классификации. Например, в главе об изучении динамики и в главе об изучении структуры излагаются различия между статическими показателями, отражающими состояние на данный момент, и динамическими показателями, отражающими изменение во времени.

Показатели конкретных свойств изучаемого объекта – это, например, уже упомянутый средний возраст работников предприятия, объем реализованной продукции предприятия, валовой внутренний продукт государства, средний надой молока на корову на ферме, объем перевозок груза автопарком, показатели рождаемости, смертности, обеспеченности населения товарами и услугами, национальное богатство, средний душевой доход жителя страны и т. д. Особенностью этих показателей является то, что они формируются не только статистикой. В построении этих показателей качественное их содержание определяется конкретной предметной наукой: показатель рождаемости – демографией, показатель внутреннего валового продукта – теорией экономики, показатели урожайности, продуктивности скота – соответствующими сельскохозяйственными науками. Статистика отвечает за методику учета или расчета количественной стороны этих показателей и их форму.

Совершенно иначе обстоит дело с *показателями статистических свойств* любых массовых явлений и процессов, не зависящих от конкретного содержания этих явлений. К таким статистическим показателям относятся: средние величины, показатели вариации, пока-

затели связи признаков, показатели структуры и характера распределения, показатели скорости и темпов изменения, показатели колеблемости в динамике. К ним же относятся статистические оценки степени точности и надежности любых конкретных статистических показателей, полученных при выборочном изучении совокупности, а также оценки надежности и точности статистических прогнозов. За качественную, как и количественную сторону этих показателей, за их построение, интерпретацию и применение отвечает не какая-либо иная научная дисциплина, а только сама статистика. Это, можно сказать, ее кровные дети! Система таких показателей создается и совершенствуется в ходе развития методов статистики, поэтому в последующих главах будут рассмотрены построение, свойства и применение именно таких статистических показателей.

Теоретическая статистика разрабатывает и изучает содержание, форму, методы расчета этих показателей в общем виде: что такое средняя арифметическая величина, коэффициент вариации, уравнение тренда ряда динамики. Если же любой из этих показателей рассчитан для определенного объекта, признака, периода времени, то он становится уже конкретным показателем, например в главе 9 «Статистическое изучение динамики» показатели сезонных колебаний импорта КНР за 1992–1995 гг. – это уже конкретные статистические показатели экономики Китая.

Статистические показатели подразделяются на абсолютные и относительные.

Абсолютным показателем является такой, который отражает либо суммарное число единиц, либо суммарное свойство объекта. Например, число крестьянских хозяйств в Ленинградской области на 1 января 1997 г., посевная площадь картофеля в районе, сумма средств, направленных на потребление за конкретный месяц или год, и т. п.

Абсолютные показатели, как правило, выражаются именованными величинами в натуральных единицах измерения: тоннах, штуках, часах, амперах и т. п., в условных единицах: условном топливе, нормо-сменах, километрах пряжи и т. д. или в стоимостных единицах: рублях, долларах, марках. Они характеризуют сумму значений первичных признаков объекта. Совершенно понятно, что наука не может ограничиваться характеристиками только изолированных отдельных свойств объекта. Поэтому статистика не ограничивается абсолютными показателями. Она измеряет и характеризует соотношение разных абсолютных величин, их изменения во времени, их взаимосвязи между собой и окружающей средой. Статистика, как и все науки, широко пользуется общенаучными методами сравнения, обобщения, синтеза.

Относительным показателем является показатель, полученный путем сравнения, сопоставления абсолютных или относительных показателей в пространстве (между объектами), во времени (по одному и тому же объекту) или сравнения показателей разных свойств изучаемого объекта.

Относительные статистические показатели, получаемые при сопоставлении абсолютных показателей, могут быть названы относительными показателями *первого порядка*, а полученные при сопоставлении относительных же показателей – *показателями высших* (второго, третьего и т. д.) *порядков*. Показатели выше четвертого порядка ввиду сложности интерпретации почти никогда не применяются. Относительные статистические показатели выражают связь между абсолютными показателями: урожайность картофеля – отношение валового сбора к посевной площади; доля городского населения в стране – отношение численности населения городов к общему числу жителей страны.

Основные виды относительных величин чаще выражаются отвлеченными числами, но могут быть также именованными относительными показателями. Построение их связано с применением различных методов статистики.

Относительные показатели можно подразделить на следующие группы:

1. *Относительные показатели, характеризующие структуру объекта.* Это доля (удельный вес) – отношение части к целому. Например, отношение площади каждой из сельхозкультур к общей посевной площади; числа женщин к общей численности населения города, республики. В эту же группу входят характеристики отношения между отдельными частями объекта; показатели, характеризующие *степень сложности структуры, степень неравномерности (вариации) долей* и др. Доли выражаются нередко в процентах или промилле (тысячных долях).

2. *Относительные показатели, характеризующие динамику процесса, изменение во времени.* Это отношения показателей, характеризующих объект в более позднее время (текущий период), к аналогичным показателям того же объекта в более ранний (базисный) период. Такие показатели называют *темпами роста*. Темп роста может быть выражен в разах или в процентах. Темп роста говорит о том, во сколько раз больше показатель текущего периода в сравнении с базисным или сколько процентов он составляет по отношению к показателю базисного периода. К относительным показателям динамики принадлежат также темпы прироста, параметры уравнений трендов, коэффициенты колеблемости и устойчивости в

динамике, индексные показатели динамики. Подробнее о них сказано в главе 9.

3. *Относительные показатели, характеризующие взаимосвязь признаков в совокупности явлений, а также взаимосвязь результативных признаков-следствий с факторными признаками-причинами,* например, связь уровня душевого дохода с размером потребления мяса или фруктов на одного человека; связь дозы удобрений с урожайностью картофеля и т.п. К таким показателям относятся рассматриваемые в главе 8 коэффициенты корреляции, эластичности, детерминации, а также в главе 10 аналитические индексы. Относительные показатели взаимосвязи могут быть как отвлеченными, так и именованными числами.

4. *Относительные показатели, характеризующие соотношение разных признаков того же объекта между собой* (иногда их называют показателями интенсивности). Эти показатели обобщают вторичные признаки объектов (например, производительность труда – отношение произведенной продукции в натуральном или стоимостном выражении к затратам труда на ее производство и др.). Показатели соотношения признаков могут быть прямыми и обратными. Например, отношение затрат труда на производство к объему продукции дает показатель трудоемкости продукции — величину, обратную прямому показателю производительности труда. И прямые, и обратные показатели выражаются именованными числами с двойными единицами измерения обоих сравниваемых признаков: в рублях за 1 час труда, в центнерах с 1 га площади. Например, продукция предприятия учитывается в миллионах рублей за год, скажем, 1800, и стоимость основных производственных фондов предприятия тоже учитывается в миллионах рублей, т. е. 4000. Если формально единицы измерения сравниваемых признаков совпадают, то неверно фондоотдачу – показатель сравнения стоимости продукции за год со стоимостью среднегодовых производственных фондов – называть отвлеченным числом (в нашем примере – 0,45). Правильно будет сказать «фондоотдача составила 45 коп. продукции на 1 рубль основных фондов за год». Стоимость продукции и стоимость фондов – разные признаки, хотя имеют одинаковую единицу измерения.

В экономике относительные показатели, характеризующие величину признака объекта, рассчитанные на единицу другого признака, используются для измерения эффективности либо интенсивности производства.

К данному классу показателей принадлежат и показатели, характеризующие степень системности признаков, например соотношение между суммой осадков и суммой эффективных температур (способствующих произрастанию сельскохозяйственных культур), так называемый гидротермический коэффициент; таково же соотношение между весом и ростом человека, характеризующее пропорциональность его тела.

5. Особым видом относительных статистических показателей являются *отношения фактически наблюдаемых величин признака к его нормативным, плановым, оптимальным или максимально возможным величинам.* Это широко распространенные на производстве показатели выполнения норм выработки, норм расхода материалов и других ресурсов. Отношения наблюдаемых величин признака к оптимальным или плановым характеризуют приближение изучаемого процесса к идеалу. Так, если оптимальная норма потребления мяса взрослым мужчиной на Северо-Западе России составляет 80 кг в год, а фактическое среднедушевое потребление составило в 1992 г. 58 кг, то ясно, что размер и структура потребления далеки от оптимальной: всего 72%. Всякое превышение или недобор до оптимальной величины, всякое отклонение от 100% такого относительного показателя (в любую сторону) означают нарушение оптимальности процесса, даже перевыполнение плана, если план не лозунг, а научно обоснованная, взаимосвязанная система объемов производства отдельных видов продукции. В этом случае превышение планового выпуска одного вида продукции, например выплавки стали без согласованного изменения производства станков, прокатных станов, других средств обработки металла, есть попросту омертвление затрат и бесполезный перерасход природных ресурсов, труда.

Отношение фактических значений признака к максимально возможным значениям часто характеризует качество процесса, агрегата, машины. Таковы, например, коэффициенты полезного действия двигателей, электромоторов. Отношения фактических показателей вариации к максимально возможным при данной численности совокупности используются при анализе вариации (см. гл. 5), при измерении степени специализации предприятия или региона на производстве определенной продукции и в ряде других задач.

Само задание в той или иной отрасли экономики может быть выражено относительной величиной динамики или структуры. Например, «снизить затраты топлива на 1 кВт·ч электроэнергии на 5% в сравнении с прошлым годом»; «увеличить долю продукции высшего качества до 85% общего выпуска». Показатели выполнения такого задания будут являться относительными показателями второго порядка.

6. Еще один вид относительных статистических показателей возникает в результате *сравнения разных объектов по одинаковым признакам.* Сравнение урожайности одной и той же культуры в том же году между хозяйствами, областями; сравнение показателей производства или уровня жизни населения в разных странах – это обычные приемы познания. При построении таких относительных показателей необходимо позаботиться, чтобы сравниваемые показатели определялись по единой методике построения, были сравнимы по единицам измерения и во всех других отношениях. В социально-экономической статистике есть специальный раздел о международных сравнениях показателей.

В качестве примера приведем сравнение производства валового внутреннего продукта на душу населения в Великобритании и в США в 1990 г.: в Великобритании на 1 жителя было произведено 12 715 долл., в США – 18 347 долл./чел. Показатель сравнения может быть построен как отношение одного душевого уровня к другому: душевое производство ВВП в Великобритании составило 69,3% душевого производства ВВП в США. Или душевое производство ВВП в США составило 18 347: 12 715 = 1,443, или 144,3% душевого производства в Великобритании. Если речь идет об исследовании по экономике Великобритании, то предпочтительнее первая форма показателя: изучаемый объект (сравниваемая величина) – числитель, а другой объект (база сравнения) – знаменатель относительного показателя. Если изучается экономика США, предпочтительнее взять в числителе показатель США.

3.3. ОБЩИЕ ПРИНЦИПЫ ПОСТРОЕНИЯ ОТНОСИТЕЛЬНЫХ СТАТИСТИЧЕСКИХ ПОКАЗАТЕЛЕЙ

Построение относительных показателей – задача, требующая сочетания конкретного знания свойств объекта и общих закономерностей статистической методологии. Остановимся на общих логико-статистических принципах построения относительных показателей.

Первый принцип. Сравниваемые в относительном показателе абсолютные (или, в свою очередь, относительные) показатели должны быть чем-то связаны в реальной жизни объективно, независимо от нашего желания. Если этого условия нет, получится согласно русской поговорке «В огороде – бузина, а в Киеве – дядька». Связать этого «дядьку» с «бузиной» чисто математически, разделив одно число на другое, можно, но никакого относительного показателя мы не построим.

Необходимо добиваться как можно большего соответствия по смыслу сравниваемых показателей. Например, мы хотим построить относительный показатель, характеризующий степень грамотности населения. Можно разделить число грамотных на общую численность населения, но это не лучший из показателей. Ведь ясно, что дети до 6 лет, некоторые категории инвалидов с детства, душевнобольных не могут наравне со здоровыми и достигшими школьного возраста людьми быть обучены грамоте. Из всего населения эти категории лиц правильнее исключить при построении относительного показателя грамотности.

Другой пример. Продуктивность молочного скота определяется делением полученного валового надоя молока на маточное поголовье (коров, коз, овец); продуктивность в производстве яиц — делением валового сбора на поголовье кур-несушек (или уток, гусынь), не включая, разумеется, самцов-петухов, селезней, гусаков. Но если продуктивность в производстве шерсти мы станем определять путем отношения валового настрига шерсти на поголовье только овцематок и козоматок, то сделаем ошибку: ведь шерсть стригут и с баранов!

Второй принцип. При построении относительного статистического показателя сравниваемые исходные показатели могут различаться только одним атрибутом: или видом признака (при одинаковом объекте, периоде времени, плановом или фактическом характере показателей), или временем (при том же признаке, объекте и т. п.), или только фактическим, плановым или нормативным характером показателей (тот же объект, признак, время) и т. д. Нельзя сопоставлять показатели, различные по двум и более атрибутам, скажем, сравнивать добычу угля в США в 1980 г. с выплавкой стали в Российской Федерации в 1992 г.

Третий принцип. Необходимо знать возможные границы существования относительного показателя. Например, как будет показано в главе о вариации, относительные показатели вариации теряют смысл и не могут применяться в тех случаях, когда их знаменатели – средние значения признаков близки к нулю, потому что при стремлении знаменателя к нулю относительный показатель стремится к абсурдному бесконечному значению. Аналогично если исходные показатели в текущем и базисном периодах имеют разные знаки, то теряет смысл и не может применяться относительная величина динамики – темп роста. Если предприятие имело в 1992 г. убыток 150 млн/руб., а в 1993 г. получило прибыль 300 млн/ руб., неверно ни то, что «финансовый результат вырос вдвое» (если отбросить знаки), ни то, что он «вырос в минус 2 раза», если делить +300 млн на – 150 млн.

Относительные показатели, измеряющие степень приближения некоторого признака к предельному значению, должны строиться так, чтобы в пределе увеличения они стремились к единице, а в другом пределе своего уменьшения – к нулю. Так строятся коэффициенты, измеряющие тесноту связи признаков, степень эффективности использования ресурсов, скажем, коэффициент полезного действия двигателя. Для многих характеристик экономической, тем более социальной и экологической, эффективности производственных процессов такие относительные показатели эффективности еще предстоит построить.

Ввиду большого значения анализа структурных сдвигов в экономике в наше время относительные и абсолютные характеристики структуры и ее изменений подробно рассматриваются в специальной главе 11.

3.4. ПОНЯТИЕ О СИСТЕМАХ СТАТИСТИЧЕСКИХ ПОКАЗАТЕЛЕЙ

Свойства, признаки изучаемых статистических объектов (совокупностей процессов) не изолированы, а связаны между собой. Поэтому и показатели этих свойств образуют более или менее полную систему. Явления и процессы никогда не могут быть познаны и охарактеризованы с абсолютной полнотой, так что системы их показателей не являются абсолютно исчерпывающими, и представляют собой лишь частные подсистемы, служащие решению определенной познавательной или управленческой задачи. Число взаимосвязанных показателей может составлять от двух-трех до нескольких сотен.

Различают жестко-детерминированные связи показателей и статистические. Примером системы жестко связанных показателей может служить система объемных и качественных показателей промышленности России за 1995 г.

Абсолютные показатели:

1. Стоимость основных производственных фондов – 4647,6 млрд руб.

2. Численность промышленно-производственного персонала – 16 037 тыс. человек.

3. Объем продукции промышленности – 989,2 млрд руб. в год.

Относительные показатели:

1. Фондовооруженность персонала: $\frac{4647{,}6 \text{ млрд руб.}}{16037 \text{ тыс. чел.}} =$

$= 298{,}8$ тыс. руб./чел.

2. Фондоотдача: $\frac{989{,}2 \text{ млрд руб. в год}}{4647{,}6 \text{ млрд руб.}} = 0{,}21284$ в год.

3. Производительность труда: $\frac{989{,}2 \text{ млрд руб. в год}}{16037 \text{ тыс. чел.}} =$

$= 61{,}681$ тыс. руб./чел. в год.

Каждый показатель этой системы может быть точно вычислен по остальным показателям, так как он является либо частным от деления других показателей, либо произведением показателей. Это означает, что жестко детерминированная система показателей может быть подвергнута арифметической проверке. Например, производительность труда должна быть равна произведению показателей фондовооруженности персонала и фондоотдачи:

$$289{,}8 \text{ тыс. руб./чел.} \cdot 0{,}21284 \text{ в год} = 61{,}681 \text{ тыс.руб. на чел. в год}.$$

Объем продукции промышленности равен произведению трех показателей: численности персонала, его фондовооруженности и фондоотдачи:

16 037 тыс.чел. • 289,8 тыс.руб. на 1 чел. • 0,21284 в год = = 989,2 млрд руб. в год.

Примером системы показателей, связанных статистической зависимостью (сущность которой подробно рассматривается в главе 8), служит система факторов, влияющих на величину заработной платы рабочего. Это

- результативный показатель – средняя месячная заработная плата руб. на 1 чел.;
- факторные показатели:
  возраст рабочего;
  стаж работы по данной специальности;
  число рабочих часов за месяц;
  выработка в числе деталей или операций за час работы;

разряд или класс рабочего;
показатели рентабельности предприятия;
отрасль промышленности.

Никакие арифметические действия над величинами факторных показателей не приводят к величине результативного показателя. Величина не может быть проверена арифметически. Однако средняя величина заработной платы в совокупности рабочих связана со стажем, с разрядом рабочего. Стаж, в свою очередь, связан с возрастом, рентабельность предприятия – с отраслью. Все показатели образуют систему, но связь их проявляется в среднем, для достаточно большой совокупности рабочих.

Система статистических показателей, как правило, должна включать как абсолютные показатели, так и относительные. Изолированный абсолютный показатель подобен человеку в пустыне: он не говорит ничего, ибо ему не с кем говорить. Положим, предприятие произвело продукцию в 1996 г. на 46 млрд руб. Из этого показателя нельзя извлечь никакого вывода, пока его не сопоставить с числом работников, затратами на производство, объемом продукции за предыдущий год и т. п., т. е. пока этот показатель не будет включен в систему и не будут построены относительные показатели. Из этого не следует делать вывод о большей информативности относительных показателей. Если известно, что в студенческой группе число отличников в данную сессию составило 200% к их числу в прошлую сессию, то это не значит, что группа резко повысила уровень знаний. Может быть в прошлую сессию был 1 отличник из 27 человек, а теперь стало 2, что и составило 200%. Только сочетание абсолютных и относительных показателей позволяет достаточно полно характеризовать объект в отношении поставленной задачи его изучения.

3.5. ФУНКЦИИ СТАТИСТИЧЕСКИХ ПОКАЗАТЕЛЕЙ

О роли и значении статистики в развитии общества, в научном познании окружающего мира и в управлении предприятием, учреждением уже сказано в предыдущих главах учебника. Конкретизируем теперь эти вопросы применительно к системам и видам статистических показателей.

Основной функцией конкретных статистических показателей и их систем является *познавательная информационная функция.* Без статистической информации невозможно познание закономерностей природных и социальных массовых явлений, их предвидение, а значит, и регулирование либо прямое управление, будь то на уровне отдельного предприятия, фермера, города или региона, на государственном или межгосударственном уровне. Отдельный человек или семья, не представляющая, сколько в среднем за месяц или за год она расходует на покупку продуктов питания, на обувь и одежду, на оплату коммунальных услуг, не может рационально расходовать средства, планировать свой бюджет. Фермеру необходимо знать показатели средней урожайности за ряд лет различных сельскохозяйственных культур на его участках земли, показатели колеблемости и устойчивости урожаев в зависимости от изменчивых условий погоды, среднюю частоту поломок деталей машин, средние цены (и темпы их роста) на покупаемые удобрения и т. д. Тем более попытки управлять государством субъективно, не опираясь на систему достаточно надежных статистических показателей – путь к социальной, экономической и экологической катастрофе. Условием выполнения статистическими показателями их информационной, познавательной функции является их научное обоснование и достаточно точное и надежное, а также своевременное количественное определение.

Прогностическая функция, т. е. роль статистических показателей в предвидении будущего, тесно связана с их информационной функцией. Конечно, данная функция присуща не всем статистическим показателям, а тем из них, которые используются при моделировании массовых процессов. О применении статистических показателей для прогнозирования будет сказано в главах 8 и 9.

Оценочная функция статистических показателей заключается в том, что на их основе люди, общество, государство оценивают деятельность предприятий, организаций, трудовых и творческих коллективов, правительств. Великий немецкий писатель, поэт и мыслитель И. В. Гете за два года до своей смерти в разговоре со своим секретарем Эккерманном сказал: «Считают, будто числа управляют миром. Но я знаю, что числа учат нас узнавать, хорошо ли мир управляется»[1]. А русский статистик, первый автор учебника статистики в России, К. Ф. Герман (1767– 1838) писал: «Статистик есть публичный провозвестник и доброго, и худого, и контролер правительства»[2]. Да, по надежным «истинным» статистическим показателям, а не по речам и рекламным роликам население должно и может оценивать деятельность руководителей всех рангов.

---

[1] *Eckermann J. P.* Gespräche mit Goethe.– Leipzig, 1902. – S. 313.
[2] *Герман К. Ф.* Всеобщая теория статистики. – СПб, 1809. – П. 78.

Но при этом недопустимо такую оценку давать по отдельному показателю, произвольно вырванному из системы. Долгое время в СССР деятельность предприятий оценивалась на основе показателя выполнения плана по валовой продукции. Поскольку в этот показатель включается и стоимость незавершенных изделий, то ради получения высокого показателя выполнения плана и премии к концу отчетного периода на предприятии аврально собирали шасси, не имея моторов, закладывали новые стройки, не достроив предыдущие, и т. д. Омертвление огромных материальных средств и труда – вот результат превращения отдельного статистического показателя в главное и единственное мерило успехов производства. Так же неверно оценивать успешность развития экономики страны только по показателю низкой инфляции или только по внешнеторговому сальдо – по любому отдельно взятому статистическому показателю.

Рекламно-пропагандистская функция статистических показателей – еще более щекотливый вопрос. С одной стороны, реклама – это одно из нормальных явлений рыночной экономики, и фирмы, компании, естественно, стремятся использовать в рекламе статистические показатели о долговечности, качественности своей продукции, зная, что цифровым данным люди больше доверяют, чем словам. Но при таком использовании статистических показателей велик риск либо подмена реального показателя планируемым, т. е. желаемым, но еще не осуществленным, либо умолчание о других показателях товара, не отвечающих целям рекламы. Поэтому к статистическим показателям, используемым в рекламных интересах, следует относиться весьма осторожно, по возможности проводить дополнительные расчеты и анализ. Например, фирма «Кудесник», рекламируя в газете «Известия» от 14 января 1997 г. кран КС-5579 на базе грузовика «КамАЗ», сообщает, что средний ресурс крана до капитального ремонта составляет 10 лет эксплуатации, или 8000 моточасов. Оба показателя впечатляют. Но если провести расчет, на каких же условиях эксплуатации рассчитан этот ресурс, то на 1 год приходится 800 часов работы, на 1 месяц при 22 рабочих днях – 66 часов, на сутки – 3 часа работы. Неудивительно, что при столь низком показателе использования по времени – всего 0,375 одной смены в сутки, кран, возможно, и на самом деле проработает 10 лет без капитального ремонта.

Так же осторожно следует подходить и к статистическим показателям, используемым государствами, политическими партиями, кандидатами на выборные должности в их пропаганде и агитации. Теоретическая статистика всегда честно указывает, как будет видно из последующих глав, на ограничения, приближенность, вероятностный характер многих своих показателей, лишь постепенно, ограниченно приближающих нас к познанию бесконечно сложного окружающего мира.

Рекомендуемая литература к главе 3

1. *Плошко Б. Г.* Группировка и системы статистических показателей. – М.: Статистика, 1971.
2. *Суслов И. П.* Теория статистических показателей. – М.: Статистика, 1975.
3. *Суслов И. П.* Основы теории достоверности статистических показателей. – Новосибирск: СО «Наука», 1979.

Глава 4

ПРЕДСТАВЛЕНИЕ СТАТИСТИЧЕСКИХ ДАННЫХ: ТАБЛИЦЫ И ГРАФИКИ

Статистические данные должны быть представлены так, чтобы ими можно было пользоваться. Существуют по крайней мере три способа представления данных: они могут быть включены в текст, представлены в таблицах или выражены графически.

4.1. СТАТИСТИЧЕСКИЕ ТАБЛИЦЫ

Если мы включим множество цифр в текст, это затруднит их восприятие. Например, имеем текст:

В 1979 г. в Российской Федерации было 999 городов, из них с численностью населения до 20 тыс. человек – 361 город, или 36%; городов с численностью населения от 500 тыс. человек до 1 млн человек – 18, а городов-миллионеров – 8; в 1993 г. общее число городов возросло на 6,5% и составило 1064, число городов с населением до 20 тыс. человек составило 355, т. е. численность этой категории городов уменьшилась на 12%, тогда как число крупных городов с населением от 500 тыс. человек до 1 млн человек возросло на 17% и достигло 18, городов-миллионеров стало 12, т.е. в 1,5 раза больше, чем в 1979 г.

Даже один абзац, включающий сравнение всего лишь трех категорий общего числа городов и характеристики изменений, плохо воспринимается, не говоря уже о том, что в этом тексте не воспроизведена вся структура городов России, ее динамика. А не сделано это только потому, что текст и без того перегружен цифровыми данными.

Более эффективной формой представления статистических данных являются таблицы.

В отличие от математических таблиц умножения, тригонометрических функций, логарифмов и других, которые по начальным условиям позволяют получить тот или иной результат, статистические таблицы рассказывают языком цифр об изучаемых объектах.

Статистическая таблица – система строк и столбцов, в которых в определенной последовательности и связи *излагается статистическая информация о социально-экономических явлениях.*

Представим в форме таблицы информацию о городах Российской Федерации (табл. 4.1).

Таблица 4.1

Распределение городов Российской Федерации по численности постоянного населения

Число жителей, тыс. чел.	Число городов				
	1979		1993		1993 г. в % к 1979 г.
	число	%	число	%	
Всего	999	100	1064	100	106,5
В том числе:					
4,9 и менее	29	2,9	24	2,2	82,7
5 – 9,9	90	9,0	80	7,3	88,8
10 – 19,9	242	24,2	251	23,6	103,7
20 – 49,9	348	34,8	372	34,9	106,9
50 – 99,9	138	13,8	167	15,7	121,0
100 – 249,9	86	8,6	91	8,5	105,8
250 – 499,9	40	4,0	46	4,3	115,0
500 – 999,9	18	1,8	21	2,0	116,7
1 млн и более	8	0,8	12	1,1	150,0

Источник. Российская Федерация в 1992 году: Статистический ежегодник. – М., 1993.– С. 98.

Таблица позволяет увидеть абсолютную и относительную численность городов каждой категории и их динамику. Различают *подлежащее и сказуемое* статистической таблицы. В подлежащем указывается характеризуемый объект – либо единицы совокупности, либо группы единиц, либо совокупность в целом. В сказуемом дается характеристика подлежащего, обычно в количественной форме – в

виде системы показателей (см. гл. 3). Обязателен *заголовок* таблицы, в котором указывается, к какой категории и какому времени относятся данные таблицы.

По характеру подлежащего статистические таблицы подразделяются на *простые, групповые, комбинационные.*

В подлежащем простой таблицы объект изучения не подразделяется на группы, а дается либо перечень всех единиц совокупности, либо указывается совокупность в целом. В первом случае таблица называется *простой перечневой.* Единицы упорядочиваются по одному-двум признакам (по возрастанию или убыванию значений). Сказуемое должно содержать данные по каждой единице совокупности. Конечно, построение такой таблицы имеет смысл для принятия каких-то оперативных решений; например, для распределения дополнительных дежурств нужно знать, сколько дней отработала каждая медсестра больницы за месяц. Такие таблицы хороши при небольшом числе единиц (10 и менее). Скажем, такую таблицу можно построить для характеристики работы метрополитена в городах России, так как метро имеется лишь в пяти городах.

При большом (несколько десятков и более) числе единиц простые перечневые таблицы составляются только как вспомогательные, например, как основа последующей группировки.

Простые таблицы, содержащие данные о совокупности в целом, можно встретить очень часто в газетах, статистических сборниках. Как правило, они представляют данные в динамике. Примером такой таблицы является табл. 4.2, в которой приведена структура макроэкономического показателя – использованного валового внутреннего продукта (ВВП) России.

Таблица 4.2

Использование валового внутреннего продукта Российской Федерации

(в процентах к итогу в фактически действующих ценах)

Год	ВВП использованный	В том числе		
		расходы на конечное потребление	валовое накопление	чистый экспорт товаров и услуг
1992	100	49,9	35,7	14,4
1996	100	70,8	23,7	5,5

Источник. Россия в цифрах. 1997: Крат. стат. сборник. – М.: Госкомстат России, 1997. – С. 140.

В подлежащем *групповой таблицы* объект изучения подразделяется на группы по *одному признаку.* В сказуемом указываются число единиц в группах (абсолютное и в процентах к итогу) и сводные показатели по группам. Примером такого рода таблицы является табл. 4.3. В этой таблице изучаемая совокупность – занятое население России – распределяется по группам по признаку – сектор экономики.

Таблица 4.3

Распределение занятого населения России по секторам экономики

(млн. чел.)

	1992	1996
Всего занято в экономике	72,1	66,0
В том числе:		
на государственных и муниципальных предприятиях и организациях	49,7	24,4
в частном секторе	13,2	25,2
в общественных организациях, фондах	0,6	0,5
на совместных предприятиях	0,2	0,4
на предприятиях и в организациях смешанной формы собственности	8,4	15,5

Источник. Россия в цифрах. 1997: Крат. стат. сборник. – М.: Госкомстат России, 1997. – С. 34.

В подлежащем *комбинационной таблицы* совокупность подразделяется на группы не по одному, а по *нескольким признакам.* Например, в табл. 4.4 изучаемая совокупность – население России – подразделяется на группы по двум признакам: возрасту и территории проживания (региону).

В табл. 4.4 при выделении групп населения точные возрастные границы не указаны, дана лишь словесная характеристика групп. Так поступают в тех случаях, когда каждая из групп имеет качественные особенности и может рассматриваться как определенный социально-экономический тип. В данном случае эти границы можно было бы указать следующим образом: моложе трудоспособного возраста – мужчины и женщины в возрасте 0-15 лет; в трудоспособ-

Таблица 4.4

Распределение населения по основным возрастным группам по регионам Российской Федерации на 1 января 1996 г.

(%)

Регионы	Моложе трудоспособного возраста	В трудоспособном возрасте	Старше трудоспособного возраста
Российская Федерация	22,5	57,0	20,5
Северный район	23,3	60,0	16,7
Северо-Западный район	19,3	57,8	22,9
Центральный район	19,3	56,6	24,1
Волго-Вятский район	21,8	55,7	22,5
Центрально-Черноземный район	20,6	54,5	24,9
Северо-Кавказский район	25,0	54,9	20,1
Уральский район	23,6	56,6	19,8
Западно-Сибирский район	24,0	58,5	17,5
Восточно-Сибирский район	26,1	58,2	15,7
Дальневосточный район	24,6	62,1	13,3

Источник. Демографический ежегодник России: Стат. сборник. – М.: Госкомстат России, 1996. – С. 46 – 48.

ном возрасте – мужчины 16–59 лет, женщины 16–54 года; старше трудоспособного возраста – мужчины 60 лет и старше, женщины – 55 лет и старше.

При построении таблиц необходимо руководствоваться следующими **общими правилами.**

Подлежащее таблицы располагается в левой части, *сказуемое* – в правой, но могут быть исключения. В простой таблице (см. табл. 4.2) подлежащее, т. е. объект изучения, указано в заголовке таблицы; в комбинационной таблице подлежащее может располагаться в левой и верхней частях таблицы (см. табл. 4.4).

В таблице не должно быть *ни одной лишней линии,* только *необходимые:* линия, отделяющая заголовок таблицы от заголовков ее граф, заголовки граф от цифровых данных. Иногда используется линия, отделяющая итоговую строку. Вертикальная разграфка может быть, а может и отсутствовать.

Заголовки граф содержат названия показателей (без сокращения слов), их единицы измерения. Последние могут указываться как в заголовке соответствующей графы, так и в заголовке таблицы или над таблицей (см., например, табл. 4.4), если все показатели таблицы выражены в одних и тех же единицах измерения и счета.

Итоговая строка завершает таблицу и располагается в конце таблицы, но иногда бывает первой: в этом случае во второй строке дается запись «в том числе», и последующие строки содержат составляющие итоговой строки, но не все, а основные.

Цифровые данные записываются с одной и той же степенью точности в пределах каждой графы; при этом обязательно разряды чисел располагаются под разрядами; целая часть числа отделяется от дробной запятой, например, 4,5, а не 4.5. Заметим, что в международных статистических публикациях используется вместо запятой «точка»; цифры целой части числа в два раза больше дробной 4.5. В таблице не должно быть ни одной пустой клетки: если данные равны нулю, ставится знак «—» (прочерк); если данные не известны, делается запись «сведений нет» или ставится знак «...» (троеточие). Если значение показателя не равно нулю, но первая значащая цифра появляется после принятой степени точности, то делается запись 0,0 (если, скажем, была принята степень точности 0,1).

Если таблица имеет много граф, то графы подлежащего обозначаются заглавными буквами («А», «Б»), а графы сказуемого – цифрами (1, 2 и т.д.). Это бывает удобно; если таблица имеет много строк и печатается на нескольких страницах, то заголовки граф не повторяются, а указываются только их обозначения.

Если таблица основана на заимствованных данных, то под таблицей *указывается источник данных* (см., например, табл. 4.2).

Если хотите, чтобы построенная вами таблица была понятна и удобна для пользования, не пренебрегайте ни одним из указанных правил.

4.2. ОСНОВНЫЕ ВИДЫ ГРАФИКОВ

Иногда статистические таблицы дополняются графиками, когда ставится цель подчеркнуть какую-то особенность данных, провести их сравнение. Графики являются самой эффективной формой представления данных с точки зрения их восприятия. Часто графики используются и вне связи с таблицей. С помощью графиков достигается наглядность характеристики структуры, динамики, взаимосвязи явлений, их сравнения.

Статистические графики представляют собой условные изображения числовых величин и их соотношений посредством линий, геометрических фигур, рисунков или *географических карт-схем.*

Графический способ облегчает рассмотрение статистических данных, делает их наглядными, выразительными, обозримыми. На графике сразу видны пределы изменения показателя, сравнительная скорость изменения разных показателей, их колеблемость. Вместе с тем графики имеют определенные ограничения: прежде всего график не может включить столько данных, сколько может войти в таблицу; кроме того, на графике показываются всегда округленные данные – не точные, а приблизительные. Таким образом, график используется только для изображения общей ситуации, а не деталей. Последний минус – трудоемкость построения графиков. Но этот недостаток может быть преодолен использованием пакетов прикладных программ для компьютерной графики, например ППП «Harvard graphics».

По способу построения графики делятся на *диаграммы, картограммы* и *картодиаграммы.*

Наиболее распространенным способом графического изображения данных являются диаграммы. Они бывают разных видов: линейные, радиальные, точечные, плоскостные, объемные, фигурные. Вид диаграмм зависит от вида представляемых данных (одна переменная или один показатель, несколько переменных или показателей, количественные или неколичественные) и задачи построения графика.

В любом случае график обязательно сопровождается *заголовком* – над или под полем графика. В заголовке указывается, какой показатель изображен, в каких единицах измерения, по какой территории и за какое время он определен.

Линейные графики используются для представления количественных переменных: характеристики вариации их значений, динамики, взаимосвязи между переменными.

Вариация данных анализируется с помощью полигона распределения, *кумуляты* (кривой «меньше, чем») и *огивы* (кривой «больше, чем»). Все эти виды графиков рассматриваются в главе 5. Линейные графики используются в решении задач классификации данных (см. гл. 6). Применение линейных графиков в анализе динамики рассмотрено в главе 9, а использование их для анализа связей – в главе 8. В этих же главах рассмотрено использование точечных диаграмм (см., например, поле корреляции в гл. 8).

Линейные графики целесообразно разделять на используемые для представления данных *по одной переменной* – одномерные или

по *двум переменным* – двумерные. Примером первого является полигон распределения, второго – линия регрессии. Но может быть такой случай, когда на графике представлено несколько переменных (показателей), а он все-таки не является многомерным. Например, на рис. 4.1 представлена динамика объема продукции промышленного и сельскохозяйственного производства в России в 1990–1996 гг.

Для того чтобы динамика двух и более показателей была сопоставимой, следует обеспечить их «единый старт», как это сделано на рис. 4.1, где объемы продукции 1990 г. приняты за 100%.

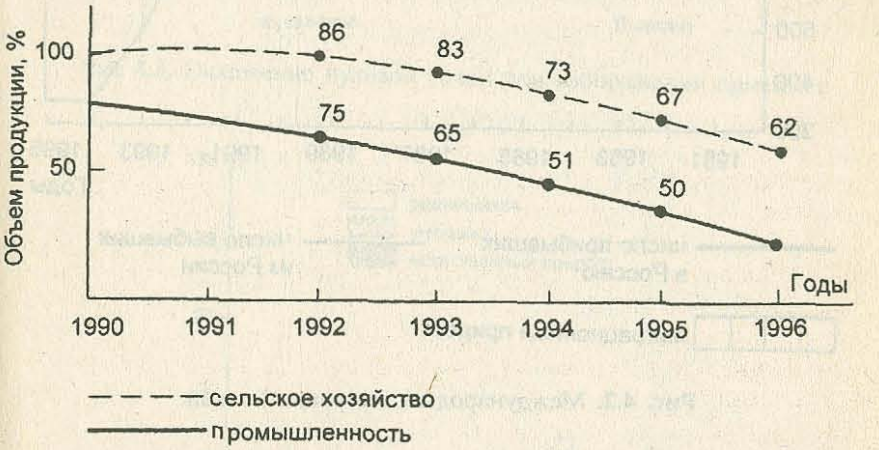


Рис. 4.1. Динамика объема продукции промышленного и сельскохозяйственного производства в России в 1990–1996 гг.

Динамика двух показателей на одном и том же графике может быть представлена и без приведения их к 100%, если эти показатели связаны каким-либо функциональным соотношением (например, представлена динамика общего показателя и показателя, который является одним из его составляющих). Примером такого графика является рис. 4.2.

При графическом изображении динамики по оси абсцисс показывается время (годы, кварталы, месяцы); по оси ординат – значения показателей или показателя. При этом ось ординат должна иметь начало в точке «нуль». Иногда вместо нулевой точки в качестве начального уровня на оси ординат показывается уровень какого-либо года. Это делается в том случае, если изменения изображаемого показателя значительны – в 8–10 и более раз в течение рассматривае-



Рис. 4.2. Международная миграция России

мого отрезка времени. Однако такой прием не рекомендуется. Правильнее указать нулевую точку, а затем (если нужно) «разорвать» ось ординат так, как это показано на рис. 4.3, б.

Иногда при больших изменениях показателя прибегают к логарифмической шкале. Предположим, значения показателя изменяются от 1 до 100 (в 100 раз); это может вызвать затруднения при построении графика. Если перейти к логарифмам, то их значения для минимальных-максимальных значений показателя будут различаться не так сильно: log 1 = 0, log 100 = 2.

Среди *плоскостных диаграмм* по частоте использования выделяются *столбиковые диаграммы,* на которых показатель представляется в виде столбика, высота которого соответствует значению показателя. Пример столбиковой диаграммы представлен на рис. 4.4. Часто на столбиковой диаграмме показываются относительные величины: при сравнении показателей по группам, по разным совокупностям, одна из которых может быть принята за 100%.

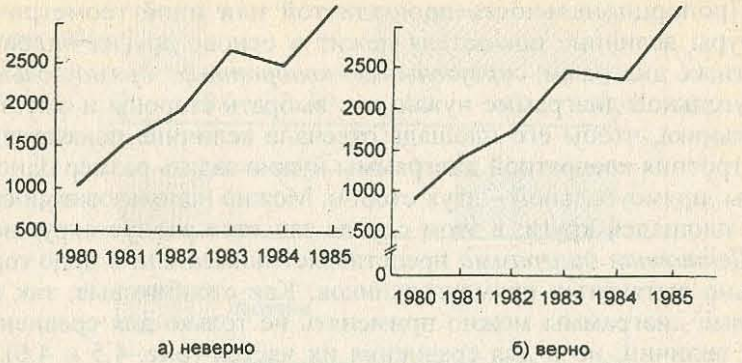


Рис 4.3. Включение нулевой точки при изображении динамики

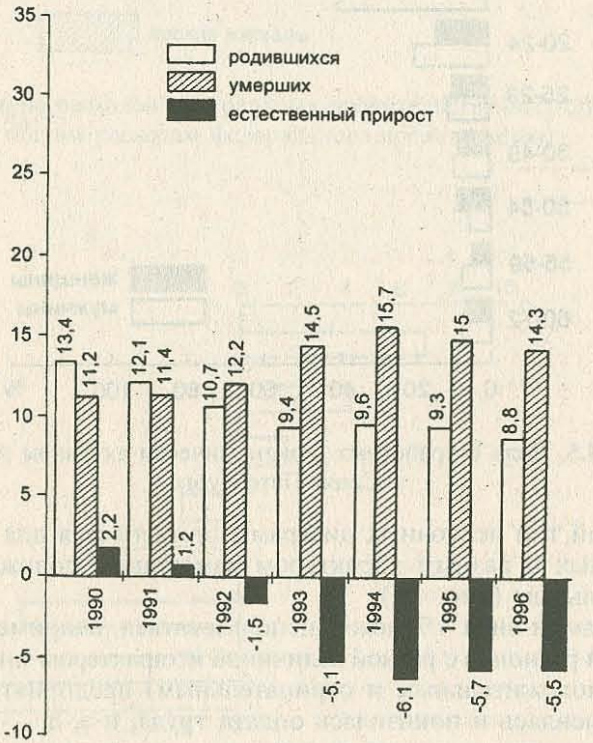


Рис. 4.4. Общие показатели рождаемости, смертности и естественного прироста населения России

Пропорциональность площади той или иной геометрической фигуры величине показателя лежит в основе других видов плоскостных диаграмм: *треугольных, квадратных, прямоугольных.* В треугольной диаграмме нужно так выбрать стороны и высоту треугольника, чтобы его площадь отвечала величине показателя. Для построения квадратной диаграммы нужно задать размер одной стороны, прямоугольной – двух сторон. Можно использовать и сравнение площадей круга; в этом случае задается радиус окружности.

Ленточная диаграмма представляет показатели в виде горизонтально вытянутых прямоугольников. Как столбиковые, так и ленточные диаграммы можно применять не только для сравнения самих величин, но и для сравнения их частей (рис. 4.5 и 4.6).

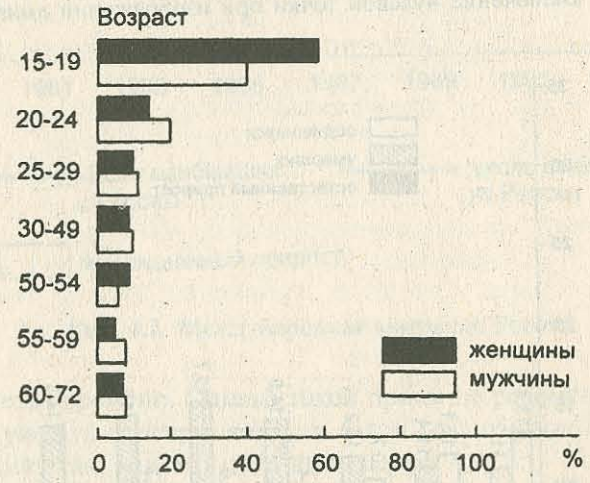


Рис. 4.5. Доля безработных в экономически активном населении Санкт-Петербурга

Особый тип ленточных диаграмм применяется для представления данных с разным характером изменений: положительным и отрицательным (рис. 4.7).

Диаграмма вида 4.7 может использоваться, например, для представления регионов с разной величиной и характером миграционного сальдо (положительным и отрицательным) предприятий, на которых повысилась и понизилась оплата труда, и т. д.

Из плоскостных диаграмм часто используется *секторная диаграмма.* Она применяется для иллюстрации структуры изучаемой совокупности. Вся совокупность принимается за 100%, ей соот-

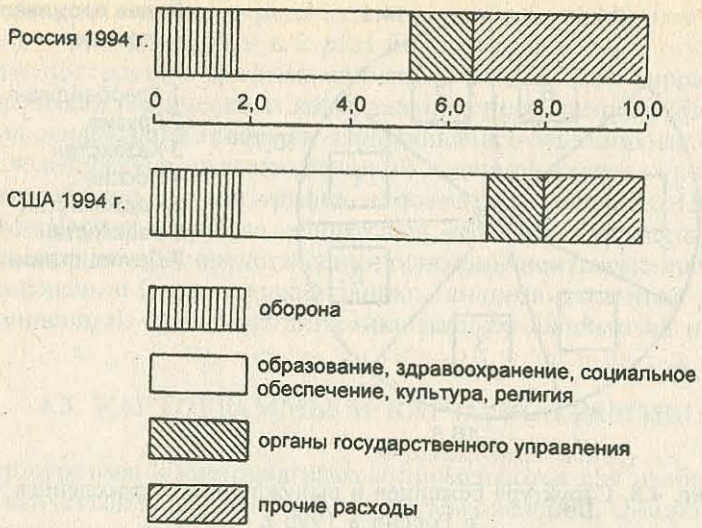


Рис. 4.6. Структура расходов центральных правительственных органов (в % к общим расходам федерального правительства)

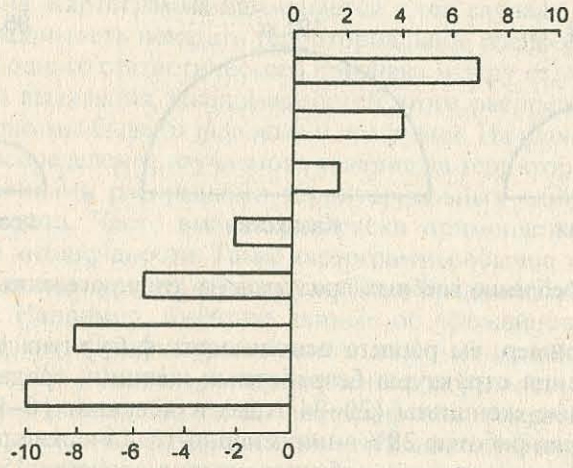


Рис. 4.7. Изменение объема производства на предприятиях текстильной промышленности города (1996 г. по сравнению с 1995 г., в %)

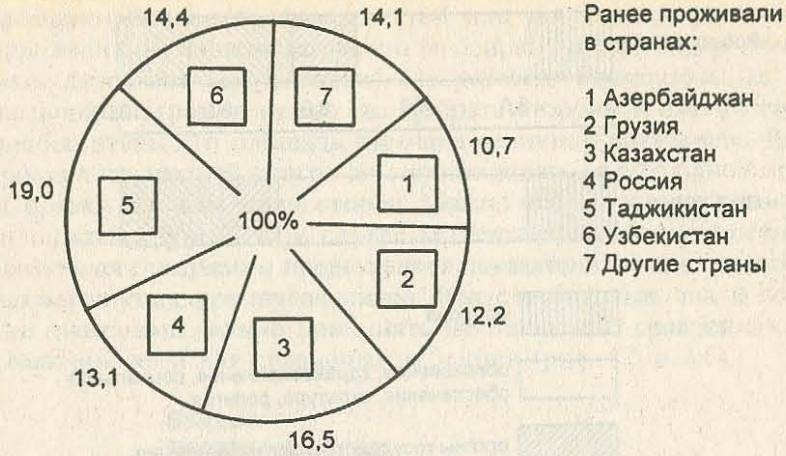


Рис. 4.8. Структура беженцев и вынужденных переселенцев в России в 1996 г.

ветствует общая площадь круга, площади секторов соответствуют частям совокупности (рис. 4.8).

Фигурные (или картинные) *диаграммы* усиливают наглядность изображения, так как включают рисунок изображаемого показателя. Размер рисунка соответствует размеру показателя (рис. 4.9).



Рис. 4.9. Потребление хлебных продуктов на душу населения в 1994 г.

Если, например, вы решите использовать фигурную диаграмму для изображения структуры безработных женщин, среди которых 57% – молодые женщины (20–24 года) и девушки 16–19 лет, не имеющие стажа работы; 28% – инженерно-технические работники и служащие со специальным образованием в возрасте 25–49 лет и 15% – работницы квалифицированного и неквалифицированного труда в возрасте 50 лет и старше, вы должны изобразить три женские фигуры, причем первая из них должна быть в 2 раза больше второй, а вторая – почти в 2 раза больше третьей.

При построении графика одинаково важно все – правильный выбор вида графического изображения, пропорций, соблюдение правил оформления графиков. Подробнее все эти вопросы освещаются в литературе, рекомендованной к данной главе.

Разнообразные виды графиков позволяют получить ППП для ПЭВМ «Harvard-graphics», «Supercalc», «Statistica», «Statgraphics» и др. На графическом представлении основаны некоторые процедуры классификации (группировки) данных, анализа динамики: выявление тенденции, сравнение динамики разных показателей и т. д.

4.3. КАРТОГРАММЫ И КАРТОДИАГРАММЫ

Картограммы и картодиаграммы применяются для изображения географической характеристики изучаемых явлений. Они показывают размещение изучаемого явления, его интенсивность на определенной территории – в республике, области, экономическом или административном районе и т. д.

На *картограмме* распределение изучаемого признака по территории изображается условными знаками (точками, штриховкой, цветом и т. д.), соответствующими определенным интервалам значений величины этого признака. Эти знаки покрывают контур каждого района. Картограмма применяется в тех случаях, когда возникает необходимость показать территориальное распределение какого-нибудь одного статистического признака между отдельными районами для выявления закономерностей этого распределения.

Картограммы бывают *фоновые* и *точечные*. На *фоновых картограммах* распределение изучаемого явления на территории изображается различными раскрасками территориальных единиц с разной густотой цвета. Часто вместо раскраски применяется штриховка различной интенсивности. Такие картограммы обычно используются для изображения уровня относительных и средних величин по территориям. Например, имеются данные об урожайности зерновых по 10 районам области: урожайность до 20 ц/га имеют три смежных района, 20–30 ц/га – четыре смежных района, свыше 30 ц/га – три смежных района. Соответствующая фоновая картограмма представлена на рис. 4.10. Чем более интенсивно явление, тем гуще штриховка (точки) или темнее окраска. Такая картограмма наглядно показывает географию урожайности зерновых культур по районам. Чем больше групп, тем точнее изображение, но большое число

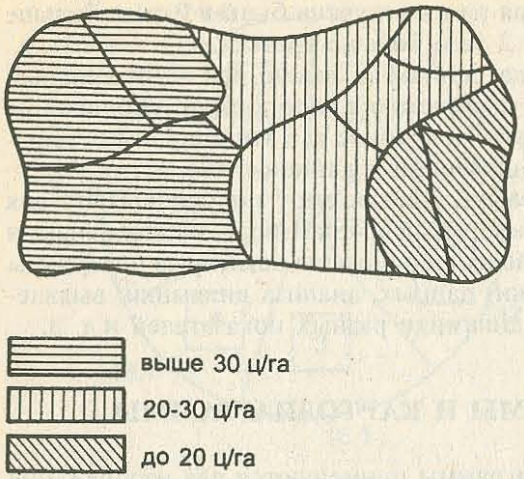


Рис. 4.10. География распределения районов по урожайности зерновых

групп создает пестроту и снижает наглядность. Поэтому практически лучше всего применять не более четырех-пяти тонов или градаций плотности штриховки.

На точечной картограмме символами графического изображения статистических данных являются точки, размещенные в пределах определенных территориальных границ. Точечная картограмма применяется для размещения абсолютных величин. Каждой точке, нанесенной на картограмму, придается числовое значение, что позволяет использовать ее для прямого счета. Например, имеются четыре района с добычей угля в 200, 500, 1000 и 1400 тыс. т в год. Для составления картограммы примем точку за 100 тыс. т и нанесем на контур каждого района соответствующее количество точек (рис. 4.11).

Картодиаграмма – это сочетание диаграммы с географической картой. В качестве изобразительных знаков в картодиаграммах используются те или иные фигуры, которые размещаются на контуре географической карты. Картодиаграммы дают возможность графически отразить более сложные статистико-географические соотношения, чем картограммы. Так, при помощи картодиаграммы можно выразить пространственную специфику в структурах изучаемых статистических совокупностей, особенности каждого района как единого целого и т. д. Например, структурная или секторная картодиаграмма, характеризующая порайонные различия в структуре посевных площадей. В качестве диаграммных знаков в картодиаграмме часто используют различные геометрические фигуры, особенно круги, которые наиболее просты и удобны для выражения сравниваемых количественных показателей на карте.

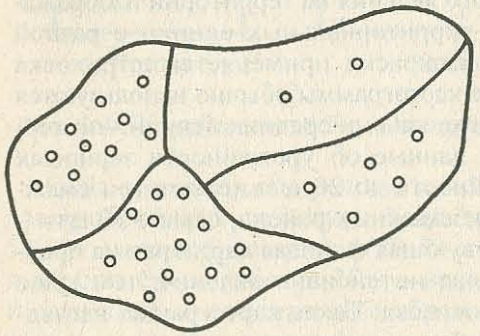


Рис. 4.11. Добыча угля по районам

Кроме рассмотренных видов диаграмм, картограмм и картодиаграмм на практике встречаются и другие, более сложные графические изображения статистических данных.

Рекомендуемая литература к главе 4

1. *Герчук Я. П.* Графические методы в статистике. – М.: Статистика, 1968.
2. *Герчук Я. П.* Графики в математико-статистическом анализе. – М.: Статистика, 1972.
3. Теория статистики /Под ред. Р. А.Шмойловой. – 3-е изд., перераб. – М.: Финансы и статистика, 1998.

5 Глава

СРЕДНИЕ ВЕЛИЧИНЫ И ИЗУЧЕНИЕ ВАРИАЦИИ

5.1. ОДНОРОДНОСТЬ И ВАРИАЦИЯ МАССОВЫХ ЯВЛЕНИЙ

Как уже сказано ранее, статистика изучает массовые явления и процессы. Каждое из таких явлений обладает как общими для всей совокупности, так и особенными, индивидуальными свойствами. Различие между индивидуальными явлениями называют *вариацией*, о ней подробно будет сказано в п. 5.5. Здесь же рассмотрим другое свойство массовых явлений – присущую им близость характеристик отдельных явлений. Если в сосуд с горячей водой добавить холодную, то температура воды во всем сосуде станет одинаковой (осреднится). Поведение детей, поступивших в одну группу детского садика или в один класс школы, тоже приобретает до какой-то степени общие, усредненные черты. Массовое промышленное производство невозможно без стандартизации, т. е. усреднения размеров деталей собираемых механизмов, узлов, агрегатов. Введение севооборота, т. е. ротация разных культур по нескольким участкам пашни, приведет к выравниванию плодородия и механических свойств почвы на этих севооборотных полях. *Итак, взаимодействие элементов совокупности приводит к ограничению вариации хотя бы части их свойств.* Эта тенденция существует объективно. Именно в ее объективности заключена причина широчайшего применения средних величин на практике и в теории.

Каждому рабочему известно, что оплата за простой не по вине рабочего производится по средним расценкам или по среднечасовому заработку. Каждому студенту известно, что такое средний балл на экзаменах. О средних величинах и серьезно, и с насмешкой говорят и пишут философы и журналисты. С помощью метода средних величин статистика решает много задач.

Главное значение средних величин состоит в их *обобщающей функции*, т. е. замене множества различных индивидуальных значений признака средней величиной, характеризующей всю совокупность явлений. Всем известны особенности развития современных людей, проявляющиеся в том числе и в более высоком росте сыновей по сравнению с отцами, дочерей в сравнении с матерями в том же возрасте. Но как измерить это явление? В разных семьях наблюдаются самые различные соотношения роста старшего и младшего поколения. Далеко не всякий сын выше отца и не каждая дочь выше матери. Но если измерить средний рост многих тысяч лиц, то по среднему росту сыновей и отцов, дочерей и матерей можно точно установить и сам факт акселерации, и типичную среднюю величину увеличения роста за одно поколение.

На производство одного и того же количества товара определенного вида и качества разные производители (заводы, фирмы) затрачивают неодинаковое количество труда и материальных ресурсов. Но рынок осредняет эти затраты, и стоимость товара определяется средним расходом ресурсов на производство.

Погода в определенном пункте земного шара в один и тот же день в разные годы может быть очень различной. Например, в Санкт-Петербурге 31 марта температура воздуха за сто с лишним лет наблюдений колебалась от –20,1° в 1883 г. до +12,24° в 1920 г. Примерно такие же колебания наблюдаются и в другие дни года. По таким индивидуальным данным о погоде в какой-то произвольно взятый год нельзя составить представление о климате Санкт-Петербурга. Характеристики климата – это средние за длительный период характеристики погоды – температуры воздуха, его влажность, скорость ветра, сумма осадков, число часов солнечного сияния за неделю, месяц и весь год и т.д. Приведем еще один пример осреднения, его роли в управлении важнейшими и опасными процессами, от которых зависит жизнь людей. Физика установила, что невозможно предсказать, когда произойдет распад ядра радиоактивного атома, например изотопа уран-235. Атом может распасться через секунду или через тысячу лет. Но в массе атомов (например, находящихся в стержнях реактора АЭС) точно можно измерить среднюю скорость распада (обычно используют показатель «время полураспада» – время, за которое распадается половина атомов). Вводя вещества-замедлители образующихся при распаде атомов урана частиц, или убирая их, можно управлять скоростью цепной реакции в урановых стержнях, регулировать мощность реактора, вводить ее в безопасные и экономически выгодные границы.

Если средняя величина обобщает качественно однородные значения признака, то она является типической характеристикой признака в данной совокупности. Так, можно говорить об измерении типичного роста русских девушек рождения 1973 г. по достижении ими 20-летнего возраста. Типичной характеристикой будет средняя величина надоя молока от коров черно-пестрой породы на первом году лактации при норме кормления 12,5 кормовой единицы в сутки. Для лиц с достаточно однородным уровнем дохода, например рабочих машиностроительной отрасли, пенсионеров по старости (исключая имеющих льготы), можно определить типичные доли расходов на покупку предметов питания в их бюджете.

Однако неправильно сводить роль средних величин только к характеристике типичных значений признаков в однородных по данному признаку совокупностях. На практике значительно чаще современная статистика использует средние величины, обобщающие явно неоднородные явления, как, например, урожайность всех зерновых культур по территории всей России, включая кукурузу, дающую по 50-60 ц/га и более, и гречиху, дающую 6-10 ц/га, и плодородные черноземы Кубани, и скудные почвы Архангельской области. Или рассмотрим такую среднюю, как среднее потребление мяса на душу населения: ведь среди этого населения и дети до одного года, вовсе не потребляющие мяса, и вегетарианцы, и северяне, и южане, шахтеры, спортсмены и пенсионеры. Еще более ясна нетипичность такого среднего показателя, как произведенный национальный доход в среднем на душу населения.

Средняя величина национального дохода на душу, средняя урожайность зерновых по всей стране, среднее потребление разных продуктов питания — это характеристики государства как единой народнохозяйственной системы, это так называемые *системные средние.*

Системные средние могут характеризовать как пространственные или объектные системы, существующие одномоментно (государство, отрасль, регион, планета Земля и т.п.), так и динамические системы, протяженные во времени (год, десятилетие, сезон и т.п.). Примером системной средней, характеризующей период времени, может служить средняя температура воздуха в Санкт-Петербурге за 1996 г., равная +5,19°С. Эта средняя величина обобщает и летние высокие температуры +20, +25°, и зимние морозы, осень и весну, дни и ночи.

С другой стороны, средняя температура воздуха за отдельный год не является типической характеристикой климата Санкт-Петербурга, потому что в разные годы средняя температура года значительно колеблется, например за последние 30 лет от +2,90° в 1976 г. до +7,44° в 1989 г. Типической характеристикой климата будет многолетняя средняя годовая температура за десятки лет, например за 1967-1996 гг. она составила +5,05°.

Итак, типическая средняя может обобщать системные средние для однородной совокупности, или системная средняя может обобщать типические средние для единой, хотя и неоднородной, системы. При этом даже типическая средняя не является раз и навсегда данной, неизменной характеристикой.

Так, многолетняя средняя температура в Санкт-Петербурге в первые десятилетия и столетие существования города была значительно ниже; она возрастает медленно, но с ускорением за последнее столетие вследствие как роста самого города и энергопотребления в нем, что повышает температуру воздуха, так и начавшегося и ускоряющегося общего потепления на Земле. Поэтому «типичность» любой средней величины – понятие относительное, ограниченное как в пространстве, так и во времени.

5.2. СРЕДНЯЯ АРИФМЕТИЧЕСКАЯ ВЕЛИЧИНА

Понятие средней арифметической

Виды средних величин различаются прежде всего тем, какое свойство, какой параметр исходной варьирующей массы индивидуальных значений признака должен быть сохранен неизменным.

Средней арифметической величиной называется такое среднее значение признака, при вычислении которого общий объем признака в совокупности сохраняется неизменным.

Иначе можно сказать, что средняя арифметическая величина – среднее слагаемое. При ее вычислении общий объем признака мысленно распределяется поровну между всеми единицами совокупности. Например, средняя заработная плата или средний доход работников предприятия – это такая сумма денег, которая приходилась бы на каждого работника, если бы весь фонд оплаты труда (или все доходы, направленные на личное потребление) был распределен между работниками поровну.

Исходя из определения, формула средней арифметической величины имеет вид:

Средняя арифметическая $\bar{x} = (x_1 + x_2 + \ldots + x_n) : n = \frac{\sum_{i=1}^{n} x_i}{n}$, (5.1)

где $\bar{x}$ – средняя величина;
n – численность совокупности.

По формуле (5.1) вычисляются средние величины *первичных (объемных) признаков*, если известны индивидуальные значения признака. Если изучаемая совокупность велика, исходная информация чаще представляет собой ряд распределения или группировку, как, например, табл. 5.1.

Таблица 5.1

Распределение футбольных матчей высшей лиги России по числу забитых за матч обеими командами мячей в 1996 г.

Число забитых мячей, x	0	1	2	3	4	5	6	7	8	9	Итого
Число матчей, f_i	30	56	71	59	49	24	12	3	0	2	306

Среднее число мячей, забитых за одну игру, должно представлять собой результат равномерного распределения общего числа забитых мячей по всем 306 матчам розыгрыша первенства. Общее число забитых мячей, согласно исходной информации табл. 5.1, можно получить как сумму произведений значений признака в каждой группе x_i, на число игр с таким количеством забитых мячей f_i (частоты). Получим формулу (5.2)

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i f_i}{\sum_{i=1}^{n} f_i}, \quad (5.2)$$

где n – число групп.

Такую форму средней арифметической величины называют *взвешенной арифметической средней* в отличие от простой средней, рассчитанной по формуле (5.1). В качестве весов выступают здесь числа единиц совокупности в разных группах. Название «вес» выражает тот факт, что разные значения признака имеют неодинаковую «важность» при расчете средней величины. «Важнее», весомее число забитых мячей, которое встречалось чаще: 1, 2, 3 мяча, а такие значения, как 7 или 9 забитых мячей, как бы ни радовались таким результативным матчам болельщики, при расчете средней не играют большой роли: их «вес» мал.

Имеем: $\bar{x}$ = 802 : 306 = 2,62 мяча за игру.

Как видим, средняя арифметическая величина может быть дробным числом, если даже индивидуальные значения признака могут принимать только целые значения (дискретный признак). Ничего «предосудительного» для метода средних в этом не заключено; из сущности средней не вытекает, что она обязана быть реальным значением признака, которое могло бы встретиться у какой-либо единицы совокупности.

Виды средней арифметической

Если при группировке значения осредняемого признака заданы интервалами, то при расчете средней арифметической величины в качестве значения признака в группах принимают середины этих интервалов, т.е. исходят из гипотезы о равномерном распределении единиц совокупности по интервалу значений признака. Для открытых интервалов в первой и последней группе, если таковые есть, значения признака надо определить экспертным путем исходя из сущности, свойств признака и совокупности. Например, по табл. 5.2 можно минимальный возраст рабочих считать 17 лет. Тогда первый интервал будет от 17 до 20 лет, а максимальный возраст – 65 лет, тогда последний интервал – 50-65 лет.

Таблица 5.2

Распределение рабочих предприятия по возрасту

Группы рабочих по возрасту, лет	Число рабочих f_j	Середина интервала x'_j	$x_j f_j$
До 20	48	18,5	888
20 – 30	120	25	3000
30 – 40	75	35	2625
40 – 50	62	45	2790
Старше 50	54	57,5	3105
Итого	359	34,56	12408

Средний возраст рабочих, рассчитанный по формуле (5.2) с заменой точных значений признака в группах серединами интервалов, составил:

$$\bar{x} = \frac{\sum_{j=1}^{k} x'_j f_j}{\sum_{j=1}^{k} f_j} = \frac{12408}{359} = 34{,}56 \text{ года},$$

что и записано в итоговую строку по графе 3 табл. 5.2. Напомним, итог объемного показателя — это сумма, итог по графе относительных показателей или средних групповых величин — средняя. Числитель дроби – это общая сумма человеко-лет, прожитых рабочими предприятия; разделив ее на число работников, получаем возраст в годах, так что логика показателя средней величины соблюдена.

Перейдем к рассмотрению средних *вторичных* (относительных) *признаков.* Сумма таких показателей сама по себе реальной величиной какого-либо признака в совокупности не является. Однако общее определение арифметической средней сохраняет силу и в этом случае. При вычислении таких *средних величин необходимо, чтобы сохранялась сумма величины объемного признака, который является числителем при построении осредняемого относительного показателя.* Например, при вычислении средней величины урожайности какой-либо сельскохозяйственной культуры (по формуле (5.2)) необходимо, чтобы общий объем валового сбора этой культуры остался неизменным при замене индивидуальных величин урожайности средней величиной. *Нельзя менять реальную величину объемного признака – она является базой расчета средней.* Чтобы выполнить указанное условие, в качестве весов при расчете средней величины относительного показателя необходимо принять значения того признака, который является знаменателем при определении относительного показателя. Так, при вычислении средней урожайности по совокупности хозяйств весами должны служить размеры площади данной культуры.

Рассмотрим пример расчета средней доли предметов народного потребления в общем выпуске промышленной продукции по совокупности предприятий (табл. 5.3). В этом случае весом должен являться общий объем всей продукции предприятия.

Тогда средняя доля предметов народного потребления в продукции четырех предприятий равна: $\bar{x}$ = (615,5: 2047) · 100% = 30,07%. Средняя доля ближе к долям у тех предприятий, которые имеют

Таблица 5.3

Объем и структура промышленной продукции

Номера предприятий	Объем всей продукции, млн руб., f_j	Доля товаров народного потребления, % x_j	Объем выпуска товаров народного потребления, млн руб., $x_j f_j$
1	138	75	103,5
2	650	38	247,0
3	1040	12	124,8
4	219	64	140,2
Итого	2047	30,07	615,5

большой объем всей продукции (предприятия № 2 и 3). Числитель средней величины $\sum_{j=1}^{4} x_j \cdot f_j$ – это объем выпуска предметов потребления всеми предприятиями – величина, которая должна сохраняться неизменной при замене разных четырех долей на среднюю долю. Расчет по данным табл. 5.3 проведен на основе известных индивидуальных значений осредняемого признака и весов.

Однако исходная информация может иметь другую форму: индивидуальные значения осредняемого признака могут быть неизвестны, зато известны индивидуальные или суммарные значения объемных признаков как числителя, так и знаменателя относительной величины. Например, известно, что в акционерном сельхозпредприятии было посажено 145 га картофеля и собрано с них 2595,5 т продукции. При этом совершенно неизвестно, сколько было собрано с каждого гектара из 145 га в отдельности, хотя на самом деле, конечно, индивидуальные величины продукции, полученные на каждом гектаре, существовали объективно. Однако никакой потребности в их раздельном учете нет; учет продукции ведется по бригадам, по отдельным полям севооборота, но не по каждому гектару. Среднюю урожайность картофеля получают попросту делением массы собранной продукции на площадь посадки, т. е. как относительную величину, характеризующую хозяйство в целом:

$$\text{Средняя урожайность} = \frac{\text{Валовой сбор, т}}{\text{Площадь посадки, га}} = \frac{2595{,}5}{145} = 17{,}9 \text{ т/га}.$$

По отношению к предприятию это относительный показатель. Но существуют и сами значения урожайности с каждого из 145 га хотя и неучтенные. По отношению к ним 17,9 т с 1 га – это средняя величина. Такую форму определения средней арифметической величины, при которой остаются неизвестными индивидуальные значения осредняемого признака, следует называть *неявной формой средней.* Формула такой средней имеет вид:

$$\bar{x} = \frac{\sum y_i}{\sum z_i}, \text{ где } x_i = \frac{y_i}{z_i}.$$

Свойства арифметической средней величины

Знание некоторых **математических свойств средней арифметической** полезно как при ее использовании, так и при ее расчете.

1. Сумма отклонений индивидуальных значений признака от его среднего значения равна нулю.

Доказательство:

$$\sum_{i=1}^{n}(x_i - \bar{x}) = (x_i - \bar{x}) + (x_2 - \bar{x}) + \ldots + (x_n - \bar{x}) =$$

$$= x_1 + x_2 + \ldots + x_n - n\bar{x} = \sum_{i=1}^{n} x_i - n\frac{\sum_{i=1}^{n} x_i}{n} = 0.$$

Примечание. Для взвешенной средней сумма взвешенных отклонений равна нулю. Попробуйте доказать это самостоятельно.

2. Если каждое индивидуальное значение признака умножить или разделить на постоянное число, то и средняя увеличится или уменьшится во столько же раз.

Доказательство:

$$\frac{\sum_{i=1}^{n}(x_i : c)}{n} = \frac{\frac{x_1}{c} + \frac{x_2}{c} + \ldots + \frac{x_n}{c}}{n} = \frac{\frac{x_1 + x_2 + \ldots + x_n}{c}}{n} = \frac{x_1 + x_2 + \ldots + x_n}{n} : c = \bar{x} : c.$$

Вследствие этого свойства индивидуальные значения признака можно сократить в c раз, произвести расчет средней и результат умножить на c.

3. Если к каждому индивидуальному значению признака прибавить или из каждого значения вычесть постоянное число, то средняя величина возрастет или уменьшится на это же число.

Доказательство:

$$\frac{\sum_{i=1}^{n}(x_i + c)}{n} = \frac{(x_1 + c) + (x_2 + c) + \ldots + (x_n + c)}{n} = \frac{\sum_{i=1}^{n} x_i + nc}{n} = \bar{x} + c.$$

Это свойство полезно использовать при расчете средней величины из многозначных и слабоварьирующих значений признака, например роста группы лиц: x_1 = 179 см; x_2 = 183 см; x_3 = 171 см; x_4 = 180 см; x_5 = 169 см. Для вычисления среднего роста из каждого значения вычитаем 170 см и находим среднюю из остатков: (9 + 13 + 1 + 10 – 1) : 5 = 6,4. Средний рост = 6,4 + 170 = 176,4 см.

4. Если веса средней взвешенной умножить или разделить на постоянное число, средняя величина не изменится.

Доказательство:

$$\frac{\sum_{j=1}^{k} x_j \frac{f_j}{c}}{\sum_{j=1}^{k} \frac{f_j}{c}} = \frac{\left(\sum_{j=1}^{k} x_j f_j\right) : c}{\left(\sum_{j=1}^{k} f_j\right) : c} = \bar{x}.$$

Используя это свойство, при расчетах следует сокращать веса на их общий сомножитель либо выражать многозначные числа весов в более крупных единицах измерения.

В табл. 5.4 приведен пример комплексного использования свойств средней арифметической для облегчения расчетов.

Таблица 5.4

Расчет средней продуктивности коров на ферме

Группы коров по надою за год, кг x_j	Число коров f_j	Середина интервала, кг x'_j	$\frac{x'_j - 4000}{100}$	$\frac{x'_j - 4000}{100} \cdot f_j$
3000 – 3400	43	3200	– 8	– 344
3400 – 3800	71	3600	– 4	– 284
3800 – 4200	102	4000	0	0
4200 – 4600	64	4400	4	256
4600 – 5000	27	4800	8	216
Итого	307	–	–	–156

Средний надой молока на корову находим так:

$$\bar{x} = \left[\sum_{j=1}^{5} \frac{x'_j - 4000}{100} \cdot f_j\right] : 307 \cdot 100 + 4000 = \frac{-156}{307} \cdot 100 + 4000 = 3949 \text{ кг.}$$

5. Сумма квадратов отклонений индивидуальных значений признака от средней арифметической меньше, чем от любого другого числа.

Доказательство.

Составим сумму квадратов отклонений от переменной a:

$$f(a) = \sum_{i=1}^{n} (x_i - a)^2.$$

Чтобы найти экстремум этой функции, нужно ее производную по a приравнять нулю:

$$\frac{df}{da} = 2\sum_{i=1}^{n} (x_i - a)(-1) = 0.$$

Отсюда имеем:

$$\sum_{i=1}^{n} (a - x_i) = 0; \quad a\sum_{i=1}^{n} (1) - \sum_{i=1}^{n} (x_i) = 0;$$

$$na = \sum_{i=1}^{n} x_i; \quad a = \frac{\sum_{i=1}^{n} x_i}{n} = \bar{x}.$$

Таким образом, экстремум суммы квадратов отклонений достигается при $a = \bar{x}$. Так как логически ясно, что максимума функция не может иметь, этот экстремум является минимумом.

Применение простой и взвешенной средней

Простая и взвешенная средние величины различаются не только по величине (не всегда), по способу вычисления, но и по своей роли в решении различных задач статистического анализа. Рассмотрим, например, среднюю величину урожайности картофеля в группе хозяйств. Если эта средняя при решении поставленной задачи входит в систему показателей площади посадки, валового сбора, себе-

стоимости, суммы затрат и других характеристик производства, то следует применять взвешенную среднюю, так как произведение невзвешенной средней на общую сумму площадей не даст суммы валового сбора.

Если же нас интересуют такие задачи, как измерение вариации урожайности между хозяйствами или связь урожайности с дозой органических удобрений, то следует применять простую среднюю величину урожайности, полностью абстрагируясь от размеров площадей посадки. Иначе на полученный результат повлияют различия площадей, совершенно не касающиеся этого признака. Точно так же, если необходимо изучить колебания урожайности за ряд лет и выявить их связь с температурой июня и суммой осадков за лето, нужно применять простую среднюю урожайность за ряд лет, абстрагируясь от различия размеров площадей в разные годы.

Чтобы правильно применять средние величины, следует знать, *от каких причин зависит различие между простой и взвешенной средними.* Рассмотрим этот вопрос на примере арифметической средней. Пусть $\bar{x}$ – простая средняя, $\bar{x}_z$ – взвешенная средняя, в которой весами выступают значения признака z, n – число единиц совокупности. Отклонения индивидуальных значений признака x_i от простой средней $\bar{x}$ обозначим $\Delta_{xi} = x_i - \bar{x}$. Отклонения признака-веса $\Delta_{zi} = z_i - \bar{z}$. Тогда индивидуальные значения признаков x и z можно выразить через их средние и отклонения: $x_i = \bar{x} + \Delta_{xi}$; $z_i = \bar{z} + \Delta_{zi}$, а взвешенную среднюю $\bar{x}_z$ представить в виде

$$\bar{x}_z = \sum_{i=1}^{n} (\bar{x} + \Delta_{x_i}) \cdot (\bar{z} + \Delta_{z_i}) : \sum_{i=1}^{n} z_i. \qquad (5.3)$$

Перемножим величины в скобках и просуммируем почленно, имея в виду, что $\sum_{i=1}^{n} \bar{x} = n\bar{x}$; $\sum_{i=1}^{n} \bar{z} = n\bar{z}$. Средние величины можно вынести за знак суммирования, как константы. Получим:

$$\bar{x}_z = \frac{n\bar{x}\bar{z} + \bar{x}\sum_{i=1}^{n} \Delta_{zi} + \bar{z}\sum_{i=1}^{n} \Delta_{xi} + \sum_{i=1}^{n} \Delta_{xi} \cdot \Delta_{zi}}{n\bar{z}}.$$

Так как суммы отклонений индивидуальных значений признака от средней арифметической согласно первому ее свойству равны нулю, то второе и третье слагаемые числителя также равны нулю.

Остается:

$$\bar{x}_z = \bar{x} + \frac{\sum_{i=1}^{n} \Delta_{xi}\Delta_{zi}}{n\bar{z}}. \qquad (5.4)$$

Числитель второго слагаемого в формуле (5.4) – это числитель коэффициента корреляции между осредняемым и весовым признаками (см. формулы 8.11 и 8.14). Подставив выражение коэффициента корреляции r_{xz} в (5.4), получим:

$$\bar{x}_z = \bar{x} + \frac{n\sigma_x\sigma_z r_{xz}}{n\bar{z}} = \bar{x} + \sigma_x v_z r_{xz}. \qquad (5.5)$$

Итак, средняя арифметическая взвешенная равна простой средней плюс произведение среднего квадратического отклонения осредняемого признака на коэффициент вариации весового признака и на коэффициент корреляции между этими признаками. Если обе части равенства (5.5) разделить на простую среднюю $\bar{x}$, получим:

$$\frac{\bar{x}_z}{\bar{x}} = 1 + v_x v_z r_{xz}.$$

(О среднем квадратическом отклонении и коэффициенте вариации см. ниже в этой главе.)

Из (5.5) следует, что взвешенная средняя равна простой **в трех случаях:**

- а) если не варьирует изучаемый признак, $\sigma_x = 0$ – тривиальная ситуация, когда и сами средние не нужны;
- б) при условии, что не варьирует признак-вес $v_z = 0$;
- в) в случаях, когда между осредняемым и признаком-весом нет линейной корреляции, $r_{xz} = 0$.

Взвешенная средняя больше простой, если эта корреляция прямая. Взвешенная средняя меньше простой средней, если эта корреляция обратная.

5.3. ДРУГИЕ ФОРМЫ СРЕДНИХ ВЕЛИЧИН

Средняя квадратическая величина

Если при замене индивидуальных величин признака на среднюю величину необходимо сохранить неизменной сумму квадратов исходных величин, то средняя будет являться *квадратической средней величиной* ($\overline{x_{кв}}$). Ее формула такова:

$$\overline{x_{кв}} = \sqrt{\frac{\sum_{i=1}^{n} x_i^2}{n}}. \qquad (5.6)$$

Например, имеются три участка земельной площади со сторонами квадрата: x_1 = 100 м; x_2 = 200 м; x_3 = 300 м. Заменяя разные значения длины сторон на среднюю, мы, очевидно, должны исходить из сохранения общей площади всех участков. Арифметическая средняя величина (100+ 200 + 300) : 3 = 200 м не удовлетворяет этому условию, так как общая площадь трех участков со стороной 200 м была бы равна: $3 \cdot (200\ м)^2 = 120\ 000\ м^2$. В то же время площадь исходных трех участков равна: $(100\ м)^2 + (200\ м)^2 + (300\ м)^2 = 140\ 000\ м^2$. Правильный ответ дает квадратическая средняя:

$$\bar{x}_{кв} = \sqrt{\frac{(100)^2 + (200)^2 + (300)^2}{3}} = 216\ м.$$

Во второй части главы будет показано, что главной сферой применения квадратической средней в силу пятого свойства средней арифметической величины является измерение вариации признака в совокупности.

Аналогично, если по условиям задачи необходимо сохранить неизменной сумму кубов индивидуальных значений признака при их замене на среднюю величину, мы приходим к *средней кубической,* имеющей вид:

$$\overline{x_{куб}} = \sqrt[3]{\frac{\sum_{i=1}^{n} x_i^3}{n}}. \qquad (5.7)$$

Средняя геометрическая величина

Если при замене индивидуальных величин признака на среднюю величину необходимо сохранить неизменным произведение индивидуальных величин, то следует применить *геометрическую среднюю величину.* Ее формула такова:

$$\bar{x}_{геом} = \sqrt[n]{x_1 \cdot x_2 \cdot \ldots \cdot x_n}. \qquad (5.8)$$

Основное применение геометрическая средняя находит при определении средних темпов роста, о чем сказано в главе 9. Пусть, например, в результате инфляции за первый год цена товара возросла в 2 раза к предыдущему году, а за второй год еще в 3 раза к уровню предыдущего года. Ясно, что за два года цена выросла в 6 раз. Каков средний темп роста цены за год? Арифметическая средняя здесь непригодна, ибо если за год цены возросли бы в $\frac{2+3}{2}=2{,}5$ раза, то за два года цена возросла бы в 2,5×2,5 = 6, 25 раза, а не в 6 раз. Геометрическая средняя дает правильный ответ: $\sqrt{6}=2{,}45$ раза.

Геометрическая средняя величина дает наиболее правильный по содержанию результат осреднения, если задача состоит в нахождении такого значения признака, который качественно был бы равно удален как от максимального, так и от минимального значения признака. Например, если максимальный размер выигрыша в лотерее составляет миллион рублей, а минимальный – сто рублей, то какую величину выигрыша можно считать средней между миллионом и сотней? Арифметическая средняя явно непригодна, она составляет 500 050 руб., а это, как и миллион, крупный, никак не средний выигрыш; он качественно однороден с максимальным и резко отличен от минимального. Не дают верного ответа ни квадратическая средняя (707 107 руб.), ни кубическая (793 699 руб.), ни рассматриваемая далее гармоническая средняя (199,98 руб.), слишком близкая к минимальному значению. Только геометрическая средняя дает верный с точки зрения экономики и логики ответ: $\sqrt{100\cdot 1\,000\,000}=10\,000$ руб. Десять тысяч — не миллион, и не сотня! Это, действительно, нечто среднее между ними.

Средняя гармоническая величина

Если по условиям задачи необходимо, чтобы неизменной оставалась при осреднении сумма величин, обратных индивидуальным значениям признака, то средняя величина является *гармонической средней.*

Формула гармонической средней величины такова:

$$\bar{x}_{гарм}=\frac{n}{\sum\limits_{i=1}^{n}\frac{1}{x_i}}. \qquad (5.9)$$

Например, автомобиль с грузом от предприятия до склада ехал со скоростью 40 км/ч, а обратно порожняком – со скоростью 60 км/ч.

Какова средняя скорость автомобиля за обе поездки? Пусть расстояние перевозки составляло s км. Никакой роли при расчете средней скорости величина s не играет. При замене индивидуальных значений скорости $x_1 = 60$ и $x_2 = 40$ на среднюю величину необходимо, чтобы неизменной величиной осталось время, затраченное на обе поездки, иначе средняя скорость может оказаться любой — от скорости черепахи до скорости света.

Время поездок есть $s/x_1 + s/x_2$. Итак, $s/\bar{x} + s/\bar{x} = s/x_1 + s/x_2$. Сократив все члены равенства на s, получим $1/\bar{x} + 1/\bar{x} = 1/x_1, + 1/x_2$, т.е. выполняется условие гармонической средней. Подставляя x_1 и x_2, получаем:

$$\frac{2}{\bar{x}}=\frac{1}{60}+\frac{1}{40},\quad \bar{x}=\frac{2}{\frac{1}{60}+\frac{1}{40}}=\frac{2\cdot 120}{5}=48 \text{ км/ч}.$$

Арифметическая средняя 50 км в час неверна, так как приводит к другому времени движения, чем на самом деле. Если расстояние равно 96 км, то реальное время движения составит:

$$\frac{96}{60}+\frac{96}{40}=1{,}6 \text{ ч}+2{,}4 \text{ ч}=4 \text{ ч}.$$

То же время дает гармоническая средняя:

$$\frac{(96\cdot 2)}{48}=4 \text{ ч}.$$

Понятие степенной средней. Соотношение между формами средних величин

Все рассмотренные выше виды средних величин принадлежат к общему типу *степенных средних.* Различаются они лишь показателем. Степенная средняя степени k есть корень k-*й* степени из частного от деления суммы индивидуальных значений признака в k-й степени на число индивидуальных значений:

$$\bar{x}=\sqrt[k]{\frac{\sum\limits_{i=1}^{n}x_i^k}{n}}. \qquad (5.10)$$

При $k = 1$ получаем арифметическую среднюю, при $k = 2$ – квадратическую, при $k = 3$ – кубическую, при $k = 0$ – геометрическую, при

$k = -1$ – гармоническую среднюю. Чем выше показатель степени k, тем больше значение средней величины (если индивидуальные значения признака варьируют). Если все исходные значения признака равны, то и все средние равны этой константе. Итак, имеем следующее соотношение, которое называется *правилом мажорантности средних:*

$$\bar{x}_{гарм} \le \bar{x}_{геом} \le \bar{x}_{арифм} \le \bar{x}_{квадр} \le \bar{x}_{куб}. \qquad (5.11)$$

Пользуясь этим правилом, статистика может в зависимости от настроения и желания ее «знатока» либо «утопить», либо «выручить» студента, получившего на сессии оценки 2 и 5. Каков его средний балл?

Если судить по средней арифметической, то средний балл равен 3,5. Но если декан желает «утопить» несчастного и вычислит среднюю гармоническую

$$\bar{x}_{гарм} = \frac{2}{\frac{1}{2}+\frac{1}{5}} = \frac{20}{7} = 2{,}86,$$

то студент остается в среднем двоечником, не дотянувшим до тройки. Однако студенческий комитет может возразить декану и представить среднюю кубическую величину:

$$\bar{x}_{куб} = \sqrt[3]{\frac{2^3+5^3}{2}} = \sqrt[3]{66{,}5} = 4{,}05.$$

Студент уже выглядит «хорошистом» и даже претендует на стипендию! И только в том случае, если лентяй провалил оба экзамена, статистика помочь не в состоянии: увы, все средние из двух двоек равны все той же двойке!

5.4. СРЕДНЯЯ ВЕЛИЧИНА КАК ВЫРАЖЕНИЕ ЗАКОНОМЕРНОСТИ

После того как мы познакомились с различными видами и формами средних величин, включая и неявную их форму, можно перейти к понятию о средних. В широком понимании термина *средней величиной* является всякий обобщающий показатель, характеризующий обобщенное значение признака, связи признаков, их динамики и структуры в совокупности массовых явлений.

Так, средними в широком смысле слова являются такие показатели, как доля мужчин в общем числе жителей страны (ведь эта доля разная в разных регионах), плотность населения, коэффициент смертности, ожидаемая продолжительность жизни родившихся в данном году и др. Рассматриваемые далее в этой главе показатели вариации признака в совокупности, а также в главе 8 показатели корреляционной связи тоже средние в широком смысле слова, так как измеряют среднее различие между значениями одного признака у разных единиц совокупности или среднюю связь вариации одного признака с вариацией другого.

В такой же степени средними являются и показатели темпов роста продукции промышленности или национального дохода страны, обобщающие темпы разных отраслей и регионов; средними являются меры колеблемости урожайности за ряд лет (гл. 9), обобщающие влияние на урожайность разных лет метеорологических и экономических условий производства.

Понятие средней в широком смысле слова сближается с такой философской категорией, как закон («закон есть общее в явлениях»), закономерность. Это далеко не случайное родство. Рассмотрим сущность процесса осреднения на примере арифметической средней согласно формуле (5.1). Среднюю считаем типической, определенной по однородной совокупности. Однородность индивидуальных значений признака — *это проявление их общих свойств, обусловленных* основными условиями и закономерностями массового процесса, порождающего данную совокупность. Однако кроме общих условий, кроме закономерности на каждую единицу совокупности влияют индивидуальные, особенные условия, случайные события, не связанные причинно с общей закономерностью. Поэтому можно индивидуальные значения признака x_i представить как состоящие из элемента, обусловленного общей закономерностью для всех единиц совокупности (обозначим этот элемент c), так и элемента Δ_i, индивидуального для каждой единицы совокупности. Итак, $x_i = c + \Delta_i$, где Δ_i может быть как положительной, так и отрицательной величиной, как малой, так и большой величиной в сравнении с c.

Теперь вычислим среднее значение признака для совокупности из n единиц:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{\sum_{i=1}^{n}(c+\Delta_i)}{n} = \frac{nc+\sum_{i=1}^{n}\Delta_i}{n} = c + \bar{\Delta}.$$

Итак, средняя величина признака слагается из элемента, выражающего закономерность, общую для всей совокупности, и из средней величины элементов, отражающих индивидуальные условия отдельных единиц этой совокупности. Элементы Δ_i могут иметь положительные и отрицательные, большие и малые значения. При осреднении они согласно закону больших чисел взаимопогашаются в зависимости от объема совокупности: тем в большей мере, чем больше объем совокупности *n*. Об этом говорит формулировка закона больших чисел, данная великим русским математиком П. Л. Чебышевым (1821-1894). Чем больше объем однородной совокупности, тем полнее взаимопогашение случайных (по отношению к совокупности в целом и ее законам) элементов признака *x;* полнее и надежнее, с большей вероятностью среднее значение признака измеряет действие общих для совокупности закономерностей.

Однако случайная вариация индивидуальных величин признаков – это не только некоторая помеха, туман, «шум» в информационном смысле, затрудняющий познание закономерности. Вариация – неотъемлемая, необходимая черта, свойство массовых явлений, имеющее громадное самостоятельное значение в развитии природы и общества.

Создатель учения о средних величинах бельгийский статистик А. Кетле по этому поводу писал следующее: «В мире существует общий закон, предназначенный как бы для того, чтобы разливать жизнь во Вселенной; в силу этого закона все живущее подлежит бесконечному разнообразию... Каждый предмет подвержен флюктуациям»[1].

В следующих разделах данной главы переходим к методам статистического изучения этого «общего закона Вселенной» – вариации массовых явлений и их признаков.

5.5. ВАРИАЦИЯ МАССОВЫХ ЯВЛЕНИЙ

Вариацией значений какого-либо признака в совокупности называется различие его значений у разных единиц данной совокупности в один и тот же период или момент времени.

В отличие от вариации различия значений признака у одного и того же объекта, у одной и той же единицы совокупности в разные моменты или периоды времени следует называть *изменениями во времени и колебаниями.* Методы их измерения и изучения отличаются принципиально от методов измерения вариации (см. гл. 9).

Причиной вариации являются разные условия существования разных единиц совокупности. Даже однояйцовые близнецы в процессе своего развития приобретают различия в росте, весе, не говоря уже о таких признаках, как специальность, образование, заработная плата (доход), число детей и т.д. Еще больше причин влияют на различия промышленных предприятий, магазинов и т. д.

Вариация присуща всем без исключения явлениям природы и общества, кроме законодательно закрепленных нормативных значений отдельных социальных признаков: не варьирует признак «число председателей правления колхоза» – все они имеют по одному председателю. Не варьирующие признаки не представляют интереса для статистики; *предметом изучения статистики* является *вариация.* Большинство методов статистики – это либо методы измерения вариации, либо методы абстрагирования от нее.

Вариация, несомненно, необходимое условие существования и развития массовых явлений. Например, вариация геномов (набора генов) родительских организмов растений и животных обеспечивает жизнеспособность потомства. Близкородственный брак, т.е. слишком малая вариация геномов родителей, ведет к неполноценному потомству. Перекрестное опыление для многих растений – обязательное условие плодоношения. Гибридизация, т.е. получение потомства от неродственных, со значительной вариацией свойств сортов сельскохозяйственных растений и пород животных — важный прием повышения урожайности и продуктивности скота.

В то же время известно, что нельзя получить потомство от организмов со слишком разными свойствами — разных видов, родов и семейств, например от кошки и собаки. Чрезмерная вариация генотипов препятствует развитию. И в промышленном производстве, особенно массовом, вариация размеров, свойств деталей, из которых собирается станок, автомашина, телевизор, должна быть введена в жесткие рамки «допусков», т. е. пренебрежимо малых величин, чтобы сборка была возможной и не страдало качество собранного агрегата.

Итак, в жизни общества, как и в природе, каждой массовой совокупности, массовому процессу присуща некоторая специфическая мера вариации ее элементов, при которой данный процесс протекает оптимально.

Чтобы руководитель предприятия, менеджер, научный работник могли управлять вариацией и изучать ее, статистикой разработаны специальные методы исследования вариации, система показателей, с помощью которой вариация измеряется, характеризуются ее свойства.

---

[1] *Кетле А.* Социальная система и законы ею управляющие: Пер. с фр. – СПб, 1866. – С. 16.

5.6. ПОСТРОЕНИЕ ВАРИАЦИОННОГО РЯДА. ВИДЫ РЯДОВ. РАНЖИРОВАНИЕ ДАННЫХ

Первым этапом статистического изучения вариации являются построение ***вариационного ряда*** – *упорядоченного распределения единиц совокупности по возрастающим (чаще) или по убывающим (реже) значениям признака и подсчет числа единиц с тем или иным значением признака.*

Существуют три формы вариационного ряда: ранжированный ряд, дискретный ряд, интервальный ряд. Вариационный ряд часто называют *рядом распределения.* Этот термин используется при изучении вариации как количественных, так и неколичественных признаков. Ряд распределения представляет собой *структурную группировку* (см. гл. 6).

Ранжированный ряд — это перечень отдельных единиц совокупности в порядке возрастания (убывания) изучаемого признака.

Примером ранжированного ряда может служить табл. 5.5.

Таблица 5.5

Крупные банки Санкт-Петербурга, ранжированные по размерам собственного капитала на 01.07.96

Название банка	Собственный капитал, млрд руб.
Петроагропромбанк	71
Петровский	146
Балтийский	196
Банк Санкт-Петербург	201
Промстройбанк	731

Если численность единиц совокупности достаточно велика, ранжированный ряд становится громоздким, а его построение, даже с помощью ЭВМ, занимает длительное время. В таких случаях вариационный ряд строится с помощью группировки единиц совокупности по значениям изучаемого признака.

Если признак принимает небольшое число значений, строится дискретный вариационный ряд. Примером такого ряда является распределение футбольных матчей по числу забитых мячей (табл. 5.1). *Дискретный вариационный ряд* – это таблица, состоящая из двух строк или граф: конкретных значений варьирующего признака x_i и числа единиц совокупности с данным значением признака f_i – частот (f - начальная буква англ. слова frequency).

Определение числа групп

Число групп в дискретном вариационном ряду определяется числом реально существующих значений варьирующего признака. Если же признак может принимать хотя и дискретные значения, но их число очень велико (например, поголовье скота на 1 января года в разных сельхозпредприятиях может составлять от нуля до десятков тысяч голов), тогда строится интервальный вариационный ряд. Интервальный вариационный ряд строится и для изучения признаков, которые могут принимать любые, как целые, так и дробные, значения в области своего существования. Таковы, например, рентабельность реализованной продукции, себестоимость единицы продукции, доход на 1 жителя города, доля лиц с высшим образованием среди населения разных территорий и вообще все вторичные признаки, значения которых рассчитываются путем деления величины одного первичного признака на величину другого (см. гл. 3).

Интервальный вариационный ряд представляет собой таблицу, состоящую из двух граф (или строк) — интервалов признака, вариация которого изучается, и числа единиц совокупности, попадающих в данный интервал (частот), или долей этого числа от общей численности совокупности (частостей).

При построении интервального вариационного ряда необходимо выбрать оптимальное число групп (интервалов признака) и установить длину интервала. Поскольку при анализе вариационного ряда сравнивают частоты в разных интервалах, необходимо, чтобы величина интервала была постоянной. Оптимальное число групп выбирается так, чтобы в достаточной мере отразилось разнообразие значений признака в совокупности и в то же время закономерность распределения, его форма не искажалась случайными колебаниями частот. Если групп будет слишком мало, не проявится закономерность вариации; если групп будет чрезмерно много, случайные скачки частот исказят форму распределения.

Чаще всего число групп в вариационном ряду устанавливают, придерживаясь формулы, рекомендованной американским статистиком Стерджессом *(Sturgess):*

$$k \approx 1 + 3{,}32 \cdot \lg n = 1{,}441\mathrm{n}\, n{+}1,$$

где k – число групп; n – численность совокупности.

Эта формула показывает, что число групп – функция объема данных.

Предположим, необходимо построить вариационный ряд распределения предприятий области по урожайности зерновых культур за какой-то год. Число сельхозпредприятий, имевших посевы зерновых культур, составило 143; наименьшее значение урожайности равно 10,7 ц/га, наибольшее – 53,1 ц/га. Имеем:

$$k \approx 1 + 3{,}32 \lg \cdot 143 = 8{,}16.$$

Так как число групп целое, следовательно, рекомендуется построить 8 или 9 групп.

Определение величины интервала

Зная число групп, рассчитывают величину интервала:

$$i = \frac{x_{max} - x_{min}}{k}.$$

В нашем примере величина интервала составляет:

а) при 8 группах

$$i = \frac{53{,}1 - 10{,}7}{8} = 5{,}3 \text{ ц/га},$$

б) при 9 группах

$$i = \frac{53{,}1 - 10{,}7}{9} = 4{,}7 \text{ ц/га}.$$

Для построения ряда и анализа вариации значительно лучше иметь по возможности округленные значения величины интервала и его границ. Поэтому наилучшим решением будет построение вариационного ряда с 9 группами с интервалом, равным 5 ц/га. Этот вариационный ряд приведен в табл. 5.6, а его графическое изображение дано на рис. 5.1.

Границы интервалов могут указываться разным образом: верхняя граница предыдущего интервала повторяет нижнюю границу следующего, как показано в табл. 5.6, или не повторяет.

В последнем случае второй интервал будет обозначен как 15,1-20, третий как 20,1-25 и т.д., т.е. предполагается, что все значения урожайности обязательно округлены до одной десятой. Кроме того, возникает нежелательное осложнение с серединой интер-

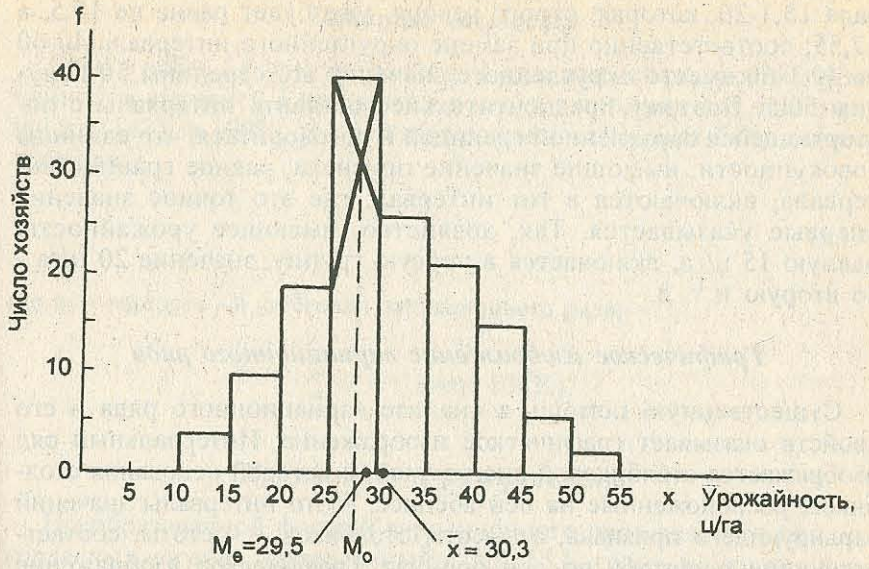


Рис. 5.1. Распределение хозяйств по урожайности

Таблица 5.6

Распределение хозяйств области по урожайности зерновых культур

Группы хозяйств по урожайности, ц/га x_j	Число хозяйств f_j	Середина интервала, ц/га x'_j	$x'_j f_j$	Накопленная частота f'_j
10 – 15	6	12,5	75,0	6
15 – 20	9	17,5	157,5	15
20 – 25	20	22,5	450,0	35
25 – 30	41	27,5	1127,5	76
30 – 35	26	32,5	845,0	102
35 – 40	21	37,5	787,5	123
40 – 45	14	42,5	595,0	137
45 – 50	5	47,5	237,5	142
50 – 55	1	52,5	52,5	143
Итого	143		4327,5	

вала 15,1-20, которая, строго говоря, уже будет равна не 17,5, а 17,55; соответственно при замене округленного интервала 40-60 на 40,1-60 вместо округленного значения его середины 50 получим 50,5. Поэтому предпочтительнее оставить интервалы с повторяющейся округленной границей и договориться, что единицы совокупности, имеющие значение признака, равное границе интервала, включаются в тот интервал, где это точное значение впервые указывается. Так, хозяйство, имеющее урожайность, равную 15 ц/га, включается в первую группу, значение 20 ц/га - во вторую и т. д.

Графическое изображение вариационного ряда

Существенную помощь в анализе вариационного ряда и его свойств оказывает графическое изображение. Интервальный ряд изображается столбиковой диаграммой, в которой основания столбиков, расположенные на оси абсцисс, – это интервалы значений варьирующего признака, а высоты столбиков – частоты, соответствующие масштабу по оси ординат. Графическое изображение распределения хозяйств области по урожайности зерновых культур приведено на рис. 5.1. Диаграмма этого рода часто называется *гистограммой* (от греческого слова «гистос» – ткань, строение).

Данные табл. 5.5 и рис. 5.1 показывают характерную для многих признаков форму распределения: чаще встречаются значения средних интервалов признака, реже – крайние; малые и большие значения признака. Форма этого распределения близка к рассматриваемому в курсе математической статистики закону нормального распределения. Великий русский математик А. М. Ляпунов (1857 – 1918) доказал, что нормальное распределение образуется, если на варьирующую переменную влияет большое число факторов, ни один из которых не имеет преобладающего влияния. Случайное сочетание множества примерно равных факторов, влияющих на вариацию урожайности зерновых культур, как природных, так и агротехнических, экономических, создает близкое к нормальному закону распределения распределение хозяйств области по урожайности.

Если имеется дискретный вариационный ряд или используются середины интервалов, то графическое изображение такого вариационного ряда называется *полигоном* (от греч. слова – многоугольник). Каждый из вас легко построит этот график, соединяя прямыми точки с координатами x_i и f_i.

Отношение высоты полигона или диаграммы к их основанию рекомендуется в пропорции примерно 5:8.

Понятие частости

Если в табл. 5.6 число хозяйств с тем или иным уровнем урожайности выразить в процентах к итогу, принимая все число хозяйств (143) за 100%, то средняя урожайность может быть вычислена так:

$$\bar{x} = \sum_{j=1}^{k} x'_j w_j, \qquad (5.12)$$

где w_j – частость j-й категории вариационного ряда;

$$\sum_{j=1}^{k} w_j = 1 \text{ (или 100\%)}.$$

Кумулятивное распределение

Преобразованной формой вариационного ряда является *ряд накопленных частот,* приведенный в табл. 5.6, графа 5. Это ряд значений числа единиц совокупности с меньшими и равными нижней границе соответствующего интервала значениями признака. Такой ряд называется *кумулятивным.* Можно построить кумулятивное распределение «не меньше, чем», а можно «больше, чем». В первом случае график кумулятивного распределения называется *кумулятой,* во втором – *огивой* (рис. 5.2).

Плотность распределения

Если приходится иметь дело с вариационным рядом с неравными интервалами, то для сопоставимости нужно частоты или частости привести к единице интервала. Полученное отношение называется *плотностью распределения*:

$$f'_j = \frac{f_j}{i_j} \text{ или } w'_j = \frac{w_j}{i_j}, \qquad (5.13)$$

где f'_j – абсолютная плотность распределения в j-м интервале;
w'_j – относительная плотность распределения в j-м интервале;
i_j – ширина j-го интервала.

Плотность распределения используется как для расчета обобщающих показателей, так и для графического изображения вариационных рядов с неравными интервалами.

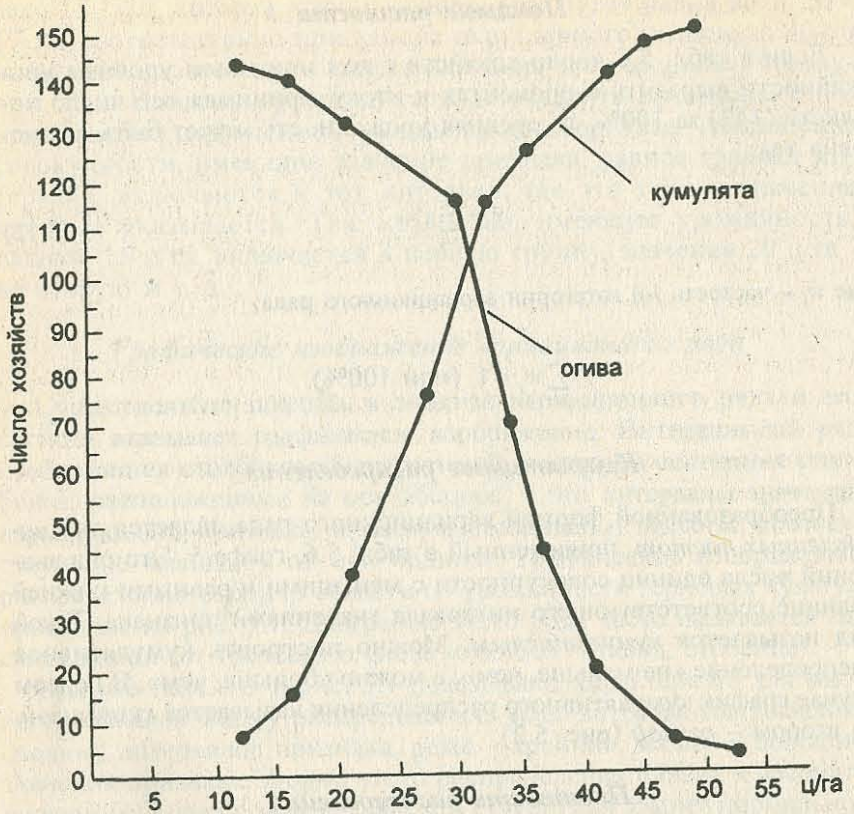


Рис. 5.2. Огива и кумулята распределения по урожайности

5.7. СТРУКТУРНЫЕ ХАРАКТЕРИСТИКИ ВАРИАЦИОННОГО РЯДА

Медиана распределения

При изучении вариации применяются такие характеристики вариационного ряда, которые описывают количественно его структуру, строение. Такова, например, *медиана* – величина варьирующего признака, делящая совокупность на две равные части – со значениями признака меньше медианы и со значениями признака больше медианы (третьего банка из пяти в табл. 5.5, т.е. 196 млрд руб.).

На примере табл. 5.5 видно принципиальное различие между медианой и средней величиной. Медиана не зависит от значений признака на краях ранжированного ряда. Если бы даже капитал крупнейшего банка Санкт-Петербурга был вдесятеро больше, величина медианы не изменилась бы. Поэтому часто медиану используют как более надежный показатель типичного значения признака, нежели арифметическая средняя, если ряд значений неоднороден, включает резкие отклонения от средней. В данном ряду средняя величина собственного капитала, равная 269 млрд руб., сложилась под большим влиянием наибольшей варианты. 80% банков имеют капитал меньше среднего и лишь 20% – больше. Вряд ли такую среднюю можно считать типичной величиной. При четном числе единиц совокупности за медиану принимают арифметическую среднюю величину из двух центральных вариант, например при десяти значениях признака – среднюю из пятого и шестого значений в ранжированном ряду.

В интервальном вариационном ряду для нахождения медианы применяется формула (5.14).

$$Me = x_0 + \frac{\frac{\sum_{j=1}^{k} f_j}{2} - f'_{Me-1}}{f_{Me}} \cdot i, \qquad (5.14)$$

где Me – медиана;
x_0 – нижняя граница интервала, в котором находится медиана;
f'_{Me-1} – накопленная частота в интервале, предшествующем медианному;
f_{Me} – частота в медианном интервале;
i – величина интервала;
k – число групп.

В табл. 5.6 медианным является среднее из 143 значений, т.е. семьдесят второе от начала ряда значение урожайности. Как видно из ряда накопленных частот, оно находится в четвертом интервале. Тогда

$$Me = 25 + \frac{72 - 35}{41} \cdot 5 = 29{,}5 \text{ ц/га}.$$

При нечетном числе единиц совокупности номер медианы, как видим, равен не $\Sigma f_j : 2$, как в формуле (5.14), а $(\Sigma f_j + 1) : 2$, но это различие несущественно и обычно игнорируется на практике.

В дискретном вариационном ряду медианой следует считать значение признака в той группе, в которой накопленная частота превышает половину численности совокупности. Например, для данных табл. 5.1 медианой числа забитых за игру мячей будет 2.

Квартили распределения

Аналогично медиане вычисляются значения признака, делящие совокупность на четыре равные по числу единиц части. Эти величины называются *квартилями* и обозначаются заглавной латинской буквой Q с подписным значком номера квартиля. Ясно, что Q_2 совпадает с Ме. Для первого и третьего квартилей приводим формулы и расчет по данным табл. 5.6.

$$Q_1 = x_0 + \frac{\left(\sum_{j=1}^{k} f_j / 4\right) - f'_{Q_1-1}}{f_{Q_1}} i = 25 + \frac{(35{,}75 - 35)}{41} \cdot 5 = 25{,}09 \text{ ц/га};$$

$$Q_3 = x_0 + \frac{\left(3\sum_{j=1}^{k} f_j / 4\right) - f'_{Q_3-1}}{f_{Q_3}} i = 35 + \frac{(107{,}25 - 102)}{21} \cdot 5 = 36{,}25 \text{ ц/га}.$$

Так как Q_2= Ме = 29,5 ц/га, видно, что различие между первым квартилем и медианой меньше, чем между медианой и третьим квартилем. Этот факт свидетельствует о наличии некоторой несимметричности в средней области распределения, что заметно и на рис. 5.1.

Значения признака, делящие ряд на пять равных частей, называют *квинтилями,* на десять частей – *децилями,* на сто частей – *перцентилями.* Поскольку эти характеристики применяются лишь при необходимости подробного изучения структуры вариационного ряда, приводить их формулы и расчет не будем.

Мода распределения

Бесспорно, важное значение имеет такая величина признака, которая встречается в изучаемом ряду, в совокупности чаще всего. Такую величину принято называть *модой* и обозначать Мо. В дискретном ряду мода определяется без вычисления как значение признака с наибольшей частотой. Например, по данным табл. 5.1 чаще всего за футбольный матч было забито 2 мяча – 71 раз. Модой

является число 2. Обычно встречаются ряды с одним модальным значением признака. Если два или несколько равных (и даже несколько различных, но больших, чем соседние) значений признака имеются в вариационном ряду, он считается соответственно бимодальным («верблюдообразным») либо мультимодальным. Это говорит о неоднородности совокупности, возможно, представляющей собой агрегат нескольких совокупностей с разными модами.

Так и в толпе туристов, приехавших из разных стран, вместо одной, преобладающей среди местных жителей модной одежды можно встретить смесь разных «мод», принятых у разных народов мира.

В интервальном вариационном ряду, тем более при непрерывной вариации признака, строго говоря, каждое значение признака встречается только один раз. Модальным интервалом является интервал с наибольшей частотой. Внутри этого интервала находят условное значение признака, вблизи которого плотность распределения, т.е. число единиц совокупности, приходящееся на единицу измерения варьирующего признака, достигает максимума. Это условное значение и считается *точечной модой.* Логично предположить, что такая точечная мода располагается ближе к той из границ интервала, за которой частота в соседнем интервале больше частоты в интервале за другой границей модального интервала. Отсюда имеем обычно применяемую формулу (5.15):

$$Mo = x_0 + \frac{f_{Mo} - f_{Mo-1}}{(f_{Mo} - f_{Mo-1}) + (f_{Mo} - f_{Mo+1})} \cdot i, \qquad (5.15)$$

где x_0 – нижняя граница модального интервала;
f_{Mo} – частота в модальном интервале;
f_{Mo-1} – частота в предыдущем интервале;
f_{Mo+1} – частота в следующем интервале за модальным;
i – величина интервала.

По данным табл. 5.6 рассчитаем моду:

$$Mo = 25 + \frac{(41 - 20)}{(41 - 20) + (41 - 26)} \cdot 5 = 27{,}9 \text{ ц/га}.$$

Вычисление моды в интервальном ряду весьма условно. Приближенно Мо может быть определена графически (см. рис. 5.1).

К изучению структуры вариационного ряда средняя арифметическая величина тоже имеет отношение, хотя основное значение этого обобщающего показателя другое. В ряду распределения хо-

зяйств по урожайности (табл. 5.6) средняя величина урожайности вычисляется как взвешенная по частоте середина интервалов x (по формуле (5.2)):

$$\bar{x} = \frac{\sum_{j=1}^{k} x'_j f_j}{\sum_{j=1}^{k} f_j} = \frac{4327{,}5}{143} = 30{,}3 \text{ ц/га}.$$

Соотношение между средней величиной, медианой и модой

Различие между средней арифметической величиной, медианой и модой в данном распределении невелико. Если распределение по форме близко к нормальному закону, то медиана находится между модой и средней величиной, причем ближе к средней, чем к моде.

При правосторонней асимметрии $\bar{x} > \text{Me} > \text{Mo}$;

при левосторонней асимметрии $\bar{x} < \text{Me} < \text{Mo}$.

Для умеренно асимметричных распределений справедливо равенство: $|\text{Mo} - \bar{x}| = 3|\text{Me} - \bar{x}|$.

5.8. ПОКАЗАТЕЛИ РАЗМЕРА И ИНТЕНСИВНОСТИ ВАРИАЦИИ

Абсолютные средние размеры вариации

Следующим этапом изучения вариации признака в совокупности является измерение характеристик силы, величины вариации. Простейшим из них может служить *размах* или *амплитуда вариации* – абсолютная разность между максимальным и минимальным значениями признака из имеющихся в изучаемой совокупности значений. Таким образом, размах вариации вычисляется по формуле

$$R = x_{\max} - x_{\min}. \qquad (5.16)$$

Поскольку величина размаха характеризует лишь максимальное различие значений признака, она не может измерять закономерную силу его вариации во всей совокупности. Предназначенный для данной цели показатель должен учитывать и обобщать все различия значений признака в совокупности без исключения. Число таких различий равно числу сочетаний по два из всех единиц совокупности; по данным табл. 5.6 оно составит: $C_{143}^{2} = 10\,153$. Однако нет

необходимости рассматривать, вычислять и осреднять все отклонения. Проще использовать среднюю из отклонений отдельных значений признака от среднего арифметического значения признака, а таковых всего 143. Но среднее отклонение значений признака от средней арифметической величины согласно известному свойству последней равно нулю. Поэтому показателем силы вариации выступает не алгебраическая средняя отклонений, а *средний модуль отклонений:*

$$a = \sum_{i=1}^{n} |x_i - \bar{x}| : n. \qquad (5.17)$$

По данным табл. 5.6 средний модуль, или *среднее линейное отклонение,* по абсолютной величине вычисляется как взвешенное по частоте отклонение по модулю середин интервалов от средней арифметической величины, т.е. по формуле

$$a = \frac{\sum_{j=1}^{k} |x'_j - \bar{x}| f_j}{\sum_{j=1}^{k} f_j} = \frac{980{,}2}{143} = 6{,}85 \text{ ц/га}. \qquad (5.18)$$

Это означает, что в среднем урожайность в изучаемой совокупности хозяйств отклонялась от средней урожайности по области на 6,85 ц/га. Простота расчета и интерпретации составляют положительные стороны данного показателя, однако математические свойства модулей «плохие»: их нельзя поставить в соответствие с каким-либо вероятностным законом, в том числе и с нормальным распределением, параметром которого является не средний модуль отклонений, а *среднее квадратическое отклонение* (в англоязычных программах для ЭВМ называемое «the standard deviation», сокращенно «s.d.» или просто «*s*», в русскоязычных – СКО). В статистической литературе среднее квадратическое отклонение от средней величины принято обозначать малой (строчной) греческой буквой сигма (σ) или s (см. гл. 7):

для ранжированного ряда

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}}, \qquad (5.19)$$

для интервального ряда

$$\sigma = \sqrt{\frac{\sum_{j=1}^{k}(x'_j - \bar{x})^2 f_j}{\sum_{j=1}^{k} f_j}}. \qquad (5.20)$$

По данным табл. 5.6 среднее квадратическое отклонение урожайности зерновых составило:

$$\sigma = \sqrt{\frac{10185}{143}} = 8{,}44 \text{ ц/га}.$$

Следует указать, что некоторое округление средней величины и середин интервалов, например до целых, мало отражается на величине σ, которая составила бы при этом 8,55 ц/га.

Среднее квадратическое отклонение по величине в реальных совокупностях всегда больше среднего модуля отклонений. Соотношение $\sigma : a$ зависит от наличия в совокупностях резких, выделяющихся отклонений и может служить индикатором «засоренности» совокупности неоднородными с основной массой элементами: чем это соотношение больше, тем сильнее подобная «засоренность». Для нормального закона распределения $\sigma : a \approx 1{,}2$.

Понятие дисперсии

Квадрат среднего квадратического отклонения дает величину *дисперсии* σ^2. Формула дисперсии:

простая (для несгруппированных данных):

$$\sigma^2 = \frac{\sum_{(i)}(x_i - \bar{x})^2}{n} = \left(\overline{x^2} - \bar{x}^2\right) \qquad (5.21)$$

или

взвешенная (для сгруппированных данных):

$$\sigma^2 = \frac{\sum_{(j)}(x'_j - \bar{x})^2 f_j}{\sum_{(j)} f_j} = \left(\overline{x^2} - \bar{x}^2\right). \qquad (5.22)$$

На дисперсии основаны практически все методы математической статистики. Большое практическое значение имеет правило сложения дисперсий (см. гл. 6).

Другие меры вариации

Еще одним показателем силы вариации, характеризующим ее не по всей совокупности, а лишь в ее центральной части, служит *среднее квартильное расстояние,* т.е. средняя величина разности между квартилями, обозначаемое далее как *q:*

$$q = \frac{(Q_3 - Me) + (Me - Q_1)}{2} = \frac{Q_3 - Q_1}{2}. \qquad (5.23)$$

Для распределения сельхозпредприятий по урожайности в табл. 5.2 $q = (36{,}25 - 25{,}09) : 2 = 5{,}58$ ц/га. Сила вариации в центральной части совокупности, как правило, меньше, чем в целом по всей совокупности. Соотношение между средним модулем отклонений и средним квартильным отклонением также служит для изучения структуры вариации: большое значение такого соотношения говорит о наличии слабоварьирующего «ядра» и сильно рассеянного вокруг этого ядра окружения, или «гало» в изучаемой совокупности. Для данных табл. 5.6 соотношение $a : q = 1{,}23$, что говорит о небольшом различии силы вариации в центральной части совокупности и на ее периферии.

Для оценки интенсивности вариации и для сравнения ее в разных совокупностях и тем более для разных признаков необходимы *относительные показатели вариации.* Они вычисляются как отношения абсолютных показателей силы вариации, рассмотренных ранее, к средней арифметической величине признака. Получаем следующие показатели:

1) относительный размах вариации ρ:

$$\rho = R : \bar{x}; \qquad (5.24)$$

2) относительное отклонение по модулю *m*:

$$m = a : \bar{x}; \qquad (5.25)$$

3) коэффициент вариации как относительное квадратическое отклонение *v*:

$$v = \sigma : \bar{x}; \qquad (5.26)$$

4) относительное квартильное расстояние d:

$$d = q : \bar{x}, \qquad (5.27)$$

где q - среднее квартильное расстояние.

Для вариации урожайности по данным табл. 5.6 эти показатели составляют:

ρ = 42,4 : 30,3 = 1,4, или 140%;

m = 6,85 : 30,3 = 0,226, или 22,6%;

v = 8,44 : 30,3 = 0,279, или 27,9%;

d = 5,58 : 30,3 = 0,184, или 18,4%.

Оценка степени интенсивности вариации возможна только для каждого отдельного признака и совокупности определенного состава. Так, для совокупности сельхозпредприятий вариация урожайности в одном и том же природном регионе может быть оценена как слабая, если $v < 10\%$, умеренная при $10\% < v < 25\%$ и сильная при $v > 25\%$.

Напротив, вариация роста в совокупности взрослых мужчин или женщин уже при коэффициенте, равном 7%, должна быть оценена и воспринимается людьми как сильная. Таким образом, оценка интенсивности вариации состоит в сравнении наблюдаемой вариации с некоторой обычной ее интенсивностью, принимаемой за норматив. Мы привыкли к тому, что урожайность, заработок или доход на душу, число жилых комнат в здании могут различаться в несколько и даже десятки раз, но различие роста людей хотя бы в полтора раза уже воспринимается как очень сильное.

Различная сила, интенсивность вариации обусловлены объективными причинами. Например, цена продажи доллара США в коммерческих банках Санкт-Петербурга на 24 января 1997 г. варьировала от 5675 до 5640 руб. при средней цене 5664 руб. Относительный размах вариации ρ=35:5664=0,6%. Такая малая вариация вызвана тем, что при значительном различии курса доллара немедленно произошел бы отлив покупателей из «дорогого» банка в более «дешевые». Напротив, цена килограмма картофеля или говядины в разных регионах России варьирует очень сильно – на десятки процентов и более. Это объясняется разными затратами на доставку товара из региона-производителя в регион-потребитель, т.е. пословицей «телушка за морем – полушка, да рубль перевоз».

5.9. МОМЕНТЫ РАСПРЕДЕЛЕНИЯ И ПОКАЗАТЕЛИ ЕГО ФОРМЫ

Центральные моменты распределения

Для дальнейшего изучения характера вариации используются средние значения разных степеней отклонений отдельных величин признака от его средней арифметической величины. Эти показатели получили название *центральных моментов* распределения порядка, соответствующего степени, в которую возводятся отклонения (табл. 5.7), или просто моментов (нецентральные моменты используются редко и здесь не будут рассматриваться). Величина третьего момента μ_3 зависит, как и его знак, от преобладания положительных кубов отклонений над отрицательными кубами либо наоборот. При нормаль-

Таблица 5.7

Центральные моменты

Порядок момента	Формула	
	по несгруппированным данным	по сгруппированным данным
первый (μ_1)	$\frac{\sum_{(i)}(x_i - \bar{x})}{n} = 0$	$\frac{\sum_{(j)}(x'_j - \bar{x})f_j}{\sum_{(j)} f_j} = 0$
второй (μ_2)	$\frac{\sum_{(i)}(x_i - \bar{x})^2}{n} = \overline{x^2} - \bar{x}^2 = \sigma^2$	$\frac{\sum_{(j)}(x'_j - \bar{x})^2 f_j}{\sum_{(j)} f_j} = \overline{x^2} - \bar{x}^2 = \sigma^2$
третий (μ_3)	$\frac{\sum_{(i)}(x_i - \bar{x})^3}{n}$	$\frac{\sum_{(j)}(x'_j - \bar{x})^3 f_j}{\sum_{(j)} f_j}$
четвертый (μ_4)	$\frac{\sum_{(i)}(x_i - \bar{x})^4}{n}$	$\frac{\sum_{(j)}(x'_j - \bar{x})^4 f_j}{\sum_{(j)} f_j}$

ном и любом другом строго симметричном распределении сумма положительных кубов строго равна сумме отрицательных кубов.

Показатели асимметрии

На основе момента третьего порядка можно построить показатель, характеризующий степень асимметричности распределения:

$$As = \frac{\mu_3}{\sigma^3}. \qquad (5.28)$$

As называют *коэффициентом асимметрии.* Он может быть рассчитан как по сгруппированным, так и по несгруппированным данным.

По данным табл. 5.6 показатель асимметрии составил:

$$As = \frac{6796}{8{,}44^3 \cdot 143} = 0{,}079,$$

т.е. асимметрия незначительна. Английский статистик К. Пирсон на основе разности между средней величиной и модой предложил другой показатель асимметрии

$$As_{П} = \frac{\bar{x} - Mo}{\sigma}. \qquad (5.29)$$

По данным табл. 5.6 показатель Пирсона составил:

$$As_{П} = \frac{30{,}3 - 27{,}9}{8{,}44} = 0{,}284.$$

Показатель Пирсона зависит от степени асимметричности в средней части ряда распределения, а показатель асимметрии, основанный на моменте третьего порядка, - от крайних значений признака. Таким образом, в нашем примере в средней части распределения асимметрия более значительна, что видно и по графику (рис. 5.1). Распределения с сильной правосторонней и левосторонней (положительной и отрицательной) асимметрией показаны на рис. 5.3.

Характеристика эксцесса распределения

С помощью момента четвертого порядка характеризуется еще более сложное свойство рядов распределения, чем асимметрия, называемое *эксцессом.*

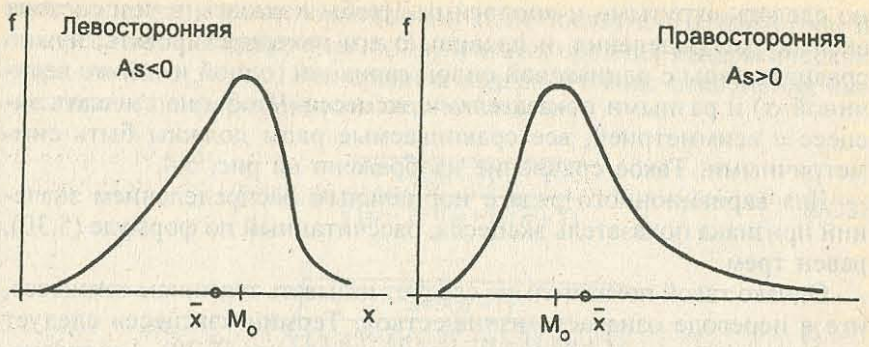


Рис. 5.3. Асимметрия распределения

Показатель эксцесса рассчитывается по формуле

$$Ex = \frac{\mu_4}{\sigma^4}. \qquad (5.30)$$

Часто эксцесс интерпретируется как «крутизна» распределения, но это неточно и неполно. График распределения может выглядеть сколь угодно крутым в зависимости от силы вариации признака: чем слабее вариация, тем круче кривая распределения при данном масштабе. Не говоря уже о том, что, изменяя масштабы по оси абсцисс и по оси ординат, любое распределение можно искусствен-

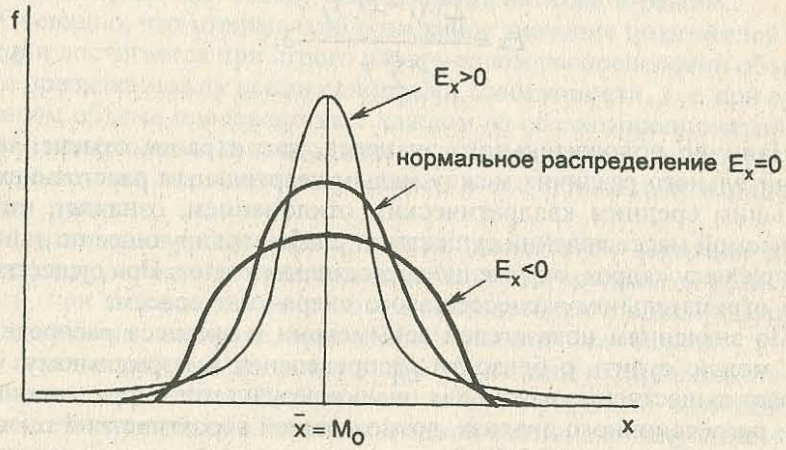


Рис. 5.4. Эксцесс распределения

но сделать «крутым» и «пологим». Чтобы показать, в чем состоит эксцесс распределения, и правильно его интерпретировать, нужно сравнить ряды с одинаковой силой вариации (одной и той же величиной σ) и разными показателями эксцесса. Чтобы не смешать эксцесс с асимметрией, все сравниваемые ряды должны быть симметричными. Такое сравнение изображено на рис. 5.4.

Для вариационного ряда с нормальным распределением значений признака показатель эксцесса, рассчитанный по формуле (5.30), равен трем.

Однако такой показатель не следует называть термином «эксцесс», что в переводе означает «излишество». Термин «эксцесс» следует применять не к самому отношению по формуле (5.30), а к сравнению такого отношения для изучаемого распределения с величиной данного отношения нормального распределения, т.е. с величиной 3. Отсюда окончательные формулы показателя эксцесса, т.е. излишества в сравнении с нормальным распределением при той же силе вариации, имеют вид:

для ранжированного ряда

$$Ex = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^4}{n\sigma^4} - 3, \qquad (5.31)$$

для интервального и дискретного вариационного ряда

$$Ex = \frac{\sum_{j=1}^{n}(x'_j - \bar{x})^4 f_j}{\sigma^4 \sum_{j=1}^{n} f_j} - 3. \qquad (5.32)$$

Наличие положительного эксцесса, как и ранее отмеченного значительного различия между малым квартильным расстоянием и большим средним квадратическим отклонением, означает, что в изучаемой массе явлений существует слабо варьирующее по данному признаку «ядро», окруженное рассеянным «гало». При существенном отрицательном эксцессе такого «ядра» нет совсем.

По значениям показателей асимметрии и эксцесса распределения можно судить о близости распределения к нормальному, что бывает существенно важно для оценки результатов корреляционного и регрессионного анализа, возможностей вероятностной оценки прогнозов (см. главы 7,8,9). Распределение можно считать нормальным, а точнее говоря – не отвергать гипотезу о сходстве фактического распределения с нормальным, если показатели асимметрии и эксцесса не превышают своих двукратных средних квадратических отклонений σ_{as} и σ_{ex}. Эти средние квадратические отклонения вычисляются по формулам:

$$\sigma_{as} = \sqrt{\frac{6(n-1)\cdot n}{(n-2)\cdot(n+1)(n+3)}}; \qquad (5.33)$$

$$\sigma_{ex} = \sqrt{\frac{24n(n-1)^2}{(n-3)\cdot(n-2)\cdot(n+3)\cdot(n+5)}}. \qquad (5.34)$$

5.10. ПРЕДЕЛЬНО ВОЗМОЖНЫЕ ЗНАЧЕНИЯ ПОКАЗАТЕЛЕЙ ВАРИАЦИИ И ИХ ПРИМЕНЕНИЕ

Применяя любой вид статистических показателей, полезно знать, каковы предельно возможные значения данного показателя для изучаемой системы и каково отношение фактически наблюдаемых значений к предельно возможным. Особенно актуальна эта проблема при изучении вариации объемных показателей, таких, как объем производства определенного вида продукции, наличие определенных ресурсов, распределение капиталовложений, доходов, прибыли. Рассмотрим теоретически и практически данный вопрос на примере распределения производства овощей между сельхозпредприятиями в районе.

Очевидно, что минимально возможное значение показателей вариации достигается при строго равномерном распределении объемного признака между всеми единицами совокупности, т. е. при одинаковом объеме производства в каждом из сельхозпредприятий. В таком предельном (конечно, весьма маловероятном на практике) распределении вариация отсутствует и все показатели вариации равны нулю.

Максимально возможное значение показателей вариации достигается при таком распределении объемного признака в совокупности, при котором весь его объем сосредоточен в одной единице совокупности; например, весь объем производства овощей – в одном сельхозпредприятии района при отсутствии их производства в остальных хозяйствах. Вероятность такого предельно возможного сосредоточения объема признака в одной единице совокупности не столь уж мала; во всяком случае она гораздо больше вероятности строго равномерного распределения.

Рассмотрим показатели вариации при указанном предельном случае ее максимальности. Обозначим число единиц совокупности n, среднюю величину признака $\bar{x}$, тогда общий объем признака в совокупности выразится как $\bar{x}n$. Весь этот объем сосредоточен у одной единицы совокупности, так что $x_{max} = \bar{x}n$, $x_{min} = 0$, откуда следует, что максимальное значение амплитуды (размаха вариации) равно:

$$p_{max} = \bar{x}n - 0 = \bar{x}n, \quad \rho_{max} = p_{max} : x = n. \tag{5.35}$$

Для вычисления максимальных значений средних отклонений по модулю и квадратического построим таблицу отклонений (табл. 5.8)[1].

Таблица 5.8

Модули и квадраты отклонений от средней при максимально возможной вариации

Номера единиц совокупности	Значения признака x_i	Отклонения от средней $x_i - \bar{x}$	Модули отклонений $\lvert x_i - \bar{x}\rvert$	Квадраты отклонений $(x_i - \bar{x})^2$
1	$\bar{x}n$	$\bar{x}(n-1)$	$\bar{x}(n-1)$	$\bar{x}^2(n-1)^2$
2	0	$-\bar{x}$	$\bar{x}$	$\bar{x}^2$
3	0	$-\bar{x}$	$\bar{x}$	$\bar{x}^2$
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
n	0	$-\bar{x}$	$\bar{x}$	$\bar{x}^2$
Итого	$\bar{x}n$	0 (нуль)	$2\bar{x}(n-1)$	$\bar{x}^2[(n-1)^2 + (n-1)]$

Исходя из выражений, стоящих в итоговой строке табл. 5.8, получаем следующие максимально возможные значения показателей вариации.

[1] См.: *Кривенкова Л. Н., Юзбашев М. М.* Область существования показателей вариации и ее применение//Вестник статистики. – 1991. – № 6. – С. 66–70.

Средний модуль отклонений, или среднее линейное отклонение:

$$a_{max} = \frac{2\bar{x}(n-1)}{n} = 2\bar{x} - \frac{2}{n}\bar{x}. \tag{5.36}$$

Среднее квадратическое отклонение:

$$\sigma_{max} = \sqrt{\frac{\bar{x}^2[(n-1)^2 + (n-1)]}{2}} = \bar{x}\sqrt{n-1}. \tag{5.37}$$

Относительное модульное (линейное) отклонение:

$$m_{max} = a_{max} : \bar{x} = 2 - \frac{2}{n}. \tag{5.38}$$

Коэффициент вариации:

$$v_{max} = \sigma_{max} : \bar{x} = \sqrt{n-1}. \tag{5.39}$$

Что касается квартильного расстояния, то система с максимально возможной вариацией обладает вырожденной структурой распределения признака, в которой не существуют («не работают») характеристики структуры: медиана, квартили и им подобные.

Исходя из полученных формул максимально возможных значений основных показателей вариации, прежде всего следует вывод о зависимости этих значений от объема совокупности n. Эта зависимость обобщена в табл. 5.9.

Наиболее узкие пределы изменения и слабую зависимость от численности совокупности обнаруживают средний модуль и относительное линейное отклонение. Напротив, среднее квадратическое отклонение и коэффициент вариации сильно зависят от численности единиц совокупности. Эту зависимость следует учитывать при сравнении силы интенсивности вариации в совокупностях разной численности. Если в совокупности шести предприятий коэффициент вариации объема продукции составил 0,58, а в совокупности из 20 предприятий он составил 0,72, то справедливо ли делать вывод о большей неравномерности объема продукции во второй совокупности? Ведь в первой, меньшей, он составил 0,58 : 2,24 = 25,9% максимально возможного, т.е. предельного, уровня концентрации

Таблица 5.9

Предельные значения показателей вариации объемного признака при разных численностях совокупности

Численность совокупностей	Максимальные значения показателей					
	R	ρ	a	m	σ	v
2	$2x$	2	$\bar{x}$	1	$\bar{x}$	1
4	$4x$	4	$1{,}5\bar{x}$	1,5	$1{,}73\bar{x}$	1,73
6	$6x$	6	$1{,}67\bar{x}$	1,67	$2{,}24\bar{x}$	2,24
10	$10x$	10	$1{,}80\bar{x}$	1,80	$3\bar{x}$	3,00
20	$20x$	20	$1{,}90\bar{x}$	1,90	$4{,}36\bar{x}$	4,36
50	$50x$	50	$1{,}96\bar{x}$	1,96	$7\bar{x}$	7,00
100	$100x$	100	$1{,}98\bar{x}$	1,98	$9{,}95\bar{x}$	9,95
∞	∞	∞	$2\bar{x}$	2	∞	∞

производства в одном предприятии из шести, а во второй, большей совокупности, наблюдаемый коэффициент вариации составил только 0,72 : 4,36 = 16,5% максимально возможного.

Имеет практическое значение и такой показатель, как отношение фактического среднего модуля отклонений к предельно возможному. Так, для совокупности шести предприятий это соотношение составило: 0,47 : 1,67 = 0,281, или 28,1%. Интерпретация полученного показателя такова: для перехода от наблюдаемого распределения объема продукции между предприятиями, к равномерному распределению потребовалось бы перераспределить

$\frac{m_{\text{факт}}}{m_{\max}} \cdot \frac{n-1}{n} = 0{,}281 \cdot \frac{6-1}{6} = 0{,}234$, или 23,4% общего объема продукции в совокупности. Если степень фактической концентрации производства (фактическая величина σ или v) составляет некоторую долю предельного значения при монополизации производства на одном предприятии, то отношение фактического показателя к предельному может характеризовать степень концентрации (или монополизации) производства.

Отношения фактических значений показателей вариации или изменения структуры к предельно возможным используются также при анализе структурных сдвигов (см. главу 11).

Рекомендуемая литература к главе 5

1. *Джини К.* Средние величины. – М.: Статистика; 1970.
2. *Кривенкова Л. Н., Юзбашев М. М.* Область существования показателей вариации и ее применение // Вестник статистики. – 1991. – №6. – С. 66-70.
3. *Пасхавер И. С.* Средние величины в статистике. – М.: Статистика. 1979.
4. *Шураков В. В., Дайитбегов Д. М. и др.* Автоматизированное рабочее место статистической обработки данных (Глава 4. Предварительная статистическая обработка данных). – М.: Финансы и статистика, 1990.

6 Глава ГРУППИРОВКА

6.1. ЗНАЧЕНИЕ И СУЩНОСТЬ ГРУППИРОВКИ

Русский статистик Д. П. Журавский (1810 – 1856) очень точно определил статистику как «счет по категориям». Действительно, среди бесконечного разнообразия явлений мы, как правило, улавливаем наличие некоторого конечного числа групп или типов.

Лицо каждого человека неповторимо, и все-таки можно классифицировать лица по типам (скуластое, продолговатое, круглое и т.д.); предприятия образуют группы по формам собственности, характеру производимой продукции, размерам (крупные, средние, мелкие), финансовому положению; государства делятся на группы по уровню экономического развития и т.д. Примеры можно продолжить, но ясно, что какую бы совокупность мы не изучали, она всегда подразделяется на группы. Это обусловлено такими объективными свойствами явлений, как вариация, наличие частных совокупностей (см. гл. 1).

Группировка – это распределение единиц по группам в соответствии со следующим принципом: различия между единицами, отнесенными к одной группе, должны быть меньше, чем между единицами, отнесенными к разным группам.

Группировка лежит в основе всей дальнейшей работы с собранной информацией. На основе группировки рассчитываются сводные показатели по группам, появляется возможность их сравнения, анализа причин различий между группами, изучения взаимосвязей между признаками. Если рассчитать сводные показатели только в целом по совокупности, то мы не сможем уловить ее структуры, роли отдельных групп, их специфики.

Однородность (гомогенность) данных является исходным условием их статистического описания и анализа – вычисления и

интерпретации обобщающих показателей, построения уравнения регрессии, измерения корреляции (см. гл. 8), статистического умозаключения (см. гл. 7).

Таким образом, значение группировки состоит в том, что этот метод обеспечивает обобщение данных, представление их в компактном, обозримом виде. Кроме того, группировка создает основу для последующей сводки и анализа данных.

Для изучения структурных изменений в экономике государственная статистика использует группировку хозяйственных субъектов по формам собственности и организационно-правовым формам, представленную в табл. 6.1.

Сводные показатели для отдельных групп являются типичными и устойчивыми, если, **во-первых**, группировка проведена правильно, **во-вторых,** группы имеют достаточную численность. Первое условие связано с тем, что деление на группы далеко не всегда очевидно. Выполнение второго условия необходимо, так как при достаточно большом числе единиц (не менее 5 единиц в группе) в сводных показателях взаимопогашаются случайные характеристики и проявляются закономерные, типичные.

Для решения задачи группировки нужно установить правила отнесения каждой единицы к той или иной группе.

В эти правила входят определения тех характеристик (признаков), по которым будет проводиться группировка (так называемых *группировочных признаков),* и их значений, отделяющих одну группу от другой *(интервалов группировки).*

Группировка называется *простой* (монотетической), если для ее построения используется один группировочный признак. Если группировка проводится по нескольким признакам, она называется *сложной* (политетической). Обычно такая группировка проводится как *комбинационная,* т.е. группы, выделенные по одному признаку, подразделяются на подгруппы по другому признаку. Казалось бы, этот метод выделения групп должен быть лучше простой группировки – ведь трудно ожидать, что различия между группами можно уловить лишь на основе одного признака. Однако комбинация признаков приводит к дроблению совокупности в геометрической прогрессии: число групп будет равно произведению числа группировочных признаков (l) на число выделенных категорий по каждому из них (m): $k = l \cdot m$. Данные становятся труднообозримыми, группы включают малое число единиц, групповые показатели становятся ненадежными.

Альтернативой является проведение *многомерных группировок* или многомерных классификаций (см. п. 6.3).

Таблица 6.1

Организационно-правовые формы и формы собственности хозяйственных субъектов Российской Федерации

Организационно-правовые формы	Форма собственности							
	государственная	муниципальная	общественных объединений	частная	смешанная (без иностранного участия)	собственность иностранных юридических лиц, граждан и лиц без гражданства	смешанная с российским и иностранным капиталом	
Государственные предприятия								
Муниципальные предприятия								
Акционерные общества и товарищества								
из них:								
акционерные общества открытого типа								
акционерные общества закрытого типа								
товарищества с ограниченной ответственностью								
смешанные товарищества								
полные товарищества								
Колхозы, совхозы, межхозяйственные предприятия (организации)								
Крестьянские (фермерские) хозяйства								
Объединения (союзы, ассоциации) крестьянских (фермерских) хозяйств								
Производственные кооперативы								
Предприятия потребительской кооперации								
Предприятия общественных и религиозных организаций								
Индивидуальные (семейные) частные предприятия (с привлечением наемного труда)								
Другие предприятия								
Другие объединения предприятий								
Некоммерческие организации:								
Государственные учреждения								
Муниципальные учреждения								
Общественные и религиозные организации, их учреждения								
Потребительские общества, учреждения потребительской кооперации								
Фонды								
Другие некоммерческие организации								
Филиалы, представительства и другие обособленные подразделения								

Остановимся на определении интервалов группировочных признаков. Используются интервалы *открытые* и *закрытые*. В первом случае указываются *верхняя* и *нижняя границы интервала*. Например, группы предприятий по численности работников, человек: 200 – 600, 600 – 1000, 1000 – 2000. Такая запись предполагает, что единица, у которой значение признака совпадает с верхней границей интервала, относится к следующей группе, т.е. интервал читается как «от – до».

Иногда границы закрытых интервалов предполагают включение единиц с нижней и верхней границами. Например, группировка населения по возрасту, лет: 0-4,5-9,10-14,15-19,20-24, 25-29 и т.д. Интервал называется открытым, если указана либо только верхняя, либо только нижняя граница: до 200 человек или 2000 человек и более.

Закрытые интервалы подразделяются на *равные* и *неравные*. Как указывалось в гл. 5, величина равного интервала находится по формуле

$$i_x = \frac{x_{max} - x_{min}}{\text{число групп}}.$$

Неравные интервалы могут определяться как *равнонаполненные*. При этом совокупность разделяется на группы равного объема с числом единиц в каждой j-й группе: $n_j = n : m$, где n – общее число единиц; m – число групп. Данные ранжируются, отсчитывается число единиц, составляющих первую группу n_1, затем – вторую n_2 и т.д. Границы интервалов будут соответствовать фактическим значениям признака в каждой группе.

Бывает, что число групп заранее неизвестно и определяется опытным путем на основе перебора вариантов группировки, выявления такого варианта, который наилучшим образом позволяет увидеть различия между группами.

При определении числа групп следует обращать внимание на то, чтобы в одну группу не попало бы свыше половины всех единиц совокупности и в средних группах было больше единиц, чем в крайних.

Если группировочный признак неколичественный, или количественный дискретный с малым числом значений, то группировка данных производится путем подсчета числа единиц с данным значением признака. Примером такой группировки является табл. 6.2.

Таблица 6.2

Группировка станкостроительных заводов по числу производимых типов станков

Число типов станков	1	2	3	4	5 и более
Число заводов	19	10	7	3	1

Очевидно, что метод группировок тесно связан с представлением данных в виде групповых или комбинационных таблиц, а также с графическим представлением структуры совокупности ее частей и соотношений между ними.

6.2. ВИДЫ ГРУППИРОВОК

Группировка производится с целью установления статистических связей и закономерностей, построения описания объекта, выявления структуры изучаемой совокупности. Различия в целевом назначении группировки выражаются в существующей в отечественной статистике классификации группировок: типологические, структурные, аналитические.

Типологическая группировка служит для выделения социально-экономических типов. Этот вид группировок в значительной степени определяется представлениями экспертов о том, какие типы могут встретиться в изучаемой совокупности. Чтобы пояснить особенность этой группировки, остановимся на последовательности действий для ее проведения:

1) называются те типы явлений, которые могут быть выделены;

2) выбираются группировочные признаки, формирующие описание типов;

3) устанавливаются границы интервалов;

4) группировка оформляется в таблицу, выделенные группы (на основе комбинации группировочных признаков) объединяются в намеченные типы, и определяется численность каждого из них.

Рассмотрим **пример**. Поставлена задача выделить типы акционерных компаний с высокими, средними и низкими дивидендами и установить распространенность каждого типа в данном регионе.

Показатель выплаты дивидендов характеризует долю прибыли на акцию или долю чистого дохода, выплачиваемого как дивиденды.

$$\text{Показатель выплаты дивидендов} = \frac{\text{Дивиденды}}{\text{Чистый доход}} \cdot 100\%.$$

Этот коэффициент зависит от структуры акционерного капитала фирмы, длительности существования фирмы и перспектив ее роста. Обычно молодые, быстрорастущие компании выплачивают низкие дивиденды, если вообще их выплачивают; тогда как зрелые компании стремятся дать более высокие дивиденды. Структура капитала и выплата дивидендов зависят от отраслевой принадлежности фирмы. Поэтому при классификации фирм по уровню выплаты дивидендов мы должны использовать в качестве группировочных признаков, во-первых, отрасль (подотрасль), во-вторых, показатель выплаты дивидендов.

Первый группировочный признак выполняет роль характеристики условий, второй непосредственно характеризует тип фирмы. Границы интервалов для второго группировочного признака могут изменяться при переходе от одной отрасли к другой, так как то, что для одной отрасли может рассматриваться как высокий уровень выплаты, для другой может оцениваться иначе.

Изменение границ интервалов группировочного признака при выделении одних и тех же типов в разных условиях называется ***специализацией интервалов группировочного признака.***

Иногда условия формирования типов приводят к различиям в их описании, в самом круге признаков. Например, выделение крупных, средних, мелких предприятий в разных отраслях должно производиться по разным характеристикам: в энергоемких отраслях – по потреблению электроэнергии; в сырьеемких – по величине товарно-материальных запасов; в трудоемких – по численности рабочих; в капиталоемких – по стоимости оборудования.

Изменение круга группировочных признаков при выделении одних и тех же типов в разных условиях называется ***специализацией группировочных признаков.***

Вернемся к нашему примеру. Предположим, что мы располагаем данными 15 фирм, представляющих три подотрасли промышленности. Проведем их группировку с учетом двух выше названных признаков (табл. 6.3).

В табл. 6.3 (гр. В) для краткости использованы условные обозначения типов компаний: н – компании с низким показателем выплаты дивидендов, с – средним, в – высоким показателем выплаты.

Таблица 6.3

Группировка акционерных компаний *n*-гс района по уровню выплаты дивидендов за 19__ г.

Подотрасль промышленности	Показатель выплаты дивидендов, %	Тип компании	Число компаний
А	Б	В	Г
Производство детских игрушек	до 30	н	–
	30 – 50	с	1
	50 и выше	в	4
Производство животного масла	до 20	н	1
	20 – 40	с	2
	40 и выше	в	–
Производство хлопчатобумажных тканей	до 10	н	2
	10 – 30	с	4
	30 и выше	в	1

Использование специализации интервалов как бы уравнивает наши оценки компаний в разных отраслях, что позволяет объединить выделенные группы в три типа независимо от отрасли (табл. 6.4). Это последний шаг типологической группировки.

Как видим, этот метод группировки позволяет избавиться от чрезмерного дробления совокупности, но он слишком субъективен: эксперт определяет, какие типы должны быть выделены, по каким признакам, какими должны быть границы интервалов. К тому же число группировочных признаков ограничено двумя-тремя. Однако, если объект исследования хорошо изучен, если имеется развитая теория, то этот метод может дать хорошо интерпретируемые результаты.

Таблица 6.4

Распределение акционерных компаний *n*-го района по уровню выплаты дивидендов в 19__ г.

Тип компаний	Число компаний	
	абсолютное	в процентах к итогу
н	3	20,0
с	7	46,7
в	5	33,3
Итого	15	100,0

В любом случае правильность проведения типологической группировки требует проверки. С этой целью рассчитываются сводные показатели по группам (средние, относительные величины); если различие между группами статистически незначимо (по *t*-критерию Стьюдента или *F*-критерию, или критерию χ^2 и т.д., см. гл. 7), то схема группировки должна быть пересмотрена – схожие группы могут быть объединены, изменены границы интервалов и т. д.

Структурная группировка характеризует структуру совокупности по какому-либо одному признаку. Пример такой группировки представлен в табл. 6.5.

Таблица 6.5

Распределение населения Российской Федерации по среднедушевому денежному доходу*

(%)

Среднедушевой денежный доход в месяц, тыс. руб.	1995	1996
400 и менее	47,4	25,3
400,1 – 600,0	22,3	22,4
600,1 – 800,0	12,6	16,9
800,1 – 1000,0	7,1	11,5
1000,1 – 1200,0	4,0	7,6
1200,1 – 1600,0	3,9	8,4
1600,1 – 2000,0	1,5	3,9
Свыше 2000,0	1,2	4,0
Всего	100	100

**Источник.* Российский статистический ежегодник. 1997. – М.: Госкомстат России, 1997. – С. 139.

Если для типологической группировки чаще используются открытые и неравные интервалы, то для структурной группировки более характерны закрытые равные интервалы. Структурная группировка позволяет изучать интенсивность вариации группировочного признака (см. гл. 5). На основе структурной группировки можно изучать динамику структуры совокупности.

Если известны структурные характеристики совокупности в одном и другом периодах: w_{i0} и w_{i1} — доли *i*-й группы в период «0» и в период «1», то можно рассчитать показатель среднего абсолютного изменения структуры:

$$d_{w_1-w_0} = \frac{\sum_{i=1}^{k} |w_{1i} - w_{0i}|}{k}, \qquad (6.1)$$

где k — число групп.

Другой сводный показатель абсолютных структурных сдвигов строится на основе формулы среднего квадратического отклонения:

$$s_{w_1-w_0} = \sqrt{\frac{\sum_{i=1}^{k} (w_{1i} - w_{0i})^2}{k}}. \qquad (6.2)$$

Если показатели структуры выразить не в долях, а в процентах, то, так же как и первый показатель, квадратичный коэффициент абсолютных структурных сдвигов оценивает на сколько процентных пунктов в среднем различаются удельные веса отдельных групп сравниваемых структур. При отсутствии структурных сдвигов оба эти показателя равны нулю; их величина тем больше, чем значительнее абсолютные изменения удельных весов групп. Квадратичный коэффициент более чутко реагирует на структурные изменения. Существуют и другие показатели для измерения структурных сдвигов (см., например, индекс структуры в гл. 10). При сравнениях предполагается, что число групп в одном и другом периодах остается одним и тем же. По данным табл. 6.5, $d_{w_1-w_0} = 5{,}5$; $s_{w_1-w_0} = 8{,}5$ процентных пункта.

Деление группировок на типологические и структурные достаточно условно. Если задать, например, границы среднедушевого дохода, соответствующие определенным типам благосостояния, то можно с полным правом назвать группировку типологической.

Аналитическая группировка характеризует взаимосвязь между двумя и более признаками, из которых один рассматривается как результат, другой (другие) — как фактор (факторы).

Пример однофакторной аналитической группировки представлен в табл. 6.6.

В данном примере оборачиваемость в днях – фактор, обозначенный x, прибыль – результат – y. Очевидно, что при одном и том же сроке оборота предприятия могут иметь разную прибыль. Чтобы установить связь между признаками, данные группируются по признаку-фактору. Затем по каждой группе рассчитывается среднее значение результата. По обобщенным данным гораздо легче увидеть, есть связь между признаками или нет, прямая связь

Таблица 6.6

Характеристика зависимости прибыли малых предприятий от оборачиваемости оборотных средств за 19__ г.

Продолжительность оборота средств в днях	Число малых предприятий	Середина интервала, дни	Средняя прибыль, млн руб.	Изменение средней прибыли, млн руб.
x	n_j	x'_j	$\bar{y}_j$	$\bar{y}_j - \bar{y}_{j-1}$
20 – 30	6	25	14,57	–
31 – 50	8	40	12,95	– 1,62
51 – 80	6	65	7,40	– 5,55
Итого	20	43	11,77	×

или обратная, линейная или нелинейная. Эти выводы делаются через сопоставление изменений средних значений результата по группам с изменениями фактора. Чтобы эти изменения были сравнимыми, следует проводить группировку с равными интервалами или рассчитывать изменения результата на единицу изменений фактора.

В примере средняя прибыль изменяется от группы к группе, следовательно, связь между оборачиваемостью и прибылью существует, причем обратная: чем медленнее оборачиваются оборотные средства, тем меньше прибыль.

Рассчитаем, насколько снижается прибыль при замедлении оборачиваемости от 40–50 до 51–70 дней и при замедлении оборачиваемости от 51–70 до 71–101 дня:

$$1)\ b_{yx} = \frac{\bar{y}_2 - \bar{y}_1}{x'_2 - x'_1} = \frac{\Delta\bar{y}}{\Delta x'} = \frac{-1{,}62}{15} = -0{,}108 \text{ млн руб./день},$$

$$2)\ b_{yx} = \frac{\bar{y}_3 - \bar{y}_2}{x'_3 - x'_2} = \frac{\Delta\bar{y}}{\Delta x'} = \frac{-5{,}55}{25} = -0{,}222 \text{ млн руб./день}.$$

Полученные значения показывают величину снижения прибыли при замедлении оборачиваемости на 1 день. Такие показатели называются показателями *силы связи.* Различие в их значениях свидетельствует, что сила влияния оборачиваемости на прибыль не явля-

ется постоянной – она возрастает при сроках оборачиваемости свыше 50 дней, т.е. связь признаков нелинейная.

В случае линейной связи важным показателем является характеристика *средней силы связи:*

$$b_{yx} = \frac{\bar{y}_m - \bar{y}_1}{x'_m - x'_1}, \qquad (6.3)$$

где $\bar{y}_m$, $\bar{y}_1$ – средние значения результативного признака в последней и первой группах соответственно; x'_m, x'_1 – середины интервалов (или средние значения) факторного признака в последней и первой группах. В случае прямой связи $b_{yx} > 0$, обратной – $b_{yx} < 0$.

По данным табл. 6.5

$$b_{yx} = \frac{7{,}4 - 14{,}57}{65 - 25} = -0{,}175 \text{ млн руб. /день.}$$

Для нелинейной связи показатель средней силы связи не имеет значения (или имеет ограниченное значение).

По аналитической группировке можно измерить связь с помощью еще одного показателя: эмпирического корреляционного отношения. Этот показатель обозначается греческой буквой η (эта). Он основан на правиле разложения дисперсии, согласно которому *общая дисперсия s^2_y равна сумме внутригрупповой и межгрупповой дисперсий.*

Дисперсия результативного признака внутри группы при относительном постоянстве признака-фактора возникает за счет других факторов (не связанных с изучаемым). Эта дисперсия называется *остаточной* (та колеблемость, которая осталась при закреплении изучаемого фактора x). Она определяется по формуле:

$$s^2_{yx} = \frac{\sum_1^{n_j} (y_{ij} - \bar{y}_j)^2}{n_j}, \qquad (6.4)$$

где y_{ij} – значение признака y для i-й единицы в j-й группе;
$\bar{y}_j$ – среднее значение признака y в j-й группе;
n_j – число единиц в j-й группе;
j = 1, 2, 3, ..., m.

Внутригрупповые дисперсии, рассчитанные для отдельных групп, объединяются в средней величине внутригрупповой дисперсии:

$$\overline{s^2_{yx}} = \frac{\sum_1^m s^2_{yx} n_j}{\sum_1^m n_j}. \qquad (6.5)$$

Межгрупповая дисперсия относится на счет изучаемого фактора (и факторов, связанных с ним), поэтому эта дисперсия называется *факторной.* Она определяется по формуле

$$s^2_{\bar{y}x} = \frac{\sum_1^m (\bar{y}_j - \bar{y})^2 n_j}{\sum_1^m n_j}. \qquad (6.6)$$

Правило сложения дисперсий может быть записано:

$$s^2_y = \overline{s^2_{yx}} + s^2_{\bar{y}x} \qquad (6.7)$$

или

$$\sum_{(j)}\sum_{(i)} (y_{ij} - \bar{y})^2 = \sum_{(j)}\sum_{(i)} (y_{ij} - \bar{y}_j)^2 + \sum_{(j)} (\bar{y}_j - \bar{y})^2 n_j. \qquad (6.8)$$

Эмпирическое корреляционное отношение измеряет, какую часть общей колеблемости результативного признака вызывает изучаемый фактор. Соответственно оно рассчитывается как отношение факторной дисперсии к общей дисперсии результативного признака:

$$\eta^2 = \frac{s^2_{\bar{y}x}}{s^2_y} \text{ – коэффициент детерминации,} \qquad (6.9)$$

$$\eta = \sqrt{\frac{s^2_{\bar{y}x}}{s^2_y}} \text{ – эмпирическое корреляционное отношение.} \qquad (6.10)$$

Этот показатель принимает значения в интервале [0,1]: чем ближе к 1, тем теснее связь, и наоборот.

По данным табл. 6.6 этот показатель равен η = 0,881, связь тесная.

Для изучения влияния нескольких факторов на результат проводится *многофакторная аналитическая группировка.* Она строится как комбинационная группировка по признакам-факторам, и для

каждой подгруппы рассчитывается среднее значение результативного признака.

Обратимся к рассмотренному выше примеру, который дополним вторым фактором формирования прибыли – величиной запаса оборотных средств (z); по этому фактору выделены три группы (табл. 6.7).

Таблица 6.7

Характеристика зависимости прибыли малых предприятий от величины запаса и оборачиваемости оборотных средств за 19__ г.

Средний запас оборотных средств, млн руб.	Оборачиваемость в днях	Число предприятий	Средняя прибыль, млн руб.	Колеблемость прибыли по группам
z_k	x_j	n_{jk}	y_{jk}	$(\bar{y}_{jk} - \bar{y})^2 . n_{jk}$
55 – 85	20 – 30	1	11,00	0,5929
	31 – 50	2	10,85	1,6928
	51 – 80	1	7,05	22,2784
85 – 115	20 – 30	2	11,85	0,0128
	31 – 50	4	11,90	0,0676
	51 – 80	2	5,75	72,4808
115 – 145	20 – 30	3	17,60	101,9667
	31 – 50	2	17,00	54,7058
	51 – 80	3	8,62	29,7675
Итого	×	20	11,77	283,5653

Эта группировка позволяет проследить колеблемость прибыли под влиянием двух факторов. Конечно, уверенность нашего заключения в том, что прибыль изменяется от группы к группе именно за счет изменений запаса оборотных средств и скорости их обращения, зависит от того, насколько обеспечено погашение влияния прочих факторов, т.е. от числа единиц в подгруппах (n_{jk}). В данном примере наполненность групп недостаточна для того, чтобы выявить «чистое» влияние изучаемых факторов.

При $n_{jk} \geq 5$ многофакторная аналитическая группировка позволяет измерить силу связи между результатом и одним из факторов при постоянстве второго фактора, т.е. получить так называемые *частные* (или чистые) показатели силы связи.

По данным табл. 6.7 рассчитаны показатели силы связи между прибылью и оборачиваемостью при закреплении уровня запасов оборотных средств. Таких показателей три (по числу групп по фактору z):

$$b_{yx,z_1} = \frac{7,05 - 11,0}{65 - 25} = -0,10 \text{ млн руб./день},$$

$$b_{yx,z_2} = \frac{5,75 - 11,85}{65 - 25} = -0,15 \text{ млн руб./день},$$

$$b_{yx,z_3} = \frac{8,62 - 17,60}{65 - 25} = -0,22 \text{ млн руб./день}.$$

Точно так же могут быть вычислены показатели силы связи между прибылью и запасом оборотных средств при закреплении оборачиваемости:

$$b_{yz,x_1} = \frac{17,6 - 11,0}{130 - 70} = 0,11 \text{ млн руб./млн руб.},$$

$$b_{yz,x_2} = \frac{17,0 - 10,85}{130 - 70} = 0,10 \text{ млн руб./млн руб.},$$

$$b_{yz,x_3} = \frac{8,62 - 7,05}{130 - 70} = 0,03 \text{ млн руб./млн руб.}$$

Чистое влияние первого фактора (обратное) возрастает при увеличении уровня закрепленного фактора (величины запасов оборотных средств), а второго – снижается.

Можно рассчитать и показатель множественной тесноты связи – совокупное эмпирическое корреляционное отношение.

Для трех признаков, как в нашем примере, его формула следующая:

$$\eta_{yxz} = \sqrt{\frac{\delta^2_{\bar{y}xz}}{\delta^2_y}}. \qquad (6.11)$$

Так же как и показатель парной связи, η_{yxz} принимает значение в интервале [0,1]. В числителе подкоренного выражения находится факторная дисперсия результативного признака:

$$\delta^2_{\bar{y}xz} = \frac{\sum\limits_{(k)}\sum\limits_{(j)}(\bar{y}_{jk} - \bar{y})^2 n_{jk}}{\sum\limits_{(k)}\sum\limits_{(j)} n_{jk}}. \qquad (6.12)$$

По данным табл. 6.7 $\eta_{yxz} = \sqrt{283{,}5653 / 20 : 15{,}38} = \sqrt{14{,}18 : 15{,}38} = 0{,}96$.

Можно с некоторыми оговорками заключить, что на 92% ($0{,}96^2$) вариация прибыли в этой совокупности предприятий определяется вариацией изучаемых факторов.

Многофакторная аналитическая группировка — очень гибкий прием изучения связей. Она позволяет уловить влияние факторов на результат с изменением условий (закреплением прочих факторов на разных уровнях).

Однако при всех отмеченных плюсах этот метод имеет огромный минус – дробление совокупности, в результате чего выделяются подгруппы с малым числом единиц. В этом случае средние значения результативного признака неустойчивы, не достигается погашение прочих факторов, соответственно, ненадежными становятся и показатели связи. Но если совокупность большого объема и распределение признаков-факторов не являются крайне асимметричными, этот метод, как никакой другой, позволяет получить много информации об отношениях между переменными.

В какой-то мере избежать дробления данных и при этом получить «чистые» характеристики связей между переменными позволяет применение *метода стандартизации распределений* в комбинационной таблице. Если в группах по одной переменной, скажем, по z в табл. 6.7, распределение по другой переменной x принять стандартным и на его основе рассчитать групповые средние величины результативного признака, то они будут отличаться за счет принадлежности к разным группам по признаку z при элиминировании признака x. В качестве стандартного применяется распределение в целом по совокупности. Так, по данным табл. 6.7 стандартное распределение по x следующее:

x_1 – 6 ед., x_2 – 8 ед., x_3 – 6 ед. или в относительном выражении: 0,3; 0,4; 0,3. Тогда средняя прибыль при заданном значении переменной z при стандартизации распределения по переменной x равна:

в первой группе:

$$\bar{y}_{1\text{станд}} = \sum_j \bar{y}_{j1} w_j = 11{,}0 \cdot 0{,}3 + 10{,}85 \cdot 0{,}4 + 7{,}05 \cdot 0{,}3 = 9{,}75;$$

во второй группе:

$$\bar{y}_{2\text{станд}} = \sum_j \bar{y}_{j2} w_j = 11{,}85 \cdot 0{,}3 + 11{,}9 \cdot 0{,}4 + 5{,}75 \cdot 0{,}3 = 10{,}04;$$

в третьей группе:

$$\bar{y}_{3\text{станд}} = \sum_j \bar{y}_{j3} w_j = 11{,}60 \cdot 0{,}3 + 17{,}0 \cdot 0{,}4 + 8{,}62 \cdot 0{,}3 = 12{,}76.$$

На основе полученных стандартизованных средних можно рассчитать показатели «чистой» связи между величиной прибыли и средним запасом оборотных средств. Попробуйте сделать такой расчет. Стандартизация распределения по переменной z, расчет стандартизованных средних результативного признака и показателей «чистой» связи между y и x при элиминировании z проводится аналогично. Заметим, что рассмотренные приемы анализа не входят пока в ППП для ЭВМ. Возможно, это сделает кто-то из вас.

6.3. МНОГОМЕРНЫЕ ГРУППИРОВКИ

Мы убедились, как трудно выбрать какой-то один признак в качестве основания группировки. Еще труднее проводить группировку по нескольким признакам. Комбинация двух признаков позволяет сохранить обозримость таблицы, но комбинация трех или четырех признаков дает совершенно неудовлетворительный результат: ведь даже при выделении трех категорий по каждому из группировочных признаков мы получим 9 или 12 подгрупп. Равномерность распределения единиц по группам в принципе невозможна. Вот и получаются группы, в которые входят 1 - 2 наблюдения. Сохранить сложность описания групп и вместе с тем преодолеть недостатки комбинационной группировки позволяют ***методы многомерных группировок.*** Часто их называют ***методами многомерной классификации.***

Эти методы получили распространение благодаря использованию ЭВМ и пакетов прикладных программ. Цель этих методов — классификация данных, иначе говоря, группировка на основе множества признаков. Такие задачи широко распространены в науках о природе и обществе, в практической деятельности по управлению массовыми процессами. Например, выделение типов предприятий по финансовому положению, по экономической эффективности деятельности производится на основе множества признаков: выделение и

изучение типов людей по степени их пригодности к определенной профессии (профпригодность); диагностика болезней на основании множества объективных признаков (симптомов) и т. д.

Простейшим вариантом многомерной классификации является группировка на основе многомерных средних.

Многомерной средней называется средняя величина нескольких признаков для одной единицы совокупности. Поскольку нельзя рассчитать среднюю величину абсолютных значений разных признаков выраженных в разных единицах измерения, то многомерная средняя вычисляется из относительных величин, как правило, – из отношений значений признаков для единицы совокупности к средним значениям этих признаков:

$$\overline{p}_i = \frac{\sum_{j=1}^{k} p_{ij}}{k} = \sum_{j=1}^{k}\left(\frac{x_{ij}}{\overline{x}_j}\right) : k, \qquad (6.13)$$

где $\overline{p}_j$ – многомерная средняя для i-единицы;
x_{ij} – значение признака x_j для i-единицы;
x_j – среднее значение признака x_j;
k – число признаков;
j – номер признака;
i – номер единицы совокупности.

Рассмотрим использование многомерных средних на примере сельскохозяйственных предприятий Всеволожского района Ленинградской области за 1995 г. (табл. 6.8). По каждому предприятию приведены четыре признака:

- среднемесячная оплата труда работника, руб., x_1;
- валовой доход на 1 га сельхозугодий, тыс. руб./га, x_2;
- среднегодовая стоимость основных производственных фондов на 1 га сельхозугодий, млн руб./га, x_3;
- отношение дебиторской задолженности к кредиторской задолженности, %, x_4.

Эти признаки можно считать однородными, так как бóльшая их величина положительно характеризует экономику предприятия. Предпочтительнее обобщать в многомерной средней признаки либо «положительные», либо «отрицательные» (чем больше, тем хуже).

Многомерные средние, приведенные в последней графе табл. 6.8, обобщают четыре признака. При этом значимость признаков для оценки предприятия полагается одинаковой, что, конечно, спорно. Можно усложнить методику, приписав признакам, на основе экс-

Таблица 6.8

Характеристики предприятий Всеволожского района Ленинградской области в 1995 г.

Предприятия	Значения признаков				В % к средней				Многомерная средняя, %
	x_1	x_2	x_3	x_4	x_1	x_2	x_3	x_4	
«Ручьи»	597	390	20,6	72	148	199	106	107	140
«Бугры»	353	96	12,1	30	88	49	62	45	61
«Пригородное»	403	84	20,6	26	100	43	106	39	72
«Авлога»	231	71	15,1	74	57	36	78	110	70
«Всеволожское»	330	114	14,8	159	82	58	76	237	113
«Выборгское»	540	235	24,0	26	134	120	184	39	104
«Приневское»	372	461	33,2	85	93	235	171	127	156
«Шеглово»	393	113	15,0	62	98	58	77	92	81
Средние величины	402	196	19,4	67	100	100	100	100	100
Средние квадратические отклонения	109	142	6,4	41	–	–	–	–	–

пертной оценки, разные веса, и вычислить взвешенные многомерные средние.

Судя по полученным оценкам, предприятия делятся на группы с *многомерными средними ниже* 100% (*четыре предприятия*), несколько выше 100% (два предприятия) и резко превышающие 100% (два предприятия).

При большом объеме совокупности для выделения групп на основе многомерной средней необходимо установить интервалы значений многомерной средней:

$$i_{\overline{p}} = \frac{\overline{p}_{\max} - \overline{p}_{\min}}{\text{число групп}}.$$

Затем следует провести группировку единиц: определить их количество в каждой группе и постараться указать, в чем состоят *качественные различия между группами*.

Более обоснованным методом многомерной классификации является *кластерный анализ*. Само название метода происходит от того

же корня, что и слово «класс», «классификация». Английское слово the cluster имеет значения: группа, пучок, куст, т. е. объединение каких-то однородных явлений. В данном контексте оно близко к математическому понятию «множества», причем, как и множество, кластер может содержать только одно явление, но не может в отличие от множества быть пустым.

Каждая единица совокупности в кластерном анализе рассматривается как точка в заданном признаковом пространстве. Значение каждого из признаков у данной единицы служит ее координатой в этом «пространстве» по аналогии с координатами точки в нашем реальном трехмерном пространстве. Таким образом, *признаковое пространство* – это область варьирования всех признаков совокупности изучаемых явлений. Если мы уподобим это пространство обычному пространству, имеющему евклидову метрику, то тем самым мы получим возможность измерять «расстояния» между точками признакового пространства. Эти расстояния называют *евклидовыми*. Их вычисляют по тем же правилам, как и в обычной евклидовой геометрии. На плоскости, т.е. в двухмерном пространстве, расстояние между точками A и B равно корню квадратному из суммы квадратов разностей координат этих точек по оси абсцисс и по оси ординат – на основании теоремы Пифагора (рис. 6.1).

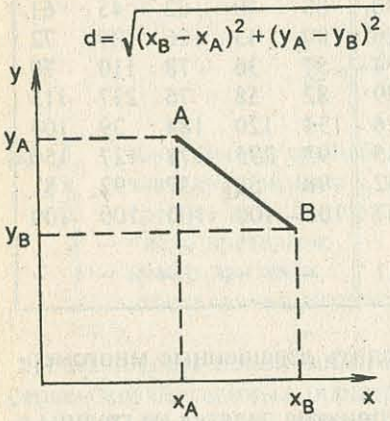


Рис. 6.1. Евклидово расстояние

В многомерном признаковом пространстве расстояние между точками p и q с k координатами, т. е. индивидуальными значениями k признаков, определяется так:

$$d_{p,q}=\sqrt{\sum_{j=1}^{k}(x_{jp}-x_{jq})^2}. \qquad (6.14)$$

Совершенно очевидно, что нельзя суммировать квадраты отклонений одной точки от другой в абсолютных значениях разнокачественных признаков. Необходимо сначала выразить различия между единицами совокупности по каждому признаку в каком-то относительно безразмерном показателе. В качестве такого показателя часто применяют «нормированную разность», т. е. величину:

$$d_{j_{p,q}}=\frac{x_{jp}-x_{jq}}{\sigma_{xj}}. \qquad (6.15)$$

где $x_{jp}-x_{jq}$ – абсолютная разность значений j-го признака у единиц совокупности с номерами p и q;

σ_{xj} – среднее квадратическое отклонение признака x_j.

Знаки нормированных разностей не имеют значения, так как «расстояние» в признаковом пространстве – скалярная, а не векторная величина.

По данным табл. 6.8 среднее квадратическое отклонение признака x_1 равно 109. Разделив все попарные разности значений этого признака на 109, получим матрицу нормированных разностей D_1 (табл. 6.9). Очевидно эта матрица размером $n\times n$ симметрична.

Таблица 6.9

Матрица нормированных разностей между предприятиями по среднемесячной оплате труда (D_1)

Предприятия	«Ручьи»	«Бугры»	«Пригородное»	«Авлога»	«Всеволожское»	«Выборгское»	«Приневское»	«Щеглово»
«Ручьи»	0							
«Бугры»	2,239	0						
«Пригородное»	1,780	0,459	0					
«Авлога»	3,358	1,119	1,578	0				
«Всеволожское»	2,450	0,211	0,670	0,908	0			
«Выборгское»	0,523	1,716	1,257	2,835	1,927	0		
«Приневское»	2,064	0,174	0,284	1,294	0,385	1,541	0	
«Щеглово»	1,872	0,367	0,092	1,486	0,518	1,349	0,193	0

Из данных табл. 6.9 видно, что величина нормированных разностей по этому признаку варьирует от 0 до 3,4. В нормально распределенной совокупности различия признака в среднем лишь в трех случаях из тысячи превосходят шесть сигм, т. е. в распределениях, близких к нормальным, величина нормированного расстояния редко превосходит 6.

Средняя нормированная разность по данным табл. 6.9 составила 1,182. В нормально распределенной совокупности $\overline{d}_{q,p}$ совпадает со средним отклонением их от средней величины, т. е. нормированная разность в нормальной совокупности в среднем равна единице. Это очень важно при установлении предельного (критического) расстояния в признаковом пространстве, при достижении которого прекращается объединение кластеров.

Аналогично вычисляются матрицы нормированных разностей по признакам x_2, x_3, x_4 (см. табл. 6.10–6.12).

Таблица 6.10

Матрица нормированных разностей между предприятиями по валовому доходу на 1 га сельхозугодий (D_2)

Предприятия	«Ручьи»	«Бугры»	«Пригородное»	«Авлога»	«Всеволожское»	«Выборгское»	«Приневское»	«Щеглово»
«Ручьи»	0							
«Бугры»	2,070	0						
«Пригородное»	2,155	0,085	0					
«Авлога»	2,246	0,176	0,092	0				
«Всеволожское»	1,944	0,127	0,211	0,303	0			
«Выборгское»	1,092	0,972	1,063	1,155	0,852	0		
«Приневское»	0,500	2,552	2,636	2,746	2,444	1,592	0	
«Щеглово»	1,951	0,119	0,203	0,296	0,007	0,859	2,451	0

Таблица 6.11

Матрица нормированных разностей между предприятиями по среднегодовой стоимости основных производственных фондов на 1 га сельхозугодий (D_3)

Предприятия	«Ручьи»	«Бугры»	«Пригородное»	«Авлога»	«Всеволожское»	«Выборгское»	«Приневское»	«Щеглово»
«Ручьи»	0							
«Бугры»	1,328	0						
«Пригородное»	0	1,328	0					
«Авлога»	0,859	0,469	0,859	0				
«Всеволожское»	0,906	0,422	0,906	0,047	0			
«Выборгское»	0,531	1,859	0,531	1,391	1,438	0		
«Приневское»	1,969	3,297	1,906	2,828	2,875	1,438	0	
«Щеглово»	0,875	0,453	0,875	0,016	0,031	1,406	2,844	0

Средняя нормированная разность $\overline{d}_3 = 1{,}11$.

Таблица 6.12

Матрица нормированных разностей между предприятиями по отношению дебиторской задолженности к кредиторской (D_4)

Предприятия	«Ручьи»	«Бугры»	«Пригородное»	«Авлога»	«Всеволожское»	«Выборгское»	«Приневское»	«Щеглово»
«Ручьи»	0							
«Бугры»	1,024	0						
«Пригородное»	1,122	0,098	0					
«Авлога»	0,049	1,073	1,171	0				
«Всеволожское»	2,122	3,146	3,244	2,073	0			
«Выборгское»	1,122	0,098	0	1,171	3,244	0		
«Приневское»	0,317	1,341	1,439	0,268	1,805	1,439	0	
«Щеглово»	0,244	0,780	0,878	0,293	2,366	0,878	0,561	0

Средняя нормированная разность $\overline{d}_4 = 1{,}086$.

На основе данных таблиц 6.9 – 6.12 формируется матрица евклидовых расстояний D (табл. 6.13).

С учетом нормировки разности признаков расстояние между двумя любыми единицами совокупности (точками в признаковом пространстве) имеет вид:

$$d_{p,q} = \sqrt{\sum_{j=1}^{k} d_{jp,q}^2}. \qquad (6.16)$$

Например, расстояние между предприятиями «Ручьи» и «Бугры», согласно формуле (6.16), составляет:

$$d_{1,2} = \sqrt{2{,}239^2 + 2{,}070^2 + 1{,}328^2 + 1{,}024^2} = 3{,}480 \text{ и т. д.}$$

Таблица 6.13

Матрица нормализованных (нормированных) евклидовых расстояний (D)

Предприятия	«Ручьи»	«Бугры»	«Пригородное»	«Авлога»	«Всеволожское»	«Выборгское»	«Приневское»	«Щеглово»
«Ручьи»	0							
«Бугры»	3,480	0						
«Пригородное»	3,012	1,411	0					
«Авлога»	4,130	1,629	2,885	0				
«Всеволожское»	3,887	3,184	3,441	2,284	0			
«Выборгское»	1,734	2,712	1,373	3,559	4,127	0		
«Приневское»	2,913	4,383	3,568	4,157	4,188	3,008	0	
«Щеглово»	2,852	0,981	1,130	1,541	2,422	2,281	3,775	0

Матрица евклидовых расстояний D служит основой *агломеративно-иерархического метода классификации*, который заключается в последовательном объединении группируемых объектов – сначала самых близких, а затем все более удаленных друг от друга. Процедура классификации состоит из последовательных шагов, на каждом из которых производится объединение двух ближайших групп объектов (кластеров). На нулевом шаге каждый

объект рассматривается как отдельный кластер. На первом шаге объединим в кластер предприятия с наименьшим евклидовым расстоянием («Бугры» и «Щеглово»). Найдем средние по всем признакам для этого кластера и евклидовы расстояния от кластера до других предприятий (табл. 6.14).

Таблица 6.14

Нормированные разности и евклидовы расстояния для кластера «Бугры + Щеглово»

Предприятия	Признаки				Евклидово расстояние
	x_1	x_2	x_3	x_4	
Средние величины по кластеру	373	104,5	13,55	46	–
«Ручьи»	2,055	2,011	1,094	0,634	3,141
«Пригородное»	0,275	0,144	1,094	0,488	1,237
«Авлога»	1,303	0,236	0,242	0,683	1,509
«Всеволожское»	0,394	0,667	0,195	2,756	2,869
«Выборгское»	1,532	0,919	1,633	0,488	2,469
«Приневское»	0,003	2,511	3,070	0,951	4,079

Заменив в матрице евклидовых расстояний (табл. 6.13) расстояния предприятий, вошедших в первый кластер, на числа последней графы табл. 6.14, видим, что теперь минимальным является расстояние между предприятием «Пригородное» и первым кластером: d =1,237 (табл. 6.15).

Следовательно, на втором шаге к первому кластеру присоединяется предприятие «Пригородное». Вычисляем средние величины, нормированные разности по каждому признаку и евклидовы расстояния от кластера, включающего три предприятия («Бугры», «Щеглово», «Пригородное») до каждого из оставшихся предприятий. Результаты представлены в табл. 6.16.

Заменив евклидовы расстояния предприятий, вошедших в кластер, данными последней графы табл. 6.16, получим новую матрицу евклидовых расстояний (табл. 6.17).

Минимальным является евклидово расстояние от кластера до предприятия «Авлога». На третьем шаге образуем кластер «Бугры + Щеглово + Пригородное + Авлога». Полученные средние величины для кластера, нормированные разности и евклидовы расстояния представлены в табл. 6.18, 6.19.

Таблица 6.15

Матрица евклидовых расстояний после образования кластера «Бугры + Щеглово»

Предприятия	«Бугры + Щеглово»	«Ручьи»	«Пригородное»	«Авлога»	«Всеволожское»	«Выборгское»	«Приневское»
Кластер «Бугры + Щеглово»	0						
«Ручьи»	3,141	0					
«Пригородное»	1,237	3,012	0				
«Авлога»	1,509	4,130	2,885	0			
«Всеволожское»	2,869	3,887	3,441	2,284	0		
«Выборгское»	2,469	1,734	1,731	3,559	4,127	0	
«Приневское»	4,079	2,913	3,568	4,157	4,183	3,008	0

Таблица 6.16

Нормированные разности и евклидовы расстояния для кластера «Бугры + Щеглово + Пригородное»

Предприятия	Признаки				Евклидово расстояние
	x_1	x_2	x_3	x_4	
Средние величины по кластеру	383	98	15,9	39	0
«Ручьи»	1,963	2,056	0,734	0,805	3,044
«Авлога»	1,394	0,190	0,125	0,854	1,651
«Всеволожское»	0,486	0,113	0,172	2,927	2,974
«Выборгское»	1,440	0,965	1,266	0,317	2,170
«Приневское»	0,101	2,556	2,703	1,122	3,887

Таблица 6.17

Матрица евклидовых расстояний после образования кластера «Бугры + Щеглово + Пригородное»

Предприятия	Кластер Б+Щ+П	«Ручьи»	«Авлога»	«Всеволожское»	«Выборгское»	«Приневское»
Кластер Б+Щ+П	0					
«Ручьи»	3,044	0				
«Авлога»	1,651	4,130	0			
«Всеволожское»	2,974	3,887	2,884	0		
«Выборгское»	2,170	1,734	3,559	4,127	0	
«Приневское»	3,887	2,913	4,157	4,188	3,008	0

Таблица 6.18

Нормированные разности и евклидовы расстояния для кластера «Бугры + Щеглово + Пригородное + Авлога»

Предприятия	Признаки				Евклидово расстояние
	x_1	x_2	x_3	x_4	
Средние величины по кластеру	345	91	15,7	48	0
«Ручьи»	2,312	2,106	0,766	0,585	3,273
«Всеволожское»	0,138	0,162	0,141	2,707	2,719
«Выборгское»	1,789	1,014	1,297	0,537	2,490
«Приневское»	0,248	2,606	2,734	0,902	3,891

Таблица 6.19

Матрица евклидовых расстояний после образования кластера «Бугры + Щеглово + Пригородное + Авлога»

Предприятия	Кластер Б+Щ+П+А	«Ручьи»	«Всеволожское»	«Выборгское»	«Приневское»
Кластер Б+Щ+П+А	0				
«Ручьи»	3,273	0			
«Всеволожское»	2,719	3,887	0		
«Выборгское»	2,490	1,734	4,127	0	
«Приневское»	3,891	2,913	4,188	3,008	0

Минимальное евклидово расстояние между предприятиями «Ручьи – Выборгское» (оно меньше 2), следовательно, эти предприятия объединяются в кластер 2 (табл. 6.20). Кластер Б+Щ+П+А будем называть кластером 1.

Таблица 6.20

Нормированные разности и евклидовы расстояния для кластеров 1 и 2

Предприятия	Признаки				Евклидово расстояние
	x_1	x_2	x_3	x_4	
Средние кластера 2	568	312	22,3	49	0
Кластер 1	2,046	1,556	1,031	0,024	2,770
«Всеволожское»	2,183	1,394	1,172	2,683	3,904
«Приневское»	1,798	1,049	1,703	0,878	2,829

После четвертого шага получаем новую матрицу евклидовых расстояний (табл. 6.21).

Согласно табл. 6.21 все расстояния больше 2. Оставляем 4 типа предприятий: предприятия, вошедшие в кластер 1, кластер 2, кластер 3 («Всеволожское») и кластер 4 («Приневское»).

Сравнивая результат кластерного анализа с многомерными средними (табл. 6.8) видим, что состав кластера 1 точно отвечает тем хозяйствам, чьи многомерные средние ниже 100%. Также выделение в самостоятельный кластер предприятия «Приневское» соответствует его высшему значению многомерной средней. А вот объединение в кластер 2 предприятий «Ручьи» и «Выборгское» не соответствует многомерным средним, по которым к предприятию «Ручьи» было ближе предприятие «Всеволожское». В результате резкого отличия по признаку x^4 предприятие «Всеволожское» выделилось в отдельный кластер 3.

Таблица 6.21

Матрица евклидовых расстояний после образования кластера 2

	Кластер 1	Кластер 2	Кластер 3 («Всеволожское»)	Кластер 4 («Приневское»)
Кластер 1	0			
Кластер 2	2,770	0		
«Всеволожское»	2,719	3,909	0	
«Приневское»	3,891	2,829	4,188	0

Обобщая рассмотренную **процедуру кластерного анализа, представим действия в виде определенной последовательности:**

1) вычисление средних величин каждого из классификационных признаков $\bar{x}_j$ в целом по совокупности;

2) вычисление средних квадратических отклонений каждого из признаков по совокупности – s_{xj} или σ_{xj};

3) вычисление матриц нормированных разностей по каждому из группировочных признаков – $d_{jp,q}$;

4) вычисление евклидовых расстояний между каждой парой сочетаний единиц совокупности – $d_{p,q}$;

5) выбор наименьшего из евклидовых расстояний – $d_{p,qmin}$;

6) объединение единиц совокупности с наименьшим евклидовым расстоянием между ними в один кластер;

7) вычисление средних значений всех признаков для единиц, объединенных в кластер;

8) вычисление новых нормированных расстояний между объединенным кластером и остальными единицами;

9) вычисление новых евклидовых расстояний между объединенным кластером и остальными единицами (или кластерами);

10) выбор наименьшего из евклидовых расстояний;

11) повторение операций (6-10) и т.д.

10–3182

Объединение в кластеры прекращается, когда все евклидовы расстояния превысят заданную критическую величину $d_{крит}$. Обычно ППП предусматривает вывод на печать состава (перечня единиц совокупности) каждого кластера, евклидовых расстояний между ними, матриц нормированных разностей по каждому признаку.

Существует много достаточно сложных алгоритмов кластерного анализа и родственных ему методов распознавания образов, таксономии и др.

Рассмотренная выше методика вычисления евклидова расстояния предполагает, что все признаки считаются равноправными. На самом же деле при выделении типов социально-экономических явлений группировочные признаки не равноправны: как правило, одни признаки имеют большее, другие — меньшее значение. Следовательно, более совершенная методика кластерного анализа должна учитывать разную значимость, разный «вес» группировочных признаков. В этом случае должно использоваться взвешенное евклидово расстояние:

$$d_{p,q} = \sqrt{\sum_{j=1}^{k} d_{jp,q} w_j}, \qquad (6.17)$$

где w_j – вес j-го признака.

Определение весов – весьма сложная задача, выходящая за пределы компетенции статистики. О том, какие признаки важнее при классификации тех или иных объектов, могут судить не статистики, а специалисты в соответствующей отрасли. Поэтому одним из способов определения весов признаков при кластерном анализе являются экспертные оценки. Опросив достаточное число специалистов-экспертов (желательно не менее 6-10), статистик сможет определить по их оценкам места (роли) каждого группировочного признака. Затем находится среднее по оценкам всех экспертов место признака или его «вес» в численном выражении. Можно просить экспертов ранжировать признаки по порядку значимости и определять «среднее место», но оценка при этом будет очень грубая: признак, поставленный на первое место, будет вдвое важнее второго и в двадцать или тридцать раз важнее последнего. Чтобы различия весов были не столь резкими, можно просить экспертов распределить между группировочными признаками, в соответствии с их значениями, общую сумму оценок (100 или 1000%). Тогда каждому из признаков будет приписана некоторая доля этой общей суммы, можно двум-трем признакам приписать одинаковые веса. Но этот

способ взвешивания требует от экспертов большей точности и напряжения, чем простое ранжирование признаков.

Субъективность экспертных оценок в какой-то мере можно компенсировать статистической обработкой. Например, по каждому признаку перед определением средней оценки его веса можно отбросить максимальную и минимальную оценки, если они резко отличаются от оценок остальных экспертов. Можно вообще исключить того эксперта, чьи оценки в среднем отличаются от средних оценок признаков более чем, например, на 2σ. Однако эти статистические коррективы небезупречны и допустимы при значительном числе экспертов для того, чтобы их средние оценки были надежны.

Существует и другая возможность оценки роли группировочных признаков, их значимости для классификации: на основе стандартизованных коэффициентов регрессии или коэффициентов раздельной детерминации (см. гл. 8).

Рассмотренный алгоритм иерархической классификации можно модифицировать, используя ***метод «ближайшего» или «дальнего соседа»*** (табл. 6.22). В этом случае в матрицу евклидовых расстояний вводятся расстояния, полученные не на основе средних величин по кластеру, в качестве представителя кластера берется входящий в него объект либо наименее удаленный от остальных объектов («ближайший сосед»), либо наиболее удаленный от остальных («дальний сосед»). Поскольку d_{min} = 0,981 (табл. 6.13) предприятия «Бугры» и «Щеглово» были объединены в кластер. При использовании метода «ближайшего соседа» в последующей после объединения этих двух предприятий матрице евклидовых расстояний кла-

Таблица 6.22

Матрица евклидовых расстояний на первом шаге (метод «ближайшего соседа»)

№ п/п	Предприятия	1	3	4	5	6	7	8, 2
1	«Ручьи»	0						
3	«Пригородное»	3,012	0					
4	«Авлога»	4,130	2,885	0				
5	«Всеволожское»	3,887	3,441	2,284	0			
6	«Выборгское»	1,734	1,373	3,559	4,127	0		
7	«Приневское»	2,913	3,568	4,157	4,188	3,008	0	
8+2	«Щеглово + Бугры»	2,852	1,130	1,541	2,422	2,281	3,775	0

стер будет представлять то «Бугры», то «Щеглово» – в зависимости от того, какое из предприятий наименее удалено от остальных. Для простоты будем использовать не названия, а порядковые номера предприятий, соответствующие их последовательности в табл. 6.8.

Минимальное евклидово расстояние между кластером и предприятием «Пригородное» $d_{8,2,3} = 1{,}130$. (табл. 6.13). Это хозяйство имеет номер 3, присоединим его к кластеру (8.2). Матрица евклидовых расстояний на втором шаге будет следующей (табл. 6.23).

Таблица 6.23

Матрица евклидовых расстояний на втором шаге (метод «ближайшего соседа»)

Предприятия	1	4	5	6	7	8, 2, 3
1	0					
4	4,130	0				
5	3,887	2,284	0			
6	1,734	3,559	4,127	0		
7	2,913	4,157	4,188	3,008	0	
8, 2, 3	2,852	1,541	2,422	1,373	3,568	0

Минимальным является расстояние между предприятием («Выборгское») и кластером: $\min d_{8,2,3,6} = 1{,}373$. При кластере из четырех предприятий матрица евклидовых расстояний представлена в табл. 6.24.

Таблица 6.24

Матрица евклидовых расстояний на третьем шаге (метод «ближайшего соседа»)

Предприятия	1	4	5	7	8, 2, 3, 6
1	0				
4	4,130	0			
5	3,887	2,284	0		
7	2,913	4,157	4,188	0	
8, 2, 3, 6	1,734	1,541	2,422	3,008	0

Поскольку $\min d_{8,2,3,6,4} < 2$, присоединим к кластеру предприятие 4 («Авлога») (табл. 6.25).

Таблица 6.25

Матрица евклидовых расстояний на четвертом шаге (метод «ближайшего соседа»)

Предприятия	1	5	7	8, 2, 3, 6, 4
1	0			
5	3,887	0		
7	2,913	4,188	0	
8, 2, 3, 6, 4	1,734	2,284	3,008	0

Сравнивая табл. 6.24 и 6.25, видим, что расстояния между кластером и объектом 1 и в обоих случаях определяются объектом 6, который является «ближайшим соседом» объекта 1. При определении расстояния между кластером и предприятием 7 представителем кластера выступает предприятие 6, которое является ближайшим к предприятию 7.

На следующем шаге к кластеру присоединяется предприятие 1 (6.26).

Таблица 6.26

Матрица евклидовых расстояний на пятом шаге (метод «ближайшего соседа»)

Предприятия	5	7	8, 2, 3, 6, 4, 1
5	0		
7	4,188	0	
8, 2, 3, 6, 4, 1	2,284	2,913	0

Расстояние между кластером и предприятием 5 также, как и на предыдущем шаге, определяется расстоянием между предприятиями 5 и 4, которое является ближайшим к 5 из всех предприятий, входящих в кластер. Расстояние между предприятием 7 и кластером в табл. 6.26 стало определяться «ближайшим соседом» – предприятием 1. Если продолжить процедуру классификации, то поскольку минимальным расстоянием в табл. 6.26 является расстояние между кластером и предприятием 5, то можно присоединить его к кластеру (табл. 6.27).

Таблица 6.27

Евклидово расстояние на шестом шаге

Предприятия	7	8, 2, 3, 6, 4, 1, 5
7	0	
8, 2, 3, 6, 4, 1, 5	2,913	0

В этом случае мы получим два кластера: один состоит из предприятия 7 («Приневское»), а другой включает остальные семь предприятий. Если придерживаться некоторого критического значения евклидового расстояния так, как в ранее рассмотренном примере, когда $d_{крит} = 2$, то предприятие 5 не присоединяется к кластеру и в итоге совокупность подразделяется на три кластера, два из которых содержат по одному предприятию (5 и 7) и один – шесть предприятий.

Представим графически процесс классификации в виде дендрограммы (рис. 6.2).

Дендрограмма – дерево объединений кластеров с порядковыми номерами объектов по горизонтальной оси и шкалой расстояний по вертикальной оси.

Решение, полученное методом «ближайшего соседа», близко к прежнему результату при описании кластеров средними показателями (табл. 6.21), но не совпадает с ним – вместо четырех кластеров здесь выделились три.

Таблица 6.28

Матрица евклидовых расстояний на первом шаге (метод «дальнего соседа»)

Предприятия	1	3	4	5	6	7	8+2
1	0						
3	3,012	0					
4	4,130	2,885	0				
5	3,887	3,441	2,284	0			
6	1,734	1,373	3,559	4,127	0		
7	2,913	3,568	4,157	4,188	3,008	0	
8+2	3,480	1,411	1,629	3,184	2,712	4,383	0

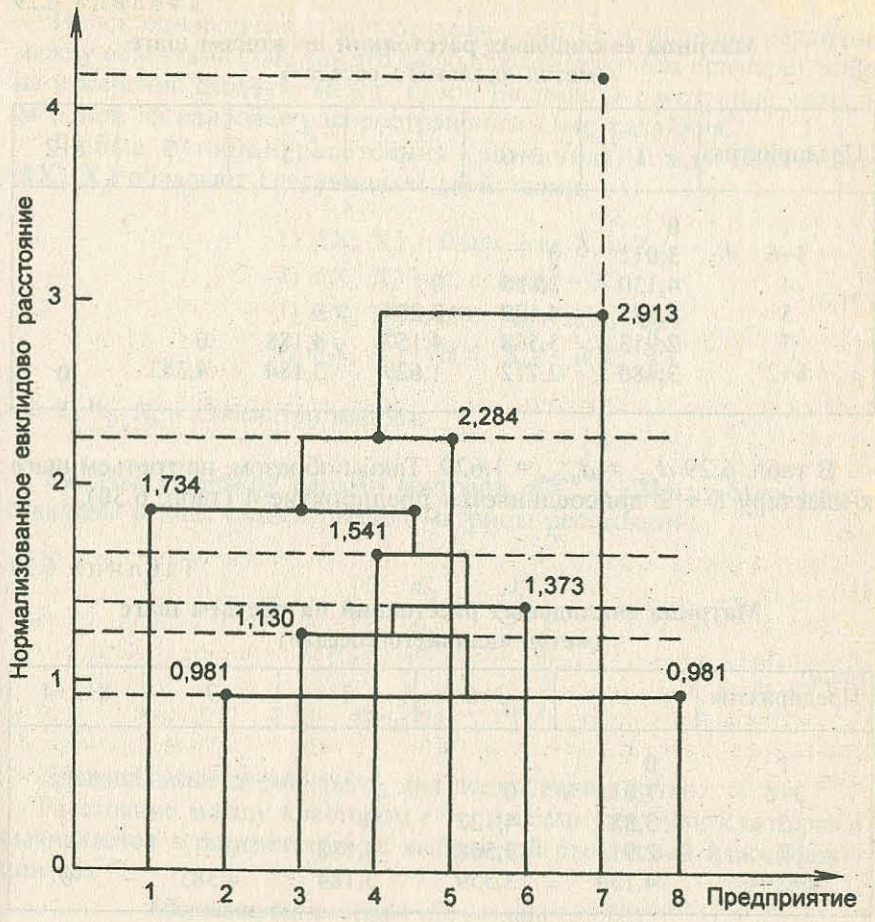


Рис. 6.2. Дендрограмма: метод «ближайшего соседа»

Если применить метод «дальнего соседа», то на первом шаге, после объединения предприятий 2 и 8, получим следующую матрицу евклидовых расстояний (табл. 6.28).

Табл. 6.28 отличается от табл. 6.21 последней строкой, в которой показаны максимальные расстояния кластера (8 + 2) от других объектов.

Затем выбирается наименьшее из $d_{p,q}$. В данном примере это расстояние между хозяйствами 3 и 6 ($d_{3,6} = 1,373$), образующими новый кластер, в котором также выделяется «дальний сосед» (табл. 6.29).

Таблица 6.29

Матрица евклидовых расстояний на втором шаге (метод «дальнего соседа»)

Предприятия	1	3+6	4	5	7	8+2
1	0					
3+6	3,012	0				
4	4,130	2,885	0			
5	3,887	4,127	2,284	0		
7	2,913	3,568	4,157	4,188	0	
8+2	3,480	2,712	1,629	3,184	4,383	0

В табл. 6.29 $d_{\min} = d_{8+2,4} = 1{,}629$. Таким образом, на третьем шаге к кластеру 8 + 2 присоединяется предприятие 4 (табл. 6.30).

Таблица 6.30

Матрица евклидовых расстояний на третьем шаге (метод «дальнего соседа»)

Предприятия	1	3+6	5	7	8+2+4
1	0				
3+6	3,012	0			
5	3,887	4,127	0		
7	2,913	3,568	4,188	0	
8+2+4	4,130	3,559	3,184	4,383	0

В табл. 6.30 все значения $d_{p,q} > 2$. Следовательно, в результате метода «дальнего соседа» получаем 5 кластеров, три из которых включают по одному предприятию.

Подведем итоги.

Все алгоритмы многомерной классификации основаны на целевой функции:

$$\sum_{1}^{m}\sum_{1}^{n_i}(x_{ij} - \bar{x}_j)^2 \to \min, \qquad (6.18)$$

т. е. выделение однородных групп при минимизации внутригрупповой колеблемости.

Поиск однородных групп основан либо на измерении различия между объектами (так, как это было в рассмотренном примере), либо на измерении сходства между ними. Евклидово расстояние является одной из наиболее распространенных мер различия.

Любые функции расстояния (различия) между объектами $d(\mathbf{X}_i, \mathbf{X}_j)$ обладают следующими свойствами:

$$\begin{aligned} &1)\ d(\mathbf{X}_i, \mathbf{X}_j) > 0 \text{ для всех } \mathbf{X}_i \text{ и } \mathbf{X}_j\ (i \neq j);\\ &2)\ d(\mathbf{X}_i, \mathbf{X}_j) = 0, \text{ если } \mathbf{X}_i = \mathbf{X}_j;\\ &3)\ d(\mathbf{X}_i, \mathbf{X}_j) = d(\mathbf{X}_j, \mathbf{X}_i) = d(\mathbf{X}_j, \mathbf{X}_i);\\ &4)\ d(\mathbf{X}_i, \mathbf{X}_j) < d(\mathbf{X}_i, \mathbf{X}_h) + d(\mathbf{X}_h, \mathbf{X}_j), \end{aligned} \qquad (6.19)$$

где $\mathbf{X}_i$, $\mathbf{X}_j$, $\mathbf{X}_h$ – любые три вектора.

Расстояния между парами векторов $d(\mathbf{X}_i, \mathbf{X}_j)$ могут быть представлены в виде симметричной матрицы расстояний:

$$D = \begin{pmatrix} 0 & d_{12} & \dots & d_{1n} \\ d_{21} & 0 & \dots & d_{2n} \\ \vdots & \vdots & & \vdots \\ d_{n1} & d_{n2} & \dots & 0 \end{pmatrix}. \qquad (6.20)$$

Диагональные элементы d_{ij} для всех i равны нулю.

Расстояние между кластером $i + j$ и всеми другими кластерами вычисляется в соответствии с выбранной стратегией классификации как

«ближайшего соседа»: $d_{i+j,\,k} = \min(d_{ik}, d_{jk})$;

«дальнего соседа»: $d_{i+j,\,k} = \max(d_{ik}, d_{jk})$;

«группового соседа»: $d_{i+j,\,k} = (d_{ik}n_i + d_{jk}n_j)/(n_i + n_j)$,

где $n_i + n_j = n_{i+j}$.

Метод «ближайшего соседа» сжимает пространство исходных переменных и рекомендуется для получения минимального дерева иерархической классификации. Метод «дальнего соседа» растягивает пространство. Метод «группового соседа» сохраняет метрику пространства.

Если классификация данных основана на мерах сходства $s(X_i, X_j)$, то следует иметь в виду общие свойства этих мер:

$$
\begin{aligned}
&1)\ 0 \le s(X_i, X_j) < 1 \text{ для } X_i \ne X_j;\\
&2)\ s(X_i, X_j) = 1;\\
&3)\ s(X_i, X_j) = s(X_j, X_i)
\end{aligned}
\qquad (6.21)
$$

Соответственно матрица мер сходства имеет вид:

$$
S = \begin{pmatrix}
1 & s_{12} & \dots & s_{1n} \\
s_{21} & 1 & \dots & s_{2n} \\
\vdots & & & \\
s_{n1} & s_{n2} & \dots & 1
\end{pmatrix}. \qquad (6.22)
$$

Диагональные элементы такой матрицы равны 1.

В качестве мер сходства чаще всего используются коэффициенты корреляции (см. гл. 8).

Основными ППП для решения задачи многомерной классификации являются «Класс-мастер», SPSS, SAS. Многие алгоритмы многомерной классификации основаны на геометрическом представлении кластера как локального скопления точек в заданном признаковом пространстве.

Большинство методов классификации основано на однозначном отнесении объекта к тому или иному классу. Но, как уже отмечалось, границы классов могут быть размытыми, нечеткими. Класс объектов, в котором нет резкой границы между объектами, входящими в него, и теми, которые в него не входят, называется *нечетким* множеством.

Для классификации данных в нечетких множествах необходимо ввести матрицу принадлежности каждого объекта к нечеткому множеству с элементами

$$\mu_{ij} \in [0,1]\ \forall\ i \in 1 \div n, j \in 1 \div k.$$

$\sum_{j=1}^{k} \mu_{ij} = 1$, т. е. объект обязательно принадлежит к тому или иному нечеткому множеству. Качество разбиения определяется как минимизацией внутриклассовой дисперсии, так и максимизацией удаленности центров классов.

Алгоритмы и программы многомерной классификации постоянно развиваются: разрабатываются ППП, учитывающие размытость границ между классами (распознавание в нечетких множествах), различную длину описаний классов и т. д. Большое значение в решении задач иерархических классификаций имеет компьютерная графика – так называемые классификационные деревья. Подробнее вопросы многомерной классификации освещаются в работах, указанных в списке рекомендуемой литературы.

Рекомендуемая литература к главе 6

1. *Айвазян С. А., Бежаева З. И., Староверов О. В.* Классификация многомерных наблюдений. – М.: Статистика, 1974.
2. *Афифи А., Эйзен С.* Статистический анализ. Подход с использованием ЭВМ: Пер. с англ. – М.: Мир, 1982.
3. *Елисеева И. И. Рукавишников В. О.* Группировка, корреляция, распознавание образов. – М.: Статистика, 1977.
4. *Енюков И. С.* Методы – алгоритмы – программы многомерного статистического анализа. – М.: Финансы и статистика, 1986.
5. *Кулаичев А. П.* Методы и средства анализа данных в среде Windows. Stadia 6.0. – М.: НПО «Информатика и компьютеры», 1996.
6. *Мандель И. Д.* Кластерный анализ. – М.: Финансы и статистика, 1988.
7. *Миркин Б. Г.* Группировки в социально-экономических исследованиях. – М.: Финансы и статистика, 1985.

7 Глава ВЫБОРОЧНОЕ НАБЛЮДЕНИЕ. ИСПЫТАНИЕ СТАТИСТИЧЕСКИХ ГИПОТЕЗ

7.1. ПРИЧИНЫ ПРИМЕНЕНИЯ ВЫБОРОЧНОГО НАБЛЮДЕНИЯ. ДИСКРИПТИВНАЯ СТАТИСТИКА И СТАТИСТИЧЕСКИЙ ВЫВОД

В главе 2 отмечалось, что статистика далеко не всегда имеет дело с данными сплошного наблюдения. Из всех видов несплошного наблюдения главным является выборочное наблюдение, так как только выборочный метод имеет статистико-математическое обоснование распространения данных, полученных по выборке, на всю совокупность.

Причин использования выборочного метода несколько.

Во-первых, как это ни парадоксально, это *повышение точности данных*; уменьшение числа единиц наблюдения в выборке резко снижает ошибки регистрации. Правда, за счет неполноты охвата единиц возникает ошибка репрезентативности, т. е. представительности выборочных данных. Но даже взятые вместе ошибка наблюдения для выборки плюс ошибка репрезентативности обеспечивают большую точность выборочных данных по сравнению с массовым сплошным наблюдением.

При ограничении объема работы можно привлечь более квалифицированных исполнителей (интервьюеров, счетчиков-регистраторов). Это положительно сказывается на качестве данных выборочного обследования.

Во-вторых, обращение к выборкам обеспечивает *экономию материальных, трудовых, финансовых ресурсов и времени*. Например, для составления баланса, денежных доходов и расходов населения, для изучения денежного обращения, выявления дифференциации населения по уровню жизни, определения черты бедности и т. д. необходимы данные о бюджетах домохозяйств. Сбор этих данных

осуществляется государственной статистикой, но один статистик в состоянии курировать ежедневные записи доходов, расходов, потребления не более чем в 20–25 домохозяйствах. Если бы решили собирать данные о бюджетах всех домохозяйств, то только для этой цели (не учитывая потребности последующей обработки) потребовалось бы примерно два миллиона статистиков. Так что использование выборочного наблюдения является единственным экономически выгодным решением, тем более что по результатам изучения сравнительно небольшой части можно получить с достаточно высокой степенью уверенности данные о всей совокупности. Подобная ситуация возникает и при аудиторских проверках крупных фирм, когда вместо детального изучения каждого платежного документа ограничиваются анализом выборки документов, и в других областях применения статистики.

В-третьих, без выборки не обойтись, когда наблюдение связано с *порчей наблюдаемых объектов*. Это относится прежде всего к изучению качества продукции, которое основано на испытаниях образцов на вибрацию, упругость, разрыв и т.д. Всю продукцию, конечно же, таким испытаниям не подвергают, только отобранные образцы. То же можно сказать об исследовании молока на жирность, зерна – на содержание белка, влажность, чистоту и всхожесть семян, электрических лампочек – на длительность горения и т.д. На выборках основаны маркетинговые исследования, оценки качества поставок.

Практика применения выборочного метода очень разнообразна. Иногда, проведя сплошное наблюдение, применяют выборочный метод при разработке данных: отбирают часть данных для более подробной разработки по расширенной программе. Так поступают, например, при разработке данных переписи населения о составе и типах семей. Нередко в процессе сбора данных применяют совместно сплошное и несплошное наблюдение. При переписях населения в нашей стране (1959, 1970, 1979 гг.) собирались сведения о каждом лице по 11 признакам, а 25% населения давали более подробную информацию (18 вопросов).

Выборки используются при опросах общественного мнения, при выяснении потребительских предпочтений, формировании доходов и расходов населения, при определении урожайности сельскохозяйственных культур и продуктивности скота. С 20-х гг. нашего века выборочный метод стал использоваться для контроля и анализа качества продукции. Сейчас методы статистической выборки все шире внедряются в самые различные области. В 1994 г. в Российской Федерации была проведена 5%-ная микроперепись населения с целью уточнения демографического и социального состава насе-

ления, уровня благосостояния, включая жилищные условия, источники дохода и др.

Та совокупность, из которой производится отбор, называется *генеральной совокупностью;* отобранные данные составляют *выборочную совокупность.* Эти данные представляют интерес постольку, поскольку дают основание для суждений о параметрах и свойствах генеральной совокупности.

Таким образом, **выборочный метод обладает следующими достоинствами:**

- относительно небольшие (по сравнению со сплошным наблюдением) материальные, трудовые и стоимостные затраты на сбор данных (включая затраты на планирование и формирование выборки);
- оперативность получения результатов;
- широкая область применения;
- высокая достоверность результатов.

Все эти достоинства проявляются лишь **при условии правильного решения проблем выборочного обследования.** К ним относятся:

1) определение границ генеральной совокупности;
2) разработка программы наблюдения и инструкций;
3) определение основы для проведения выборки – списка единиц генеральной совокупности, сведений об их размещении и т.д.;
4) установление допустимого размера погрешности и определение объема выборки;
5) определение вида выборочного наблюдения;
6) установление сроков проведения наблюдения;
7) определение потребности в кадрах для проведения выборочного наблюдения, их подготовка;
8) оценка точности и достоверности данных выборки, определение порядка их распространения на генеральную совокупность.

Представление о статистических данных как о выборочных может относиться не только к собственно выборке, но и к данным сплошного наблюдения, которые иногда рассматриваются как выборка из всех возможных реализаций изучаемого процесса. Это имеет смысл в случае малого числа единиц совокупности. Кроме того, трактовка данных как выборочных используется применительно к результатам эксперимента, которые рассматриваются как некая выборка из потенциально бесконечного числа повторений экспериментальных наблюдений.

Трактовка данных как выборочных является основой деления статистики на *описательную* (дискриптивную) и *выводную.* **Методы описательной статистики** включают сбор данных по всем

единицам изучаемой совокупности, их обработку, получение сводных показателей, которые являются характеристиками только наблюдаемой совокупности. Например, если наша задача состоит в изучении успеваемости группы студентов, включающей 25 человек, вычисленный средний балл по этой группе, процент отличных оценок и т. д. являются описаниями этой совокупности. Если же мы будем рассматривать эту группу студентов с точки зрения оценки успеваемости всех студентов данного колледжа или университета, то эта группа предстанет как выборка из общего числа студентов. В этом случае средний балл для группы будет являться оценкой средней успеваемости студентов колледжа в целом.

Генеральная совокупность может быть *реальной,* а может быть *гипотетической,* включающей случаи, которые реально не существуют, например все возможные результаты эксперимента.

В выводной статистике принято строго различать *параметры* и свойства генеральной совокупности и их *оценки* по данным выборки. С этой целью принята следующая система обозначений: генеральные параметры обозначаются греческими буквами, выборочные показатели, которые рассматриваются как оценки генеральных параметров, обозначаются латинскими буквами. Например,

	Генеральная совокупность	Выборка
Средняя величина	μ	$\bar{x}$
Относительная величина	π	p
Дисперсия	σ^2	s^2
Коэффициент корреляции	ρ	r

Объем генеральной совокупности обозначают N, объем выборочной совокупности – k.

Выборочные оценки отличаются от генеральных параметров за счет ошибки наблюдения и ошибки выборки:

Выборочная оценка = Генеральный параметр ± Ошибка наблюдения ± Ошибка выборки.

Подводя итоги, можно сказать, что описательная статистика является инструментом описания совокупности, по которой у нас полностью имеются исходные данные. **Метод статистического вывода** позволяет по данным выборок делать заключение о более большой совокупности, по которой мы не имеем исчерпывающих наблюдений.

7.2 СПОСОБЫ ОТБОРА, ОБЕСПЕЧИВАЮЩИЕ РЕПРЕЗЕНТАТИВНОСТЬ ВЫБОРКИ. ВИДЫ ВЫБОРКИ

Для того чтобы можно было по выборке делать вывод о свойствах генеральной совокупности, выборка должна быть *репрезентативной* (представительной), т. е. она должна полно и адекватно представлять свойства генеральной совокупности. Репрезентативность выборки может быть обеспечена только при *объективности отбора* данных.

Выборочная совокупность формируется по принципу массовых вероятностных процессов без каких бы то ни было исключений от принятой схемы отбора; необходимо обеспечить относительную однородность выборочной совокупности или ее разделение на однородные группы единиц. При формировании выборочной совокупности должно быть дано четкое определение единицы отбора. Желателен приблизительно одинаковый размер единиц отбора, причем результаты будут тем точнее, чем меньше единица отбора.

Возможны три способа отбора: *случайный отбор, отбор единиц по определенной схеме, сочетание первого и второго способов.*

Если отбор в соответствии с принятой схемой проводится из генеральной совокупности, предварительно разделенной на типы (слои или страты), то такая выборка называется *типической* (или расслоенной, или стратифицированной, или районированной). Еще одно деление выборки по видам определяется тем, что является единицей отбора: единица наблюдения или серия единиц (иногда используют термин «гнездо»). В последнем случае выборка называется *серийной,* или *гнездовой.* На практике часто используется сочетание типической выборки с отбором сериями. В *математической статистике,* обсуждая проблему отбора данных, обязательно вводят деление выборки на *повторную* и *бесповторную.* Первая соответствует схеме возвратного шара, вторая – безвозвратного (при рассмотрении процесса отбора данных на примере отбора шаров разного цвета из урны). В социально-экономической статистике нет смысла применять повторную выборку, поэтому, как правило, имеется в виду *бесповторный отбор.* Если выборка производится по схеме возвращенного шара, то вероятность попадания любой единицы в выборку равна $1/N$, и она остается той же самой на протяжении всей процедуры отбора. Если выборка производится по схеме невозвращенного шара, то вероятность попадания единицы в выборку изменяется от $1/N$ – для первой отбираемой единицы, до $\frac{1}{N-n+1}$ – для последней.

Так как социально-экономические объекты имеют сложную структуру, то выборку бывает довольно трудно организовать. Например, чтобы провести отбор домохозяйств при изучении потребления населением крупного города, легче произвести сначала отбор территориальных ячеек, жилых домов, потом квартир или домохозяйств, затем респондента. Такая выборка называется *многоступенчатой.* На каждой ступени используются разные единицы отбора: более крупные – на начальных ступенях, на последней ступени единица отбора совпадает с единицей наблюдения.

Еще один вид выборочного наблюдения – *многофазовая выборка.* Такая выборка включает определенное количество фаз, каждая из которых отличается подробностью программы наблюдения. Например, 25% всей генеральной совокупности обследуются по краткой программе, каждая 4-я единица из этой выборки обследуется по более полной программе и т.д.

При любом виде выборки отбор единиц производится тремя отмеченными способами. Рассмотрим процедуру *случайного отбора.* Прежде всего составляется список единиц совокупности, в котором каждой единице присваивается цифровой код (номер или метка). Затем производится жеребьевка. Закладываются в барабан шары с соответствующими номерами, они перемешиваются и проводится отбор шаров. Выпавшие номера соответствуют единицам, попавшим в выборку; число номеров равно запланированному объему выборки.

Отбор жеребьевкой может быть подвержен смещениям, вызванным недостатками техники (качеством шаров, барабана) и другими причинами. Более надежен с точки зрения объективности отбор по таблице случайных чисел. Такая таблица содержит серии цифр, чередующихся случайным образом, отобранных путем электронных сигналов. Так как мы пользуемся десятичной цифровой системой 0, 1, 2, ..., 9, вероятность появления любой цифры равна 1/10. Следовательно, если бы нужно было создать таблицу случайных чисел, включающую 500 знаков, то из них около 50 были бы 0, столько же – 1 и т.д. Ввиду того, что каждая цифра и их последовательность являются случайными, можно использовать таблицу, перемещаясь либо по ее вертикали, либо по горизонтали. Цифры сгруппированы по 5 для лучшей обозримости таблицы и пользования ею (см. Приложение, табл. 7).

Предположим, что нам нужно из 9540 студентов университета произвести 5%-ную выборку: $n = 5\% \cdot N = 477$ студентов. Ввиду того, что объем генеральной совокупности выражается четырехзначным числом, код каждого студента должен быть четырех-

значным: от 0001 – для первого студента до 9540 – для последнего студента в списке. Чтобы произвести отбор по таблице случайных чисел, нужно выбрать начальную точку: можно закрыть глаза и поставить наугад точку в таблице карандашом. Предположим, мы попали в 13-ю строку в 1-й столбец (табл. 7.1).

Таблица 7.1

Пример использования таблицы случайных чисел

Строки	Столбцы							
	1	2	3	4	5	6	7	8
13	90822	60280	88925	99610	42772	60561	76873	04117
14	72121	79152	96591	90305	10189	79778	68016	13747
15	95268	41377	25684	08151	61816	58555	54305	86189
16	92603	09091	75884	93424	72586	88903	30061	14457
17	18813	90291	05275	01223	79607	95426	34900	09778
18	38840	26903	28624	67157	51986	42865	14508	49315

Следовательно, единица с номером 9082 является первой в выборке. Если двигаться по строке, то единица с номером 2602 будет второй, 8088 – третьей, 9259 – четвертой. Следующий код 9610 пропускаем, так как у нас нет студента с таким номером. Далее в выборку попадают номера 4277, 2605, 6176, 8730, 4117, 7212, 1791, 5296, 5919, 0305, 1018. Код 9797 пропускается. Следующие отобранные номера 7868, 0161, 3747, 9526, 8413, 7725 и т.д.

Процедура продолжается, пока число отобранных номеров не составит требуемый объем выборки (n = 477).

Часто используется отбор по какой-либо схеме (так называемая направленная выборка). Схема отбора принимается такой, чтобы отразить основные свойства и пропорции генеральной совокупности. Простейший способ: по спискам единиц генеральной совокупности, составленным так, чтобы упорядочивание единиц было бы не связано с изучаемыми свойствами, проводится *механический отбор* единиц с шагом, равным $N : n$. Обычно отбор начинают не с первой единицы, а отступив полшага, чтобы уменьшить возможность смещения выборки. Частота появления единиц с теми или иными особенностями, например студентов с тем или иным уровнем успеваемости, живущих в общежитии, и т.д. будет определяться той структурой, которая сложилась в генеральной совокупности.

Для большей уверенности в том, что выборка отразит структуру генеральной совокупности, последняя подразделяется на типы (страты или районы), и проводится случайный или механический отбор из каждого типа (района, страта). Общее число единиц, отобранных из разных типов, должно соответствовать объему выборки.

Особые трудности возникают, когда нет списка единиц, а отбор нужно произвести либо на местности, либо из образцов продукции на складе готовой продукции. В этих случаях важно детально разработать схему ориентации на местности и схему отбора и следовать ей, не допуская отклонений. Например, счетчик имеет указание двигаться от определенной автобусной остановки на север по четной стороне улицы и, отсчитав два дома от первого угла, войти в третий и провести опрос в каждом 5-м жилом помещении. Неукоснительное следование принятой схеме обеспечивает выполнение главного условия формирования репрезентативной выборки – объективности отбора единиц.

От случайной выборки следует отличать ***квотный отбор***, когда выборка конструируется из единиц определенных категорий (квот), которые должны быть представлены в заданных пропорциях. Например, при опросе покупателей универмага может быть запланировано провести отбор 150 респондентов, в том числе 90 женщин, из них 25 – девушек, 20 – молодых женщин с маленькими детьми, 35 – женщин среднего возраста, одетых в деловой костюм, 10 – женщин 50 лет и старше; кроме того, планировался опрос 70 мужчин, из них 25 – подростков и юношей, 20 – молодых мужчин с детьми, 15 – мужчин, которые одеты в костюмы, 10 – мужчин, одетых в спортивную одежду. Для определения потребительских ориентаций и предпочтений такая выборка, может быть, и хороша, но если мы захотим по ней установить среднюю сумму покупок, их структуру, мы получим непредставительные результаты. Это происходит потому, что квотная выборка нацелена на отбор определенных категорий.

Выборка может быть нерепрезентативной, даже если она формируется в соответствии с известными пропорциями генеральной совокупности, но отбор проводится без какой-либо схемы – единицы набираются как угодно, лишь бы обеспечить соотношение их категорий в тех же пропорциях, что и в генеральной совокупности (например, соотношение мужчин и женщин, респондентов в возрасте моложе и старше трудоспособного и в трудоспособном и т.д.).

Эти замечания должны предостеречь вас от подобных подходов к формированию выборки и еще раз подчеркнуть необходимость объективного отбора.

7.3. ОШИБКА ВЫБОРКИ

Все ошибки выборочного наблюдения подразделяются на ошибки выборки (случайные); ошибки, вызванные отклонением от схемы отбора (неслучайные); ошибки наблюдения (случайные и неслучайные). Плохо, когда ошибка выборки превышает допустимый размер погрешности, но слишком высокая точность также подозрительна и, как правило, свидетельствует об ошибках отбора.

Ошибки отбора приводят к неслучайным ошибкам. Так бывает, если объективный отбор подменяется «удобной» выборкой. Например, когда появляются добровольные респонденты – те, кто сами предлагают, чтобы их опросили. Очевидно, что характеристики таких добровольцев и недобровольцев могут быть отличны и это приведет к ошибочному заключению о генеральной совокупности.

Такая же опасность возникает при замене по какой-либо причине единиц, попавших в выборку, другими единицами (например, вместо отобранного домохозяйства, где в момент прихода интервьюера никто не открыл дверь, был проведен опрос в соседней квартире; или интервьюер встретил решительный отказ участвовать в опросе и был вынужден пойти на замену домохозяйства). Как отмечает социолог В. И. Паниотто, систематические ошибки представляют собой некоторое постоянное смещение, которое не уменьшается с увеличением числа опрошенных и вызвано недостатками и просчетами в системе отбора респондентов. Если, например, для изучения общественного мнения жителей города в архитектурном управлении получить сведения о жилом фонде и из всех имеющихся в городе квартир отобрать случайным образом 400 квартир, а затем предложить интервьюерам опросить всех, кого они застанут в момент посещения в этих квартирах, то полученные данные не будут репрезентативны. Допущена систематическая ошибка: более подвижная часть населения попадает в выборку в меньшей пропорции, а менее подвижная – в большей пропорции, чем в генеральной совокупности. Пенсионеров, например, можно чаще застать дома, чем студентов-вечерников. При увеличении выборки эта ошибка не устраняется: если мы проведем опрос в 800 квартирах или даже во всех квартирах города (сплошной опрос), то полученные данные будут репрезентативны для населения, находящегося дома в момент прихода интервьюера, а не для всех жителей города.

Неслучайные ошибки могут возникнуть из-за методов сбора данных: вопросов, слишком болезненных для опрашиваемых (об отношении к властям, если опрашиваются беженцы или пострадавшие от стихийных бедствий и т.д.) или формы задания вопроса (очень трудно, чтобы всем было все понятно), или времени опроса (например, на вопрос молодым родителям, не жалеют ли они о том, что у них есть дети, можно получить разное распределение ответов в зависимости от того, проводился ли опрос долгим зимним вечером, когда все утомлены приготовлением уроков, простудами и т.д., или прекрасным летним днем, когда дети находятся на даче, в оздоровительном лагере).

Случайные ошибки – те, которые изменяются по вероятностным законам. К случайным относится ошибка выборки.

Ошибка выборки или, иначе говоря, *ошибка репрезентативности* – это разница между значением показателя, полученного по выборке, и генеральным параметром. Так, ошибка репрезентативности выборочной средней равна $\varepsilon_{\bar{x}} = \bar{x} - \mu$, выборочной относительной величины $\varepsilon_p = p - \pi$, дисперсии $\varepsilon_{s^2} = s^2 - \sigma^2$, коэффициента корреляции $\varepsilon_r = r - \rho$.

Если представить, что было проведено бесконечное число выборок равного объема из одной и той же генеральной совокупности, то показатели отдельных выборок образовали бы ряд возможных значений: выборочных средних величин $\bar{x}_1$, $\bar{x}_2$, $\bar{x}_3$, ..., относительных величин p_1, p_2, p_3 ..., дисперсий s^2_1, s^2_2, s^2_3, ..., и т. д. Каждая выборка имеет свою ошибку репрезентативности. Следовательно, можно построить ряды распределения выборок по величине ошибки репрезентативности для каждого показателя: для средней, относительной величины и т.д. В таких распределениях улавливается тенденция к концентрации ошибок около центрального значения. Число выборок с той или иной величиной ошибки репрезентативности может быть симметрично или асимметрично относительно этого центрального значения. При бесконечно большом числе выборок получится кривая частот, которая представляет *кривую выборочного распределения.* Свойства таких распределений используются для получения статистических заключений, установления вероятности той или иной величины ошибки репрезентативности.

Рассмотрим *выборочное распределение средней величины.* Такое распределение будет являться нормальным или приближаться к нему по мере увеличения объема выборки, независимо от того, имеет или нет нормальное распределение та генеральная совокупность, из которой взяты выборки. С увеличением числа выборок средняя для всех выборок будет приближаться к генеральной средней. По вы-

борочному распределению может быть рассчитана средняя квадратическая ошибка репрезентативности:

$$s = \sqrt{\frac{\sum \varepsilon_i^2 f_i}{\sum f_i}}, \qquad (7.1)$$

где ε_i^2 – квадрат ошибки репрезентативности для i-й выборки;
f_i – число выборок с одинаковым значением выборочной средней.

Среднее квадратическое отклонение выборочных средних от генеральной средней называется *средней ошибкой выборочной средней:*

$$s_{\bar{x}} = \sqrt{\frac{\sum (\bar{x}_i - \mu)^2 f_i}{\sum f_i}}. \qquad (7.2)$$

Поскольку, как правило, генеральная средняя μ неизвестна, этой формулой нельзя воспользоваться. Кроме того, в социально-экономических исследованиях из одной и той же совокупности выборки не проводятся многократно. Используют следующее соотношение: квадрат средней ошибки (дисперсия выборочных средних) прямо пропорционален дисперсии признака x в генеральной совокупности σ^2 и обратно пропорционален объему выборки n:

$$s_{\bar{x}}^2 = \frac{\sigma^2}{n}. \qquad (7.3)$$

Соответственно средняя ошибка выборочной средней равна:

$$s_{\bar{x}} = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}. \qquad (7.4)$$

Следовательно, средняя ошибка выборки тем больше, чем больше вариация в генеральной совокупности, и тем меньше, чем больше объем выборки.

Таким образом, можно утверждать, что отклонение выборочной средней $\bar{x}$ от генеральной средней μ в среднем равно $\pm s_{\bar{x}}$. Ошибка конкретной выборки может принимать различные значения, но отношение ее к средней ошибке практически не превышает ±3, если величина n достаточно большая ($n > 100$). Отношение ошибки конкретной выборки к средней квадратической ошибке называется *нормированным отклонением* и обозначается как t:

$$t = \frac{\bar{x} - \mu}{s_{\bar{x}}}. \qquad (7.5)$$

Распределение нормированного отклонения выборочной средней от генеральной средней при численности выборки $n \to \infty$ определяется уравнением Лапласа–Гаусса:

$$f(t) = \frac{1}{\sigma\sqrt{2\pi}} e^{-t^2/2}, \qquad (7.6)$$

где $f(t)$ – плотность вероятности;
σ – среднее квадратическое отклонение значений переменной x от средней в генеральной совокупности;
π и e – математические константы ($\pi = 3{,}14$, $e = 2{,}718$).

Так как средняя нормированных отклонений $t = 0$, дисперсия $\sigma_t^2 = 1$, т. е. $\sigma = 1$, то выражение (7.6) может быть записано так:

$$f(t) = \frac{1}{\sqrt{2\pi}} e^{-t^2/2}. \qquad (7.7)$$

Уравнение (7.7) называют стандартным уравнением нормальной кривой. Величина $f(t)$ достигает максимума при $t = 0$, в этом случае $e^{-t^2/2} = 1$. По мере увеличения t величина $e^{-t^2/2}$ уменьшается, и соответственно уменьшается $f(t)$. На рис. 7.1 приведен график кривой нормального распределения вероятностей. Ординаты на графике соответствуют вероятностям при том или ином значении t. Чтобы определить вероятность значений в интервале от t_1 до t_2, следует найти отношение части площади кривой, заключенной между орди-

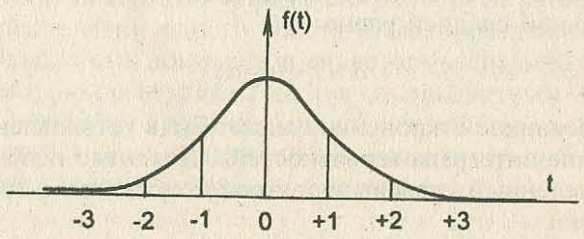


Рис. 7.1. Распределение ошибок выборочных средних

натами, соответствующими t_1 и t_2, ко всей площади кривой. Вся площадь под кривой нормального распределения вероятностей принимается за единицу.

Уравнение Лапласа – Гаусса предполагает непрерывное изменение t и неограниченное возрастание n. Поэтому площадь нормальной кривой, заключенную между ординатами t_1 и t_2, определяют, интегрируя функцию (7.7).

Имеются таблицы, которые содержат значения вероятностей для нормированных отклонений t или для интервалов от t_1 до t_2. Одна из таких таблиц приведена в приложении «Значение интеграла вероятностей». Эта таблица содержит пропорциональные доли площадей, заключенных между ординатами, соответствующими $\pm t$. Зная нормированное отклонение t, можно определить вероятность или на основе определенной вероятности установить величину t.

На пересечении строк и граф таблицы находится значение вероятности $F(t)$, соответствующее данному значению t. Для краткости записи в таблице приводятся только десятичные знаки вероятности, следовательно, к табличному значению $F(t)$ надо приписывать ноль целых. Например, чтобы определить, какая вероятность соответствует t = 1,96, надо взять строку 1,9 и графу 6 и на их пересечении прочитать значение вероятности, добавив перед первым знаком ноль целых. Если t = 1,96, то $F(t)$ = 0,9500. По мере увеличения t (уже при $t = \pm 3$) значение интеграла вероятностей приближается к единице. Чем шире пределы t, тем большая площадь под кривой охватывается ординатами, восстановленными из соответствующих значений t. Поскольку вероятность — это отношение части площади под кривой, заключенной между ординатами, ко всей площади, соответственно возрастает и вероятность.

Распределение ошибок выборочных средних имеет характер нормального распределения или приближается к нему даже в случаях, когда генеральная совокупность имеет иную форму распределения.

Из формулы (7.5) следует, что отклонение выборочной средней от генеральной средней равно:

$$\Delta = \bar{x} - \mu = ts_{\bar{x}}. \qquad (7.8)$$

Нормированное отклонение t может быть установлено по таблице «Значение интеграла вероятностей». Для этого необходимо принять определенный уровень вероятности суждения о точности данной выборки.

Вероятность, которая принимается при расчете ошибки выборочной характеристики, называют *доверительной*. Чаще всего принимают доверительную вероятность равной 0,95, 0,954, 0,997 или даже 0,999. Доверительный уровень вероятности 0,95 означает, что только в 5 случаях из 100 ошибка может выйти за установленные границы; вероятности 0,954 – в 46 случаях из 1000, при 0,997 – в 3 случаях, а при 0,999 – в 1 случае из 1000.

Чтобы вычислить ошибку выборки при принятой доверительной вероятности, нужно рассчитать величину средней ошибки $s_{\bar{x}}$. Формула для ее определения (7.4) включает дисперсию признака в генеральной совокупности σ^2, которая, как правило, неизвестна. Может быть определена только выборочная дисперсия s^2. Доказано, что соотношение между σ^2 и s^2 определяется следующим равенством:

$$s^2 = \frac{n-1}{n} \cdot \sigma^2. \qquad (7.9)$$

Отсюда

$$\sigma^2 = s^2 \frac{n}{n-1}. \qquad (7.10)$$

Если n велико, то сомножитель $n/(n-1) \approx 1$ и можно принять выборочную дисперсию в качестве оценки величины генеральной дисперсии. Подставив выражение (7.10) в формулу средней ошибки выборочной средней, получим:

$$s_{\bar{x}} = \sqrt{\frac{s^2}{n-1}} = \frac{s}{\sqrt{n-1}} \quad \text{или} \quad s_{\bar{x}} = \frac{s}{\sqrt{n}}. \qquad (7.11)$$

Соответственно

$$\Delta_{\bar{x}} = ts_{\bar{x}} = t\frac{s}{\sqrt{n}}. \qquad (7.12)$$

Рассмотрим **пример.** Для определения скорости расчетов с кредиторами предприятий одного треста была проведена случайная выборка 50 платежных документов, по которым средний срок перечисления денег оказался равен 28,2 дня со стандартным отклонением 5,4 дня. Определим средний срок прохождения всех платежей в течение данного года с доверительной вероятностью $F(t)$ = 0,95. Тогда t = 1,96; скорректированная дисперсия

$$s^2 = 5{,}4^2 \frac{50}{50-1} = 29{,}755;$$

средняя ошибка выборки

$$s_{\bar{x}} = \sqrt{\frac{29{,}755}{50}} = \sqrt{0{,}5951} = \pm 0{,}77 \text{ дня.}$$

Отклонение выборочной средней от генеральной с вероятностью 0,95 составит $\Delta_{\bar{x}} = 1{,}96 \cdot 0{,}77 = \pm 1{,}51$ дня.

Δ называется доверительной ошибкой выборки или *предельной ошибкой выборки.* Рассчитав величину Δ, мы можем записать следующее неравенство:

$$28{,}2 - 1{,}51 \le \mu \le 28{,}2 + 1{,}51;$$

$$26{,}7 \text{ дня} \le \mu \le 29{,}7 \text{ дня.}$$

Таким образом, с вероятностью 0,95 можно утверждать, что средняя продолжительность расчетов предприятия данного треста с кредиторами составляет не менее 26,7 дня и не более 29,7 дня.

Ошибка выборки для *выборочной относительной величины (доли)* определяется аналогично. Дисперсия относительной величины по данным выборки

$$s^2 = p(1 - p), \qquad (7.13)$$

где p – доля тех или иных единиц в выборке.

Выражение (7.13) получено в соответствии с обычной формулой дисперсии. Поскольку имеется в виду альтернативная или дихотомическая переменная, обозначим ее значение в одной категории единиц 0, в другой – 1. Тогда среднее значение переменной составит:

$$\frac{f_1 \cdot 0 + f_2 \cdot 1}{f_1 + f_2} = \frac{f_2}{n} = p;$$

квадрат отклонения от средней

$$s^2 = \frac{(0-p)^2 f_1 + (1-p)^2 f^2}{f_1 + f_2} = \frac{p^2 f_1 + f_2 - 2pf_2 + p^2 f_2}{n} = p^2(f_1 + f_2) +$$

$$+ f_2(1-2p) = \frac{f_2}{n^2} n + f_2(1-2p) = f_2(p+1-2p) = \frac{f_2(1-p)}{n} = p(1-p),$$

что соответствует выражению (7.13).

Средняя ошибка выборочной доли

$$s_p = \sqrt{\frac{p(1-p)}{n}}. \qquad (7.14)$$

Предельная ошибка выборочной доли с принятой доверительной вероятностью имеет вид:

$$\Delta_p = t \cdot s_p = t\sqrt{\frac{p(1-p)}{n}}. \qquad (7.15)$$

Рассмотрим **пример.** По данным выборочного изучения 100 платежных документов предприятий одного треста оказалось, что в 6 случаях сроки расчетов с кредиторами были превышены. С вероятностью 0,954 требуется установить доверительный интервал доли платежных документов треста без нарушения сроков:

$$q = \frac{6}{100} = 0{,}06, \text{ или } 6\%, \text{ p} = 0{,}94;$$

$$s_p = \sqrt{\frac{0{,}94 \cdot 0{,}06}{100}} = \pm 0{,}024;$$

$$\Delta_p = 2 \cdot 0{,}024 = \pm 0{,}048.$$

Генеральная доля платежных документов π, не выходящих за установленные сроки, с вероятностью 0,954 находится в интервале

$$0{,}892 \le \pi \le 0{,}988, \text{ или } 89{,}2\% \le \pi \le 98{,}8\%.$$

7.4. ВЛИЯНИЕ ВИДА ВЫБОРКИ НА ВЕЛИЧИНУ ОШИБКИ ВЫБОРКИ

Как указывалось в п. 7.2, при проведении выборочного наблюдения используются различные способы формирования выборочной совокупности: случайный отбор – повторный или бесповторный, механический, серийный, типический. Вид выборки влияет на величину ошибки выборки. При бесповторном отборе формула средней ошибки выборки дополняется множителем

$$\sqrt{\frac{N-n}{N-1}},$$

который корректирует величину ошибки выборки и в связи с изменением состава совокупности и вероятности попадания единиц в выборку. В серийной выборке дисперсия определяется как колеблемость между сериями:

$$s_{\breve{x}}^2 = \frac{\sum_1^r (\breve{x}_j - \bar{x})^2}{r}, \qquad (7.14')$$

где $\breve{x}_j$ – среднее значение признака x в j-й серии;
$\bar{x}$ – среднее значение в целом по выборке;
r – число отобранных серий.

Формула (7.14′) предполагает равенство серий по числу единиц, если это условие не выполняется, то в числитель выражения (7.14′) вводится вес – число единиц в j-й серии, f_j; тогда в знаменателе указывается не r, а Σf_j. Межсерийная дисперсия представляет часть общей дисперсии признака x, и потому ее использование направлено на уменьшение ошибки выборки. Однако значение r намного меньше n, так как число отобранных гнезд намного меньше числа единиц наблюдения. Этот фактор увеличивает ошибку выборки. Его действие более значительно, нежели понижающее влияние межсерийной дисперсии – в результате ошибка серийной выборки в среднем больше ошибки выборки при отборе единицами.

При типическом отборе (стратифицированная или районированная выборка) дисперсия рассчитывается как средняя из внутрирайонных дисперсий:

$$\overline{s^2} = \frac{\sum_1^m s_{x_j}^2 n_j}{\sum_1^m n_j}, \qquad (7.15')$$

где $s_{x_j}^2$ – выборочная дисперсия признака x в j-м районе;

$$s_{x_j}^2 = \frac{\sum_{i=1}^n (x_{ij} - \bar{x}_j)^2}{n_j};$$

n_j – объем выборки в j-м районе;
m – число районов.

Очевидно, что по правилу сложения дисперсий величина $\overline{s^2}$ меньше, чем величина общей дисперсии.

Величина ошибки районированной выборки меньше величины ошибки простой (нерайонированной выборки).

Часто используется сочетание районированного отбора с отбором сериями. Такой вид выборки обеспечивает преимущества в организации выборки и уменьшение ошибки выборки. Дисперсия такой выборки представляет среднюю из межсерийных дисперсий для каждого j-го района:

$$\overline{s_{\breve{x}}^2} = \frac{\sum_{j=1}^m s_{\breve{x}_j}^2 r_j}{\sum_1^m r_j}, \qquad (7.16)$$

где $s_{\breve{x}_j}^2$ – межсерийная дисперсия в j-м районе;

$$s_{\breve{x}_j}^2 = \frac{\sum_1^r (\breve{x}_{ij} - \bar{x}_j)^2}{r_j};$$

$\breve{x}_{ij}$ – средняя в i-й серии j-го района;
$\bar{x}_j$ – средняя в j-м районе;
r_j – число серий, отобранных в j-м районе;
m – число районов.

Табл. 7.2 содержит формулы средней ошибки выборки для выборочной средней и выборочной относительной величины для разных видов выборки. В приведенных формулах требуют пояснения выражения дисперсий выборочной относительной величины.

При нерайонированной серийной выборке

$$s_p^2 = \frac{\sum_j^r (p_j - p)^2}{r}, \qquad (7.17)$$

где p_j – доля единиц определенной категории в j-й серии;
p – доля единиц этой категории в выборке.

Таблица 7.2

Формулы средней ошибки выборочной средней и выборочной относительной величины

Вид выборки	Средняя ошибка	
	выборочной средней	выборочной относительной величины (доли)
Повторная – отбор единицами	$\sqrt{\frac{s^2}{n}}$	$\sqrt{\frac{p(1-p)}{n}}$
Бесповторная – отбор единицами	$\sqrt{\frac{s^2}{n}\left(\frac{N-n}{N-1}\right)}$	$\sqrt{\frac{p(1-p)}{n}\left(\frac{N-n}{N-1}\right)}$
Серийная (нерайонированная)	$\sqrt{\frac{s_{\bar{x}}^2}{r}\left(\frac{R-r}{R-1}\right)}$	$\sqrt{\frac{s_p^2}{r}\left(\frac{R-r}{R-1}\right)}$
Районированная – отбор единицами, бесповторная	$\sqrt{\frac{\overline{s^2}}{n}\left(\frac{N-n}{N-1}\right)}$	$\sqrt{\frac{\overline{s_p^2}}{n}\left(\frac{N-n}{N-1}\right)}$
Районированная – отбор сериями, бесповторная	$\sqrt{\frac{\overline{s_{\bar{x}}^2}}{r}\left(\frac{R-r}{R-1}\right)}$	$\sqrt{\frac{\overline{s_p^2}}{r}\left(\frac{R-r}{R-1}\right)}$

При районированной серийной выборке

$$\overline{s_p^2} = \frac{\sum_{j=1}^{m} s_{p_j}^2 r_j}{\sum_{j=1}^{m} r_j}, \qquad (7.18)$$

где $s_{p_j}^2$ – межсерийная дисперсия доли в j-м районе;
r_j – число серий, отобранных в j-м районе;
m – число районов.

Рассмотрим на **примере** влияние вида выборки на величину ошибки выборки. Исходные данные представлены в табл. 7.3.

Таблица 7.3

Показатели 60 предприятий легкой промышленности Санкт-Петербурга (по данным статистической отчетности за I полугодие 1995 г.)

№ п/п	Форма собственности	Оборачиваемость запасов x_1	Коэффициент покрытия x_2	№ п/п	Форма собственности	Оборачиваемость запасов x_1	Коэффициент покрытия x_2
1	государственная	5,65	0,22	31	частная	1,23	1,18
2	»	2,86	0,35	32	»	0,82	1,59
3	»	1,61	1,06	33	»	2,63	0,74
4	»	3,99	1,01	34	»	1,83	1,52
5	»	2,17	8,88	35	»	2,26	2,43
6	»	1,52	1,06	36	»	2,33	3,28
7	»	0,40	0,99	37	»	2,35	1,13
8	»	2,18	1,07	38	»	1,68	0,89
9	»	1,36	4,62	39	»	2,00	1,67
10	»	3,69	1,40	40	»	2,64	1,48
11	»	0,45	1,24	41	»	2,75	1,51
12	частная	1,00	1,16	42	»	3,29	5,96
13	»	2,05	2,00	43	»	1,26	1,38
14	»	2,36	1,43	44	»	1,90	2,39
15	»	4,90	1,76	45	»	3,27	3,62
16	»	3,12	1,26	46	»	3,49	0,46
17	»	1,36	1,89	47	»	2,92	1,26
18	»	1,56	12,36	48	смешанная	3,22	0,78
19	»	4,84	1,23	49	»	2,61	1,67
20	»	1,23	3,26	50	»	5,17	0,95
21	»	0,81	2,22	51	»	8,63	0,96
22	»	0,70	1,16	52	»	1,06	2,51
23	»	0,87	1,21	53	»	2,13	3,49
24	»	0,20	1,45	54	»	2,03	1,22
25	»	1,71	4,04	55	»	1,82	2,92
26	»	1,83	2,07	56	»	3,12	1,54
27	»	1,32	0,69	57	»	0,77	0,97
28	»	1,95	1,97	58	»	4,15	0,93
29	»	1,46	1,31	59	»	3,62	1,34
30	»	2,96	5,32	60	»	3,89	3,51

Предприятия легкой промышленности примем за генеральную совокупность. Ее характеристики:

численность	$N = 60$;
генеральные средние:	$\mu_1 = 2{,}40$ число оборотов;
	$\mu_2 = 1{,}424$;
генеральные дисперсии:	$\sigma_1^2 = 2{,}24$;
	$\sigma_2^2 = 4{,}38$;
средние квадратические	$\sigma_1 = 1{,}49$ оборотов;
отклонения:	$\sigma_2 = 2{,}09$.

Остановимся на смысле характеристик предприятий: оборачиваемость запасов рассчитывается делением продолжительности периода (полгода) на среднюю продолжительность одного периода оборота запасов. Очевидно, чем скорее оборачиваются запасы, тем выше их отдача. Коэффициент покрытия рассчитывается как отношение суммы всех источников покрытия запасов к стоимости запасов. Если значение этого показателя меньше единицы, то текущее финансовое состояние предприятия рассматривается как неустойчивое. В нашем примере вариация этого признака примерно в 2 раза превосходит вариацию предприятий по уровню оборачиваемости запасов: $v_2 = 147\%$, $v_1 = 62\%$.

Произведем 30%-ную выборку. Объем выборки составит $n = 20$ предприятий. При формировании выборки методом механического отбора каждое третье предприятие попадет в выборку. Отбор начинаем с полушага отбора, т. е. первым предприятием, попавшим в выборку, является второе по списку. Средние по выборке равны: оборачиваемость запасов $\bar{x}_1 = 2{,}16$ оборотов, коэффициент покрытия $\bar{x}_2 = 2{,}01$.

Средняя ошибка выборочной средней оборачиваемости запасов

$$s_{\bar{x}_1} = \sqrt{\frac{2{,}24}{20}\left(\frac{60-20}{60-1}\right)} = \pm 0{,}275 \text{ оборотов.}$$

Средняя ошибка выборочного среднего коэффициента покрытия

$$s_{\bar{x}_2} = \sqrt{\frac{4{,}38}{20}\left(\frac{60-20}{60-1}\right)} = \pm 0{,}385.$$

С вероятностью 0,954 можно утверждать, что средняя оборачиваемость запасов на предприятиях легкой промышленности не ниже

$\bar{x}_1 - 2s_{\bar{x}_1} = 2{,}16 - 0{,}55 = 1{,}61$ оборотов и не выше $\bar{x}_1 + 2s_{\bar{x}} = 2{,}16 + 0{,}55 = 2{,}71$ оборотов.

Действительно генеральная средняя ($\mu_1 = 2{,}40$) попадает в этот интервал.

Фактическая ошибка репрезентативности

$$\varepsilon_1 = |\bar{x}_1 - \mu_1| = |2{,}16 - 2{,}40| = 0{,}36 \text{ оборотов.}$$

Эта величина меньше предельной ошибки выборки, гарантированной с принятой доверительной вероятностью, $0{,}36 < 0{,}55$. Следовательно, выборка репрезентативна по этому признаку.

Вычислим предельную ошибку выборки коэффициента покрытия и определим доверительный интервал для этой характеристики. Его нижняя граница с той же вероятностью

$$\bar{x}_2 - 2s_{\bar{x}_2} = 2{,}01 - 0{,}77 = 1{,}24;$$

верхняя граница:

$$\bar{x}_2 + 2s_{\bar{x}_2} = 2{,}01 + 0{,}77 = 2{,}78.$$

Генеральная средняя ($\mu_2 = 1{,}424$) так же попадает в доверительный интервал.

Фактическая ошибка репрезентативности составляет:

$$\varepsilon_2 = |\bar{x}_2 - \mu| = |2{,}01 - 1{,}424| = 0{,}586.$$

Эта величина меньше предельной ошибки выборки (0,77), что дает основание считать выборку репрезентативной и по этому признаку.

В генеральной совокупности доля единиц с неустойчивым финансовым положением ($x_2 < 1$) составила $\pi = \frac{12}{60} = 0{,}20$, в выборке

$$p = \frac{3}{20} = 0{,}15.$$

Доверительный интервал для оценки доли таких предприятий в генеральной совокупности составляет с вероятностью 0,954:

$$0{,}15 \pm 2\sqrt{\frac{0{,}15 \cdot 0{,}85}{60} \cdot \left(\frac{60-20}{60-1}\right)};$$

$$0{,}15 \pm 0{,}076,$$

т. е. таких предприятий должно быть не меньше 7,4% и не больше 22,6%. Фактически их оказалось 20% от общего числа предприятий, т. е. выборка дает репрезентативный результат и по этому показателю.

Выполненная выборка формировалась как простая бесповторная механическая. Однако, наверняка статистик будет стремиться учесть структуру генеральной совокупности, поэтому более естественной была бы выборка, учитывающая выделение предприятий разных форм собственности. Тогда выборка должна быть районированной.

Рассмотрим **пример.** Генеральная совокупность состоит из 11 государственных предприятий, 36 частных, 13 смешанных. В выборке эти пропорции соблюдаются следующим образом: отобраны по 4 предприятия государственных и смешанных и 12 – частных:

Предприятия	Генеральные характеристики		Выборочные характеристики	
	средние	доли	средние	доли
Государственные	$\mu_1 = 2{,}35$	$\pi_1 = 0{,}27$	$\bar{x}_1 = 1{,}92$	$P_1 = 0{,}25$
Частные	$\mu_1 = 2{,}11$	$\pi_2 = 0{,}11$	$\bar{x}_1 = 1{,}79$	$P_2 = 0{,}08$
Смешанные	$\mu_1 = 3{,}25$	$\pi_3 = 0{,}38$	$\bar{x}_1 = 3{,}51$	$P_3 = 0{,}25$

Средняя из внутрирайонных дисперсий, рассчитанных по каждой группе предприятий в генеральной совокупности:

$$\overline{\sigma^2} = \frac{\sum \sigma_j^2 \cdot N_j}{\sum N_j} = \frac{25{,}32 + 39{,}16 + 50{,}7}{60} = 1{,}9197.$$

Эта величина меньше общей дисперсии без учета районирования ($\sigma^2 = 2{,}24$). Следовательно, и величина ошибки выборки при районированном отборе будет меньше:

$$\Delta_{\bar{x}_1} = 2\sqrt{\frac{1{,}9197}{20} \cdot \left(\frac{60-20}{60-1}\right)} = 2 \cdot (\pm 0{,}147) = \pm 0{,}294.$$

Итак, с вероятностью 0,954 генеральная средняя оборачиваемости запасов находится в интервале $2{,}16 \pm 0{,}294$; $1{,}866 \le \mu \le 2{,}454$.

Чтобы понять, насколько целесообразно в том или ином случае применение районированного отбора, можно воспользоваться корреляционным отношением η. Согласно правилу сложения дисперсий средняя из внутригрупповых дисперсий может быть представлена как

$$\overline{s^2} = s^2(1 - \eta^2), \qquad (7.17')$$

где η^2 – квадрат корреляционного отношения, равный $\delta^2 : s^2$.

Следовательно, применение районированной (типической) выборки изменяет предельную ошибку на $\sqrt{1-\eta^2}$. В нашем примере для первой переменной (оборачиваемость) имеем:

$$\sqrt{1-\eta^2} = \sqrt{1 - \frac{0{,}3181}{2{,}2378}} = 0{,}37.$$

Сопоставим полученный результат с изменением предельной ошибки выборки: Δ_{x_1} (без учета районирования) = 0,55;

Δ_x (при районировании) = 0,294, т. е. ошибка уменьшилась примерно вполовину.

Корреляционное отношение используется и при корректировке величины

$$t^* = t\sqrt{1-\eta^2}. \qquad (7.18')$$

Тогда при вероятности 0,954 и $t = 2$; $t^* = 2 \cdot \sqrt{0{,}86} = 1{,}85$, т. е. вместо $t = 2$ достаточно взять $t = 1{,}85$.

Многие выборки формируются как многоступенчатые. *Ошибка многоступенчатой выборки* может быть представлена как

$$\bar{x}_k - \mu = (\bar{x}_k - \bar{x}_{k-1}) + \ldots + (\bar{x}_2 - \bar{x}_1) + (\bar{x}_1 - \mu).$$

Она складывается из ошибок отдельных ступеней. Поэтому практически используется не больше 2–3 ступеней отбора.

Средняя ошибка выборки при двухступенчатом отборе рассчитывается по формуле

$$s_{\bar{x}} = \sqrt{\frac{s_{x_1}^2}{m} + \sum_{1}^{m} \frac{s_{x_2}^2}{n_i} \frac{1}{m}}, \qquad (7.19)$$

где $s_{x_1}^2$ – дисперсия признака x по совокупности «крупных» единиц;

$s_{x_2}^2$ – дисперсия признака x в каждой из отобранных «крупных» единиц;

n_i – число отобранных единиц наблюдения в i-й «крупной» единице;

m – число отобранных «крупных» единиц.

Таким образом, применение многоступенчатой выборки улучшает организацию выборки, но увеличивает ее ошибку.

Кроме рассмотренных, применяется *многофазовая выборка*, когда одни сведения собираются на основе изучения всех единиц выборки, а другие – только на основании изучения некоторых из этих единиц, отобранных так, что они составляют подвыборки из единиц первоначальной выборки.

При периодическом повторении выборочных обследований с целью изучения динамики явлений применяются либо *независимые выборки* – через определенные промежутки времени отбор каждый раз производится независимо от предыдущих выборок; либо *фиксированные выборки* – в этом случае повторные обследования проводятся по одной и той же выборке. В связи с тем, что в фиксированной выборке могут происходить изменения (прежде всего за счет выбытия единиц) практикуют периодическую адаптацию фиксированной выборки происходящим изменениям. Чаще для целей изучения динамики используется промежуточный вариант – *ротационная выборка* (частичное замещение). При этом нужно следовать определенному плану замещения, например, каждый раз замещать четверть выборки, тогда каждая первоначальная единица останется в выборке в четырех следующих друг за другом обследованиях.

Названные виды выборок ориентированы на отбор конкретных материальных явлений. Кроме них следует назвать как особый вид выборки *метод моментных наблюдений*. Сущность метода моментных наблюдений состоит в периодической фиксации состояний наблюдаемых единиц в отобранные моменты времени. Расчет объема такой выборки дает количество моментов. Этот вид выборочного наблюдения применяется при изучении использования производственного оборудования, либо рабочего времени (см. п. 7.13).

7.5. ЗАДАЧИ, РЕШАЕМЫЕ ПРИ ПРИМЕНЕНИИ ВЫБОРОЧНОГО МЕТОДА

При применении выборочного наблюдения возникают три основные задачи:

- определение объема выборки, необходимого для получения требуемой точности результатов с заданной вероятностью;
- определение возможного предела ошибки репрезентативности, гарантированного с заданной вероятностью, и сравнение его с величиной допустимой погрешности.
- определение вероятности того, что ошибка выборки не превысит допустимой погрешности.

Все эти задачи решаются на основе теоремы Чебышева, согласно которой $P\{|\bar{x} - \mu| < \varepsilon\} \geq 1 - h$, когда n – достаточно большое число; ε и h – сколь угодно малые положительные числа. Это соотношение, как было показано в п. 7.3, может быть выражено через формулу предельной ошибки выборки $\Delta_x = ts_x$ или $\Delta_p = ts_p$. Решение указанных задач зависит от того, какие величины в формуле предельной ошибки заданы, а какие нужно найти.

Объем выборки рассчитывается на стадии проектирования выборочного обследования. Так как

$$\Delta = t\sqrt{\frac{\sigma^2}{n}},$$

то

$$n = \frac{t^2\sigma^2}{\Delta^2}, \qquad (7.20)$$

где Δ – допустимая погрешность, которая задается исследователем исходя из требуемой точности результатов проектируемой выборки;

t – табличная величина, соответствующая заданной доверительной вероятности $F(t)$, с которой будут гарантированы оценки генеральной совокупности по данным выборочного обследования;

σ^2 – генеральная дисперсия.

Последняя величина, как правило, неизвестна. Используются какие-либо ее оценки: результаты прошлых обследований той же совокупности, если ее структура и условия развития достаточно стабильны, или же зная примерную величину средней, находят дисперсию из соотношения

$$\sigma \approx \frac{1}{3}\bar{x};$$

если известны x_{max} и x_{min}, то можно определить среднее квадратическое отклонение в соответствии с правилом «трех сигм»

$$\sigma = \frac{1}{6}(x_{max} - x_{min}),$$

так как в нормальном распределении в размахе вариации «укладывается» $6\sigma(\pm 3\sigma)$. Если распределение заведомо асимметричное, то

$$\sigma \approx \frac{1}{5}(x_{max} - x_{min}).$$

Для относительной величины принимают максимальную величину дисперсии $\sigma^2_{max} = 0{,}5 \cdot 0{,}5 = 0{,}25$.

При расчете n не следует гнаться за большими значениями t и малыми значениями Δ, так как это приведет к увеличению объема выборки, а следовательно, к увеличению затрат средств, труда и времени, вовсе не являющемуся необходимым.

Формула (7.20) не учитывает бесповторности отбора и дает максимальную величину выборки, которую можно скорректировать «на бесповторность». Так как

$$\Delta = t\frac{s}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}},$$

то на основе (7.20) получаем выражение скорректированного объема выборки (n):

$$n = \frac{n_0}{\frac{n_0 + (N-1)}{N}}, \qquad (7.21)$$

где

$$n_0 = \frac{t^2\sigma^2}{\Delta^2}.$$

При больших размерах генеральной совокупности скорректированный объем выборки незначительно отличается от n_0.

Например, для изучения структуры и стоимости покупок в универмаге из 10 000 покупателей следует отобрать определенное число человек, которое бы обеспечивало с вероятностью 0,95 определение средней стоимости покупок с точностью не менее 2 тыс. руб. Дисперсию примем по прошлому обследованию равной 625.

$$n_0 = \frac{1{,}96^2 \cdot 625}{4} = 600{,}25 \approx 601 \text{ человек};$$

тогда скорректированная численность

$$n = \frac{601}{\frac{601 + (10\,000 + 1)}{10\,000}} = 567 \text{ человек } (\approx 570 \text{ человек}).$$

При проектировании районированной выборки рассчитанный объем выборки распределяют пропорционально численности районов (*пропорциональный отбор*):

$$n_i = n \cdot \frac{N_i}{N}, \qquad (7.22)$$

где n_i – объем выборки для i-го района;
N_i – объем i-го района в генеральной совокупности;
n – общий объем выборки;
N – общий объем генеральной совокупности.

При различиях в однородности выделенных районов лучшие результаты дает распределение запланированного объема выборки между районами не только с учетом их объема, но и с учетом дисперсии признака (*оптимальный отбор*). В этом случае объем выборки в i-м районе определяется как

$$n_i = n \cdot \frac{N_i\sigma_i^2}{\sum_{i=1}^{m} N_i\sigma_i^2}, \qquad (7.23)$$

где σ_i^2 – дисперсия признака x в i-м районе.

При любом виде проектируемой выборки расчет объема выборки начинают по формуле повторного отбора (7.20). Если в резуль-

тате расчета n доля отбора превысит 5%, проводят второй вариант расчета по формуле бесповторного отбора, либо по формуле (7.21), либо как

$$n = \frac{Nt^2\sigma^2}{N\Delta^2 + t^2\sigma^2}.$$

Если доля отбора меньше 5%, к формуле бесповторного отбора не переходят, так как это не скажется существенно на величине n.

Выборка должна быть такой, чтобы выборочные показатели по всем основным характеристикам были репрезентативны. Поэтому численность выборки рассчитывают многократно исходя из допустимых ошибок разных показателей, значения которых в генеральной совокупности известны.

Например, при выборочном учете детей школьного возраста требуется определить число семей, которые надо обследовать. При этом надо учесть: а) число детей в возрасте 6 – 7 лет; б) число детей в возрасте 6–15 лет; в) число детей в возрасте 16–17 лет; г) среднедушевой доход (например, для решения вопроса о строительстве базы отдыха).

Так как репрезентируемые признаки могут иметь разную размерность, то допустимая погрешность для каждого их них задается в виде относительной величины ($\Delta : \bar{x}$) (например, планируется, что в определении среднего размера семьи ошибка должна быть не больше 2%, в определении дохода – не больше 3% и т.д.). В этом случае вместо дисперсии в формуле (7.20) берется квадрат коэффициента вариации.

Вычислив значение n, на основе каждой из характеристик получаем разные объемы выборки: 1200; 300; 700; 100. Обследовать необходимо 1200 семей, т.е. из рассчитанных численностей берется максимальная. При резких различиях необходимых объемов выборки для разных вопросов программы проводится многофазный отбор. В рассмотренном примере среднедушевой доход достаточно учитывать в одной из каждых 12 семей, попавших в выборку.

Многофазный отбор, как правило, довольно сложно организовать, может быть нарушен принцип случайности отбора. Поэтому для обеспечения репрезентативности оказывается выгоднее затратить больше средств на учет большего числа единиц совокупности. Многофазный отбор целесообразно применять, если соотношение между рассчитанными объемами выборки по крайней мере 1:6.

Поскольку расчет необходимой численности выборки основан не на точных, а на предположительных данных о колеблемости в совокупности, следует соблюдать следующие рекомендации: абсо-

лютную величину n округлять только вверх; долю отбора округлять только вниз, т.е. из предосторожности планировать несколько больший объем выборки, чем показывают расчеты.

Объем многоступенчатой выборки рекомендуется увеличить не менее чем на 10% от рассчитанной численности, поскольку, как было показано в предыдущем параграфе, многоступенчатость отбора увеличивает ошибку выборки.

После проведения выборки рассчитывают *возможные ошибки выборочных показателей* (ошибки репрезентативности), которые используются для оценки результатов выборки и для получения характеристик генеральной совокупности.

Пример. На электроламповом заводе взято для проверки 100 ламп. Средняя продолжительность их горения оказалась 1420 ч со средним квадратическим отклонением 61,03 ч. Поскольку приемщика продукции интересует качество всей партии (50 тыс. электроламп), оценивают точность полученной средней. Средняя возможная ошибка вычисленной выборочной средней

$$s_{\bar{x}} = \frac{61{,}03}{\sqrt{100}} = \pm 6{,}1 \text{ ч}.$$

С вероятностью 0,954 предел возможной ошибки

$$\Delta_{\bar{x}} = 2 \cdot 6{,}1 = \pm\ 12{,}2 \text{ ч}.$$

С вероятностью 0,954 можно утверждать, что средняя продолжительность горения 1 электролампы во всей партии будет находиться в пределах от 1408 до 1432 ч; 46 электроламп из 1000 могут иметь срок горения, выходящий за эти пределы.

Приемщика продукции интересуют отклонения от вычисленных пределов только в сторону сокращения продолжительности горения. Меньше чем 1408 ч могут гореть 23 лампы из 1000. На основании этого приемщик продукции решает вопрос о годности всей партии электроламп.

Решение вопроса может быть уточнено: определим, у какой доли ламп срок службы окажется меньше установленного лимита. Для потребителя продукции таким лимитом являются 1410 ч, продукция с меньшим сроком горения неприемлема.

При контрольной проверке 100 ламп 100 ламп горели менее 1410 ч, их удельный вес $p = 0{,}1$, или 10%. Средняя возможная ошибка этой доли

$$s_p = \sqrt{\frac{0{,}1 \cdot 0{,}9}{100}} = \pm 0{,}03, \text{ или } \pm 3\%.$$

С вероятностью 0,954 предел ошибки доли $\Delta_p = 2 \cdot 0{,}03 = \pm 0{,}06$, или ±6%. Следовательно, во всей партии можно ожидать от 4 до 16% некачественных электроламп.

Чаще всего делают заключение об удовлетворительности выборки, сопоставляя получившиеся пределы ошибок выборочных показателей с величинами допустимых погрешностей. Может получиться, что предел ошибки, рассчитанный с заданной вероятностью, окажется выше допустимого размера погрешности. В этих случаях определяют *вероятность того, что ошибка выборки не превзойдет допускаемую погрешность.* Решение этой задачи и заключается в отыскании $F(t)$ на основе формулы предела ошибки выборки:

$$\Delta = t\sqrt{\frac{s^2}{n}},$$

где Δ – допустимый размер погрешности оцениваемого показателя;
s^2 – дисперсия показателя, рассчитанная по данным выборочного наблюдения;
n – объем проведенной выборки.

Продолжим **пример** с оценкой качества электроламп. Если при приемке партии электроламп ставится условие, что минимальный срок горения электроламп 1410 ч, то, учитывая среднюю продолжительность горения по выборке ($\bar{x} = 1420$ ч), допустимая погрешность равна 10 ч: 1410 – 1420 = – 10 ч.

Как было установлено выше, с вероятностью 0,954 предел возможной ошибки выборочной средней составил 12,2 ч, что превосходит допустимую погрешность. Является ли это основанием для браковки всей партии? Для ответа на этот вопрос определяют вероятность риска при приемке продукции:

$$10 = t\sqrt{\frac{61{,}03^2}{100}}, \text{ отсюда } t = 1{,}64.$$

Соответствующая доверительная вероятность 0,899 (см. приложение, табл. 1). Вероятность того, что средний срок горения лампы меньше 1410 ч, равна:

$$1 - \frac{1 + F(t)}{2} = 1 - \left(0{,}5 + \frac{0{,}899}{2}\right) = 0{,}05.$$

Следовательно, из 100 ламп 5 могут гореть менее 1410 ч – риск появления некачественной продукции достаточно высок.

Аналогично можно определить вероятность того, что предел ошибки доли не превысит допускаемую погрешность доли.

Оценки надежности выборочных показателей, как показано на примере, позволяют принять обоснованные решения в отношении генеральной совокупности.

7.6. РАСПРОСТРАНЕНИЕ ДАННЫХ ВЫБОРОЧНОГО НАБЛЮДЕНИЯ НА ГЕНЕРАЛЬНУЮ СОВОКУПНОСТЬ

Конечной целью выборочного наблюдения является характеристика генеральной совокупности на основе данных, полученных по выборке. При этом исходят из того, что все средние и относительные показатели, полученные по выборке, являются несмещенными и эффективными характеристиками генеральной совокупности.

Выборочные средние и относительные величины распространяются на генеральную совокупность обязательно с учетом предела их возможной ошибки. Приводится выборочный показатель со справкой о пределах ошибки с указанием доверительной вероятности: $\bar{x} \pm \Delta_{\bar{x}}$, $p \pm \Delta_p$. Или же указывают границы значений генеральной характеристики с определенной вероятностью $F(t)$:

$$\bar{x} - \Delta_{\bar{x}} \le \mu \le \bar{x} + \Delta_{\bar{x}};$$

$$p - \Delta_p \le \pi \le p + \Delta_p.$$

Последняя форма записи является основной.

Иногда требуется указать только один (верхний или нижний) предел характеристики генеральной совокупности. При испытании качества продукции часто нас не интересуют положительные ошибки выборки (качество фактически выше, чем получилось по выборке), беспокоит нижний предел, как в примере, рассмотренном в предыдущем параграфе. В некоторых случаях, напротив, интерес вызывают верхние границы оцениваемых показателей, например при анализе расхода материалов. Так что при характеристике генеральной совокупности всегда указывают неблагоприятный предел.

На основе выборки могут быть получены и значения объемных показателей, т. е. подсчетов для генеральной совокупности. Такой расчет осуществляется двумя способами: путем прямого расчета и способом коэффициентов. *Прямой расчет* заключается в том, что

выборочная средняя или доля умножается на объем генеральной совокупности:

$$\bar{x}N = \sum_{1}^{N} x_i.$$

Так как средняя величина имеет ошибку репрезентативности $\pm \Delta_x$, то можно считать, что итоговый подсчет в генеральной совокупности находится в пределах

$$(\bar{x} - \Delta_{\bar{x}})N \le \sum_{i=1}^{N} x_i \le (\bar{x} + \Delta_{\bar{x}})N. \qquad (7.24)$$

Итоговый подсчет по генеральной совокупности можно получить на основе итогового подсчета по выборке, разделив его на долю отбора единиц совокупности

$$\sum_{1}^{n} x / (n/N).$$

Прежде чем проводить расчет объемных показателей для генеральной совокупности, нужно убедиться, что структура выборки соответствует структуре генеральной совокупности. При наличии значительных смещений в структуре выборки в долях отдельных групп (0,03 и выше) следует применить *метод перевзвешивания*, т. е. рассчитывать генеральную среднюю на основе выборочных средних по группам и удельного веса этих групп в генеральной совокупности:

$$\mu = \bar{x}_1 w_1 + \ldots + \bar{x}_m w_m,$$

где $w_i = N_i / N$.

При *способе коэффициентов* также используются не только выборочные данные, но и сведения о генеральной совокупности.

Этот способ основан на связи признаков друг с другом. Например, в результате выборочного обследования семей города получены размер среднедушевого дохода ($\bar{x}$), средний доход семьи ($\bar{y}$) и среднее число человек в семье ($\bar{z}$). Так что $\bar{x} = \bar{y}/\bar{z}$.

Зная численность населения города, требуется рассчитать общую величину денежного дохода населения. Очевидно, это можно сделать, умножив душевой доход на общее число жителей в городе: $\bar{x}N$. Общий доход можно получить, суммируя доход отдельных семей; численность населения можно получить, суммируя данные о числе членов семей.

Тогда

$$\sum_{1}^{N} x = \frac{\sum_{1}^{n} y}{\sum_{1}^{n} z} \sum_{1}^{N} z.$$

Средний душевой расход $\sum_{1}^{n} y / \sum_{1}^{n} z$ представляет собой коэффициент, подсчитанный по выборке, который связывает две характеристики. Этот коэффициент рассчитывается как отношение двух итоговых подсчетов по выборке:

$$\sum_{1}^{n} y : \sum_{1}^{n} z.$$

Следовательно,

$$\sum_{1}^{N} x_i = \frac{\sum_{1}^{n} y}{\sum_{1}^{n} z} \sum_{1}^{N} z_i = \sum_{1}^{n} y \frac{\sum_{1}^{N} z_1}{\sum_{1}^{n} z_i}.$$

Последний сомножитель не что иное, как обратная величина доли отбора, рассчитанной по значениям признака z.

Итак, итоговый подсчет по генеральной совокупности может быть получен делением соответствующего итогового подсчета по выборке на долю отбора. При прямом расчете берется доля отбора единиц совокупности, при способе коэффициентов – доля отбора по значению какого-либо признака.

Эффективность способа коэффициентов по сравнению с методом прямого расчета зависит от того, насколько тесно связаны между собой признаки, лежащие в основе расчета коэффициента, т.е. признак, по которому подсчитывается итог, и признак, по которому определяется доля отбора. Эффект проявляется, если коэффициент корреляции между ними больше 0,8.

Способ коэффициентов используется для корректировки данных сплошного наблюдения. Например, перепись скота дала сведения, что поголовье свиней в районе составляет 10 000, в том числе в тех

хозяйствах, которые потом были охвачены контрольным обходом, сплошное наблюдение показало число свиней 1100. Контрольный обход дал уточненную цифру: не 1100, а 1107 свиней. Тогда поправочный коэффициент

$$k = \frac{7}{1100} = 0{,}0064.$$

Отсюда скорректированная численность поголовья свиней во всем районе

$$N = N' + \Delta N;\ \Delta N = kN' = 0{,}064 \cdot 10\ 000 = 64.$$

$$\text{N} = 10\ 000 + 64 = 10\ 064 \text{ голов.}$$

7.7. МАЛАЯ ВЫБОРКА

Таблицы интеграла вероятностей используются для выборок большого объема из бесконечно большой генеральной совокупности. Но уже при $n < 100$ получается несоответствие между табличными данными и вероятностью предела; при $n \leq 30$ погрешность становится значительной. Несоответствие вызывается главным образом характером распределения единиц генеральной совокупности. При большом объеме выборки особенность распределения в генеральной совокупности не имеет значения, так как распределение отклонений выборочного показателя от генеральной характеристики при большой выборке всегда оказывается нормальным.

В выборках небольшого объема $n \leq 30$ характер распределения генеральной совокупности сказывается на распределении ошибок выборки. Поэтому для расчета ошибки выборки при небольшом объеме наблюдения (уже менее 100 единиц) отбор должен проводиться из совокупности, имеющей нормальное распределение.

Теория малых выборок разработана английским статистиком В. Госсетом (писавшим под псевдонимом Стьюдент) в начале XX в. В 1908 г. им построено специальное распределение, которое позволяет и при малых выборках соотносить t и доверительную вероятность $F(t)$. При $n > 100$ таблицы распределения Стьюдента дают те же результаты, что и таблицы интеграла вероятностей Лапласа, при $30 \leq n \leq 100$ различия незначительны. Поэтому практически к малым выборкам относят выборки объемом менее 30 единиц (безусловно, большой считается выборка с объемом более 100 единиц).

Использование малых выборок в ряде случаев обусловлено характером обследуемой совокупности. Так, в селекционной работе «чистого» опыта легче добиться на небольшом числе делянок. Производственный и экономический эксперимент, связанный с экономическими затратами, также проводится на небольшом числе испытаний.

Как уже отмечалось, в случае малой выборки только для нормально распределенной генеральной совокупности могут быть рассчитаны и доверительные вероятности, и доверительные пределы генеральной средней.

Плотность вероятностей распределения Стьюдента описывается функцией

$$f(t,n) = B_n\left(1+\frac{t^2}{n-1}\right)^{-n/2}, \qquad (7.25)$$

где t – текущая переменная;
n – объем выборки;
B – величина, зависящая лишь от n.

Распределение Стьюдента имеет только один параметр: *d.f.* – число степеней свободы (иногда обозначается k).

Это распределение, как и нормальное, симметрично относительно точки $t = 0$, но оно более пологое. При увеличении объема выборки, а следовательно, и числа степеней свободы распределение Стьюдента быстро приближается к нормальному. Число степеней свободы равно числу тех индивидуальных значений признаков, которыми нужно располагать для определения искомой характеристики.

Так, для расчета дисперсии должна быть известна средняя величина. Поэтому при расчете дисперсии $d.f. = n - 1$.

Таблицы распределения Стьюдента публикуются в двух вариантах:

1) аналогично таблицам интеграла вероятностей приводятся значения t и соответствующие вероятности $F(t)$ при разном числе степеней свободы;

2) значения t приводятся для наиболее употребимых доверительных вероятностей 0,90; 0,95 и 0,99 или для $1 - 0{,}9 = 0{,}1$, $1 - 0{,}95 = 0{,}05$ и $1 - 0{,}99 = 0{,}01$ при разном числе степеней свободы. Такого рода таблица приведена в приложении (табл. 2), а также значение

t-критерия Стьюдента при уровне значимости 0,10; 0,05; 0,01.

При малых выборках расчет средней возможной ошибки основан на выборочных дисперсиях, поэтому

$$s_{\bar{x}} = \sqrt{\frac{s_x^2}{n-1}}. \qquad (7.26)$$

Приведенная формула используется для определения предела возможной ошибки выборочного показателя:

$$\Delta_x = ts_x .$$

Порядок расчетов тот же, что и при больших выборках.

Пример. Для изучения интенсивности труда было организовано наблюдение за 10 отобранными рабочими. Доля работавших все время оказалась равной 0,40, дисперсия $0,4 \cdot 0,6 = 0,24$. По табл. 2 приложения находим для $F(t) = 0,95$ и $d.f. = n - 1 = 9$, $t = 2,26$. Рассчитаем среднюю ошибку выборки доли работавших все время:

$$s_p = \sqrt{\frac{0,24}{10-1}} = \pm 0,16.$$

Тогда предельная ошибка выборки $\Delta_p = 2,26 \cdot 0,16 = \pm 0,36$. Таким образом, с вероятностью 0,95 доля рабочих, работавших без простоев, в данном цехе предприятия находится в пределах

$$39,64\% \le \pi \le 40,36\%$$

или

$$39,6\% \le \pi \le 40,4\%.$$

Если бы мы использовали для расчета доверительных границ генерального параметра таблицу интеграла вероятностей, то *t* было бы равно 1,96 и $\Delta_p = \pm 0,31$, т. е. доверительный интервал был бы несколько уже.

Малые выборки широко используются для решения задач, связанных с испытанием статистических гипотез, особенно гипотез о средних величинах.

7.8. СТАТИСТИЧЕСКАЯ ПРОВЕРКА ГИПОТЕЗ (ОБЩИЕ ПОНЯТИЯ)

Оценку генерального параметра получают на основе выборочного показателя с учетом ошибки репрезентативности. В другом случае в отношении свойств генеральной совокупности выдвигается некоторая гипотеза о величине средней, дисперсии, характере распределения, форме и тесноте связи между переменными. Проверка гипотезы осуществляется на основе выявления согласованности эмпирических данных с гипотетическими (теоретическими). Если расхождение между сравниваемыми величинами не выходит за пределы случайных ошибок, гипотезу принимают. При этом не делается никаких заключений о правильности самой гипотезы, речь идет лишь о согласованности сравниваемых данных. Основой проверки статистических гипотез являются данные случайных выборок. При этом безразлично, оцениваются ли гипотезы в отношении реальной или гипотетической генеральной совокупности. Последнее открывает путь применения этого метода за пределами собственно выборки: при анализе результатов эксперимента, данных сплошного наблюдения, но малой численности. В этом случае рекомендуется проверить, не вызвана ли установленная закономерность стечением случайных обстоятельств, насколько она характерна для того комплекса условий, в которых находится изучаемая совокупность.

Особенно часто процедура проверки статистических гипотез применяется для оценки существенности расхождений сводных характеристик отдельных совокупностей (групп): средних, относительных величин. Такого рода задачи, как правило, возникают в социальной статистике. Трудоемкость статистико-социологических исследований приводит к тому, что почти все они строятся на несплошном учете. Поэтому проблема доказательности выводов в социальной статистике стоит особенно остро. Применяя процедуру проверки статистических гипотез, следует помнить, что она может гарантировать результаты с определенной вероятностью лишь по «беспристрастным» выборкам, на основе объективных данных.

Статистической гипотезой называется предположение о свойстве генеральной совокупности, которое можно проверить, опираясь на данные выборки. Обозначается гипотеза буквой *H* от латинского слова hypothesis. Так, может быть выдвинута гипотеза о том, что средняя в генеральной совокупности равна некоторой величине $H: \mu = a$, или о том, что генеральная средняя больше некоторой величины $H: \mu > b$.

Различают *простые* и *сложные* гипотезы. Гипотеза называется *простой,* если она однозначно характеризует параметр распределения случайной величины. Например, $H : \mu = a$. *Сложная* гипотеза состоит из конечного или бесконечного числа простых гипотез, при этом указывается некоторая область вероятных значений параметра. Например, $H : \mu > b$. Эта гипотеза состоит из множества простых гипотез $H : \mu = c$, где c — любое число, большее b.

Гипотезы о параметрах генеральной совокупности называются *параметрическими,* о распределениях – *непараметрическими.*

Гипотеза о том, что две совокупности, сравниваемые по одному или нескольким признакам, не отличаются, называется *нулевой гипотезой* (или нуль-гипотезой). Она обозначается H_0. При этом предполагается, что действительное различие сравниваемых величин равно нулю, а выявленное по данным отличие от нуля носит случайный характер. Например, $H_0 : \mu_1 = \mu_2$ и т. д.

Нулевая гипотеза отвергается тогда, когда по выборке получается результат, который при истинности выдвинутой нулевой гипотезы маловероятен. Границей невозможного или маловероятного обычно считают $\alpha = 0{,}05$, т.е. 5%, или 0,01, 0,001. Если ориентироваться на правило «трех сигм», то вероятность ошибки α должна быть равна 0,0027. Однако для этого уровня вероятности ошибки значения критериев редко табулируются: как правило, значения критериев в статистико-математических таблицах рассчитаны для вероятностей ошибки 0,05; 0,01; 0,001.

Статистическим критерием называют определенное правило, устанавливающее условия, при которых проверяемую нулевую гипотезу следует либо отклонить, либо не отклонить. Критерий проверки статистической гипотезы определяет, противоречит ли выдвинутая гипотеза фактическим данным или нет.

Проверка статистических гипотез складывается из следующих этапов:

- формулируется в виде статистической гипотезы задача исследования;
- выбирается статистическая характеристика гипотезы;
- выбираются испытуемая и альтернативная гипотезы на основе анализа возможных ошибочных решений и их последствий;
- определяются область допустимых значений, критическая область, а также критическое значение статистического критерия *(t, F, χ^2)* по соответствующей таблице;
- вычисляется фактическое значение статистического критерия;
- проверяется испытуемая гипотеза на основе сравнения фактического и критического значений критерия, и в зависимости от

результатов проверки гипотеза либо отклоняется, либо не отклоняется.

При проверке гипотез по одному из критериев возможны *два ошибочных решения:*

1) неправильное отклонение нулевой гипотезы: ошибка 1-го рода;

2) неправильное принятие нулевой гипотезы: ошибка 2-го рода.

В то время, как фактически нулевая гипотеза верна (1) и нулевая гипотеза не верна (2), принимают два ошибочных решения: 1) нулевая гипотеза отклоняется и принимается альтернативная гипотеза; 2) нулевая гипотеза не отклоняется. Возможные решения представлены в табл. 7.4.

Таблица 7.4

Возможные выводы при проверке гипотез

Решение по критерию	Фактически	
	H_0 верна	H_0 не верна
H_0 отклоняется	Ошибка 1-го рода	Правильное решение
H_0 не отклоняется	Правильное решение	Ошибка 2-го рода

Если, например, установлено, что новое минеральное удобрение лучше, хотя на самом деле его действие не отличается от старого, то это ошибка 1-го рода. Если мы решили, что оба вида удобрений одинаковы, то допущена ошибка 2-го рода.

Вероятности, соответствующие неверным решениям, называются *риском 1* и *риском 2. Риск 1* равен вероятности ошибки α *(уровню значимости), риск 2* равен вероятности ошибки β. Поскольку α всегда больше нуля, то всегда есть риск ошибки β. При заданных α и объеме выборки n значение β будет тем больше, чем меньше принятое α. Если n велико, то α и β могут быть сколь угодно малыми, т.е. решения будут более обоснованными. При малом объеме выборки и малом α возможность установить фактически существующие различия мала.

Обычно задают значение α и пытаются сделать возможно β малым. Вероятность $1 - \beta$ называется *мощностью критерия:* чем она больше, тем меньше вероятность ошибки второго рода.

Альтернативная гипотеза H_1 может быть сформулирована поразному в зависимости от того, какие отклонения от гипотетической величины нас особенно беспокоят: положительные, отрицатель-

ные либо и те, и другие. Соответственно альтернативные гипотезы могут быть записаны как

$$H_1 : \mu > a, \quad H_1 : \mu < a, \; H_1 : \mu \neq a.$$

От того, как формулируется альтернативная гипотеза, зависят границы критической области и области допустимых значений.

Критической областью называется область, попадание значения статистического критерия в которую приводит к отклонению H_0. Вероятность попадания значения критерия в эту область равна принятому уровню значимости.

Область допустимых значений дополняет критическую область. Если значение критерия попадает в область допустимых значений, это свидетельствует о том, что выдвинутая гипотеза H_0 не противоречит фактическим данным (H_0 не отклоняется).

Точки, разделяющие критическую область и область допустимых значений, называются *критическими точками* или *границами критической области.* В зависимости от формулировки альтернативной гипотезы критическая область может быть *двухсторонняя* или *односторонняя* (левосторонняя либо правосторонняя).

Если вычисляемое значение критерия попадает в критическую область, нулевая гипотеза отклоняется, она противоречит фактическим данным.

7.9. ПРОВЕРКА ГИПОТЕЗЫ О ЗАКОНЕ РАСПРЕДЕЛЕНИЯ

Одна из важнейших задач анализа вариационных рядов заключается в выявлении закономерности распределения и определении ее характера. Основной путь в выявлении закономерности распределения – построение вариационных рядов для достаточно больших совокупностей. Большое значение для выявления закономерностей распределения имеет правильное построение самого вариационного ряда: выбор числа групп и размера интервала варьирующего признака.

Когда мы говорим о характере, типе закономерности распределения, то имеем в виду отражение в нем общих условий, определяющих вариацию. При этом речь всегда идет о распределениях качественно однородных явлений. Общие условия, определяющие тип закономерности распределения, познаются анализом сущности явления, тех его свойств, которые определяют вариацию изучаемого признака. Следовательно, должна быть выдвинута какая-то научная

гипотеза, обосновывающая определенный тип теоретической кривой распределения.

Под теоретической кривой распределения понимается графическое изображение ряда в виде непрерывной линии изменения частот в вариационном ряду, функционально связанного с изменением вариантов (значений признака). Теоретическое распределение может быть выражено аналитически – формулой, которая связывает частоты вариационного ряда и соответствующие значения признака. Такие алгебраические формулы носят название ***законов распределения.***

Большое познавательное значение имеет сопоставление фактических кривых распределения с теоретическими.

Как уже отмечалось, часто пользуются типом распределения, которое называется *нормальным.* Формула функции плотности нормального распределения:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{(x-\mu)^2}{2\sigma^2}}.$$

Следовательно, кривая нормального распределения может быть построена по двум параметрам – средней арифметической μ и среднему квадратическому отклонению σ.

Гипотезы о распределениях заключаются в том, что выдвигается предположение о том, что распределение в генеральной совокупности подчиняется какому-то определенному закону. Проверка гипотезы состоит в том, чтобы на основании сравнения фактических (эмпирических) частот с предполагаемыми (теоретическими) частотами сделать вывод о соответствии фактического распределения гипотетическому распределению. Может проводиться и сравнение частостей.

Под гипотетическим распределением необязательно понимается нормальное распределение. Может быть выдвинута гипотеза о биномиальном распределении, распределении Пуассона и т.д. Причина частого обращения к нормальному распределению в том, что в этом типе распределения выражается закономерность, возникающая при взаимодействии множества случайных причин, когда ни одна из них не имеет преобладающего влияния. Закон нормального распределения лежит в основе многих теорем математической статистики, применяемых для оценки репрезентативности выборок, при измерении связей и т. д. В социально-экономической статистике нормальное распределение встречается редко, но сравнение с ним важно для выяснения степени и характера отклонения от него фактического распределения.

В главе 5 отмечалось, что близость средней арифметической величины, медианы и моды указывает на вероятное соответствие изучаемого распределения нормальному закону. Но более полная и точная проверка соответствия распределения гипотезе о нормальном законе производится с использованием специальных критериев, из которых рассмотрим наиболее употребимый критерий χ^2 (хи-квадрат) К. Пирсона.

Для проверки гипотезы о соответствии эмпирического распределения закону нормального распределения необходимо частоты (частости) фактического распределения сравнить с частотами (частостями) нормального распределения. Значит, нужно по фактическим данным вычислить теоретические частоты кривой нормального распределения $\hat{f}$ по формуле (для дискретных рядов):

$$\hat{f} = \frac{ni}{s}\frac{1}{\sqrt{2\pi}}e^{-t^2/2} = \frac{ni}{s}f(t), \qquad (7.27)$$

где n – объем выборки;
i – величина интервала вариационного ряда.

Значение ординат кривой нормального распределения $f(t)$ можно получить по таблицам значения функции:

$$f(t) = \frac{1}{\sqrt{2\pi}}e^{-t^2/2}.$$

Проверяемая гипотеза формулируется как H_0: $f_j = \hat{f}_j$, альтернативная – как H_1: $f_j \neq \hat{f}_j$.

Проверка гипотезы требует, чтобы был построен теоретический ряд распределения с частотами $\hat{f}_j$, соответствующими нормальному закону, при тех же значениях параметров распределения

$$\sum\hat{f}_j = \sum f_j = n;\ s;\ x.$$

Методика построения теоретического ряда такова:

1. По фактическому интервальному ряду (табл. 5.6) вычисляются значения t для каждой группы хозяйств по формуле (для интервальных рядов):

$$t_j = \frac{x_j - x}{s}$$ – для начала и конца интервала.

2. Вычисляется вероятность попадания единицы наблюдения в данный интервал при выполнении гипотезы о нормальном законе:

$$P_j = \frac{F(t_j) - F(t_{j+1})}{2},$$

где $|t_j| > |t_{j+1}|$.

3. Определяется теоретическая частота в данной группе, равная произведению объема совокупности на вероятность попадания в данный интервал:

$$\hat{f}_j = \sum_{j=1}^{k} f_j P_j.$$

4. Находится значение критерия χ^2 по формуле

$$\chi^2 = \sum_{j=1}^{k}\frac{(f_j - \hat{f}_j)^2}{\hat{f}_j}, \qquad (7.28)$$

где k – число категорий ряда распределения;
j – номер категории;
f_j – частота эмпирического распределения;
$\hat{f}_j$ – частота теоретического распределения.

При расчете χ^2 частоты можно заменить частостями:

$$\chi^2 = n\sum_{j=1}^{k}\frac{(p_j - \pi_j)^2}{\pi_j}, \qquad (7.29)$$

где p_j – частости эмпирического распределения;
π_j – вероятности теоретического распределения.

При этом, согласно Ф. Йейтсу (Jates), группы с теоретическими частотами менее 5 принято объединять, что снижает влияние случайных ошибок (см. [6]).

Если все эмпирические частоты равны соответствующим теоретическим частотам, то χ^2 равно нулю. Очевидно, что чем больше отличаются эмпирические и теоретические частоты, тем χ^2 больше; если расхождение несущественно, то χ^2 должно быть малым. Имеются специальные таблицы критических значений χ^2 при 5%-ном и 1%-ном уровнях значимости. Критические значения зависят от чис-

ла степеней свободы (d.f. – degrees of freedom) и уровня значимости.

Число степеней свободы рассчитывается так: если эмпирический ряд распределения имеет k категорий, то k эмпирических частот $f_1, f_2, ..., f_k$ должны быть связаны следующим соотношением: $\sum_{1}^{k} f_j = n$. Если параметры теоретического распределения известны, то только $k - 1$ частот могут принимать произвольные значения, т. е. свободно варьировать, а последняя частота может быть найдена из указанного соотношения. Поэтому говорят, что система из k частот благодаря наличию одной связи теряет одну «степень свободы» и имеет только $k - 1$ степеней свободы. Кроме того, если при нахождении теоретических частот p параметров теоретического распределения неизвестны, то они должны быть найдены по данным эмпирического ряда. Это накладывает на эмпирические частоты еще p связей, благодаря чему система теряет еще p степеней свободы. Таким образом, число свободно варьируемых частот (а значит, и число степеней свободы) становится равным:

$$d.f. = (k - 1) - p = k - (p + 1). \qquad (7.30)$$

Полученное значение критерия χ^2 сравнивается с табличным при числе степеней свободы, равном числу групп (с условием Ф. Йейтса), за минусом трех – по числу фиксированных параметров в формуле нормального закона распределения и с учетом равенства сумм теоретических и фактических частот (см. приложение, табл. 4).

В первой графе этой таблицы дано число степеней свободы, а в заголовках граф – уровни значимости. Если фактическое значение χ^2 превышает табличное при том же числе степеней свободы, то вероятность соответствия распределения нормальному закону меньше указанной. Результаты расчета χ^2 по данным табл. 5.6 (глава 5) приведены в табл. 7.5 при $\bar{x} = 30{,}3$; $s = 8{,}44$.

Сумма теоретических частот нормального распределения меньше суммы фактических частот, так как нормальный закон не ограничен рамками фактических минимума и максимума.

Число групп после объединения малочисленных составило 7. Критическое значение χ^2 по табл. 4 приложения при 7–3 = 4 степеням свободы и значимости 0,05 составляет 9,49. Значит, вероятность расхождения распределения с нормальным меньше 0,05, и вероятность соответствия его нормальному закону больше 0,95. Табличное значение χ^2 для значимости 0,1 равно 7,78, что также больше фактического.

Таблица 7.5

Проверка соответствия распределения хозяйств по урожайности зерновых культур нормальному закону

Группы хозяйств	f_j	t_j	$t_j + i$	P_j	$\hat{f}_j$	$(f_j - \hat{f}_j)^2 / \hat{f}_j^2$
1	6	–2,41	–1,81	0,0235	3	0,071
2	9	–1,81	–1,22	0,0798	11	
3	20	–1,22	–0,63	0,1531	22	0,182
4	41	–0,63	–0,04	0,2197	32	2,531
5	26	–0,04	0,56	0,2282	33	1,485
6	21	0,56	1,15	0,1627	23	0,174
7	14	1,15	1,74	0,0842	12	0,333
8	5	1,74	2,33	0,0310	4	0,200
9	1	2,33	2,93	0,0082	1	
Σ	143	×	×	0,9904	141	4,976

Ясно, что гипотеза о соответствии распределения хозяйств по урожайности нормальному закону не может быть отклонена.

Какое практическое значение может иметь произведенная проверка гипотезы? Во-первых, соответствие нормальному закону позволяет прогнозировать, какое число хозяйств (или доля совокупности) попадает в тот или иной интервал значений признака. Во-вторых, нормальное распределение возникает при действии на вариацию изучаемого показателя множества независимых факторов. Из этого следует, что нельзя существенно снизить вариацию урожайности, воздействуя только на один-два управляемых фактора, скажем удобрения или энергозатраты.

С помощью критерия χ^2 можно проверять не только гипотезу о согласии эмпирического распределения с нормальным законом, но и с любым другим известным законом распределения – равномерным распределением, распределением Пуассона и т. д. Например, суд рассматривает жалобу посетителей казино на то, что, по их мнению, игральная кость, которой там пользуются, фальшива, некоторые числа очков, якобы, выпадают чаще, чем другие, и этим пользуются крупье, обирающие игроков.

Суд назначает экспертизу игральной кости: эксперт делает 600 бросков и записывает число выпавших единиц, двоек, троек и т. д.

Полученное эмпирическое распределение сравнивается с теоретическим, т. е. равномерным: в правильной кости вероятность выпадения каждого числа очков должна быть равна 1/6, при 600 бросках это даст по 100 выпадений каждого числа очков. С помощью критерия χ^2 проверяется нулевая гипотеза о том, что различия эмпирического и теоретического распределений случайны, т. е. не являются систематическим результатом фальсификации формы кости или положения центра тяжести в ней; H_0: $f_{факт} = f_{теор}$. Результаты испытания и расчет χ^2 приводятся в табл. 7.6.

Таблица 7.6

Результаты испытания игральной кости

Число очков	Количество выпадений, $f_{факт}$	$f_{теор}$	$f_{факт} - f_{теор}$	$\frac{(f_{факт} - f_{теор})^2}{f_{теор}}$
1	101	100	1	0,01
2	86	100	–14	1,96
3	107	100	7	0,49
4	94	100	–6	0,36
5	97	100	–3	0,09
6	117	100	17	2,89
Итого	600	600	0	5,80

Табличное значение χ^2 при уровне значимости 0,05 (это вероятность ошибочного отклонения нулевой гипотезы при условии, что она верна) и при 6 – 2 = 4 степенях свободы (фиксировано 2 параметра: сумма числа бросков 600 и вероятность каждого числа очков – 1/6) составляет 9,49. Вычисленное значение χ^2 = 5,8, что значительно ниже табличного. Следовательно, нулевая гипотеза не отклоняется: распределение бросков по числу выпавших очков нельзя считать неравномерным. Обвинение игроков против служащих казино не подтверждено достаточно надежно, но не доказано и то, что кость правильная. Можно назначить более дорогую экспертизу – сделать 100 000 бросков кости, но можно и согласиться, что вероятность ошибочного признания правильности кости мала – всего 5% – и отклонить обвинение.

Выбор закона распределения проводится на основе теоретического анализа. Кроме того, целесообразно руководствоваться следующей рекомендацией: выражение, определяющее функцию плотности распределения, должно зависеть от возможно меньшего числа параметров. Например, экспоненциальное распределение зависит от одного параметра – средней величины; нормальное и логнормальное распределение – от двух параметров.

7.10. ПРОВЕРКА ГИПОТЕЗЫ О СВЯЗИ НА ОСНОВЕ КРИТЕРИЯ χ^2 (ХИ-КВАДРАТ)

Одним из основных приложений критерия χ^2 является его использование при анализе таблиц сопряженности двух переменных для установления факта наличия и уровня значимости взаимосвязи. Как правило, критерий χ^2 применяется для анализа таблиц сопряженности номинальных признаков, однако он может быть использован и при анализе взаимосвязи порядковых или интервальных (количественных) переменных, несмотря на то, что для последних случаев существуют более мощные тесты.

Рассмотрим общий случай – таблицу сопряженности двух переменных размером $r \times s$. Обозначим:

n_{ij} – наблюдаемая частота (число объектов) в ячейке (i, j) таблицы, так называемая фактическая клеточная частота; $\hat{n}_{ij}$ – теоретически ожидаемая (по H_0) частота в этой ячейке, $i = 1, 2, ..., r$; $j = 1, 2, ..., s$; r – число строк, s – число столбцов.

$$\left.\begin{aligned} n_i &= \sum_{j=1}^{s} n_{ij} \quad \text{– сумма по } j\text{-й строке} \\ n_j &= \sum_{i=1}^{r} n_{ij} \quad \text{– сумма по } j\text{-у столбцу} \end{aligned}\right\} \text{маргинальные частоты;} \tag{7.31}$$

$$n = \sum_{i=1}^{r} n_j = \sum_{j=1}^{s} n_i = \sum_{i=1}^{r}\sum_{j=1}^{s} n_{ij} \quad \text{– общее число объектов или объем выборки.}$$

В этом случае испытуемая гипотеза H_0: $n_{ij} = \hat{n}_{ij}$ или $H_0 : \chi^2 = 0$, альтернативная гипотеза $H_1 : n_{ij} \neq \hat{n}_{ij}$. Критерий χ^2 для проверки H_0 имеет вид:

$$\chi^2 = \sum_{(i)}\sum_{(j)} \frac{(n_{ij} - \hat{n}_{ij})^2}{\hat{n}_{ij}}. \tag{7.32}$$

Расчет теоретически ожидаемых частот в ячейках таблицы сопряженности должен производиться, как мы уже указывали выше, в предположении справедливости нулевой гипотезы. Нуль-гипотеза (H_0) в данном случае есть *предположение о статистической независимости* рассматриваемых переменных. Как известно из теории вероятностей, две случайные величины (события) являются статистически независимыми, если вероятность их совместной реализации равна произведению вероятностей реализации каждой из них по отдельности, т. е.

$$\pi(x_i, x_j) = \pi(x_i)\,\pi(x_j),$$

где

$$\pi(x_i, x_j) = \pi(\mathbf{X}_i = x_i, \mathbf{X}_j = x_j).$$

В нашем случае выборочными оценками соответствующих вероятностей π будут являться величины $p(x_i, x_j) = n_{ij}/n$,

$$p(x_i) = \frac{n_i}{n},\quad p(x_j) = \frac{n_j}{n},$$

и поэтому расчет теоретически ожидаемой по H_0 частоты $\hat{n}_{ij}$ следует производить по формуле

$$\hat{n}_{ij} = n\frac{n_i}{n}\cdot\frac{n_j}{n} = \frac{n_i n_j}{n}, \qquad (7.33)$$

т. е. произведение итогов по столбцу и строке, деленное на общий объем данных.

Если подставить выражение $\hat{n}_{ij}$ в формулу (7.32), то получим

$$\chi^2 = n\left(\sum_{i=1}^{r}\sum_{j=1}^{s}\frac{n_{ij}^2}{n_i n_j} - 1\right). \qquad (7.34)$$

Используя эту формулу, мы можем находить эмпирические значения критерия χ^2 без промежуточного вычисления теоретических частот в явном виде.

Очевидно, что для определения эмпирического значения критерия χ^2 нет необходимости рассчитывать все s теоретических частот

в каждой строке, а достаточно найти лишь $s - 1$ значение частоты в $r - 1$ строке, так как оставшиеся частоты могут быть получены как разности между маргинальными суммами эмпирических частот и суммами известных теоретических частот, т. е. значения теоретических частот в последних строке и столбце таблицы всегда полностью детерминированы. Поэтому число степеней свободы для $r \times s$ таблицы сопряженности равно:

$$d.f. = (r - 1)\,(s - 1). \qquad (7.35)$$

Заметим, что для таблицы 2×2 число степеней свободы равно 1.

В таблице распределения статистики $\chi^2_{d.f.,\alpha}$ приведены значения этой величины для различных уровней значимости при различных числах степеней свободы (см. приложение, табл. 4). Например, на уровне $\alpha = 0{,}01$ для $d.f. = 1$ мы находим $\chi^2 = 10{,}827$. Это означает, что равное или большее значение этой величины χ^2 может встретиться только один раз из ста при условии, что все сделанные допущения (нуль-гипотеза) справедливы. Другими словами, если выполняется предложение об отсутствии взаимосвязи между переменными, то крайне маловероятно ($P < 0{,}001$), что наблюдаемые и ожидаемые частоты будут отличаться настолько, что фактическая величина χ^2 будет равной или большей 10,827. Если же $\chi^2_{\phi} \geq \chi^2_{d.f.,\alpha}$, то гипотеза H_0 на данном уровне значимости α может быть отвергнута.

Вероятность того, что, отвергая нулевую гипотезу, мы совершаем ошибку (первого рода), которая численно равна уровню значимости α, задаваемому при проверке гипотезы.

Интерпретация χ^2 теста зачастую усложняется, когда в таблице сопряженности имеются ячейки с нулевыми значениями наблюдаемых частот. Дело в том, что если пара (x_i, x_j) значений переменных не наблюдалась в выборке, то это может означать, что объем выборки не столь велик, чтобы зафиксировать такую редкую комбинацию, либо что данная комбинация невозможна по каким-то объективным причинам. В последнем случае действительное число степеней свободы анализируемой системы меньше числа степеней свободы таблицы сопряженности, на основании которого произведена оценка уровня значимости χ^2 теста.

Корректировка применения χ^2 теста возможна лишь в том случае, если эмпирические данные, наполняющие таблицу сопряженности, есть результаты независимой случайной выборки относительно большого объема n. Последнее требование вызвано тем, что выборочное распределение χ^2 аппроксимирует табличное распределение статистики χ^2 только при больших n. Естественно,

возникает вопрос о том, насколько велико должно быть n, чтобы иметь возможность использовать данный тест. Ответ на этот вопрос зависит от числа ячеек и величин маргинальных сумм. Вообще говоря, чем меньше число ячеек и чем более близки между собой по величине маргиналы, тем меньше может быть n. Существует, однако, практическое число, позволяющее оценить снизу по n диапазон возможного применения критерия χ^2 : если в данной таблице сопряженности любая из теоретических ожидаемых частот $\hat{n}_{ij}$ в ячейке *(i, j)* не больше 5, то рекомендуется произвести, если это возможно, модификацию таблицы либо воспользоваться другим критерием.

В общем случае корректировка таблицы размера $r \times s$ затруднительна. Практика показала, что если число ячеек велико, а ожидаемые частоты, равные или меньше пяти, встречаются лишь в одной-двух ячейках, то проведение корректировки нецелесообразно; во всех иных случаях разумной альтернативой является объединение категорий (градаций) с тем, чтобы элиминировать подобные ячейки. Естественно, такое объединение должно быть таким, чтобы получаемая в результате комбинация не была содержательно бессмысленной.

Пример. Согласно опросу 157 предпринимателей, работающих в приватизированных кафе и ресторанах, относительно оценки возможностей деятельности при разных формах собственности получены следующие данные (табл. 7.7).

Таблица 7.7

Исходные данные

Формы собственности	Оценка возможностей деятельности					
	крайне неблагоприятно	неблагоприятно	трудно сказать	благоприятно	исключительно благоприятно	итого
Один владелец	18	16	5	13	5	57
Товарищество	4	4	10	11	11	40
Товарищество с ограниченной ответственностью	10	15	8	23	4	60
Итого	32	35	23	47	20	157

Испытаем гипотезу о независимости переменных H_0: $n_{ij} = \hat{n}_{ij}$, где n_{ij} – генеральные частоты, оценками которых выступают выборочные частоты n_{ij}. Теоретические частоты, рассчитанные в соответствии с нуль-гипотезой как $n_{ij} = \frac{n_i n_j}{n}$, представлены в табл. 7.8.

Таблица 7.8

Теоретические частоты

Формы собственности	Оценка возможностей деятельности					
	крайне неблагоприятно	неблагоприятно	трудно сказать	благоприятно	исключительно благоприятно	итого
Один владелец	11,6	12,7	8,3	17,1	7,3	57
Товарищество	8,2	8,9	5,9	11,9	5,1	40
Товарищество с ограниченной ответственностью	12,2	13,4	8,8	18,0	7,6	60
Итого	32	35	23	47	20	157

Таким было бы распределение ответов о возможностях деятельности, если бы формы собственности никак не сказывались. Задавая уровень значимости $\alpha = 0{,}05$, находим по табл. 4 приложения критическое значение критерия $\chi^2_{\alpha,\, df}$ при числе степеней свободы $d.f. = (3 - 1)(5-1) = 8$. Отсюда $\chi^2_{\alpha,\, df} = 15{,}51$.

Различия между фактическими и теоретическими клеточными частостями обобщаются в величине χ^2:

$$\chi^2 = \frac{(18-11{,}6)^2}{11{,}6} + \frac{(16-12{,}7)^2}{12{,}7} + \frac{(5-8{,}3)^2}{8{,}3} + \frac{(13-17{,}1)^2}{17{,}1} +$$

$$+ \frac{(5-7{,}3)^2}{7{,}3} + \frac{(4-8{,}2)^2}{8{,}2} + \ldots + \frac{(23-18)^2}{18} + \frac{(4-7{,}6)^2}{7{,}6} = 31{,}045.$$

Так как $\chi^2_{факт} > \chi^2_{крит}$, H_0 отклоняется, т.е. форма собственности небезразлична для деятельности кафе и ресторанов. Таким образом, наблюдаемое значение χ^2 является *значимым* на 5%-ном уров-

не значимости, и нулевая гипотеза может быть отвергнута в пользу альтернативной.

Итак, мы рассмотрели один из возможных способов ответа на вопрос: существует ли связь между двумя переменными? Для этого мы выдвинули нулевую гипотезу, что такой связи нет, а затем рассмотрели способ статистического испытания этой гипотезы. Мы можем оценить величину риска в принятии предположения о *существовании* связи. Но означает ли это, что данная связь существенна с точки зрения ее силы? Вовсе не обязательно. Вопрос о силе или степени, тесноте зависимости — это иной вопрос, отличный от вопроса о существовании взаимосвязи.

В социально-экономических исследованиях, как правило, установление факта наличия связи между переменными не самоцель. Установив наличие связи, исследователь должен измерить ее силу (тесноту) с тем, чтобы иметь возможность сравнивать взаимосвязи между различными характеристиками, выделять наиболее сильные из них (см. гл. 8).

7.11. ПРОВЕРКА ГИПОТЕЗЫ О СРЕДНИХ ВЕЛИЧИНАХ

Основные гипотезы о средних величинах следующие: гипотезы о значении генеральной средней (при известной генеральной дисперсии или при неизвестной генеральной дисперсии); гипотезы о равенстве генеральных средних нормально распределенных совокупностей (при известных генеральных дисперсиях, при неизвестных равных генеральных дисперсиях, при неизвестных неравных генеральных дисперсиях).

Первая задача чаще всего решается при неизвестной генеральной дисперсии. Испытуемая гипотеза H_0: $\mu = \mu_0$, альтернативная гипотеза H_1: $\mu \neq \mu_0$. Испытание гипотезы проводят с помощью t- критерия. При большом числе наблюдений критическое значение критерия определяется по таблице интеграла вероятностей, при малом – по таблице распределения Стьюдента с заданным уровнем значимости и числом степеней свободы, $n - 1$.

Если испытуемая гипотеза H_0: $\mu = a$, то фактическое значение критерия представляет отношение оцениваемой разности к средней возможной ошибке выборочной средней.

$$t = \frac{\bar{x} - a}{s_{\bar{x}}}, \qquad (7.36)$$

где $s_{\bar{x}} = \sqrt{\frac{s^2}{n}}$ – при большой выборке;

$s_{\bar{x}} = \sqrt{\frac{s^2}{n-1}}$ – при малой выборке.

Если $t_{факт} < t_{крит}$, H_0 не отклоняется, если $t_{факт} > t_{крит}$, H_0 отклоняется.

Рассмотрим **пример.** Часовая выработка забойщика при добыче угля в шахте по норме составляет 400 кг. Фактическая выработка соответствовала норме. При переходе в новый забой условия работы забойщиков усложнились. Для проверки обоснованности нормы в новых условиях был проведен учет работы 9 забойщиков: их средняя часовая выработка составила 388 кг с дисперсией, равной $s^2 = 171$.

Выдвигается гипотеза о том, что норму выработки пересматривать не нужно, т.е. H_0: $\mu = 400$ кг. Проверим эту гипотезу на 5%-ном уровне значимости. Критическое значение t-критерия определяется по таблице распределения Стьюдента при доверительной вероятности 0,95 (1 – 0,05) и числе степеней свободы $d.f. = n - 1 = 8$. Критическое значение составит $t_{крит} = 2{,}3$. Фактические значения t-критерия вычисляются по формуле (7.36):

$$t = \frac{388 - 400}{\sqrt{\frac{171}{8}}} = 2{,}6.$$

Поскольку $t_{факт} > t_{крит}$, H_0 отклоняется. Норма выработки в новых условиях должна быть пересмотрена, так как производительность труда стала существенно ниже нормативной.

В рассмотренном примере различие между фактическим и табличным значениями t-критерия невелико, поэтому вывод недостаточно надежен. Надежность вывода вообще понижается, если нет уверенности в нормальном распределении генеральной совокупности.

Гипотеза о равенстве средних может рассматриваться как гипотеза о связи, если сопоставляются средние величины, обусловленные действием какого-либо фактора. Например, сравнивается средняя заработная плата рабочих двух специальностей. Нулевая гипотеза состоит в том, что специальность рабочего не влияет на заработок. Если окажется, что $t_{факт} > t_{крит}$, нулевую гипотезу отклоняют

и делают вывод о том, что специальность оказывает влияние на заработную плату.

Рассмотрим решение этой задачи при условии, что генеральные дисперсии неизвестны, но принимаются разными. При сравнении средних величин выдвигается гипотеза, что обе выборки принадлежат одной и той же генеральной совокупности со средней μ и дисперсией σ^2.

При неизвестной генеральной дисперсии формула t-критерия имеет вид:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}. \qquad (7.37)$$

Поскольку s^2_1 и s^2_2 рассматриваются как выборочные оценки общей дисперсии σ^2, то формула (7.37) может быть записана так:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s^2(n_1^{-1} + n_2^{-1})}}, \qquad (7.38)$$

где $\bar{x}_1$, $\bar{x}_2$ – выборочные средние; s^2 – выборочная оценка общей дисперсии;

$$s^2 = \frac{\sum(x_{i1} - \bar{x}_1)^2 + \sum(x_{i2} - \bar{x}_2)^2}{n_1 + n_2 - 2}. \qquad (7.39)$$

Гипотеза H_0 отклоняется, если $|t| > t_{1-\alpha(n_1+n_2-2)}$.

Рассмотрим **пример.** Для проверки устойчивости цен на яблоки в летний период на двух рынках города проведено выборочное обследование: на первом рынке по данным 15 продавцов определена средняя цена, равная 2 тыс. руб./кг, при среднем квадратическом отклонении $s_1 = 0{,}5$ тыс. руб.; на втором рынке обследовано 17 продавцов, средняя цена оказалась равной 2,5 тыс. руб./кг, $s_2 = 0{,}4$ тыс. руб.

$$H_0: \mu_1 = \mu_2, \quad H_1: \mu_1 \neq \mu_2.$$

$$t = \frac{|2 - 2{,}5|}{\sqrt{\frac{15 \cdot 0{,}5^2 + 17 \cdot 0{,}4^2}{15 + 17 - 2}}} \cdot \sqrt{\frac{15 \cdot 17}{15 + 17}} = 3{,}02.$$

При $\alpha = 0{,}05$ и $d.f. = 30$, $t_{крит} = 2{,}042$, $t_{факт} > t_{крит}$, H_0 отклоняется, т. е. различия в ценах на двух рынках нельзя объяснить лишь случайностями выборки.

Проверка той же нулевой гипотезы при односторонней критической области будет проводиться на следующих условиях определения: $t_{крит}$: $1 - 2\alpha$ и $d.f. = n_1 + n_2 - 2$. Следовательно, если H_1: $\mu_1 = \mu_2$, $t_{крит} = 1{,}697$ ($2\alpha = 0{,}1$, $d.f. = 30$), так что H_0 опять-таки отклоняется.

Случай проверки гипотезы о средних величинах при неизвестных дисперсиях, равенство которых не предполагается, здесь не рассматривается ввиду его недостаточной теоретической разработанности[1].

7.12. ОСНОВЫ ДИСПЕРСИОННОГО АНАЛИЗА

Может быть поставлена задача сравнения двух выборочных дисперсий. Для ее решения применяется критерий, названный в честь английского статистика Рональда Фишера (1890 – 1968) F- критерием. Этот критерий представляет собой отношение выборочных дисперсий s_1^2 и s_2^2, которые рассматриваются как оценки одной и той же генеральной дисперсии σ^2:

$$F = \frac{s_1^2}{s_2^2}. \qquad (7.40)$$

Испытуемая гипотеза является нулевой гипотезой H_0: $\sigma^2_1 = \sigma^2_2 = \sigma^2$, альтернативная гипотеза H_1: $\sigma^2_1 \neq \sigma^2_2 \neq \sigma^2$.

F-критерий строится так, что в числителе стоит бóльшая дисперсия. $F_{min} = 1$, $F_{max} \to \infty$. Критические значения критерия F берутся из таблиц F-распределения. F-распределение зависит от уровня значимости и от числа степеней свободы сравниваемых дисперсий $d.f._1$ и $d.f._2$ (см. приложение, табл. 3).

В дисперсионном анализе общая вариация подразделяется на составляющие и производится сравнение этих составляющих. Испытуемая гипотеза состоит в том, что если данные каждой группы представляют случайную выборку из нормально распределенной генеральной совокупности, то величины всех частных дисперсий должны быть пропорциональны своим степеням свободы и

[1] *Афифи А., Эйзен С.* Статистический анализ. Подход с использованием ЭВМ/Пер. с англ.; Под ред. Г. П. Башарина. – М.: Мир, 1982.

каждую из них можно рассматривать как оценку генеральной дисперсии.

Дисперсионный анализ часто применяется совместно с аналитической группировкой (см. гл. 6). В этом случае данные подразделяются на группы по значениям признака-фактора, вычисляются значения средних величин результативного признака в группах, считается, что различия в их значениях определяются различиями в значениях фактора. Задача состоит в оценке существенности различий между средними значениями результативного признака в группах. Итак, испытуемая гипотеза может быть записана как гипотеза о средних величинах H_0: $\mu_1 = \mu_2 = \mu_3 = \ldots$ Как было показано в предыдущем параграфе, когда выделяются две группы, эта задача решается с помощью *t*-критерия. Если же число сравниваемых групп больше двух, то существенность различий между группами доказывается с помощью дисперсионного анализа, на основе *F*-критерия. Заметим, что результаты дисперсионного анализа, так же как и выводы о характере связи, значения показателей ее силы и тесноты, зависят от числа групп, выделенных по признаку-фактору.

В случае выделения групп по одному фактору мы имеем так называемый *однофакторный дисперсионный комплекс.* Разложение дисперсии при этом производится в соответствии с правилом сложения дисперсий (см. гл. 6):

$$\sum_{j=1}^{m}\sum_{i=1}^{n_j}(y_{ij}-\bar{y})^2 = \sum_{j=1}^{m}(\bar{y}_j-\bar{y})^2 n_j + \sum_{j=1}^{m}\sum_{i=1}^{n_j}(y_{ij}-\bar{y}_j)^2,$$

где y_{ij} – значение результативного признака у i-й единицы в j-й группе;
i – номер единицы, $i = 1, \ldots, n_j$;
j – номер группы;
n_j – численность j-й группы;
$\bar{y}_j$ – средняя величина результативного признака в j-й группе;
$\bar{y}$ – общая средняя результативного признака.

Если обозначить суммы квадратов отклонений буквой D, получим равенство:

$$D_{общ} = D_{факт} + D_{ост}. \qquad (7.41)$$

На основе разложения дисперсии (7.41) в соответствии с гипотезой отсутствия различий между группами могут быть получены три оценки генеральной дисперсии, пропорциональные степени свободы: на основе общей вариации, межгрупповой (факторной) и внутригрупповой (остаточной). Число степеней свободы равно:

для общей вариации $d.f._{общ} = (mn_j - 1) = n - 1$;
для межгрупповой вариации $d.f._{факт} = m - 1$;
для внутригрупповой вариации $d.f._{ост} = (n_j - 1)\, m = n - m$.

Как и суммы квадратов отклонений, числа степеней свободы связаны между собой равенством:

$$d.f._{общ} = d.f._{факт} + d.f._{ост}$$

или

$$n - 1 = (m - 1) + (n - m). \qquad (7.42)$$

Деление сумм квадратов отклонений на соответствующее число степеней свободы дает три оценки генеральной дисперсии σ^2.

$$s^2_{общ} = \frac{D_{общ}}{n-1},$$

$$s^2_{факт} = \frac{D_{факт}}{m-1}, \qquad (7.43)$$

$$s^2_{ост} = \frac{D_{ост}}{n-m}.$$

Поскольку $D_{факт}$ измеряет вариацию результативного признака, связанную с изменением фактора, по которому произведена группировка, а $D_{ост}$ – вариацию, связанную с изменением всех прочих факторов, сравнение этих величин, рассчитанных на одну степень свободы, дает возможность оценить существенность влияния признака-фактора на результативный признак с помощью *F*-критерия:

$$F = \frac{s^2_{факт}}{s^2_{ост}}.$$

Эта запись предполагает, что $s^2_{факт} \geq s^2_{ост}$. Как правило, мы получаем именно такое соотношение. Если $F_{факт} > F_{табл\ (\alpha,\ d.f.1,\ d.f.2)}$, можно утверждать, что нуль-гипотеза не соответствует фактическим данным, влияние признака-фактора является существенным или, иначе говоря, статистически значимым.

Рассмотренные этапы однофакторного дисперсионного анализа представлены в табл. 7.9.

Таблица 7.9

Схема однофакторного дисперсионного анализа

Источник вариации	Сумма квадратов отклонений D	Число степеней свободы $d.f.$	Средний квадрат отклонений $s^2 = D:d.f.$	F-критерий
Между группами	$\sum_{1}^{m}(\bar{y}_j - \bar{y})^2 n_j$	$m-1$	s^2_1	$F = \frac{s_1^2}{s_2^2}$
Внутри групп	$\sum_{1}^{m}\sum_{1}^{n_j}(y_{ij} - \bar{y}_j)^2$	$n-m$	s^2_2	
Общая	$\sum_{1}^{m}\sum_{1}^{n_j}(y_{ij} - \bar{y})^2$	$n-1$	s^2	

По данным табл. 6.6 проверим гипотезу H_0: $\mu_1 = \mu_2 = \ldots$, т. е. предположим, что оборачиваемость средств никак не влияет на прибыль.

$$D_{факт} = 172{,}76,\quad d.f._{факт} = 3 - 1 = 2,\quad \sigma_1^2 = 86{,}38;$$
$$D_{ост} = D_{общ} - D_{факт} = 224{,}4 - 172{,}76 = 51{,}64;$$
$$d.f._{ост} = 20 - 3 = 17;\quad s_2^2 = 3{,}03.$$

Тогда $F = 28{,}5$. Критическое значение F-критерия из табл. 3 приложения $F_{(\alpha\,=\,0{,}05,\ d.f.1\,=\,2,\ d.f.2\,=\,17)} = 3{,}59$. Таким образом $F_{факт} > F_{крит}$, следовательно, H_0 отклоняется. Действительно, скорость оборота средств является очень важным фактором формирования прибыли, на это указывало и значение эмпирического корреляционного отношения $\eta = 0{,}881$.

Рассмотрим *двухфакторный дисперсионный анализ*, основой проведения которого служит комбинационная группировка по двум факторам x и z, с последующим разложением дисперсии результативного признака y:

$$\sum_{j=1}^{m}\sum_{k=1}^{p}\sum_{i=1}^{n_{jk}}(y_{ijk} - \bar{y})^2 = \sum_{j=1}^{m}(\bar{y}_j - \bar{y})^2 n_j + \sum_{k=1}^{p}(\bar{y}_k - \bar{y})^2 n_k + \tag{7.44}$$
$$+ \sum_{j=1}^{m}\sum_{k=1}^{p}(\bar{y}_{jk} - \bar{y}_j - \bar{y}_k + \bar{y})^2 n_{jk} + \sum_{j=1}^{m}\sum_{k=1}^{p}\sum_{i=1}^{n_{jk}}(y_{ijk} - \bar{y}_{ik})^2,$$

где i – номер единицы в j-й группе по признаку x и k-й по признаку z;

$j = \overline{1, m}$,

$k = \overline{1, p}$, $\bar{y}_{jk}$ – среднее значение признака $\bar{y}$ в группе, образованной комбинацией j-го значения признака x и k-го значения признака z;

$\bar{y}_j$ – среднее значение признака y в j-й группе по признаку x;

$\bar{y}_k$ – среднее значение признака y в k-й группе по признаку z;

$\bar{y}$ – общая средняя признака y в целом по выборке;

n_{jk} – число единиц в группе, образованной комбинацией j-го значения признака x и k-го значения признака z;

n_j – число единиц в j-й группе по признаку x;

n_k – число единиц в k-й группе по признаку z;

n – общее число единиц, $n = \sum_{j=1}^{m} n_j = \sum_{k=1}^{p} n_k = \sum_{j=1}^{m}\sum_{k=1}^{p} n_{jk}$.

Равенство (7.44) можно записать так:

$$D_{общ} = D_x + D_z + D_{xz} + D_{ост} \tag{7.45}$$

где D_x – вариация y под влиянием фактора x;

D_z – вариация y под влиянием фактора z;

D_{xz} – вариация y, обусловленная взаимодействием факторов x и z;

$D_{ост}$ – вариация y под влиянием прочих факторов.

Первые три слагаемые составляют вариацию признака y, вызванную изучаемыми факторами, поэтому равенство (7.45) можно записать в виде:

$$D_{общ} = D_{факт} + D_{ост} \tag{7.46}$$

где

$$D_{факт} = D_x + D_z + Dxz. \tag{7.47}$$

Величина $D_{факт}$ может быть рассчитана не через составляющие, а непосредственно как

$$D_{факт} = \sum_{j=1}^{m}\sum_{k=1}^{p}(\bar{y}_{jk} - \bar{y})^2 n_{ik}. \tag{7.48}$$

Однако при неравенстве численностей подгрупп n_{jk} и групп n_j и n_k равенство нарушается (за счет взвешивания при неравных весах).

Поэтому рассчитываются невзвешенные величины:

$$D'_{факт} = \sum_{j=1}^{m}\sum_{k=1}^{p}(\bar{y}_{jk} - \bar{y})^2;$$

$$D'_x = \sum_{j=1}^{m}(\bar{y}_j - \bar{y})^2, \qquad (7.49)$$

$$D'_z = \sum_{k=1}^{p}(\bar{y}_k - \bar{y})^2,$$

$$D'_{xz} = D'_{факт} - D'_x - D'_z.$$

Затем на основе сравнения взвешенной (7.48) и невзвешенной величин факторной дисперсии находят поправочный коэффициент:

$$K = D_{факт} : D'_{факт}. \qquad (7.50)$$

Этот коэффициент используется для корректировки невзвешенных сумм квадратов отклонений D'_x, D'_z, D'_{xz}, на основе которых проводят расчет F-критериев:

$$D_{x(корр)} = D'_x K;\ D_{z(корр)} = D'_z K;\ D_{xz(корр)} = D'_{xz} K.$$

Число степеней свободы для каждой суммы квадратов отклонений составляет:

$$d.f._x = m - 1;\ d.f._z = p - 1;\ d.f._{xz} = (m-1)(p-1) = mp - m - p + 1,$$

в целом

$$d.f._{факт} = d.f._x + d.f._z + d.f._{xz} = mp - 1;$$

$$d.f._{общ} = n - 1;\ d.f._{ост} = d.f._{общ}\ d.f._{факт} = n - mp. \qquad (7.51)$$

В двухфакторном дисперсионном анализе испытуемые гипотезы формулируются следующим образом:

1) $H_0 : \mu_1. = \mu_2. = \ldots = \mu_m$;
2) $H_0 : \mu_1. = \mu_2. = \ldots = \mu_p$;
3) $H_0 : \mu_1. = \mu_2. = \ldots = \mu_{mp}$.

Вся процедура двухфакторного дисперсионного анализа обобщается в табл. 7.10.

Таблица 7.10

Схема двухфакторного дисперсионного анализа

Источник вариации	Сумма квадратов отклонений D	Число степеней свободы $d.f.$	Средний квадрат отклонений $s^2 = D/d.f.$	F-критерий
Факторы x и z	$D'_{факт} \cdot K$	$mp - 1$	$s^2_{факт}$	$F = \frac{s^2_x}{s^2_{ост}}$
Фактор x	$D'_x \cdot K$	$m - 1$	s^2_x	
Фактор z	$D'_z \cdot K$	$p - 1$	s^2_z	$F = \frac{s^2_z}{s^2_{ост}}$
Взаимодействие факторов x и z	$(D'_{факт} - D'_x - D'_z) K$	$mp - p - m + 1$	s^2_{xz}	$F = \frac{s^2_{xz}}{s^2_{ост}}$
Остаточная	$D_{общ} - D'_{факт} \cdot K$	$n - mp$	$s^2_{ост}$	
Общая	$D_{общ}$	$n - 1$	s^2	

Решение о первой гипотезе принимается на основе сравнения $F = s_x^2/s^2_{ост}$ с $F_{крит(\alpha d.f._1 = d.f._x\ d.f._2 = d.f._{ост})}$. Если $F_{факт} > F_{крит}$, то H_0 отклоняется.

Вторая гипотеза испытывается на основе сравнения $F = s_z^2/s^2_{ост}$ с $F_{крит(\alpha d.f._1 = d.f._x\ d.f._2 = d.f._{ост})}$.

Третья – на основе сравнения $F = s^2_{xz}/x^2_{ост}$ с $F_{крит(\alpha d.f._1 = d.f._x\ d.f._2 = d.f._{ост})}$.

Во всех случаях, если $F_{факт} > F_{крит}$, H_0 отклоняется.

На основе F-критерия принимаются решения о форме уравнения регрессии, о статистической значимости той или иной объясняющей переменной при построении многофакторного уравнения регрессии (см. гл. 8) и др.

Рассмотренные направления проверки статистических гипотез охватывают лишь важнейшие из них. Процедура испытания статистических гипотез применяется для определения того, случайно или нет полученное значение коэффициента корреляции, коэффициента вариации и т. д., случайны или нет различия в значениях показателей (медиан, коэффициентов корреляции, регрессии и т.д.) в разных

совокупностях. Во всех случаях результатом является *вероятностное суждение,* которое составляет сущность анализа данных в разнообразных сферах: в медицине, биологии, технике, политике, спорте, экономике, психологии и социологии.

7.13. ПРИМЕРЫ ПРИМЕНЕНИЯ ВЫБОРОЧНОГО МЕТОДА И ПРОВЕРКИ СТАТИСТИЧЕСКИХ ГИПОТЕЗ

Потребность в использовании выборочного метода, выработке вероятностных суждений в современной отечественной практике непрерывно расширяется. В государственной статистике основными направлениями использования выборочного метода традиционно являются бюджетные обследования семей, выборочные переписи населения, контрольные обходы и проверки после проведения сплошных обследований.

Создание единого государственного регистра предприятий и организаций (ЕГРПО), в котором фиксируются все хозяйствующие субъекты на территории Российской Федерации всех форм собственности, открывает возможность проведения разнообразных выборочных обследований в области экономики.

В области социальных исследований для государственной статистики главным является *бюджетное обследование,* которое охватывает примерно 48 тыс. домохозяйств. Оно основано на многоступенчатом отборе. Общий объем выборки распределяется по сферам занятости (для работающих) и территориям. Затем для работающих производится отбор предприятий в пределах каждой отрасли в отобранной территории. Если, например, нужно отобрать 100 рабочих, занятых в определенной отрасли, для обследования семейных бюджетов так, чтобы на каждом отобранном предприятии было не менее 20 бюджетов, включающих рабочих с разным уровнем заработной платы, то, значит, должно быть отобрано 100 : : 20 = 5 предприятий. Отбор предприятий проводят по списку, в котором предприятия располагаются в порядке убывания средней заработной платы рабочих, указываются общее число рабочих, их суммарная заработная плата. Шаг отбора определяется делением общего числа рабочих на предприятиях данной отрасли на число отбираемых предприятий. Если всего на предприятиях данной отрасли в области занято 30525 человек, то шаг отбора равен 30 525 : : 5 = 6105. По данным кумулятивной численности рабочих с данным шагом отбора производится отбор предприятий, которые затем проверяются на репрезентативность по показателю средней месячной заработной платы. Следующая стадия связана с отбором рабочих на выбранных предприятиях: среди 20 бюджетов должны быть пропорционально представлены бюджеты семей малоквалифицированных и высококвалифицированных рабочих, а среди этих категорий отбор проводится механически по спискам рабочих, составленным в порядке убывания среднемесячной заработной платы. Выборочная совокупность при бюджетных обследованиях включает и семьи неработающих (пенсионеров, студентов, инвалидов) и одиночек.

Задачей статистики в области бюджетных обследований является обеспечение представительства всех социальных групп и учет всех источников дохода. Наиболее общим показателем уровня благосостояния населения являются денежные доходы, поступающие в семью в виде заработной платы, премий, единовременных выплат, гонораров, предпринимательского дохода или дохода от собственности, компенсационных выплат и дотаций. В совокупные доходы семьи включаются также натуральная оплата труда, доходы, полученные от реализации и потребления продукции личного подсобного хозяйства (садового участка, коллективного огорода). Для характеристики обеспеченности семей следует учитывать их накопления, а также валютные поступления. Возрастает значение анализа личного потребления.

Большое значение имеет применение выборочного метода на промышленных предприятиях для статистического контроля качества продукции и использования внутрисменного времени рабочих.

Контроль качества продукции проводится для готовой продукции и в процессе ее изготовления. Выборочный контроль качества готовой продукции осуществляется так: отбирается на пробу некоторое число изделий и оценивается качество каждого из них. По доле дефектных изделий среди отобранных судят о качестве всей партии изделий. Если доля брака не превышает некоторого допустимого предела, то вся партия принимается без сплошного контроля. Если же доля брака больше допустимого предела, то проводится сплошная проверка всех остальных изделий в партии, конечно, если она не связана с уничтожением или порчей изделий.

При проведении контроля на стадии производства продукции машиностроения металлообработки основное внимание уделяется контролю положения центра настройки станков и вариации размеров деталей, обработанных на металлорежущем оборудовании.

Для изучения структуры рабочего времени разных категорий работников, особенно рабочих, а также для характеристики использования машин и оборудования используется *метод моментных наблюдений.* Этот метод состоит в регистрации вида затрат вре-

мени в определенные, заранее выбранные моменты. Заранее составляется список всех возможных состояний или видов затрат времени. Подсчитывается доля отметок о каждом состоянии, и оценивается доверительный интервал доли времени, затраченного на тот или иной вид работы. Отбор моментов выборки может быть проведен либо по схеме механической выборки — через равные промежутки времени, либо по схеме случайной выборки с использованием таблицы случайных чисел. Необходимая численность моментов наблюдения рассчитывается как

$$n = \frac{0{,}25t^2}{\Delta^2}.$$

Если принять доверительную вероятность равной 0,954 и допустимую ошибку 0,005, т.е. 0,5%, то $n = \frac{0{,}254}{0{,}000\,025} = 40\,000$ наблюдений. При продолжительности наблюдений 10 дней и охвате наблюдением 100 рабочих, в день должно проводиться 40 наблюдений за каждым рабочим. Если продолжительность смены составляет 8 ч, то интервал между обходами должен составлять 12 мин. [(8 ч · 60 мин.) : 40 наблюдений]. За начальный момент времени можно принять момент начала смены плюс пол-интервала: 7 ч 30 мин. + 6 мин. = 7 ч 36 мин., тогда второй обход будет проводиться в 7 ч 48 мин. и т.д. По итогам моментного наблюдения рассчитываются доверительные интервалы для каждого *i*-го вида потерь рабочего времени:

$$p_i - \Delta_{p_i} \le n_i \le p_i + \Delta_{p_i}.$$

Результаты наблюдений используются для анализа потерь рабочего времени, статистической оценки напряженности труда рабочих.

Выборочный метод используется в *аудиторской практике* при проверке бухгалтерских документов. При этом решаются две задачи: 1) дать оценку количества документов в данной фирме (предприятии, объединении и т. д.), в оформлении которых не соблюдались принятые правила; 2) оценить правильность указанных в документах сумм денежных средств. Первую задачу решают с помощью так называемой *атрибутивной выборки,* вторую – с помощью *монетарной выборки.* В первой выборке единицей отбора является учетный документ, во второй – денежная единица.

При организации атрибутивной выборки в качестве генеральной совокупности выступает вся совокупность расчетных

документов фирмы за проверяемый период. Обычно она предварительно разбивается на однородные массивы: по характеру документов, по центрам ответственности, по географическому признаку, по временной последовательности, по интенсивности запросов на данный вид информации и т.д. Каждому документу присваивается числовая метка, и по таблице случайных чисел проводится отбор номеров в количестве, соответствующем объему выборки. Можно провести и механический отбор с шагом отбора, равным $N : n$, где N – объем генеральной совокупности, n – объем выборки. Обычно начинают отбор не с первого документа, а отступив полшага.

Объем атрибутивной выборки находится из соотношения:

$$n = \frac{\text{Коэффициент надежности}}{\text{Максимально допустимая частота отклонений от стандартов оформления документов}} = \frac{R}{\Delta}.$$

Коэффициент надежности определяется по таблице распределения Пуассона, поскольку появление ошибки в оформлении расчетных документов относится к классу редких событий. При этом предполагаемая средняя частота ошибок закрепляется на определенном уровне, например 1; 1,5 или 2.

Если фактическая частота несоответствий в оформлении документов меньше максимально допустимой, то вычисляют коэффициент надежности как произведение объема выборки на величину фактической частоты несоответствий, после чего по таблице распределения Пуассона определяют вероятность, соответствующую рассчитанной величине коэффициента надежности, чтобы убедиться, что доверительная вероятность результатов выборки достаточно высока.

Если фактически выявленная частота несоответствия принятым правилам превышает максимально допустимую величину, то обязательно проводят монетарную выборку.

При монетарной выборке генеральной совокупностью является сумма денежных средств, зафиксированных во всех проверяемых документах. В качестве единицы отбора выступает денежная единица (1 руб.), а единицей наблюдения является расчетный документ. Требуемая точность результатов задается как допустимая относительная сумма ошибки. Объем монетарной выборки рассчитывается как

$$n = \frac{\text{Коэффициент надежности}}{\text{Максимально возможная относительная сумма ошибок в документах}} = \frac{R}{\Delta/N} = \frac{RN}{\Delta}.$$

Например, если аудитор исходит из 1%-ного риска (при односторонней критической области — опасения, что суммарная ошибка будет не больше принятой величины), т. е. при 98%-ной доверительной вероятности наличия суммарной ошибки 50 000 руб. при объеме генеральной совокупности, равном 60 млн/ руб., то объем выборки

$$n = \frac{2{,}31 \cdot 60\,000\,000}{50\,000} = 2772 \text{ денежные единицы.}$$

Определяется шаг отбора, равный $N : n = 60\,000\,000 : 2772 = 21645$ руб. Все расчетные документы, в которых зафиксирована сумма, равная или превышающая величину шага отбора, обязательно попадут в выборку. Начало отбора устанавливается произвольно.

Рассмотрим в качестве **примера** записи по счету «Расчеты с покупателями» (табл. 7.11).

Таблица 7.11

Формирование монетарной выборки (в качестве начала отбора принято 25 000 руб., шаг отбора равен 21 645)

Номер операции	Сумма, руб.	Нарастающий итог, руб.	Отбираемая единица, руб.
1	22000	22000	–
2	10000	32000	25000
3	18500	50500	46645
4	10275	60775	–
5	126850	187625	68290
			89935
			111580
			133225
			154870
			176515
6	12590	200215	198160
.	.	.	.
.	.	.	.
.	.	.	.

Приведенный пример показывает, что число отобранных документов может быть значительно меньше объема выборки по числу отбираемых денежных единиц. Если сумма операций многократно превышает шаг отбора, мы получаем несколько раз указание на необходимость проверки этой операции (в примере операция 5 получила представительство в выборке шесть раз), и, наоборот, если сумма операции меньше шага отбора, она может не попасть в выборку (в примере это произошло с операцией 4). В целом чем крупнее операции по сравнению с шагом отбора, тем меньше будет совокупность отобранных документов – единиц наблюдения по сравнению с числом отобранных единиц.

Особенности решения всех вопросов по определению репрезентативности выборки и распространению ее результатов на генеральную совокупность зависят от того, были ли выявлены ошибки в выборке или нет. Это влияет на значение коэффициента надежности: сохранится оно или нет. Исходя из этого проводится проверка соответствия фактической точности тому значению максимально допустимой суммарной величины ошибки, которое закладывалось при проектировании выборки. Если фактическая точность меньше или равна принятой, то выборка признается репрезентативной, если превышает ее, то применяются специальные методы оценки данных. Проверка производится на основе соотношения

$$\text{Шаг отбора} = \frac{\text{Допустимый размер ошибки } (\Delta)}{\text{Коэффициент надежности } (R) \text{ с учетом фактического обнаружения ошибок}}$$

отсюда $\Delta_{факт} = R \times$ шаг отбора.

Если при проверке отобранных документов ошибок не обнаружено, то с принятой доверительной вероятностью мы можем распространить результаты выборки на всю генеральную совокупность и считать, что итог по генеральной совокупности не завышен более чем на величину предельно допустимой ошибки. Если же обнаружена по крайней мере одна ошибка, то первоначальная гипотеза относительно отсутствия ошибок, которая закладывалась при планировании выборки, оказывается несостоятельной. В этом случае должны быть пересмотрены либо значение коэффициента надежности, либо величина предельно допустимой ошибки (точность), либо и то, и другое. Если ошибки выявлены в операциях, значение которых превышает величину шага отбора, то можно быть уверенным в отношении аб-

солютного размера ошибок в таких операциях, так как каждая из них проверялась полностью. В этом случае нужно решить вопрос о распространении абсолютного размера выявленных ошибок на операции, значение которых меньше шага отбора.

Все ошибки группируются в два класса: завышение суммы и ее занижение. Для всех операций, значение которых превышает шаг отбора, выявленная ошибка является точным размером завышения или занижения. Для операций, значение которых меньше шага отбора, размер выявленной ошибки относится к значению операции, и полученная относительная ошибка умножается на шаг отбора, т. е. распространяется на весь интервал. Приведем пример (табл. 7.12).

Таблица 7.12

Расчет суммарной ошибки на основе распространения результатов выборки

Характер и размер ошибки, руб.	Шаг отбора, руб.	Значение операции, руб.	Ожидаемый размер ошибки в интервале выборки
Завышение			
2500	21 645	126 850	2500
250	21 645	18 500	291,21 (установленный размер ошибки 1,35% распространен на весь интервал)
Итого 2750			2792,21
Занижение			
200	21 645	10 000	432,90
.	.	.	.
.	.	.	.
.	.	.	.

После определения суммарного размера ожидаемой ошибки по всем интервалам выборки (т. е. шагам отбора) производится сравнение с допустимым размером суммарной ошибки, и если рассчитанная суммарная ошибка превосходит допустимую величину, то, подставляя первую в формулу объема выборки, определяют, с каким коэффициентом надежности и соответственно с какой доверительной вероятностью могут гарантироваться результаты данного выборочного исследования:

$$R = \frac{n\Delta}{N}.$$

Как известно, в экономических исследованиях обычно принимают доверительную вероятность не ниже 90%.

Использование выборного метода в работе аудитора резко повышает эффективность получения результатов и приводит к экономии финансовых и трудовых затрат.

Рекомендуемая литература к главе 7

1. *Афифи А., Эйзен С.* Статистический анализ. Подход с использованием ЭВМ/Пер. с англ.; Под ред. Г. П. Башарина. – М.: Мир, 1982.
2. *Бокун Н. Ч., Чернышова Н. М.* Методы выборочных обследований. – Минск: Министерство статистики и анализа Республики Беларусь. НИИ статистики, 1997.
3. *Головач А. В., Ерина А. М., Трофимов В. П.* Критерии математической статистики в экономических исследованиях. – М.: Статистика, 1973.
4. *Джессен Р.* Методы статистических обследований/Пер. с англ.; Под ред. и с предисл. Е. М. Четыркина. – М.: Финансы и статистика, 1985.
5. *Дружинин Н. К.* Математическая статистика в экономике. – М.: Статистика, 1971.
6. Информатика в статистике: словарь-справочник. – М., Финансы и статистика, 1994.
7. *Йейтс Ф.* Выборочный метод в переписях и обследованиях. – М.: Статистика, 1965.
8. *Закс Л.* Статистическое оценивание / Пер. с нем.; Под ред. и с предисл. Ю. П. Адлера и В. Г. Горского. – М.: Статистика, 1976.
9. *Кокрен У.* Методы выборочного исследования/Пер. с англ.; Под ред. А. Г. Волкова. – М.: Статистика, 1976.
10. *Паниотто В. И.* Качество социологической информации (Методы оценки и процедуры обеспечения). – Киев: Наукова думка, 1986.
11. *Фишер Р. А.* Статистические методы для исследователей: Пер. с англ. – М.: Госстатиздат, 1958.

Глава 8

КОРРЕЛЯЦИОННО-РЕГРЕССИОННЫЙ АНАЛИЗ И МОДЕЛИРОВАНИЕ СТАТИСТИЧЕСКИХ СВЯЗЕЙ

8.1. ПОНЯТИЕ О СТАТИСТИЧЕСКОЙ И КОРРЕЛЯЦИОННОЙ СВЯЗИ

Современная наука исходит из взаимосвязи всех явлений природы и общества. Объем продукции предприятия связан с численностью работников, мощностью двигателей, стоимостью производственных фондов и еще многими признаками.

Невозможно управлять явлениями, предсказывать их развитие без изучения характера, силы и других особенностей связей. Поэтому методы исследования, измерения связей составляют чрезвычайно важную часть методологии научного исследования, в том числе и статистического.

Различают два типа связей между различными явлениями и их признаками: *функциональную* или *жестко детерминированную,* с одной стороны, и *статистическую* или *стохастически детерминированную* – с другой. Строго определить различие этих типов связи можно тогда, когда они получают математическую формулировку. Для простоты будем говорить о связи двух явлений или двух признаков, математически отображаемой в форме уравнения связи двух переменных.

Если с изменением значения одной из переменных вторая изменяется строго определенным образом, т.е. значению одной переменной обязательно соответствует одно или несколько точно заданных значений другой переменной, связь между ними является функциональной.

Нередко говорят о строгом соответствии лишь одного значения второй из переменных каждому значению первой из них, но это неверно. Например, связь между y и x является строго функциональной, если $y = \sqrt{x}$; но значению $x = 4$ соответствует не одно, а два значения: $y_1 = +2$; $y_2 = -2$. Уравнения более высоких степеней

могут иметь несколько корней, связь, разумеется, остается функциональной.

Функциональная связь двух величин возможна лишь при условии, что вторая из них зависит *только* от первой и ни от чего более. В реальной природе (и тем более в обществе) таких связей нет; они являются лишь абстракциями, полезными и необходимыми при анализе явлений, но упрощающими реальность. Функциональная зависимость данной величины y от многих факторов $x_1, x_2, ..., x_k$ возможна только в том случае, если величина y всегда зависит *только* от перечисленного набора факторов $x_1, x_2 ..., x_k$ и ни от чего более. Между тем все явления и процессы безграничного реального мира связаны между собой, и нет такого конечного числа переменных k, которые абсолютно полно определяли бы собою зависимую величину y. Следовательно, множественная функциональная зависимость переменных есть тоже абстракция, упрощающая реальность.

Однако такие науки, как механика, электротехника, акустика, политическая экономия и другие, успешно используют представление связей как функциональных не только в аналитических целях, но нередко и в целях прогнозирования. Это возможно потому, что в простых системах интересующая нас переменная величина зависит в основном (скажем, на 99% или даже на 99,99%) от немногих других переменных или только от одной переменной. То есть связь в такой несложной системе является хотя и не абсолютно функциональной, но практически очень близкой к таковой. Например, длина года (период обращения Земли вокруг Солнца) почти функционально зависит только от массы Солнца и расстояния Земли от него. На самом деле она зависит в очень слабой степени и от масс, и расстояния других планет от Земли, но вносимые ими (и тем более в миллионы раз более далекими звездами) искажения функциональной связи для всех практических целей, кроме космонавтики, пренебрежимо малы.

Стохастически детерминированная связь не имеет ограничений и условий, присущих функциональной связи. Если с изменением значения одной из переменных вторая может в определенных пределах принимать любые значения с некоторыми вероятностями, но ее среднее значение или иные статистические (массовые) характеристики изменяются по определенному закону – связь является статистической. Иными словами, *при статистической связи разным значениям одной переменной соответствуют разные распределения значений другой переменной.*

В настоящее время наука не знает более широкого определения связи. Все связи, которые могут быть измерены и выражены численно, подходят под определение «статистические связи», в том

числе и функциональные. Последние представляют собой частный случай статистических связей, когда значениям одной переменной соответствуют «распределения» значений второй, состоящие из одного или нескольких значений и имеющие вероятность, равную единице. Конечно, качественное различие действительно вероятностных распределений и отдельных значений, имеющих вероятность единицы (достоверных), настолько велико, что хотя функциональные связи и подходят в широком смысле под определение статистической связи, все же с полным основанием можно говорить о двух типах связей.

Корреляционной связью называют важнейший частный случай статистической связи, состоящий в том, что разным значениям одной переменной соответствуют различные *средние значения* другой. С изменением значения признака x закономерным образом изменяется среднее значение признака y; в то время как в каждом отдельном случае значение признака y (с различными вероятностями) может принимать множество различных значений.

Если же с изменением значения признака x среднее значение признака y не изменяется закономерным образом, но закономерно изменяется другая статистическая характеристика (показатели вариации, асимметрии, эксцесса и т.п.), то связь является не корреляционной, хотя и статистической.

Статистическая связь между двумя признаками (переменными величинами) предполагает, что каждый из них имеет случайную вариацию индивидуальных значений относительно средней величины. Если же такую вариацию имеет лишь один из признаков, а значения другого являются жестко детерминированными, то говорят лишь о регрессии, но не о статистической (тем более корреляционной) связи. Например, при анализе динамических рядов можно измерять регрессию уровней ряда урожайности (имеющих случайную колеблемость) на номера лет. Но нельзя говорить о корреляции между ними и применять показатели корреляции с соответствующей им интерпретацией (см. гл. 9).

Само слово *корреляция* ввел в употребление в статистику английский биолог и статистик Френсис Гальтон в конце XIX в. Тогда оно писалось как *«corelation»* (соответствие), но не просто «связь» *(relation)*, а «как бы связь», т. е. связь, но не в привычной в то время функциональной форме. В науке вообще, а именно в палеонтологии, термин «корреляция» применил еще раньше, в конце XYIII в., знаменитый французский палеонтолог (специалист по ископаемым останкам животных и растений прошлых эпох) Жорж Кювье. Он

ввел даже «закон корреляции» частей и органов животных. «Закон корреляции» помогает восстановить по найденным в раскопках черепу, костям и т. д. облик всего животного и его место в системе: если череп с рогами, то это было травоядное животное, а его конечности имели копыта; если же лапа с когтями – то хищное животное без рогов, но с крупными клыками.

Известен следующий рассказ о Кювье и «законе корреляции». В дни университетского праздника студенты решили подшутить над профессором Кювье. Они вырядили одного из студентов в козлиную шкуру с рогами и копытами и подсадили его в окно спальни Кювье. Ряженый загремел копытами и завопил: «Я тебя съем!». Кювье проснулся, увидел силуэт с рогами и спокойно отвечал: «Если у тебя рога и копыта, то по закону корреляции ты травоядное, и съесть меня не можешь. А за то, что не знаешь закона корреляции, получишь двойку!».

Корреляционная связь между признаками может возникать разными путями. Важнейший путь – *причинная зависимость* результативного признака (его вариации) от вариации факторного признака. Например, признак x – балл оценки плодородия почв, признак y – урожайность сельскохозяйственной культуры. Здесь совершенно ясно логически, какой признак выступает как независимая переменная (фактор) x, какой – как зависимая переменная (результат) y.

Совершенно иная интерпретация необходима при изучении корреляционной связи между двумя следствиями общей причины. Известен классический пример, приведенный крупнейшим статистиком России начала XX в. А. А. Чупровым: если в качестве признака x взять число пожарных команд в городе, а за признак y – сумму убытков за год в городе от пожаров, то между признаками x и y в совокупности городов России существенна прямая корреляция; в среднем, чем больше пожарников в городе, тем больше и убытков от пожаров! Уж не занимались ли пожарники поджигательством из боязни потерять работу? Но дело в другом. Данную корреляцию нельзя интерпретировать как связь причины и следствия; оба признака – следствия общей причины – размера города. Вполне логично, что в крупных городах больше пожарных частей, но больше и пожаров, и убытков от них за год, чем в мелких городах.

Третий путь возникновения корреляции – взаимосвязь признаков, каждый из которых и причина, и следствие. Такова, например, корреляция между уровнями производительности труда рабочих и уровнем оплаты 1 ч труда (тарифной ставкой). С одной стороны, уровень зарплаты – следствие производительности труда: чем она выше, тем выше и оплата. Но с другой стороны, установленные

тарифные ставки и расценки играют стимулирующую роль: при правильной системе оплаты они выступают в качестве фактора, от которого зависит производительность труда. В такой системе признаков допустимы обе постановки задачи; каждый признак может выступать и в роли независимой переменной x, и в качестве зависимой переменной y.

8.2. УСЛОВИЯ ПРИМЕНЕНИЯ И ОГРАНИЧЕНИЯ КОРРЕЛЯЦИОННО-РЕГРЕССИОННОГО МЕТОДА

Поскольку корреляционная связь является статистической, **первым условием** возможности ее изучения является общее условие всякого статистического исследования: наличие данных по достаточно большой совокупности явлений. По отдельным явлениям можно получить совершенно превратное представление о связи признаков, ибо в каждом отдельном явлении значения признаков кроме закономерной составляющей имеют случайное отклонение (вариацию). Например, сравнивая два хозяйства, одно из которых имеет лучшее качество почв, по уровню урожайности, можно обнаружить, что урожайность выше в хозяйстве с худшими почвами. Ведь урожайность зависит от сотен факторов и при том же самом качестве почв может быть и выше, и ниже. Но если сравнивать большое число хозяйств с лучшими почвами и большое число – с худшими, то средняя урожайность в первой группе окажется выше и станет возможным измерить достаточно точно параметры корреляционной связи.

Какое именно число явлений достаточно для анализа корреляционной и вообще статистической связи, зависит от цели анализа, требуемой точности и надежности параметров связи, от числа факторов, корреляция с которыми изучается. Обычно считают, что число наблюдений должно быть не менее чем в 5-6, а лучше – не менее чем в 10 раз больше числа факторов. Еще лучше, если число наблюдений в несколько десятков или в сотни раз больше числа факторов, тогда закон больших чисел, действуя в полную силу, обеспечивает эффективное взаимопогашение случайных отклонений от закономерного характера связи признаков.

Вторым условием закономерного проявления корреляционной связи служит условие, обеспечивающее надежное выражение закономерности в средней величине. Кроме уже указанного большого числа единиц совокупности для этого необходима достаточная качественная однородность совокупности. Нарушение этого условия может извратить параметры корреляции. Например, в массе зерновых хозяйств уровень продукции с гектара растет по мере концентрации площадей, т.е. он выше в крупных хозяйствах. В массе овощных и овоще-молочных хозяйств (пригородный тип) наблюдается та же прямая связь уровня продукции с размером хозяйства. Но если соединить в общую неоднородную совокупность те и другие хозяйства, то связь уровня продукции с размером площади пашни (или посевной площади) получится обратной. Причина в том, что овощные и овоще-молочные хозяйства, имея меньшую площадь, чем зерновые, производят больше продукции с гектара ввиду большей интенсивности производства в данных отраслях, чем в производстве зерна.

Иногда как **условие** корреляционного анализа выдвигают необходимость подчинения распределения совокупности по результативному и факторным признакам нормальному закону распределения вероятностей. Это условие связано с применением метода наименьших квадратов при расчете параметров корреляции: только при нормальном распределении метод наименьших квадратов дает оценку параметров, отвечающую принципам максимального правдоподобия. На практике эта предпосылка чаще всего выполняется приближенно, но и тогда метод наименьших квадратов дает неплохие результаты[1].

Однако при значительном отклонении распределений признаков от нормального закона нельзя оценивать надежность выборочного коэффициента корреляции, используя параметры нормального распределения вероятностей или распределения Стьюдента.

Еще одним **спорным вопросом** является допустимость применения корреляционного анализа к функционально связанным признакам. Можно ли, например, построить уравнение корреляционной зависимости размеров выручки от продажи картофеля, от объема продажи и цены? Ведь произведение объема продажи и цены равно выручке в каждом отдельном случае. Как правило, к таким жестко детерминированным связям применяют только индексный метод анализа. Однако на этот вопрос можно взглянуть и с другой точки зрения. При индексном анализе выручки предполагается, что количество проданного картофеля и его цена независимы друг от друга, потому-то и допустима абстракция от изменения одного фактора при измерении влияния другого, как это принято в индексном методе (см. гл. 10). В реальности количество и цена не являются вполне независимыми друг от друга.

---

[1] *Крастинь О. П.* Разработка и интерпретация моделей корреляционных связей в экономике. – Рига: Зинатне, 1983. – С. 14.

Корреляционно-регрессионный анализ учитывает межфакторные связи, следовательно, дает нам более полное измерение роли каждого фактора: прямое, непосредственное его влияние на результативный признак; косвенное влияние фактора через его влияние на другие факторы; влияние всех факторов на результативный признак. Если связь между факторами несущественна, индексным анализом можно ограничиться. В противном случае его полезно дополнить корреляционно-регрессионным измерением влияния факторов, даже если они функционально связаны с результативным признаком.

8.3. ЗАДАЧИ КОРРЕЛЯЦИОННО-РЕГРЕССИОННОГО АНАЛИЗА И МОДЕЛИРОВАНИЯ

В соответствии с сущностью корреляционной связи ее изучение имеет две цели:

1) измерение параметров уравнения, выражающего связь средних значений зависимой переменной со значениями независимой переменной (зависимость средних величин результативного признака от значений одного или нескольких факторных признаков);

2) измерение тесноты связи двух (или большего числа) признаков между собой.

Вторая задача специфична для статистических связей, а первая разработана для функциональных связей и является общей. Основным методом решения задачи нахождения параметров уравнения связи является метод наименьших квадратов (МНК), разработанный К. Ф. Гауссом (1777–1855). Он состоит в минимизации суммы квадратов отклонений фактически измеренных значений зависимой переменной y от ее значений, вычисленных по уравнению связи с факторным признаком (многими признаками) x.

Для измерения тесноты связи применяется несколько показателей. При парной связи теснота связи измеряется прежде всего корреляционным отношением, которое обозначается греческой буквой η. Квадрат корреляционного отношения – это отношение межгрупповой дисперсии результативного признака, которая выражает влияние различий группировочного факторного признака на среднюю величину результативного признака, к общей дисперсии результативного признака, выражающей влияние на него всех причин и условий. Квадрат корреляционного отношения называется *коэффициентом детерминации:*

$$\eta^2 = \frac{\sum_{j=1}^{k} (\bar{y}_j - \bar{y})^2 \cdot f_j}{\sum_{i=1}^{N} (y_i - \bar{y})^2}, \qquad (8.1)$$

где k – число групп по факторному признаку;

N – число единиц совокупности;

y_i – индивидуальные значения результативного признака;

$\bar{y}_j$ – его средние групповые значения;

$\bar{y}$ – его общее среднее значение;

f_j – частота в j-й группе.

Формула (8.1) применяется при расчете показателя тесноты связи по аналитической группировке (см. гл. 6). При вычислении корреляционного отношения по уравнению связи (уравнению парной или множественной регрессии) применяется формула (8.2):

$$\eta^2 = \frac{\sum_{i=1}^{N} (\hat{y}_i - \bar{y})^2 : N}{\sum_{i=1}^{N} (y_i - \bar{y})^2 : N}, \qquad (8.2)$$

где $\hat{y}_i$ – индивидуальные значения y по уравнению связи.

Сумма квадратов в числителе – это объясненная связью с фактором x (факторами) дисперсия результативного признака y. Она вычисляется по индивидуальным данным, полученным для каждой единицы совокупности на основе уравнения регрессии.

Если уравнение выбрано неверно или сделана ошибка при расчете его параметров, то сумма квадратов в числителе может оказаться большей, чем в знаменателе, и отношение утратит тот смысл, который оно должно иметь, а именно какова доля общей вариации результативного признака, объясняемая на основе выбранного уравнения связи его с факторным признаком (признаками). Чтобы избежать ошибочного результата, лучше вычислять корреляционное отношение по другой формуле (8.3), не столь наглядно выявляющей сущность показателя, но зато полностью гарантирующей от возможного искажения:

$$\eta^2 = \sqrt{1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}}. \qquad (8.3)$$

В числителе формулы (8.3) стоит сумма квадратов отклонений фактических значений признака y от его индивидуальных расчетных значений, т. е. доля вариации этого признака, не объясняемая за счет входящих в уравнение связи признаков-факторов. Эта сумма не может стать равной нулю, если связь не является функциональной. При неверной формуле уравнения связи или ошибке в расчетах возрастают расхождения фактических и расчетных значений, и корреляционное отношение снижается, как логически и должно быть.

В основе перехода от формулы (8.2) к формуле (8.3) лежит известное правило разложения сумм квадратов отклонений при группировке совокупности:

$$D_{общ} = D_{межгр} + D_{внутригр}$$

Согласно этому правилу можно вместо межгрупповой (факторной) дисперсии использовать разность:

$$D_{общ} - D_{внутригр}$$

что дает:

$$\eta^2 = \frac{D_{общ} - D_{внутригр}}{D_{общ}} = 1 - \frac{D_{внутригр}}{D_{общ}}.$$

При расчете η не по группировке, а по уравнению корреляционной связи (уравнению регрессии) мы используем формулу (8.3). В этом случае правило разложения суммы квадратов отклонений результативного признака записывается как

$$D_{общ} = D_{\substack{объяснена \\ уравнением \\ регрессии}} + D_{ост},$$

Важнейшее положение, которое следует теперь усвоить любому, желающему правильно применять метод корреляционно-регрессионного анализа, состоит в интерпретации формул (8.2) и (8.3). Это положение гласит:

Уравнение корреляционной связи измеряет зависимость между вариацией результативного признака и вариацией факторного признака (признаков). Меры тесноты связи измеряют долю вариации результативного признака, которая связана корреляционно с вариацией факторного признака (признаков).

Интерпретировать корреляционные показатели строго следует лишь в терминах вариации (различий в пространстве) отклонений от средней величины. Если же задача исследования состоит в измерении связи не между вариацией двух признаков в совокупности, а между изменениями признаков объекта во времени, то метод корреляционно-регрессионного анализа требует значительного изменения (см. гл. 9).

Из вышеприведенного положения об интерпретации показателей корреляции следует, что нельзя трактовать корреляцию признаков как связь их уровней. Это ясно хотя бы из следующего примера. Если бы все крестьяне области внесли под картофель одинаковую дозу удобрений, то вариация этой дозы была бы равна нулю, а следовательно, она абсолютно не могла бы влиять на вариацию урожайности картофеля. Параметры корреляции дозы удобрений с урожайностью будут тогда строго равны нулю. Но ведь и в этом случае уровень урожайности зависел бы от дозы удобрений – он был бы выше, чем без удобрений.

Итак, строго говоря, метод корреляционно-регрессионного анализа не может объяснить роли факторных признаков в создании результативного признака. Это очень серьезное ограничение метода, о котором не следует забывать.

Следующий общий вопрос – это уже рассмотренный в разделе о группировке вопрос о «чистоте» измерения влияния каждого отдельного факторного признака. Как отмечалось в главе 6, группировка совокупности по одному факторному признаку может отразить влияние именно данного фактора на результативный признак при условии, что все другие факторы не связаны с изучаемым, а случайные отклонения и ошибки взаимопогасились в большой совокупности. Если же изучаемый фактор связан с другими факторами, влияющими на результативный признак, будет получена не «чистая» характеристика влияния только одного фактора, а сложный комплекс, состоящий как из непосредственного влияния фактора, так и из его косвенных влияний, через его связь с другими факторами и их влияние на результативный признак. Данное положение полностью относится и к парной корреляционной связи.

Однако коренное отличие метода корреляционно-регрессионного анализа от аналитической группировки состоит в том, что корреля-

ционно-регрессионный анализ позволяет разделить влияние комплекса факторных признаков, анализировать различные стороны сложной системы взаимосвязей. Если метод комбинированной аналитической группировки, как правило, не дает возможность анализировать более 3 факторов, то корреляционный метод при объеме совокупности около ста единиц позволяет вести анализ системы с 8-10 факторами и разделить их влияние.

Наконец, развивающиеся на базе корреляционно-регрессионного анализа многомерные методы (метод главных компонент, факторный анализ) позволяют синтезировать влияние признаков (первичных факторов), выделяя из них непосредственно не учитываемые глубинные факторы (компоненты). Например, изучая корреляцию ряда признаков интенсификации сельскохозяйственного производства, таких, как фондообеспеченность, затраты труда на единицу площади, энергообеспеченность, внесение удобрений на единицу площади, плотность поголовья скота, можно синтезировать общую часть их влияния на уровень продукции с единицы площади или на производительность труда, получив обобщенный фактор «интенсификация производства», непосредственно не измеримый, не отражаемый единым показателем.

Правильное применение и интерпретация результатов корреляционно-регрессионного анализа возможны лишь при понимании всех специфических черт, достоинств и ограничений метода. Поэтому нужно рекомендовать вернуться к данному разделу заново после изучения остальных разделов данной главы и после приобретения некоторой практики применения метода к решению различных задач.

Необходимо сказать и о других задачах применения корреляционно-регрессионного метода, имеющих не формально математический, а содержательный характер.

1. *Задача выделения важнейших факторов,* влияющих на результативный признак (т.е. на вариацию его значений в совокупности). Эта задача решается в основном на базе мер тесноты связи факторов с результативным признаком.

2. *Задача оценки хозяйственной деятельности по эффективности использования имеющихся факторов производства.* Эта задача решается путем расчета для каждой единицы совокупности тех величин результативного признака, которые были бы получены при средней по совокупности эффективности использования факторов и сравнения их с фактическими результатами производства,

3. *Задача прогнозирования возможных значений результативного признака при задаваемых значениях факторных признаков.*

Такая задача решается путем подстановки ожидаемых, или планируемых, или возможных значений факторных признаков в уравнение связи и вычисления ожидаемых значений результативного признака.

Приходится решать и обратную задачу: вычисление необходимых значений факторных признаков для обеспечения планового или желаемого значения результативного признака в среднем по совокупности. Эта задача обычно не имеет единственного решения в рамках данного метода и должна дополняться постановкой и решением оптимизационной задачи на нахождение наилучшего из возможных вариантов ее решения (например, варианта, позволяющего достичь требуемого результата с минимальными затратами).

4. *Задача подготовки данных, необходимых в качестве исходных для решения оптимизационных задач.* Например, для нахождения оптимальной структуры производства в районе на перспективу исходная информация должна включать показатели производительности на предприятиях разных отраслей и форм собственности. В свою очередь, эти показатели могут быть получены на основе корреляционно-регрессионной модели либо на основании тренда динамического ряда (а тренд – это тоже уравнение регрессии).

При решении каждой из названных задач нужно учитывать особенности и ограничения корреляционно-регрессионного метода. Всякий раз необходимо специально обосновать возможность причинной интерпретации уравнения как *объясняющего* связь между вариацией фактора и результата. Трудно обеспечить раздельную оценку влияния каждого из факторов. В этом отношении корреляционные методы глубоко противоречивы. С одной стороны, их идеал – измерение чистого влияния каждого фактора. С другой стороны, такое измерение возможно при отсутствии связи между факторами и случайной вариации признаков. А тогда связь является функциональной, и корреляционные методы анализа излишни. В реальных системах связь всегда имеет статистический характер, и тогда идеал методов корреляции становится недостижимым. Но это не значит, что эти методы не нужны.

Данное противоречие означает попросту недостижимость абсолютной истины в познании реальных связей. Приближенный характер любых результатов корреляционно-регрессионного анализа не является поводом для отрицания их полезности. Всякая научная истина – относительна. Забыть об этом и абсолютизировать параметры регрессионных уравнений, меры корреляции было бы ошибкой, так же как и отказаться от использования этих мер.

8.4. ВЫЧИСЛЕНИЕ И ИНТЕРПРЕТАЦИЯ ПАРАМЕТРОВ ПАРНОЙ ЛИНЕЙНОЙ КОРРЕЛЯЦИИ

Простейшей системой корреляционной связи является линейная связь между двумя признаками – *парная линейная корреляция.*

Практическое значение ее в том, что есть системы, в которых среди всех факторов, влияющих на результативный признак, выделяется один важнейший фактор, который в основном определяет вариацию результативного признака. Измерение парных корреляций составляет необходимый этап в изучении сложных, многофакторных связей. Есть такие системы связей, при изучении которых следует предпочесть парную корреляцию. Внимание к линейным связям объясняется ограниченной вариацией переменных и тем, что в большинстве случаев нелинейные формы связей для выполнения расчетов преобразуются в линейную форму.

Уравнение парной линейной корреляционной связи называется *уравнением парной регрессии* и имеет вид:

$$\tilde{y} = a + bx, \qquad (8.4)$$

где $\tilde{y}$ – среднее значение результативного признака y при определенном значении факторного признака x;

a – свободный член уравнения;

b – коэффициент регрессии, измеряющий среднее отношение отклонения результативного признака от его средней величины к отклонению факторного признака от его средней величины на одну единицу его измерения – вариация y, приходящаяся на единицу вариации x.

Что касается термина регрессия, его происхождение таково: создатели корреляционного анализа Ф. Гальтон (1822 – 1911) и К. Пирсон (1857 – 1936) интересовались связью между ростом отцов и их сыновей. Ф. Гальтон изучил более 200 семей и обнаружил, что в группе семей с высокорослыми отцами сыновья в среднем ниже ростом, чем их отцы, а в группе семей с низкорослыми отцами сыновья в среднем выше отцов. Таким образом, отклонение роста от средней в следующем поколении уменьшается – *регрессирует.* Причина в том, что на рост сыновей влияет не только рост отцов, но и рост матерей и много других факторов развития ребенка, и эти факторы, случайно направленные как в сторону увеличения, так и снижения роста, приближают рост сыновей к среднему росту. В целом же вариация роста, конечно, не уменьшается, а в наше время «акселерации» сам средний рост увеличивается из поколения в поколение.

Уравнение (8.4) определяется по данным о значениях признаков x и y в изучаемой совокупности, состоящей из n единиц. Параметры уравнения a и b находятся методом наименьших квадратов (МНК).

Исходное условие МНК для прямой линии имеет вид:

$$f(a,b) = \sum_{i=1}^{n}[y_i - (a + bx_i)]^2 \to \min. \qquad (8.5)$$

Для отыскания значений параметров a и b, при которых $f(a,b)$ принимает минимальное значение, частные производные функции приравниваем нулю и преобразуем получаемые уравнения, которые называются *нормальными уравнениями МНК для прямой:*

$$\frac{df}{da} = 2\sum_{i=1}^{n}\left(y_i - a - bx_i\right)(-1) = 0;$$

$$\frac{df}{db} = 2\sum_{i=1}^{n}\left(y_i - a - bx_i\right)(-x) = 0.$$

Отсюда система нормальных уравнений имеет вид:

$$na + b\sum_{i=1}^{n}x_i = \sum_{i=1}^{n}y_i;$$

$$a\sum_{i=1}^{n}x_i + b\sum_{i=1}^{n}(x_i)^2 = \sum_{i=1}^{n}x_i y_i.$$

Нормальные уравнения МНК для прямой линии регрессии являются системой двух уравнений с двумя неизвестными a и b. Все остальные величины, входящие в систему, определяются по исходной информации. Таким образом, однозначно вычисляются при решении этой системы уравнений оба параметра уравнения линейной регрессии.

Если первое нормальное уравнение разделить на n, получим:

$$a + b\bar{x} = \bar{y}, \text{ откуда } a = \bar{y} - b\bar{x}. \qquad (8.6)$$

По уравнению (8.6) обычно на практике вычисляется свободный член уравнения регрессии a. Параметр b вычисляется по преобразованной формуле, которую можно вывести, решая систему нормальных уравнений относительно b:

$$b = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\overline{x^2} - \bar{x}^2}. \qquad (8.7)$$

Так как знаменатель этого выражения есть не что иное, как дисперсия признака x, т. е. σ^2_x, то можно записать формулу коэффициента регрессии в виде:

$$b = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\sigma_x^2}. \qquad (8.8)$$

Подставив в (8.8) выражение для σ^2_x, получим:

$$b = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\left[\sum_{i=1}^{n}(x_i - \bar{x})^2\right] : n} = \frac{n\overline{xy} - n\bar{x} \cdot \bar{y}}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}. \qquad (8.9)$$

Параметры уравнения регрессии можно вычислить через определители:

$$a = \frac{\Delta_a}{\Delta}, \quad b = \frac{\Delta_b}{\Delta}, \qquad (8.10)$$

где Δ – определитель системы;

Δ_a – частный определитель, получаемый в результате замены коэффициентов при a свободными членами из правой части системы уравнений;

Δ_b – частный определитель, получаемый в результате замены коэффициентов при b свободными членами из правой части системы уравнений.

Формулы (8.10) соответствуют самому общему подходу к определению параметров уравнения регрессии и могут применяться в случае как парной, так и множественной регрессии.

Применение одной из формул (8.7), (8.8) или (8.9) зависит от характера данных и наличия уже вычисленных на предыдущих этапах анализа показателей. Если были вычислены $\bar{x}$, $\bar{y}$, σ_x, σ_y, то проще применить формулу (8.7) или (8.8). Если расчет параметров уравнения корреляционной связи ведется исходя из первичных данных x_i, y_i, то удобнее формула (8.9). Особенно существенно она сокращает объем вычислений при слабой вариации признаков, ибо тогда отклонения их индивидуальных значений от средних величин

на порядок или два меньше самих индивидуальных и средних величин. Кроме того, формула (8.9) явно выражает указанную в п. 8.1 особенность корреляционного анализа связей: параметры корреляции зависят не от уровней признаков, а только от их отклонений от средних значений.

Если значение признака увеличить в 10 раз, корреляция не изменится, также не изменятся параметры корреляции, кроме свободного члена, если ко всем значениям каждого признака прибавить постоянное число.

Коэффициент парной линейной регрессии, обозначенный b, имеет смысл показателя *силы* связи между вариацией факторного признака x и вариацией результативного признака y. Он измеряет среднее по совокупности отклонение y от его средней величины при отклонении признака x от своей средней величины на принятую единицу измерения.

Например, по данным табл. 8.1 при отклонении затрат на 1 корову от средней величины на 1 руб. надой молока на корову отклоняется от своего среднего значения на 3,47 кг в среднем по совокупности. При отклонении фактора на $\bar{x}_i - \bar{x}$ результативный признак отклоняется в среднем на $\bar{y}_i - \bar{y}$.

Теснота парной линейной корреляционной связи, как и любой другой показатель, может быть измерена корреляционным отношением η. Кроме того, при линейной форме уравнения применяется другой показатель тесноты связи – коэффициент корреляции r_{xy}. Этот показатель представляет собой *стандартизованный коэффициент регрессии,* т. е. коэффициент, выраженный не в абсолютных единицах измерения признаков, а в долях среднего квадратического отклонения результативного признака:

$$r_{xy} = b\frac{\sigma_x}{\sigma_y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 : n}}{\sum_{i=1}^{n}(x_i - \bar{x})^2\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2 : n}} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2\sum_{i=1}^{n}(y_i - \bar{y})^2}}. \qquad (8.11)$$

Коэффициент корреляции был предложен английским статистиком и философом Карлом Пирсоном (1857 – 1936). Его интерпретация такова: отклонение признака-фактора от его среднего значения на величину своего среднего квадратического отклонения в среднем по совокупности приводит к отклонению признака-результата от своего среднего значения на r_{xy} его среднего квадратического отклонения.

В отличие от коэффициента регрессии b коэффициент корреляции не зависит от принятых единиц измерения признаков, а стало быть, он сравним для любых признаков.

Обычно считают связь сильной, если $r \geq 0{,}7$; средней тесноты, при $0{,}5 \leq r \leq 0{,}7$; слабой при $r < 0{,}5$. Не следует, особенно работая с ЭВМ, гнаться за большим числом знаков коэффициента корреляции. Во-первых, исходная информация редко имеет более трех значащих точных цифр, во-вторых, оценка тесноты связи не требует более двух значащих цифр.

Квадрат коэффициента корреляции называется *коэффициентом детерминации:*

$$r_{xy}^2 = \frac{\left[\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})\right]^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{b\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(y_i - \bar{y})^2}. \qquad (8.12)$$

Эта формула понадобится при анализе множественной корреляции. Умножив числитель и знаменатель (8.12) на $\sum_{i=1}^{n}(x_i - \bar{x})^2$, получим:

$$r_{xy}^2 = \frac{b^2\sum_{i=1}^{n}(x_i - \bar{x})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}.$$

Поскольку $b\,(x_i - \bar{x}) = \hat{y}_i - \bar{y}$, имеем:

$$r_{xy}^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}. \qquad (8.13)$$

Это выражение соответствует выражению η^2 (см. формулу (8.2)). Тождество коэффициента детерминации и квадрата корреляционного отношения служит основанием для интерпретации величины r^2_{xy} как доли общей дисперсии результативного признака y, которая объясняется вариацией признака-фактора x (и связью между вариацией обоих признаков). Собственно говоря, основным показателем тесноты связи и следовало бы считать коэффициент детерминации

(для линейной формулы связи) или квадрат корреляционного отношения. Но исторически раньше был введен коэффициент корреляции, который долгое время и рассматривался как основной показатель.

Аналогично разным «рабочим» формулам для вычисления коэффициента регрессии можно на основе исходной формулы (8.10) получить разные «рабочие» формулы коэффициента корреляции.

1. Разделив числитель и знаменатель формулы (8.11) на n, получим:

$$r_{xy} = \frac{\left[\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})\right] : n}{\sqrt{\sigma_x^2\sigma_y^2}} = \frac{\overline{xy} - \bar{x}\cdot\bar{y}}{\sigma_x\sigma_y}. \qquad (8.14)$$

Эта формула соответствует формуле (8.8) для коэффициента регрессии.

2. Средние квадратические отклонения можно выразить через средние величины признака:

$$\sigma_x = \sqrt{\overline{x^2} - (\bar{x})^2}; \quad \sigma_y = \sqrt{\overline{y^2} - (\bar{y})^2}.$$

Подставив эти выражения в (8.14), получим:

$$r_{xy} = \frac{\overline{xy} - \bar{x}\cdot\bar{y}}{\sqrt{[\overline{x^2} - (\bar{x})^2]\,[\overline{y^2} - (\bar{y})^2]}}. \qquad (8.15)$$

Эта формула (8.15) удобнее для расчетов, если средние величины признаков и средние квадраты индивидуальных величин вычислены ранее. Смысл же коэффициента корреляции раскрывается исходной формулой (8.11). В преобразованных формулах этот смысл не столь ясен.

Рассмотрим фактический пример анализа корреляционной парной линейной связи по данным 16 сельхозпредприятий о затратах на 1 корову и о надое молока на корову. Ограниченный объем совокупности принят только в учебных целях, чтобы избежать приведения громоздких таблиц (табл. 8.1).

Средние значения признаков: $\bar{x} = 1605$ руб.; $\bar{y} = 35{,}2$ ц/голов.

Сопоставляя знаки отклонений признаков x и y от средних величин, видим явное преобладание совпадающих по знакам пар отклонений: их 14 и только 2 пары несовпадающих знаков.

Таблица 8.1

Корреляция между затратами на корову и надоем молока в среднем от коровы

Номера единиц совокупности	Затраты на 1 корову, руб./голов x_i	Надой от 1 коровы, ц y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x}) \times (y_i - \bar{y})$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$	Расчетные значения надоя, ц $\tilde{y}_i$
1	1602	34,2	–3	–1,0	+3,0	9	1,00	35,1
2	1199	19,6	–406	–15,6	+6333,6	164836	243,36	21,1
3	1321	27,3	–283	–7,9	+2235,7	80089	62,41	25,3
4	1678	32,5	+73	–2,7	–197,1	5329	7,29	37,7
5	1600	33,2	–5	–2,0	+10,0	25	4,00	35,0
6	1355	31,8	–250	–3,4	+850,0	62500	11,56	26,5
7	1413	30,7	–192	–4,5	+864,0	36864	20,25	28,5
8	1490	32,6	–115	–2,6	+299,0	13225	6,76	31,2
9	1616	26,7	+11	–8,5	–93,5	121	72,25	35,6
10	1693	42,4	+88	+7,2	+633,6	7744	51,84	38,2
11	1665	37,9	+60	+2,7	+162,0	3600	7,29	37,3
12	1666	36,6	+61	+1,4	+85,4	3721	1,96	37,3
13	1628	38,0	+23	+2,8	+64,4	529	7,84	36,0
14	1604	32,7	–1	–2,5	+2,5	1	6,25	35,2
15	2077	51,7	+472	+16,5	+7788	222784	272,25	51,6
16	2071	55,3	+466	+20,1	+9366,6	217156	404,01	51,4
Σ	25678	563,2	–	–	+28473,7	818533	1180,32	563,0

Немецкий психиатр Г. Т. Фехнер (1801 – 1887) предложил меру тесноты связи в виде отношения разности числа пар совпадающих и несовпадающих пар знаков к сумме этих чисел:

$$K_{Фехнера} = \frac{C - H}{C + H} = \frac{14 - 2}{14 + 2} = 0{,}75.$$

Конечно, коэффициент Фехнера – очень грубый показатель тесноты связи, не учитывающий величину отклонений признаков от средних значений, но он может служить некоторым ориентиром в оценке интенсивности связи. В данном случае он указывает на тесную связь признаков.

Вычислим на основе итоговой строки табл. 8.1 параметр парной линейной корреляции:

$$b = \frac{+28473{,}7}{818533} = +0{,}0347.$$

Он означает, что в среднем по изучаемой совокупности отклонение затрат на 1 корову от средней величины на 1 руб. приводило к отклонению с тем же знаком среднего надоя молока на 0,0347 ц, т. е. на 3,47 кг на корову. При нестрогой интерпретации говорят: «С увеличением затрат на корову на 1 руб. в среднем надой молока возрастал на 3,47 кг». Поскольку и до начала резкой инфляции стоимость 3,47 кг молока значительно превосходила рубль, увеличение затрат на корову было экономически целесообразным.

Свободный член уравнения регрессии вычислим по формуле (8.6):

$$a = 35{,}2 - 0{,}0347 \cdot 1605 = -20{,}49.$$

Уравнение регрессии в целом имеет вид:

$$\tilde{y} = 0{,}0347x - 20{,}49.$$

Отрицательная величина свободного члена уравнения означает, что область существования признака y не включает нулевого значения признака x и близких значений. Можно рассчитать минимально возможную величину фактора x, при которой обеспечивается наименьшее значение признака y (разумеется, положительное).

$$x_{min} = a : b = 20{,}49 : 0{,}0347 = 590{,}5 \text{ руб./голов/}$$

– это наименьшая сумма затрат на 1 корову, при которых корова способна давать молоко. Если же область существования результативного признака включает нулевое значение признака-фактора, то свободный член является положительным и означает среднее значение результативного признака при отсутствии данного фактора, например среднюю урожайность картофеля при отсутствии органических удобрений.

Графическое изображение корреляционной связи по данным табл. 8.1. приведено на рис. 8.1.

Коэффициент корреляции, рассчитанный на основе табл. 8.1,

$$r_{xy} = \frac{+28473{,}7}{\sqrt{818533 \cdot 1180{,}32}} = +0{,}916.$$

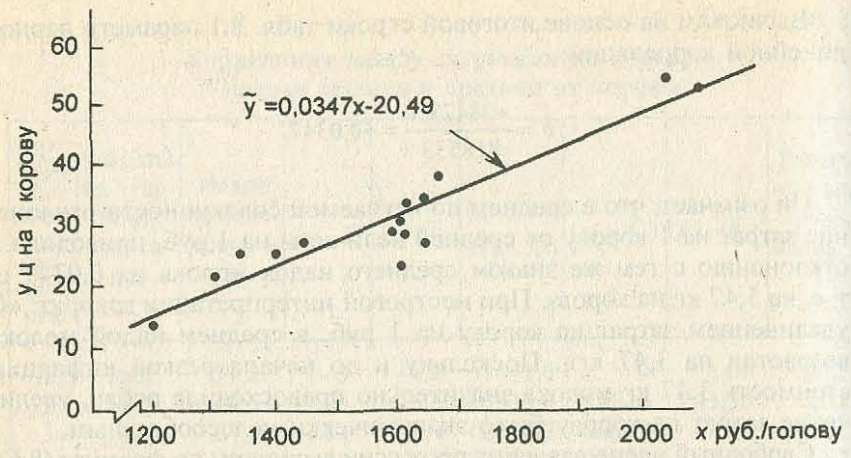


Рис. 8.1. Корреляция затрат на корову с продуктивностью

Полученное значение гораздо больше коэффициента Фехнера. Квадрат коэффициента корреляции, т. е. коэффициент детерминации, составил 0,839, или 83,9%. Вариации надоев молока на корову связаны с вариацией затрат в хозяйствах, произведенных в среднем на 1 корову.

Для интерпретации коэффициента корреляции необходимо знать область его существования $0 \le |r| \le 1$. Как ясно из формулы (8.11), минимальное, именно нулевое, значение коэффициента корреляции может быть достигнуто, если положительные и отрицательные произведения отклонений признаков от их средних величин в числителе полностью уравновесят друг друга. Это свидетельствовало бы о полном отсутствии связи, но вероятность такого абсолютно точного взаимопогашения крайне мала для любой реальной, не бесконечно большой совокупности. Поэтому и при отсутствии реальной связи коэффициент корреляции на практике не равен нулю. Например, коэффициент корреляции между надоем молока от коров и числом букв в названии предприятия в совокупности хозяйств, указанных в табл. 8.1, равен +0,216. Как отделить реальные, надежно установленные связи от таких случайных, незначимых величин коэффициента корреляции, рассматривается в следующем разделе этой главы.

Максимально тесная связь – это связь функциональная, когда каждое индивидуальное значение результативного признака y_i, может быть однозначно поставлено в соответствие значению x_i, например когда $y_i = x_i \cdot c$, где c – константа. Подставив это выражение y_i в формулу коэффициента корреляции (8.11), получим:

$$r_{xy\max} = \frac{\sum_1^n (x_i - \bar{x})(x_i c - \bar{x}c)}{\sqrt{\sum_1^n (x_i - \bar{x})^2 \sum (x_i c - \bar{x}c)^2}} = \frac{c\sum_1^n (x_i - \bar{x})^2}{\sqrt{c^2\left(\sum (x_i - \bar{x})^2\right)^2}} = \frac{c\sum_1^n (x_i - \bar{x})^2}{c\sum_1^n (x_i - \bar{x})^2} = 1$$

Если связь обратная и $y_i = -cx_i$, то коэффициент корреляции будет равен минус единице. Чем ближе коэффициент корреляции к единице, тем ближе связь к функциональной. Полученное в примере значение +0,916 свидетельствует об очень тесной связи надоев молока с затратами в расчете на 1 корову. Об этом же говорит и рис. 8.1, где реальные значения для отдельных хозяйств (точки корреляционного поля) близко расположены к линии регрессии, выражающей среднюю закономерность связи.

8.5. СТАТИСТИЧЕСКАЯ ОЦЕНКА НАДЕЖНОСТИ ПАРАМЕТРОВ ПАРНОЙ КОРРЕЛЯЦИИ

Показатели корреляционной связи, вычисленные по ограниченной совокупности (по выборке), являются лишь оценками той или иной статистической закономерности, поскольку в любом параметре сохраняется элемент не полностью погасившейся случайности, присущей индивидуальным значениям признаков. Поэтому необходима статистическая оценка степени точности и надежности параметров корреляции. Под надежностью здесь понимается вероятность того, что значение проверяемого параметра не равно нулю, не включает в себя величины противоположных знаков.

Вероятностная оценка параметров корреляции производится по общим правилам проверки статистических гипотез, разработанным математической статистикой, в частности путем сравнения оцениваемой величины со средней случайной ошибкой оценки. Для коэффициента парной регрессии b средняя ошибка оценки вычисляется как:

$$m_b = \sqrt{\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2 : (n-2)}{\sum_{i=1}^n (x_i - \bar{x})^2}}, \qquad (8.16)$$

где $\tilde{y}_i$ – расчетные значения результативного признака;
n–2 – число степеней свободы.

Числитель подкоренного выражения есть остаточная дисперсия результативного признака.

В примере по данным табл. 8.1 средняя ошибка оценки коэффициента регрессии

$$m_b = \sqrt{\frac{195{,}4 : 14}{818533}} = 0{,}00413.$$

Зная среднюю ошибку оценки коэффициента регрессии, можно вычислить вероятность того, что нулевое значение коэффициента входит в интервал возможных с учетом ошибки значений. С этой целью находится отношение коэффициента к его средней ошибке, т. е. t-критерий Стьюдента:

$$t = \frac{b}{m_b} = \frac{0{,}0347}{0{,}00413} = 8{,}4.$$

Табличное значение t-критерия Стьюдента при 16 – 2 степенях свободы и уровне значимости 0,01 составляет 2,98 (см. приложение, табл. 2). Полученное значение критерия много больше, следовательно, вероятность нулевого значения коэффициента регрессии менее 0,01. Гипотезу о несущественности этого коэффициента можно отклонить: данные табл. 8.1 надежно говорят о влиянии вариации затрат на корову на вариацию надоя молока от коров. Расчет критерия Стьюдента для коэффициентов регрессии входит в программы ЭВМ и ПЭВМ для корреляционного анализа, например «Mikrostat», MAKR-4, «Statgraphics» и др.

Надежность установления связи можно проверить и по средней случайной ошибке коэффициента корреляции, вычисляемой по формуле:

$$m_r = \frac{1 - r^2}{\sqrt{n-2}}. \qquad (8.17)$$

Проверим значимость заведомо бессодержательного коэффициента корреляции надоя от коров с числом букв в названии сельхозпредприятия:

$$m_r = \frac{1 - 0{,}216^2}{16 - 2} = 0{,}255; \quad t = \frac{0{,}216}{0{,}255} = 0{,}85.$$

Полученное значение t намного ниже его критического значения даже для значимости 0,1, составляющего 1,76. Следовательно, вероятность того, что нулевое значение коэффициента входит в возможный интервал его оценок значительно больше 0,1 и нулевая гипотеза не может быть отброшена. Конечно, анекдотический характер фактора «число букв» позволяет сделать решительный вывод об отсутствии связи. Если же проверяемый фактор на самом деле мог влиять на результативный признак, то вывод следует формулировать не в терминах отсутствия связи, а в том, что по изучаемой информации связь надежно не установлена.

Если коэффициент корреляции близок к единице, то распределение его оценок отличается от нормального или распределения Стьюдента, так как он ограничен величиной 1. В таких случаях Р. Фишер предложил для оценки надежности коэффициента преобразовывать его величину в форму, не имеющую такого ограничения:

$$z = 0{,}5 \ln \frac{1 + r_{xy}}{1 - r_{xy}}. \qquad (8.18)$$

Средняя ошибка величины z определяется по формуле

$$m_z = \frac{1}{\sqrt{n-3}}. \qquad (8.19)$$

Величину z можно взять из табл. 6 приложения.

Проверим этим способом надежность коэффициента корреляции надоя молока с затратами на 1 корову:

$$z = 0{,}5 \ln \frac{1 + 0{,}916}{1 - 0{,}916} = 1{,}564;$$

$$m_z = \frac{1}{\sqrt{16-3}} = 0{,}2774; \quad t = \frac{1{,}564}{0{,}2774} = 5{,}64.$$

Значение критерия Стьюдента намного больше его критического значения для значимости 0,01. Следовательно, коэффициент корреляции с очень большой вероятностью больше нуля; связь установлена надежно. Для оценки надежности коэффициента корреляции можно воспользоваться таблицей критических значений для заданных уровней значимости (0,05 или 0,01) и числа степеней свободы (см. приложение, табл. 5).

Например, по выборке объемом 32 единицы получен парный коэффициент корреляции 0,319. Число степеней свободы для него равно 30, поскольку в расчете r участвуют две величины, значения которых закреплены – $\bar{x}$ и $\bar{y}$. За счет этого мы теряем две степени свободы: 32 – 2. Так как критическое значение для 30 степеней свободы равно (при уровне значимости 0,05) 0,3494, то полученное значение ниже критического по модулю. Соответственно, гипотеза о связи признаков надежно не доказана. Неверен вывод и об отсутствии связи – он также надежно не доказан. Из табл. 5 приложения видно, что при малой выборке надежно можно установить только тесные связи, а при большой численности совокупности, например, 102 единицы, надежно измеряются и слабые связи. Этот вывод важен для практической работы по корреляционному анализу.

Можно рассчитать доверительный интервал оценки коэффициента корреляции с заданной вероятностью, скажем, 0,95. При этих условиях и 13 степенях свободы вариации значение t-критерия Стьюдента равно 2,16. Тогда доверительный интервал для z составит: 1,564 ± 2,16 · 0,2774, т. е. от 0,965 до 2,163. Подставив эти граничные значения z в формулу (8.18), получаем границы интервала значений коэффициента корреляции: от 0,974 до 0,747. Как видим, с большой вероятностью связь на самом деле является весьма тесной, коэффициент корреляции не ниже 0,7.

8.6. ПРИМЕНЕНИЕ ПАРНОГО ЛИНЕЙНОГО УРАВНЕНИЯ РЕГРЕССИИ

Прежде чем обсуждать вопросы использования уравнений парной регрессии, напомним, что парный корреляционный анализ не дает чистых мер влияния только одного изучаемого фактора. Если факторы взаимосвязаны, то парная связь измеряет влияние данного фактора и часть влияния прочих факторов, связанных с ним. И все же при тесной связи уравнение регрессии может стать полезным орудием анализа экономических, технологических, социальных или природных процессов.

Сравнивая фактические уровни надоя в табл. 8.1 с расчетными, т. е. такими, которые были бы получены при фактических затратах средств на корову и средней по совокупности эффективности, измеряемой коэффициентом регрессии, можно найти отклонения $y_i - \tilde{y}_i$. Они показывают, насколько хозяйство получило от коров больше или меньше молока в условиях фактической эффективности использования средств, чем при средней по совокупности эффективности

использования средств. Так, в хозяйстве № 6 получено от коровы в среднем 31,8 ц молока, хотя при низком уровне затрат 1355 руб. на корову и средней эффективности затрат было бы получено только по 26,5 ц молока. Фактически надой составил 120% к расчетному. Наоборот, хозяйство № 9 получило по 26,7 ц вместо расчетных 35,6 ц. Следовательно, эффективность использования средств на производство молока в этом хозяйстве (1616 руб. на корову) составила только 26,7 : 35,7 = 75% средней по совокупности.

Оценка хозяйственной деятельности по отклонениям от расчетных значений показателей на основе уравнений регрессии (тем более на основе многофакторных регрессионных моделей) гораздо более оправдана и содержательна, чем оценка результатов производства по отклонениям от среднего значения результативного признака в совокупности без учета факторов – характеристик возможностей и природных условий предприятия.

Уравнения регрессии применимо и для прогнозирования возможных ожидаемых значений результативного признака. При этом следует учесть, что перенос закономерности связи, измеренной в варьирующей совокупности, в статике на динамику не является, строго говоря, корректным и требует проверки условий допустимости такого переноса (экстраполяции), что выходит за рамки статистики и может быть сделано только специалистом, хорошо знающим объект (систему) и возможности его развития в будущем.

Ограничением прогнозирования на основании регрессионного уравнения, тем более парного, служит условие стабильности или по крайней мере малой изменчивости других факторов и условий изучаемого процесса, не связанных с ними. Если резко изменится "внешняя среда" протекающего процесса, прежнее уравнение регрессии результативного признака на факторный потеряет свое значение. В сильно засушливый год доза удобрений может не оказать влияния на урожайность сельскохозяйственной культуры, так как последнюю лимитирует недостаточная влагообеспеченность.

Прогнозируемое значение результативного показателя получается при подстановке в уравнение регрессии ожидаемой величины факторного признака. Так, если подставить в уравнение (8.14) расход средств на корову, равный 2200 руб., то получим ожидаемый надой молока от коровы, равный 55,85 ц. При таком прогнозировании следует соблюдать еще одно ограничение: нельзя подставлять значения факторного признака, значительно отличающиеся от входящих в базисную информацию, по которой вычислено уравнение регрессии. При качественно иных уровнях фактора, если они даже возможны в принципе, были бы другими параметры уравнения.

Можно рекомендовать при определении значений факторов не выходить за пределы трети размаха вариации как за минимальное, так и за максимальное значение признака-фактора, имевшееся в исходной информации.

Прогноз, полученный подстановкой в уравнение регрессии ожидаемого значения фактора, называют *точечным прогнозом*. Вероятность точной реализации такого прогноза крайне мала. Необходимо сопроводить его значением средней ошибки прогноза или *доверительным интервалом прогноза* с достаточно большой вероятностью. Средняя ошибка положения линии регрессии в генеральной совокупности при значении факторного признака, равном x_k, вычисляется для линии регрессии по формуле (8.20)

$$m_{\tilde{y}_k} = s_{y_{ост}} \sqrt{\frac{1}{n} + \frac{(x_k - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}, \qquad (8.20)$$

где $m_{\tilde{y}_k}$ – средняя ошибка положения линии регрессии в генеральной совокупности при $x = x_k$;

n – объем выборки;

x_k – ожидаемое значение фактора;

$s_{y_{ост}}$ – оценка среднего квадратического отклонения результативного признака от линии регрессии в генеральной совокупности с учетом степеней свободы вариации.

По данным табл. 8.1 находим $s_{y_{ост}}$.

$$s_{y_{ост}} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2}{n-2}} = \sqrt{\frac{195{,}4}{16-2}} = 3{,}736 \text{ ц на 1 корову.}$$

При x_k = 2200 руб. на 1 голову имеем:

$$m_{\hat{y}(x_k)} = 3{,}736\sqrt{\frac{1}{16} + \frac{(2200-1605)^2}{818\,533}} = 2{,}629 \text{ ц на 1 корову.}$$

Для вычисления доверительных границ прогноза линии регрессии нужно умножить ее среднюю ошибку на t-критерий Стьюдента. При 14 степенях свободы и доверительной вероятности 0,95 ($\alpha = 0{,}05$) значение t-критерия равно 2,14. Получаем доверительные границы: 55,85 ± 2,629 · 2,14, или от 50,22 до 61,48 ц от 1 коровы. Интервал

довольно широкий. Значительная неопределенность прогноза линии регрессии связана с малым объемом выборки. При объеме совокупности, равном 400, и той же вариации надоев ошибка прогноза была бы в 5 раз меньше и доверительный интервал был бы уже.

Средняя ошибка прогноза для индивидуального значения по правилу о дисперсии суммы независимых переменных образуется из ошибки прогноза положения линии регрессии и среднего квадратического отклонения индивидуальных значений от линии регрессии (остаточной вариации), т. е.

$$m_{y(x_k)} = \sqrt{m_{\tilde{y}_k}^2 + s_{y_{ост}}^2}. \qquad (8.21)$$

В нашем примере имеем:

$$m_{y(x_k)} = \sqrt{2{,}629^2 + 3{,}736^2} = 4{,}568 \text{ ц на 1 корову.}$$

Доверительные границы прогноза индивидуальных значений надоя молока на корову при расходе 2200 руб. на 1 голову составляют с вероятностью нахождения внутри границ, равной 0,95:

$$55{,}85 \pm 4{,}568 \cdot 2{,}14, \text{ или от } 46{,}07 \text{ до } 65{,}63 \text{ ц.}$$

Главным источником ошибки неопределенности прогноза индивидуальных значений служит не столько неопределенность прогноза линии регрессии, сколько значительная вариация надоев за счет других факторов, кроме входящих в уравнение регрессии.

8.7. ВЫЧИСЛЕНИЕ ПАРАМЕТРОВ ПАРНОЙ ЛИНЕЙНОЙ КОРРЕЛЯЦИИ НА ОСНОВЕ АНАЛИТИЧЕСКОЙ ГРУППИРОВКИ

В главе 6 рассмотрен метод аналитической группировки, позволяющий установить наличие, вид и форму связи признаков. Но группировка не дает меры тесноты связи и уравнения регрессии. Теперь, пользуясь методикой корреляционно-регрессионного анализа, можно дополнить аналитическую группировку вычислением этих мер связи.

Возьмем в качестве примера приведенную в главе 6 группировку и рассчитаем необходимые показатели (см. табл. 8.2).

Таблица 8.2

Расчет корреляции по аналитической группировке

Группа предприятий по оборачиваемости в днях	Число предприятий f_j	Среднее число дней x'_j	Средняя прибыль, млн руб. $\bar{y}_j$	$\bar{y}_j - \bar{y}$	$(\bar{y}_j - \bar{y})^2 f_j$	$x' - \bar{x}$	$(x'_j - \bar{x}) \times (\bar{y}_j - \bar{y}) f_j$	$(x'_j - \bar{x})^2 f_j$	$\tilde{y}_{x_j}$	$(\tilde{y}_{x_j} - \bar{y})^2 f_j$
А	1	2	3	4	5	6	7	8	9	10
40 – 50	6	45	14,57	2,80	47,04	–18	–302,4	1944	15,06	64,94
51 – 70	8	60	12,95	1,18	11,14	–3	–28,0	72	12,36	2,78
71 – 101	6	86	7,40	–4,37	114,58	+23	–603,0	3174	7,69	99,88
Итого	20	63	11,77	–	172,76	–	–933,4	5190	11,77	167,60

Коэффициент линейной регрессии

$$b = \frac{\sum_{j=1}^{3}(x'_j - \bar{x})(\bar{y}_j - \bar{y}) f_j}{\sum_{1}^{3}(\bar{x}_j - \bar{x})^2 f_j} = \frac{-933,4}{5190} = -0,18,$$

свободный член уравнения регрессии

$$a = \bar{y} - b\bar{x} = 11,77 - (-0,18 \cdot 63) = 23,15.$$

Итак, имеем уравнение связи: $\tilde{y} = 23,15 - 0,18x$. Вычислим по этому уравнению расчетные значения прибыли $\tilde{y}_i$ для каждой группы. Подставив в уравнение середины интервалов групп $\bar{x}'$, запишем $\tilde{y}_i$ в графу 9 табл. 8.2. Вариация расчетных значений прибыли связана с влиянием оборачиваемости x. Найдем сумму квадратов отклонений прибыли за счет вариации оборачиваемости – факторную вариацию (графа 10 табл. 8.2). Для расчета общей вариации результативного признака была вычислена сумма квадратов отклонений по индивидуальным данным:

$$\sum_{1}^{20}(y_i - \bar{y})^2.$$

Эта сумма квадратов – общая вариация объема прибыли – равна 222,4. Теперь можем построить меры тесноты связи:

теоретическое корреляционное отношение

$$\eta_T^2 = \frac{\sum_{j=1}^{k}(\bar{y}_j - \bar{y})^2 f_j}{\sum_{1}^{20}(y_j - \bar{y})^2} = \frac{167,6}{222,4} = 0,754; \qquad \eta_T = \sqrt{0,754} = 0,868;$$

эмпирическое корреляционное отношение (рассчитанное в гл. 6)

$$\eta_Э = \sqrt{0,777} = 0,881.$$

Оба квадрата корреляционных отношений соответствуют по содержанию ранее рассмотренному коэффициенту детерминации (8.1) и (8.2) и интерпретируются как показатели доли вариации результативного признака, объясняемой за счет вариации группировочного, факторного признака (и, конечно, связанных с ним прочих факторов). В данном примере связь является тесной. Различие в том, что в эмпирическом корреляционном отношении связь признаков не абстрагирована от случайных влияний прочих факторов на вариацию y, не связанных с вариацией x.

Наиболее рациональным приемом анализа и расчета параметров корреляционной связи с помощью группировки является построение так называемой «корреляционной решетки» (табл. 8.3). Это таблица, в которой изучаемая совокупность сгруппирована одновременно по обоим признакам, связь между которыми изучается (двумерное распределение). Число групп по признакам может быть как равным, так и неравным. Если наибольшие числа частот каждой строки и каждого столбца располагаются на первой диагонали (в табл. 8.3 эти цифры подчеркнуты), связь является прямой и близкой к линейной; если наибольшие числа частот располагаются вдоль второй диагонали (в табл. 8.3 эти цифры также подчеркнуты), связь обратная, линейная. Если частоты во всех клетках таблицы примерно равны, связи нет; если наибольшие числа расположены по дуге, связь криволинейная. В табл. 8.3 кроме частот приведены строки и графы

Таблица 8.3

Корреляция между возрастом женихов и невест, вступивших в брак в Санкт-Петербурге в 1991 г.

Возраст женихов, лет x	Середины интервалов x'_i	Возраст невест y					Итого f_i	$x'_i - \bar{x}$	$(x'_i - \bar{x})^2$	$(x'_i - \bar{x})^2 f_i$
		до 25 $y'_1 = 21$	25 – 34 $y'_2 = 30$	35 – 44 $y'_3 = 40$	45 – 54 $y'_4 = 50$	55 и старше $y'_5 = 62$				
До 25	22	18212	1914	147	8	4	20285	–9,2	84,64	1716922,4
25 – 34	30	5574	6677	1112	85	18	13466	–1,2	1,44	19391,04
35 – 44	40	498	2171	2595	419	43	5726	+8,8	77,44	443421,44
45 – 54	50	98	368	1177	1280	308	3231	+18,8	353,44	1141964,6
55 и старше	63	19	75	271	840	1701	2906	+31,8	1011,24	2938663,4
Итого f_j		24401	11205	5302	2632	2074	45614	–		6260362,8
$y'_j - \bar{y}$		–9	+1	+11	+21	+33				
$(y'_j - \bar{y})^2$		81	1	121	441	1089				
$(y'_j - \bar{y})^2 f_j$		1976481	11205	641542	1160712	2258586	6048526			5196031,6

Средний возраст составил для женихов: $\bar{x} = 31{,}2$ года, для невест: $\bar{y} = 29{,}0$ года.

для расчета необходимых сумм при вычислении параметров корреляционной связи.

В табл. 8.3 наибольшие частоты в строках и графах расположены вдоль первой диагонали, что говорит в соответствии с логикой о прямой линейной связи возрастов женихов и невест. Связь эта далеко не полная; как видим, «любви все возрасты покорны», все клетки таблицы заполнены, значит, существуют браки между лицами любых возрастов.

Как средние величины признаков, так и все суммы, входящие в расчет параметров корреляции, при группировке взвешиваются на соответствующие частоты, поэтому формулы (8.9) и (8.11) приобретают следующий вид:

$$b = \frac{\sum_{i=1}^{k_1}\sum_{j=1}^{k_2}(x'_i - \bar{x})(y'_i - \bar{y}) f_{ij}}{\sum_{i=1}^{k_1}(x'_i - \bar{x})^2 f_i}, \tag{8.22}$$

$$r_{x_1 y} = \frac{\sum_{i=1}^{k_1}\sum_{j=1}^{k_2}(x'_i - \bar{x})(y'_j - \bar{y}) f_{ij}}{\sqrt{\sum_{i=1}^{k_1}(x'_j - \bar{x})^2 f_i \sum_{j=1}^{k_2}(y'_j - \bar{y})^2 f_j}}, \tag{8.23}$$

где x'_i, y'_j – 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 (это числа в клетках «корреляционной решетки»).

Взвешенные суммы квадратов отклонений подсчитаны и приведены в последней графе и в последней строке табл. 8.3. Для вычисления числителя в (8.22) и (8.23) необходимо умножить отклонения по обоим признакам (с учетом их знаков) на частоты совместного распределения и сложить все 25 произведений:
$(-9)\cdot(-9{,}2)\cdot 18212 + 1\cdot(-9{,}2)\cdot 1914 + \ldots + 33\cdot 31{,}8\cdot 1701 = 5196031{,}6$.

Это число записано в правом нижнем углу табл. 8.3. Рассчитаем параметры уравнения регрессии. Согласно (8.22)

$$b = \frac{5196031{,}6}{6260362{,}8} = 0{,}830.$$

Это означает, что в среднем с увеличением возраста женихов на 1 год возраст их невест возрастал на 0,83 года. Свободный член уравнения согласно (8.6)

$$a = 29{,}0 - 0{,}83 \cdot 31{,}2 = 3{,}1.$$

Уравнение имеет вид:

$$\hat{y} = 3{,}1 + 0{,}83 \cdot x.$$

Так как оба признака равноправны, то можно получить уравнение зависимости среднего возраста жениха от возраста невесты. Поменяв местами x и y, получаем:

$$b = \frac{5196031{,}6}{6048526} = 0{,}859; \quad a = 31{,}2 - 0{,}859 \cdot 29 = 6{,}3; \quad \hat{x} = 6{,}3 + 0{,}859y.$$

Коэффициент корреляции согласно формуле (8.23) составляет:

$$r_{x,y} = \frac{5196031{,}6}{\sqrt{6260362{,}8 \cdot 6048526}} = 0{,}844.$$

Коэффициент детерминации $r^2 = 71{,}3\%$, т. е. вариация возраста супруга или супруги на 71% зависит от вариации возраста «второй половины». Связь весьма тесная.

Конечно, расчет параметров корреляции на основе группировки является приближенным: реальные значения признаков заменяются серединами интервалов, а при открытых интервалах – их экспертными оценками. Не учитывается неравномерность изменения частот внутри интервалов. Казалось бы, с появлением программ для ЭВМ этот метод должен отмереть. Однако для больших совокупностей в десятки и сотни тысяч единиц большинство программ ввиду ограничений на объем оперативной памяти непригодно. Да и сам процесс занесения в память ЭВМ сотни тысяч чисел занял бы столь громадное время, что выигрыш во времени расчета на ЭВМ был бы многократно превышен. Таким образом, иногда трудоемкость расчета с помощью группировки и простого калькулятора оказывается намного меньше, чем на ЭВМ, а степень точности достаточна для большинства задач анализа связи.

8.8. КОЭФФИЦИЕНТ КОРРЕЛЯЦИИ РАНГОВ

К мерам тесноты парной связи относится и предложенный английским психологом Ч. Спирменом (1863 – 1945) *коэффициент корреляции рангов. Ранги* – это порядковые номера единиц совокупности в ранжированном ряду. Если проранжировать совокупность по двум признакам, связь между которыми изучается, то полное совпадение рангов означает максимально тесную прямую связь, а полная противоположность рангов – максимально тесную обратную связь. Ранжировать оба признака необходимо в одном и том же порядке: либо от меньших значений признака к большим, либо наоборот. Если ранги единиц совокупности по признакам x и y обозначить как p_{xi}, p_{yi}, то коэффициент корреляции рангов согласно (8.11) имеет вид:

$$r_{p_x p_y} = \frac{\sum_{i=1}^{n}(p_{x_i} - \bar{p}_x)(p_{y_i} - \bar{p}_y)}{\sqrt{\sum_{i=1}^{n}(p_{x_i} - \bar{p}_x)^2 \sum_{i=1}^{n}(p_{y_i} - \bar{p}_y)^2}}, \qquad (8.24)$$

где $\bar{p}_x = \bar{p}_y$ – средние ранги в ряду натуральных чисел от 1 до n, равные, как известно, $(n+1)/2$. Также известно, что сумма квадратов отклонений чисел натурального ряда от их средней величины $\sum_{i=1}^{n}(p_{x_i} - \bar{p}_x)^2$ и $\sum_{i=1}^{n}(p_{y_i} - \bar{p}_y)^2$ равна $(n^3 - n)/12$. Следовательно, знаменатель формулы (8.23) есть $(n^3 - n)/12$.

Рассмотрим далее разности рангов $d_i = p_{x_i} - p_{y_i}$ и сумму их квадратов:

$$\sum_{i=1}^{n} d_i^2 = \sum_{i=1}^{n}(p_{x_i} - p_{y_i})^2 = \sum_{i=1}^{n}\left[(p_{x_i} - \bar{p}_x) - (p_{y_i} - \bar{p}_y)\right]^2 =$$

$$= \sum_{i=1}^{n}(p_{x_i} - \bar{p}_x)^2 + \sum_{i=1}^{n}(p_{y_i} - \bar{p}_y)^2 - 2\sum_{i=1}^{n}(p_{x_i} - \bar{p}_x)(p_{y_i} - \bar{p}_y) =$$

$$= \frac{2(n^3 - n)}{12} - 2\sum_{i=1}^{n}(p_{x_i} - \bar{p}_x)(p_{y_i} - \bar{p}_y).$$

Отсюда

$$\sum_{i=1}^{n}(p_{x_i} - \bar{p}_x)(p_{y_i} - \bar{p}_y) = \left(\frac{2(n^3 - n)}{12} - \sum_{i=1}^{n} d_i^2\right) : 2 = \frac{n^3 - n}{12} - \frac{\sum_{i=1}^{n} d_i^2}{2}.$$

Это числитель коэффициента корреляции рангов. Подставив в (8.24) найденные выражения для числителя и для знаменателя, имеем:

$$r_{p_x p_y} = \frac{\frac{(n^3-n)}{12} - \frac{\sum_{i=1}^{n} d_i^2}{2}}{\frac{n^3-n}{12}} = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n^3-n}. \qquad (8.25)$$

Это и есть *формула Спирмена.*

Преимущество коэффициента корреляции рангов состоит в том, что ранжировать можно и по таким признакам, которые нельзя выразить численно: можно проранжировать кандидатов на занятие определенной должности по профессиональному уровню, по умению руководить коллективом, по личному обаянию и т. п. При экспертных оценках можно ранжировать оценки разных экспертов и найти их корреляции друг с другом, чтобы затем исключить из рассмотрения оценки эксперта, слабо коррелированные с оценками других экспертов. Коэффициент корреляции рангов, как будет показано в гл. 9, применяется для оценки устойчивости тенденции динамики.

Недостатком коэффициента корреляции рангов является то, что одинаковым разностям рангов могут соответствовать совершенно отличные разности значений признаков (в случае количественных признаков). Поэтому для последних следует считать корреляцию рангов, как и коэффициент знаков Фехнера, приближенными мерами тесноты связи, обладающими меньшей информативностью, чем коэффициент корреляции числовых значений признаков.

В качестве примера рассчитаем коэффициент корреляции рангов по данным табл. 8.1 (табл. 8.4).

Коэффициент корреляции рангов по формуле Спирмена

$$r_{p_x p_y} = 1 - \frac{6 \cdot 138}{16^3 - 16} = 0{,}797.$$

Полученное значение больше коэффициента Фехнера, но намного ниже обычного коэффициента корреляции, составившего 0,916. Как видим, недоучет размеров отклонений признаков от их средних величин занижает меру тесноты связи.

Если среди значений признаков x и y встречается несколько одинаковых, образуются *связанные ранги,* т. е. одинаковые средние номера; например, вместо одинаковых по порядку третьего и четвертого значений признака будут два ранга по 3,5. В таком случае коэффициент Спирмена вычисляется как

$$r_{p_x p_y} = 1 - \frac{6\sum d^2 - A - B}{\sqrt{(n^3 - n - 12A)(n^3 - n - 12B)}}, \qquad (8.26)$$

где:

$$A = \frac{1}{12}\sum_j (A_j^3 - A_j); \quad B = \frac{1}{12}\sum_k (B_k^3 - B_k);$$

j – номера связок по порядку для признака x;
A_j – число одинаковых рангов в j-й связке по x;
k – номера связок по порядку для признака y;
B_k – число одинаковых рангов в k-й связке по y.

Таблица 8.4

Расчет коэффициента корреляции рангов по данным табл. 8.1

Номера хозяйств	Ранг по затратам на 1 голову p_x	Ранг по надою молока p_y	$d = p_x - p_y$	d^2
1	7	10	–3	9
2	1	1	0	0
3	2	3	–1	1
4	13	6	7	49
5	6	9	–3	9
6	3	5	–2	4
7	4	4	0	0
8	5	7	–2	4
9	9	2	7	49
10	14	14	0	0
11	11	12	–1	1
12	12	11	1	1
13	10	13	–3	9
14	8	8	0	0
15	16	15	1	1
16	15	16	–1	1
Σ	136	136	0	138

Коэффициент корреляции рангов может быть рассчитан и по формуле, предложенной английским статистиком М. Кендаллом:

$$\tau = \frac{s}{\frac{1}{2}n(n-1)}, \qquad (8.27)$$

где s – фактическая сумма рангов;

$\frac{1}{2}n(n-1)$ – максимальная сумма рангов.

Этот коэффициент также изменяется в пределах $-1 < \tau < 1$. Он дает несколько более строгую оценку связи нежели коэффициент Спирмена: $r_s \approx \frac{3}{2}\tau$. Это соотношение выполняется при большом числе наблюдений, $n > 30$, и слабых либо умеренно тесных связях. Для расчета τ все единицы ранжируются по признаку x; по ряду другого признака y подсчитывается для каждого ранга число последующих рангов, превышающих данный (их сумму обозначим P), и число последующих рангов ниже данного (их сумму обозначим Q). Тогда $S = P - Q$. Можно показать, что $P+Q=\frac{1}{2}n(n-1)$, так что τ может быть представлен как

$$\tau = \frac{P-Q}{P+Q}. \qquad (8.28)$$

Вычислим коэффициент корреляции рангов Кендалла по данным табл. 8.4:

Ранги по x	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Ранги по y	1	3	5	4	7	9	10	8	2	13	12	11	6	14	16	15

$S = (15 - 0) + (13 - 1) + (11 - 2) + (11 - 1) + (9 - 2) + (7 - 3) + (6 - 3) + (6 - 2) + (7 - 0) + (3 - 3) + (3 - 2) + (3 - 1) + (3 - 0) + (2 - 0) + (0 - 1) = 99 - 21 = 78;$

$\tau = \frac{78}{\frac{1}{2}16 \cdot 15}$. Хотя $\tau < r_s$ ($r_s = 0{,}797$), но поскольку связь тесная со-

отношение между этими двумя коэффициентами не вполне соответствует упомянутому: коэффициент Спирмена в нашем примере превосходит τ не в 1,5 раза, а на 23%.

8.9. ПАРАБОЛИЧЕСКАЯ КОРРЕЛЯЦИЯ

Линейные связи являются основными. Однако встречаются и нелинейные связи, хорошо описываемые параболой, гиперболой и т. д.

Уравнение регрессии в форме параболы 2-го порядка имеет следующий вид:

$$\hat{y} = a + bx + cx^2. \qquad (8.29)$$

Если при линейной связи среднее изменение результативного признака на единицу фактора постоянно по всей области вариации фактора, то при параболической корреляции изменение признака x на единицу признака y меняется равномерно с изменением величины фактора. В результате связь может даже поменять знак на противоположный, из прямой превратится в обратную, из обратной в прямую. Такой характер связи объективно присущ многим системам. Например, с увеличением дозы удобрений урожайность сельхозкультур сначала повышается, но если превысить оптимальную величину дозы, то при дальнейшем росте дозы удобрений растения угнетаются и урожайность снижается.

Нормальные уравнения метода наименьших квадратов для параболы 2-го порядка таковы:

$$na + b\sum_{1}^{n} x_i + c\sum_{1}^{n} x_i^2 = \sum_{1}^{n} y_i,$$

$$a\sum_{1}^{n} x_i + b\sum_{1}^{n} x_i^2 + c\sum_{1}^{n} x_i^3 = \sum_{1}^{n} y_i x_i,$$

$$a\sum_{1}^{n} x_i^2 + b\sum_{1}^{n} x_i^3 + c\sum_{1}^{n} x_i^4 = \sum_{1}^{n} y_i x_i^2.$$

Если расчет производится не по индивидуальным данным, а на основе аналитической группировки, то уравнения МНК приобретают следующий вид:

$$a\sum_{j=1}^{k} f_j + b\sum_{j=1}^{k} x'_j f_j + c\sum_{j=1}^{k} x'^2 f_j = \sum_{j=1}^{k} y'_j f_j,$$

$$a\sum_{j=1}^{k} x'_j f_j + b\sum_{j=1}^{k} x'^2_j f_j + c\sum_{j=1}^{k} x'^3_j f_j = \sum_{j=1}^{k} y'_j f_j x'_j,$$

$$a\sum_{j=1}^{k} x'^2_j f_j + b\sum_{j=1}^{k} x'^3_j f_j + c\sum_{j=1}^{k} x'^4_j f_j = \sum_{j=1}^{k} y'_j f_j x'^2_j.$$

Решая эту систему, получаем значения параметров a, b и c. Показателем тесноты параболической корреляции является корреляционное отношение, вычисляемое как корень квадратный из выражения (8.2).

В качестве примера параболической корреляционной связи рассмотрим зависимость себестоимости молока от продуктивности коров по данным аналитической группировки сельхозпредприятий области (табл. 8.5). В этой же таблице приведены расчетные величины, входящие в уравнения МНК для параболы.

Были получены нормальные уравнения МНК:

$$136a + 525b + 2123{,}4c = 4585{,}1,$$
$$525a + 2123{,}4b + 9017{,}1c = 17318{,}1,$$
$$2123{,}4a + 9017{,}1b + 40199{,}3c = 68586{,}4.$$

Решив систему уравнений, получили:

$$c = 2{,}3249;\ b = -23{,}64;\ a = 88{,}68.$$

Уравнение регрессии имеет вид:

$$\hat{y} = 88{,}68 - 23{,}641x + 2{,}3249x^2 \qquad (8.30)$$

Эта парабола имеет точку минимума в фактической области вариации факторного признака. Для нахождения значения фактора, при котором достигается минимальное значение результативного признака, следует приравнять нулю первую производную по x уравнения (8.30):

$$\frac{d\hat{y}}{dx} = -23{,}641 + 2 \cdot 2{,}3249x,$$

откуда x = 23,641/4,6498 = 5,084 т молока на корову.

Итак, минимальная себестоимость молока в совокупности предприятий, в условиях периода, к которому относятся данные, достигалась в среднем при надое молока на корову 5084 кг. Значение фактора x при

Таблица 8.5

Параболическая корреляционная связь продуктивности коров и себестоимости молока

Группы предприятий по надою x_j т	Число предприятий f_j	Середина интервалов x'_j т	Средняя себестоимость молока, руб. за 1 т $\bar{y}_j$	$x'_j f_j$	$x'^2_j f_j$	$x'^3_j f_j$	$x'^4_j f_j$	$\bar{y}_j f_j$	$\bar{y}_j f_j x'_j$	$\bar{y}_j f_j x'^2_j$	$\tilde{y}_j$	$\bar{y}_j - \bar{y}$	$(\bar{y}_j - \bar{y})^2 f_j$
До 3,0	16	2,80	41,25	44,8	125,4	351,2	983,4	660,0	1848,0	5174,4	40,71	+7,00	784,0
3,0–3,5	42	3,25	36,67	136,5	443,6	1441,8	4685,8	1540,1	5005,5	16267,7	36,40	+2,69	303,9
3,5–4,0	31	3,75	32,50	116,2	435,9	1634,8	6130,4	1007,5	3778,1	14168,0	32,72	–0,99	30,4
4,0–4,5	18	4,25	28,89	76,5	325,1	1381,8	5872,6	520,0	2210,1	9392,9	30,20	–3,51	221,8
4,5–5,0	13	4,75	29,04	61,8	293,3	1393,2	6617,9	377,5	1793,2	8517,8	28,84	–4,87	308,3
5,0–5,5	9	5,25	29,17	47,2	248,1	1302,3	6837,2	262,5	1378,3	7236,3	28,64	–5,07	321,3
Более 5,5	7	6,00	31,07	42,0	252,0	1512,0	9072,0	217,5	1304,9	7829,6	30,53	–3,18	70,8
Итого	136	–	33,71	525,0	2123,4	9017,1	40199,3	4585,1	17318,1	68586,4	–	–	1950,5

достижении минимума себестоимости можно назвать оптимальной продуктивностью коров, а сама задача его поиска – это одна из оптимизационных задач, решаемая математико-статистическим методом.

Для измерения тесноты параболической корреляционной связи находим вариацию результативного признака y, объясняемую вариацией фактора x как сумму квадратов отклонений расчетных величин $\tilde{y}$ от средней величины $\bar{y}$, взвешенных на число предприятий. Общая сумма квадратов отклонений всех 136 значений y_i от средней величины составляет 4624,7. Таким образом согласно формуле (8.1), корреляционное отношение

$$\eta = \sqrt{\frac{1950,5}{4624,7}} = \sqrt{0,422} = 0,650.$$

Поскольку η^2 – аналог коэффициента детерминации, можно сделать вывод, что 42,2% вариации себестоимости молока в совокупности 136 предприятий были связаны с вариацией продуктивности коров (и с факторами, варьирующими согласованно с продуктивностью в соответствии с ранее сделанной оговоркой об интерпретации парных связей).

8.10. ГИПЕРБОЛИЧЕСКАЯ КОРРЕЛЯЦИЯ

Уравнение регрессии в форме гиперболы имеет следующий вид:

$$\hat{y} = a + \frac{b}{x}. \qquad (8.31)$$

Если величина b положительна, то при увеличении значений факторного признака x значения результативного признака уменьшаются, причем это уменьшение все время замедляется, и при $x \to \infty$ средняя величина признака y будет равна a. Если же параметр b отрицателен, то значения результативного признака с ростом фактора возрастают, причем их рост замедляется, и в пределе при $x \to \infty$ $\tilde{y} = a$. Таким образом, гиперболические зависимости характерны для связей, в которых результативный признак не может варьировать неограниченно, его вариация имеет односторонний предел. Например, при освоении нового оборудования его производительность возрастет, но рост замедлится по мере приближения к конструктивно-технологическому пределу производственной мощности агрегата. Совершенствуя двигатель, можно увеличивать его КПД, но тоже не выше предела, допускаемого данным видом преобразования энергии. Таков же характер связи между уровнем душевого дохода x в семье и долей семей, имеющих телевизоры, y; он приближен к пределу (100%) в наиболее обеспеченной группе семей.

Нормальные уравнения метода наименьших квадратов для гиперболы таковы:

$$\begin{cases} na + b\sum\limits_{1}^{n}\frac{1}{x_i} = \sum\limits_{1}^{n} y_i; \\ a\sum\limits_{1}^{n}\frac{1}{x_i} + b\sum\limits_{1}^{n}\frac{1}{x_i^2} = \sum\limits_{1}^{n}\frac{y_i}{x_i}. \end{cases}$$

Легко видеть, что эти уравнения, по существу, те же, что и для линейной связи. Линеаризация гиперболического уравнения достигается заменой $1/x$ на новую переменную, которую можно обозначить z. Тогда уравнение (8.27) примет вид $\tilde{y} = a + bz$. Это и следует сделать, вычисляя гиперболу на компьютере, если программа для него не предусматривает автоматического вычисления гиперболических регрессий.

В качестве примера расчета уравнения гиперболической связи рассмотрим влияние среднесуточного прироста живой массы крупного рогатого скота на откорме на себестоимость прироста живой массы в совокупности предприятий области, занимавшихся откормом скота (табл. 8.6).

Таблица 8.6

Гиперболическая связь себестоимости прироста со скоростью прироста массы скота

Группы предприятий по среднесуточному приросту массы граммов на 1 голову x_i	Число предприятий f_j	Средняя себестоимость прироста руб./ц $\bar{y}_j$	Середина интервалов x'_j в сотнях граммов на голову	$\frac{f_j}{x'_j}$	$\frac{f_j}{(x'_j)^2}$	$\bar{y}_j f_j$	$\frac{\bar{y}_j \cdot f_j}{x'_j}$	$\tilde{y}_j$
334–425	22	496	3,8	5,79	1,52	10912	2872	513
425–516	37	425	4,7	7,87	1,67	15725	3346	419
516–607	28	360	5,6	5,00	0,89	10080	1800	356
607–698	27	310	6,5	4,15	0,64	8370	1288	310
698–789	9	283	7,4	1,22	0,16	2547	344	275
Итого	123	387	–	24,03	4,88	47634	9650	–

Нормальные уравнения МНК имеют вид:

$$\begin{cases} 123a + 24{,}03b = 47634; \\ 24{,}03a + 4{,}88b = 9650. \end{cases}$$

Решение системы уравнений дает значения показателей:

$$a = 24{,}44;\ b = 1857.$$

Уравнение регрессии имеет вид:

$$y = 24{,}44 + \frac{1857}{x},$$

где x в сотнях граммов.

Точечный прогноз по уравнению регрессии при среднесуточном приросте массы животных, равном 900 г, уже достигнутом передовыми хозяйствами, приводит к ожидаемой средней себестоимости, $y = 24{,}44 + \frac{1857}{9} = 230{,}77$ руб./ц.

Корреляционное отношение для данной связи, рассчитанное по формуле (8.1):

$$\eta = \sqrt{\frac{687047}{1023451}} = \sqrt{0{,}671} = 0{,}819.$$

Следовательно, 67% вариации себестоимости прироста массы скота объяснились вариацией скорости роста массы животных и связанных с ней других факторов, например, чем быстрее растет масса, тем меньше расход кормов на единицу прироста массы.

8.11. МНОЖЕСТВЕННОЕ УРАВНЕНИЕ РЕГРЕССИИ

Проблемы множественного корреляционно-регрессионного анализа и моделирования подробно изучаются в специальном курсе того же названия. В курсе «Общая теория статистики» рассматриваются только самые общие вопросы этой сложной проблемы и дается начальное представление о методике построения уравнения множественной регрессии и показателей связи. Рассмотрим линейную

форму многофакторных связей не только как наиболее простую, но и как форму, предусмотренную пакетами прикладных программ для ПЭВМ. Если же связь отдельного фактора с результативным признаком не является линейной, то производят линеаризацию уравнения путем замены или преобразования величины факторного признака.

Общий вид многофакторного уравнения регрессии имеет вид:

$$\hat{y} = a + b_1 x_1 + \ldots + b_k x_k = a + \sum_{j=1}^{k} b_j x_j, \qquad (8.32)$$

где k – число факторных признаков.

Чтобы упростить систему уравнений МНК, необходимую для вычисления параметров уравнения (8.32), обычно вводят величины отклонений индивидуальных значений всех признаков от средних величин этих признаков.

$$\Delta y_i = y_i - \bar{y};\ \ \Delta x_{ij} = x_{ji} - \bar{x}_j.$$

Получаем систему k уравнений МНК:

$$b_1 \sum_{i=1}^{n} \Delta^2 x_{1i} + b_2 \sum_{i=1}^{n} \Delta x_{1i} \Delta x_{2i} + \ldots + b_k \sum_{i=1}^{n} \Delta x_{1i} \Delta x_{ki} = \sum_{i=1}^{n} \Delta y_i \Delta x_{1i},$$

$$b_1 \sum_{i=1}^{n} \Delta x_{1i} \Delta x_{2i} + b_2 \sum_{i=1}^{n} \Delta^2 x_{2i} + \ldots + b_k \sum_{i=1}^{n} \Delta x_{2i} \Delta x_{ki} = \sum_{i=1}^{n} \Delta y_i \Delta x_{2i},$$

$$\cdots \qquad \cdots \qquad \cdots \qquad \cdots$$

$$b_1 \sum_{i=1}^{n} \Delta x_{1i} \Delta x_{ki} + b_2 \sum_{i=1}^{n} \Delta x_{2i} \Delta x_{ki} + \ldots + b_k \sum_{i=1}^{n} \Delta^2 x_{ki} = \sum_{i=1}^{n} \Delta y_i \Delta x_{ki}.$$

Решая эту систему, получаем значения коэффициентов условно-чистой регрессии b_j. Свободный член уравнения вычисляется по формуле

$$a = \bar{y} - \sum_{j=1}^{k} b_j \bar{x}_j. \qquad (8.33)$$

Термин «*коэффициент условно-чистой регрессии*» означает, что каждая из величин b_j измеряет среднее по совокупности отклонение результативного признака от его средней величины при откло-

нении данного фактора x_j от своей средней величины на единицу его измерения и *при условии, что все прочие факторы, входящие в уравнение регрессии, закреплены на средних значениях,* не изменяются, не варьируют.

Таким образом, в отличие от коэффициента парной регрессии коэффициент условно-чистой регрессии измеряет влияние фактора, абстрагируясь от связи вариации этого фактора с вариацией остальных факторов. Если было бы возможным включить в уравнение регрессии все факторы, влияющие на вариацию результативного признака, то величины b_j можно было бы считать мерами *чистого* влияния факторов. Но так как реально невозможно включить все факторы в уравнение, то коэффициенты b_j не свободны от примеси влияния факторов, не входящих в уравнение.

Включить все факторы в уравнение регрессии невозможно по одной из трех причин или сразу по ним всем, так как: 1) часть факторов может быть неизвестна современной науке, познание любого процесса всегда неполное; 2) по части известных теоретических факторов нет информации либо таковая ненадежна; 3) численность изучаемой совокупности (выборки) ограничена, что позволяет включить в уравнение регрессии ограниченное число факторов.

Коэффициенты условно-чистой регрессии b_j являются именованными числами, выраженными в разных единицах измерения, и поэтому несравнимы друг с другом. Для преобразования их в сравнимые относительные показатели применяется то же преобразование, что и для получения коэффициента парной корреляции. Полученную величину называют *стандартизованным коэффициентом регрессии или* β*-коэффициентом.*

$$\beta_j = b_j \frac{\sigma_{x_j}}{\sigma_y}. \qquad (8.34)$$

β-коэффициент при факторе x_j, определяет меру влияния вариации фактора x_j на вариацию результативного признака y при отвлечении от сопутствующей вариации других факторов, входящих в уравнение регрессии.

Коэффициенты условно-чистой регрессии полезно выразить в виде относительных сравнимых показателей связи, *коэффициентов эластичности:*

$$e_j = b_j \frac{\bar{x}_j}{\bar{y}}. \qquad (8.35)$$

Коэффициент эластичности фактора x_j, говорит о том, что при отклонении величины данного фактора от его средней величины на 1% и при отвлечении от сопутствующего отклонения других факторов, входящих в уравнение, результативный признак отклонится от своего среднего значения на e_j процентов от $\bar{y}$. Чаще интерпретируют и применяют коэффициенты эластичности в терминах динамики: при увеличении фактора x_j на 1% его средней величины результативный признак увеличится на e_j процентов его средней величины.

Рассмотрим расчет и интерпретацию уравнения многофакторной регрессии на примере тех же 16 хозяйств (табл. 8.1). Результативный признак – уровень валового дохода и три фактора, влияющих на него, представлены в табл. 8.7.

Напомним еще раз, что для получения надежных и достаточно точных показателей корреляционной связи необходима более многочисленная совокупность.

Таблица 8.7

Уровень валового дохода и его факторы

Номера хозяйств	Валовой доход, руб./га y	Затраты труда, чел.-дни/га x_1	Доля пашни, % x_2	Надой молока на 1 корову, кг, x_3
1	704	265	45,1	3422
2	293	193	35,1	1956
3	346	229	69,4	2733
4	420	193	60,2	3254
5	691	225	59,0	3323
6	679	255	63,4	3179
7	457	201	58,1	3073
8	503	208	51,8	3257
9	314	170	73,2	2669
10	803	276	59,0	4235
11	691	188	42,5	3790
12	775	232	50,5	3658
13	584	173	48,6	3801
14	504	183	51,9	3266
15	777	236	58,9	5173
16	1138	265	38,8	5526
Сумма	9679	3492	865,5	56315
Средняя	604,9	218,2	54,1	3520
s	221,9	34,6	10,6	887
v, %	36,7	15,9	19,6	25,2

Таблица 8.8

Показатели уравнения регрессии

Dependent variable: *y*					
Var.	Regression coefficient	Std. error	T (DF = 12)	Prob.	Partial r^2
*x*1	2,260978	,680030	3,325	,00606	,4795
*x*2	–4,307303	1,982283	–2,173	,05053	,2824
*x*3	,166091	,027050	6,140	,00005	,7586
Constant	–240,112905				
Std. error of est. = 79,243276					

Решение проведено по программе «Microstat» для ПЭВМ. Приведем таблицы из распечатки: табл. 8.7 дает средние величины и средние квадратические отклонения всех признаков. Табл. 8.8 содержит коэффициенты регрессии и их вероятностную оценку: первая графа «var» – переменные, т. е. факторы; вторая графа «regression coefficient» – коэффициенты условно-чистой регрессии *bj*; третья графа «std. errror» – средние ошибки оценок коэффициентов регрессии; четвертая графа – значения *t*-критерия Стьюдента при 12 степенях свободы вариации; пятая графа «prob» – вероятности нулевой гипотезы относительно коэффициентов регрессии; шестая графа «partial r^2» – частные коэффициенты детерминации. Содержание и методика расчета показателей в графах 3 – 6 рассматриваются далее в главе 8. «Constant» – свободный член уравнения регрессии *a*; «Std. error of est.» – средняя квадратическая ошибка оценки результативного признака по уравнению регрессии.

Было получено уравнение множественной регрессии:

$$\hat{y} = 2{,}26x_1 - 4{,}31x_2 + 0{,}166x_3 - 240.$$

Это означает, что величина валового дохода на 1 га сельхозугодий в среднем по совокупности возрастала на 2,26 руб. при увеличении затрат труда на 1 ч/га; уменьшалась в среднем на 4,31 руб. при возрастании доли пашни в сельхозугодиях на 1% и увеличивалась на 0,166 руб. при росте надоя молока на корову на 1 кг. Отрицательная величина свободного члена вполне закономерна, и, как уже отмечено в п. 8.2, результативный признак – валовой доход –

становится нулевым задолго до достижения нулевых значений факторов, которое в производстве невозможно.

Отрицательное значение коэффициента при x_2 – сигнал о существенном неблагополучии в экономике изучаемых хозяйств, где растениеводство убыточно, а прибыльно только животноводство. При рациональных методах ведения сельского хозяйства и нормальных ценах (равновесных или близких к ним) на продукцию всех отраслей, доход должен не уменьшаться, а возрастать с увеличением наиболее плодородной доли в сельхозугодиях – пашни.

На основе данных предпоследних двух строк табл. 8.7 и табл. 8.8 рассчитаем β-коэффициенты и коэффициенты эластичности согласно формулам (8.34) и (8.35).

Как на вариацию уровня дохода, так и на его возможное изменение в динамике самое сильное влияние оказывает фактор x_3 – продуктивность коров, а самое слабое – x_2 – доля пашни. Значения β^2_j будут использоваться в дальнейшем (табл. 8.9).

Таблица 8.9

Сравнительное влияние факторов на уровень дохода

Факторы x_j	β_j	e_j	β_j^2
x_1	0,352	0,816	0,138
x_2	–0,206	–0,385	0,042
x_3	0,664	0,966	0,441

Между β-коэффициентом и коэффициентом эластичности существует следующее соотношение:

$$\beta_j = b_j \frac{\sigma_{x_j}}{\sigma_y};$$

в качестве σ_{x_j}, σ_y приняты их оценки s_x, s_y (табл. 8.7).

$$e_j = b_j \frac{\bar{x}_j}{\bar{y}}.$$

Разделим β_j на e_j: $\dfrac{\beta_j}{e_j} = b_j \dfrac{s_{x_j}}{s_{x_y}} : b_j \dfrac{\bar{x}_j}{\bar{y}} = \dfrac{s_{x_j}}{\bar{x}_j} : \dfrac{s_y}{\bar{y}} = v_{x_j} : v_y.$ (8.36)

Итак, мы получили, что β-коэффициент фактора x_j относится к коэффициенту эластичности этого фактора, как коэффициент вариации фактора к коэффициенту вариации результативного признака. Поскольку, как видно по последней строке табл. 8.7, коэффициенты вариации всех факторов меньше коэффициента вариации результативного признака; все β-коэффициенты меньше коэффициентов эластичности.

Рассмотрим соотношение между парным и условно-чистым коэффициентом регрессии на примере фактора x_1. Парное линейное уравнение связи y с x_j имеет вид:

$$\hat{y} = 3{,}886x_1 - 243{,}2.$$

Условно-чистый коэффициент регрессии при x_1 составляет только 58% парного. Остальные 42% связаны с тем, что вариации x_1 сопутствует вариация факторов x_2 x_3, которая, в свою очередь, влияет на результативный признак y. Связи всех признаков и их коэффициенты парных регрессий представлены на графе связей (рис. 8.2).

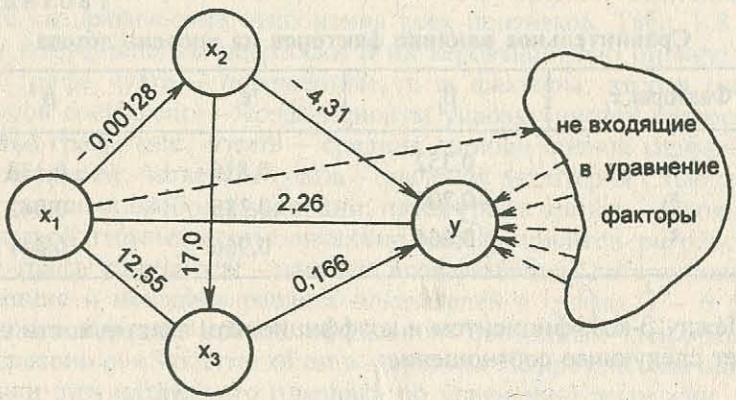


Рис. 8.2. Граф связей

Если сложить оценки прямого и опосредованного влияния вариации x_1 на y, т. е. произведения коэффициентов парных регрессий по всем «путям» (рис. 8.2), получим: $2{,}26 + 12{,}55 \cdot 0{,}166 + (-0{,}00128) \cdot (-4{,}31) + (-0{,}00128) \cdot 17{,}00 \cdot 0{,}166 = 4{,}344$.

Эта величина даже больше парного коэффициента связи x_1 с y. Следовательно, косвенное влияние вариации x_1 через не входящие в уравнение признаки-факторы – обратное, дающее в сумме:

$$3{,}886 - 4{,}344 = -0{,}458.$$

8.12. МЕРЫ ТЕСНОТЫ СВЯЗЕЙ В МНОГОФАКТОРНОЙ СИСТЕМЕ

Многофакторная система требует уже не одного, а множества показателей тесноты связей, имеющих разный смысл и применение. Основой измерения связей является матрица парных коэффициентов корреляции (табл. 8.10).

Таблица 8.10

Матрица парных коэффициентов корреляции

Признаки	y	x_1	...	x_k
y	1			
x_1	r_{yx_1}	1		
.				
.	r_{yx_j}	$r_{x_1x_j}$ ...	1	
.			...	
x_k	r_{yx_k}	$r_{x_1x_k}$	$r_{x_jx_k}$ ...	1

По этой матрице можно судить о тесноте связи факторов с результативным признаком и между собой. Хотя все эти показатели относятся к парным связям, все же матрицу можно использовать для предварительного отбора факторов для включения в уравнение регрессии. Не рекомендуется включать в уравнение факторы слабо связанные с результативными признаками, но тесно связанные с другими факторами. Если, например, имеем: $r_{yx1} = 0{,}8$; $r_{yx2} = 0{,}65$; $r_{x1x2} = 0{,}88$, то в регрессионное уравнение следует включить фактор x_1, а фактор x_2 не включать, так как он тесно связан с x_1 (коллинеарен с x_1), и его корреляция с y слабее, чем корреляция фактора x_1. Совершенно недопустимо включать в анализ факторы, функционально связанные друг с другом, т. е. с коэффициентом корреляции, равным единице. Включение таких пар признаков приводит к вырожденной матрице коэффициентов и неопределенности решения. В этом случае решение задачи на ПЭВМ прекращается.

Матрица парных коэффициентов для нашего примера (табл. 8.11) говорит об отсутствии коллинеарных (т. е. линейно связанных) факторов, что позволяет включить все эти факторы в уравнении регрессии.

На основе этой матрицы вычисляется наиболее общий показатель тесноты связи всех входящих в уравнение регрессии факторов

Таблица 8.11

Матрица парных коэффициентов корреляции

Признаки	y	x_1	x_2	x_3
y	1			
x_1	0,687	1		
x_2	–0,355	–0,044	1	
x_3	0,878	0,490	–0,203	1

с результативным признаком – коэффициент множественной детерминации $R^2_{y,x_1\ldots x_k}$ как частное от деления определителя матрицы Δ^* на определитель матрицы Δ, где

$$\Delta^* = \begin{vmatrix} r_{yx_1} & r_{yx_2} & \ldots & r_{yx_k} & 0 \\ 1 & r_{x_1x_2} & \ldots & r_{x_1x_k} & r_{yx_1} \\ r_{x_2x_1} & 1 & \ldots & r_{x_2x_k} & r_{yx_2} \\ \vdots & \vdots & & \vdots & \vdots \\ r_{x_1x_k} & r_{x_2x_k} & \ldots & 1 & r_{yx_k} \end{vmatrix};$$

$$\Delta = \begin{vmatrix} 1 & r_{x_1x_2} & \ldots & r_{x_1x_k} \\ r_{x_2x_1} & 1 & \ldots & r_{x_2x_k} \\ \vdots & \vdots & & \vdots \\ r_{x_1x_k} & r_{x_2x_k} & \ldots & 1 \end{vmatrix}.$$

Этим способом можно определить величину R^2, не вычисляя расчетных значений результативного признака $\hat{y}_i$ для всех единиц совокупности. Если полученная величина R^2 не удовлетворяет исследователя, то можно прекратить дальнейшие вычисления и не рассчитывать $\hat{y}_i$ (это имеет значение, если совокупность состоит из сотен и тысяч единиц).

Принципиальное *содержание множественного коэффициента детерминации*, как и парного, раскрывается формулой (8.2). *Это отношение части вариации результативного признака, объясняемой за счет вариации входящих в уравнение факторов, к общей вариации результативного признака за счет всех факторов.* Здесь под «вариацией» понимается сумма квадратов отклонений индивидуальных расчетных по уравнению величин от средней («объясненная вариация») и первичных индивидуальных величин от средней («общая вариация»).

В нашем примере значение сумм квадратов отклонений и коэффициенты детерминации и корреляции приведены по распечатке программы «Microstat» в табл. 8.12.

Таблица 8.12

Показатели множественной корреляционной связи

Adjusted R squared = .872390
R squared = .897912
Multiple R = .947582

Analysis of variance table

	Sum of squares	D. F	Mean square	F-Ratio	Prob
Regression	662772,975894	3	220924,325298	35.182	3.171E-06
Residual	75353.961606	12	6279.496801		
Total	738126.937500	15			

Верхняя строка: корректированный R-квадрат = 0,872390; вторая строка: R-квадрат = 0,897912; третья строка: множественный R = 0,947582. Затем приводится таблица дисперсионного анализа, в которой указываются источники вариации: *объясненная* сумма квадратов отклонений значений, рассчитанных по уравнению регрессии, от среднего значения $D_{объясн} = \Sigma(\hat{y}_i - \bar{y})^2 = 662\,772{,}98$ при числе степеней свободы, равном числу объясняющих переменных $dfk = 3$; *остаточная* – отклонения фактических значений от расчетных $D_{ост} = \Sigma(y_i - \hat{y}_i)^2 = 75353{,}96$ при числе степеней свободы, равном $df = n - k - 1$, $df = 12$; *общая* – $\Sigma(y_i - \bar{y})^2 = 738\,126{,}94$, при числе степеней свободы $df = n - 1$, $df = 15$. Затем приводится средний квадрат отклонений: $s^2_1 = D_{объясн} : df_{объясн} = 662772{,}98 : 3 = 220924{,}3$; $s^2_2 = D_{ост} : df_{ост} = 75353{,}96 : 12 = 6279{,}5$. Далее указано их отношение, т. е. s^2_1/s^2_2 = F-критерию. Наконец, указывается вероятность ошибочного решения, т. е. нулевого R^2, равная 0,000003171.

Три фактора, включенные в уравнение регрессии, объясняют 89,8% вариации уровня валового дохода, если рассматривать 16 хозяйств как генеральную совокупность, не считаясь с ее ограниченной численностью (некорректированный коэффициент детерминации равен 0,8979). Если же учесть конечность объема совокупно-

сти n, число факторов k, а также свойство метода, по которому по мере приближения числа k к числу n коэффициент детерминации автоматически приближается к единице и достигает ее при $k = n - 1$ независимо от реальной роли факторов, то необходимо корректировать коэффициент множественной детерминации на потерю степеней свободы вариации:

$$R^2_{корр} = 1 - (1 - R^2)\left(\frac{n-1}{n-k-1}\right). \qquad (8.37)$$

Корректированный коэффициент детерминации всегда ниже, чем некорректированный, причем разность их значений тем меньше, чем меньше факторов входит в уравнение регрессии. Если из числа факторов исключить факторы, слабо связанные с результативным признаком (т. е. с низким значением β_j, например, $\beta_j < 0{,}1$), то некорректированный коэффициент детерминации немного уменьшится (он всегда уменьшается при исключении части факторов), но корректированный коэффициент может даже возрасти за счет уменьшения разности между R^2 и корректированным R^2. Что касается множественного коэффициента корреляции R, то программа «Microstat» рассчитывает его, как корень квадратный из некорректированного R^2, а другие программы, например «Statgraphics», – как корень квадратный из $R^2_{корр}$.

Для случая двух факторов коэффициент множественной детерминации легко вычисляется по рекуррентной формуле из парных коэффициентов детерминации:

$$R^2_{yx_1x_2} = \frac{r^2_{yx_1} + r^2_{yx_2} - 2r_{yx_1}r_{yx_2}r_{x_1x_2}}{1 - r^2_{x_1x_2}}. \qquad (8.38)$$

Используя матрицу парных коэффициентов корреляции (табл. 8.11), получим:

$$R^2_{yx_1x_2} = \frac{0{,}687^2 + (-0{,}355)^2 - 2 \cdot 0{,}687 \cdot (-0{,}355) \cdot (-0{,}044)}{1 - (-0{,}044)^2} = 0{,}5765.$$

Таким образом, за счет вариации факторов x_1 и x_2 объясняется 57,65% общей вариации валового дохода с 1 га сельхозугодий.

Вернемся к табл. 8.12. Дисперсионный анализ системы связей предназначен для оценки того, насколько надежно доказывают исходные данные наличие связи результативного признака со всеми факторами, входящими в уравнение. Для этого сравниваются дис-

персии y – объясненная и остаточная: суммы соответствующих квадратов отклонений, приходящиеся на одну степень свободы вариации. Отношение дисперсии за счет факторов к остаточной дисперсии есть критерий Фишера F; в нашем примере он равен 35,18. Табличное критическое значение для 3 и 12 степеней свободы при вероятности нулевой гипотезы 0,01 составляет 5,95. Следовательно, вероятность нулевой гипотезы много меньше 0,01. Программа «Microstat» дает значение вероятности нулевой гипотезы, т. е. вероятность случайного отклонения от нуля коэффициента детерминации при отсутствии связи в генеральной совокупности; она равна $3{,}17 \cdot 10^{-6}$, т. е. три миллионных! Ясно, что эту ничтожную вероятность можно игнорировать и сделать вывод, что имеющаяся информация надежно свидетельствует о наличии связи.

Кроме показателя общей тесноты связи вариации результативного признака со всеми факторами, входящими в регрессионное уравнение, необходимы и показатели, измеряющие тесноту связи с каждым фактором. К таким показателям относятся коэффициенты раздельной детерминации.

Коэффициентом раздельной детерминации, обозначаемым далее как d_j^2, называется *произведение парного коэффициента корреляции фактора* x_j на его β-коэффициент.

$$d_j^2 = r_{x_j y}\beta_j; \quad \sum_{j=1}^{k} d_j^2 = R^2. \qquad (8.39)$$

Формула (8.39) дает еще один метод вычисления коэффициента множественной детерминации, используемый в некоторых программах для ЭВМ. В нашем примере получаем следующие значения коэффициентов раздельной детерминации:

$$d^2_1 = 0{,}687 \cdot 0{,}352 = 0{,}2418;$$
$$d^2_2 = -0{,}355 \cdot (-0{,}206) = 0{,}0731;$$
$$d^2_3 = 0{,}878 \cdot 0{,}664 = 0{,}5830.$$

Таким образом, за счет вариации x_1 объясняется 24,2% вариации y, за счет вариации x_2 – всего 7,3%; за счет вариации x_3 – более половины – 58,3% вариации уровня дохода. Сумма коэффициентов раздельной детерминации равна некорректированному коэффициенту R^2.

Недостатком коэффициентов раздельной детерминации является их гетерогенный характер: то, что они объединяют коэффициент *парной* корреляции, измеряющий нечистое влияние фактора, с β-коэффициентом, измеряющим условно чистое влияние фактора,

абстрагированное от влияния других факторов, входящих в уравнение связи. Из-за этого могут возникнуть неинтерпретируемые отрицательные величины коэффициентов d_j^2, если знаки парного коэффициента корреляции и β-коэффициента не совпадают при существенной взаимосвязи между факторами. Кроме того, сама идея о том, что совокупное влияние всех факторов равно сумме влияния каждого из них, противоречит системному подходу к исследованию.

Рассмотрим разложение R^2 с учетом системного эффекта. Система факторов – это не простая их сумма, так как система предполагает внутренние связи, взаимодействие составляющих ее элементов. Действие системы не равно сумме воздействий составляющих ее элементов. К последним добавляется «системный эффект» «Emergency». Методом, полностью отвечающим системному подходу, является метод разложения коэффициента множественной детерминации на сумму чистых влияний каждого фактора, выражаемых величинами β^2_1, и показатель влияния системного эффекта факторов η_s.

Так как расчетные значения результативного признака $\hat{y}_i$ можно представить как $a+\sum_{j=1}^{k} b_j x_{ji}$, то вариацию $\tilde{y}_{ji}$ только за счет влияния фактора x_m можно представить при условии, что все остальные факторы, входящие в уравнение, закреплены на своих средних уровнях:

$$\tilde{y}(x_m)_i = a + b_m x_{mi} + \sum_{j=1\ldots m-1,\, m+1,\ldots,\, k} b_j \bar{x}_j. \qquad (8.40)$$

Подставим в (8.40) значение фактора $x_{m\,i} = x_m + \Delta x_{m_i}$:

$$\tilde{y}(x_m)_i = a + b_m(\bar{x}_m + \Delta x_{m_i}) + \sum_{j=1\ldots m-1,\, m+1\ldots k} b_j \bar{x}_j =$$

$$= a + \sum_{j=1}^{k} b_j \bar{x}_j + b_m \Delta x_{m_i} = \bar{y} + b_m \Delta x_{m_i}.$$

Теперь измерим сумму квадратов отклонений y только за счет вариации признака x_m.

$$\sum_{i=1}^{n} (\tilde{y}(x_m)_i - \bar{y})^2 = \sum_{i=1}^{n} (\bar{y} + b_m \Delta x_{m_i} - \bar{y})^2 = \qquad (8.41)$$

$$= \sum_{i=1}^{n} (b_m \Delta x_{m_i})^2 = b_m^2 \sum_{i=1}^{n} (x_{m_i} - \bar{x}_m)^2 = b_m^2 n \sigma^2_{x_m}.$$

Мерой вариации результативного признака за счет изолированного влияния вариации фактора x_m является доля объясняемой этим влиянием вариации y. Соответственно получаем:

$$\frac{b_m^2 n \sigma^2_{x_m}}{\sum_{i=1}^{n} (y_i - \bar{y})^2} = \frac{b_m^2 n \sigma^2_{x_m}}{n \sigma_y^2} = \left(b_m \frac{\sigma_{x_m}}{\sigma_y} \right)^2 = \beta_m^2.$$

Сумма изолированных долей влияния каждого фактора в отдельности на вариацию y есть $\sum_{j=1}^{k} \beta_j^2$, а системный эффект

$$\eta_s = R^2 - \sum_{j=1}^{k} \beta_j^2. \qquad (8.42)$$

Проведем разложение коэффициента множественной детерминации по данным нашего примера:

за счет вариации x_1 : $\beta^2_1 = 0{,}3522^2 = 0{,}1239$, или 12,39%;

за счет вариации x_2 : $\beta^2_2 = (-0{,}206)^2 = 0{,}0424$, или 4,24%;

за счет вариации x_3 : $\beta^2_3 = 0{,}664^2 = 0{,}4409$, или 44,09%.

Суммарное влияние трех факторов составило $\sum_{j=1}^{3} \beta_j^2 = 60{,}72\%$ системный эффект:

$$\eta_s = R^2 - \sum_{i=1}^{3} \beta_i^2 = 0{,}8979 - 0{,}6072 = 0{,}2907, \text{ или } 29{,}07\%.$$

Как видим, роль системного эффекта связей между факторами довольно велика: он на втором месте после влияния третьего фактора.

Системный эффект может, в свою очередь, быть разложен на влияние ковариации каждой пары факторов или на влияние совместной вариации отдельных групп факторов; если число последних велико. Если исследователь все же желает отказаться от выделения системного эффекта, свести коэффициент множественной детерминации к сумме по отдельным факторам, можно разделить величину η_s пропорционально величине β^2_j.

Программы анализа связей на ЭВМ обычно предусматривают вычисление *коэффициентов частной детерминации*. Они приведены выше в последней графе табл. 8.8. *Коэффициент частной де-*

терминации фактора x_m – это доля вариации y, дополнительно объясняемой при включении фактора x_m после остальных факторов в уравнение регрессии, в величине вариации y, не объясненной ранее включенными факторами. Наиболее ясно суть частных коэффициентов детерминации выражается формулой их расчета через коэффициенты множественной детерминации. *Частный коэффициент детерминации* для фактора x_m обозначим как

$$r^2_{yx_m(x_1 \ldots x_{m-1},\, x_{m+1} \ldots x_k)}.$$

Тогда

$$r^2_{yx_m(x_1 \ldots x_{m-1},\, x_{m+1} \ldots x_k)} = \frac{R^2_y - R^2_{yx_1 \ldots x_{m-1},\, x_{m+1} \ldots x_k}}{1 - R^2_{yx_1 \ldots x_{m-1},\, x_{m+1} \ldots x_k}}. \qquad (8.43)$$

Здесь R_y^2 – коэффициент детерминации для уравнения со всеми k факторами. Числитель (8.43) и есть дополнительно объясняемая часть вариации y при включении фактора x_m в уравнение *после всех остальных факторов*. В нашем примере, используя ранее рассчитанную величину $R^2_{yx_1x_2} = 0{,}5765$, при включении в анализ фактора x_3 получаем:

$$r^2_{yx_3(x_1x_2)} = \frac{R^2_y - R^2_{yx_1x_2}}{1 - R^2_{yx_1x_2}} = \frac{0{,}8979 - 0{,}5765}{1 - 0{,}5765} = \frac{0{,}3214}{0{,}4235} = 0{,}7589.$$

Некоторое расхождение в четвертой значащей цифре с табл. 8.8 объясняется округлением промежуточных расчетных показателей.

Следует усвоить, что коэффициенты частной детерминации – это доли от разных величин, поэтому они несравнимы; по этим долям нельзя судить о роли факторов. Их главное практическое значение – определить, имеет ли смысл добавить в уравнение регрессии новый фактор или нет. Если при его включении ранее необъясненная вариация уменьшится на три четверти, как в примере при введении фактора x_3, его включение оправдано; если же коэффициент частной детерминации мал, то дополнительный фактор включать не следует. Сумма частных коэффициентов детерминации смысла не имеет и растет с ростом числа факторов и ростом R^2 без ограничения.

При последовательном вводе факторов в уравнение регрессии объясняемая часть вариации результативного признака возрастает с каждым новым фактором, вводимым в уравнение. При вводе последнего фактора эта часть достигает величины R^2. Доли вариации y, объясняемые

вводом каждого следующего фактора, и называют *коэффициентами последовательной детерминации*. Обозначим их как p_j^2. Для первого фактора этот коэффициент равен коэффициенту парной детерминации первого фактора, для второго – разности между коэффициентом детерминации при двух факторах и парным коэффициентом детерминации первого фактора и так далее. По данным нашего примера имеем:

$$p_1^2 = r^2_{yx_1} = 0{,}687^2 = 0{,}4720;$$

$$p_2^2 = R^2_{yx_1x_2} - r^2_{yx_1} = 0{,}5765 - 0{,}4720 = 0{,}1045;$$

$$p_3^2 = R^2_{yx_1x_2x_3} - R^2_{yx_1x_2} = 0{,}8979 - 0{,}5765 = 0{,}3214;$$

$$\sum_{j=1}^{k} p_j^2 = R^2.$$

Однако крупнейшим недостатком такого способа разложения R^2 является зависимость величин p_j^2 от принятого порядка включения факторов в уравнение регрессии. Первый включаемый фактор «забирает в свою пользу» львиную часть системного эффекта, а на долю последнего фактора остается ничтожная часть. Например, если переставить местами факторы x_1 и x_3, а также вычислить по рекуррентной формуле двухфакторный коэффициент детерминации $R^2_{yx_3x_2} = 0{,}8035$, то получим результаты, отличные от предыдущих:

p_1^2 (для фактора x_3) $= r^2_{yx_3} = 0{,}8782 = 0{,}7709$;

p_2^2 (для фактора x_2) $= R^2_{yx_3x_2} - r^2_{yx_3} = 0{,}8035 - 0{,}7709 = 0{,}0326$;

p_3^2 (для фактора x_1) $= R^2_{yx_1x_2x_3} - R^2_{yx_3x_2} = 0{,}8979 - 0{,}8035 = 0{,}0944$.

Доля фактора x_3 возросла более чем вдвое, а доля фактора x_1 уменьшилась более чем втрое.

8.13. ВЕРОЯТНОСТНЫЕ ОЦЕНКИ ПАРАМЕТРОВ МНОЖЕСТВЕННОЙ РЕГРЕССИИ И КОРРЕЛЯЦИИ

Если показатели многофакторной системы связи используются как оценки генеральных параметров, экстраполируются на другие значения факторов, как при прогнозировании, то значения параметров необходимо сопроводить вероятностными оценками, указать

среднюю ошибку и доверительные границы параметра с заданной вероятностью. Для парной корреляции эта проблема изложена в п. 8.5. В этом параграфе приводятся формулы средних ошибок репрезентативности для специфических параметров многофакторной системы.

Средняя ошибка условно чистого коэффициента регрессии b_p для фактора x_p, обозначаемая m_{bp}, имеет вид:

$$m_{bp} = \frac{s_{y_{ост}}}{s_{x_p}\sqrt{n}}\sqrt{\frac{1}{1-R^2_{x_p, x_1 \ldots x_{p-1}, x_{p+1} \ldots x_k}}}, \qquad (8.44)$$

где $s_{y_{ост}}$ – оценка остаточного (не объясненного факторами) среднего квадратического отклонения результативного признака с учетом степеней свободы вариации:

$$s_{y_{ост}} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2}{n-k-1}},$$

$$s_{x_p} = \sqrt{\frac{\sum_{i=1}^{n}(x_{p_i} - \bar{x}_p)^2}{n-1}},$$

где s_{x_p} – оценка среднего квадратического отклонения признака x_p,

$R^2_{x_p, x_1 \ldots x_{p-1}, x_{p+1} \ldots x_k}$ – коэффициент множественной детерминации для фактора x_p, доля вариации фактора x_p, связанная с вариацией других факторов.

Например, для фактора x_1 имеем:

$$s_{y_{ост}} = 79{,}24.$$
$$s_{x_1} = 34{,}6.$$

$R^2_{x_1 x_2 x_3} = 0{,}2433$ – вычислен по рекуррентной формуле по данным табл. 8.11. Отсюда:

$$m_{b_1} = \frac{79{,}24}{34{,}6\sqrt{16}}\sqrt{\frac{1}{1-0{,}2433}} = 0{,}6582.$$

Отношение величины коэффициента регрессии к его средней ошибке есть t-критерий Стьюдента. В данном случае имеем:

$b_1/m_{b1} = 2{,}26/0{,}6582 = 3{,}43$. Критическое значение t для вероятности нулевой гипотезы 0,01 при 12 степенях свободы равно 3,05. Следовательно, надежно установлено, что генеральное значение коэффициента b_1 не является нулевым, влияние (условно чистое) фактора x_1 на вариацию y существенно.

Доверительные границы коэффициента регрессии b_1 с вероятностью 0,95, для которой значение критерия Стьюдента равно 2,18, составляют $2{,}26 \pm 2{,}18 \cdot 0{,}658$ или от 0,826 до 3,694.

Очень широкие границы объясняются малой численностью единиц совокупности. Из (8.44) следует, что при росте объема совокупности в q раз ошибка коэффициента регрессии, как и ошибка выборочной оценки средней величины, уменьшится в $\sqrt{q}$ раз. При 400 единицах совокупности ошибка была бы меньше в 5 раз.

Если значение критерия t оказывается ниже критического для вероятности нулевой гипотезы 0,05, влияние фактора считается не доказанным надежно, и при работе программ ЭВМ с отсевом несущественных факторов по t-критерию данный фактор автоматически исключается из уравнения регрессии.

Средняя ошибка оценки коэффициента множественной корреляции m_R определяется по формуле

$$m_R = \sqrt{\frac{1-R^2}{n-k-1}}. \qquad (8.45)$$

Оценка существенности и расчет доверительных границ генерального коэффициента корреляции осуществляются так же, как и для коэффициента регрессии. Если значение R близко к единице, необходимо использовать преобразование Фишера, рассмотренное ранее в п. 8.2. Существуют также специальные таблицы критических значений коэффициента корреляции для заданного числа степеней свободы и вероятности нулевой гипотезы (см. приложение, табл. 5).

8.14. КОРРЕЛЯЦИОННО-РЕГРЕССИОННЫЕ МОДЕЛИ (КРМ) И ИХ ПРИМЕНЕНИЕ В АНАЛИЗЕ И ПРОГНОЗЕ

*Корреляционно-регрессионной моделью системы взаимосвязанных признаков является такое уравнение регрессии, которое включает основные факторы, влияющие на вариацию результативного признака, обладает высоким (*не ниже 0,5*) коэффициентом детерминации и коэффициентами регрессии, интерпретируемыми в соответствии с теоретическим знанием о природе связей в изучаемой системе.*

Приведенное определение КРМ включает достаточно строгие условия: далеко не всякое уравнение регрессии можно считать моделью. В частности, полученное выше по 16 хозяйствам уравнение не отвечает последнему требованию из-за противоречащего экономике сельского хозяйства знака при факторе x_2 – доля пашни. Однако в учебных целях используем его как модель.

Теория и практика выработали ряд рекомендаций для построения корреляционно-регрессионной модели.

1. Признаки-факторы должны находиться в причинной связи с результативным признаком (следствием). Поэтому, недопустимо, например, в модель себестоимости y вводить в качестве одного из факторов x_j коэффициент рентабельности, хотя включение такого «фактора» значительно повышает коэффициент детерминации.

2. Признаки-факторы не должны быть составными частями результативного признака или его функциями, о чем уже сказано ранее.

3. Признаки-факторы не должны дублировать друг друга, т. е. быть коллинеарными (с коэффициентом корреляции более 0,8). Так, не следует в модель производительности труда включать и энерговооруженность рабочих, и их фондовооруженность, так как эти факторы тесно связаны друг с другом в большинстве объектов.

4. Не следует включать в модель факторы разных уровней иерархии, т. е. фактор ближайшего порядка и его субфакторы. Например, в моделях себестоимости зерна не следует включать и урожайность зерновых культур, и дозу удобрений под них или затраты на обработку гектара, показатели качества семян, плодородия почвы, т. е. субфакторы самой урожайности.

5. Желательно, чтобы между результативным признаком и факторами соблюдалось единство единицы совокупности, к которой они отнесены. Например, если y – валовой доход предприятия, то и все факторы должны относиться к предприятию: стоимость производственных фондов, уровень специализации, численность работников и т. д. Если же y – средняя зарплата рабочего на предприятии, то факторы должны относиться к рабочему: разряд или классность, стаж работы, возраст, уровень образования, энерговооруженность и т. д. Правило это не категорическое, в модель зарплаты рабочего можно включить, например и уровень специализации предприятия.

6. Математическая форма уравнения регрессии должна соответствовать логике связи факторов с результатом в реальном объекте. Например, такие факторы урожайности, как дозы разных удобрений, уровень плодородия, число прополок и т. п., создают прибавки величины урожайности, мало зависящие друг от друга; урожайность может существовать и без любого из этих факторов. Такому характеру связей отвечает аддитивное уравнение регрессии:

$$\hat{y} = a + b_1x_2 + b_2x_2 + \ldots + b_{k_1}x_{k_1}.$$

Наоборот, если y – объем валовой продукции завода, x_1 – число работников, x_2 – стоимость основных производственных фондов, x_3 – затраты на энергию, топливо, сырье и материалы (комплектующие изделия), то результат без любого из факторов не существует, поэтому большинство экономистов-статистиков строят КРМ (называемую *производственной функцией,* что весьма неудачно терминологически) в мультипликативной форме:

$$y = a \cdot x_1^{b_1} \cdot x_2^{b_2} \cdot x_3^{b_3} \ldots x_k^{b_k}, \qquad (8.46)$$

где коэффициенты b_j соответствуют коэффициентам эластичности факторов при стремлении прироста фактора к бесконечно малой величине: $b_j \to e_j$ при $\Delta x_j \to 0$. Для конечных приростов факторов коэффициенты уравнения (8.44) не равны коэффициентам эластичности, как иногда утверждается в литературе.

Уравнение (8.46) линеаризуется логарифмированием

$$\ln\hat{y} = b_1\ln x_1 + b_2\ln x_2 + b_3\ln x_3 + \ln a,$$

и далее решается как линейная модель.

7. Принцип простоты: предпочтительнее модель с меньшим числом факторов при том же коэффициенте детерминации или даже при несущественно меньшем коэффициенте.

Для анализа степени эффективности управления производством можно использовать сравнение единиц совокупности по показателям отклонений результативного признака от средней величины и от значения, рассчитанного по уравнению регрессии

$$y_i - \bar{y} = (\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i). \qquad (8.47)$$

Первое слагаемое в правой части равенства – это отклонение, которое возникает за счет отличия индивидуальных значений факторов y данной единицы совокупности от их средних значений по совокупности. Его можно назвать *эффектом факторообеспеченности.* Второе слагаемое – отклонение, которое возникает за счет не входящих в модель факторов и отличия индивидуальной эффективности факторов по данной единице совокупности от средней эффективности факторов в совокупности, измеряемой коэффициента-

ми условно-чистой регрессии. Его можно назвать *эффектом фактороотдачи.*

Рассмотрим **пример** расчета и анализа отклонений по ранее построенной модели уровня валового дохода в 16 хозяйствах. Знаки тех и других отклонений 8 раз совпадают и 8 раз не совпадают. Коэффициент корреляции рангов отклонений двух видов составил 0,156. Это означает, что связь вариации факторообеспеченности с вариацией фактороотдачи слабая, несущественная (табл. 8.13).

Обратим внимание на хозяйство № 15 с высокой факторообеспеченностью (15-е место) и самой худшей фактороотдачей (1-й ранг), из-за которой хозяйство недополучило по 122 руб. дохода с 1 га. Напротив, хозяйство № 5 имеет факторообеспеченность ниже средней, но благодаря более эффективному использованию факторов получило на 125 руб. дохода с 1 га больше, чем было бы получено при средней по совокупности эффективности факторов. Более высокая эффективность фактора x_1 (затраты труда) может означать более высокую квалификацию работников, лучшую заинтересованность работников в качестве выполняемой работы. Более высокая эффективность фактора x_3 с точки зрения доходности может состоять в высоком качестве молока (жирности, охлажденности), ввиду которого оно реализовано по более высоким ценам. Коэффициент регрессии при x_2, как уже отмечено, экономически не обоснован.

Таблица 8.13

Анализ факторообеспеченности и фактороотдачи по регрессионной модели уровня валового дохода

Номера хозяйств	Уровни дохода, руб./га		Отклонения уровней, руб./га		Анализ связи отклонений			
	y_i	$\hat{y}_i$	$\hat{y}_i - \bar{y}$	$y_i - \hat{y}_i$	ранг по $\hat{y}_i - \bar{y}$	ранг по $y_i - \hat{y}_i$	разность Δp	Δ_p^2
1	704	733	+128	–29	13	7	6	36
2	293	370	–235	–77	2	3	–1	1
3	346	433	–172	–87	3	2	1	1
4	420	477	–128	–57	5	4	1	1
5	691	566	–39	+125	8	16	–8	64
6	679	591	–14	+88	10	14	–4	16
7	457	474	–131	–17	4	8	–4	16
8	503	548	–57	–45	7	5	2	4
9	314	272	–333	+42	1	12	–11	121
10	803	833	+228	–30	14	6	8	64
11	691	631	+26	+60	11	13	–2	4
12	775	674	+69	+101	12	15	–3	9
13	584	573	–32	+11	9	9	0	0
14	504	493	–112	+11	6	10	–4	16
15	777	899	+294	–122	15	1	14	196
16	1138	1110	+505	+28	16	11	5	25
Итого	605	605	0	0	–	–	0	574

Использование регрессионной модели для прогнозирования состоит в подстановке в уравнение регрессии ожидаемых значений факторных признаков для расчета точечного прогноза результативного признака или (и) его доверительного интервала с заданной вероятностью, как уже сказано в 8.2. Сформулированные там же ограничения прогнозирования по уравнению регрессии сохраняют свое значение и для многофакторных моделей. Кроме того, необходимо соблюдать системность между подставляемыми в модель значениями факторных признаков.

Формулы для расчета средних ошибок оценки положения гиперплоскости регрессии в заданной многомерной точке и для индивидуальной величины результативного признака весьма сложны, требуют применения матричной алгебры и здесь не рассматриваются. Средняя ошибка оценки значения результативного признака, рассчитанная по программе ПЭВМ «Microstat» и приведенная в табл. 8.8, равна 79,2 руб. на 1 га. Это лишь среднее квадратическое отклонение фактических значений дохода от расчетных по уравнению, не учитывающее ошибки положения самой гиперплоскости регрессии при экстраполяции значений факторных признаков. Поэтому ограничимся точечными прогнозами в нескольких вариантах (табл. 8.14).

Для сравнения прогнозов с базисным уровнем средних по совокупности значений признаков введена первая строка таблицы. Краткосрочный прогноз рассчитан на малые изменения факторов за короткое время и снижение трудообеспеченности.

Результат неблагоприятен, доход снижается. Долгосрочный прогноз А – «осторожный», он предполагает весьма умеренный прогресс факторов и соответственно небольшое увеличение дохода. Вариант Б – «оптимистический», рассчитан на существенное изменение факторов. Вариант № 5 построен по способу, которым Агафья Тихоновна в комедии Н. В. Гоголя «Женитьба» мысленно конструирует портрет «идеального жениха»: нос взять от одного претендента, подбородок от другого, рост от третьего, характер от

Таблица 8.14

Прогнозы валового дохода по регрессионной модели

Варианты прогнозов	Возможные значения факторов			Ожидаемое по модели значение результативного признака	
	x_1	x_2	x_3	$\tilde{y}$, руб./га	в % к базисному
1. Фактический, базисный	218,2	54,1	3520	604,9	100
2. Краткосрочный	200	50	3600	594,1	98,2
3. Долгосрочный А	220	55	4000	684,2	113,1
4. « Б	250	55	4500	834,9	138,0
5. «Идеального жениха»	276	35,1	5526	1149,8	190,1

четвертого... вот если бы соединить все нравящиеся ей качества в одном человеке, она бы не колеблясь вышла замуж... Так и при прогнозировании мы объединяем лучшие (с точки зрения модели дохода) наблюдаемые значения факторов: берем значение x_1 от хозяйства № 10, значение x_2 от хозяйства № 2, значение x_3 от хозяйства №16. Все значения факторов уже существуют реально в изучаемой совокупности, они не «ожидаемые», не «взятые с потолка», это хорошо. Однако могут ли эти значения факторов сочетаться в одном предприятии, системны ли эти значения? Решение данного спорного вопроса выходит за рамки статистики, оно требует конкретных знаний об объекте прогнозирования.

8.15. ИЗМЕРЕНИЕ СВЯЗИ НЕКОЛИЧЕСТВЕННЫХ ПРИЗНАКОВ

Корреляционно-регрессионный метод применим только к количественным признакам. Однако задача измерения связи ставится перед статистикой и по отношению к таким признакам, как пол, образование, занятие, семейное состояние человека, отрасль, форма собственности предприятия, т. е. признакам, не имеющим количественного выражения.

Учеными разных стран за последние сто лет разработано несколько методов измерения связей таких признаков. Отметим прежде всего уже рассмотренный ранее коэффициент корреляции рангов Спирмена, применимый и к количественным, и неколичественным, но поддающимся ранжированию признакам. Так, например, можно при помощи одной группы экспертов проранжировать кандидатов на занятие какой-либо должности по степени профессиональной подготовленности, а другую группу экспертов просить проранжировать тех же кандидатов по личностным и этическим качествам, а затем измерить связь между рангами.

Важным частным случаем задачи является измерение связи при альтернативной вариации двух признаков, один из которых имеет характер причины, а другой – следствия. Например, при социологическом обследовании 1000 жителей города были поставлены два вопроса: 1. Считаете ли вы, что ваши доходы позволяют обеспечивать удовлетворение основных потребностей? 2. Удовлетворяет ли вас деятельность мэра города? Можно предположить, что причиной отрицательного ответа на второй вопрос у части населения является неудовлетворенность их потребностей доходами, т.е. имеется связь между ответами на оба вопроса. Для измерения этой связи составляют двухмерное (дихотомическое) распределение ответов 2 × 2, приведенное в табл. 8.15.

Таблица 8.15

Взаимосвязь между ответами на два вопроса социологического обследования

Ответы на 1-й вопрос		Ответы на 2-й вопрос		Итого
		Да, a	Нет, $в$	
Да	A	170	80	$\Sigma A = 250$
Нет	B	230	520	$\Sigma B = 750$
Итого		$\Sigma a = 400$	$\Sigma в = 600$	$N = 1000$

Если бы все, ответившие «да» на 1-й вопрос, отвечали бы «да» на 2-й вопрос, и так же совпадали ответы «нет», то связь была бы предельно тесной, функциональной. Но на самом деле распределение ответов на оба вопроса не совпадает. Большая часть ответивших «да» на 1-й вопрос ответила «да» и на 2-й вопрос, но часть ответила «нет». То же относится к ответившим «да» на 2-й вопрос. Связь есть, но неполная, типа корреляционной, и нужно измерить тесноту этой связи.

К. Пирсон предложил показатель, названный *коэффициентом ассоциации*. В числителе этого относительного показателя разность произведения чисел с одинаковыми ответами на оба вопроса: да-да и нет-нет и произведения чисел с неодинаковыми ответами: да-нет и нет-да. В знаменателе коэффициента ассоциации – корень квадратный из произведения всех четырех *частных* итогов. В буквенных обозначениях по табл. 8.13 имеем:

$$K_{ассоциации} = \frac{Aa \cdot Bb - Ab \cdot Ba}{\sqrt{\sum A \sum B \sum a \sum b}}; \qquad (8.48)$$

$$K_{ассоциации} = \frac{170 \cdot 520 - 80 \cdot 230}{\sqrt{250 \cdot 750 \cdot 400 \cdot 600}} = 0{,}330.$$

Свойства коэффициента ассоциации такие же, как и у коэффициента корреляции: коэффициент ассоциации обращается в нуль, если оба произведения в числителе точно уравновешиваются (что крайне маловероятно); он равен плюс единице, если отсутствуют оба гетерогенных сочетания *Аб* и *Ба*; равен минус единице, если отсутствуют гомогенные сочетания ответов *Аа* и *Бб*.

Другой метод измерения связи по «четырехклеточной таблице» предложен английскими статистиками Эдни Дж. Юлом (1871-1951) и Морисом Дж. Кендэлом (1907). Числитель этого коэффициента, называемого *коэффициентом контингенции*, совпадает с числителем коэффициента ассоциации Пирсона, а в знаменателе – сумма тех же произведений, разность которых стоит в числителе:

$$K_{контингенции} = \frac{Aa \cdot Bb - Ab \cdot Ba}{Aa \cdot Bb + Ab \cdot Ba}; \quad K_{контингенции} = \frac{170 \cdot 520 - 80 \cdot 230}{170 \cdot 520 + 80 \cdot 230} = 0{,}655.$$

Как видим, коэффициент Юла-Кендэла значительно выше, чем коэффициент Пирсона. Крупный недостаток данного коэффициента в том, что уже при равенстве нулю только одного из двух гетерогенных сочетаний – либо *Аб*, либо *Ба* коэффициент Юла – Кендэла обращается в единицу. Можно сказать, что этот показатель очень либерально оценивает тесноту связи, завышает ее.

Наконец, вполне возможно предложить показатель тесноты связи в форме отношения избытка суммы гомогенных сочетаний над их пропорциональной суммой к предельно возможному избытку.

Для этого необходимо вначале вычислить, каковы были бы пропорциональные числа гомогенных сочетаний *Аа* и *Бб*? Пропорциональные числа – это доли от общей численности совокупности «*N*», которые были бы получены при полном отсутствии взаимосвязи группировок по двум признакам (ответам на два вопроса), т. е. числа $(\sum A \cdot \sum a : N)$ и $(\sum B \cdot \sum в : N)$, составляющие по данным табл. 8.13:

$$Aa' = \frac{250 \cdot 400}{1000} = 100 \quad \text{и} \quad Bb' = \frac{750 \cdot 600}{1000} = 450.$$

При отсутствии связи на первой диагонали таблицы в сумме было бы 100 + 450 = 550 единиц совокупности, а на самом деле их 170 + 520 = 690. Избыток, образовавшийся ввиду прямой связи между ответами, составил 690–550 = 140.

Предельно возможный избыток был бы в том случае, если бы не было гетерогенных сочетаний, т. е. *Аб* и *Ба*. Он составляет 140 + 80 + 230 = 450. Сам же показатель тесноты связи – отношение фактического излишка к предельному: 140 : 450 = 0,311. Как видим, этот показатель близок к коэффициенту ассоциации, но обладает чрезвычайно логичной и ясной интерпретацией: связь составляет 0,311 или 31,1%, от предельно возможной функциональной. Этот показатель – аналог не коэффициента корреляции, а коэффициента детерминации. Поэтому правомерно обозначить его как R^2 или η^2. Он имеет вид:

$$\eta^2 = \frac{Aa + bB - |Aa' + Bb'|}{N - (Aa + Bb)}, \qquad (8.49)$$

где

$$Aa' = \frac{\sum A \sum a}{N}; \qquad Bb' = \frac{\sum B \sum b}{N}.$$

Подставляя эти выражения в (8.49), получим:

$$\eta^2 = \frac{Aa + Bb - \frac{\sum A \sum a + \sum B \sum b}{N}}{N - \frac{\sum A \sum a + \sum B \sum b}{N}} = \frac{N(Aa + Bb) - \left(\sum A \sum a + \sum B \sum b\right)}{N^2 - \left(\sum A \sum a + \sum B \sum b\right)}. \qquad (8.50)^*$$

*Эта мера связи предложена М. Юзбашевым в 1986 г. в статье «О новом показателе тесноты связи описательных признаков» // Вестник статистики. – 1986. – № 3. – С. 65 – 66.

При наличии не двух, а более возможных значений каждого из взаимосвязанных признаков также разработаны разные методы измерения тесноты связи.

Рассмотрим некоторые из этих мер на примере изучения влияния религиозной принадлежности на формирование супружеских пар. Воспользуемся данными ФРГ, где такой учет ведется постоянно. Статистический ежегодник Федеративной Республики Германии приводит распределение живорожденных младенцев по религиозной принадлежности отца и матери. При этом выделены 5 групп по религиозной принадлежности граждан: евангелическая (в России их чаще именуют протестантами); 2) римско-католическая; 3) прочие христиане (включая и православных); 4) других религий; 5) неверующие или не указавшие религиозную принадлежность (табл. 8.16).

Таблица 8.16

Распределение новорожденных в ФРГ по религиозной принадлежности отца и матери в 1993 г.

(тыс. чел.)

Религия отца	Религия матери					Итого Σf_i
	Евангелическая	Римско-католическая	Прочие христиане	Другие религии	Неверующие и не указавшие	
Евангелическая	146,1	57,6	1,1	0,5	8,8	214,1
Римско-католическая	57,3	195,9	1,1	0,7	5,2	260,2
Прочие христиане	1,3	1,4	10,5	0,1	0,3	13,6
Другие религии	1,8	2,0	0,1	62,8	1,1	67,8
Неверующие и не указавшие	29,1	16,1	0,7	0,8	77,7	124,4
Итого Σf_i	235,6	273,0	13,5	64,9	93,1	680,1

Источник: Statistisches Jahrbuch für die BRD. – 1995. – С. 74.

В табл. 8.16 представлена «решетка» 5 × 5, и все ее клетки не пусты: встречаются детные браки между лицами любых вероисповеданий. Но при этом наибольшие числа располагаются вдоль «главной диагонали», т. е. явно преобладают случаи, когда и отец и мать

ребенка принадлежат к одной и той же религии. Такое явление ярче всего проявляется среди лиц «других религий» – магометан, иудеев, буддистов, индуистов. Среди них в 92,6% оба родителя относятся к той же вере. Среди лиц евангелического вероисповедания только 68,2% родительских пар относятся к той же вере.

Гипотеза о связи заключается в том, что существует предпочтение к заключению брака между лицами одинаковой религиозной принадлежности.

Соответствующие этой гипотезе частоты f_{ij} расположены вдоль первой диагонали таблицы (из верхнего левого угла в правый нижний).

Пирсон предложил для измерения связи в распределениях со множеством групп по обоим признакам показатель

$$C = \sqrt{\frac{\varphi^2}{1+\varphi^2}}, \tag{8.51}$$

где

$$\varphi^2 = \left(\sum_{i=1}^{k_1} \sum_{j=1}^{k_2} \frac{f_{ij}^2}{f_i f_j} \right) - 1, \tag{8.52}$$

где k_1 и k_2 – число групп по признакам I и II.

$$\varphi^2 = \frac{146{,}1^2}{214{,}1 \cdot 235{,}6} + \frac{57{,}6^2}{214{,}1 \cdot 273} + \ldots + \frac{77{,}7^2}{124{,}4 \cdot 93{,}1} - 1 = 2{,}1364;$$

$$C = \sqrt{\frac{2{,}1364}{3{,}1364}} = 0{,}825.$$

Недостаток коэффициента Пирсона в том, что он не достигает единицы и при полной связи признаков, а лишь стремится к единице при увеличении числа групп. Полезно поэтому провести корректировку коэффициента Пирсона, разделив его величину на предельно возможное значение, которое легко получается при подстановке в (8.52) значений $f_{ij} = f_i = f_j$, что имеет место при полной связи признаков. Имеем

$$C_{max} = \sqrt{\frac{k-1}{k}}$$

и вычисляем значения C_{max} (табл. 8.17).

Таблица 8.17

Предельные значения коэффициента Пирсона

K	2	3	4	5	6	7	8	9	10
C_{max}	0,707	0,816	0,866	0,894	0,913	0,926	0,935	0,943	0,949

В рассмотренной табл. 8.16 при 5 группах корректированный показатель связи Пирсона составит:

$$C_{корр} = 0{,}8253 : 0{,}894 = 0{,}923.$$

Русским статистиком А. А. Чупровым (1874 – 1926) предложен другой показатель связи T.

$$T = \sqrt{\frac{\varphi^2}{\sqrt{(k_1 - 1)(k_2 - 1)}}}. \qquad (8.53)$$

Так как при полной связи $\varphi^2 = k - 1$, то при равенстве $k_1 = k_2$ показатель Чупрова будет равен единице. Для случая, когда $k_1 \neq k_2$, шведский математик и статистик Г. Крамер в 1946 г. предложил дополнение:

$$T_{kp} = \sqrt{\frac{\varphi^2}{k_{\min} - 1}}. \qquad (8.54)$$

С одной стороны, действительно, любая диагональ в таблице $k_1 \cdot k_2$ при $k_1 \neq k_2$ может содержать только $k_{\min}$ число клеток. Но, с другой стороны, вовсе не обязательно связывать гипотезу о полной связи признаков в случае $k_1 \neq k_2$ только с частотами на какой-либо диагонали. Случай этот сложен и подлежит рассмотрению только по каждой конкретной задаче. Далее поэтому рассматривается только случай, когда $k_1 = k_2$. По данным табл. 8.16 показатель Чупрова составил:

$$T = \sqrt{\frac{2{,}1364}{\sqrt{(5-1)\cdot(5-1)}}} = 0{,}7308.$$

Как правило, показатель Чупрова гораздо строже оценивает тесноту связи, чем показатель Пирсона, слишком быстро приближающийся к единице.

Модифицируя для любого числа групп ранее предложенный для двух групп способ и формулы (8.49) и (8.50) с учетом принятых в табл. 8.16 обозначений частот, получаем:

$$\eta^2 = \frac{\sum_{i=1}^{k} f_{ij}(i=j) - \sum_{i=1}^{k} f'_{ij}(i=j)}{\sum_{i=1}^{k}\sum_{j=1}^{k} f_{ij} - \sum f'_{ij}(i=j)}, \qquad (8.55)$$

где

$$f'_{ij}(i=j) = \frac{f_i \cdot f_j(i=j)}{\sum_{i=1}^{k}\sum_{j=1}^{k} f_{ij}},$$

т. е. частоты в клетках первой диагонали при отсутствии связи признаков. Подставив значения f'_{ij} в (8.55), получаем формулу, аналогичную (8.50):

$$\eta^2 = \frac{\left(\sum_{i=1}^{k}\sum_{j=1}^{k} f_{ij}\right)\sum_{i=1}^{k} f_{ij}(i=j) - \sum_{i=1}^{k} f_i \cdot f_j(i=j)}{\left(\sum_{i=1}^{k}\sum_{j=1}^{k} f_{ij}\right)^2 - \sum_{i=1}^{k} f_i f_j(i=j)}, \qquad (8.56)$$

где

$$f'_{ij}(i=j) = \frac{\sum f_i \cdot \sum f_j(i=j)}{\sum_{i=1}^{k_1}\sum_{j=1}^{k_2} f_{ij}}.$$

т. е. $f'_{ij}(i=j)$ – это такие частоты в клетках первой диагонали таблицы, которые были бы при отсутствии связи: пропорциональные частоты.

По данным табл. 8.16 имеем:

$$\sum_{i=1}^{k_1}\sum_{j=1}^{k_2} f'_{ij}(i=j) =$$

$$= \frac{(214{,}1\cdot 235{,}6)+(260{,}2\cdot 273)+(13{,}6\cdot 13{,}5)+(67{,}8\cdot 64{,}9)+(124{,}4\cdot 93{,}1)}{680{,}1} =$$

$$= 202{,}17.$$

$$\sum_{i=1}^{k_1}\sum_{j=1}^{k_2} f_{ij}(i=j) = 146{,}1+195{,}9+10{,}5+62{,}8+77{,}7 = 493{,}0.$$

$$\eta^2 = \frac{493-202{,}17}{680{,}1-202{,}17} = 0{,}6085; \quad \eta = 0{,}780.$$

Таким образом, за счет предпочтения браков между лицами одинаковых религий на главную диагональ «собралось» 60,85% возможных родительских пар сверх равномерного распределения: связь составила 60,85% предельно тесной. Итак, все способы измерения показали, что влияние религии на формирование супружеских пар в ФРГ в 1993 году было значительное.

Если кроме количественных факторов при многофакторном регрессионном анализе включается и неколичественный, то применяют следующую методику: наличие неколичественного фактора у единиц совокупности обозначают единицей, его отсутствие –нулем. Если таких факторов, или градаций неколичественного фактора несколько, в уравнение регрессии вводятся несколько так называемых «фиктивных переменных», принимающих значения либо единицы, либо нуля. Например, пусть имеется три количественных фактора урожайности (x_1, x_2, x_3) и три природных зоны. В ЭВМ вводятся переменные в порядке их принадлежности к той или иной зоне (табл. 8.18).

Линейное уравнение регрессии будет иметь вид:

$$\hat{y} = a + b_1x_1 + b_2x_2 + b_3x_3 + b_4u_1 + b_5u_2. \qquad (8.57)$$

Величина коэффициента b_4 означает, что все единицы II зоны при тех же значениях количественных факторов, как и единицы I зоны, будут в среднем иметь значение $\hat{y}$ на b_4 больше (или меньше, если $b_4 < 0$), чем единицы совокупности I зоны. Величина b_5 означает то же для единиц совокупности III зоны. Иначе говоря, мы получаем сразу три зональных регрессионных модели:

I зона: $\hat{y} = a + b_1x_1 + b_2x_2 + b_3x_3 \quad (u_1 = 0;\ u_2 = 0)$;

II зона: $\hat{y} = a + b_1x_1 + b_2x_2 + b_3x_3 + b_4u_1 \quad (u_2 = 0)$;

III зона: $\hat{y} = a + b_1x_1 + b_2x_2 + b_3x_3 + b_5u_2 \quad (u_1 = 0)$.

Число фиктивных переменных должно быть на единицу меньше числа градаций качественного (неколичественного) фактора. С помощью данного приема можно измерять влияние уровня образования, местожительства, типа жилища и других социальных или природных не измеряемых количественно факторов, изолируя их от влияния количественных факторов.

Таблица 8.18

Зоны	Результативный признак y_i	Количественные факторы			Фиктивные переменные	
		x_1	x_2	x_3	u_1	u_2
I	y_1	x_{11}	x_{21}	x_{31}	0	0
	y_2	x_{12}	x_{22}	x_{32}	0	0
	.	.	.	.	.	.
	.	.	.	.	.	.
	.	.	.	.	.	.
	y_{n_1}	x_{n_1}	x_{2n_1}	x_{3n_1}	0	0
II	y_{n_1+1}	x_{1n_1+1}	x_{2n_1+1}	x_{3n_1+1}	1	0
	.	.	.	.	.	.
	.	.	.	.	.	.
	.	.	.	.	.	.
	$y_{n_1+n_2}$	$x_{1n_1+n_2}$	$x_{2n_1+n_2}$	$x_{3n_1+n_2}$	1	0
III	$y_{n_1+n_2+1}$	$x_{1n_1+n_2+1}$	$x_{2n_1+n_2+1}$	$x_{3n_1+n_2+1}$	0	1
	.	.	.	.	.	.
	.	.	.	.	.	.
	.	.	.	.	.	.
	$y_{n_1+n_2+n_3}$	$x_{1n_1+n_2+n_3}$	$x_{2n_1+n_2+n_3}$	$x_{3n_1+n_2+n_3}$	0	1

Рекомендуемая литература к главе 8

1. *Аптон Г.* Анализ таблиц сопряженности / Пер. с англ. – М.: Финансы и статистика, 1982.
2. *Елисеева И. И.* Статистические методы измерения связей. – Л.: Изд-во Ленинградского ун-та, 1982.
3. *Елисеева И. И., Рукавишников В. О.* Логика прикладного статистического анализа. – М.: Финансы и статистика, 1982.
4. *Крастинь О. П.* Разработка и интерпретация моделей корреляционных связей в экономике. – Рига: Занатне, 1983.
5. *Кулаичев А. П.* Методы и средства анализа данных в среде Windows. Stadia 6.0 – М.: НПО Информатика и компьютеры, 1996.
6. Статистическое моделирование и прогнозирование: Учебное пособие / Под ред. А. Г. Гранберга. – М.: Финансы и статистика, 1990.
7. *Ферстер Э., Ренц Б.* Методы корреляционного и регрессионного анализа. Руководство для экономистов / Пер. с нем. – М.: Финансы и статистика, 1983.
8. *Шураков В. В.* и др. Автоматизированное рабочее место для статистической обработки данных. – М.: Финансы и статистика, 1990.

Глава 9

СТАТИСТИЧЕСКОЕ ИЗУЧЕНИЕ ДИНАМИКИ

Одно из основных положений научной методологии – необходимость изучать все явления в развитии, во времени. Это относится и к статистике: она должна дать характеристику изменений статистических показателей во времени. Как изменяются год за годом валовой национальный продукт и национальный доход страны? Как возрастает или снижается уровень оплаты труда? Велики ли колебания урожайности зерновых культур и существует ли тенденция ее роста? На все аналогичные вопросы ответ может дать только специальная система статистических методов, предназначенная для изучения развития, изменений во времени или, как принято в статистике говорить, изучения динамики.

9.1. СОСТАВЛЯЮЩИЕ ЭЛЕМЕНТЫ ДИНАМИКИ: ОСНОВНАЯ ТЕНДЕНЦИЯ И КОЛЕБАНИЯ

Рассмотрим данные табл. 9.1. Условимся, что относящиеся к отдельным годам значения урожайности картофеля, будем называть *уровнями,* а всю их последовательность с 1986 по 1996 г. – *рядом динамики* (динамическим рядом, временным рядом, английский термин «Time series»).

Таблица 9.1

Динамика урожайности картофеля в хозяйстве

Годы / Показатели	1986	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996
Урожайность, ц с 1 га	149	145	168	146	177	176	190	186	176	211	170

Ряд динамики состоит из двух строк или столбцов: промежутков или моментов времени, к которым относятся уровни, и самих уровней признака (показателя). Ряд, в котором время задано в виде промежутков – лет, месяцев, суток, называется *интервальным динамическим рядом.* В табл. 9.1 приведен такой ряд. Ряд, в котором время задано в виде конкретных дат (моментов времени), называется *моментным динамическим рядом.* Например, ряд численности населения по оценке на 1 января каждого года.

Вернемся к табл. 9.1. Сравнивая уровни разных лет, мы замечаем, что в целом урожайность возрастает. Однако нередко уровень урожайности следующего года оказывается ниже предыдущего. Иногда рост по сравнению с предыдущим годом велик, как в 1990 г., а иногда мал. Следовательно, рост урожайности наблюдается лишь в среднем, как *тенденция.* В отдельные же годы уровни испытывают *колебания,* отклоняясь от основной тенденции. Эти колебания урожайности связаны в основном с различием метеорологических условий в разные годы.

Если рассматривать динамические ряды месячных уровней производства мяса или молока, ряды объема продажи разных видов одежды и обуви, ряды заболеваемости населения, выявятся регулярно повторяющиеся из года в год *сезонные* колебания уровней. В силу солнечно-земных связей частота полярных сияний, интенсивность гроз, те же изменения урожайности отдельных сельхозкультур и ряд других процессов имеют циклическую 10 – 11- летнюю колеблемость. Колебания числа рождений, связанные с потерями в войне, повторяются с угасающей амплитудой через поколение, т.е. через 20 - 25 лет.

Тенденция динамики связана с действием *долговременно* существующих причин и условий развития, хотя, конечно, после какого-то периода эти причины и условия тоже могут измениться и породить уже другую тенденцию развития изучаемого объекта. Колебания же, напротив, связаны с действием *краткосрочных* или циклических факторов, влияющих на отдельные уровни динамического ряда, и отклоняющих уровни от тенденции то в одном, то в другом направлении. Например, тенденция динамики урожайности связана с прогрессом агротехники, с укреплением экономики данной совокупности хозяйств, совершенствованием организации производства. Колеблемость урожайности вызвана чередованием благоприятных по погоде и неблагоприятных лет, циклами солнечной активности, колебаниями в развитии вредных насекомых и болезней растений.

При статистическом изучении динамики необходимо четко разделить ее два основных элемента – тенденцию и колеблемость, чтобы дать каждому из них количественную характеристику с помощью специальных показателей. Смешение тенденции и колеблемости ведет к неверным выводам о динамике. Если из табл. 9.1 произвольно взять данные за отдельные годы и сравнить их друг с другом, можно получить «выводы», прямо противоположные истине. Например, если сравнить урожайность в 1995 г. с урожайностью в 1987 г., то получим, что за 8 лет она возросла на 66 ц с 1 га, т.е. более чем по 8 ц с 1 га за год. Если же урожайность в 1996 г. сравнить с ее уровнем в 1988 г., то получим, что за 8 лет, из которых 7 лет те же, что и в предыдущем сравнении, урожайность возросла всего лишь на 2 ц с1 га.

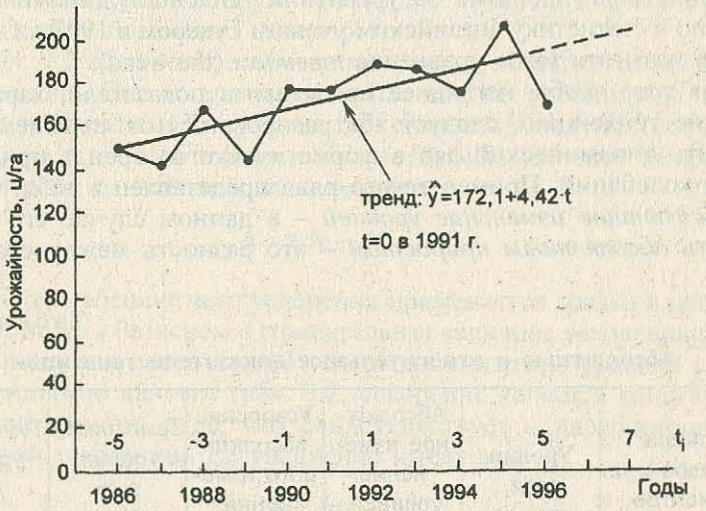


Рис. 9.1. Динамика урожайности картофеля

Тенденцию и колебания наглядно показывает график (рис. 9.1). По оси абсцисс всегда отражается время, по оси ординат – уровни. По обеим осям строго соблюдается масштаб, иначе характер динамики будет искажен.

На рис 9.1 хорошо заметно, что рост урожайности в 1986 – 1996 гг. характеризовался линейной тенденцией, а колеблемость была хаотической, без явной цикличности. О линии тренда и ее уравнении будет сказано далее, в п. 9.5 и 9.6.

9.2. ПОКАЗАТЕЛИ, ХАРАКТЕРИЗУЮЩИЕ ТЕНДЕНЦИЮ ДИНАМИКИ

Чтобы построить систему показателей, характеризующих тенденцию динамики, нужно ответить на вопрос: какие черты, свойства этой тенденции нужно измерить и выразить в статистических показателях? Очевидно, нас интересует величина изменений уровня как в абсолютном, так и в относительном выражении (на какую долю, процент уровня, принятого за базу, произошло изменение?). Далее нас интересует: является ли изменение равномерным или неравномерным, ускоренным (замедленным?). Наконец, нас интересует выражение тенденции в форме некоторого достаточно простого уравнения, наилучшим образом аппроксимирующего фактическую тенденцию динамики. Понятие об уравнении тенденции динамики было введено в статистику английским ученым Гукером в 1902 г. Он предложил называть такое уравнение *трендом* (the trend).

Для того чтобы нагляднее представить показатели, характеризующие тенденцию, следует абстрагироваться от колеблемости и выявить динамический ряд в форме «чистого» тренда при отсутствии колебаний. Пример такого ряда представлен в табл. 9.2.

Абсолютное изменение уровней – в данном случае его можно назвать *абсолютным приростом* – это разность между сравниваемым уровнем и уровнем более раннего периода, принятым за базу сравнения. Если эта база непосредственно предыдущий уровень, показатель называют *цепным*, если за базу взят, например, начальный уровень, показатель называют *базисным*. Формулы абсолютного изменения уровня:

Таблица 9.2

Абсолютные и относительные показатели тенденции

Номера периодов или моментов времени	Уровни ряда, тыс. т	Абсолютное изменение уровней, тыс. т в год	Ускорение абсолютного изменения, тыс. т за год	Темп роста уровня, в % к предыдущему	Темп роста уровня, в % к начальному
0 (начальный)	100	–	–	–	100
1	112	12	–	112	112
2	128	16	4	114,3	128
3	148	20	4	115,6	148
4	172	24	4	116,2	172
5	200	28	4	116,3	200
6	232	32	4	116,0	232

$$\text{цепное: } \Delta_i = y_i - y_{i-1}$$
$$\text{базисное: } \Delta_{0i} = y_i - y_0. \qquad (9.1)$$

Если абсолютное изменение отрицательно, его следует называть абсолютным сокращением. Абсолютное изменение имеет ту же единицу измерения, что и уровни ряда с добавлением единицы времени, за которую определено изменение: 22 тысячи тонн в год (или 1,83 тыс. т в месяц, или 110 тыс. т в пятилетие). Без указания единицы времени, за которую произошло измерение, абсолютный прирост нельзя правильно интерпретировать.

В табл. 9.2 абсолютное изменение уровня не является константой тенденции. Оно со временем возрастает, т.е. уровни ряда изменяются с ускорением. *Ускорение* – это разность между абсолютным изменением за данный период и абсолютным изменением за предыдущий период одинаковой длительности:

$$\Delta'_i = \Delta_i - \Delta_{i-1}. \qquad (9.2)$$

Показатель абсолютного ускорения применяется только в цепном варианте, но не в базисном. Отрицательная величина ускорения говорит о замедлении роста или об ускорении снижения уровней ряда.

Как видно по данным табл. 9.2, ускорение является константой тенденции данного ряда, что свидетельствует о параболической форме этой тенденции. Ее уравнение имеет вид:

$$\hat{y}_i = a + bt_i + ct_i^2, \qquad (9.3)$$

где t_i – номер периода;
a – уровень ряда в начальный период, при $t = 0$;
b – средний по ряду абсолютный прирост;
c – половина ускорения.

По данным табл. 9.2 имеем:

$$y_i = 100 + 10t_i + 2t_i^2.$$

Показатель ускорения абсолютного изменения уровней выражается в единицах измерения уровня, деленных на квадрат длины

периода. В нашем случае ускорение составило 4 тыс. т в год за год или 4 тыс. т год$^{-2}$. Смысл показателя следующий: объем производства (или добычи угля, руды) имел абсолютный прирост, возрастающий на 4 тыс. т в год ежегодно.

Усвоить рассмотренные показатели поможет следующая аналогия с механическим движением: уровень – это аналог пройденного пути, причем начало его отсчета не в нулевой точке. Абсолютный прирост – аналог скорости движения тела, а ускорение абсолютного прироста – аналог ускорения движения. Пройденный путь, считая и тот, который уже был пройден до начала отсчета времени в данной задаче, равен:

$$s = s_0 + v_0 t + \frac{at^2}{2},$$

где s_0 – путь, пройденный до начала отсчета времени;
v_0 – начальная скорость;
a – ускорение;
t – время, прошедшее от начала его отсчета в задаче.

Сравнивая с формулой (9.3), видим, что s_0 – аналог свободного члена a; v_0 – аналог абсолютного изменения $в$; $a/2$ – аналог ускорения прироста С.

Как показано в гл. 3, система показателей должна содержать не только абсолютные, но и относительные статистические показатели.

Относительные показатели динамики необходимы для сравнения развития разных объектов, особенно если их абсолютные характеристики различны. Предположим, другое предприятие увеличивало производство аналогичной продукции с тенденцией, выраженной уравнением тренда: $y_t = 20 + 4t + 0{,}5t_i^2$. И абсолютный прирост, и ускорение роста объема продукции во втором предприятии гораздо меньше, чем в первом. Но можно ли ограничиться этими показателями и сделать вывод, что развитие второго предприятия более медленное, чем первого? Меньший уровень еще не есть меньший темп развития, и это покажет относительная характеристика тенденции динамики – темп роста.

Темп роста – это отношение сравниваемого уровня (более позднего) к уровню, принятому за базу сравнения (более раннему). Темп роста исчисляется в *цепном* варианте – к уровню предыдущего года и в *базисном* варианте – к одному и тому же, обычно начальному уровню (см. формулы (9.4). Он говорит о том, сколько процентов составляет сравниваемый уровень по отношению к

уровню, принятому за базу, или во сколько раз сравниваемый уровень больше уровня, принятого за базу. При этом если уровни снижаются со временем, то сказать, что последующий уровень «больше в 0,33 раза», или составляет 33,3% базового уровня, это, разумеется, означает, что уровень уменьшился в 3 раза. Но сказать что «уровень меньше в 0,33 раза», это неверно. Темп изменения в разах всегда говорит о том, во сколько раз сравниваемый уровень больше.

Теперь можно сказать, что относительная характеристика роста объема продукции на первом предприятии в среднем за год близка к 115% (рост приблизительно на 15% за год), и за шесть лет продукция увеличилась в 2,32 раза, а на втором предприятии, вычислив также шесть уровней параболического тренда, читатель убедится, что в среднем за год объем продукции возрастал примерно на 20%, а за шесть лет объем ее возрос в 3,1 раза. Следовательно, в относительном выражении объем продукции на втором предприятии развивался, возрастал быстрее. Только в сочетании абсолютных и относительных характеристик динамики можно правильно отразить процесс развития совокупности (объекта).

Рассмотрим связь абсолютных и относительных показателей динамики. Обозначим темп изменения через k. Тогда имеем:

$$\text{цепной темп роста } K_{i/i-1} = \frac{y_i}{y_{i-1}};$$
$$\text{базисный темп роста } K_{i/0} = \frac{y_i}{y_0}. \qquad (9.4)$$

Если сравниваемый уровень y выразить через уровень предыдущего года плюс прирост или через уровень базисного года плюс базисное абсолютное изменение, получим:

$$k_{i/i-1} = \frac{y_{i-1} + \Delta_i}{y_{i-1}} = 1 + \frac{\Delta_i}{y_{i-1}} \text{ или } 100\% + \frac{\Delta_i}{y_{i-1}} \cdot 100;$$
$$k_{i/0} = \frac{y_0 + \Delta_{0i}}{y_0} = 1 + \frac{\Delta_{0i}}{y_0} \text{ или } 100\% + \frac{\Delta_{0i}}{y_0} \cdot 100. \qquad (9.5)$$

Величину $\Delta_i : y_{i1}$ или $\Delta_{0i} : y_0$, т. е. отношение абсолютного изменения к предыдущему или к базисному уровню часто называют *относительным приростом* (относительным изменением)

или же *темпом прироста*. Он равен k-1 или k-100%. Темп прироста (относительное изменение) может иметь как положительные значения, так и отрицательные. Наоборот, темп изменения – величина всегда положительная. Если уровень ряда динамики принимает положительные и отрицательные значения, например финансовый результат от реализации продукции предприятием может быть прибылью (+), а может быть убытком (–), тогда темп изменения и темп прироста применять нельзя. В этом случае такие показатели теряют смысл и не имеют экономической интерпретации. Сохраняют смысл только абсолютные показатели динамики.

Рассмотрим соотношения между цепными и базисными показателями на примере данных табл. 9.2:

1) сумма цепных абсолютных изменений равна базисному абсолютному изменению

$$\sum \Delta_{i(\text{цепн})} = \Delta_{i(\text{баз})}, \qquad (9.6)$$

По данным табл. 9.2 получим:

$$12 + 16 + 20 + 24 + 28 + 32 = 232 - 100 = 132;$$

2) произведение цепных темпов изменения равно базисному темпу изменения

$$\prod_{(i)} k_{i(\text{цепн})} = k_{i(\text{баз})}, \qquad (9.7)$$

По данным табл. 9.2 получим:

$$1{,}12 \cdot 1{,}143 \cdot 1{,}156 \cdot 1{,}162 \cdot 1{,}163 \cdot 1{,}16 = 2{,}32.$$

Сумма цепных темпов прироста не равна базисному темпу прироста.

$$12 + 13{,}3 + 15{,}6 + 16{,}2 + 16{,}3 + 16 \neq 132 \text{ (в процентах)}.$$

Значения цепных темпов прироста, рассчитанных каждый к своей базе, различаются не только числом процентов, но и величиной абсолютного изменения, составляющей каждый процент. Поэтому складывать или вычитать цепные темпы прироста нельзя. *Абсолютное значение 1% прироста равно сотой части предыдущего уровня, или базисного уровня.*

9.3. ОСОБЕННОСТИ ПОКАЗАТЕЛЕЙ ДИНАМИКИ ДЛЯ РЯДОВ, СОСТОЯЩИХ ИЗ ОТНОСИТЕЛЬНЫХ УРОВНЕЙ

Уровнями динамического ряда могут быть не только абсолютные показатели. Ряды динамики могут отражать развитие структуры совокупности, изменение со временем вариации признака в совокупности, взаимосвязи между признаками, соотношения значений признака для разных объектов. В этих случаях уровни динамического ряда сами являются относительными показателями, нередко выражаются в процентах. Следовательно, абсолютные изменения (и ускорения) тоже окажутся относительными величинами, могут быть выражены в процентах. В процентах, разумеется, будут выражены темпы изменения и относительные приросты. Все это создает нередко путаницу в интерпретации и использовании показателей динамики в печати и даже в специальной экономической литературе.

Рассмотрим **пример.** В США с конца XIX в. для группы ведущих акционерных компаний исчисляется так называемый *индекс Доу Джонса – арифметическая средняя величина котировок акций на фондовых биржах.* Этот показатель характеризует хозяйственную конъюнктуру: если индекс Доу Джонса повышается, т.е. растет относительная цена акций, значит, вкладчики капитала рассчитывают получить по акциям бóльший дивиденд (распределяемая часть прибыли). Это говорит о росте деловой активности. Падение индекса Доу Джонса говорит о снижении деловой активности в стране. Величина этого показателя есть отношение в процентах цены акций на бирже к их номиналу (первоначальной цене при выпуске акций). Это отношение зависит не только от колебаний деловой активности, но имеет также общую тенденцию роста ввиду инфляции – падения покупательной силы доллара США. С начала века этот рост значителен, поэтому в наше время индекс Доу Джонса составляет более 2000% (акция, когда – то выпущенная на сумму 100 долл., теперь стоит более 2000 современных долларов).

Биржевые новости за 5.05.1990 г. сообщают: индекс Доу Джонса на 3.05.1990 г. составил 2689,64% в сравнении с 2759,55% на 29.04.1990 г. Если вычислить показатель абсолютного изменения индекса, т.е. 2689,64% – 2759,55% = – 69,91%, и сказать, что индекс Доу Джонса за неделю понизился почти на 70%, создается ложное впечатление о чудовищном крахе на биржах США, потому что сни-

жение на 70% воспринимается как темп изменения – будто от прежней цены акций осталось только 30%.

На самом деле снижение показателей с 2760 до 2690% никакой катастрофой экономике США не грозит: это обычная на рынке ценных бумаг колеблемость курсов. «Биржевые ведомости» далее сообщали, что индекс Доу Джонса на 7.06.1990 г. достиг 2911,6%, т.е. с 5.05.1990 г. возрос на 222 единицы, которые во избежании путаницы принято именовать «пунктами». В первом рассмотренном случае индекс снизился на 70 пунктов, во втором – возрос на 222 пункта, а не процента. В процентах рост составил: 222 : 2690 = 8,25% – это и есть темп прироста курса акций.

Аналогичные термины должны применяться и к динамике показателей структуры. Например, общее производство электроэнергии в Российской Федерации в 1980 г. составляло 805 млрд кВ·ч, в том числе на атомных электростанциях 54 млрд кВт·ч, т. е. их доля была равна 6,7%. В 1995 г. общее производство электроэнергии составило 860 млрд кВт·ч, в том числе на АЭС 99,5 млрд кВт·ч, или 11,6%. Доля АЭС возросла, за 15 лет на 11,6 – 6,7 = 4,9 пункта. А темп роста доли АЭС составил 11,6% : 6,7% = 1,73. Доля АЭС возросла на 73%.

Показатели динамики долей имеют еще одну особенность, вытекающую из того, что сумма всех долей в любой период времени равна единице, или 100%. Изменение, произошедшее с одной из долей, поэтому, неизбежно меняет и доли всех других частей целого, если даже по абсолютной величине эти части не изменились. Казалось бы, это положение самоочевидно, однако нередко в печати встречаются рассуждения о том, что увеличение доли пшеницы и ячменя среди зерновых культур – это хорошо, но вот плохо, что уменьшились доли ржи, овса и гречихи. Как будто все доли сразу могут увеличиться!

Если признак варьирует альтернативно, то увеличение доли одной группы равно уменьшению доли другой группы в пунктах, но темпы изменения долей в процентах при этом могут сильно различаться. Темп больше у той доли, которая в базисном периоде была меньше – темп прироста (изменения) понимается по абсолютной величине, по модулю. Например, в 1992 г. оплата труда составила 69,9% всех денежных доходов населения России, а прочие доходы 30,1%. В 1995 г. оплата труда составила только 39,3% всех денежных доходов населения, а доля прочих доходов возросла до 60,7%. Темп прироста доли прочих доходов составил 201,7%, т. е. их доля

возросла на 101,7%. Доля же оплаты труда сократилась в относительном выражении на 43,8% $\left[\frac{39,3}{69,9} - 1 = 0,562 - 1 = -0,438\right]$[1].

В общем виде темп роста одной из альтернативных долей зависит от темпа роста другой доли и величины этой доли следующим образом:

$$k_2 = \frac{1 - k_1 x_0}{1 - x_0}, \tag{9.8}$$

где x_0 – доля в базисном периоде одного из альтернативных значений признака;
k_1 – темп роста этой доли;
k_2 – темп изменения доли второго альтернативного значения признака.

Абсолютное изменение долей в пунктах зависит от величины доли и темпа роста таким образом:

$$\Delta_1 = -\Delta_2 = x_0 (k_1 - 1) \cdot 100. \tag{9.9}$$

При наличии в совокупности не двух, а более групп абсолютное изменение каждой из долей в пунктах зависит от доли этой группы в базисный период и от соотношения темпа роста абсолютной величины объемного признака этой группы со средним темпом роста объемного признака во всей совокупности. Доля i-й группы в сравниваемый (текущий) период определяется как

$$d_{i1} = d_{i0} \frac{k_i}{\bar{k}} = \frac{d_{i0} k_i}{\sum_{i=1}^{m} d_{i0} k_i}, \tag{9.10}$$

где d_{0i}, d_{i1} – доли i-й группы в базисный и текущий периоды;
k_i – темп роста объемного признака в i-й группе;
$\bar{k}$ – средний темп роста;
m – число групп.

Рассмотрим распределение занятого населения России по секторам экономики и его изменение (табл. 9.3).

---

[1] Россия в цифрах. 1996: Статистический сборник / Госкомстат России. – М.: Финансы и статистика, 1996. – С. 53.

Таблица 9.3

Занятое население России по секторам экономики в организациях по формам собственности[1]

Организации с формой собственности	Доля в 1992 г., %	Темп изменения численности в 1995 г. к 1992 году, %
Государственная и муниципальная	68,9	50,7
Частная	18,3	184,8
Общественная	0,8	83,3
Совместная и смешанная	12,0	198,8
Всего занятых	100,0	93,3

[1] *Источник:* Россия в цифрах. 1996: Статистический сборник. Госкомстат России. – М.: Финансы и статистика, 1996. – С. 34.

Согласно формуле (9.10) доля работающих в организациях с государственной и муниципальной формами собственности в 1995 г. составит:

$$0{,}689 \cdot \frac{0{,}507}{0{,}689 \cdot 0{,}507 + 0{,}183 \cdot 1{,}848 + 0{,}8 \cdot 0{,}833 + 12 \cdot 1{,}988} =$$

$$= \frac{0{,}3493}{0{,}9327} = 0{,}3745 \text{ или } 37{,}45\%.$$

Доля работающих в частном секторе: $0{,}183 \cdot \frac{1{,}848}{0{,}9327} = 0{,}3626$ или 36,26%.

Доля работающих в общественных организациях: $0{,}008 \cdot \frac{0{,}833}{0{,}9327} =$ = 0,007 или 0,7%.

Доля работающих в совместных и предприятиях смешанной формы собственности:

$$\frac{0{,}12 \cdot 1{,}988}{0{,}9327} = 0{,}2558 \text{ или } 25{,}58\%.$$

Знаменатели обеих дробей – 0,9327 – это средний (общий) темп изменения численности всех занятых.

Особенностью показателей динамики относительных величин интенсивности является то, что темпы роста и темпы прироста (или сокращения) прямого и обратного показателей не совпадают.

Пример. Трудоемкость производственной операции на старом станке составляла 10 мин., а производительность труда – 48 операций за смену. После замены станка на новый трудоемкость операции снизилась в 5 раз (до 2 мин.), а производительность возросла в те же 5 раз – до 240 операций за смену. Относительное изменение трудоемкости составило (2 – 10) : 10 = –0,8, т. е. трудоемкость снизилась на 80%. Относительное изменение производительности труда составило (240 – 48) : 48 = 4 или 400%, т. е. производительность труда возросла на 400%. Причина состоит в том, что пределом, к которому стремятся по мере прогресса показатели ресурсоотдачи, является бесконечность, а пределом, к которому стремятся обратные им показатели ресурсоемкости, является нуль. Понимание разного поведения показателей динамики прямых и обратных мер эффективности очень важно для экономиста и статистика.

По мере приближения относительного показателя к пределу одно и то же абсолютное изменение в пунктах приобретает иное качественное содержание. Например, если показатель тесноты связи – коэффициент детерминации – возрос с 40 до 65% (на 25 пунктов), то система факторов в регрессионном уравнении как была, так и осталась неполной, хорошей модели не получено. Но если после изменения состава факторов коэффициент детерминации возрос с 65 до 90% – на те же 25 пунктов, это изменение имеет другое качественное содержание: получена хорошая регрессионная модель, в основном объясняющая вариацию результативного признака с достаточно полной системой факторов.

9.4. СРЕДНИЕ ПОКАЗАТЕЛИ ТЕНДЕНЦИИ ДИНАМИКИ

Средние показатели динамики – средний уровень ряда, средние абсолютные изменения и ускорения, средние темпы роста – характеризуют тенденцию. Они необходимы при обобщении характеристик тенденции за длительный период, по различным периодам и незаменимы при сравнении развития за неодинаковые по длительности отрезки времени, при выборе аналитического выражения тренда. При наличии в динамическом ряду существенных колебаний уровней определение средних показателей тенденции требует

применения специальных методов статистики, которые излагаются в последующих разделах. В данном разделе рассматривается только форма, математические свойства средних показателей динамики и простейшие приемы их вычисления, применимые на практике к рядам со слабой колеблемостью.

Средний уровень интервального ряда динамики определяется как простая арифметическая средняя из уровней за равные промежутки времени:

$$\bar{y} = \frac{\sum_{(i)}^{n} y_i}{n} \qquad (9.11)$$

или как взвешенная арифметическая средняя из уровней за неравные промежутки времени, длительность которых и является весами.

По данным табл. 9.1 определим среднегодовые уровни урожайности картофеля по пяти-шестилетиям:

Периоды		Среднегодовые уровни, ц/га
Пятилетие	1986–1990	157,0
Шестилетие	1991–1996	184,8

Средние уровни принято условно относить к середине интервала времени, т. е. для пятилетия 1986–1990 гг. – к 1988 г., для шестилетия 1991–1996 – к середине между 1993 и 1994 гг., т. е. к 1993,5.

Если, например, с 1-го числа месяца по 18-е число на предприятии работали 45 человек, с 19-го по 27-е – 48 человек, а с 28-го по 31-е число – 54 человека, то среднее списочное число работников за месяц составит:

$$\bar{y} = \frac{45 \cdot 18 + 48 \cdot 9 + 54 \cdot 4}{31} = 47{,}03 \text{ чел.}$$

В моментном ряду роль, смысл среднего уровня в том, что он характеризует уже не состояние объекта в отдельные моменты, а его среднее, обобщенное состояние между начальным и конечным моментом учета. Из этого следует, что роль уровней, относящихся к начальному и конечному моменту, существенно иная, чем роль уровней на моменты внутри изучаемого отрезка времени. Начальный и конечный уровни находятся на границе изучаемого интервала, они наполовину относятся к предыдущему и последующему интервалам и лишь наполовину к изучаемому. Уровни, относящиеся к моментам внутри осредняемого интервала, целиком относятся только к нему. Отсюда получаем особую форму средней арифметической величины, называемой *хронологической средней:*

$$\bar{y}_{хрон} = \left(\frac{y_1}{2} + \sum_{i=2}^{n-1} y_i + \frac{y_n}{2} \right) : (n-1). \qquad (9.12)$$

Проблема вычисления среднего уровня моментного ряда при неравных промежутках между моментами является спорной и здесь не рассматривается.

Если известны точные даты изменения уровней моментного ряда то средний уровень определяется как

$$\bar{y} = \frac{\sum_{(i)} y_i t_i}{\sum t_i}. \qquad (9.13)$$

где t_i – время, в течение которого сохранялся уровень.

Средний абсолютный прирост (абсолютное изменение) определяется как простая арифметическая средняя из абсолютных изменений за равные промежутки времени (цепных абсолютных изменений) или как частное от деления базисного абсолютного изменения на число осредняемых отрезков времени от базисного до сравниваемого периода:

$$\bar{\Delta} = \frac{\sum \Delta_i}{n} = \frac{y_n - y_0}{n}. \qquad (9.14)$$

Как уже сказано в п. 9.1, при наличии существенной колеблемости уровней средний абсолютный прирост (изменение), как и средний темп следует вычислять, отделив сначала тренд от колебаний (соответствующая методика будет изложена ниже). Прямое определение среднего абсолютного прироста по крайним уровням ряда допустимо, если нет существенных колебаний уровней. Например,

добыча угля в России довольно равномерно снижалась с 337 млн т в 1992 г. до 262 млн т в 1995 г.[1].

По формуле (9.14) среднее годовое сокращение добычи угля составило: $\bar{\Delta} = \frac{262-337}{3} = -25$ млн т в год. Итак, добыча угля в период 1992 – 1995 гг. в среднем за год снижалась на 25 млн т в год, или на 2,08 млн т в месяц.

Для правильной интерпретации показатель среднего абсолютного изменения должен сопровождаться указанием двух единиц времени: 1) время, за которое он вычислен, к которому относится и которое он характеризует (в нашем примере это трехлетие – 1992 - 1995); 2) время, на которое показатель рассчитан, время, входящее в его единицу измерения, – 1 год. Можно рассчитать среднемесячный прирост за пятилетие, среднесуточное изменение за год, за месяц, за квартал.

Среднее ускорение абсолютного изменения применяется реже. Для его надежного расчета даже при слабых колебаниях уровней требуется применять методику аналитического выравнивания по параболе II порядка (см. п. 9.5 и 9.6). Не рекомендуется измерять среднее ускорение без абстрагирования от колебаний уровней. Для более грубого, приближенного расчета среднего ускорения можно воспользоваться средними годовыми уровнями, сглаживающими колебания. Например, среднегодовое производство мяса в Российской Федерации составляло:

Годы	1976 – 1980	1981 – 1985	1986 – 1990
Млн т	7,40	8,09	9,68

Абсолютный прирост за второе пятилетие в сравнении с первым составил 0,69 млн т, за третье в сравнении со вторым – 1,59 млн т. Следовательно, ускорение в третьем пятилетии по сравнению со вторым составило 1,59 – 0,69 = 0,90 млн т в год за пять лет, а среднегодовое ускорение прироста равно: 0,90 : 5 = 0,18 млн т в год за год. Среднее ускорение требует указания трех единиц времени, хотя, как правило, две из них одинаковы: период, на который рассчитан прирост, и время, на которое рассчитано ускорение.

Средний темп изменения определяется наиболее точно при аналитическом выравнивании динамического ряда по экспоненте (см. п. 9.5 и 9.6). Если можно пренебречь колеблемостью, то средний

[1] Россия в цифрах. 1996: Статистический сборник / Госкомстат России. – М.: Финансы и статистика, 1996. – С. 297.

темп определяют как *геометрическую среднюю* (см. гл. 5) из цепных темпов роста за n лет или из общего (базисного) темпа роста за n лет:

$$\bar{k} = \sqrt[n]{\prod_{i=1}^{n} k_i} = \sqrt[n]{\frac{y_n}{y_0}}. \qquad (9.15)$$

Например, стоимость потребительской корзины за год в результате инфляции возросла в 6 раз. Каков средний месячный темп инфляции?

$$\bar{k} = \sqrt[12]{6} = 1{,}16, \text{ или } 116\%,$$

т.е. в среднем за месяц цена увеличивалась на 16% к уровню предыдущего месяца.

Средний темп роста так же, как средний прирост, следует сопровождать указанием двух единиц времени: 1) периода, который им характеризуется; 2) периода, на который рассчитан темп. Например, среднегодовой темп за последнее десятилетие; среднемесячный темп за полугодие и т.п.

Если исходной информацией служат темпы прироста и нужно вычислить их среднегодовую величину, то предварительно следует все темпы прироста превратить в темпы роста, прибавив 1, или 100%, вычислить их среднюю геометрическую и снова вычесть 1, или 100%. Интересно, что ввиду асимметрии темпа прироста и темпа сокращения при равных их величинах общий темп прироста всегда отрицателен. Так, если за первый год объем производства вырос на 20%, а за второй снизился на 20% (темпы цепные), то за два года имеем:

средний темп роста $\bar{k} = \sqrt{1{,}2 \cdot 0{,}8} = 0{,}9798$, или 97,98%;

средний темп прироста $\bar{k} - 1 = -0{,}0202$, или – 2,02%.

Как отмечалось в главе 5, применяя для вычисления среднего темпа среднюю геометрическую, мы опираемся на соблюдение фактического отношение конечного уровня к начальному при замене фактических темпов на средние. В практических задачах может потребоваться вычисление среднего уровня при условии соблюдения отношения суммы уровней за период к уровню, принятому за базу. Например, если общий выпуск продукции за пятилетие должен составить 800% к базисному (среднегодовому за предыдущие 5 лет выпуску), или, что то же самое, среднегодовой уровень дол-

жен составить 160% к базовому уровню, каков должен быть среднегодовой темп роста выпуска продукции? В 1974 г. украинские статистики А. и И. Соляники предложили следующую приближенную формулу для среднего темпа роста, удовлетворяющую этому условию:

$$\bar{k}_{пар} = 1 + \frac{-3}{2(m-1)} + \sqrt{\frac{9}{4(m-1)^2} + \frac{6}{m(m^2-1)}\left(\frac{\sum_{i=1}^{m} y_i}{y_0} - m\right)}, \qquad (9.16)$$

где m – число суммируемых уровней;
y_0 – базисный уровень.

Темп роста данного вида называют параболическим (отсюда обозначение $\bar{k}_{пар}$), так как он вычисляется по уравнению параболы порядка m. При $m = 5$ имеем:

$$k_{пар} = 1 - \frac{3}{8} + \sqrt{\frac{9}{64} + \frac{1}{20}\left(\frac{\sum y_i}{y_0} - 5\right)} =$$

$$= 1 - 0{,}375 + \sqrt{0{,}1406 + 0{,}05(8-5)} = 1{,}16407 \simeq 116{,}4\%.$$

Расчет по этому среднегодовому темпу дает сумму выпуска за 5 лет в 8,069 раза больше базисной, т.е. приближение хорошее. В общем виде проблема параболических темпов исследована саратовским статистиком Л. С. Казинцом в книге «Темпы роста и абсолютные приросты» (М.: Статистика, 1975). Им составлены таблицы, с помощью которых, зная отношение суммы уровней к базисному уровню и число суммируемых уровней m, можно получить $\bar{k}$пар. Таблица Л. С. Казинца рассчитана на основе нахождения корней уравнения:

$$\bar{k} + \bar{k}^2 + \bar{k}^3 + \ldots + \bar{k}^n = \left(\sum_1^n y_i : y_0\right).$$

Для нашего примера таблица Л. С. Казинца дает среднегодовой темп роста 116,1% и сумму выпуска в 8,00016 раза больше базисной.

Если необходимо определить средний темп изменения, исходя из заданной на n периодов суммы абсолютных изменений, то следует использовать формулу (9.17):

$$\bar{k} = \sqrt[n]{\frac{y_0 + \sum_{i=1}^{n} \Delta_i}{y_0}} = \sqrt[n]{1 + \frac{\sum_{i=1}^{n} \Delta_i}{y_0}}, \qquad (9.17)$$

где y_0 – начальный уровень;

$\sum_{i=1}^{n} \Delta_i$ – сумма абсолютных изменений;

$y_n = y_0 + \sum_{i=1}^{n} \Delta_i$ – конечный уровень.

Рассмотрим применение этой формулы. В 1995 г. в России добыто 262 млн т угля. Каким должен быть среднегодовой темп роста добычи, чтобы к 2000 г. (за 5 лет) сумма абсолютных приростов достигла 150 млн т? Или иначе: чтобы к 2000 г. добыча достигла 412 млн т? Делим 150 млн т на базисный уровень 262 млн т, получаем 57,25% прироста, далее по формуле (9.17) вычисляем среднегодовой темп: $\bar{k} = \sqrt[5]{1{,}5725} = 1{,}09476$ или 109,476%.

Вычисления проверим в следующей таблице:

Годы	Добыча, млн т	Абсолютный прирост, млн т/год
1995	262 –	–
1996	262 · 1,09476 = 286,8	24,8
1997	286,8 · 1,09476 = 314,0	27,2
1998	314,0 · 1,09476 = 343,8	29,8
1999	343,8 · 1,09476 = 376,3	32,5
2000	376,3 · 1,09476 = 412,0	35,7
Итого	1732,9	150,0

Интересную задачу представляет определение срока, за который ряд с бóльшим средним показателем динамики, но меньшим начальным уровнем догонит другой ряд с бóльшим начальным уровнем, но меньшим показателем динамики.

Если в качестве показателя динамики берутся средние абсолютные приросты, то задача решается просто. Пусть имеется первый ряд с базисным уровнем y_{01} и средним абсолютным приростом $\bar{\Delta}_1$ и второй ряд с соответствующими показателями y_{02} и $\bar{\Delta}_2$; причем $y_{02} > y_{01}$, $\bar{\Delta}_2 < \bar{\Delta}_1$. Тогда уровень первого ряда сравняется с уровнем второго ряда через

$$\frac{y_{02}-y_{01}}{\bar{\Delta}_1-\bar{\Delta}_2} \text{ лет}.$$

Та же задача может быть решена на основе *ускорений.* Имеем первый ряд с базисным уровнем y_{01}, базисным абсолютным изменением a_{01} и средним ускорением b_1; второй ряд – с показателями y_{02}, a_{02}, b_2. При каком числе n периодов (лет) после базисного уровня рядов сравняются?

Тенденции рядов параболические:

$$y_{n1} = y_{01} + a_{01}n + \frac{b_1 n^2}{2};$$

$$y_{n2} = y_{02} + a_{02}n + \frac{b_2 n^2}{2}.$$

Приравняв правые части уравнений, получим:

$$y_{02} + a_{02}n + \frac{b_2 n^2}{2} = y_{01} + a_{01}n + \frac{b_1 n^2}{2}$$

или

$$(y_{02}-y_{01}) + n(a_{02}-a_{01}) + n^2\left(\frac{b_2-b_1}{2}\right) = 0. \qquad (9.18)$$

Искомый срок n является корнем этого квадратного уравнения. Если, например, имеем:

$y_{01} = 500$; $a_{01} = 40$; $b_1 = 2$;

$y_{02} = 300$; $a_{02} = 26$; $b_2 = 3$;

то – $200 + (-14)\,n + 0{,}5n^2 = 0$.

Откуда

$$n = \frac{14 \pm \sqrt{196+400}}{2 \cdot 0{,}5} = 14 \pm 24{,}4.$$

Второй ряд догонит первый по уровню через 38,4 года; в прошлом уровни рядов были одинаковы 10,4 года назад. Будущие равные уровни составляют 3510, а прошлые были равны 192.

Если мы хотим найти срок n, через который уровни рядов сравняются, то эту задачу можно решить и на основе средних *темпов динамики.*

Имеем:

$$y_{02} \cdot \bar{k}_2^{\,n} = y_{01}\bar{k}_1^{\,n}.$$

Логарифмируя это равенство получаем:

$$n\log\bar{k}_2 + \log y_{02} = n\log\bar{k}_1 + \log y_{01}.$$

Откуда

$$n(\log\bar{k}_2 - \log\bar{k}_1) = \log y_{01} - \log y_{02}, \qquad (9.19)$$

$$n = \frac{\log y_{01} - \log y_{02}}{\log\bar{k}_2 - \log\bar{k}_1},$$

т. е. искомый срок равен частному от деления разности логарифмов уровней рядов в базисном периоде на разность логарифмов темпов изменения, только переставленных при вычитании. Обычно и в числителе, и в знаменателе от большего логарифма вычитается меньший. Например, первый ряд имеет $y^{\mathrm{I}}_0 = 300$; $k^{\mathrm{I}} = 1{,}09$; второй ряд имеет $y_0^{\mathrm{II}} = 100$; $k^{\mathrm{II}} = 1{,}2$. Тогда:

$$n = \frac{\ln 300 - \ln 100}{\ln 1{,}2 - \ln 1{,}09} = \frac{5{,}70382 - 4{,}60517}{0{,}18232 - 0{,}08618} = 11{,}43.$$

Через 11,43 года уровень второго ряда сравняется с первым при сохранении экспоненциальных трендов обоих рядов.

9.5. МЕТОДЫ ВЫЯВЛЕНИЯ ТИПА ТЕНДЕНЦИИ ДИНАМИКИ

Прежде чем применить методы математического анализа для вычисления параметров уравнения тренда, необходимо выявить тип тенденции, а эта задача не является чисто математической. Наличие колебаний уровней крайне усложняет выявление типа тенденции и требует всестороннего подхода к этой проблеме, прежде всего качественного изучения характера развития объекта. При этом нужно дать ответ на такие вопросы:

1. Были ли условия развития объекта достаточно однородными в изучаемый период?

2. Каков характер действия основных факторов развития?

3. Не произошло ли качественное, существенное изменение условий развития объекта внутри изучаемого периода времени?

Если, например, часть периода предприятие работало по старой технологии, а затем произошло техническое перевооружение – введены новые цехи, поточные линии, то единой тенденции показателей за весь период не будет, скорее всего нужна «периодизация» ряда, т.е. его дробление на отдельные подпериоды: до реконструкции, во время таковой (если она длительна) и после освоения новой технологии.

Чем крупнее изучаемая система, чем больше факторов влияют на динамику изучаемого признака, тем реже возможны резкие, скачкообразные изменения в ряду динамики (не колебания, а именно изменения в тенденции). Большие и сложные системы обладают значительной инерцией, и для скачкообразного, резкого изменения тенденции такой системы требуются большие затраты ресурсов, которые общество выделить не в состоянии. Поэтому такое столь коренное изменение в экономике, как переход от командно-административного планового хозяйства к рыночной регулируемой экономике, в масштабе нашей страны неизбежно займет достаточно большое время, за которое сформируются новые тенденции народнохозяйственных показателей. Чтобы разглядеть эти новые тенденции, понадобится время.

Напротив, в масштабе отдельных предприятий вполне возможны резкие изменения, переходы от одной тенденции к другой.

Рассмотрим некоторые основные типы уравнений тренда, выражающие те или иные качественные свойства развития.

А. Линейная форма тренда:

$$\hat{y} = a + bt, \qquad (9.20)$$

где $\hat{y}$ – уровни, освобожденные от колебаний, выравненные по прямой;
a – начальный уровень тренда в момент или период, принятый за начало отсчета времени t;
b – среднегодовой абсолютный прирост (среднее изменение за единицу времени); константа тренда.

Линейный тренд хорошо отражает тенденцию изменений при действии множества разнообразных факторов, изменяющихся различным образом по разным закономерностям. Равнодействующая этих факторов при взаимопогашении особенностей отдельных факторов (ускорение, замедление, нелинейность) часто выражается в примерно постоянной абсолютной скорости изменения, т.е. в прямолинейном тренде. Таковы, например, тенденции динамики урожайности для масштаба области, республики, крупного региона, страны в целом.

Б. Параболическая форма тренда:

$$\hat{y} = a + bt + ct^2, \qquad (9.21)$$

где c – квадратический параметр, равный половине ускорения; константа параболического тренда. Остальные обозначения прежние.

Параболическая форма тренда выражает ускоренное или замедленное изменение уровней ряда с постоянным ускорением. Такой характер развития можно ожидать при наличии важных факторов прогрессивного развития (прогрессирующее поступление нового высокопроизводительного оборудования, увеличение среднесуточного прироста живого веса поросят с возрастом и т.п.). Ускоренное возрастание может происходить в период после снятия каких-то сдерживающих развитие преград – ограничений в распределении дохода, в уровне оплаты труда, при повышении цены реализации на дефицитную продукцию.

Параболическая форма тренда с отрицательным ускорением ($c < 0$) приводит со временем не только к приостановке роста уровня, но и к его снижению со все большей скоростью. Такой характер развития может быть свойствен производству устаревшей продукции, ликвидируемой отрасли сельского хозяйства на предприятии (ферме) и т.п.

Парабола 2-го порядка (квадратическая) имеет либо максимум (если $c < 0$ и $b > 0$), либо минимум ($b < 0$, $c > 0$). Для нахождения экстремума производную параболы по времени t следует приравнять нулю и решить полученное уравнение относительно t. Например, если население города (тыс. чел.) возрастает по параболе

$$y = 1800 + 80t - 2t^2,$$

то производная по времени df/dt будет иметь вид: $80 - 4t = 0$, откуда $t = 20$. Максимум населения будет достигнут через 20 лет после начала отсчета времени, и это максимальное население составит:

$$\hat{y}_{max} = 1800 + 80 \cdot 20 - 2 \cdot 20^2 = 2600 \text{ тыс. человек.}$$

В. Экспоненциальная форма тренда:

$$\tilde{y} = ak^t \qquad (9.22)$$

где k – темп изменения в разах; константа тренда.

Если $k > 1$, экспоненциальный тренд выражает тенденцию ускоренного и все более ускоряющегося возрастания уровней. Такой характер свойствен, например, размножению организмов при отсутствии ограничения со стороны среды: кормов, пространства, хищников, болезней. При росте по экспоненте абсолютный прирост пропорционален достигнутому уровню. Так росло население Земли в эпоху «демографического взрыва» в XX столетии; сейчас этот период заканчивается и темп роста населения стал уменьшаться. Если бы он остался на уровне 1960 – 1970 гг. т. е. около 2% прироста в год от 1985 г., когда население составило 5 млрд чел., то к 2500 г. население Земли достигло бы уровня: 5 млрд $\cdot$ $1{,}02^{515} = 134$ трлн 286 млрд человек; на 1 человека приходилось бы примерно 1 м^2 всей площади суши. Ясно, что рост любого объекта по экспоненциальному закону может продолжаться только небольшой исторический период времени, ибо ресурсы для любого процесса развития всегда встретят ограничения.

При $k < 1$ экспоненциальный тренд означает тенденцию постоянно все более замедляющегося снижения уровней динамического ряда. Такая тенденция может быть присуща динамике трудоемкости продукции, удельных затрат топлива, металла на единицу полезного эффекта (на 1 кВт ч, на 1 м^2 жилой площади и т.д.) при технологическом прогрессе; экстремальных точек экспонента не имеет.

Г. Логарифмическая форма тренда:

$$\hat{y} = a + b\log t. \qquad (9.23)$$

Логарифмический тренд пригоден для отображения тенденции замедляющегося роста уровней при отсутствии предельного возможного значения. Замедление роста становится все меньше и меньше, и при достаточно большом t логарифмическая кривая становится малоотличимой от прямой линии. Логарифмический тренд пригоден для отображения роста спортивных достижений (чем они выше, тем труднее их улучшать), роста производительности агрегата по мере его освоения и совершенствования, повышения продуктивности скота или вообще эффективности системы при ее совершенствовании без качественных, коренных преобразований. Экстремума логарифмическая кривая не имеет.

Д. Тренд в форме степенной кривой:

$$\hat{y} = at^b, \qquad (9.24)$$

где b – константа тренда.

При $b = 1$ имеем линейный тренд, $b = 2$ – параболический и т.п. Степенная форма – гибкая, пригодная для отображения изменений с разной мерой пропорциональности изменений во времени. Жестким условием является обязательное прохождение через начало координат: при $t = 0$, $\tilde{y} = 0$. Можно усложнить форму тренда: $\tilde{y} = a + t^b$ или $\tilde{y} = a + ct^b$, но эти уравнения нельзя логарифмировать, трудно вычислять параметры, и они крайне редко применяются.

Е. Гиперболическая форма тренда:

$$\hat{y} = a + \frac{b}{t}. \qquad (9.25)$$

Если $b > 0$, гиперболический тренд выражает тенденцию замедляющегося снижения уровня, стремящегося к пределу a. Если $b < 0$, тренд выражает тенденцию замедляющегося роста уровней, стремящихся в пределе к a. Следовательно, гиперболическая форма тренда подходит для отображения тенденции, процессов, ограниченных предельным значением уровня (предельным коэффициентом полезного действия двигателя, пределом 100%-ной грамотности населения и т.п.).

Ж. Логистическая форма тренда:

$$\hat{y} = \frac{y_{max} - y_{min}}{e^{a+bt} + 1} + y_{min}, \qquad (9.26)$$

где e – основание натуральных логарифмов;
y_{max}, y_{min} – максимальное и минимальное из возможных значений уровня;
a, b – параметры тренда.

Логистическая кривая имеет форму латинской буквы s положенной на бок, отчего еще называется *эсобразной кривой.* Она имеет два перегиба: от ускоряющегося роста к равномерному (вогнутость) и от равномерного роста посреди периода к замедляющемуся (выпуклость). Она подходит для отображения развития в течение длительного периода, проходящего все фазы, например процесса насыщения потребителей каким-то новым товаром, скажем, телевизорами: сначала медленный, но все ускоряющийся рост доли семей,

имеющих телевизор, затем рост равномерный (примерно от 30 – 40% семей до 70 – 80%). Затем рост доли семей, имеющих телевизор, замедляется по мере приближения доли к 100%. Если $y_{min} = 0$, $y_{max} = 100\%$, или 1, уравнение упрощается до формы

$$\hat{y} = \frac{1}{e^{a+bt}+1}. \qquad (9.27)$$

После теоретического исследования особенностей разных форм тренда необходимо обратиться к фактическому ряду динамики, тем более что далеко не всегда можно надежно установить, какой должна быть форма тренда из чисто теоретических соображений. По фактическому динамическому ряду тип тренда устанавливают на основе графического изображения, путем осреднения показателей динамики, на основе статистической проверки гипотезы о постоянстве параметра тренда.

На рис. 9.1 достаточно хорошо видно, что тренд урожайности выражен прямой линией. Исходный ряд уровней короткий, поэтому на данном примере нельзя использовать другие приемы. Применим их к анализу динамики индекса цен на нетопливные товары развивающихся стран за 1979 – 1995 гг.[1]

Скользящая пятилетняя средняя, сглаживая колебания отдельных уровней, довольно отчетливо показывает тенденцию равномерного снижения уровней. Если разбить ряд на три части, то средние уровни также подтверждают этот вывод: за 1979 – 1983 гг. средний уровень равен 112,3; за 1984 – 1989 гг. – 103,0; за 1990 – 1995 гг. – 97,0. Существенного различия в величине снижения среднегодовых уровней нет. Оба приема – скользящая средняя и средние уровни по частям ряда – не свободны от субъективных факторов. Можно скользящую среднюю вычислять не за 5 лет, а за 6 или 7; можно иначе разбить ряд на три части или на другое число частей.

Более обоснованным приемом выявления тренда является проверка статистической гипотезы о постоянстве того или иного показателя динамики[2]. Рассмотрим этот прием по данным табл. 9.4.

[1] International Monetary fund // World Economic Outlook. – Washington: D. C., 1996. – P. 68.

[2] Прием предложен М. С. Каяйкиной в статье «Выбор типа линии при аналитическом выравнивании динамических рядов урожайности сельскохозяйственных культур» // Записки ЛСХИ, – Ленинград – Пушкин, 1972. – Т. 196.

Таблица 9.4

Проверка гипотезы о линейном тренде индекса цен (1990 г. = 100%)

Годы	Уровни y_i	Суммы за 5 лет	Средние скользящие за 5 лет, $\bar{y}_i$	Цепные приросты, $\bar{y}_i - \bar{y}_{i-1}$
1979	105	...	...	...
1980	111	...	...	...
1981	110	550	110,0	...
1982	106	569	113,8	+3,8
1983	118	571	114,2	+0,4
1984	124	553	110,6	–3,6
1985	113	538	107,6	–3,0
1986	92	529	105,8	–1,8
1987	91	518	103,6	–2,2
1988	109	505	101,0	–2,6
1989	113	507	101,4	+0,4
1990	100	507	101,4	0,0
1991	94	490	98,0	–3,4
1992	91	479	95,8	–2,2
1993	92	485	97,0	+1,2
1994	102	...	...	...
1995	106	...	...	...
	1777	–	–	–

В первую очередь проверяется гипотеза о наиболее простой – линейной форме уравнения тренда, т. е. о несущественности различий цепных абсолютных изменений. Имеем 12 абсолютных изменений скользящей средней, которая хотя и сгладила сильные колебания уровней ряда, но как видим, ее абсолютные изменения далеко не одинаковы. Разбиваем эти 12 цепных приростов на два подпериода: по 6 приростов в каждом, и для каждого подпериода вычисляем среднюю $\bar{\Delta}_k$, среднее квадратическое отклонение (СКО) как оценку генерального СКО с учетом потери одной степени свободы вариации, s

$$s_{\Delta k} = \sqrt{\frac{\sum_{i=1}^{5}(\Delta_i - \bar{\Delta})^2}{6-1}}$$

и среднюю ошибку среднего изменения $m_{\Delta k}$ по правилам, рассмотренным в главе 7:

$$m_{\Delta k} = \frac{s_{\Delta k}}{\sqrt{6}}.$$

$$\overline{\Delta}_1 = \frac{-6,4}{6} = -1,07; \qquad s_1 = \sqrt{\frac{37,82}{6-1}} = 2,75;$$

$$m_1 = s_1 : \sqrt{6} = 1,123; \qquad \overline{\Delta}_2 = \frac{-6,6}{6} = -1,1;$$

$$s_2 = \sqrt{\frac{17,50}{6-1}} = 1,87; \qquad m_2 = s_2 : \sqrt{6} = 0,763.$$

Для проверки гипотезы о несущественности различий между средними абсолютными изменениями по подпериодам $\overline{\Delta}_1$, $\overline{\Delta}_2$. М. С. Каяйкина предложила проверять существенность их различий попарно по *t*-критерию Стьюдента. Затем методика была дополнена и усовершенствована А. И. Манеллей, предложившим проверять существенность всех различий сразу по критерию Фишера.

Средняя случайная ошибка разностей двух выборочных средних оценок, как показано в гл. 7, есть корень квадратный из суммы квадратов ошибок каждой из средних, т. е.

$$m_{d(I-II)} = \sqrt{m_1^2 + m_2^2} = \sqrt{1,123^2 + 0,763^2} = 1,356.$$

Критерий Стьюдента для существенности различия двух среднегодовых приростов (изменений) составит:

$$t = \frac{-1,07 - (-1,10)}{1,356} = 0,022.$$

Критическое значение критерия при уровне значимости 0,05 и при (6–1) + (6–1) = 10 степенях свободы равно 2,23 (см. Приложение 2). Фактическое значение много меньше. Следовательно вероятность того, что различие среднегодовых приростов в разные подпериоды случайно, превышает 0,05 и гипотеза о равенстве приростов не отклоняется. А значит, тенденцию динамики на всем протяжении ряда можно считать линейной.

Если же гипотеза о линейности отклоняется, по скользящим средним и их цепным приростам вычисляют ускорения приростов и аналогичным методом проверяют существенность различия ускорения в подпериодах. Если несущественно различие ускорений, принимается гипотеза о том, что тренд – парабола II порядка. Если и гипотеза о постоянстве ускорений отклоняется, то по скользящей средней вычисляют цепные темпы роста и проверяют гипотезу об их постоянстве по подпериодам. Подтверждение (неотклонение) этой гипотезы означает принятие гипотезы о том, что тренд экспоненциальный.

Проверка гипотез о других типах тенденций динамики, рассмотренных в п. 9.4, сложнее и здесь излагаться не будет. Итак, в нашем примере принято решение считать тренд линейным, и следует приступить к вычислению его параметров.

9.6. МЕТОДИКА ИЗМЕРЕНИЯ ПАРАМЕТРОВ ТРЕНДА

Когда тип тренда установлен, необходимо вычислить оптимальные значения параметров тренда исходя из фактических уровней. Для этого обычно используют метод наименьших квадратов (МНК). Его значение уже рассмотрено в предыдущих главах учебного пособия, в данном случае оптимизация состоит в минимизации суммы квадратов отклонений фактических уровней ряда от выравненных уровней (от тренда). Для каждого типа тренда МНК дает систему нормальных уравнений, решая которую вычисляют параметры тренда. Рассмотрим лишь три такие системы: для прямой, для параболы 2-го порядка и для экспоненты. Приемы определения параметров других типов тренда рассматриваются в специальной монографической литературе.

Для *линейного тренда* нормальные уравнения МНК имеют вид:

$$\begin{aligned} na + b\sum_{i=1}^{n} t_i &= \sum_{i=1}^{n} y_i, \\ a\sum_{i=1}^{n} t_i + b\sum_{i=1}^{n} t_i^2 &= \sum_{i=1}^{n} y_i t_i, \end{aligned} \qquad (9.28)$$

где y_i – уровни исходного ряда динамики;
t_i – номера периодов или моментов времени;
n – число уровней ряда.

Систему можно упростить, перенеся начало отсчета времени t_i в середину ряда. Тогда $\sum t_i$ (а также суммы всех нечетных степеней t_i) будет равна нулю и система приобретет вид:

$$na = \sum_i y_i,$$
$$b\sum_i t_i^2 = \sum y_i t_i, \qquad (9.29)$$

откуда $a = \bar{y}$; $\quad b = \dfrac{\sum y_i t_i}{\sum t_i^2}$.

Суммы $\sum y_i t_i$ и $\sum t_i^2$ для значений t_i – от $-(n-1)/2$ до $+(n-1)/2$ (см. табл. 9.5), при этом $\sum t_i^2$ можно вычислить как $(n^3 - n)/12$.

Нормальные уравнения МНК для *параболы 2-го порядка* имеют следующий вид:

$$na + b\sum_{(i)} t_i + c\sum_{(i)} t_i^2 = \sum_{(i)} y_i,$$
$$a\sum_{(i)} t_i + b\sum_{(i)} t_i^2 + c\sum_{(i)} t_i^3 = \sum_{(i)} y_i t_i, \qquad (9.30)$$
$$a\sum_{(i)} t_i^2 + b\sum_{(i)} t_i^3 + c\sum_{(i)} t_i^4 = \sum_{(i)} y_i t_i^2.$$

После переноса начала отсчета t в середину ряда имеем:

$$na + c\sum_{(i)} t_i^2 = \sum_{(i)} y_i,$$
$$b\sum_{(i)} t_i^2 = \sum_{(i)} y_i t_i, \qquad (9.31)$$
$$a\sum_{(i)} t_i^2 + c\sum_{(i)} t_i^4 = \sum_{(i)} y_i t_i^2.$$

Суммы включают значения t_i от $-(n-1):2$ до $+(n-1):2$, при этом сумма биквадратов может быть вычислена по формуле

$$\sum t^4 = \frac{3n^5 - 10n^3 + 7n}{240}.$$

Нормальные уравнения МНК для *экспоненты* имеют следующий вид:

$$n\ln a + \ln k\sum_{i=1}^{n} t_i = \sum_{i=1}^{n} \ln y_i,$$
$$\ln a\sum_{i=1}^{n} t_i + \ln k\sum_{i=1}^{n} t_i^2 = \sum_{i=1}^{n} \ln y_i t_i. \qquad (9.32)$$

После переноса начала отсчета t_i в середину ряда получим:

$$n\ln a = \sum \ln y_i \text{ откуда } \ln a = \overline{\ln y_i},$$
$$\ln k\sum t_i^2 = \sum \ln y_i t_i, \text{ откуда } \ln k = \frac{\sum \ln y_i t_i}{\sum t_i^2}. \qquad (9.33)$$

В формуле (9.33) суммирование от $t = -(n-1):2$ до $t = (n-1):2$; в целом формула (9.33) аналогична формуле для линейного тренда (9.29).

По данным табл. 9.1 рассчитаем все три перечисленных тренда для динамического ряда урожайности картофеля с целью их сравнения (см. табл. 9.5).

Таблица 9.5

Расчет параметров трендов

Годы	Уровни y_i	t_i	$y_i t_i$	t_i^2	$y_i t_i^2$	t_i^4	$\ln y_i$	$\ln y_i t_i$	Уровни трендов $\hat{y}_i$		
									линейного	параболы II порядка	экспоненты
1986	149	–5	–745	25	3725	625	5,0039	–25,020	150	141,7	150,3
1987	145	–4	–580	16	2320	256	4,9767	–19,907	155	151,2	154,2
1988	168	–3	–504	9	1512	81	5,1240	–15,372	159	159,5	158,3
1989	146	–2	–292	4	584	16	4,9836	–9,967	163	166,7	162,4
1990	177	–1	–177	1	177	1	5,1762	–5,176	168	172,8	166,7
1991	176	0	0	0	0	0	5,1705	0	172	177,7	171,1
1992	190	1	190	1	190	1	5,2470	5,247	177	181,6	175,6
1993	186	2	372	4	744	16	5,2257	10,451	181	184,4	180,2
1994	176	3	528	9	1584	81	5,1705	15,511	185	186,0	184,9
1995	211	4	844	16	3376	256	5,3519	21,407	190	186,5	189,8
1996	170	5	850	25	4250	625	5,1358	25,679	194	185,9	194,8
Σ	1894	0	486	110	18462	1958	56,5658	2,853	1894	1894	1888,3

Согласно формуле (9.29) параметры линейного тренда равны a = 1894/11 = 172,2 ц/га; b = 486/110 = 4,418 ц/га. Уравнение линейного тренда имеет вид: $\hat{y}$ = 172,2 + 4,418t, где t = 0 в 1987 г. Это означает, что средний фактический и выравненный уровень, отнесенный к середине периода, т.е. к 1991 г., равен 172 ц с 1 га, а среднегодовой прирост составляет 4,418 ц/га в год.

Параметры параболического тренда согласно (9.23) равны: b = 4,418; a = 177,75; c = –0,5571. Уравнение параболического тренда имеет вид $\bar{y}$ = 177,75 + 4,418t – 0,5571 t^2; t = 0 в 1991 г. Это означает, что абсолютный прирост урожайности замедляется в среднем на 2 · 0,56 ц/га в год за год. Сам же абсолютный прирост уже не является константой параболического тренда, а является средней величиной за период. В год, принятый за начало отсчета, т. е. 1991 г., тренд проходит через точку с ординатой 177,75 ц/га. Свободный член параболического тренда не является средним уровнем за период. Параметры экспоненциального тренда вычисляются по формулам (9.32) и (9.33) ln$a$ = 56,5658/11 = 5,1423; потенцируя, получаем: a = 171,1; ln$k$ = 2,853 : 110 = 0,025936; потенцируя, получаем: k = 1,02628.

Уравнение экспоненциального тренда имеет вид: $\bar{y}$ = 171,1·1,02628^t. Это означает, что среднегодовой темп роста урожайности за период составил 102,63%. В точке, принятой за начало отсчета, тренд проходит точку с ординатой 171,1 ц/га.

Рассчитанные по уравнениям трендов уровни записаны в трех последних графах табл. 9.5. Как видно по этим данным, расчетные значения уровней по всем трем видам трендов различаются ненамного, так как и ускорение параболы, и темп роста экспоненты невелики. Существенное отличие имеет парабола – рост уровней с 1995 г. прекращается, в то время как при линейном тренде уровни растут и далее, а при экспоненте их рост ускоряется. Поэтому для прогнозов на будущее эти три тренда неравноправны: при экстраполяции параболы на будущие годы уровни резко разойдутся с прямой и экспонентой, что видно из табл. 9.6. В этой таблице представлена распечатка решения на ПЭВМ по программе «Statgraphics» тех же трех трендов. Отличие их свободных членов от приведенных выше объясняется тем, что программа нумерует года не от середины, а от начала, так что свободные члены трендов относятся к 1986 г., для которого t = 0. Уравнение экспоненты на распечатке оставлено в логарифмированном виде. Прогноз сделан на 5 лет вперед, т.е. до 2001 г.. При изменении начала координат (отсчета времени) в уравнении параболы меняется и средний абсолютный прирост, параметр b, так как в результате отрицательного ускорения

Таблица 9.6

Распечатка решения на ПЭВМ по программе «Statgraphics»

Data	149	145	168	146	177	176	190	186	176	211	170
Forecast summary						ME	MSE	MAE	MAPE	MPE	
145,673 + 4,41818 × T						0,00000	173,127	11,2099	6,48645	–0,56996	
131,188 + 11,1035 × T – 0,55711 × T ^ 2						0,00000	148,918	9,91218	5,79005	–0,47525	
EXP (4,98665 + 0,02594498 × T)						0,48852	178,764	11,5423	6,66106	–0,28789	
Period	Period		Period		Period						
12	13		14		15		16				
198,691	203,109		207,527		211,945		216,364				
184,206	181,382		177,443		172,391		166,224				
199,945	205,202		210,596		216,133		221,815				

прирост все время сокращается, а его максимум – в начале периода. Константой параболы является только ускорение.

В строке «Data» приводятся уровни исходного ряда; «Forecast summary» означает сводные данные для прогноза. В следующих строках – уравнения прямой, параболы, экспоненты – в логарифмическом виде. Графа ME означает среднее расхождение между уровнями исходного ряда и уровнями тренда (выравненными). Для прямой и параболы это расхождение всегда равно нулю. Уровни экспоненты в среднем на 0,48852 ниже уровней исходного ряда. Точное совпадение возможно, если истинный тренд – экспонента; в данном случае совпадения нет, но различие , мало. Графа MAE – это дисперсия s^2 – мера колеблемости фактических уровней относительно тренда, о чем сказано в п. 9.7. Графа MAE – среднее линейное отклонение уровней от тренда по модулю (см. параграф 5.8); графа MAPE – относительное линейное отклонение в процентах. Здесь они приведены как показатели пригодности выбранного вида тренда. Меньшую дисперсию и модуль отклонения имеет парабола: она за период 1986 – 1996 гг. ближе к фактическим уровням. Но выбор типа тренда нельзя сводить лишь к этому критерию. На самом деле замедление прироста есть результат большого отрицательного отклонения, т. е. неурожая в 1996 г.

Вторая половина таблицы – это прогноз уровней урожайности по трем видам трендов на годы; t_i = 12, 13, 14, 15 и 16 от начала отсчета (1986 г.). Прогнозируемые уровни по экспоненте вплоть до 16-го года ненамного выше, чем по прямой. Уровни тренда-параболы – снижаются, все более расходясь с другими трендами.

Как видно в табл. 9.4, при вычислении параметров тренда уровни исходного ряда входят с разными весами – значениями t_i и их квадратов. Поэтому влияние колебаний уровней на параметры тренда зависит от того, на какой номер года приходится урожайный либо неурожайный год. Если резкое отклонение приходится на год с нулевым номером (t_i = 0), то оно никакого влияния на параметры тренда не окажет, а если попадет на начало и конец ряда, то повлияет сильно. Следовательно, однократное аналитическое выравнивание неполно освобождает параметры тренда от влияния колеблемости, и при сильных колебаниях они могут быть сильно искажены, что в нашем примере случилось с параболой. Для дальнейшего исключения искажающего влияния колебаний на параметры тренда следует применить *метод многократного скользящего выравнивания.*

Этот прием состоит в том, что параметры тренда вычисляются не сразу по всему ряду, а скользящим методом, сначала за первые m периодов времени или моментов, затем за период от 2-го до m + 1,

от 3-го до $(m + 2)$-го уровня и т.п. Если число исходных уровней ряда равно n, а длина каждой скользящей базы расчета параметров равна m, то число таких скользящих баз l или отдельных значений параметров, которые будут по ним определены, составит:

$$L = n + 1 - m.$$

Применение методики скользящего многократного выравнивания рассматривать, как видно из приведенных расчетов, возможно только при достаточно большом числе уровней ряда, как правило 15 и более. Рассмотрим эту методику на примере данных табл. 9.4 – динамики цен на нетопливные товары развивающихся стран, что опять же дает возможность читателю участвовать в небольшом научном исследовании. На этом же примере продолжим и методику прогнозирования в разделе 9.10.

Если вычислять в нашем ряду параметры по 11-летним периодам (по 11 уровням), то $l = 17 + 1 - 11 = 7$. Смысл многократного скользящего выравнивания в том, что при последовательных сдвигах базы расчета параметров на концах ее и в середине окажутся разные уровни с разными по знаку и величине отклонениями от тренда. Поэтому при одних сдвигах базы параметры будут завышаться, при других – занижаться, а при последующем усреднении значений параметров по всем сдвигам базы расчета произойдет дальнейшее взаимопогашение искажений параметров тренда колебаниями уровней.

Многократное скользящее выравнивание не только позволяет получить более точную и надежную оценку параметров тренда, но и осуществить контроль правильности выбора типа уравнения тренда. Если окажется, что ведущий параметр тренда, его константа при расчете по скользящим базам не беспорядочно колеблется, а систематически изменяет свою величину существенным образом, значит, тип тренда был выбран неверно, данный параметр константой не является.

Что касается свободного члена при многократном выравнивании, то нет необходимости и, более того, просто неверно вычислять его величину как среднюю по всем сдвигам базы, ибо при таком способе отдельные уровни исходного ряда входили бы в расчет средней с разными весами, и сумма выравненных уровней разошлась бы с суммой членов исходного ряда. Свободный член тренда – это средняя величина уровня за период, при условии отсчета времени от середины периода. При отсчете от начала, если первый уровень t_i = 1, свободный член будет равен: $a_0 = \bar{y} - b((N-l)/2)$. Рекомендуется длину скользящей базы расчета параметров тренда выбирать

не менее 9 – 11 уровней, чтобы в достаточной мере погасить колебания уровней. Если исходный ряд очень длинный, база может составлять до 0,7 – 0,8 его длины. Для устранения влияния долгопериодических (циклических) колебаний на параметры тренда, число сдвигов базы должно быть равно или кратно длине цикла колебаний. Тогда начало и конец базы будут последовательно «пробегать» все фазы цикла и при усреднении параметра по всем сдвигам его искажения от циклических колебаний будут взаимопогашаться. Другой способ – взять длину скользящей базы, равной длине цикла, чтобы начало базы и конец базы всегда приходились на одну и ту же фазу цикла колебаний.

Поскольку по данным табл. 9.4, уже было установлено, что тренд имеет линейную форму, проводим расчет среднегодового абсолютного прироста, т. е. параметра b уравнения линейного тренда скользящим способом по 11-летним базам (см. табл. 9.7). В ней же приведен расчет данных, необходимых для последующего изучения колеблемости в параграфе 9.7. Остановимся подробнее на методике многократного выравнивания по скользящим базам.

Рассчитаем параметр b по всем базам:

$$b_1 = \frac{-58}{110} = -0{,}527; \quad b_2 = \frac{-120}{110} = -1{,}091; \quad b_3 = \frac{-171}{110} = -1{,}535;$$

$$b_4 = \frac{-226}{110} = -2{,}055; \quad b_5 = \frac{-281}{110} = -2{,}555; \quad b_6 = \frac{-200}{111} = -1{,}818;$$

$$b_7 = \frac{-47}{110} = -0{,}427; \quad \bar{b} = -1{,}433; \quad a = \frac{1777}{17} = 104{,}53.$$

Многократное скользящее

Годы	Уровни y_i	I база		II база		III база		IV база		V база	
		t_i	$y_i t_i$	t_i	$y_i t_i$	t_i	$y_i t_i$	t_i	$y_i t_i$	t_i	$y_i t_i$
1979	105	–5	–525								
1980	111	–4	–444	–5	–555						
1981	110	–3	–330	–4	–440	–5	–550				
1982	106	–2	–212	–3	–318	–4	–424	–5	–530		
1983	118	–1	–118	–2	–236	–3	–354	–4	–472	–5	–590
1984	124	0	0	–1	–124	–2	–248	–3	–372	–4	–496
1985	113	1	113	0	0	–1	–113	–2	–226	–3	–339
1986	92	2	184	1	92	0	0	–1	–92	–2	–184
1987	91	3	273	2	182	1	91	0	0	–1	–91
1988	109	4	436	3	327	2	218	1	109	0	0
1989	113	5	565	4	452	3	339	2	226	1	113
1990	100			5	500	4	400	3	300	2	200
1991	94					5	470	4	376	3	282
1992	91							5	455	4	364
1993	92									5	460
1994	102										
1995	106										
Σ	1777	–	–58	–	–120	–	–171	–	–226	–	–281

Таблица 9.7

выравнивание по прямой

VI база		VII база		Тренд, $\hat{y}_i$	Колебания, $y_i - \hat{y}_i = u_i$	u_i^2	$u_i \cdot u_{i+1}$	Число поворотных точек
t_i	$y_i t_i$	t_i	$y_i t_i$					
				116,0	–11,0	121,00	...	
				114,6	–3,6	12,96	+39,6	
				113,1	–3,1	9,61	+11,16	•
				111,7	–5,7	32,49	+17,67	•
				110,3	+7,7	59,29	–43,89	
–5	–620			108,8	+15,2	231,04	+117,04	•
–4	–452	–5	–565	107,4	+5,6	31,36	+85,12	
–3	–276	–4	–368	106,0	–14,0	196,00	–78,40	•
–2	–182	–3	–273	104,5	–13,5	182,25	+189,00	
–1	–109	–2	–218	103,1	+5,9	34,81	–79,65	
0	0	–1	–113	101,7	+11,3	127,69	+66,67	•
1	100	0	0	100,2	–0,2	0,04	–2,26	
2	188	1	94	98,8	–4,8	23,04	+0,96	
3	273	2	182	97,4	–6,4	40,96	+30,72	•
4	368	3	276	95,9	–3,9	15,21	+24,96	
5	510	4	408	94,5	+7,5	56,25	–29,25	
		5	530	93,1	+12,9	166,41	+96,75	
–	–200	–	–47	1777,1	–	1340,41	+446,20	6

Уравнение тренда: $\hat{y} = 104{,}53 - 1{,}433t$; $t = 0$ в 1987 г.

Итак, индекс цен в среднем за год снижался на 1,433 пункта. Однократное выравнивание по всем 17 уровням может исказить этот параметр, ибо начальный уровень содержит значительное отрицательное отклонение, а конечный уровень – положительное. В самом деле, однократное выравнивание дает величину среднегодового изменения индекса всего на 0,953 пункта.

9.7. МЕТОДИКА ИЗУЧЕНИЯ И ПОКАЗАТЕЛИ КОЛЕБЛЕМОСТИ

Если при изучении и измерении тенденции динамики колебания уровней играли лишь роль помех, «информационного шума», от которого следовало по возможности абстрагироваться, то в дальнейшем сама колеблемость становится предметом статистического исследования. Значение изучения колебаний уровней динамического ряда очевидно: колебания урожайности, продуктивности скота, производства мяса экономически нежелательны, так как потребность в продукции агрокомплекса постоянна. Эти колебания следует уменьшать, применяя прогрессивную технологию и другие меры. Напротив, сезонные колебания объемов производства зимней и летней обуви, одежды, мороженого, зонтиков, коньков – необходимы и закономерны, так как спрос на эти товары тоже колеблется по сезонам и равномерное производство требует лишних затрат на хранение запасов. Регулирование рыночной экономики как со стороны государства, так и производителей в значительной мере состоит в регулировании колебаний экономических процессов.

Типы колебаний статистических показателей весьма разнообразны, но все же можно выделить три основных: пилообразную или маятниковую колеблемость, циклическую долгопериодическую и случайно распределенную во времени колеблемость. Их свойства и отличия друг от друга хорошо видны при графическом изображении рис. 9.2).

Пилообразная или *маятниковая колеблемость* состоит в попеременных отклонениях уровней от тренда в одну и в другую сторону. Таковы автоколебания маятника. Такие автоколебания можно наблюдать в динамике урожайности при невысоком уровне агротехники: высокий урожай при благоприятных условиях погоды выносит из почвы больше питательных веществ, чем их образуется естественным путем за год; почва обедняется, что вызывает снижение следу-

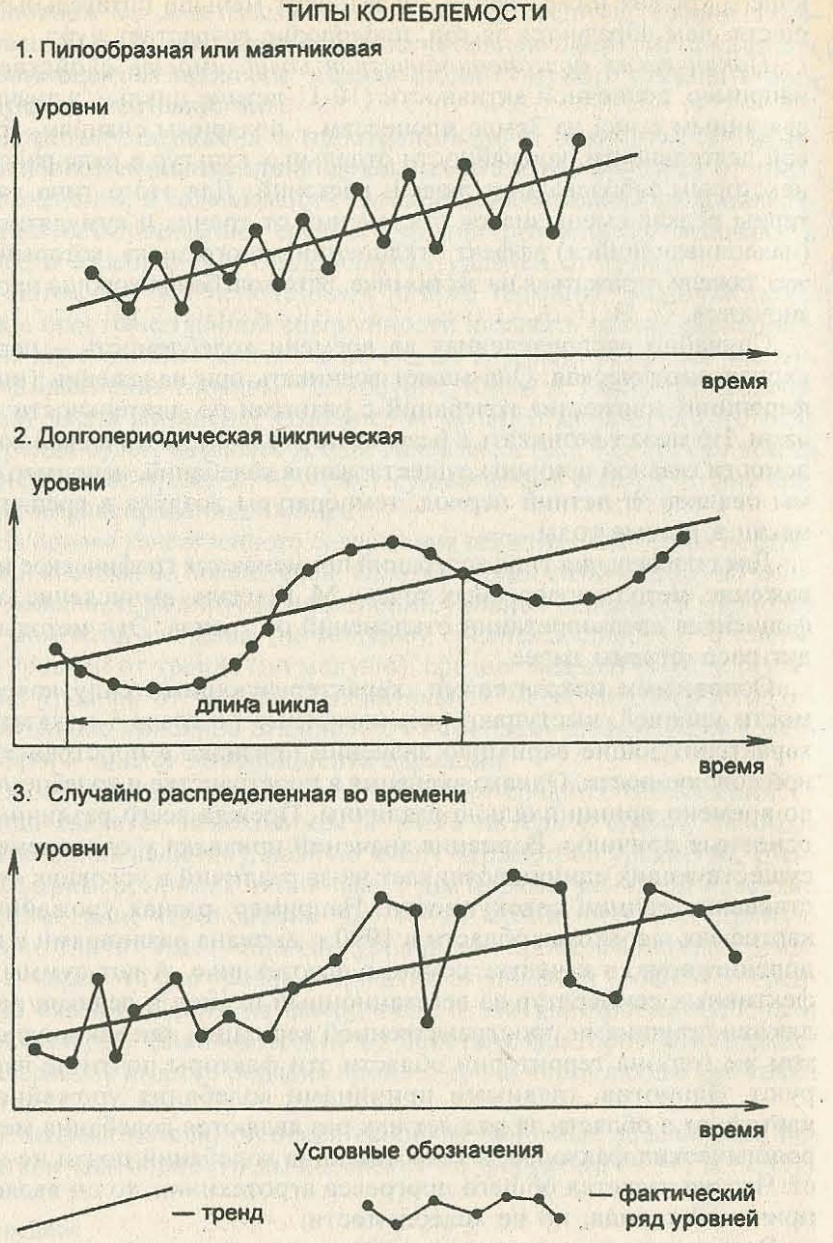


Рис. 9.2. Виды колебаний

ющего урожая ниже тренда, он выносит меньше питательных веществ, чем образуется за год, плодородие возрастает и т.д.

Циклическая долгопериодическая колеблемость свойственна, например, солнечной активности (10-11-летние циклы), а значит, и связанным с ней на Земле процессам – полярным сияниям, грозовой деятельности, урожайности отдельных культур в ряде районов, некоторым заболеваниям людей, растений. Для этого типа характерны редкая смена знаков отклонений от тренда и кумулятивный (накапливающийся) эффект отклонений одного знака, который может тяжело отражаться на экономике. Зато колебания хорошо прогнозируются.

Случайно распределенная во времени колеблемость – нерегулярная, хаотическая. Она может возникать при наложении (интерференции) множества колебаний с разными по длительности циклами. Но может возникать в результате столь же хаотической колеблемости главной причины существования колебаний, например суммы осадков за летний период, температуры воздуха в среднем за месяц в разные годы.

Для определения типа колебаний применяются графическое изображение, метод «поворотных точек» М. Кендэла, вычисление коэффициентов автокорреляции отклонений от тренда. Эти методы будут рассмотрены далее.

Основными показателями, характеризующими силу колеблемости уровней, выступают уже известные по главе 5 показатели, характеризующие вариацию значений признака в пространственной совокупности. Однако вариация в пространстве и колеблемость во времени принципиально различны. Прежде всего различны их основные причины. Вариация значений признака у одновременно существующих единиц возникает из-за различий в условиях существования единиц совокупности. Например, разная урожайность картофеля в совхозах области в 1990 г. вызвана различиями в плодородии почв, в качестве семян, в агротехнике. А вот суммы эффективных температур за вегетационный период и осадков не являются причинами пространственной вариации, так как в одном и том же году на территории области эти факторы почти не варьируют. Напротив, главными причинами колебания урожайности картофеля в области за ряд лет как раз являются колебания метеорологических факторов, а качество почв колебаний почти не имеет. Что же касается общего прогресса агротехники, то он является причиной тренда, но не колеблемости.

Второе коренное отличие состоит в том, что значения варьирующего признака в пространственной совокупности можно считать

в основном не зависимыми друг от друга, напротив, уровни динамического ряда, как правило, являются зависимыми: это показатели развивающегося процесса, каждая стадия которого связана с предыдущими состояниями.

В-третьих, вариация в пространственной совокупности измеряется отклонениями индивидуальных значений признака от среднего значения, а колеблемость уровней динамического ряда измеряется не их отличиями от среднего уровня (эти отличия включают и тренд, и колебания), а отклонениями уровней от тренда.

Поэтому лучше использовать разные термины: различия признака в пространственной совокупности называть только вариацией, но не колебаниями: никто же не станет называть различия численности населения Москвы, Петербурга, Киева и Ташкента «колебаниями числа жителей»! Отклонения уровней динамического ряда от тренда будем называть всегда колеблемостью. Колебания всегда происходят во времени, не может существовать колебаний вне времени, в фиксированный момент.

На основе качественного содержания понятия колеблемости строится и система ее показателей. *Показателями силы колебаний уровней являются*: амплитуда отклонений уровней отдельных периодов или моментов от тренда (по модулю), среднее абсолютное отклонение уровней от тренда (по модулю), среднее квадратическое отклонение уровней от тренда. Относительные меры колеблемости: относительное линейное отклонение от тренда и коэффициент колеблемости – аналог коэффициента вариации.

Особенностью методики вычисления средних отклонений от тренда является необходимость учета потерь степеней свободы колебаний на величину, равную числу параметров уравнения тренда. Например, прямая линия имеет два параметра, и, как известно из геометрии, через любые две точки можно провести прямую линию. Значит, имея лишь два уровня, мы проведем линию тренда точно через эти два уровня, и никаких отклонений уровней от тренда не окажется, хотя на самом деле и эти два уровня включали колебания, не были свободны от действия факторов колеблемости. Парабола второго порядка пройдет точно через любые три точки и т.п.

Учитывая потерю степеней свободы, основные абсолютные показатели колеблемости вычисляются по формулам (9.34) и (9.35):

среднее линейное отклонение

$$a(t) = \frac{\sum_{i=1}^{n} | y_i - \hat{y}_i |}{n - p}. \qquad (9.34)$$

среднее квадратическое отклонение

$$s(t)=\sqrt{\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{n-p}}, \qquad (9.35)$$

где y_i – фактический уровень;

$\hat{y}_i$ – выравненный уровень, тренд;

n – число уровней;

p – число параметров тренда.

Знак времени «t» в скобках после показателя означает, что это показатель не обычной пространственной вариации, как в главе V, а показатель колеблемости во времени.

Относительные показатели колеблемости вычисляются делением абсолютных показателей на средний уровень за весь изучаемый период. Расчет показателей колеблемости проведем по результатам анализа динамики индекса цен (см. табл. 9.7). Тренд примем по результатам многократного скользящего выравнивания, т. е. $\hat{y}$ = 104,53 – 1,433 t; t = 0 в 1987 г.

1. Амплитуда колебаний составила от –14,0 в 1986 г. до +15,2 в 1984 г., т. е. 29,2 пункта.

2. Среднее линейное отклонение по модулю найдем, сложив модули $|u_i|$ (их сумма равна 132,3), и разделив на $(n - p)$, согласно формуле (9.34):

$$a(t)=\frac{132{,}3}{17-2}=8{,}82 \text{ пункта.}$$

3. Среднее квадратическое отклонение уровней от тренда по формуле (9.35) составило:

$$s(t)=\sqrt{\frac{1340{,}41}{17-2}}=9{,}45 \text{ пункта.}$$

Небольшое превышение среднего квадратического отклонения над линейным указывает на отсутствие среди отклонений резко выделяющихся по абсолютной величине.

4. Коэффициент колеблемости: $v_{(t)}=\frac{s(t)}{\bar{y}}=\frac{9{,}45}{104{,}53}=0{,}0904$ или 9,04%. Колеблемость умеренная, не сильная. Для сравнения приводим показатели (без расчета) по колебаниям урожайности картофеля, данные таблиц 9.1 и 9.5 – отклонение от линейного тренда: $s(t)$ = 14,38 ц с 1 га, $v(t)$ = 8,35%.

Для выявления типа колебаний воспользуемся приемом, предложенным М. Кендэлом. Он состоит в подсчете так называемых «поворотных точек» в ряду отклонений от тренда u_i, т. е. локальных экстремумов. Отклонение, либо большее по алгебраической величине, либо меньшее двух соседних, отмечается точкой. Обратимся к рис. 9.2. При маятниковой колеблемости все отклонения, кроме двух крайних, будут «поворотными», следовательно, их число составит n – 2. При долгопериодических циклах на цикл приходятся один минимум и один максимум, а общее число точек составит $2(n : l)$, где l – длительность цикла. При случайно распределенной во времени колеблемости, как доказал М. Кендэл, число поворотных точек в среднем составит: 2/3 $(n - 2)$. В нашем примере при маятниковой колеблемости было бы 15 точек, при связанной с 11-летним циклом было бы $2 \cdot (17 : 11) \approx 3$ точки, при случайно распределенной во времени в среднем было бы $\frac{2}{3}(17-2)=10$ точек. Фактическое число точек 6 выходит за границы двукратного среднего квадратического отклонения числа поворотных точек, которое по Кендэлу равно $\sqrt{\frac{16-29}{90}}$ *, в нашем случае $\sqrt{\frac{16-17-29}{90}}=1{,}6$.

Наличие 6 точек, при 2 точках за цикл, означает, что в ряду могут быть примерно 3 цикла, продолжительность периода которых 5,5 – 6 лет. Возможно сочетание таких циклических колебаний со случайными.

Другой метод анализа типа колеблемости и поиска длины цикла основан на вычислении коэффициентов **автокорреляции отклонений от тренда.**

Автокорреляция – это корреляция между уровнями ряда или отклонениями от тренда, взятыми со сдвигом во времени: на 1 период (год), на 2, на 3 и т. д., поэтому говорят о коэффициентах автокорреляции разных **порядков:** первого, второго и т. д. Рассмотрим сначала коэффициент автокорреляции отклонений от тренда первого порядка.

Одна из основных формул для расчета коэффициента автокорреляции отклонений от тренда имеет вид:

$$r_u^{al}=\frac{\sum_{i=1}^{n-1}u_iu_{i+1}}{\frac{u_1^2}{2}+\sum_{i=2}^{n-1}u_i^2+\frac{u_n^2}{2}}. \qquad (9.36)$$

---

* *Юл. Дж. Э. Кендэл М.* Теория статистики. – М.: Госстатиздат, 1960. – С. 708.

Как легко видеть по табл. 9.7, первое и последнее в ряду отклонения участвуют только в одном произведении в числителе, а все прочие отклонения от второго до $(n-1)$-го – в двух. Поэтому и в знаменателе квадраты первого и последнего отклонений следует взять с половинным весом, как в хронологической средней. По данным табл. 9.7 имеем:

$$r_n^{a_1} = \frac{446{,}2}{60{,}5 + 1051{,}1 + 83{,}2} = 0{,}373.$$

Теперь обратимся к рис. 9.2. При маятниковой колеблемости все произведения в числителе будут отрицательными величинами, и коэффициент автокорреляции первого порядка будет близок к –1. При долгопериодических циклах будут преобладать положительные произведения соседних отклонений, а смена знака происходит лишь дважды за цикл. Чем длиннее цикл, тем больше перевес положительных произведений в числителе, и коэффициент автокорреляции первого порядка ближе к +1. При случайно распределенной во времени колеблемости знаки отклонений чередуются хаотически, число положительных произведений близко к числу отрицательных, ввиду чего коэффициент автокорреляции близок к нулю. Полученное значение говорит о наличии как случайно распределенных во времени колебаний, так и циклических. Коэффициенты автокорреляции следующих порядков: II = – 0,577; III = –0,611; IV = –0,095; V = +0,376; VI = +0,404; VII = +0,044. Следовательно, противофаза цикла ближе всего к 3 годам (наибольший отрицательный коэффициент при сдвиге на 3 года), а совпадающие фазы ближе к 6 годам, что и дает длину цикла колебаний. Эти максимальные по абсолютной величине коэффициенты не близки к единице. Это означает, что циклическая колеблемость смешана со значительной случайной колеблемостью. Таким образом, подробный автокорреляционный анализ в целом дал те же результаты, что и выводы по автокорреляции первого порядка.

Если динамический ряд достаточно длинен, можно поставить и решить задачу об изменении показателей колеблемости с течением времени. Для этого рассчитывают эти показатели по подпериодам, но длительностью не менее 9 – 11 лет, иначе измерения колеблемости ненадежны. Кроме того, можно рассчитывать показатели колеблемости скользящим способом, а затем произвести их выравнивание, т. е. вычислить тренд показателей колеблемости. Это полезно, чтобы сделать вывод о действенности мер, применявшихся для уменьшения колебаний урожайности и других нежелательных колебаний, а также для того, чтобы по тренду сделать прогноз ожидаемых в будущем размеров колебаний.

9.8. ИЗМЕРЕНИЕ УСТОЙЧИВОСТИ В ДИНАМИКЕ

Понятие «устойчивость» используется в весьма различных смыслах. По отношению к статистическому изучению динамики мы рассмотрим два аспекта этого понятия: 1) устойчивость как категория, противоположная колеблемости; 2) устойчивость направленности изменений, т. е. устойчивость тенденции.

В первом понимании показатель устойчивости, который может быть только относительным, должен изменяться от нуля до единицы (100%). Это разность между единицей и относительным показателем колеблемости. Коэффициент колеблемости составил 9,0%. Следовательно, коэффициент устойчивости равен 100% – 9,0% = 91,0%. Этот показатель характеризует близость фактических уровней к тренду и совершенно не зависит от характера последнего. Слабая колеблемость и высокая устойчивость уровней в данном смысле могут существовать даже при полном застое в развитии, когда тренд выражен горизонтальной прямой.

Устойчивость во втором смысле характеризует не сами по себе уровни, а процесс их направленного изменения. Можно узнать, например, насколько устойчив процесс сокращения удельных затрат ресурсов на производство единицы продукции, является ли устойчивой тенденция снижения детской смертности и т. д. С этой точки зрения полной устойчивостью направленного изменения уровней динамического ряда следует считать такое изменение, в процессе которого каждый следующий уровень либо выше всех предшествующих (устойчивый рост), либо ниже всех предшествующих (устойчивое снижение). Всякое нарушение строго ранжированной последовательности уровней свидетельствует о неполной устойчивости изменений.

Из определения понятия устойчивости тенденции вытекает и метод построения ее показателя. В качестве показателя устойчивости можно использовать *коэффициент корреляции рангов* Ч. Спирмэна (Spearman) – r_s

$$r_s = 1 - \frac{6\sum_{i=1}^{n}\Delta_i^2}{n^3 - n}, \qquad (9.37)$$

где n – число уровней;

Δ_i – разность рангов уровней и номеров периодов времени.

При полном совпадении рангов уровней, начиная с наименьшего, и номеров периодов (моментов) времени по их хронологическому порядку коэффициент корреляции рангов равен +1. Это значение

соответствует случаю полной устойчивости возрастания уровней. При полной противоположности рангов уровней рангам лет коэффициент Спирмэна равен –1, что означает полную устойчивость процесса сокращения уровней. При хаотическом чередовании рангов уровней коэффициент близок к нулю, это означает неустойчивость какой-либо тенденции. Приведем расчет коэффициента корреляции Спирмэна по данным о динамике индекса цен (табл. 9.7) в табл. 9.8.

Таблица 9.8

Расчет коэффициентов корреляции рангов Спирмена

Годы	Уровни, y_i	Ранг лет, P_x	Ранг уровней, P_y	$P_x - P_y$	$(P_x - P_y)^2$
1979	105	1	8	7	49
1980	111	2	13	11	121
1981	110	3	12	9	81
1982	106	4	9,5	5,5	30,25
1983	118	5	16	11	121
1984	124	6	17	11	121
1985	113	7	14,5	7,5	56,25
1986	92	8	3,5	4,5	20,25
1987	91	9	1,5	7,5	56,25
1988	109	10	11	1	1
1989	113	11	14,5	3,5	12,25
1990	100	12	6	6	36
1991	94	13	5	8	64
1992	91	14	1,5	12,5	156,25
1993	92	15	3,5	11,5	132,25
1994	102	16	7	9	81
1995	106	17	9,5	7,5	56,25
Σ	1777	–	–	–	1141

Ввиду наличия трех пар «связанных рангов» применяем формулу (8.26):

$$r_s = 1 - \frac{6\sum_{i=1}^{n}(R_x - R_y)^2 - A}{\sqrt{(n^3 - n)\cdot(n^3 - n - A)}} = 1 - \frac{6\cdot 1141 - 2}{\sqrt{(17^3 - 17)\cdot(17^3 - 17 - 2)}} = -0{,}398.$$

Отрицательное значение r_s, указывает на наличие тенденции снижения уровней, причем устойчивость этой тенденции ниже средней.

При этом следует иметь в виду, что даже при 100%-ной устойчивости тенденции в ряду динамики может быть колеблемость уровней, и коэффициент их устойчивости будет ниже 100%. При слабой колеблемости, но еще более слабой тенденции, напротив, возможен высокий коэффициент устойчивости уровней, но близкий к нулю коэффициент устойчивости тренда. В целом же оба показателя связаны, конечно, прямой зависимостью: чаще всего большая устойчивость уровней наблюдается одновременно с большей устойчивостью тренда.

Устойчивость тенденции развития или комплексная устойчивость в динамике может быть охарактеризована соотношением между среднегодовым абсолютным изменением и средним квадратическим (либо линейным) отклонением уровней от тренда:

$$c = \frac{b}{s(t)}. \qquad (9.38)$$

Если, как нередко бывает, распределение отклонений уровней ряда от тренда близко к нормальному, то с вероятностью 0,95 отклонение от тренда вниз не превысит $1{,}645s(t)$ по величине. Следовательно, если в ряду динамики $c > 1{,}64$, то уровни, более низкие, чем предыдущие, в среднем будут встречаться менее 5 раз за 100 периодов, или 1 раз из 20, т. е. устойчивость тренда будет высока. При $c = 1$ нарушения ранжированности уровней будут встречаться в среднем 16 раз из 100, а при $c = 0{,}5$ – уже 31 раз из 100, т. е. устойчивость тенденции будет низкой. Можно также пользоваться отношением среднего темпа прироста к коэффициенту колеблемости, что дает показатель, близкий к c – показателю устойчивости. Этот показатель более пригоден для экспоненциального тренда. О показателях устойчивости нелинейных трендов и об общих проблемах устойчивости экономических и социальных процессов можно подробнее прочесть в рекомендуемой к данной главе литературе [2].

9.9. СЕЗОННЫЕ КОЛЕБАНИЯ И ПОЛНОЕ РАЗЛОЖЕНИЕ ДИСПЕРСИИ УРОВНЕЙ ДИНАМИЧЕСКОГО РЯДА

Сезонными называют периодические колебания, возникающие под влиянием смены времени года. Их роль очень велика в агропромышленном комплексе, торговле многими товарами, заболеваемости, строительстве, деятельности рекреационных учреждений, на

транспорте. Сезонные колебания строго цикличны – повторяются через каждый год, хотя сама длительность времен года имеет колебания. Для изучения сезонных колебаний необходимо иметь уровни за каждый квартал, а лучше за каждый месяц, иногда даже за декады, хотя декадные уровни могут уже сильно исказиться мелкомасштабной случайной колеблемостью.

Следует еще раз указать, что не всякие различия в месячных или квартальных уровнях являются сезонными колебаниями, а только регулярно повторяющиеся год за годом. Если же различия месячных уровней или любых внутригодичных уровней в один год распределены совершенно иначе, чем в другой год, то это – не сезонные, а случайные колебания т. е. колебания, вызванные причинами, не связанными со сменой времен года. Например, такими могут быть колебания курсов акций, обменных курсов валют, вызванные изменением финансовой политики государства, научно-техническими открытиями, политическими кризисами в стране и мире, слиянием и разделением компаний и т. п.

Поскольку интервальные уровни зависят от длительности интервалов времени, а длина месяцев не равная, точнее проводить анализ сезонных колебаний не по фактическим месячным уровням, а по уровням, пересчитанным на равную (30-дневную) длительность всех месяцев или среднесуточным. Если изучаются сезонные колебания за отдельный год, то обычно тренд не принимается во внимание, и отклонения месячных (30-дневных) уровней исчисляются от среднемесячного уровня за год. Кроме рассмотренных показателей колеблемости для характеристики сезонных колебаний важное значение имеет форма сезонной «волны», изучаемая с помощью относительных показателей – отношений месячных уровней к среднемесячному (так называемый «индекс сезонности»). Лучше, конечно, изучать сезонные колебания за несколько лет, чтобы сгладить случайные колебания и точнее измерить сезонные. Рассмотрим сезонность смертности в Санкт-Петербурге за 1994 – 1996 гг. (табл. 9.9).

Итак, по данным табл. 9.9 смертность возрастает зимой – в январе, феврале, затем убывает и достигает минимума в августе, после чего возрастает, со странным исключением в ноябре.

Среднее линейное отклонение по модулю составляет 8,76; среднее квадратическое отклонение $s(t)\sqrt{\frac{1792,24}{12}}=12,22$; коэффициент колеблемости $v(t)=\frac{12,22}{208,8}=0,0585$. Таким образом, сезонная колеблемость смертности в Санкт-Петербурге слабая. Примененная методика не является оптимальной: не учтено наличие тренда, в данном случае – тенденции снижения числа умерших.

Таблица 9.9

Сезонность смертности в Санкт-Петербурге

Месяц	Число умерших в среднем за сутки				В % к среднемесячному показателю	Отклонения от средней, $y_i-\bar{y}$	$(y_i-\bar{y})^2$
	1994	1995	1996	в среднем за 3 года			
Январь	246,8	229,5	243,0	239,8	114,8	31,0	961,00
Февраль	258,9	212,3	209,7	227,0	108,7	18,2	331,24
Март	225,5	220,7	187,6	211,3	101,2	2,5	6,25
Апрель	211,9	212,3	182,8	202,3	96,9	–7,2	51,84
Май	231,0	208,7	180,6	206,8	99,0	–2,0	4,00
Июнь	235,4	205,7	173,1	204,7	98,0	–4,1	16,81
Июль	227,4	211,3	181,8	206,8	99,0	–2,0	4,00
Август	220,6	204,5	171,4	198,8	95,2	–10,0	100,00
Сентябрь	233,5	193,0	178,8	201,8	96,6	–7,0	49,00
Октябрь	229,3	196,0	186,5	203,9	97,7	–4,9	24,01
Ноябрь	212,5	196,7	170,4	193,2	92,5	–15,6	243,36
Декабрь	218,3	228,6	181,5	209,5	100,3	0,7	0,49
Средняя	229,1	210,0	187,3	208,8	100,0	–	1792,24

Наиболее точную и полную методику анализа с разложением ряда динамики на три компонента: тренд, сезонную колеблемость и случайную колеблемость рассмотрим на примере динамики импорта КНР по кварталам за 1993 – 1995 гг. (табл. 9.10).

Анализ в табл. 9.10 проводится по следующей методике:

1. По месячным или квартальным данным за все годы вычисляется уравнение тренда и выравненные по нему уровни, обозначаемые $\hat{y}_{ij}$, где i – номер года; j – номер квартала или месяца.

2. Каждый фактический уровень делится на соответствующий выравненный для расчета индексов сезонности C_{ij}.

Таблица 9.10

Анализ сезонности импорта КНР*

(млрд долл. США)

Годы	Кварталы	y_{ij}	t_{ij}	$y_{ij} \cdot t_{ij}$	$\hat{y}_{ij}$	C_{ij}	$\hat{y}'_{ij}$	$u_{сезij}$	$u_{случij}$	$u_{сез-случij}$
	I	15,8	–6	–94,8	21,0	0,7524	16,1	–4,9	–0,3	–5,2
	II	21,5	–5	–107,5	22,0	0,9773	22,0	0	–0,5	–0,5
1993	III	24,8	–4	–99,2	22,9	1,0830	23,6	+0,7	+1,2	+1,9
	IV	33,1	–3	–99,3	23,9	1,3849	30,5	+6,6	+2,6	+9,2
	I	18,7	–2	–37,4	24,9	0,7510	19,1	–5,8	–0,4	–6,2
	II	26,4	–1	–26,4	25,9	1,0193	25,9	0	+0,5	+0,5
1994	III	26,5	0	0	26,8	0,9888	27,6	+0,8	–1,1	–0,3
	IV	34,5	1	34,5	27,8	1,2410	35,5	+7,7	–1,0	+6,7
	I	21,9	2	43,8	28,8	0,7604	22,1	–6,7	–0,2	–6,9
	II	30,0	3	90,0	29,8	1,0067	29,8	0	+0,2	+0,2
1995	III	31,2	4	124,8	30,7	1,0163	31,6	+0,9	–0,4	+0,5
	IV	38,1	5	190,5	31,7	1,2019	40,4	+8,7	–2,3	+6,4
1996	I	26,4	6	158,4	32,7	0,8073	25,1	–7,6	+1,3	–6,3
Σ	–	348,9	0	177,4	348,9	1	349,3*	0,4	0,4	0,4

* Расхождение на 0,4 с суммой фактических уровней за счет округления до 0,1

3. Индексы сезонности усредняются за все годы, получаем средние индексы сезонности для каждого квартала или месяца:

$$\bar{c}_j = \frac{\sum_{i=1}^{k} c_{ij}}{k}, \tag{9.39}$$

где i – номер года;
k – число лет;
j – номер квартала, месяца.

4. Выравненные уровни умножаются на средние индексы сезонности для соответствующих кварталов или месяцев, получаем выравненные уровни с учетом сезонности $\hat{y}'_{ij}$

$$\hat{y}'_{ij} = \hat{y}_{ij} \cdot \bar{c}_j. \tag{9.40}$$

5. Вычисляются отклонения (и их квадраты) за счет сезонности:

$$u_{сезij} = \hat{y}'_{ij} - \hat{y}_{ij}. \tag{9.41}$$

6. Вычисляются отклонения (и их квадраты) за счет случайной колеблемости:

$$u_{случij} = y_{ij} - \hat{y}'_{ij}. \tag{9.42}$$

7. Вычисляются общие отклонения:

$$u_{сез} + u_{случij} = y_{ij} - \hat{y}_{ij}. \tag{9.43}$$

Уровни в табл. 9.10 – это объем импорта в КНР по кварталам (в ценах *fob*, т. е. «franco board» – с учетом затрат на погрузку на борт корабля или в вагоны, на грузовики, но не включая стоимость перевозки, фрахта)[1].

Прежде всего отметим, что при наличии существенных сезонных колебаний параметры тренда будут вычислены правильно только при условии, что первый и конечный уровни относятся к одному и тому же кварталу (месяцу), иначе сезонность исказит параметры тренда. Поэтому, в расчет включаем и I квартал 1996 г.

Графическое изображение динамического ряда с наличием сезонных колебаний возможно двумя способами. **Первый способ** –

[1] International Financial Statistics. – Washington: Издание. International Monatary Fund. – 1996. – P. 178–179.

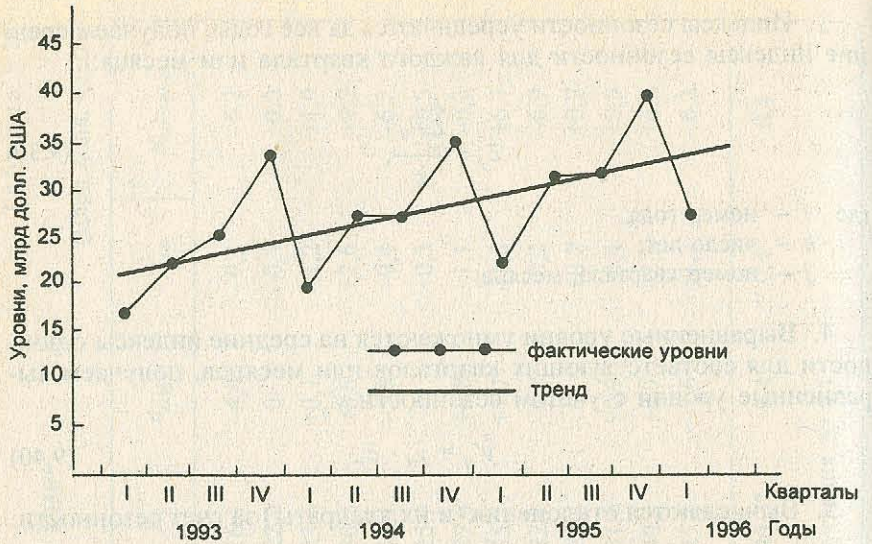


Рис. 9.3. Динамика импорта КНР, млрд долл. США

обычная линейная диаграмма в декартовой системе координат, аналогичная рис. 9.3.

Второй способ – изображение в полярных координатах. Величина уровня изображается расстоянием от центра, между месяцами угол 30°, между кварталами – 90°. График имеет вид разворачивающейся спирали, если тренд направлен к возрастанию, и сворачивающейся, если тренд направлен к уменьшению уровней. Сезонные колебания выражены тем, что точки четвертых кварталов далеко выходят за окружность, радиус которой – средний уровень 1993 г., а точки первых кварталов – внутри нее (рис. 9.4).

Средний уровень ряда

$$\bar{y} = \sum_{i=1}^{k}\sum_{j=1}^{n} y_{ij} : k \cdot n = 348{,}9 : 13 = 26{,}84.$$

Тренд имеет линейную форму со среднегодовым абсолютным приростом

$$b = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n} y_{ij} \cdot t_{ij}}{\sum_{i=1}^{k}\sum_{j=1}^{n} t_{ij}^2} = \frac{177{,}4}{182} = 0{,}9747.$$

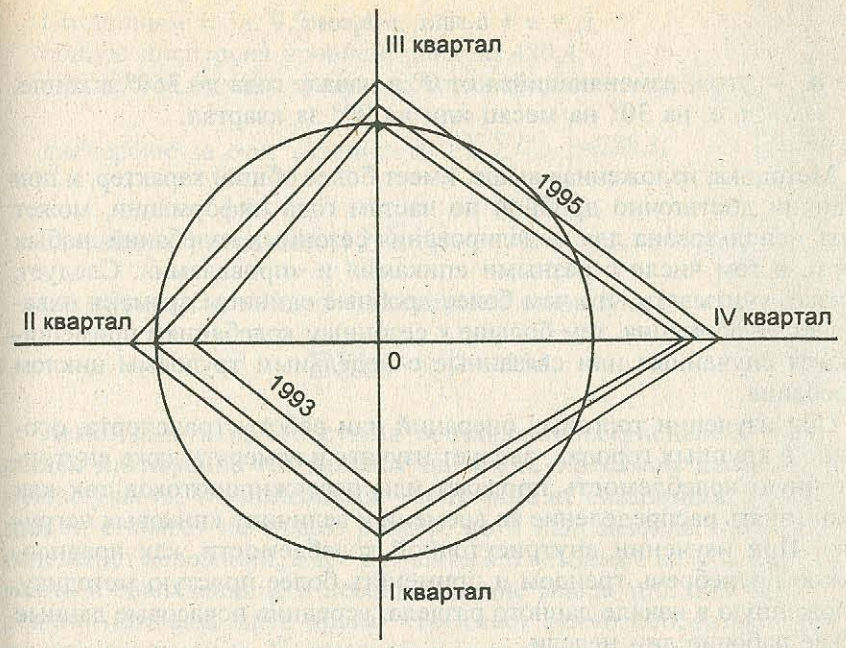


Рис. 9.4. Сезонность импорта КНР, млрд долл. США

Средние индексы сезонности за 3 года (для I квартала – за 4 года) составляют:

Квартал	c_j
I	0,7678
II	1,0011
III	1,0294
IV	1,2759

Таким образом, наблюдается сезонный спад импорта в I квартале на 23 с лишним процента и подъем в конце года, а уровни II и III кварталов почти равны средним квартальным значениям. Такая форма сезонных колебаний весьма далека от плавной синусоидальной кривой и к ней неприменима модель тригонометрического вида:

$$\tilde{y}_j = a + b_1 \sin\alpha_j + b_2 \cos\alpha_j,$$

где α_j – угол, изменяющийся от 0° в начале года до 360° в конце, т. е. на 30° на месяц или на 90° за квартал.

Методика, изложенная выше, имеет более общий характер, и при наличии достаточно дробной по частям года информации, может быть использована для моделирования сезонных колебаний любых форм, в том числе с разными «пиками» и «провалами». Следует, однако, учитывать, что чем более дробные единицы времени охватывает информация, тем больше к сезонным колебаниям примешиваются случайные или связанные с недельным трудовым циклом колебания.

При изучении торговых операций или работы транспорта, особенно в крупных городах, следует изучать и измерять даже внутрисуточную колеблемость продажи или пассажиропотоков так как важно знать распределение во времени и величину «пиковых нагрузок». При изучении внутрисуточной колеблемости, как правило, можно пренебречь трендом и применять более простую методику, изложенную в начале данного раздела, усреднив почасовые данные за все рабочие дни недели.

Общую дисперсию уровней динамического ряда, измеряемую суммой квадратов отклонений этих уровней от их средней величины $\sum_{j=1}^{m}\sum_{i=1}^{k}(y_{ij}-\bar{y})^2$ можно разложить на составляющие:

1. Дисперсия за счет тренда:

$$\sum_{j=1}^{m}\sum_{i=1}^{k}(\tilde{y}_{ij}-\bar{y}_{ij})^2 = D_t. \qquad (9.44)$$

2. Дисперсия за счет сезонных колебаний:

$$\sum_{j=1}^{m}\sum_{i=1}^{k}(\tilde{y}'_{ij}-\tilde{y}_{ij})^2 = D_{сез}. \qquad (9.45)$$

3. Дисперсия за счет случайных колебаний (остаточная):

$$\sum_{i=1}^{k}\sum_{j=1}^{m}(y_{ij}-\tilde{y}'_{ij})^2 = D_{ост}. \qquad (9.46)$$

$$D_{случ+сез} = \sum_{i=1}^{k}\sum_{j=1}^{m}(y_{ij}-\tilde{y}_{ij})^2. \qquad (9.47)$$

По данным табл. 9.9 имеем:
общую дисперсия уровней, равную 480,4;
дисперсию за счет тренда, которая составляет 168,1;

дисперсию за счет сезонности $\left(\sum_i\sum_j U^2_{сез}\right) = 389{,}3$;

случайную дисперсию $\left(\sum_i\sum_j U^2_{случ}\right) = 18{,}4$;

дисперсию случайную + дисперсию сезонную:

$$\sum_i\sum_j(y_{ij}-\hat{y}_{ij})^2 = 288{,}2.$$

Легко заметить, что сумма составляющих дисперсий больше общей дисперсии, что кажется ошибкой. На самом деле, однако, нужно учесть, что **колебания – величина не скалярная, а векторная,** т. е. имеет не только размер, но и направление, знак. Тренд отделен от колебаний, а все случайные и сезонные колебания могут иметь и совпадающие и несовпадающие знаки, т. е. они могут частично погашать друг друга, что имеет место особенно в конце изучаемого периода. Поэтому общая колеблемость, измеряемая суммой квадратов отклонений (9.47) значительно меньше, чем сумма дисперсий за счет сезонной и случайной колеблемости. По данным табл. 9.10 общая колеблемость составила 288,2. Находим отношение этой величины к сумме сезонной и случайной дисперсий:

$$288{,}2 : (389{,}3 + 18{,}4) = 0{,}70706.$$

На эту величину корректируем сезонную и случайную суммы квадратов отклонений и окончательно получаем следующее разложение общей дисперсии уровней ряда (табл. 9.11)

Из табл. 9.11 следует ряд выводов: основным источником различия квартальных уровней импорта КНР за изучаемый период времени являлась сезонная колеблемость. Случайная колеблемость существенной роли не играла. Проверка существенности различий по критерию Фишера показала, что и тренд и сезонная колеблемость существенны, как и различия уровней в целом. Табличное значение F в несколько раз меньше фактических, так что вероятность существенности различий много ближе к единице, чем к 0,95, для которой приведены табличные значения F. Отметим, что при изучении сезонных колебаний по месячным уровням, сезонная дисперсия будет иметь (12–1) степень свободы. Сумма степеней

Таблица 9.11

Разложение суммы квадратов отклонений уровней динамического ряда от средней на составляющие

Источник дисперсии	Сумма квадратов отклонений		Число степеней свободы	Дисперсия на 1 степень свободы	Значения F-критерия	
	величина	доля, %			фактическое	табличное с вероятностью 0,95
Общая дисперсия	480,4	100,0	12	40	21,1	3,28
в том числе:						
тренд	168,1	36,8	1	168,1	103,7	5,32
сезонность	275,2	60,3	3	91,7	56,6	4,07
случайная (остаточная) колеблемость	13,0	2,9	8	1,62	1	–

свободы сезонной и случайной дисперсий равна числу уровней ряда за вычетом числа параметров тренда.

Сделаем прогноз объема импорта КНР с учетом тренда и сезонности на IV квартал 1997 г. Уровень тренда

$$\hat{y}_{1997,\,IV} = 26{,}84 + 0{,}9747 \cdot 13 = 39{,}51.$$

Умножим уровень тренда на средний индекс сезонности IV квартала:

$$39{,}51 \cdot 1{,}2759 = 50{,}41 \text{ (млрд долл. США).}$$

Смысл прогноза в том, что при сохранении до конца 1997 г. измеренного за 1993 – 1995 гг. тренда и характера сезонных колебаний, импорт составит 50,41 млрд долл. США. Это точечный прогноз. Проблема измерения средней ошибки прогноза с учетом тренда и сезонности сложна и здесь не излагается.

Иногда полученные удельные веса составляющих в общей сумме квадратов рассматривают как характеристики роли разных комплексов факторов в развитии изучаемого объекта. К таким оценкам следует подходить очень осторожно. Дело в том, что различия уровней за счет тренда с течением времени накапливаются, и чем больший период времени подвергается анализу, тем более значительной становится роль комплекса факторов, обусловливающих тенденцию динамики в сравнении с факторами колеблемости, не имеющей кумулятивного эффекта.

9.10. ПРОГНОЗИРОВАНИЕ НА ОСНОВЕ ТРЕНДА И КОЛЕБЛЕМОСТИ

Прогнозирование возможных в будущем значений признаков изучаемого объекта – одна и основных задач науки. В ее решении роль статистических методов очень значительна. Одним из статистических методов прогнозирования является расчет прогнозов на основе тренда и колеблемости динамического ряда до настоящего времени. Если мы будем знать, как быстро и в каком направлении изменились уровни какого-то признака, то сможем узнать, какого значения достигнет уровень через известное время. Методика статистического прогноза по тренду и колеблемости основана на их **экстраполяции,** т.е. на предположении, что параметры тренда и колебаний сохраняются до прогнозируемого периода. Такая экстраполяция справедлива, если система развивается эволюционно в достаточно стабильных условиях. Чем крупнее система, тем более вероятно сохранение параметров ее изменения, конечно, на срок не слишком большой! Обычно рекомендуют, чтобы срок прогноза не превышал одной трети длительности базы расчета тренда.

В отличие от прогноза на основе регрессионного уравнения прогноз по тренду учитывает факторы развития только в неявном виде, и это не позволяет «проигрывать» разные варианты прогнозов при разных возможных значениях факторов, влияющих на изучаемый признак. Зато прогноз по тренду охватывает все факторы, в то время как в регрессионную модель невозможно включить в явном виде более 10-20 факторов в самом лучшем случае.

Сущность прогноза на основе тренда хорошо иллюстрируется следующим рассказом о греческом философе Диогене, жившем в большой бочке на берегу Саронического залива недалеко от афинского порта Пирея. Как-то вечером Диогена стал окликать снаружи неизвестный. Диоген вышел в нему. – «Скажи, мудрый человек», – спросил путник, – дойду ли я к закату в Афины?» Диоген посмотрел на него и сказал: – «Иди!» Путник повторил свой вопрос... – «Иди!» – закричал Диоген, и путник, пожав плечами, побрел по берегу. – «Вернись!» – снова закричал Диоген, и путник вернулся к нему. – «Вот теперь я тебе скажу, что до заката ты не дойдешь до

Афин. Оставайся у меня». – «А почему же ты сразу мне это не сказал, а прогнал меня?» Диоген усмехнулся: – «А как же я скажу, дойдешь ли ты до Афин, если я не видел, как быстро ты ходишь?» Прогноз по тренду – это и есть Диогенов прогноз на основании знания того, как изучаемая система «шла» до настоящего времени.

Рассмотрим методику прогнозирования по тренду с учетом колеблемости на примере цен на нетопливные товары развивающихся стран, тренд и колеблемость которых была измерена в параграфах 9.6 и 9.7 (табл. 9.4 и 9.7). За основу прогнозов возьмем параметры, полученные методом многократного скользящего выравнивания. Параллельно будет показана и методика расчетов при однократном выравнивании.

Итак, имеем уравнение тренда $\hat{y} = 104{,}53 - 1{,}433\ t$, где $t = 0$ в 1987 г., оценку генеральной величины среднего квадратического отклонения от тренда $s(t) = 9{,}45$. Эти значения получены при анализе динамики цен весьма значительного сектора мировой торговли, т. е. очень большой и сложной системы. Маловероятно, что условия развития этой системы существенно изменятся, скажем, до 1998 г. Поэтому, прогноз на 1998 г. по измеренному тренду можно теоретически считать достаточно обоснованным. Обычно рекомендуется, чтобы период упреждения (от конца базы расчета до прогнозируемого периода) составлял не более трети длины базы расчета.

Прежде всего, вычисляется «*точечный прогноз*» – значение уровня тренда при подстановке в его уравнение номера 1998 г., считая от 1987 г., т. е. $t_k = 11$.

$$\hat{y}_{1988} = 104{,}53 - 1,\ 433 \cdot 11 = 88{,}77.$$

Это означает, что наиболее вероятное значение индекса цен на нетопливные товары развивающихся стран в 1998 г. составит около 89% к уровню цен 1990 г., принятому за 100%. Однако, параметры тренда, полученные по ограниченному числу уровней ряда – это лишь выборочные средние оценки, не свободные от влияния распределения колебаний отдельных уровней во времени, как уже сказано ранее. При изменении базы расчета тренда, если, скажем взять 1977 – 1993 гг. или 1981–1997 гг., были бы получены несколько иные значения параметров, а значит, и другие значения $\hat{y}_{1998}$. Прогноз должен иметь вероятностную форму, как всякое суждение о будущем.

Средняя ошибка прогноза положения линейного тренда на год (момент) с номером t_k вычисляется по формулам:

А) • Для однократного выравнивания:

$$m_{\hat{y}_k} = s(t) \cdot \sqrt{\frac{1}{N} + \frac{t_k^2}{\sum t_i^2}} \qquad (9.48)$$

где t_k – номер года прогноза,

$\sum t_i^2$ – по всей длине ряда N, т. е. $\dfrac{N^3 - N}{12}$.

Б) • Для многократного скользящего выравнивания при l сдвигах базы и длине ее n:

$$m_{\hat{y}_k} = s(t) \cdot \sqrt{\frac{1}{N} + \frac{t_k^2}{l \cdot \sum t_i^2}}, \qquad (9.49)$$

где $\sum t_i^2 = \dfrac{n^3 - n}{12}$.

При $N = 17$, $n = 11$, $l = 7$ получаем:

$m_{\hat{y}_k}$ для однократного выравнивания $= 9{,}133 \cdot \sqrt{\dfrac{1}{17} + \dfrac{11^2}{408}} = 5{,}44$

$m_{\hat{y}_k}$ для многократного выравнивания $= 9{,}45 \cdot \sqrt{\dfrac{1}{17} + \dfrac{11^2}{7 \cdot 110}} = 4{,}39.$

Как видим, метод многократного выравнивания на 20% снизил среднюю ошибку прогноза положения тренда.

Для получения достаточно надежных границ прогноза положения тренда, скажем, с вероятностью 0,9 того, что ошибка будет не более указанной, следует среднюю ошибку умножить на величину t-критерия Стьюдента при указанной вероятности (или значимости $1 - 0{,}9 = 0{,}1$) и при числе степеней свободы, равном, для линейного тренда, $N - 2$, т. е. 15. Эта величина равна 1,753. Получаем предельную с данной вероятностью ошибку

$$\alpha = m \cdot t_{Стьюдента} = 4{,}39 \cdot 1{,}753 = 7{,}70.$$

Следовательно, с вероятностью 0,9 можно ожидать, что тренд индекса цен в 1988 г. пройдет между значениями $\hat{y}_{1998+\alpha}$ и $\hat{y}_{1998-\alpha}$, т. е. 88,77 + 7,70 и 88,77 – 7,70; от 81,07 до 96,47 в процентах к

уровню цен 1990 г. и, конечно, в одинаковой валюте, без учета ее инфляции.

Однако, фактические уровни ряда отклоняются от тренда. Уровень цен в 1998 г. также может быть вовсе не равен уровню положения тренда в этом году. Ошибка прогноза конкретного уровня включает две неопределенности: во-первых, мы не знаем точно, где окажется тренд в 1998 г., а во-вторых – в какую сторону и на сколько уровень ряда отклонится в 1998 г. от положения тренда. Считая, как уже было сказано, колебания случайно (в основном, случайно) распределенными во времени, т. е. независимыми от тренда, определим ошибку прогноза уровня конкретного года по правилу сложения независимых дисперсий.

$$m_{y_k} = \sqrt{m_{\hat{y}_k}^2 + s_{(t)}^2} \qquad (9.50)$$

$$m_{y_k} \sqrt{4{,}39^2 + 9{,}45^2} = 10{,}42.$$

С вероятностью 0,9 ошибка прогноза уровня цен не превзойдет величины 18,27 (10,42 · 1,753) и доверительные границы прогноза. составят от 70,5 до 107,0% к уровню 1990 г. Как видим, точность прогноза, невелика, разброс возможных значений достиг 37 пунктов, а вероятная ошибка составила 0,206 или 20,6% $\left[\frac{18{,}27}{88{,}77}\right]$ от средней величины (точечного прогноза). Можно уменьшить значение, вероятной ошибки, если сделать прогноз с меньшей надежностью, скажем, с вероятностью 0,75. Тогда значение *t*-критерия Стьюдента составит 1,197, вероятная ошибка составит 12,47 пункта, [10,42 · 1,197] доверительные границы – от 76,30 до 101,24 % к уровню 1990 г. За уменьшение вероятной ошибки, однако, пришлось заплатить снижением надежности прогноза.

Из имеющейся информации нельзя извлечь больше, чем в ней содержится: как в физике действует закон сохранения массы и энергии, импульса («количества движения»), так здесь действует закон сохранения информации: увеличивая точность, мы понижаем надежность, увеличивая надежность – понижаем точность. Методика анализа и прогнозирования тоже имеет значение. Она определяет степень полноты извлечения информации, содержащейся в исходном ряду динамики. С помощью методики многократного выравнивания удается более полно извлечь информацию о тренде и уменьшить среднюю ошибку прогноза его положения в прогнозируемом периоде с 5,44 до 4,39. Однако, как видно из (9.50), главной

составляющей ошибки прогноза конкретного уровня в нашем расчете является не ошибка прогноза положения тренда, а колеблемость уровней около тренда. Поэтому ошибка прогноза конкретного уровня незначительно сократилась за счет многократного выравнивания. При слабой колеблемости уровней и прогнозировании на значительное удаление от базы, главную роль станет играть ошибка положения тренда. Тогда многократное выравнивание даст значительное сокращение средней ошибки прогноза конкретных уровней. Но в любом случае эта ошибка всегда больше показателя колеблемости уровней – среднего квадратического отклонения $S_{y(t)}$[1]. В частности, в указанной литературе содержатся формулы для вычисления средней ошибки прогноза положения линии тренда при параболической и экспоненциальной его формах[2]. Если средняя ошибка положения тренда вычислена, ошибку конкретного уровня при любой форме тренда вычисляют по формуле (9.50).

9.11. КОРРЕЛЯЦИЯ РЯДОВ ДИНАМИКИ

В главах, посвященных статистическому изучению взаимосвязей методом аналитической группировки и методом корреляционного анализа, рассматривались зависимости между признаками, варьирующими в пространственной совокупности. Но необходимо изучать и связи, проявляющиеся в развитии, во времени. Например, есть ли связь между изменениями урожайности сельскохозяйственных культур и изменениями ее себестоимости, рентабельности? Есть ли связь между динамикой рождаемости и динамикой обеспеченности населения жильем? К сожалению, проблема изучения причинных связей во времени очень сложна, и полное решение всех задач такого изучения до сих пор не разработано.

Характерным примером для иллюстрации особенностей методики анализа корреляции в рядах динамики служит связь динамики урожайности сельскохозяйственных культур с себестоимостью продукции в 70 – 80-е гг. в СССР. Официально тогда, не признавалось наличие инфляции. Однако, даже в тех хозяйствах, где агротехника прогрессировала и урожайность имела тенденцию роста, себестоимость продукции тоже возрастала. Такой пример представлен в табл.9.12.

---

[1] *Четыркин Е. М.* Статистические методы прогнозирования. – М.: Статистика, 1977.

[2] *Юзбашев М. М., Манелля А. И.* Статистический анализ тенденций и колеблемости. – М.: Финансы и статистика, 1983.

Таблица 9.12

Корреляция урожайности картофеля с себестоимостью
Совхоз им. Ленина Волосовского р-на Ленинградской области

Годы	Урожайность, ц/га x_i	Себестоимость, руб./ц y_i	Отклонения от средних		$\Delta^2 x_i$	$\Delta^2 y_i$	$\Delta x_i \Delta y_i$	Тренды		Отклонения от трендов		$u^2_{x_i}$	$u^2_{y_i}$	$u_{x_i} u_{y_i}$
			Δx_i	Δy_i				$\hat{x}$	$\hat{y}$	u_{x_i}	u_{y_i}			
1977	108	11,8	–12	–7,2	144	51,84	+86,4	97	+11,7	+11	+0,1	121	0,01	+1,1
1978	81	15,4	–39	–3,6	1521	12,96	+140,4	101	12,9	–20	+2,5	400	6,25	–50,0
1979	106	13,0	–14	–6,0	196	36,00	+84,0	105	14,1	+1	–1,1	1	1,21	–1,1
1980	124	13,9	+4	–5,1	16	26,01	–20,4	108	15,3	+16	–1,4	256	1,96	–22,4
1981	103	15,1	–17	–3,9	289	15,21	+66,3	112	16,6	–9	–1,5	81	2,25	+13,5
1982	106	19,6	–14	+0,6	196	0,36	–8,4	116	17,8	–10	+1,8	100	3,24	–18,0
1983	149	16,2	+29	–2,8	841	7,84	–81,2	120	19,0	+29	–2,8	841	7,84	–81,2
1984	148	17,2	+28	–1,8	784	3,24	50,4	124	20,2	+24	–3,0	576	9,00	–72,0
1985	102	24,0	–18	+5,0	324	25,0	–90,00	128	21,4	–26	+2,6	676	6,76	–67,6
1986	130	22,4	+10	+3,4	100	11,56	+34,0	131	22,7	–1	–0,3	1	0,09	+0,3
1987	80	32,3	–40	+13,3	1600	176,89	–532,0	135	23,9	–55	+8,4	3025	70,56	–462,0
1988	139	24,7	+19	+5,7	361	32,49	+108,3	139	25,1	0	–0,4	0	0,16	0
1989	183	21,4	+63	+2,4	3969	5,76	+151,2	143	26,3	+40	–4,9	1600	24,01	–196,0
Σ	1559	247,0	–	–	10341	405,16	–111,8	1559	247,0	–	–	7678	133,34	–952,7

Основная сложность состоит в том, что, как показано в предыдущем разделе главы, при наличии тренда за достаточно длительный период большая часть суммы квадратов отклонений связана с трендом. Если два признака имеют тренды с одинаковым направлением изменения уровней, то между уровнями этих признаков будет наблюдаться положительная ковариация. И в одном, и в другом ряду уровни более поздних лет будут либо больше, либо меньше уровней более ранних периодов. Коэффициент корреляции уровней окажется положительным. При разной направленности трендов ковариация уровней и коэффициент их корреляции окажутся отрицательными.

Но ведь одинаковая направленность трендов вовсе не означает причинной зависимости. Например, рост производства ракет не причина происходившего в тот же период роста производства мяса. Гораздо вероятнее, что при отсутствии гонки производства ракетного оружия производство мяса росло бы значительно быстрее. А коэффициенты корреляции уровней высоки! Таким образом, не только, возникает масса «ложных корреляций», за которыми нет причинной зависимости, но искажаются (преувеличиваются или преуменьшаются) и те показатели корреляции, за которыми стоят реальные причинные зависимости.

Рассмотрим табл. 9.12. Корреляция уровней урожайности с уровнями себестоимости картофеля отсутствует: коэффициент корреляции равен –0,055, т. е. незначимо отличен от нуля. Но ведь на самом деле по законам экономики, при пространственной корреляции в совокупности хозяйств связь урожайности и себестоимости сильная, обратная.

Среднее значение урожайности по данным табл. 9.12 составило $\bar{x}$ = 119,92 ц/га, себестоимость $\bar{y}$ = 19,0 руб./ц. Уравнения трендов урожайности $\hat{x}$ = 119,9 + 3,81t, себестоимости $\hat{y}$ = 19,0 + 1,22t, t = 0 в 1983 г.

Всесторонний экономический и статистико-математический анализ ситуации показывает, что причина отсутствия корреляции уровней в том, что оба признака имеют одинаково направленные тренды – возрастание урожайности происходило параллельно с возрастанием себестоимости, вовсе не являясь причиной последнего! Себестоимость росла из-за инфляции в стране, влияние которой оказалось сильнее, чем направленное на снижение себестоимости влияние роста урожайности.

Если же рассматривать уровни признаков год за годом, легко заметить, что без исключений снижению урожайности в сравнении с предыдущим годом соответствовал рост себестоимости, а повышению

урожайности – ее снижение, т. е. связь обратная, которая и должна быть. Следовательно, чтобы получить реальные показатели корреляции, необходимо абстрагироваться от искажающего влияния трендов: вычислить отклонения уровней урожайности и себестоимости от трендов и измерить корреляцию не уровней, а колебаний двух признаков.

Подставляя в формулу парного коэффициента корреляции (8.11) вместо уровней признаков их отклонения от трендов u_{x_i}, u_{y_i}, получаем:

$$r_{u_x u_y} = \frac{\sum\limits_{(i)} (u_{x_i} - \overline{u}_x)(u_{y_i} - \overline{u}_y)}{\sqrt{\sum\limits_{(i)} (u_{x_i} - \overline{u}_x)^2 \sum\limits_{(i)} (u_{y_i} - \overline{u}_y)^2}}. \qquad (9.51)$$

Однако среднее отклонение от тренда равно нулю (для прямой и параболы всегда, а для других типов тренда лишь в том случае, если правильно отражают тенденцию), $\overline{u}_x = \overline{u}_y = 0$. Подставив в (9.51), получим:

$$r_{u_x u_y} = \frac{\sum\limits_{(i)} u_{x_i} u_{y_i}}{\sqrt{\sum\limits_{(i)} u_{x_i}^2 \sum\limits_{(i)} u_{y_i}^2}}. \qquad (9.52)$$

Коэффициент регрессии для линейной зависимости принимает вид:

$$b = \frac{\sum\limits_{(i)} u_{x_i} u_{y_i}}{\sum\limits_{(i)} u_{x_i}^2}. \qquad (9.53)$$

Свободный член линейного уравнения регрессии

$$a = \overline{u}_y = b\overline{u}_x = 0.$$

Регрессионное уравнение отклонений от тренда имеет вид:

$$\tilde{u}_y = bu_x \qquad (9.54)$$

По данным табл. 9.12 коэффициент корреляции уровней урожайности и себестоимости

$$r_{x,y} = \frac{-111{,}8}{\sqrt{10341 \cdot 405{,}16}} = -0{,}055.$$

Прямая связь одинаково направленных трендов почти полностью компенсировала обратную связь между колебаниями признаков. Из 13 произведений $\Delta_{x_i}\Delta_{y_i}$ семь положительны. Прежде всего в начале и в конце ряда, где сильнее всего сказались тренды. Если бы не страшный неурожай в 1987 г., вызвавший огромные отклонения уровней, коэффициент корреляции был бы даже положителен.

Напротив, корреляция отклонений от трендов дает результат, соответствующий экономическому содержанию связи урожайности с себестоимостью. **Коэффициент корреляции отклонений от трендов** по формуле (9.52) составил:

$$r_{u_x u_y} = \frac{-952{,}7}{\sqrt{7678 \cdot 133{,}3}} = -0{,}941.$$

Коэффициент детерминации $r^2_{u_x u_y}$ равен 0,88, или 88% колебаний себестоимости картофеля связаны с колебаниями урожайности. Положительны лишь три произведения отклонения $u_{x_i} \cdot u_{y_i}$, притом наименьшие.

Коэффициент регрессии по формуле (9.53)

$$b = \frac{-952{,}7}{7678} = -0{,}124.$$

Уравнение регрессии:

$$\tilde{u}_{y_i} = -\,0{,}124u_{x_i}.$$

Это означает, что в среднем за период отклонение себестоимости от тренда было противоположно по знаку и составляло 0,124 отклонения урожайности от своего тренда. Если, например, урожайность в 1993 г. окажется на 20 ц/га ниже уровня тренда для этого года, составляющего 119,9 + 3,81 · 10 = 158 ц/га, то себестоимость надо ожидать на –20 (–0,124) = 2,48 руб. за 1 ц выше уровня тренда, который для 1993 г. равен 31,2 руб. за 1 ц, т. е., учитывая и тренды, и предполагаемый плохой урожай в 1993 г., себестоимость картофеля составила бы 31,2 + 2,48 = 33,66 руб./ц. Естественно, что этот прогноз всего лишь пример, как пользоваться уравнением регрессии отклонений от тренда. В нашем случае метеорология не дает оснований для прогноза урожайности, а сильнейшая инфляция делает вообще невозможным любой прогноз себестоимости без использования дефлятора (см. гл. 10).

Данные табл. 9.12 позволяют сделать интересное заключение о различии характера динамики признаков. Если из общей дисперсии (суммы квадратов отклонений от среднего уровня) урожайности 10341 большую часть составляет дисперсия за счет колеблемости 7678, то для себестоимости преобладающим моментом общей дисперсии, равной 405,16, является не колеблемость, дающая только 133,34, а тренд; это эффект скрытой инфляции до 1989 г.

Другим приемом измерения корреляции в рядах динамики может служить корреляция между теми из цепных показателей рядов, которые являются константами их трендов. При линейных трендах – это цепные абсолютные приросты. Вычислив их по исходным рядам динамики (a_x, a_y), находим коэффициент корреляции между абсолютными изменениями по формуле (9.52) или, что более точно, по формуле (9.51), так как средние изменения не равны нулю в отличие от средних отклонений от трендов. Допустимость данного способа основана на том, что разность между соседними уровнями в основном состоит из колебаний, а доля тренда в них невелика, следовательно, искажение корреляции от тренда очень большое при кумулятивном эффекте на протяжении длительного периода, весьма мало – за каждый год в отдельности. Однако нужно помнить, что это справедливо лишь для рядов с *c*-показателем, существенно меньшим единицы. В нашем примере для ряда урожайности *c*-показатель равен 0,144, для себестоимости он равен 0,350. Коэффициент корреляции цепных абсолютных изменений составил 0,928, что очень близко к коэффициенту корреляции отклонений от трендов.

Для рядов с тенденцией, близкой к экспоненте, следует рекомендовать корреляцию цепных темпов роста. Вычисление корреляции рядов динамики по цепным показателям не требует предварительного вычисления трендов, но все же желательно иметь о характере тенденции приближенное представление. Для параболических трендов с не очень большими ускорениями можно коррелировать цепные абсолютные изменения; при больших ускорениях лучше их не коррелировать. Если коррелируемые ряды имеют разные типы тенденций, вполне допустимо коррелировать соответствующие разные цепные показатели: абсолютные изменения в одном ряду с темпами изменений в другом и т. д.

К сожалению, все вышеизложенные приемы по существу решают только задачу измерения связи между колебаниями признаков, а не между тенденциями их изменений. Насколько допустимо переносить выводы о тесноте связи между колебаниями на связь динамических рядов в целом, зависит от материального, качественного содержания процесса и причинного механизма связи. Это проблема, выходящая далеко за пределы статистической науки. Если колебания урожайности являются на самом деле следствиями колебания суммы осадков за лето, т. е. корреляция именно колебаний отвечает сущности причинной связи, то, например, причинную связь между дозой удобрений и урожайностью нельзя свести к зависимости только между колебаниями. Здесь главное – причинная связь тенденций, а ее измерять мы так и не научились.

Завершая этим признанием главу о статистическом анализе рядов динамики, дадим последние методологические советы изучающим статистику.

Всякая наука – это процесс продолжающегося познания природы и общества. Нет наук законченных, которые следует лишь выучить наизусть, чтобы все знать.

Учебники и учебные пособия – лишь сжатые и неполные изложения уже достигнутого наукой уровня познания. Изучайте специальную литературу, если хотите больше знать, а также новейшие достижения ученых всего мира.

Не считайте и себя только «сосудами для вливания» знаний. Познав известное, вы тоже можете (и должны!) внести свой вклад в дальнейшее развитие теории статистики. «Если не я – то кто же?»

Рекомендуемая литература к главе 9

1. *Андерсон Т.* Статистический анализ временных рядов / Пер. с англ. – М.: Мир. – 1976.
2. *Афанасьев В. Н.* Статистическое обеспечение проблемы устойчивости сельскохозяйственного производства. – М.: Финансы и статистика, 1996.
3. *Вайну Я. Я.-Ф.* Корреляция рядов динамики. – М.: Статистика, 1977.
4. *Казинец Л. С.* Темпы роста и абсолютные приросты. – М.: Статистика, 1975.
5. *Четыркин Е. М.* Статистические методы прогнозирования. – Изд. 2-е, – М.: Статистика, 1977.
6. *Юзбашев М. М., Манелля А. И.* Статистический анализ тенденций и колеблемости. – М.: Финансы и статистика, 1983.

10 Глава

ИНДЕКСЫ

Само слово индекс (index) означает показатель. Обычно этот термин используется для некоей обобщающей характеристики изменений. Например, уже знакомый вам индекс Доу Джонса, индекс деловой активности, индекс объема промышленного производства и т. д. Гораздо реже термин «индекс» используется как обобщенный показатель состояния, например, известный индекс интеллектуального развития *IQ*.

В этой главе мы рассмотрим индексы прежде всего как показатели изменений. Очевидно, что сфера использования таких показателей безгранична: спортсмены стремятся улучшить свои достижения, предприниматель желает увеличить прибыль и т.д. Во всех этих случаях необходимо выразить изменения количественно. Как изменились цены, уровень жизни, покупательная сила денег и пр.? Ответы на все эти вопросы позволяют дать *индексы*.

10.1. ПОНЯТИЕ ИНДЕКСА

В предыдущей главе вы познакомились с показателями, которые измеряют абсолютные и относительные изменения: темпы роста, прироста, абсолютный прирост, цепные и базисные показатели, показатели средних изменений за период. В чем же **специфика индексов?** Принципиальных отличий три.

Во-первых индексы позволяют измерить изменение сложных явлений. Например, нужно определить, как изменились за год расходы жителей Москвы на городской транспорт. Для ответа на этот вопрос вы должны иметь численность пассажиров, перевезенных за год каждым видом городского транспорта, рассчитать среднемесячную численность пассажиров или взять точные данные из отчетов по месяцам, умножить численность на тариф перевозки (и число месяцев его действия – в случае использования среднемесячной численности) и полученные величины просуммировать. То же нужно сделать по данным за прошлый год. Затем сопоставить сумму расходов за последний год с суммой за прошлый год. То есть это не просто сравнение двух чисел, как при расчете темпов динамики или приростов, а получение и сравнение некоторых агрегированных величин.

Во-вторых, индексы позволяют проанализировать изменение – выявить роль отдельных факторов. Например, можно определить, как изменилась сумма выручки городского транспорта за счет изменения численности пассажиров и тарифов, наконец, за счет соотношения в объеме перевозок разными видами транспорта.

В-третьих, индексы являются показателями сравнений не только с прошлым периодом (сравнение во времени), но и с другой территорией (сравнение в пространстве), а также с нормативами. Например, интересно знать, не только как изменилось среднедушевое потребление мяса в России в данном году по сравнению с прошлым годом (или с каким-либо другим периодом), но и сравнить показатели среднедушевого потребления мяса в России и развитых странах Запада, Востока, а также провести сравнение с нормативной величиной, отвечающей нормам рационального питания. Очевидно, что каждое направление сравнения вносит что-то новое. Так, удой молока на одну корову в хозяйствах Российской Федерации в 1990 г. составил 2781 кг, а в 1989 – 2773 кг. Индекс равен 100,3% (+0,3) ≠ [2781 : 2773 = 1,0029 · 100%]. Повышение – 8 кг на одну корову, такое сравнение вроде внушает хотя и умеренный, но оптимизм. Если же сравнить с удоем в других странах, то те же данные выглядят так: в Великобритании в 1990 г. этот показатель был равен 5213 кг/корову, Польше – 3234, Швеции – 6213 кг/корову. Соответствующие индексы составили 53,3; 86; 44,8%.

Существует несколько определений индекса. Приведем одно из них, может быть самое краткое.

Индекс – это показатель сравнения двух состояний одного и того же явления (простого или сложного, состоящего из соизмеримых или несоизмеримых элементов).

Каждый индекс включает два вида данных: оцениваемые данные, которые принято называть ***отчетными*** и обозначать значком «1», и данные, которые используются в качестве базы сравнения – **базисные,** обозначаемые знаком «0».

Индекс, который строится как сравнение обобщенных величин, называется **сводным** или **общим,** и обозначается *I*. Если же сравни-

ваются необобщенные величины, то индекс называется **индивидуальным** и обозначается *i*. Как правило, подстрочно дается значок, который указывает, для оценки какой величины построен индекс. Например, $I_{q_{1/0}}$ или $i_{q_{1/0}}$, т. е. сводный и индивидуальный индексы для величины q.

Сравнения во времени могут охватывать короткий период: выработка за этот день и за вчерашний день, цены в сентябре по сравнению с августом и т. д. Но сравнение может проводиться и с отдаленным периодом: современные данные с довоенным 1940 г. или с 1986 г. – годом начала перестройки, когда экономика еще не была затронута структурными изменениями и т. д. Выбор базисного периода всегда аргументирован той задачей, для которой строится индекс. Обычно руководствуются двумя правилами: либо база сравнения представляет стабильный уровень, либо экстремальное значение – высшее достижение или низший уровень (в случае падения экономических показателей). Конечно, сравнение с отдаленным периодом вносит дополнительные трудности, что уже отмечалось в предыдущей главе. Некоторые специфические для построения индексов проблемы будут затронуты ниже.

10.2. ИНДЕКС КАК ПОКАЗАТЕЛЬ ЦЕНТРАЛЬНОЙ ТЕНДЕНЦИИ (ИНДЕКС СРЕДНИЙ ИЗ ИНДИВИДУАЛЬНЫХ)

Вы можете услышать, что уровень потребительских цен понизился или повысился. Речь в этом случае идет об индексе цен на потребительские товары. Общее изменение образуется под влиянием изменений цен на отдельные товары. Таким образом, мы имеем ряд отношений:

$$\frac{p_{11}}{p_{01}}, \frac{p_{12}}{p_{02}}, \frac{p_{13}}{p_{03}}, \ldots, \frac{p_{1n}}{p_{0n}} \text{ и т. д.}$$

Эти отношения есть не что иное, как индивидуальные индексы, и сводный индекс представляет собой средний из них:

$$I_p = \overline{\left(\frac{p_{1j}}{p_{0j}}\right)},$$

где j – номер товара.

Так как средняя есть показатель центра распределения, то и сводный индекс можно назвать показателем центральной тенденции. Проблема состоит в том, как получить этот сводный индекс. Впервые она возникла при попытке оценить совокупное изменение цен либо в виде отношения сумм цен:

$$\frac{p_{11} + p_{12} + \ldots + p_{1n}}{p_{01} + p_{02} + \ldots + p_{0n}} = \frac{\sum_{j=1}^{n} p_{1j}}{\sum_{j=1}^{n} p_{0j}},$$

либо как среднее из изменений цен на отдельные товары:

$$\frac{\frac{p_{11}}{p_{01}} + \frac{p_{12}}{p_{02}} + \ldots + \frac{p_{1n}}{p_{0n}}}{n} = \frac{\sum_{j=1}^{n} \frac{p_{1j}}{p_{0j}}}{n}. \qquad (10.1)$$

В том и другом варианте представлены невзвешенные средние. Первый вариант исходит из того, что цена рассчитывается за единицу товара, например за 1 кг, и сумма цен может рассматриваться как набор слагаемых с равными весами. Однако этот вариант не отвечает задаче осреднения показателей изменений цен на отдельные товары. Второй вариант настораживает тем, что согласно общему правилу средняя из относительных величин должна вычисляться как средняя взвешенная. Действительно, если говорить конкретно об измерении динамики цен на все продовольственные или непродовольственные товары, то ясно, что если цены на ювелирные изделия из золота удвоятся, а цены на хлеб останутся неизменными, это не значит, что в целом цены выросли на 50% ((2+ 1)/2 = 1,5). Приведенный пример показывает, что индекс цен для каждого товара должен сопровождаться неким «весом», который позволяет оценить относительную значимость этого индекса для потребителя. В качестве веса используют удельный вес в общей стоимости покупок: в базисном периоде:

$$d_0 = \frac{q_{0j} p_{0j}}{\sum_{(i)} q_{oj} p_{oj}}.$$

Если обозначить удельный вес отдельных затрат d_{0j}, то получим общий индекс цен как средний арифметический взвешенный из индивидуальных индексов цен:

$$I_p = \frac{\sum\limits_{(j)} i_{pj} d_{0j}}{\sum\limits_{(j)} d_{0j}} = \frac{\sum\limits_{(j)} i_{pj} q_{0j} p_{0j}}{\sum\limits_{(j)} q_{0j} p_{0j}}, \qquad (10.2)$$

т. е. $I_p = \bar{i}_p$.

Используя формулу (10.2) можно получить общее изменение цен на продукты по данным табл. 10.1.

Часто можно встретить утверждение, что чем сильнее варьируют веса средней, тем значительнее отличие невзвешенной средней от взвешенной. Покажем ошибочность этого утверждения применительно к индексу среднему из индивидуальных. Рассмотрим два примера А и Б.

А. Равенство взвешенной и простой средних при сильной вариации весов.

В табл. 10.1 представлены данные примера А.

Таблица 10.1

№ товара	Цены		Индекс i_p	Доля в базисной выручке d_0	$i_p d_0$	Вариация долей	
	P_0	P_1				$d_{j0} - \bar{d}_0$	$(d_{j0} - \bar{d}_0)^2$
1	10	11	1,1	0,40	0,44	0,20	0,0400
2	15	30	2,0	0,25	0,50	0,05	0,0025
3	20	28	1,4	0,15	0,21	–0,05	0,0025
4	25	40	1,6	0,10	0,16	–0,10	0,0100
5	30	27	0,9	0,10	0,09	–0,10	0,0100
Итого			1,4	1,00	1,40	0	0,0650

Невзвешенный средний индекс цен $\bar{i}_p = \frac{\sum i_p}{5} = \frac{7}{5} = 1{,}4$.

Среднее значение веса $\bar{d}_0 = \frac{1}{5} = 0{,}2$.

Взвешенный средний индекс цен $\bar{i}_p = \frac{\sum i_p \cdot d_0}{\sum d_0} = \sum i_p \cdot d_0 = 1{,}4$.

Результат совпадает с простой средней. Между тем вариация весов значительна, стандартное отклонение

$$s_d = \sqrt{\frac{0{,}0650}{5}} = 0{,}1140175;$$

Коэффициент вариации весов

$$v_d = \frac{0{,}1140175}{0{,}2} = 0{,}5700, \text{ т. е. } 57\%.$$

Б. Неравенство взвешенной и простой средних при слабой вариации весов.

В табл. 10.2 представлены данные примера Б.

Таблица 10.2

№ товара	Цены		Индекс i_p	Доля в базисной выручке d_0	$i_p d_0$	Вариация долей	
	P_0	P_1				$d_{j0} - \bar{d}_0$	$(d_{j0} - \bar{d}_0)^2$
1	10	11	1,1	0,15	0,165	–0,05	0,0025
2	15	30	2,0	0,26	0,520	0,06	0,0036
3	20	28	1,4	0,19	0,266	–0,01	0,0001
4	25	40	1,6	0,25	0,400	0,05	0,0025
5	30	27	0,9	0,15	0,135	–0,05	0,0025
Итого	×	×	1,4	1,00	1,486	0	0,0112

невзвешенный средний индекс цен: $\bar{i}_p = \frac{7}{5} = 1{,}4$;

взвешенный средний индекс цен $\bar{i}_p = \frac{1{,}486}{1} = 1{,}486$;

вариация весов $s_d = \sqrt{\frac{0{,}0112}{5}} = 0{,}04733$;

$v_d = 0{,}2366$ или 23,7%, т. е. вариация весов намного слабее, чем в примере А.

Рассмотрим, в чем секрет таких соотношений? Обратимся к формуле взвешенной средней:

$$\bar{x}_{взвеш.} = \frac{\sum_{i=1}^{n} x_i f_i}{\sum_{i=1}^{n} f_i} = \frac{\sum_{i=1}^{n} (\bar{x} + \Delta x_i) \cdot (\bar{f} + \Delta f_i)}{\sum_{i=1}^{n} (f_i + \Delta f_i)}.$$

где x, $\bar{f}$ – простые средние;
Δx, Δf – отклонения от них.

Представим последнее выражение как:

$$\frac{\sum \bar{x}\bar{f} + \sum \bar{x}\Delta f_i + \sum \bar{f}\,\Delta x_i + \sum \Delta x_i \Delta f_i}{\sum \bar{f} + \sum \Delta f_i} =$$

$$= \frac{n\bar{x}\bar{f} + \bar{x}\sum \Delta f_i + \bar{f}\sum \Delta x_i + \sum \Delta x_i \Delta f_i}{n\bar{f} + 0} =$$

$$= \frac{n\bar{x}\bar{f} + \bar{x}\cdot 0 + \bar{f}\cdot 0 + \sum \Delta x_i \Delta f_i}{n\bar{f}} = \bar{x} + \frac{\sum \Delta x_i \Delta f_i}{n\bar{f}}.$$

Числитель второго слагаемого можно представить через коэффициент корреляции между x и f:

$$\bar{x}_{взвеш.} = \bar{x} + \frac{n\sigma_x \sigma_f r_{xf}}{n\bar{f}} = \bar{x} + \sigma_x \cdot v_f \cdot r_{xf}. \qquad (10.3)$$

Эта формула аналогична формуле (5.6).

Следовательно, средняя взвешенная равна простой средней, если:

- вариация признака x_i отсутствует, т. е. $\sigma_x = 0$;
- вариация весов f_i отсутствует, т. е. $v_f = 0$;
- нет корреляции между вариациями признака и весов, т. е. $r_{xf} = 0$ (хотя бы сами x_i и f_i варьировали как угодно сильно).

Отношение взвешенной средней и простой можно выразить следующим образом:

$$\frac{\bar{x}_{взвеш.}}{\bar{x}_{прост}} = 1 + v_x \cdot v_f \cdot r_{xf}. \qquad (10.4)$$

Поскольку различие взвешенной и простой средних зависит от корреляции значений признака и веса, постольку оно может оказаться большим при слабой вариации весов, чем при их сильной вариации (см. главу 5).

Рассмотрим соотношения между индексами (10.1) и (10.2) на примере табл. 10.3.

Таблица 10.3

Данные розничной торговли города *N*

	Выручка в мае		Отношение цен в июне к ценам в мае, %	Выручка с учетом изменения цен, млн руб.
	абс. млн. руб.	относит.		
	$q_0 p_0$	d_0	$i_p = p_1 : p_0$	$q_0 p_1 = q_0 p_0 \cdot i_p$
1	2	3	4	5
Мясо и мясопродукты	2352,0	0,271	110,5	2599,0
Рыба и рыбопродукты	735,0	0,085	112,2	824,7
Масло животное	2058,0	0,237	103,2	2123,8
Масло растительное	9,8	0,001	105,6	10,4
Молоко и молочные продукты	882,0	0,102	102,4	903,2
Сахар	2644,0	0,304	107,3	2837,0
Итого	8680,8	1,000	641,2*	9298,1

* Обычно i_p не суммируются.

Обратите внимание на данные гр. 5 табл. 10.3: произведение $q_0 p_0 i_p$ имеет не просто техническое значение взвешивания индивидуального индекса, но дает определенный содержательный результат – показатель условных затрат на покупку с учетом изменения цен $q_0 \cdot p_0 \cdot i_p = q_0 \cdot p_1$.

Это дает право представить формулу (10.2) в виде:

$$I_p = \frac{\sum_{(j)} q_{0j} p_{1j}}{\sum_{(j)} q_{0j} p_{0j}}. \qquad (10.5)$$

Выражение (10.5) получило известность как индекс Ласпейреса, предложившего эту формулу в 1864 г.

По данным табл. 10.3

$$I_p = \frac{9298{,}1 \text{ млн руб.}}{8680{,}8 \text{ млн руб.}} = 100\% = 107{,}1\%,$$

т. е. цены возросли в среднем на 7,1%. Если воспользоваться формулой (10.1), то $I_p = 641{,}2/6 = 1{,}069 \cdot 100 = 106{,}9\%$, т. е. в среднем цены возросли на 6,9%. Отличие от среднего взвешенного арифметического индекса составляет 0,2%.

Мы рассмотрели определение среднего изменения на основе средней арифметической из индивидуальных, но ведь могут использоваться и другие виды средних: средняя геометрическая, средняя гармоническая и т. д. – невзвешенные и взвешенные. Используя среднюю геометрическую невзвешенную, получаем:

$$I_p = \sqrt[6]{1{,}105 \cdot 1{,}122 \cdot 1{,}032 \cdot 1{,}056 \cdot 1{,}024 \cdot 1{,}073} =$$

$$= \sqrt[6]{1{,}486} \cdot 100\% = 106{,}8\%.$$

Средняя гармоническая всегда дает результат, меньший средней арифметической. Применяя среднюю гармоническую невзвешенную, получаем:

$$I_p = \frac{6}{\frac{1}{1{,}105} + \frac{1}{1{,}122} + \frac{1}{1{,}032} + \frac{1}{1{,}056} + \frac{1}{1{,}024} + \frac{1}{1{,}073}} = 1{,}068 \cdot 100\% = 106{,}8\%.$$

Опять-таки деление единицы на каждый индекс предполагает равное значение изменения цен на товары, что не соответствует практике.

Используя в качестве весов затраты на покупку в отчетном периоде, получаем сводный индекс цен как средний гармонический взвешенный из индивидуальных индексов цен:

$$I_p = \frac{\sum q_1 p_1}{\sum \frac{q_1 p_1}{i_p}}. \qquad (10.6)$$

В формуле (10.6) и далее для простоты мы опустили подстрочный значок, соответствующий номеру товара (элемента), хотя, конечно же, суммирование и в числителе, и в знаменателе производится по всему набору товаров (элементов).

Рассчитаем этот индекс по данным табл. 10.3. Кроме того, нам потребуются дополнительные данные. Как всегда, лучшей формой представления цифровых данных является таблица. Представим все необходимые данные в табл. 10.4, используя вместо названий номера продуктов.

Таблица 10.4

Данные розничной торговли города

№ п/п	Относительное изменение количества купленных продуктов в июне по сравнению с маем, % $i_p = q_1 : q_0$	Выручка в июне, млн руб., $q_1 \cdot p_1$	Условная выручка без учета изменения цен, млн руб., $q_1 p_0 = q_1 p_1 : i_p$
1	98,5	2560,0	2316,7
2	100,3	827,2	737,3
3	97,8	2077,1	2012,7
4	102,0	10,6	10,0
5	100,0	903,2	882,0
6	98,0	2780,3	2591,1
Итого	596,6*	9158,4	8549,8

* Обычно i_q не суммируются.

$$I_p = \frac{9158{,}4 \text{ млн руб.}}{8549{,}8 \text{ млн руб.}} \cdot 100\% = 107{,}1\%.$$

Результат совпал с тем значением I_{p_1}, которое было получено по формуле (10.2). Но это случайное совпадение, которое оказалось возможным из-за слабой корреляции между изменением уровня цен и объема продаж отдельных товаров. Это может быть при сравнении за короткий период. В рыночной экономике взаимосвязь между колебаниями цен и объема продаж проявляется при сравнении за более длительный период. Ниже будет показано, как измерить величину этой корреляции (см. формулу (10.17).

Знаменатель формулы (10.6) имеет смысл затрат на покупку «отчетного» количества товаров по базисным ценам:

$$\left(\frac{q_1 \cdot p_1}{p_1 / p_0}\right) = q_1 p_0.$$

Тогда формула (10.6) может быть представлена как

$$I_p = \frac{\sum\limits_{(J)} q_1 p_1}{\sum\limits_{(J)} q_1 p_1}. \qquad (10.7)$$

Эта формула индекса цен была предложена Пааше в 1874 г. Различие между индексами Пааше и Ласпейреса, их использование обсуждаются ниже в данной главе.

Итак, мы рассмотрели применение разных форм и видов средних величин для определения среднего изменения цен по всем товарам. Люди всегда в первую очередь интересовались ценами и их изменениями. Но такой же подход может быть применен к оценке сводных изменений других характеристик, например объема (количества) покупок товаров. Кстати заметим, что используемые нами обозначения цен *(p)*, количества *(q)* неслучайны и соответствуют начальным буквам английских слов price (цена) и quantity (количество). Это закрепленные обозначения в статистике.

Таким образом, общее изменение количества проданных товаров формируется как среднее по отношению к изменениям объема покупок отдельных товаров, т. е.

$$Iq = \bar{i}_q, \text{ где } i_q = \frac{q_1}{q_0}.$$

Возникает вопрос о порядке расчета средней из i_q: средняя арифметическая – простая или взвешенная – или другая форма средней. Ограничимся рассмотрением только средней арифметической.

По данным табл. 10.4 простая средняя арифметическая из индивидуальных индексов количества равна:

$$I_q = \frac{\sum\limits_{(J)} i_q}{n} = \frac{596{,}6}{6} = 0{,}994 \cdot 100\% = 99{,}4\%(-0{,}6\%).$$

Используя в качестве весов для изменений объема покупок удельный вес покупок в общей сумме затрат, получаем:

$$I_p = \frac{\sum\limits_{(J)} i_q \cdot d_0}{\sum\limits_{(J)} d_0} = \frac{\sum\limits_{(J)} i_q \cdot q_0 \cdot p_0}{\sum\limits_{(J)} q_0 \cdot p_0}, \qquad (10.8)$$

т. е. индекс I_q – средний арифметический взвешенный из индивидуальных i_q.

По данным нашего примера (табл. 10.3 и 10.4) общий индекс количества равен:

$$I_q = \frac{8549{,}8 \text{ млн руб.}}{8680{,}8 \text{ млн руб.}} \cdot 100\% = 98{,}5\%.$$

Получилось, что объем покупок продовольственных товаров сократился в среднем на 1,5%. Это более значительная оценка снижения, нежели полученная при расчетах по простой средней арифметической (–0,6%). Так что мы еще раз получили подтверждение зависимости результата от использованной формулы.

Зная среднюю величину изменения показателя и индивидуальные индексы, можно проводить анализ методами вариационной статистики: анализировать распределение товаров по изменению цен, объема покупок, сравнивать модальное и среднее изменение, максимальное и минимальное; по показателям эксцесса распределений делать выводы о том, насколько однородны изменения цен и количества по отдельным товарам, группировать товары по уровню цен и степени их изменения и т. д.

10.3. АГРЕГАТНЫЕ ИНДЕКСЫ. СИСТЕМА ИНДЕКСОВ

Мы познакомились с построением сводных индексов на основе индивидуальных. Однако возможен и другой путь. Обратимся к формулам индексов Ласпейреса (10.5) и Пааше (10.7). Эти индексы могут быть рассчитаны на основе данных о количестве проданных товаров в базисном и отчетном периоде (по каждому j-му товару) q_{0j} и q_{1j} и ценах – p_{1j} и p_{0j}. Такие индексы принято называть агрегатными. Так же можно построить и I_q не через осреднение индивидуальных индексов, а на основе сравнения двух сумм (агрегатов), см. (10.7).

Агрегатные индексы считаются основной формой индексов. Они выполняют две функции: **синтетическую и аналитическую.** Первая функция обеспечивается тем, что в одном индексе обобщаются (синтезируются) непосредственно несоизмеримые явления. Например, цены на разные товары или разные товары, абсолютно не сопоставимые между собой в натуральном выражении. Когда мы записываем

$$I_p = \frac{\sum\limits_{(j)} q_1 p_0}{\sum\limits_{(j)} q_0 p_0},$$

то благодаря использованию ценового соизмерителя можно агрегировать данные по различным товарам.

Вторая функция – аналитическая – следует из взаимосвязи индексов. Дело в том, что практически каждый индекс можно рассматривать как составляющую некоей системы индексов, в которой его роль сводится к измерению одного из факторов общего изменения сложного явления и вклада этого фактора в совокупное изменение. Так, например, индекс цен можно рассматривать как показатель влияния изменения цен на выручку от продажи. Такая трактовка опирается на следующую связь признаков:

количество × цена = выручка (или затраты на покупку), т. е.

$$qp = w. \tag{10.9}$$

Системе признаков соответствует система индексов (т. е. показателей их изменений). Исходя из этого можно записать:

$$I_q = I_{w(q)} = \frac{\sum\limits_{(j)} q_1 p_0}{\sum\limits_{(j)} q_0 p_0}; \quad I_p = I_{w(p)} = \frac{\sum\limits_{(j)} q_0 p_1}{\sum\limits_{(j)} q_0 p_0}. \tag{10.10}$$

Обратите внимание: эта запись соответствует трактовке индекса как метода анализа. Когда мы указываем $I_{w(q)}$ или $I_{w(p)}$, то имеем в виду измерение общего изменения результативного явления (в данном случае w) за счет одного из факторов (q или p). Конечно, можно ограничиться записью I_q и i_p – ничего не изменится по существу.

При построении агрегатных индексов удобно пользоваться такими понятиями, как «индексируемый признак» и «признак-вес». **Индексируемый** – это признак, изменение которого характеризует данный индекс. Например, в I_q – это q, в i_p – это p. Значение индексируемого признака изменяется: отчетное значение сопоставляется с базисным.

Признак-вес выполняет функцию веса по отношению к индексируемому признаку; его значение в данном индексе принимается неизменным, так как он не должен искажать оценку изменения индексируемого признака. В I_q признаком-весом является p, а в I_p – q.

Индексируемый признак можно назвать **фактором** изменения общего результата, а признак-вес – **характеристикой условий,** в которых оценивается это изменение.

Если индексы рассматриваются в системе, то должна обеспечиваться взаимосвязь между ними. Например, в соответствии с (10.9) должно выполняться равенство

$$I_q \cdot I_p = I_w. \tag{10.11}$$

Обратимся к формулам (10.11). Каждый из индексов показывает, как изменился тот или иной фактор при неизменности прочих условий: и в формуле индекса I_q и в формуле I_p веса закреплены на базисном уровне. Это обеспечивает сопоставимость оценок изменений факторов. Однако равенство (10.11) не обеспечивается или, как говорят иначе, не обеспечивается увязка индексов в систему:

$$\frac{\sum\limits_{(j)} q_1 p_0}{\sum\limits_{(j)} q_0 p_0} \cdot \frac{\sum\limits_{(j)} q_0 p_1}{\sum\limits_{(j)} q_0 p_0} \neq \frac{\sum\limits_{(j)} q_1 p_1}{\sum\limits_{(j)} q_0 p_0}.$$

То же происходит, если все индексы будут построены с отчетными весами:

$$\frac{\sum\limits_{(j)} q_1 p_1}{\sum\limits_{(j)} q_0 p_1} \cdot \frac{\sum\limits_{(j)} q_1 p_1}{\sum\limits_{(j)} q_1 p_0} \neq \frac{\sum\limits_{(j)} q_1 p_1}{\sum\limits_{(j)} q_0 p_0}.$$

Только когда взаимосвязанные индексы строятся с весами разных периодов, увязка их в систему выполняется:

$$\frac{\sum\limits_{(j)} q_1 p_1}{\sum\limits_{(j)} q_0 p_1} \cdot \frac{\sum\limits_{(j)} q_0 p_1}{\sum\limits_{(j)} q_0 p_0} = \frac{\sum\limits_{(j)} q_1 p_1}{\sum\limits_{(j)} q_0 p_0}. \tag{10.12}$$

или

$$\frac{\sum\limits_{(j)} q_1 p_0}{\sum\limits_{(j)} q_0 p_0} \cdot \frac{\sum\limits_{(j)} q_1 p_1}{\sum\limits_{(j)} q_1 p_0} = \frac{\sum\limits_{(j)} q_1 p_1}{\sum\limits_{(j)} q_0 p_0}. \tag{10.13}$$

Из этих двух вариантов отечественная статистика долгое время отдавала предпочтение второму. Соответственно существовало правило определения периода весов: индексы первичных признаков строятся на весах базисного периода, вторичных – на весах отчетного периода. Это правило признавало неравное значение признаков в системе: первичный признак выступает как основа формирования нового (отчетного) значения результативного признака w_1. Этим объясняется то, что индекс первичного признака (например,

I_q) оценивает изменение этого признака при сохранении базисных условий, тогда как изменение вторичного признака оценивается уже в изменившихся условиях, когда первичный признак принял значение отчетного периода.

Рассмотрим на примере, как влияет использование разных значений признака-веса на величину индекса (табл. 10.5).

Таблица 10.5

Данные о продаже продуктов на городском рынке за месяц

	Цена, тыс. руб./кг		Продано, т		Выручка, млн руб.			
	май	июнь	май	июнь	май	июнь	условная	условная
	p_0	p_1	q_0	q_1	$w_0 = q_0p_0$	$w_1 = q_1p_1$	q_1p_0	q_0p_1
Говядина	15,05	15,95	3,0	2,98	45,15	47,53	44,85	47,85
Свинина	16,30	17,54	2,8	2,75	45,64	48,24	44,82	49,11
Масло животное	13,75	14,19	1,5	1,45	20,62	20,58	19,94	21,28
Масло растительное	5,56	5,77	1,0	0,90	5,56	5,19	5,00	5,77
Картофель	1,02	1,13	10,2	10,80	10,40	12,20	11,02	11,53
Капуста	4,12	4,08	8,5	8,8	35,02	35,90	36,26	34,68
Яблоки	9,24	9,26	5,7	4,9	52,67	45,37	45,28	52,78
Итого	–	–	–	–	215,06	215,01	207,17	223,00

1) $I_q = \dfrac{215{,}01}{223{,}00} = 0{,}9642;$ $\qquad I_p = \dfrac{223{,}00}{215{,}06} = 1{,}037;$

с весами p_1 $\qquad$ с весами q_0

$$I_w = \frac{215{,}01}{215{,}06} = 0{,}9998; \quad 0{,}9642 \cdot 1{,}037 = 0{,}9998;$$

2) $I_q = \dfrac{207{,}17}{215{,}06} = 0{,}9633;$ $\qquad I_p = \dfrac{215{,}01}{207{,}17} = 1{,}0378;$

с весами p_1 $\qquad$ с весами q_0

$$I_w = \frac{215{,}01}{215{,}06} = 0{,}9998 = 0{,}9633 \cdot 1{,}0378 \approx 0{,}9998.$$

В обоих вариантах получены показатели снижения объема продажи и роста цен, но в первом случае объем продажи снизился на 3,58%, цены повысились на 3,7%, а во втором – снижение объема продажи на 3,7% и рост цен на 3,78%. Следуя статистической логике, можно сказать, что точечные оценки в принципе невозможны; можно говорить лишь о поле или интервале оценок: для объема продажи – снижение от –3,58% до – 3,7%; для цен – рост от 3,7% до 3,78%.

Однако в практическом использовании индексов стремятся получить однозначное решение тем или иным способом. Первый путь – получение средних оценок изменений: либо в форме индексов, построенных на средних весах:

$$I_p = \frac{\sum\limits_{(j)} p_1 \dfrac{q_1 + q_0}{2}}{\sum\limits_{(j)} p_0 \dfrac{q_1 + q_0}{2}},$$

либо через осреднение разновзвешенных индексов. При этом предпочтение отдается средней геометрической:

$$I_p = \sqrt{\frac{\sum q_1 p_1}{\sum q_1 p_0} \cdot \frac{\sum q_0 p_1}{\sum q_0 p_0}}. \qquad (10.14)$$

Второй путь основан на предпочтении какого-то одного варианта построения взаимосвязанных индексов. Как уже отмечалось, в отечественной статистике был принят второй вариант. Но при этом возникала несопоставимость оценок изменений признаков. Поэтому делалась попытка построения всех взаимосвязанных индексов на весах одного периода – базисного:

$$I_q = \frac{\sum\limits_{(j)} q_1 p_0}{\sum\limits_{(j)} q_0 p_0} \cdot I_p = \frac{\sum\limits_{(j)} q_0 p_1}{\sum\limits_{(j)} q_0 p_0}. \qquad (10.15)$$

Понятно, что в этом случае не выполняется увязка индексов в систему:

$$I_q \cdot I_p \neq I_w.$$

Изолированная оценка изменения каждого фактора при неизменности другого приводит к недоучету эффекта совместного измене-

ния факторов. Скажем, вы смотрите движущееся изображение без звука или слушаете звуковое сопровождение без изображения, и в том, и в другом случае воздействие меньше, чем при соединении изображения и звука. Наглядно это можно показать с помощью особого вида плоскостной диаграммы, известной в отечественной статистике как «знак Варзара» (по имени русского статистика В. Е. Варзара (1851–1940) (см. рис. 10.1).

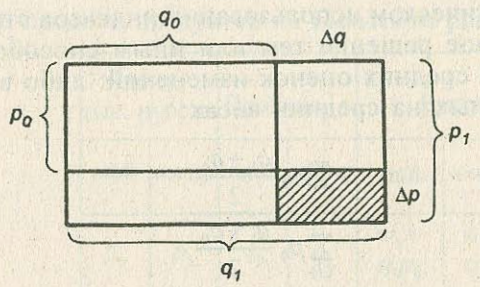


Рис. 10.1. Знак Варзара

Результативное явление представлено здесь в виде прямоугольника, площадь которого в базисном периоде $s_{q_0p_0}$, в отчетном – $s_{q_1p_1}$. Переход от базисного состояния к отчетному формируется за счет изменения фактора $q(s_{\Delta qp_0})$, изменения фактора $p(s_{\Delta pq_0})$ и совместного изменения обоих факторов $(s_{\Delta q\Delta p})$:

$$S_{q_1p_1} = s_{q_0p_0} + s_{\Delta qp_0} + s_{\Delta pq_0} + s_{\Delta q\Delta p}. \qquad (10.16)$$

В статистической науке выработано множество версий такого разложения: 1) выделение эффекта взаимодействия факторов в самостоятельный член; 2) присоединение его к какому-либо одному фактору (т. е. построение какого-либо из индексов на весах отчетного периода); 3) разделение эффекта взаимодействия факторов и присоединение к изменениям факторов – поровну, либо пропорционально значениям индексов факторов, либо еще по какому-то принципу. Вы можете тоже попытаться предложить свое решение – актуальность проблемы сохраняется.

В. И. Борткевич (1868–1931) вывел формулу, объясняющую различие между индексами с разными весами:

$$\frac{\sum\limits_{(j)} q_1p_1}{\sum\limits_{(j)} q_1p_0} : \frac{\sum\limits_{(j)} q_0p_1}{\sum\limits_{(j)} q_0p_0} = \underbrace{1 + r_{i_q i_p} v_{i_q} v_{i_p}}_{\text{эффект совместного изменения факторов q и p}}. \qquad (10.17)$$

Точно так же можно выразить соотношение между индексами фактора q с разными весами. Из формулы (10.17) ясно, что индексы с отчетными и базисными весами будут равны, если выполняется хотя бы одно из условий: или корреляция между изменениями цен и объема продажи на отдельные товары отсутствует, $r_{i_p i_q} = 0$; или темпы изменения объемов товаров всех видов будут одинаковы, $v_{i_q} = 0$; или темпы изменений цен на все товары будут одинаковы, $v_{i_p} = 0$. Чем большая дистанция разделяет сравнимые периоды, тем сильнее проявляются все отмеченные факторы различий между индексами с разными весами.

Ничего не меняется, если результативный признак включает *более двух факторов,* т. е. в случае мультипликативной модели:

$$y = x_1 \cdot x_2 \ldots x_k.$$

Если придерживаться концепции неравноправия факторов и строить индексы с разными весами, то все зависит от принятой последовательности факторов в системе. Например, общие затраты на кожу для изготовления женских туфель можно представить как $w = qlp$, где q – количество пар туфель; l – средний расход кожи на одну пару; p – цена кожи. Первым стоит фактор q как первичный, с которого и начинаются все изменения. Тогда индексы будут иметь вид:

$$\frac{\sum q_1l_0p_0}{\sum q_0l_0p_0} \cdot \frac{\sum q_1l_1p_0}{\sum q_1l_0p_0} \cdot \frac{\sum q_1l_1p_1}{\sum q_1l_1p_0} = \frac{\sum q_1l_1p_1}{\sum q_0l_0p_0} = \frac{\sum w_1}{\sum w_0}. \qquad (10.18)$$

Здесь используется то же правило выбора весов, которое было сформулировано выше. Признаки, стоящие слева от индексируемого признака, трактуются по отношению к нему как первичные и закрепляются на отчетном уровне (они «уже» изменились), стоящие справа от него трактуются как вторичные и закрепляются на базисном уровне (они как бы «еще» не изменились). К этому добавляется условие содержательной интерпретации при последовательном объе-

динении признаков слева направо. Скажем, произведение ql имеет экономический смысл – это расход кожи на весь объем производства туфель, при перестановке признаков q, p, 1 произведение qp экономического смысла не имеет. На таком подходе основан метод цепных подстановок, широко используемый в экономическом анализе.

Если же все индексы строятся на весах одного и того же (базисного) периода, то последовательность признаков не имеет значения. Система индексов будет иметь вид:

$$\frac{\sum q_1 l_0 p_0}{\sum q_0 l_0 p_0} \cdot \frac{\sum q_0 l_1 p_0}{\sum q_0 l_0 p_0} \cdot \frac{\sum q_0 l_0 p_1}{\sum q_0 l_0 p_0} \cdot I_{\substack{\text{совместных} \\ \text{изменений}}} = I_w, \qquad (10.19)$$

$$I_{\substack{\text{совместных} \\ \text{изменений}}} = \frac{I_w}{I_q \cdot I_l \cdot I_p}.$$

И в этом случае многофакторной модели эффект совместных изменений можно либо сохранить как самостоятельный член разложения, либо распределить между изменениями факторов. Это зависит от поставленной задачи и от пристрастий исследователя.

Сравнение данных отчетного и базисного периодов неявно предполагает представление экономических процессов в виде дискретной последовательности периодов времени, что особенно проблематично при сравнении в длительном периоде. Экономические индексы для моментов непрерывного времени были предложены в 1928 г. французским статистиком Ф. Девизиа. Это привело к использованию в индексном анализе дифференциального исчисления. Данный подход до сих пор не вошел в статистическую практику, однако теоретически он более обоснован, нежели традиционные методы.

10.4. СВОЙСТВА ИНДЕКСОВ

Как было показано, в построении индексов возникает много дискуссионных вопросов. Индексы считаются построенными правильно, если они удовлетворяют ряду тестов. Эти тесты были сформулированы американским статистиком И. Фишером (1867 – 1947). Основные тесты таковы:

1. Тест обратимости во времени. Индексы, исчисленные в «прямом» и «обратном» направлениях, должны быть взаимообратными

числами. Например, если индекс показывает, что уровень цен в отчетном периоде по сравнению с базисным повысился в два раза, то он должен отражать, что в базисном периоде цены были вполовину ниже, чем в отчетном, т. е.

$$I_{ab} \cdot I_{ba} = 1, \qquad (10.20)$$

где a и b – сравниваемые периоды.

Очевидно, что наличие этого свойства желательно у любого индекса, ибо в таком случае сравнение между двумя состояниями не будет зависеть от того, какое из них принято за базу, особенно это важно при территориальных сравнениях.

2. Тест обратимости по факторам. Если поменять местами в индексе цен символы для цен и для количества, то мы должны получить индекс количества, который, будучи умножен на индекс цен, должен дать изменение общей стоимости товаров. Например, имеем:

$$I_p = \frac{\sum p_1 q_0}{\sum p_0 q_0}.$$

Если теперь поменять местами p и q, то получим:

$$I_q = \frac{\sum q_1 p_0}{\sum q_0 p_0}.$$

Произведение этих индексов

$$\frac{\sum p_1 q_0}{\sum p_0 q_0} \cdot \frac{\sum q_1 p_0}{\sum q_0 p_0},$$

не равно индексу общей стоимости $\sum q_1 p_1 / \sum q_0 p_0$. Следовательно, индексы этого типа не отвечают тесту обратимости факторов. Тесту обратимости отвечает средний геометрический индекс (10.14). По этой причине он был назван И. Фишером идеальным индексом.

3. Тест кружного испытания (циркулярность). Если построен некоторый индекс для года a при базисном годе b и для года b при базисном годе c, то из них можно получить индекс года a при базисном годе c. Тест кружного испытания требует, чтобы $I_{a/c}$, основанный на промежуточных сравнениях, совпал с тем, какой мы получили бы при непосредственном сравнении a с c, т. е.

$$I_{a/b} \cdot I_{b/c} = I_{a/c} \qquad (10.21)$$

Это требование принято называть в статистике **«цепным тестом»**.

В случае взвешенных индексов этот тест выполняется только для индексов с постоянными весами. Особенно трудно обеспечить выполнение этого теста при сравнении с отдаленной базой. Легко сравнивать каждый из ряда лет с предыдущим, но нелегко сравнивать удаленные годы: произведение цепных сравнений (между прилежащими годами) может отличаться от результатов непосредственного сравнения лет в начале и конце периода. Тут возникает много экономических проблем – и постоянство весов (проблема выбора неизменных цен при построении индексов объема производства), и выделение сравнимого круга элементов на протяжении всего периода (сравнимого круга товаров, видов продукции, труда и т. д.) при анализе изменений цен, заработной платы и т. п.

В этот же тест Фишер вводил условие **круговой сходимости,** которое гласит: если условия начального и конечного моментов времени совпадают по уровням цен и объемов товаров, то произведение индексов цен и объемов товаров за все подпериоды должно быть равно единице.

4. Соизмеримость. Численные значения индексов не должны зависеть от выбора единиц измерения объема товаров и цен.

5. Пропорциональность. Согласно данному тесту, если темпы роста всех цен (или объемов товаров) равны одному и тому же числу, то этому же числу должен быть равен индекс цен (или индекс объема).

6. Включение – исключение. Если к набору товаров, по которым вычисляются индексы, и объему товаров, добавить еще один товар, темпы роста цены (или объема) которого совпадают с первоначальным индексом, то первоначальный индекс цен (или объема) не должен измениться.

Как видим, формулировка всех тестов основана на логике построения экономико-статистических показателей.

Тесты сыграли большую роль в развитии методологии экономических индексов.

10.5 ИНДЕКСНЫЙ АНАЛИЗ ВЗВЕШЕННОЙ СРЕДНЕЙ. ИНДЕКС СТРУКТУРЫ

Индексы позволяют анализировать изменения не только агрегатов, но и средних величин. Предположим, изучается динамика средней цены товара на трех рынках города, расположенных в разных районах – центральном и двух периферийных – старой и новой

застройки. Уровень цен в этих районах разный, соответственно на среднюю цену продажи на колхозных рынках влияют не только цены на каждом из них, но и доля каждого рынка в общем объеме продажи.

Формула средней цены:

$$\bar{p} = \frac{\sum_{i=1}^{n} p_i q_i}{\sum_{i=1}^{n} q_1},$$

где p_i – цена товара на i-м рынке, $q_i / \sum_{(i)} q_i$ – структура продажи.

Изменение средней цены (как и любой взвешенной средней) выражается индексом:

$$I_{\bar{p}} = \frac{\bar{p}_1}{\bar{p}_0} = \frac{\sum_{(i)} q_i p_i}{\sum_{(i)} q_1} : \frac{\sum q_0 p_0}{\sum q_0}. \qquad (10.22)$$

Этот индекс получил название индекса *переменного состава*, так как отражает не только изменение осредняемого признака p, но и структуры совокупности $q_i / \sum_{(i)} q_i$. На основе индекса средней величины могут быть построены индекс самого осредняемого признака при постоянстве структуры совокупности и индекс структуры:

$$I_{\bar{p}} = \frac{\sum q_1 p_1}{\sum q_1} : \frac{\sum q_1 p_0}{\sum q_1} = \frac{\sum q_1 p_1}{\sum q_1 p_0}. \qquad (10.23)$$

Этот индекс получил название **индекса постоянного состава.** Соответственно

$$I_{структуры} = \frac{\sum q_1 p_0}{\sum q_1} : \frac{\sum q_0 p_0}{\sum q_0}. \qquad (10.24)$$

Формулы индексов (10.23) и (10.24) основаны на общепринятом правиле, по которому структура совокупности как первичная характеристика при индексации цен закрепляется на уровне отчетного периода, а цены как вторичная характеристика при индексации структуры закрепляются на уровне базисного периода. Очевидно,

что применение весов разных периодов и в этом случае обеспечивает выполнение равенства:

$$I_p \cdot I_{\text{структуры}} = I_{\bar{p}} \text{ или } I_{\bar{p}(p)} \cdot I_{\bar{p}\left(\frac{q}{\sum q}\right)} = I_{\bar{p}}. \qquad (10.25)$$

Конечно, можно все индексы построить на весах базисного периода, и это будет правильнее с точки зрения оценки изменения каждого из факторов, но тогда равенство (10.25) будет нарушено.

Рассмотрим построение этих индексов на примере. На трех рынках города продается картофель. Данные о продаже за день в зарегистрированных ценах приведены в табл. 10.6.

Таблица 10.6

Дневная продажа картофеля на колхозных рынках города

Рынки	Объем дневной продажи, кг		Цена, руб./кг		Изменение цены, %	Удельный вес каждого рынка, %		Выручка от продажи, тыс. руб.		
	август	сентябрь	август	сентябрь				август	сентябрь	условная
	q_0	q_1	p_0	p_1	i_p	d_0	d_1	q_0p_0	q_1p_1	q_1p_0
Центральный	160	150	1,60	2,00	125,0	38,1	30,6	256	300	240
Старый	100	90	1,50	1,60	106,7	23,8	18,4	150	144	135
Новый	160	250	1,80	2,30	127,8	38,1	51,0	288	575	450
Итого	420	490	1,65	2,08	123,5	100,0	100,0	694	1019	825

Средняя цена картофеля в августе составила $\bar{p}_0$ = 1,65 тыс. руб./кг, в сентябре $\bar{p}_1$= 2,08 тыс. руб./кг. Наибольший рост цен произошел на рынке в новом районе, но здесь же и наибольшее увеличение объема продаж, в результате чего доля этого рынка в общей дневной реализации картофеля в сентябре стала превышать половину всего объема. Индекс средней цены составил:

$$\bar{I}_p = 2{,}08 \text{ тыс. руб./кг} : 1{,}65 \text{ тыс. руб./кг} = 1{,}259 \cdot 100\% = 125{,}9\%$$

Изменение самой цены в условиях структуры продажи, сложившейся в отчетном периоде, составило:

$$I_p = 208 : \frac{825}{490} = 208 : 1{,}684 = 1{,}235 \cdot 100\% = 123{,}5\%,$$

т. е. среднее повышение цен на рынках было несколько меньшим, чем повышение средней цены (+23,5% против +25,9%). Эту величину мы получили делением средней цены в отчетном периоде на среднюю условную цену, которая была бы при базисном уровне цен на рынках и отчетной структуре продаж. Этот же индекс можно было получить как отношение сумм выручки в отчетном периоде к условной выручке:

$$I_p = \frac{1019 \text{ тыс руб.}}{825 \text{ тыс руб.}} = 1{,}235.$$

Различие между индексом постоянного состава I_p и индексом переменного состава $I_{\bar{p}}$ вызвано изменением структуры:

$$I_{\text{структуры}} = 168{,}4 : 165{,}2 = 1{,}019 \cdot 100\% = 101{,}9\%.$$

За счет изменения структуры продажи средняя цена картофеля на колхозных рынках повысилась на 1,9%. Это связано с повышением удельного веса нового рынка, на котором цены выше. Очевидно, что выполняется равенство 1,235 · 1,019 = 1,259.

Если использовать обозначение структуры продажи d, то индексы (10.22), (10.23), (10.24) будут иметь вид:

$$\frac{\sum p_1 d_1}{\sum p_0 d_0} = \frac{\sum p_1 d_1}{\sum p_0 d_1} \cdot \frac{\sum p_0 d_1}{\sum p_0 d_0}. \qquad (10.26)$$

Можно выразить и абсолютное изменение средней величины с учетом изменения факторов – самого осредняемого признака и структуры (т. е. признака-веса):

$$\Delta\bar{p} = \Delta\bar{p}(p) + \Delta\bar{p}(d). \qquad (10.27)$$

По данным табл. 10.6 средняя цена картофеля повысилась на 43 руб./кг: $\Delta\bar{p}$ = 2,08 – 1,65 = 0,43 руб./кг; в том числе за счет самой цены $\Delta\bar{p}(p)$ = 2,08 – 1,68 = 0,4 руб./кг и за счет структурного фактора $\Delta\bar{p}(d)$ = 1,68 – 1,65 = 0,03 руб./кг.

И при относительном, и при абсолютном разложении эффект взаимодействия факторов – цены и структуры продажи – присоеди-

нился к оценке изменения цен. Если получить эту оценку в условиях базисного периода, то сравнение индексов

$$I_p = \frac{\sum p_1 d_1}{\sum p_0 d_1} \quad \text{и} \quad I_p = \frac{\sum p_1 d_0}{\sum p_0 d_0}$$

позволит выделить эффект совместного изменения факторов. По данным табл. 10.6 получаем:

$$I_p = \frac{\sum p_1 d_0}{\sum p_0 d_0} = \frac{2{,}079}{1{,}652} = 1{,}258.$$

Этот результат мало отличается от того, который был получен в условиях структуры продажи отчетного периода (I_p = 1,259), так что эффект взаимодействия факторов оказался незначителен и направлен на повышение средней цены.

Влияние структурных сдвигов может привести к неожиданным результатам: изменение себестоимости в целом по отрасли может оказаться большим, чем на отдельных предприятиях; или при выполнении производственной программы всеми предприятиями региона может оказаться, что регион в целом с программой не справился. Этот вопрос подробнее освещен в п. 10.7.

10.6. ПОСТРОЕНИЕ ИНДЕКСОВ ПРИ ОБОБЩЕНИИ ДАННЫХ ПО ЕДИНИЦАМ СОВОКУПНОСТИ И ПО ЭЛЕМЕНТАМ

Мы обсудили построение индексов при обобщении данных по многим товарам или «элементам» и при обобщении данных по единицам при наличии одного элемента (одного вида товара). В экономических расчетах приходится иметь дело с задачами построения индексов, объединяющих данные по единицам и по элементам.

Обозначим число элементов m, число единиц n. Обобщение данных при построении индексов можно подразделить на **три уровня:**

1) $n = 1$, $m > 1$ – индексный анализ проводится по одной единице (предприятию, магазину и т.д.) и группе элементов;

2) $n > 1, m = 1$ – индексный анализ по группе единиц и одному элементу (товару, виду продукции);

3) $n > 1$, $m > 1$— индексный анализ по группе единиц и элементов.

Таблица 10.7

Данные о покупках продовольственных товаров в районе города

Вид торговли	Товары	Базисный период			Отчетный период			Расчетные показатели	
		тонн q_0	руб./кг p_0	тыс. руб. q_0p_0	тонн q_1	руб./кг p_1	тыс. руб. q_1p_1	руб./кг q_1p_0	тыс. руб. q_0p_1
Рынок	Говядина	1,2	25,00	30,00	0,5	32,00	16,00	12,50	38,40
	Свинина	1,0	58,00	58,00	1,5	64,00	96,00	87,00	64,00
	Картофель	3,5	1,80	6,30	3,2	2,20	7,04	5,76	7,70
	Яблоки	4,2	4,00	16,80	3,0	6,00	18,00	12,00	25,20
	Крупа гречневая	0,5	5,50	2,75	0,3	6,50	1,95	1,65	3,25
Итого		–	–	113,85	–	–	138,99	118,91	138,55
Магазин 1	Говядина	–	–	–	2,0	18,00	36,00	29,752	–
	Картофель	3,0	1,60	4,80	4,5	1,80	8,10	7,20	5,40
	Яблоки	4,5	3,50	15,75	3,8	4,00	15,20	13,30	18,00
	Крупа гречневая	2,5	4,00	10,00	1,5	4,80	7,20	6,00	12,00
	Сахарный песок	4,0	5,00	20,00	3,5	6,50	22,75	17,50	26,00
Итого	по полному кругу	–	–	50,55	–	–	89,25	73,752	61,40
в том числе	по сопоставимому кругу	–	–	50,55	–	–	53,25	*44,00*	*61,40*

Продолжение

Вид торговли	Товары		Базисный период			Отчетный период			Расчетные показатели	
			тонн q_0	руб./кг p_0	тыс. руб. q_0p_0	тонн q_1	руб./кг p_1	тыс. руб. q_1p_1	руб./кг q_1p_0	тыс. руб. q_0p_1
Магазин 2	Говядина		2,0	18,00	36,00	2,0	19,00	38,00	36,00	38,00
	Яблоки		4,0	3,80	15,20	4,5	4,00	18,00	17,10	16,00
	Крупа гречневая		2,0	4,50	9,00	1,7	4,60	7,82	7,65	9,20
	Сахарный песок		3,2	5,80	18,56	–	–	–	–	(...)
Итого	по полному кругу		–	–	78,76	–	–	63,82	60,75	(...)
в том числе	по сопоставимому кругу				60,20			63,82	60,75	63,20
			Q_0	$\bar{P}_0$	$Q_0\bar{P}_0$	Q_1	$\bar{P}_1$	$Q_1\bar{P}_1$	$Q_1\bar{P}_0$	$Q_0\bar{P}_1$
Всего по всем видам торговли	Говядина		3,2	20,625	66,00	4,5	20,000	90,00	92,813	64,000
	Свинина		1,0	58,000	58,00	1,5	64,000	96,00	87,000	64,000
	Картофель		6,5	1,708	11,10	7,7	1,966	15,14	13,152	12,779
	Яблоки		12,7	3,760	47,75	11,3	4,531	51,20	42,488	57,544
	Крупа гречневая		5,0	4,35	21,75	3,5	4,849	16,97	15,225	24,245
	Сахарный песок		7,2	5,356	38,56	3,5	6,500	22,75	18,746	46,800
Итого			–	–	243,16	–	–	292,06	269,424	269,368

Последний тип задач характерен для муниципального управления, аналитической работы в региональных, ведомственных и федеральных статистических службах.

Предположим, нам нужно изучить потребление продовольствия в районе города. Собраны данные об объеме покупок товаров и ценах на рынках и в приватизированных магазинах (табл. 10.7).

По данным табл. 10.7 можно построить индексы объема продажи и цен для каждого вида торговли в отдельности, что соответствует данным типа 1; можно определить изменение объема продажи и цен на каждый из товаров по всем видам торговли, что соответствует данным типа 2; наконец, можно получить индексы объема продаж и цен по всем видам торговли и всем товарам. Произведем последовательно расчеты всех индексов, используя базисные веса в индексах объема и отчетные – в индексах цен.

Рынок.

$$I_q = \frac{118,91}{113,85} = 1,044 \cdot 100\% = 104,4\%;$$

$$I_p = \frac{138,99}{118,91} = 1,169 \cdot 100\% = 116,9\%;$$

$$I_w = \frac{138,99}{113,85} = 1,221 \cdot 100\% = 122,1\%.$$

Магазин 1. В этом магазине ассортимент товаров изменялся: в базисном периоде говядины не было в продаже, в отчетном – она появилась. В этом случае изменение цен определяется по сопоставимому кругу товаров, т.е. по четырем товарам. Обозначим сопоставимый круг товаров l, полный круг в базисном периоде – m_0, в отчетном – m_1. Тогда

$$I_p = \frac{\sum_{1}^{l} q_1 p_1}{\sum_{1}^{l} q_1 p_0}, \qquad (10.28)$$

$$I_p = \frac{53,25}{44,00} = 1,21 \cdot 100\% = 121\%.$$

Индекс объема продаж должен отразить изменение объема продаж тех товаров, которые были в базисном и продолжали продаваться в отчетном периоде, и, кроме того, изменение в объеме про-

дажи в связи с появлением нового товара (несопоставимого). Так что

$$I_q = \frac{\sum_{1}^{m_1} q_1 p_0}{\sum_{1}^{m_0} q_0 p_0}, \qquad (10.29)$$

т. е. этот индекс должен включать данные по полному кругу элементов (товаров): сопоставимым и несопоставимым. В нашем примере

$$I_q = \frac{\sum_{1}^{5} q_1 p_0}{\sum_{1}^{4} q_0 p_0}.$$

Возникает проблема **определения базисной цены для товара,** который имелся **только в отчетном периоде.** Возможны по крайней мере три ее решения.

1. Использование для несопоставимых элементов цены отчетного периода, т. е. числитель I_q представляется как

$$\left(\sum_{1}^{I} q_1 p_0 + \sum_{1}^{m_1 - 1} q_1 p_1\right).$$

По данным табл. 10.7

$$I_q = \frac{4400 + 3600}{5055} = 1{,}583 \cdot 100\% = 158{,}3\%.$$

Расчет I_q этим методом нарушает увязку индексов в систему: 1,21 · 1,583 = 1,9154 · 100% = 191,54%, тогда как по данным табл. 10.7 индекс выручки составил:

$$I_w = I_{qp} = \frac{89{,}25}{50{,}55} = 1{,}7656 \cdot 100\% = 176{,}6\%.$$

Однако выполняется увязка абсолютных изменений:

$\Delta w = \Delta_{qp} = \Delta_q p_0 + \Delta_p q_1 = (80{,}00 - 50{,}55) + (53{,}25 - 44{,}00) =$ = 38,70 тыс. руб.; если сравним непосредственно суммы выручки отчетного периода, то получим ту же величину: Δ_w = 89,25 – 50,55 = = 38,70 тыс. руб.

2. Использование условных значений базисных цен, которые определяются расчетным путем. Логично предположить, что если бы говядина была в базисном периоде, то цены на нее повысились примерно так же, как и на остальные товары. Это предположение можно записать в виде равенства: $i^{г}_{p(усл.)} = I_p$.

Так как

$$i^{г}_{p(усл)} = \frac{p^{г}_1}{p^{г}_{0(усл)}}, \quad \text{отсюда} \quad p^{г}_{0(усл)} = \frac{p^{г}_1}{i^{г}_{p(усл)}} = \frac{p^{г}_1}{I_p}.$$

В нашем примере условная цена говядины в магазине 1 в базисном периоде составила $p^{г}_{0(усл)} = \frac{18{,}00}{1{,}21} = 14{,}876$ тыс. руб./кг. Тогда

$I_q = \frac{73{,}752}{50{,}55} = 1{,}459 \cdot 100\% = 145{,}9\%$. Обратим внимание на то, что этот метод обеспечивает увязку индексов в систему, т. е. $I_q \cdot I_p = Iqp$ или I_w = 1,459 · 1,21 = 1,765. Однако взаимосвязь абсолютных изменений не выполняется: (73,752 – 50,55) + (53,25 – 44,00) = 23,202 + 9,25 = 32452 тыс. руб. вместо фактической величины изменения 38,70 тыс. руб.

3. Использование базисных цен других единиц или средних по совокупности. Например, при расчете I_q для магазина 1 использовать базисную цену на говядину в магазине 2. Тогда

$$I_q = \frac{44{,}00 + 2{,}0 \cdot 18{,}00}{50{,}55} = 1{,}583.$$

В данном случае из-за совпадения $p_0^{г}$ в магазине 2 с $p_1^{г}$ в магазине 1 этот метод привел к тому же результату, что и первый.

Использование средней базисной цены на говядину по всем видам торговли дает менее реальные результаты ввиду того, что значительный объем продаж говядины осуществляется на рынке, где цены были выше. Поэтому $\bar{p}^{г} > \bar{p}^{г}_0$. Значение индекса объема продаж составило:

$$I_q = \frac{44{,}00 + 2{,}0 \cdot 20{,}625}{50{,}55} = \frac{85{,}25}{50{,}55} = 1{,}686 \cdot 100\% = 168{,}6\%.$$

Это значение превышает все предыдущие.

Магазин 2. Здесь тоже ассортимент менялся: в отчетном периоде не торговали сахарным песком. Но это не вызывает трудностей в построении индексов, так как изменение объема продаж обусловлено как изменением продаж сопоставимых товаров, так и отсутствием продажи несопоставимого товара.

$$I_p \frac{\sum_1^3 q_1 p_1}{\sum_1^3 q_1 p_0} = \frac{63{,}82}{60{,}75} = 1{,}051 \cdot 100\% = 105{,}1\%;$$

$$I_q = \frac{\sum_1^3 q_1 p_0}{\sum_1^4 q_0 p_0} = \frac{60{,}75}{78{,}76} = 0{,}771 \cdot 100\% = 77{,}1\%.$$

Если нужно рассчитать $q_0 p_1$ для несопоставимого товара, то можно воспользоваться одним из рассмотренных способов.

Чтобы получить результаты по всем видам торговли, данные обобщаются определенным образом. При этом возможны два подхода.

Первый подход основан на суммировании данных по видам торговли (или отдельным предприятиям). Этот метод основан на данных отдельных хозяйственных единиц и поэтому называется **заводским методом.**

$$I_q = \frac{\sum_1^n \sum_1^{m_1} q_1 p_0}{\sum_1^n \sum_1^{m_1} q_0 p_0}; \qquad (10.30)$$

$$I_p = \frac{\sum_1^n \sum_1^l q_1 p_1}{\sum_1^n \sum_1^l q_1 p_0}, \qquad (10.31)$$

где n – число единиц совокупности;
m – число элементов, всего;
l – число сопоставимых элементов.

В этом случае один и тот же товар или вид продукции взвешивается по разным ценам в зависимости от того, где он учитывается.

Второй подход основан на обобщении данных по отдельным товарам независимо от места реализации. Рассчитываются для каждого товара сводные показатели количества и цены:

$$\sum_1^n q_1 = Q_1; \quad \sum_1^n q_0 = Q_0;$$

$$\overline{p}_{1i} = \frac{\sum_1^n q_1 p_1}{\sum_1^n q_1}; \quad \overline{p}_{0i} = \frac{\sum_1^n q_0 p_0}{\sum_1^n q_0}.$$

Затем данные обобщаются по всем товарам, при этом каждый из них взвешивается по средней цене для данного товара. Этот метод основан на обобщении с позиций совокупности, а не отдельных единиц и потому получил название **отраслевой метод:**

$$I_Q = \frac{\sum_1^{m_1} Q_1 \overline{p}_0}{\sum_1^{m_0} Q_0 \overline{p}_0}; \qquad (10.32)$$

$$I_p = \frac{\sum_1^l Q_1 \overline{p}_1}{\sum_1^l Q_1 \overline{p}_0}. \qquad (10.33)$$

При этом сопоставимость элементов определяется, исходя из условий совокупности в целом.

В нашем примере (табл. 10.7) были товары, не сопоставимые с позиций отдельных видов торговли, но в целом для торговли все они сопоставимы.

Вычислим индексы тем и другим методом:

по данным отдельных видов торговли

$$I_p = \frac{138{,}99 + 89{,}25 + 63{,}82}{118{,}91 + 73{,}752 + 60{,}75} = \frac{292{,}06}{253{,}412} = 1{,}153 \cdot 100\% = 115{,}3\%;$$

$$I_q = \frac{118{,}91 + 73{,}752 + 60{,}75}{113{,}85 + 50{,}55 + 78{,}76} = \frac{253{,}412}{243{,}16} = 1{,}042 \cdot 100\% = 104{,}2\%;$$

$$Iw = Iq \cdot Ip = 1{,}042 \cdot 1{,}153 = 1{,}2014 \cdot 100\% = 120{,}14\%;$$

по данным всех видов торговли, вместе взятых,

$$I_Q = \frac{269{,}422}{243{,}16} = 1{,}108 \cdot 100\% = 110{,}8\%;$$

$$I_{\bar{p}} = \frac{292{,}06}{269{,}424} = 1{,}084 \cdot 100\% = 108{,}4\%.$$

Получилось, что средние цены повысились в меньшей степени, чем в отдельных видах торговли: +8,4% против +14,5%. Это соотношение отражает влияние структурного фактора, изменение удельного веса продажи того или иного товара разными видами торговли. В частности, большое значение имела продажа говядины в отчетном периоде не только на рынке и в магазине 1, но и в магазине 2, где цены ниже.

Соотношение индекса средних цен и индекса цен без учета структурного фактора дает оценку структурных сдвигов:

$$I_{\text{структуры}} = I_{\bar{p}} : I_p = 1{,}084 : 1{,}153 = 0{,}941 \cdot 100\% = 94{,}1\%,$$

т. е. за счет изменения соотношения разных видов торговли в общем объеме реализации средние цены снизились на 5,9%.

Мы получили индекс структуры исходя из взаимосвязи индексов. Можно рассчитать его значение непосредственно по формуле

$$I_{структуры} = \frac{\sum_1^n \sum_1^m q_1 p_0}{\sum_1^m Q_1 \bar{p}_0} = \frac{118{,}91 + 73{,}752 + 60{,}75}{243{,}16} =$$

$$= \frac{253{,}412}{269{,}424} = 0{,}941 \cdot 100\% = 94{,}1\%.$$

Если стоит задача измерить влияние объема продаж на величину выручки от продажи, т. е. найти $I_{w(q)}$, то это влияние измеряет индекс, найденный по заводскому методу. Его величина отражает одновременно и изменение объема продажи во всех видах торговли, и изменение структуры продажи.

Алгебраически это можно представить в виде равенства:

$$I_q = I_Q \cdot \underbrace{I_{\bar{p}\left(\frac{q}{\Sigma q}\right)}}_{\substack{\text{индекс} \\ \text{структуры}}}, \tag{10.34}$$

$$\frac{\sum_1^n \sum_1^m q_1 p_0}{\sum_1^n \sum_1^m q_0 p_0} = \frac{\sum_1^m Q_1 \bar{p}_0}{\sum_1^m Q_0 \bar{p}_0} \cdot \frac{\sum_1^n \sum_1^m q_1 \bar{p}_0}{\sum_1^m Q_1 \bar{p}_0}. \tag{10.35}$$

где

$$\sum_1^n \sum_1^m q_0 p_0 = \sum_1^m Q_0 \bar{p}_0 = \sum_1^m \frac{\sum_1^n q_0 p_0}{\sum_1^n q_0} \cdot \sum_1^n q_0.$$

По данным табл. 10.7 $I_Q = 1{,}108$, $I_{\bar{p}\left(\frac{q}{\Sigma q}\right)} = 0{,}941$, тогда

$$I_q = 1{,}108 \cdot 0{,}941 = 1{,}042 \cdot 100\% = 104{,}2\%.$$

Мы рассмотрели систему индексов, в которой использовались разновзвешенные индексы: индексы объемного фактора (количества) с базисными весами, индексы качественного фактора (цен) – с отчетными весами. Плюсы и минусы такого метода построения индексов уже обсуждались.

Если необходимо все индексы построить на базисных весах, то в системе индексов появляется индекс совместных изменений. С экономической точки зрения его часто называют индексом смещения ассортимента, так как он показывает изменение в реализации доли товаров с разным темпом снижения цен:

$$\frac{\sum_1^m Q_1 \bar{p}_1}{\sum_1^m Q_1 \bar{p}_0} = \frac{\sum_1^m Q_1 \bar{p}_0}{\sum_1^m Q_0 \bar{p}_0} \cdot \frac{\sum_1^m Q_0 \bar{p}_1}{\sum_1^m Q_0 \bar{p}_0} \left(\frac{\sum_1^m Q_1 \bar{p}_1}{\sum_1^m Q_0 \bar{p}_0} : \frac{\sum_1^m Q_0 \bar{p}_1}{\sum_1^m Q_0 \bar{p}_0} \right). \tag{10.36}$$

Можно оценить эффект совместного изменения признаков q и p в системе индексов, построенных по заводскому методу:

$$\frac{\sum\sum q_1 p_1}{\sum\sum q_0 p_0} = \frac{\sum\sum q_1 p_0}{\sum\sum q_0 p_0} \cdot \frac{\sum\sum q_0 p_1}{\sum\sum q_0 p_0} \left(\frac{\sum\sum q_1 p_1}{\sum\sum q_1 p_0} : \frac{\sum\sum q_0 p_1}{\sum\sum q_0 p_0} \right). \quad (10.37)$$

Вполне возможны различия в значениях индексов совместных изменений, полученных по формуле (10.36) или (10.37). Это различие может возникнуть из-за разного охвата элементов: в первом случае сопоставимость определяется с общеотраслевых позиций, во втором – с позиций отдельного предприятия.

Итак, вы получили представление о способах построения индексов при обобщении данных и по многим товарам, видам продукции и по магазинам, рынкам, предприятиям. Какой метод выбрать в каждом конкретном случае, вам часто придется решать самим, ведь далеко не всегда имеется инструкция по проведению расчетов.

10.7. ГРАНИЦЫ И УСЛОВИЯ ПРИМЕНЕНИЯ ИНДЕКСНОГО МЕТОДА

Каждый метод ориентирован на особые представления изучаемого объекта, на особую его модель. Индексный метод предполагает, что связь между признаками является **жестко детерминированной,** которая проявляется как в каждом отдельном случае (для отдельного товара, вида продукции, предприятия и т. д.), так и в совокупности. Связь, изучаемая с помощью индексов, выражается в виде **уравнения связи:**

либо мультипликативного

$$y = x_1 x_2 \ldots x_k,$$

либо аддитивного

$$y = x_1 + x_2 + \ldots + x_k.$$

Вид функции, число переменных факторов (сомножителей или слагаемых) определяются нашими представлениями о логике изучаемой связи. Многофакторная мультипликативная модель строится путем последовательного расчленения одного из факторов на составляющие.

Например, можно записать следующее уравнение связи:

Объем произведенной продукции	=	Число отработанных человеко-часов	·	Средняя часовая выработка.

Эту модель можно детализировать. Она будет включать не два, а пять факторов:

Объем произведенной продукции	=	Средняя списочная численность работников	·	Доля рабочих в среднесписочной численности работников	·	Среднее число дней работы	·	Средняя продолжительность рабочего дня	·	Средняя часовая выработка.

Если мультипликативная модель имеет в качестве результативного первичный признак, то она называется полной[1].Примером такой модели является вышеуказанная модель. Разделив обе части равенства на первый фактор, получим *неполную* модель среднечасовой выработки работника.

Представление связи как жестко детерминированной является условным, так как связи социально-экономических явлений носят стохастический характер.

Если представить мультипликативную модель как двухфакторную, т.е. $y = x_1 x_2$, то в целом по совокупности уравнение имеет вид $y = ax$. Коэффициент a является коэффициентом связи между y и x. Он передает *прямое* влияние фактора x на результат y. Для нашего примера величина отработанных человеко-часов передает влияние среднечасовой выработки на объем продукции. Однако выработка влияет на результат не только непосредственно, но и через другие факторы: уровень выработки может определять численность рабочих, их долю в списочном составе, фактическую продолжительность рабочего дня. В корреляционном анализе, измеряя корреляцию между результатом и фактором, мы получаем полную меру корреляции независимо от того, как реализуется связь – непосредственно или опосредованно. В индексном анализе мы измеряем только прямое влияние изменения фактора на изменение результата.

При построении уравнения связи иногда допускаются отступления от логики ради обеспечения увязки признаков, получения жестко детерминированного выражения связи. Поэтому можно встретить уравнения связи, в которых не все составляющие элементы экономически обоснованы, нередки случаи появления среди факторов обратных величин.

---

[1] *Адамов В. Е.* Факторный индексный анализ. Методология и проблемы. – М.: Статистика, 1977. – С. 101.

Приведем пример недостаточно обоснованного уравнения связи:[1]

$$\frac{\text{Балансовая прибыль}}{\text{Производственные фонды}} = \frac{\text{Балансовая прибыль}}{\text{Прибыль от реализации}} \cdot \frac{\text{Прибыль от реализации}}{\text{Стоимость реализованной продукции}} \cdot \frac{\text{Стоимость реализованной продукции}}{\text{Основные фонды}} \cdot \frac{\text{Основные фонды}}{\text{Производственные фонды}}.$$

Трудно представить, чтобы рост доли основных фондов вызывал рост балансовой прибыли.

При мультипликативной связи индексов относительные выражения приростов факторов связаны аддитивно. Например,

$$\frac{\sum q_1 z_1}{\sum q_0 z_0} = \frac{\sum q_1 z_0}{\sum q_0 z_0} \cdot \frac{\sum q_1 z_1}{\sum q_1 z_0},$$

однако

$$\frac{\Delta q x_0 = \sum q_1 z_0 - \sum q_0 z_0}{\sum q_0 z_0} + \frac{\Delta z q_1 = \sum q_1 z_1 - \sum q_1 z_0}{\sum q_0 z_0} = \frac{\sum q_1 z_1 - \sum q_0 z_0}{\sum q_0 z_0}.$$

Чем больше различаются индексы отдельных факторов, тем больше сумма относительных приростов отличается от темпа прироста результата. Например,

$$I_{\substack{\text{производ.}\\ \text{труда}}} \cdot I_{\substack{\text{затрат}\\ \text{времени}}} = I_{\substack{\text{объема}\\ \text{продукции}}}.$$

Если известно, что значение первого индекса 1,2, второго – 1,05, то

$$I_{\substack{\text{объема}\\ \text{продукции}}} = 1{,}26, \text{ тогда как } \frac{20\%}{100} + \frac{5\%}{100} = 25\%,$$

т. е. переход от темпов роста к темпам прироста приводит к определенным трудностям в интерпретации количественного влияния факторов на результат.

Наконец, решение вопроса об измерении эффекта отдельных факторов и их совместного изменения всегда условно.

---

[1] *Адамов В. Е.* Факторный анализ. Методология и проблемы. – М.: Статистика, 1977.

Все предыдущее изложение было ориентировано на мультипликативную модель. При аддитивной связи признаков индексный анализ проводится по следующей формуле:

$$I_y = \frac{y_1}{y_0} = \frac{x_1 + z_1}{x_0 + z_0} = \frac{x_1}{x_0} \cdot \frac{x_0}{x_0 + z_0} + \frac{z_1}{z_0} \cdot \frac{z_0}{x_0 + z_0} = I_{x_{1/0}} d_{x_0} + I_{z_{1/0}} d_{z_0},$$

т. е. общее изменение результата зависит от изменения каждого фактора и его доли в базисной величине результата. Приведем пример (табл. 10.8).

Таблица 10.8

Численность работников на заводе

Период	Всего, чел.	В том числе	
		заняты физическим трудом	заняты умственным трудом
Базисный	1000	700	300
Отчетный	800	640	160

Общее изменение численности работников может быть представлено как результат изменения численности занятых умственным и физическим трудом и их доли в общей численности работников:

$$\frac{640}{700} \cdot 0{,}7 + \frac{160}{300} \cdot 0{,}3 = 0{,}914 \cdot 0{,}7 + 0{,}533 \cdot 0{,}3 = 0{,}7997.$$

Этот результат отличается от 0,8 (800 : 1000) только за счет округления в расчетах.

Одним из сложнейших вопросов индексного анализа является оценка структурных сдвигов. Этот фактор может приводить к парадоксальным результатам в индексах.

Возьмем для примера условные данные о работе трех химических предприятий одного района (табл. 10.9).

На каждом из этих предприятий рост объема производства сопровождался повышением производительности труда. Средняя выработка на одного работника по предприятиям составила в отчетном квартале соответственно 12,6 тыс. руб., 21,26 тыс. руб. и 50,40 тыс. руб. В прошлом квартале средняя выработка составляла 12 тыс. руб., 20,25 тыс. руб. и 48 тыс. руб. Сравнение этих данных

Таблица 10.9

Показатели работы химических предприятий района

Предприятия	Отчетный квартал		Прошлый квартал	
	товарная продукция в сопоставимых ценах, тыс. руб.	средняя списочная численность работающих, чел.	товарная продукция в сопоставимых ценах, тыс. руб.	средняя списочная численность работающих, чел.
Завод по производству минеральных удобрений	40320	3200	14400	1200
Фабрика искусственного меха	14882	700	12150	600
Завод пластиков	10080	200	9600	200
В целом по химическим предприятиям	65282	4100	36150	2000

показывает, что выработка росла равномерно на всех предприятиях: на первом заводе (12,6 : 12 = 1,05, или 105%); на втором – (21,26 : 20,25 = 1,0498, или 104,98%); на третьем – (50,4 : 48 = 1,05, или 105%). Если рассчитать среднюю выработку по всем трем заводам и определить ее динамику, то результат покажется невероятным. В отчетном квартале средняя выработка в целом составила 15,92 тыс. руб. (65282 тыс. руб. : 4100 чел.), а в прошлом квартале – 18,075 тыс. руб. (36150 тыс. руб. : 2000 чел.), т. е. средняя выработка по отрасли снизилась на 12% (15,92 тыс. руб. : 18,075 тыс. руб. = 0,88 · 100% = 88%).

Этот результат объясняется тем, что динамика среднеотраслевой выработки учитывает не только, какой была динамика выработки на отдельных предприятиях, но и как изменялось распределение работников между ними. Ведь уровень средней выработки на одного работника на отдельных предприятиях различается достаточно сильно: максимален он на заводе пластиков, минимален — на заводе минеральных удобрений. Именно на этом заводе численность работников возросла почти в 3 раза. Доля этого завода в численности работающих составляла 60% в прошлом квартале и 78,1% – в отчетном квартале. Отсюда и совокупный результат.

Можно измерить общее изменение выработки без учета изменения соотношений между предприятиями: если сравнить общий

объем товарной продукции в сопоставимых ценах в отчетном квартале с тем объемом, который был бы получен, если бы выработка на каждом заводе оставалась прежней. Величина такой «условной» товарной продукции составит: 12 тыс. руб. 3200 чел.+ 20,25 тыс. руб. • 700 чел. + 50,4 тыс. руб. • 200 чел.= 62175 тыс. руб. Суммарные показатели товарной продукции 65282 тыс. руб. и 62175 тыс. руб. различаются только за счет выработки, значит их сравнение покажет динамику средней выработки по всем трем предприятиям без учета динамики численности работников. Действительно, получаем, что и в целом рост выработки составил +5% (65282 тыс. руб : 62175 тыс. руб. = 1,05, или 105%). Вот теперь нет никакого противоречия между результатами работы отдельных предприятий и отрасли. Но чтобы разобраться в этом, нужно знать, какой методикой пользовался статистик, как он получил те или иные результаты.

10.8. КОМПЛЕКСНОЕ ИСПОЛЬЗОВАНИЕ ИНДЕКСНОГО И РЕГРЕССИОННОГО МЕТОДОВ АНАЛИЗА

Применяя индексный анализ часто ощущается недостаточность использования только этого метода прежде всего из-за того, что уравнение связи как жестко детерминированная функция может быть построено лишь для «ближайшего» круга факторов, тех, которые непосредственно составляют результат. Такие факторы могут оказаться недостаточными для объяснения его динамики. Эта особенность анализа связи на основе жестко детерминированного выражения результата очевидна, например, при постатейном анализе себестоимости продукции. Вроде бы такой анализ обеспечивает точность показателей связи. Так, если изменится норма расхода того или иного материала и заготовительные расходы на него, можно точно указать, на какую величину снизится (повысится) себестоимость продукции данного вида. Вместе с тем «функциональный» анализ себестоимости продукции вскрывает лишь непосредственное различие себестоимости из-за различий величин, прямо входящих в ее расчет, но не вскрывает причин самих этих различий. Можно установить, насколько на предприятиях, производящих однородную продукцию, различаются нормы расхода сырья, сдельные расценки и т.п. Но само по себе выяснение этих факторов еще ничего не говорит об их причинах, которые зависят от уровня технического оснащения предприятия, квалификации его работников, организации производства и т.п. Эти факторы воздействуют на ве-

личину себестоимости не непосредственно, а через величины, прямо учитываемые в ее расчетах: через нормы расхода материалов, расценки и ставки заработной платы, суммы амортизации и другие виды производственных затрат. В отличие от ближайших факторов такие факторы принадлежат к другому, так сказать, «глубинному уровню изучаемой структуры.

Далеко не всегда можно выявить механизмы связи между глубинными причинами и результатом в силу их большей отдаленности, многоплановости влияния. Не всегда можно включить их в жестко детерминированное уравнение связи путем последовательного развертывания признаков. Это приводит к комплексному использованию методов, основанных на жесткой детерминации признаков, и методов, не ориентированных на такой характер связей.

Понять в полной мере задачи интеграции разных методов статистического изучения связей можно с помощью графа связей. **Граф связей** учитывает непосредственные, т. е. причинные связи, которые предполагают изменение x_j при изменении влияющего на него x_j при постоянстве всех прочих факторов. Асимметричность причинных связей отражается в направленности дуг графа (**дуга** – соединение вершин графа, т. е. точек, соответствующих элементам структуры).

Нередко оказывается, что разобраться в системе связей можно только тогда, когда граф связей будет включать не только факторы – признаки данной единицы совокупности, непосредственно определяемые в процессе ее функционирования (**эндогенные**), но и факторы, не зависящие от нее, но влияющие на изучаемый результат (**экзогенные**). Если первые образуют систему признаков и могут находиться в жестко детерминированной связи с изучаемой результативной переменной вследствие устойчивости связи в рамках единицы совокупности, то вторые не являются признаками изучаемой единицы, потому их связь с результатом неустойчива, стохастична. Как правило, действие экзогенных факторов опосредовано эндогенными переменными, формирующими результат. Потребность сочетания разных уровней анализа – «вышележащего», на котором могут иметь место жестко детерминированные связи, и «нижележащего», на котором они отсутствуют, вызывает **интеграцию разных методов анализа.** Например, изучая, почему произведен тот или иной объем валовой продукции, весьма важно не останавливаться на анализе уравнения связи, подобного приведенному в п. 10.7, включающего признаки, определяемые на уровне предприятия, а перейти на другой уровень анализа. Выявить, например, чем обусловлена та или иная величина среднечасовой выработки рабочих. Для этого необходимо перейти к совокупности рабочих и их признакам (уровню квалификации, стажу, умению организовывать процесс труда и т. д.).

На рис. 10.2 изображен гипотетический граф связей, в котором элементы «высшего» уровня структуры обозначены как x_j ($j = 1,2,\ldots, m$). Это факторы, находящиеся в жесткой связи с результатом (y); z_l – глубинные факторы, принадлежащие другому уровню изучаемой структуры связей ($l = 1,2,\ldots,z$). Эти факторы находятся в стохастической связи с x_j и y.

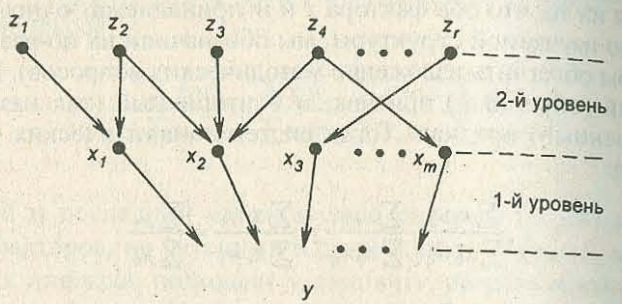


Рис. 10.2. Граф связей

Рис. 10.2 показывает, что используя индексы (или другой метод анализа, основанный на жестко детерминированных связях) мы ограничиваемся только одним уровнем структуры связей, включающим отношения между y и x_j и не затрагиваем связи между z_l и x_j. Используя только этот путь анализа, мы можем не выяснить причины изменения результата. Кроме того, в анализе только жестко детерминированных связей [$x_j \rightarrow y$], каждый из x_j выступает как независимая величина, тогда как они могут быть связаны, как непосредственно, так и через общие детерминирующие факторы. Эта связь является стохастической и может быть измерена с помощью соответствующих методов.

Методика комплексного использования индексного и регрессионного анализа такова. Определяется жестко детерминированное уравнение связей $y = f(x_1, \ldots, x_m)$, на основе графа связей строится уравнение регрессии для каждой компоненты (x_j):

$$\hat{x}_j = a_0 + a_1 z_1 + \ldots + a_2 z_2,$$

где z_l – так называемые глубинные причины.

Оценив значимость параметров отдельных регрессий, устанавливается круг причин для каждого из x_j, общий круг причин для x_i и x_j $(i \neq j)$. Используя полученное на основе регрессии значение $\hat{x}_j$, мы получаем возможность измерить влияние каждого из учтенных в регрессии факторов на y. Таким образом в анализе участвуют функционально и нефункционально связанные факторы.

Остановимся подробнее на методике комплексного использования методов. Рассмотрим простейший случай. Пусть изучаемый результативный признак может быть представлен как жестко детерминированная двухфакторная мультипликативная функция $y = xw$ (несмотря на то, что оба фактора x и w принадлежат одному и тому же уровню изучаемой структуры, мы обозначили их по-разному для того, чтобы облегчить изложение методических вопросов). Пусть x – первичный (объемный) признак, w – вторичный (так называемый количественный) признак. Тогда система аналитических индексов имеет вид:

$$\frac{\sum x_1 w_1}{\sum x_0 w_0} \cdot \frac{\sum x_1 w_1}{\sum x_1 w_0} = \frac{\sum x_1 w_1}{\sum x_0 w_1} = \frac{\sum y_1}{\sum y_0}$$

или

$$I_{y(x)} \cdot I_{y(w)} = I_y.$$

На следующем этапе анализа перейдем на другой уровень структуры связей. Введем различные обозначения для факторов, влияющих на x и на w:

$$x = \hat{f}(u_1, u_2, \ldots, u_m), \; w = \hat{f}/v_1, \ldots, v_2.$$

Заметим, что в принципе, как уже отмечалось, круг факторов для x и w может частично совпадать. В случае непосредственной связи между x и w, та из переменных, которая является независимой, может включаться в регрессию другой (зависимой) переменной. Положим, что круг объясняющих переменных для x и w остался неизменным в отчетном периоде по сравнению с базисным. Принимая регрессии линейными, имеем по две регрессии для x и w, описывающих базисное и отчетное состояние x и w. .

Для базисного периода:

$$\hat{x}_0 = a_{00} + a_{01} \cdot u_{01} + \ldots + a_{0m} \cdot u_{0m};$$

$$\hat{w}_0 = b_{00} + b_{01} \cdot v_{01} + \ldots + b_{0r} \cdot v_{0r}$$

Для отчетного периода:

$$\hat{x}_1 = a_{10} + a_{11} \cdot u_{11} + \ldots + a_{1m} \cdot u_{1m};$$

$$\hat{w}_1 = b_{10} + b_{11} \cdot v_{11} + \ldots + b_{1r} \cdot v_{1r},$$

где первый подстрочный значок в каждой регрессии обозначает период, к которому она относится, второй – номер параметра или переменной, соответственно.

Введем в индекс $I_{y(x)}$ расчетное значение $\hat{x}$, получим следующую систему индексов:

$$I_{y(x)} = \frac{\sum x_1 w_0}{\sum x_0 w_0} = \frac{\sum x_1 w_0}{\sum \hat{x}_1 w_0} \cdot \frac{\sum \hat{x}_1 w_0}{\sum \hat{x}_0 w_0} \cdot \frac{\sum \hat{x}_0 w_0}{\sum x_0 w_0}. \qquad (10.38)$$

Первый и последний индексы этой системы (10.38) измеряют влияние факторов, не учтенных в регрессии $x = \hat{f}(u_1, \ldots, u_m)$. Сравнение этих индексов позволяет установить, регрессия какого периода точнее описывает фактические данные. Если регрессии построены правильно, то расхождения фактических и расчетных значений x и для базисного и для отчетного периодов будут незначительны, и оба индекса будут близки к единице.

Центральная роль принадлежит второму индексу системы – он измеряет влияние на y изменений в расчетных значениях $\hat{x}$. Расхождение между $\sum \hat{x}_1 w_0$ и $\sum \hat{x}_0 w_0$ может возникнуть как вследствие изменений значений переменных $u_1, \ldots, u_m$, так и в результате изменений силы их влияния на x – коэффициентов регрессий $a_{11}, a_{12}, \ldots, a_{1m}$ по сравнению с $a_{01}, a_{02}, \ldots, a_{0m}$. Раздельную оценку влияния на y глубинных факторов u и силы их воздействия a можно получить на основе специальной системы индексов. При этом рекомендуется первичным считать значение переменной, а вторичным – коэффициент регрессии[1].

Получим:

а) систему индексов, измеряющих влияние на y изменения значений переменных u:

[1] *Юзбашев М., Рудакова Р.* Регрессионные модели и индексы в анализе сельскохозяйственных предприятий // Вестник статистики. – 1976. – № 5. – С. 56 – 66.

$$I_{y\{\hat{x}(u_1)\}} = \frac{\sum(a_{00} + a_{01}\cdot u_{11} + \ldots + a_{0m}\cdot u_{01m})W_0}{\sum(a_{00} + a_{01}\cdot u_{01} + \ldots + a_{0m}\cdot u_{0m})W_0}$$

$$\cdot \quad \cdot \quad \cdot$$

$$I_{y\{\hat{x}(u_m)\}} = \frac{\sum(a_{00} + a_{01}\cdot u_{01} + \ldots + a_{0m}\cdot u_{1m})W_0}{\sum(a_{00} + a_{01}\cdot u_{01} + \ldots + a_{0m}\cdot u_{0m})W_0};$$

б) систему индексов, измеряющих влияние на y изменений интенсивности связей между x и u (a_l):

$$I_{y\{\hat{x}(a_{01})\}} = \frac{\sum(a_{00} + a_{11}\cdot u_{11} + \ldots + a_{0m}\cdot u_{1m})W_0}{\sum(a_{00} + a_{01}\cdot u_{11} + \ldots + a_{0m}\cdot u_{1m})W_0}$$

$$\cdot \quad \cdot \quad \cdot$$

$$I_{y\{\hat{x}(a_{01})\}} = \frac{\sum(a_{01} + a_{01}\cdot u_{11} + \ldots + a_{1m}\cdot u_{1m})W_0}{\sum(a_{00} + a_{01}\cdot u_{11} + \ldots + a_{0m}\cdot u_{1m})W_0};$$

в) индекс, учитывающий изменение свободного члена уравнения регрессии (a_0):

$$I_{y\{\hat{x}(a_0)\}} = \frac{\sum(a_{10} + a_{11}\cdot u_{11} + \ldots + a_{1m}\cdot u_{1m})W_0}{\sum(a_{00} + a_{11}\cdot u_{11} + \ldots + a_{1m}\cdot u_{1m})W_0}.$$

Тогда

$$I_{y\{\hat{x}\}} = I_{y\{\hat{x}(a_0)\}}\cdot\prod_{(l)} I_{y\{\hat{x}(u_l\}}\cdot\prod_{(l)} I_{y\{\hat{x}(a_l)\}}.$$

Очевидно, что на основе приведенных формул могут быть получены и соответствующие абсолютные эффекты

$$\Delta y\{\hat{x}(a_0)\},\ \Delta y\{\hat{x}(u_l)\},\ \Delta y\{\hat{x}(a_l)\}.$$

Точно так же может быть проанализировано влияние на факторов, детерминирующих W.

Покажем применение описанной методики анализа на условном **примере.** Предположим, что изучается работа каменноугольных шахт одного треста. В качестве результативного признака выступает среднесменный объем добычи угля (W), который может быть

представлен как произведение двух факторов: численности рабочих на подземных работах (N) и среднесменной добычи угля на одного подземного рабочего (V) – $W = N\cdot V$. По данным табл. 10.10 определим, как изменился среднесменный объем добычи угля в целом по тресту в отчетном периоде по сравнению с базисным и как на это изменение повлияло изменение численности подземных рабочих и среднесменной добычи на одного рабочего.

Таблица 10.10

№ шахты	Базисный период			Отчетный период			Условный среднесменный объем добычи, т
	среднесменная добыча на 1 подземного рабочего, т	число подземных рабочих	среднесменный объем добычи, т	среднесменная добыча на 1 подземного рабочего, т	число подземных рабочих	среднесменный объем добычи, т	
A	V_0	N_0	W_0	V_1	N_1	W_1	N_1V_0
1	5	80	400	5	90	450	450
2	10	100	1000	12	110	1320	1100
3	9	120	1080	12	120	1440	1080
4	7	110	770	9	130	1170	910
5	6	85	510	7	90	630	540
6	9	80	720	10	85	850	765
7	6	90	540	7	90	630	540
8	11	150	1650	12	135	1620	1485
9	5	80	400	5	85	425	425
10	7	95	665	6	100	600	700
Σ	×	990	7735	×	1035	9135	7995

Среднесменный объем добычи угля в целом по тресту увеличился:

$$I_{w(1/0)} = \frac{\sum W_1}{\sum W_0} = \frac{9135}{7735} = 1{,}18,$$

объем добычи вырос на 18%, что составило в абсолютном выражении 1400 т.

Численность подземных рабочих в данный период увеличилась и за счет этого среднесменный объем добычи вырос следующим образом:

$$I_{w(N1/0)} = \frac{\sum N_1 \cdot \sum V_0}{\sum N_0 \cdot \sum V_0} = \frac{7935}{7735} = 1{,}033.$$

Как видим, этот фактор в меньшей степени способствовал росту общего объема добычи. За счет него объем добычи вырос только на 3,3% или на 260 т.

Основная роль в общем изменении результата принадлежит интенсивному фактору – росту среднесменной добычи угля на одного подземного рабочего:

$$I_{w(V1/0)} = \frac{\sum N_1 V_1}{\sum N_1 V_0} = \frac{9135}{7995} = 1{,}142,$$

т. е. за счет роста среднесменной добычи угля на одного подземного рабочего общий объем добычи вырос на 14,2% или на 1140 т.

Вычисленные индексы образуют систему индексов:

$$I_{w(N)} \cdot I_{w(v)} = I_w;$$

$$1{,}033 \cdot 1{,}142 = 1{,}18.$$

Как отмечает Г. И. Бакланов, влияние каждого фактора на относительное изменение общей абсолютной величины можно получить, выразив соответствующую абсолютную разность в процентах к общей абсолютной величине в базисном периоде[1]. Вычисляя относительное влияние факторов таким образом, мы получим аддитивное разложение относительного изменения результативного признака. При этом относительная оценка влияния первичного признака (N) будет той же самой (+3,3%), а относительная оценка влияния вторичного признака (v) изменится и составит:

$$\frac{\sum N_1 V_1 - \sum N_1 V_0}{\sum N_0 V_0} = \frac{1140}{7735} \cdot 100 = 14{,}7\%.$$

Тогда 3,3% + 14,7% = 18,0%.

[1] *Бакланов Г. И.* Некоторые вопросы индексного метода. – М.: Статистика, 1972. – С. 15 – 16.

Ввиду того, что основная роль в общей динамике объема добычи принадлежит производительности труда – среднесменной добыче одного подземного рабочего, на следующем этапе анализа рассмотрим, за счет каких факторов сложится тот или иной уровень производительности труда и как изменение этих факторов сказалось на величине общего объема среднесменной добычи угля по тресту.

Среднесменная добыча подземного рабочего определяется многими факторами, среди которых можно назвать как характеристики рабочих (стаж, квалификация и т. д.), так и характеристики условий труда (используемая техника, степень механизации производственных процессов и др.) и разрабатываемого угольного пласта (длина лавы, мощность пласта и т. д.). Все эти факторы имеют стохастическую связь со среднесменной добычей рабочего, через нее оказывая влияние на общий объем добычи. Предположим, что из всего множества факторов главными и в отчетном и в базисном периодах оказались только два: мощность пласта и уровень механизации навалки угля. Данные по этим факторам по каждой из десяти шахт приведены в табл. 10.11.

Таблица 10.11

№ шахты	Базисный период			Отчетный период		
	среднесменная добыча на 1 подземного рабочего, т	мощность пласта, м.	уровень механизации навалки угля, %	среднесменная добыча на 1 подземного рабочего, т	мощность пласта, м	уровень механизации навалки угля, %
A	V_0	M_0	K_0	V_1	M_1	K_1
1	5	75	48	5	74	45
2	10	108	82	12	108	90
3	9	116	85	12	114	90
4	7	98	82	9	95	95
5	6	91	45	7	88	55
6	9	125	100	10	125	100
7	6	85	35	7	85	45
8	11	14	82	12	115	85
9	5	75	67	5	70	70
10	7	101	64	6	90	65

Прежде всего определим средние значения признаков, средние квадратические отклонения и коэффициенты вариации (табл. 10.12).

Таблица 10.12

Период	Средние величины			Средние квадратические отклонения			Коэффициенты вариации, %		
	V	$\bar{M}$	$\bar{K}$	S_v	S_M	S_k	V_v	V_M	V_k
Базисный	7,5	99,0	69,0	2,01	15,80	19,81	26,8	16,0	28,7
Отчетный	8,5	96,4	74,0	2,73	17,46	19,73	32,1	18,1	26,7

Сравнение отчетных данных с базисными свидетельствует: о возрастании средних значений тех признаков, которые отражают функционирование шахт (эндогенных); среднее значение экзогенного признака – мощности пласта – несколько снизилось. Возросла вариация шахт по величине среднесменной выработки одного подземного рабочего и по мощности пласта, тогда как по уровню механизации навалки угля наметилось некоторое выравнивание данных.

Вычисленные значения коэффициентов парной корреляции указывают на тесную связь между признаками (табл. 10.13).

Таблица 10.13

Период	r_{vM}	r_{vM}	r_{Mk}
Базисный	0,903	0,726	0,819
Отчетный	0,893	0,761	0,780

И в том и в другом периоде среднесменная добыча рабочего теснее коррелирует с мощностью пласта, нежели с уровнем механизации навалки угля. Однако наметилось некоторое снижение величины *rvm* при повышении r_{vM}. Сравнение коэффициентов парной корреляции зависимой переменной (V) с независимыми переменными и корреляции последних между собой свидетельствует о колинеарности факторов – их тесной линейной связи. При таком соотношении нецелесообразно построение множественной регрессии, куда бы входили оба названных фактора – и мощность пласта и коэффициент механизированной навалки угля. Поэтому построим

парную регрессию, описывающую зависимость среднесменной добычи одного рабочего только от мощности пласта: $\hat{v} = a + b$.

Для базисного периода уравнение парной регрессии:

$$\hat{V}_0 = a_0 + b_0 M_0 = -3{,}885 + 0{,}115\ V_0;$$

для отчетного периода:

$$\hat{V}_1 = a_1 + b_1 M_1 = -5{,}0 + 0{,}140\ M_1.$$

Так как вариация зависимой переменной превосходит вариацию независимой переменной ($v_v > v_M$), свободный член уравнения регрессии в обоих периодах – отрицательная величина ($a < 0$). Сравнение коэффициентов регрессии b_0 и b_1 показывает, что сила влияния данного фактора на среднесменную добычу рабочего растет, а теснота связи падает ($r_{v_1 M_1} < r_{v_0 M_0}$). Если коэффициент детерминации в базисном периоде составил $r^2_{v_0 M_0} = 81{,}54\%$, то в отчетном – $r^2_{v_1 M_1} = 79{,}74\%$.

Мощность пласта не входит в жестко детерминированное выражение общего среднесменного объема добычи угля, которое мы анализировали с помощью индексов. Однако этот фактор так же может быть учтен в анализе через регрессию $v = \hat{f}(M)$ и включение в индексы расчетных значений $\hat{v}$. В этом случае индекс, характеризующий влияние изменения среднесменной добычи одного подземного рабочего на величину общей среднесменной добычи угля, должен быть представлен как произведение трех индексов:

$$I_{w(v)} = \frac{\sum N_1 V_1}{\sum N_1 \hat{V}_1} \cdot \frac{\sum N_1 \hat{V}_1}{\sum N_1 \hat{V}_0} \cdot \frac{\sum N_1 \hat{V}_0}{\sum N_1 V_0}.$$

Расчетные значения среднесменной добычи угля на одного подземного рабочего и соответствующие расчетные значения общей среднесменной добычи представлены в табл. 10.14.

Подставляя в записанную систему индексов расчетные значения среднесменного объема добычи, получаем:

$$I_{w(v)} = \frac{9135}{9003} \cdot \frac{9003}{7912} \cdot \frac{7912}{7995} = 1{,}0146 \cdot 1{,}1378 \cdot 0{,}9896 = 1{,}1425.$$

Сопоставление первого и последнего индексов показывает, что базисная регрессия $v = \hat{f}(M)$ точнее описывает исходные данные. Этого следовало ожидать, так как $r_{v_0 M_0} > r_{v_1 M_1}$. Средний из трех ин-

Таблица 10.14

№ шахты	Расчетные значения							
	среднесменной добычи одного рабочего, т*				среднесменного объема добычи, т			
A	$\hat{V}_0$	$\hat{V}_1$	$\hat{V}'$	$\hat{V}''$	$\hat{V}_0 N_1$	$\hat{V}_1 N_1$	$\hat{V}'_{N_1}$	$\hat{V}''_{N_1}$
1	4,9	5,4	4,6	6,5	441	486	414	585
2	8,5	10,1	8,5	11,2	935	1111	935	1232
3	9,4	11,0	9,2	12,1	1128	1320	1104	1452
4	7,4	8,3	7,0	9,4	962	1079	910	1222
5	6,6	7,3	6,2	8,4	594	657	658	756
6	10,5	12,5	10,5	13,6	893	1062	892	1156
7	5,9	6,9	5,9	8,0	531	621	531	720
8	9,2	11,1	9,3	12,2	1242	1499	1256	1647
9	4,9	4,8	4,2	5,9	416	408	357	502
10	7,7	7,6	6,5	8,7	770	760	650	870
Итого	×	×	×	×	7912	9003	7607	10142

* $\hat{V}_0 = a_0 + в_0 M_0$; $\hat{V}_1 = a_1 + в_1 M_1$;
$\hat{V}' = a_0 + в_0 M_1$; $\hat{V}'' = a_0 + в_1 M_1$.

дексов отражает динамику среднесменного объема добычи под влиянием мощности пласта. В соответствии с изложенной выше методикой этот индекс можно разложить на частные индексы, отражающие влияние изменения величины мощности пласта:

$$I_{w\{v(m)\}} = \frac{\sum(a_0 + b_0 m_1)N_1}{\sum(a_0 + b_0 m_0)N_1} = \frac{\sum V' N_1}{\sum V_0 N_1},$$

Изменение силы воздействия этого признака на выработку и, соответственно, на общий объем добычи:

$$I_{w\{v(b)\}} = \frac{\sum(a_0 + b_1 m_1)N_1}{\sum(a_0 + b_0 m_1)N_1} = \frac{\sum V'' N_1}{\sum V' N_1}.$$

Для увязки этих частных индексов следует ввести корректирующий индекс, отражающий изменение свободного члена уравнения регрессии v по M:

$$I_{w\{v(a)\}} = \frac{\sum(a_1 + b_1 M_1)N_1}{\sum(a_0 + b_1 M_1)N_1} = \frac{\sum V_1 N_1}{\sum V'' N_1}.$$

Все величины, требуемые для расчетов этих индексов, представлены в табл. 10.14. С учетом этого

$$I_{w\{v(M)\}} = \frac{7607}{7912} = 0{,}9614\,(-3{,}86\%),$$

т. е. за счет наблюдаемого в отчетном периоде снижения мощности пласта среднесменная добыча угля сократилась в целом по тресту на 3,86% или на 305 т. Сокращение мощности пласта происходило, как уже было выявлено, наряду с усилением влияния этого фактора – коэффициент регрессии в отчетном периоде выше, чем в базисном ($b_1 = 0{,}140$, $b_0 = 0{,}115$). Повышение силы влияния мощности пласта на среднесменную выработку, а через нее на объем добычи характеризует следующий индекс:

$$I_{w\{v(b)\}} = \frac{10142}{7607} = 1{,}333,$$

т. е. за счет роста силы связи общий объем среднесменной добычи вырос на 33,3% или на 2535 т. Влияние изменения свободного члена уравнения регрессии – параметра a – оценивается следующим индексом:

$$I_{w\{v(a)\}} = \frac{9003}{10142} = 0{,}8876.$$

Этот результат никак не комментируется, как и сам параметр a, он не может быть содержательно интерпретирован.

Рассмотренный пример показывает, что подобный анализ основан на определенной условности. Так, оценку влияния изменения коэффициента регрессии мы проводим при базисном значении свободного члена уравнения, тогда как параметры уравнения регрессии связаны друг с другом. Все они получаются в результате решения одной и той же системы уравнений. То же можно сказать в отношении раздельной оценки изменения значения фактора и силы

его влияния. Тем не менее, соединение индексного и регрессионного методов обогащает анализ, позволяет ввести в него нефункционально связанные факторы.

Рассмотренная методика анализа позволяет измерить влияние факторов, непосредственно не входящих в жестко детерминированное выражение результативного признака, не только в целом по совокупности, но и по каждому единичному явлению.

Проведение анализа по отдельным единицам с использованием уравнения регрессии обычно основывается на разложении величины отклонения от общей средней $(y_i - \bar{y})$ на две составляющие $(\hat{y}_i - \bar{y})$ и $(y_i - \hat{y}_i)$. Если в уравнение регрессии входят все важные и существенные факторы, от которых зависит величина результативного признака, и коэффициент детерминации близок к единице, то остальные, не включенные в уравнение факторы, характеризуют индивидуальные, несущественные особенности, зачастую не имеющие количественного выражения. В этом случае разница $(y_i - \hat{y}_i)$ образуется за счет несовпадения интенсивности воздействия на y всех учтенных факторов в условиях данной i-й единицы и средней интенсивности их воздействия, выраженной в величинах коэффициентов регрессии, входящих в расчетное значение $\hat{y}_i$. Это дает право интерпретировать разницу $(y_i - \hat{y}_i)$ или отношение $y_i/\hat{y}_i$ как показатель того, как эффективность использования учтенных факторов у i-й единицы соотносится со средней эффективностью их использования. Разница $(\hat{y}_i - \bar{y})$ возникает за счет различия в значениях учтенных факторов для данной i-й единицы и в среднем по совокупности. Такое разложение дает возможность выявить резервы, имеющиеся у каждой отдельной единицы, в части эффективности использования факторов и в части их уровня.

При анализе взаимосвязей в сочетании с изучением динамики явлений нас интересует в первую очередь не соотношение индивидуального и среднего по совокупности значений результативного признака, а изменение его состояния в отчетном периоде по сравнению с базисным – $(y_1 - y_0)$. В случае использования регрессионного анализа эта разница может быть представлена следующим образом:

$$(y_1 - y_0) = [(y_1 - \hat{y}_1) - (y_0 - \hat{y}_0)] + (\hat{y}_1 - \hat{y}_0).$$

Первый член разложения характеризует изменение в величине y, вызванное как изменением влияния тех неучтенных в регрессии факторов, которые не коррелируют с учтенными, так и изменением соотношения индивидуальной и средней силы влияния на y учтенных в регрессии факторов. Второй член этого разложения характеризует изменение в величине y, вызванное изменением в значениях факторов, учтенных в регрессии, и изменением средней силы их воздействия на y.

Продолжая наш пример, проведем анализ изменения среднесменной добычи угля, приходящейся на одного подземного рабочего (v), по данным отдельных шахт. Все необходимые величины приведены в табл. 10.15.

Таблица 10.15

№ шахты	ΔV	$V_0 - \hat{V}_0$	$V_1 - \hat{V}_1$	$\Delta(V - \hat{V})$	$\hat{V}_1 - \hat{V}_0$	$\Delta\hat{V}$(м)	$\Delta\hat{V}(a,в)$
1	0	0,1	–0,4	–0,5	0,5	–0,3	0,8
2	2	1,5	1,9	0,4	1,6	0	1,6
3	3	–0,4	1,0	1,4	1,6	–0,2	1,8
4	2	–0,4	0,7	1,1	0,9	–0,4	1,5
5	1	–0,5	–0,3	0,9	0,7	–0,4	1,1
6	1	–1,5	–2,5	–1,0	2,0	0	2,0
7	1	0,1	0,1	0	1,0	0	1,0
8	1	1,8	0,9	–0,9	1,9	0,1	1,8
9	0	0,1	0,2	0,1	–0,1	–0,7	0,6
10	–1	–0,7	–1,6	–0,9	–0,1	–1,2	1,1

Учитывая сравнительно низкие значения отчетного и базисного коэффициентов детерминации ($r^2_0 = 0{,}8154$, $r^2_1 = 0{,}7974$), разница фактической и расчетной величин $(V_i - \hat{V}_i)$ выражает не только различия в эффективности использования учтенного фактора – мощности пласта – на данной конкретной шахте по сравнению со средней эффективностью по тресту, но и влияние неучтенных в уравнении регрессии факторов.

В среднем среднесменная добыча одного подземного рабочего увеличилась в отчетном периоде по сравнению с базисным на 1 т. Мощность пласта снизилась в среднем на 2,6 м. Если бы действовал только этот фактор, то средняя добыча снизилась бы на 0,3 т. Таким образом, весь прирост среднесменной добычи вызван действием прочих факторов.

Увеличение абсолютной величины – свободного члена уравнения регрессии параметра a – является следствием снижения тесноты прямолинейной связи между мощностью пласта и среднесменной добычей угля на одного подземного рабочего. Данные табл. 10.15 позволяют определить значимость изменения мощности пласта и

прочих факторов в общем изменении величины среднесменной добычи, приходящейся на одного подземного рабочего на каждой шахте. Так, нулевое приращение среднесменной выработки на первой шахте явилось результатом равнодействия отрицательного влияния снижения мощности пласта и других факторов и общей тенденции повышения «съема» угля с одного метра мощности пласта. На второй шахте прирост среднесменной добычи одного рабочего обусловлен, с одной стороны, более эффективным использованием мощности пласта, нежели в среднем по тресту, с другой стороны – положительным влиянием изменения прочих факторов, как коррелирующих с мощностью пласта, так и не связанных с ним. Подобные заключения можно сделать по третьей, четвертой и т. д. шахтам.

Таким образом, введение в жестко детерминированное уравнение связи величин, найденных на основе уравнения регрессии, позволяет учесть в комплексе как жестко детерминированные, так и стохастические связи.

В экономическом анализе часто решаются задачи, связанные с изучением средних величин, их уровня и динамики – какова средняя цена 1 кг ржаного хлеба, говядины, средняя заработная плата в промышленности, в экономике в целом и т. д. Изменение средней величины отражает индекс

$$I_{\bar{y}1/0} = \bar{y}_1 : \bar{y}_0.$$

По данным отчетного и базисного периодов можно построить регрессии – обязательно с одним и тем же набором объясняющих переменных:

$$\hat{y}_1 = a_1 + b_{11}x_1 + \ldots + b_{1k}x_k;$$

$$\hat{y}_0 = a_0 + b_{01}x_1 + \ldots + b_{0k}x_k.$$

Поскольку, как известно из гл. 8, $\bar{y} = a + b_1\bar{x}_1 + \ldots + b_k\bar{x}_k$, то отчетная и базисная регрессии могут быть использованы для анализа изменения среднего уровня, $\bar{y}$.

$$I_{\bar{y}} = \frac{\bar{y}_1}{\bar{y}_0} = \frac{a_1 + b_{11}\bar{x}_{11} + b_{12}\bar{x}_{12} + \ldots + b_{1k}\bar{x}_k}{a_0 + b_{01}\bar{x}_{01} + b_{02}\bar{x}_{02} + \ldots + b_{0k}\bar{x}_k}. \qquad (10.39)$$

Средняя величина, $\bar{y}$, может изменяться, во-первых, за счет изменений средних значений объясняющих переменных $\bar{x}_j$ в отчетном периоде по сравнению с базисным, во-вторых, за счет изменения значений коэффициентов регрессии b_j, в-третьих, за счет изменения значения свободного члена уравнения регрессии a_0.

Соответственно, получаем систему индексов:

$$I_{\bar{y}} = \frac{\bar{y}_1}{\bar{y}_0} = \prod_{(J)} I_{\bar{y}(\bar{x}j)} \cdot \prod_{(J)} I_{\bar{y}(bj)} \cdot I_{\bar{y}(a)}. \qquad (10.40)$$

Чтобы обеспечить это равенство, нужно принять какое-то правило индексации. Например, в соответствии с уже высказанным положением сначала индексируются все $\bar{x}_j$ при постоянных (базисных) значениях коэффициентов регрессии и свободного члена, затем индексируются коэффициенты регрессии при постоянных (отчетных) средних значениях $\bar{x}_j$, затем индексируется свободный член уравнения регрессии при постоянных (отчетных) значениях как x_j, так и b_j.

Применим рассмотренную методику к анализу среднесменной добычи угля одним рабочим. Среднесменная добыча одного подземного рабочего: базисная – 7,5 т/чел; отчетная – 8,5 т/чел. (табл. 10.12). Были построены базисная и отчетная регрессии, описывающие связь между среднесменной добычей (y) и мощностью пласта (x). Подставляя средние значения $\bar{x}$ и $\bar{y}$, получим:

$$-3{,}885 + 0{,}115 \cdot 99{,}0 = 7{,}5 \text{ (т/чел.)};$$

$$-5{,}0 + 0{,}140 \cdot 96{,}4 = 8{,}5 \text{ (т/чел.)}.$$

Измерим, как изменилась среднесменная добыча рабочего и как на нее повлияло изменение средней мощности пласта ($\bar{x}$), силы влияния этого фактора на добычу (b) и корректирующего параметра, т. е. свободного члена уравнения регрессии (a):

$$I_{\bar{y}} = 8{,}5 : 7{,}5 = 1{,}133 \quad (+13{,}3\%);$$

$$I_{\bar{y}(\bar{x})} = \frac{-3{,}885 + 0{,}115 \cdot 96{,}4}{-3{,}885 + 0{,}115 \cdot 99{,}0} = \frac{7{,}201}{7{,}500} = 0{,}96 \quad (-4\%);$$

$$I_{\bar{y}(b)} = \frac{-3{,}885 + 0{,}140 \cdot 96{,}4}{-3{,}885 + 0{,}115 \cdot 96{,}4} = \frac{9{,}611}{7{,}201} = 1{,}335 \quad (+33{,}5\%);$$

$$I_{\bar{y}(a)} = \frac{-5{,}0 + 0{,}140 \cdot 96{,}4}{-3{,}885 + 0{,}140 \cdot 96{,}4} = \frac{8{,}5}{9{,}611} = 0{,}884 \quad (-11{,}6\%).$$

Таким образом, рост среднесменной добычи угля был обусловлен усилением использования такого фактора как мощность пласта – ростом его воздействия на добычу на 33,5%. Сама мощность пласта несколько уменьшилась, что привело к снижению среднесменной добычи на 4%. Изменение свободного члена тоже оказало негативное влияние на среднесменную выработку (–11,6%).

Все полученные индексы образуют систему индексов: их произведение равно индексу среднесменной добычи.

10.9. ПРИМЕРЫ ИСПОЛЬЗОВАНИЯ ИНДЕКСОВ В ЭКОНОМИКО-СТАТИСТИЧЕСКИХ РАСЧЕТАХ

Практически в любом аналитическом обзоре, публикациях итогов развития экономики страны, региона за месяц, квартал, год, в перспективных расчетах обязательно используются индексы. Широкое использование индексов в экономико-статистической практике объясняется свойствами этих показателей: во-первых, взаимосвязью частных и общих индексов, что обеспечивает возможность последовательного агрегирования расчетов – по товарам и товарным группам, по территориям, по стране в целом и т. д.; во-вторых, взаимосвязями между индексами разных показателей –урожайности и валового сбора, производительности труда и фондовооруженности и т. д.

Зная изменение одного из взаимосвязанных показателей, всегда можно определить рассчетным путем изменение другого показателя. Например, по данным отчетности промышленных предприятий одного района известно, что численность занятых в промышленности сократилась в IV квартале по сравнению с I кварталом на 1,5%, объем промышленной продукции снизился на 3%, средняя зарплата возросла на 15%. Как изменились производительность труда и фонд зарплаты?

$$I_{\text{п.т}} = 0{,}97 : 0{,}985 = 0{,}985;$$

$$I_{\text{ф.з.}} = 1{,}15 \cdot 0{,}985 = 1{,}133,$$

т. е. производительность труда снизилась на 1,5%, хотя заработная плата росла, что привело к увеличению фонда заработной платы на 13,3%.

С помощью индексов измеряют динамику производительности труда. Производительность труда может измеряться либо количеством продукции, вырабатываемой в единицу времени q, либо

затратами рабочего времени на единицу продукции t. Причем эти показатели находятся в соотношении $q = 1/t$. Первый из них называют прямым показателем производительности труда, а второй – обратным. Сводный индекс производительности труда определяется как средний из индивидуальных индексов: либо $i_q = q_1/q_0$, либо $i_t = t_0/t_1$ (то, что базисное значение в числителе, объясняется обратным характером показателя трудоемкости). Отсюда

$$I_{\substack{\text{производит.}\\ \text{труда}}} = \frac{\sum i_q(q_1 t_1)}{\sum q_1 t_1} = \frac{\sum iT_1}{\sum T_1},$$

где i – индивидуальные индексы часовой, дневной или месячной производительности труда (по видам продукции);

T_1 – общие затраты времени в отчетном периоде соответственно в человеко-часах, человеко-днях или человеко-месяцах.

В последнем случае в качестве T_1 используется численность рабочих.

Важное значение для анализа и прогноза экономических процессов в стране, для международных сравнений имеет индекс *физического объема промышленной продукции.* Методика его построения основана на последовательном обобщении данных: индексы для более крупных совокупностей представляют собой средние из составных элементов этих совокупностей. Этим определяется порядок расчета индекса физического объема, который включает:

определение структурных показателей промышленности по отраслям, которые затем используются в качестве веса при агрегировании индивидуальных индексов в общепромышленный;

отбор товаров-представителей для каждой отрасли, по которым определяется динамика промышленной продукции в каждой отрасли;

агрегирование отраслевых индексов в общепромышленный.

В соответствии с международной практикой структура промышленного производства должна определяться по показателю добавленной стоимости, но ввиду новизны этого показателя для нашей страны для характеристики структуры промышленного производства используется показатель условно-чистой продукции (табл. 10.16).

«Стандартный» набор товаров-представителей включает профильные для каждой отрасли изделия, занимающие значительный удельный вес в общем объеме промышленного производства. По машиностроению и ряду других отраслей товары-представители учиты-

Таблица 10.16

Структура промышленного производства Российской Федерации в 1991 г. по показателю условно-чистой продукции

Отрасли	В процентах к итогу
Вся промышленность	100
в том числе:	
Электроэнергетика	6,0
Топливная промышленность	10,8
Черная металлургия	5,1
Цветная металлургия	5,8
Химическая и нефтехимическая промышленность	7,0
Машиностроение	28,5
Лесная, деревообрабатывающая и целлюлозно-бумажная промышленность	7,1
Промышленность строительных материалов	3,5
Стекольная и фарфоро-фаянсовая промышленность	0,6
Легкая промышленность	9,8
Пищевая промышленность	13,0
Микробиологическая промышленность	0,2
Мукомольно-крупяная и комбикормовая промышленность	0,9

вают выпуск этими отраслями товаров народного потребления. Набор товаров учитывает и качественную дифференциацию продукции, направления ее использования (уголь подразделяется на энергетический и коксующийся, прокат – на сортовой и листовой и т. д.). Всего для построения индекса физического объема промышленного производства используются данные примерно по 400 товарам-представителям в разрезе 120 отраслей и производств. В отраслевых индексах натуральный выпуск продукции по товарам-представителям обобщается по средним оптовым ценам базисного года:

$$I_{q_i} = \frac{\sum q_1 \cdot \overline{p}_0}{\sum q_0 \cdot \overline{p}_0}.$$

Сводный индекс промышленного производства равен:

$$I_q = \sum I_{q_1} \cdot w_i,$$

где w_i – отдельный вес i-й отрасли по показателю условно-чистой продукции.

Не менее важное значение для социально-экономического анализа и международных сравнений имеет *индекс потребительских цен* (ИПЦ). С 1992 г. на всей территории России осуществляется наблюдение за изменением цен и тарифов, которое ведет специально созданная Госкомстатом государственная служба.

Вторым источником информации служат данные бюджетной статистики. Свыше 48 тыс. домохозяйств в России ведут подробный учет своих доходов и расходов.

На основе этих двух информационных потоков производится расчет ИПЦ по фиксированному набору основных потребительских товаров и услуг по методологии, принятой в международной практике.

ИПЦ измеряет изменение стоимости фиксированной потребительской корзины товаров и услуг, используемых семьями. Корзина товаров и услуг фиксирована с тем, чтобы данному уровню жизни соответствовало одно и то же значение индекса. При таком подходе изменения ИПЦ могут вызываться только изменением цен, но не переменами в структуре потребления в результате изменения доходов или появления новых товаров. По этой причине ИПЦ называют *индексом стоимости жизни.* Он широко используется в качестве показателя инфляции.

Национальный ИПЦ рассчитывается на основе данных по 27 крупным городам России, представляющих все 11 экономических районов страны. Каждый из этих городов имеет население более 200 тыс. человек, в их число входят 13 городов-миллионеров. В сумме население отобранных городов составляет примерно одну треть городского населения Российской Федерации. Информация о ценах, собранная по этим городам, используется для расчета средних цен с использованием в качестве весов суммы расходов всех домохозяйств каждого города. На основе этих данных строятся и региональные ИПЦ – для 11 экономических районов и, если необходимо, конечно же, строятся ИПЦ для отдельных товаров и товарных групп.

Общегосударственный ИПЦ рассчитывается на основе отношений цен на 262 товара, зарегистрированных в 27 городах. Для каждого города отношения цен агрегируются в общегосударственные средние с использованием общих расходов в каждом городе в качестве весов (численность населения города умножается на душевое потребление, данные о котором берутся из бюджетного обследования).

Расчетная формула ИПЦ:

$$I_{n/0} = \frac{\sum\limits_{(j)} \left[p_{0j} Q_{0j} \frac{p_{nj}}{p_{0j}} \right]}{\sum\limits_{(j)} p_{0j} Q_{0j}} \cdot 100\%, \qquad (10.41)$$

где p_{0j} – цена товара j в базисном периоде;
p_{nj} – цена товара j в периоде n;
Q_{0j} – количество товара j в базисном периоде.

Очевидно, что эта формула тождественна формуле индекса цен Ласпейреса. Ее можно представить как

$$I_{n/0} = \sum_{(j)} w_{0j} i_{0j} \cdot 100, \qquad (10.42)$$

где

$$w_{0j} = \frac{p_{nj} Q_{0j}}{\sum p_{0j} Q_{0j}}$$

– доля расходов на товар j в общих расходах.

Однако практически трудно использовать и первое, и второе выражение ИПЦ, так как оба варианта включают отношение цены для периода к цене в базисный период (p_n/p_0) и предполагают сравнение изменений цен для каждого товара за длительные периоды с сохранением характеристик данных товаров. Эти условия трудно выполнить при изменении круга продаваемых товаров, замещении товаров, изменении структуры товарных потоков.

Поэтому применяется вариант ИПЦ с использованием отношения цены товара в периоде n к цене в предыдущем периоде $n-1$ (p_n/p_{n-1}):

$$I_{n/0} = \frac{\sum\limits_{(j)} (p_{nj} / p_{n-1,j} (p_{n-1} Q_{0j}))}{\sum\limits_{(j)} p_{0j} Q_{0j}} \cdot 100, \qquad (10.43)$$

где

$$p_{n-1j} Q_{0j} = p_{0j} Q_{0j} \frac{p_{1j}}{p_{0j}} \frac{p_{2j}}{p_{1j}} \cdots \frac{p_{n-1j}}{p_{n-2j}}.$$

Последняя формула ИПЦ тождественна двум предыдущим, но использование цепных сравнений цен облегчает введение новых товаров или их замещение, когда возникает такая необходимость.

ИПЦ строится путем последовательного агрегирования данных. Сначала определяются потоварные индексы цен, охватывающие все виды торговли, затем определяются индексы цен по товарным группам, затем строится сводный ИПЦ.

Например, в состав ИПЦ входит индекс потребительских цен на мясо и мясопродукты:

$$I_{p \text{ мясо и мясопродукты}} = \frac{\sum I_j w_0}{\sum w_0},$$

где w_0 – удельный вес расходов на покупку данных товаров в потребительских расходах населения (по данным бюджетных обследований).

Исходные данные для построения индекса потребительских цен для этой товарной группы представлены в табл. 10.17.

Отсюда сводный индекс цен по товарной группе

$$I_{p \text{ мясо и мясопродукты}} = \frac{1{,}15 \cdot 0{,}022 + 1{,}12 \cdot 0{,}0012 + \ldots + 1{,}06 \cdot 0{,}002}{0{,}2613} \approx 1{,}16.$$

В свою очередь, каждый из индексов цен по товарам этой группы обобщают динамику цен на данный товар по всем видам торговли. Например, индекс цен на говядину в городе N определялся по данным табл. 10.18.

Общенациональный ИПЦ строится как средний из территориальных индексов, взвешенных по численности населения:

$$I_{\text{ИПЦ}} = \frac{\Sigma I_{\text{ПЦ}}^{\text{численность населения}}}{\Sigma^{\text{численность населения}}}.$$

Трудно перечислить все индексы, используемые в социально-экономической статистике. Это и индексы урожайности, структуры посевных площадей, валового сбора, индексы себестоимости продукции, рентабельности и т. д. В условиях инфляции особенно большое значение приобретают индексы цен. Кроме индекса потребительских цен службы государственной статистики рассчитывают индексы оптовых цен (цен производства) и др. Индексы цен выполняют роль дефлятора, т. е. используются для пересчета показателей, выраженных в текущих ценах в базисные цены, т. е. в цены года,

Таблица 10.17

Данные по товарной группе «мясо и мясопродукты» в городе *N* за октябрь 1993 г.

Товары	Удельный вес в потребительских расходах населения	Индексы цен
Говядина	0,0220	1,15
Баранина	0,0012	1,12
Свинина	0,1090	1,23
Мясо птицы	0,1240	1,11
Мясные полуфабрикаты	0,0011	1,09
Пельмени	0,0020	1,15
Субпродукты	0,0020	1,06
Итого	0,2613	–

Таблица 10.18

Цены на говядину в городе *N* в октябре 1993 г.

Цена говядины, руб./кг	Государственная торговля	Потребкооперация	Частная торговля	Колхозный рынок
p_0	2000	2500	2200	3000
p_1	2500	2800	2800	3200
Удельный вес расходов на покупку	0,23	0,20	0,25	0,32

$$I_{\text{р говядина}} = \frac{\overline{p}_1}{\overline{p}_0} = \frac{2550 \cdot 0{,}23 + 2800 \cdot 0{,}2 + 2800 \cdot 0{,}25 + 3200 \cdot 0{,}32}{2000 \cdot 0{,}18 + 2500 \cdot 0{,}2 + 2200 \cdot 0{,}3 + 3000 \cdot 0{,}32} =$$

$$= \frac{2870{,}5}{2470} = 1{,}16$$

принятого в качестве базисного. С помощью дефляторов исчисляется динамика сводных статистико-экономических показателей – валового внутреннего продукта, валового национального продукта,

объема капитальных вложений и т. д. С помощью ИПЦ решаются вопросы индексации доходов населения. В практических расчетах строятся как изолированные индексы, так и системы взаимосвязанных индексов. На их основе проводится анализ изменения сложных явлений по факторам. Однако, проводя аналитические расчеты с помощью индексов, помните, что строгость их формул, взаимные увязки, количественные оценки (относительные и абсолютные) вкладов отдельных факторов в совокупное изменение нельзя воспринимать как абсолютную истину. Это всего лишь приближение к истине, которое получено при той или иной методике построения индексов (система выбора весов, базы сравнения, построения исходного уравнения связи между признаками). Не обольщайтесь кажущейся точностью, отнеситесь к результатам критически!

Большое значение в экономической практике имеют соотношения в изменениях показателей, т. е. соотношения между величинами их индексов. Например, известно, что в эффективной экономике темпы роста производительности труда должны опережать темпы роста заработной платы:

$$I_{\text{произв. труда}} > I_{\text{заработной платы}}.$$

Или же – для развития предприятий оптимально следующее соотношение динамики основных показателей:

$$I_{\text{балансовой прибыли}} > I_{\text{реализации}} > I_{\text{авансированного капитала}} > 100\%.$$

Такого рода соотношение принято называть «экономической нормалью» или «динамическим нормативом».

Сравнение с нормалью используется в аудиторской деятельности для заключения о финансовом положении предприятия, его потенциале. Например, для трудоемкого производства в качестве нормали формулируется следующее неравенство:

$$I_{\text{объема реализации}} > I_{\text{материальные затраты на производство}} > I_{\text{численность промышленно-производственного персонала}} > I_{\text{средняя стоимость основных производственных фондов}}.$$

Чтобы определить соответствие фактической динамики нормали, нужно иметь данные об изменении показателей (индексы) за несколько периодов. Например, оказалось, что поквартальные индексы за два года показывают следующее (табл. 10.19).

Таблица 10.19

1-й год					2-й год			
Кварталы	1	2	3	4	1	2	3	4
I объема реализации	×	0,994	0,985	1,041	1,055	0,990	1,002	1,036
I материальных затрат на производство	×	0,955	0,967	1,096	1,007	0,960	0,983	1,056
I численности промышленно-производственного персонала	×	0,972	0,995	1,016	1,000	0,989	1,009	0,998
I стоимости основных производственных фондов	×	1,001	0,999	1,001	1,001	1,000	1,002	1,006

Только в трех кварталах соотношение в изменении показателей было близко к нормали. Аудитор обязан указать на это в своем заключении и рекомендовать менеджерам обратить внимание на причины: высокие цены поставщиков, избыточная численность персонала, неэффективная структура и использование основных фондов и т. д.

Динамика, соответствующая экономической нормали, обычно определяет стратегию развития предприятий и для управления компанией (фирмой) важно проводить сравнение фактического соотношения темпов изменения показателей с «нормальным», выявлять, в каком звене нормали возникли нарушения и вносить коррективы в деятельность предприятия.

Рекомендуемая литература к главе 10

1. *Адамов В. Е.* Факторный индексный анализ. Методология и проблемы. – М.: Статистика, 1977.
2. *Аллен Р.* Экономические индексы / Пер. с англ. – М.: Статистика, 1980.
3. *Зоркальцев В. М.* Индексы цен и инфляционные процессы. – Новосибирск: Наука-Сибирская издательская фирма РАН, 1996.
4. *Казинец Л. С.* Теория индексов. – М.: Госстатиздат, 1963.
5. *Ковалевский Г. В.* Индексный метод в экономике. – М.: Финансы и статистика, 1989.
6. *Плошко Б. Г.* Индексы. – Л.: ЛГУ, 1958.
7. *Фишер И.* Построение индексов / Пер. с англ. – М.: Изд-во ЦСУ СССР, 1928.

11 Глава СТАТИСТИЧЕСКОЕ ИЗУЧЕНИЕ СТРУКТУРЫ СОВОКУПНОСТИ И ЕЕ ИЗМЕНЕНИЙ

Развитие статистической совокупности проявляется не только в количественном росте или уменьшении элементов этой системы, но также и в изменении ее структуры. **Структура** – это строение, форма организации системы, состоящей из отдельных элементов и связей между ними. Так, человеческий организм представляет собой систему, состоящую из различных тканей, органов, закономерным образом взаимодействующих друг с другом. Экономика страны состоит из отраслей, предприятий, учреждений, связанных материально, информационно, энергетически.

Решающим условием дальнейшего развития человеческого общества в современную эпоху стало не простое расширение, количественное возрастание его параметров (численности населения, объемов производства и потребления ресурсов), а структурное изменение –переход от ресурсопотребляющей к ресурсосберегающей стратегии. На смену экспоненциальному росту потребления невозобновляемых ресурсов приходит экологически ориентированное производство, ограничение роста населения, как условия повышения качества жизни. Соответственно возрастает роль методов и показателей статистики, характеризующих структуру социальных, производственных, технологических систем, ее изменений.

11.1. ПОКАЗАТЕЛИ ПРОСТОЙ (ОДНОМЕРНОЙ) СТРУКТУРЫ

В п. 3.2 были рассмотрены относительные показатели, характеризующие простую (одномерную) структуру: доля или удельный вес отдельных элементов в итоге абсолютного признака совокупности. В гл. 5 рассмотрена система показателей и методика анализа распределения совокупности по значениям какого-либо отдельного признака. В данном разделе рассмотрены особенности изучения

структуры по такому признаку, который способен принимать как положительные, так и отрицательные значения (например, финансовый результат деятельности фирмы, предприятия, сальдо миграции).

Примером такого рода являются данные табл. 11.1.

Таблица 11.1

Структура финансового результата деятельности сельскохозяйственных предприятий района

Виды деятельности	Прибыль (+), убыток (–) млн. руб.	В % к итогу			Модули сумм, млн. руб.	В % к сумме модулей
		Прибыль	Убыток	Сальдо финансового результата		
А	1	2	3	4	5	6
Растениеводство	+150	50	–	+75	150	37,5
Животноводство	–40	–	40	–20	40	10,0
Промышленная переработка	+120	40	–	+60	120	30,0
Услуги на сторону	+21	7	–	+10,5	21	5,25
Жилищно-коммунальное хозяйство	–60	–	60	–30	60	15,0
Прочая деятельность	+9	3	–	+4,5	9	2,25
Итоги: Прибыли	+300	100	–	–	400	100,0
Итоги: Убытка	–100	–	100	–		
Итоги: Сальдо	+200	–	–	100		

Половину полученной прибыли обеспечило растениеводство; значительна доля в прибыли продукции промышленной переработки сельхозпродукции. Показатели графы 3 (табл. 11.1) говорят о структуре убытков, указывая на неблагополучное финансовое состояние жилищно-коммунального хозяйства этих предприятий. По данным графы 4 рассчитаны показатели структуры знакопеременного признака – финансового результата. Эти показатели структуры также имеют разные знаки. Никакого запрета на отрицательную величину долей статистика не налагает; требуется, чтобы сумма долей была равна 100% и только. Экономический смысл показателей структуры финансового результата совершенно очевиден: растениеводство дало 50% прибыли, но не 50, а 75% финансового результата от всех видов деятельности. Животноводство снизило финансовый результат не на 40% данного им убытка, а на 20%. Знакопеременные доли в графе 5 реально отражают «вклады» каждой из отраслей деятельности в конечный общий финансовый результат. Другой способ получить аналогичные по экономическому значению показатели – построение показателей структуры по модулям финансовых результатов. В этом случае нужно, отбросив знаки, сложить прибыли и убытки, а затем вычислить доли каждой отрасли. Показатели графы 6 абстрагированы от знака, они характеризуют не направление, а только сравнительную величину влияния, «вклада» каждой отрасли в образование финансового результата. Ранжируя отрасли по этим долям, можем сделать вывод, что самое большое влияние оказало растениеводство, на втором месте – промышленная переработка, на третьем по силе влияния – жилищно-коммунальное хозяйство и т. д. Показатели графы 6 пропорциональны показателям графы 4, но последние, кроме того, характеризуют и направление «вклада» отраслей, а потому более информативны; именно им нужно отдать предпочтение.

11.2. ПОКАЗАТЕЛИ ИЕРАРХИЧЕСКОЙ «ДРЕВОВИДНОЙ» СТРУКТУРЫ

Древовидной иерархической структурой называется сложная структура, образуемая при последовательном дроблении системы на все более однородные группы элементов. Она состоит из нескольких уровней («шагов» дробления). Таковы, например, административно-управленческая структура предприятия – двух-трехзвенная или более сложная, структура народного хозяйства по крупным отраслям, подотраслям и группам однородных предприятий; структура товарооборота по группам товаров, их видам, сортам, размерам и т. д.

Рассмотрим пример иерархической структуры (схема 11.1.)[1].

[1] *Агапова Т. Н.* Методы статистического изучения структуры сложных систем и ее изменения. – М.: Финансы и статистика, 1996. – С. 59 – 62.

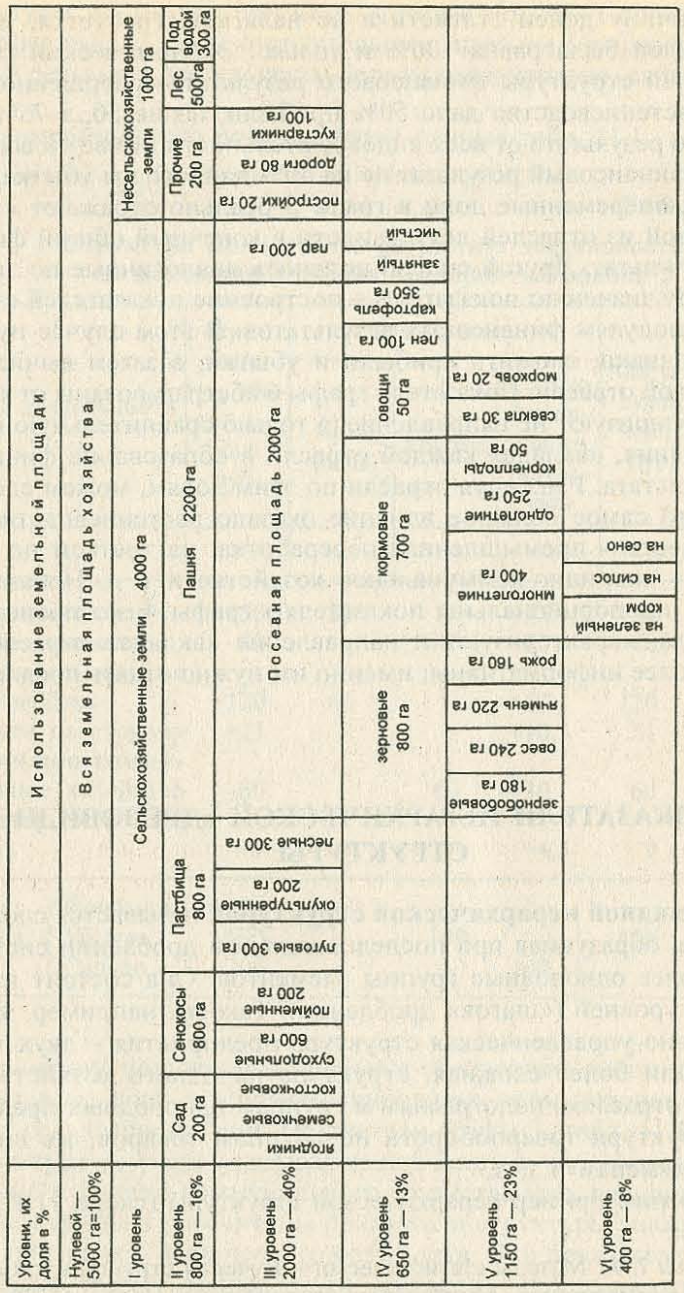


Схема 11.1. Иерархическая структура земельной площади хозяйства

В схеме 11.1 указаны шесть уровней и площадь каждой иерархии (дробления), доля этой площади в общей величине земельной площади хозяйства. Отметим, что вся иерархическая структура насчитывает шесть уровней, не считая нулевого, на котором еще нет дробления.

Иерархическая структура характеризуется не только долями объема признака, но и дополнительными показателями:

I. Характеристикой степени сложности структуры:

1. Числом уровней дробления («порядок» структуры). На схеме 11.1 приведена структура шестого порядка.

2. Средним порядком структуры, т. е. средним номером уровня, взвешенным по долям объема признака, дробление которых завершилось на данном уровне:

$$\overline{\Pi} = \sum_{i=1}^{k} \Pi_i \cdot d_i, \tag{11.1}$$

где $\overline{\Pi}$ – средний порядок;
Π_i – номер уровня (порядок дробления);
k – число уровней;
d_i – доля признака на i-ом уровне.

По данным схемы 11.1 $\overline{\Pi} = 6 \cdot 0{,}08 + 5 \cdot 0{,}23 + 4 \cdot 0{,}13 + 3 \cdot 0{,}4 + 2 \cdot 0{,}16 = 3{,}67$.

Эта величина характеризует среднее число дроблений объема признака.

3. Общим числом конечных (т. е. не дробящихся далее) ветвей структуры. В данном примере имеем 28 конечных ветвей.

4. Средним числом конечных ветвей, приходящимся на один уровень.

$$\overline{b} = \frac{\sum_{i=1}^{k} b_i}{k}. \tag{11.2}$$

В примере $\overline{b} = \frac{28}{6} = 4{,}67$.

Этот показатель характеризует «насыщенность» уровней, как бы «густоту» дерева иерархической структуры, а число уровней – «высоту» этого дерева.

При анализе иерархической структуры вычисляются цепные и базисные доли. **Цепная доля** – это отношение объема признака на

вышележащем уровне иерархии к объему признака на непосредственно нижележащем уровне, из которого вышла ветвь вышележащего уровня. Например:

доля ржи в посевной площади зерновых: $d_{5/4} = \frac{160}{800} = 0{,}2$;

доля зерновых в общей посевной площади: $d_{4/3} = \frac{800}{2000} = 0{,}4$;

доля посевной площади в площади пашни: $d_{3/2} = \frac{2000}{2200} = 0{,}909$;

доля пашни в площади сельхозугодий: $d_{2/1} = \frac{2200}{4000} = 0{,}55$;

доля сельхозугодий в общей земельной площади: $d_{1/0} = \frac{4000}{5000} = 0{,}8$.

Базисная доля равна произведению цепных долей и выражает отношение величины вышележащего уровня к величине уровня, лежащего ниже на два и более порядка или к нулевому исходному уровню, например:

Произведение двух первых цепных долей дает долю ржи в общей посевной площади всех культур:

$$d_{5/4} \cdot d_{4/3} = d_{5/3} = 0{,}2 \cdot 0{,}4 = 0{,}08.$$

Произведение трех цепных долей дает долю ржи в площади пашни:

$$d_{5/3} \cdot d_{4/3} \cdot d_{3/2} = d_{5/2} = 0{,}2 \cdot 0{,}4 \cdot 0{,}909 = 0{,}07273.$$

Произведение четырех цепных долей дает долю площади ржи в сельхозугодьях:

$$d_{5/4} \cdot d_{4/3} \cdot d_{3/2} \cdot d_{2/1} = 0{,}2 \cdot 0{,}4 \cdot 0{,}909 \cdot 0{,}55 = 0{,}04.$$

Наконец, произведение всех цепных долей дает долю площади ржи в общей земельной площади хозяйства:

$$d_{5/4} \cdot d_{4/3} \cdot d_{3/2} \cdot d_{2/1} \cdot d_{1/0} = 0{,}2 \cdot 0{,}4 \cdot 0{,}909 \cdot 0{,}55 \cdot 0{,}8 = 0{,}032.$$

Очевидна аналогия этих показателей с цепными и базисными темпами роста при анализе динамического ряда.

11.3. ПОКАЗАТЕЛИ БАЛАНСОВОЙ СТРУКТУРЫ

Баланс – в буквальном переводе – равновесие, – это особая форма сопоставления структуры одной и той же величины признака, характеризуемой с двух разных сторон или в двух различных аспектах. Например, наиболее известный читателям бухгалтерский баланс – это характеристика структуры средств предприятия, банка, фирмы с одной стороны (пассив) – по источникам этих средств, с другой стороны (актив) – по их вещественной форме. Бухгалтерский баланс на определенную дату – пример статического баланса. Динамические балансы отражают движение изучаемых натуральных, стоимостных или информационных объектов за некоторый период. В наиболее общей форме динамический баланс состоит из четырех составляющих: запас на начало периода, приход за период, расход за период, запас на конец периода. Запас на начало + приход = расход + запас на конец. Для аналитических целей каждая из четырех составляющих делится по различным классификационным признакам на части, группы или подгруппы.

Каждая из «сторон» или «половинок» динамического баланса состоит из двух разнокачественных уровней: запас – это моментный уровень, не зависящий от длительности интервала времени, отражаемого в балансе, а приход и расход, часто называемые потоками, это интервальные показатели, зависящие, как показано в гл. 9, от длительности интервала времени. В случае равномерного во времени процесса потоки пропорциональны величине интервала времени. Поэтому соотношение между запасами и потоками зависит от этого интервала, и, в пределе, при интервале, стремящемся к нулю, отношение запаса к потоку стремится к бесконечности, а при интервале, стремящемся к бесконечности, отношение запаса к потоку стремится к нулю. Данное свойство непременно должно учитываться при анализе балансовых структур.

Но при заданной величине интервала времени, например, равной одному году, отношение запаса к потоку является очень существенным структурным показателем, характеризующим изучаемый объект. Если «запас» значительно превышает величину «потока» за год, объект можно условно назвать «консервативным». Таковы, как правило, основные фонды предприятия. Их поступление за год и выбытие за год обычно не достигают и 50% «запаса», т. е. наличия на 1 января или среднегодового. Напротив, если «поток» за год существенно превышает «запас», объект можно условно назвать «мобильным». Таковы оборотные средства предприятий, товары в розничной торговле, денежные средства большей части населения.

Остановимся на показателях соотношения между «запасом» и «потоком». Примем такой вариант, когда показателем запаса считается его среднегодовой уровень (средняя из величин запаса на начало и конец года или точнее – хронологическая средняя из данных на начало каждого месяца или квартала, как показано в гл. 9), а величиной потока будем считать меньшую из величин входящего потока (поступление) и исходящего потока (выбытие). Это допущение позволяет отделить поток, проходящий через объект, от прироста или уменьшения запаса за год.

Рассмотрим условный **пример.** Пусть начальный запас данного материала составил 2000 единиц, приход за год 5000 единиц, выбытие – 4500 единиц, конечный запас 2500 единиц. Тогда среднегодовой запас составит 2250 единиц, поток 4500 единиц. Отношение потока к среднегодовому запасу равно двум. Эту величину обычно интерпретируют, как «число оборотов» данного материала за год, т. е. величина имеет единицу измерения «год в минус первой степени», что и вытекает из отношения:

$$\frac{\text{поток} = 4500 \text{ единиц/год}}{\text{запас} = 2250 \text{ единиц}} = 2 \text{ год}^{-1}, \text{ т. е. два оборота в год.}$$

Если предположить, что поток был в течение года равномерным, то за квартал он составит 1125 единиц, тогда средний запас за I квартал составит:

$$\frac{2000 + \left(2000 + \frac{500}{4}\right)}{2} = 2062{,}5 \text{ единиц,}$$

а число оборотов:

$$\frac{1125 \text{ ед./квартал}}{2062{,}5 \text{ единиц}} = 0{,}545 \text{ оборота в квартал}$$

или 0,545 квартал$^{-1}$.

За IV квартал имеем средний запас

$$\frac{\left(2000 + 3\frac{500}{4}\right) + 2500}{2} = 2437{,}5 \text{ единиц,}$$

а число оборотов: 1125 единиц/квартал : 2437,5 ед. = 0,462 квартал$^{-1}$. Как видим, при равномерном потоке и росте запаса, отношение потока к запасу постепенно уменьшается, скорость оборота замедляется. При равномерном потоке и сокращении запаса, наоборот, скорость оборота будет возрастать. Обратная величина – отношение запаса к потоку за год – составит:

$$\frac{2250 \text{ единиц}}{4500 \text{ единиц/год}} = 0{,}5 \text{ года или } 182{,}5 \text{ дня.}$$

При указанных выше единицах измерения прямого и обратного показателя их произведение равно единице. В нашем примере при двух оборотах за год, средства можно считать умеренно мобильными.

Конечно, изучение структуры динамического баланса не ограничивается приведенными общими показателями. Значительный интерес представляет изучение структуры входящего и исходящего потока, например, долей импорта и собственного производства в приходе товара, реализации и потерь в исходящем потоке и других отношений. Поскольку они выражаются обычными долями нет необходимости рассматривать методику их определения.

Одним из важнейших следствий деятельности человечества на планете Земля в настоящее время является возрастание содержания в атмосфере окиси углерода. В результате увеличивается «парниковый эффект» атмосферы, повышается средняя температура воздуха, что может привести к очень серьезным и неблагоприятным для человечества последствиям. Рассмотрим структуру динамического баланса содержания углекислого газа в атмосфере (табл. 11.2).

Отношение «потока» – величины выбытия к среднегодовому запасу составляет 0,521, что характеризует систему, как весьма мобильную. Доля антропогенных выбросов в поступлении CO^2 невелика, только 4,5%. Однако быстрый рост антропогенных выбросов в поступлении CO^2 в XIX в. и особенно в XX в. превысил его выбытие. При сохранении выбросов на нынешнем уровне запас содержания CO^2 в атмосфере возрастает примерно на 1% в год, что ведет к удвоению доли CO^2 среди всех компонентов атмосферы за столетие и резкому возрастанию «парникового эффекта», так как именно молекулы CO^2 (а также метана) задерживают уходящее с поверхности Земли низкочастотное тепловое излучение. Таким образом, казалось бы небольшое нарушение структуры баланса за достаточно длительное время может привести к очень крупным изменениям системы.

Перейдем к специфическим показателям, характеризующим структурные соотношения между различными сторонами бухгалтерского баланса. В качестве примера рассмотрим баланс (без подробного состава статей) мясокомбината АО «Самсон» на 1.01.1994 (табл. 11.3).

Таблица 11.2

Годовой баланс CO_2 в атмосфере Земли[1]

Наличие на начало и поступление		Выбытие и наличие на конец года	
Статьи	млрд т	Статьи	млрд т
Наличие на начало	1540	Выбытие	
Поступление:			
1. Дыхание растений и животных	220	1. Фотосинтез растений на суше	440
2. Выделяется из почв	220	2. Поглощение в океане	367
3. Выделяется из океана	341	в том числе осаждение в известняке	11
4. От сгорания лесов	5		
5. Выбросы промышленности, транспорта, бытовые	37	Итого выбытие	807
Итого поступление	823	Наличие на конец года	1556
Баланс	2363	Баланс	2363

[1] *Источник данных:* Добровольцев Г. В., Куст Г. В. Деградация почв – «Тихий кризис планеты» // Природа. – 1996. – № 10. – С. 53–63;

При анализе структуры баланса используется ряд показателей, из которых рассмотрим важнейшие.

1. Величина собственных оборотных средств и их доля в валюте баланса, определяемых как разность между текущими активами (ТА) и текущими обязательствами (ТО).

ТА – ТО = 13184 млн руб. – 10 434 млн руб. = 2750 млн руб, или 71,22% – 56,34% = 14,88%.

Такая величина говорит об удовлетворительном состоянии собственных оборотных средств.

2. Показатель маневренности капитала – доля собственных оборотных средств, находящихся в денежной форме:

ДС : (ТА – ТО) = 1238 : 2750 = 0,45 или 45%.

Иначе: 6,69% : 14,88% = 0,45.

Нормально действующее предприятие имеет этот показатель в интервале от 0 до 1.

3. Общий коэффициент покрытия обязательств – отношение текущих активов к текущим обязательствам, т. е.:

ТА : ТО = 13 184 : 10 434 или 71,22% : 56,34% = 1,26.

Таблица 11.3

Баланс АО «Самсон» на 1.01.1994 г.

Статьи баланса	Условные обозначения	Сумма, млн руб.	Доля в валюте баланса, %
Актив			
I. Текущие активы			
1. Денежные средства	ДС	1238	6,69
2. Расчеты с дебиторами	ДБ	3321	17,94
3. Запасы и затраты	ЗЗ	8595	46,43
4. Прочие активы	ПА	30	0,16
Итого раздел I	ТА	13184	71,22
II. Основные средства и прочие внеоборотные активы			
1. Основные средства и капитальные вложения	ОС	5075	27,41
2. Долгосрочные финансовые вложения	ФВ	231	1,25
3. Прочие внеоборотные активы	ПВ	22	0,12
Итого раздел II	ОС, ВА	5328	28,78
Баланс	Б	18512	100
Пассив			
I. Привлеченный капитал			
1. Текущие краткосрочные обязательства	ТО	10434	56,34
2. Долгосрочные обязательства	–	–	–
Итого раздел I	ПК	10434	56,34
II. Собственный капитал			
1. Фонды собственных средств	СС	8078	43,66
2. Прочие источники собственных средств	ПИ	0	0
3. Нераспределенная прибыль	–	–	–
Итого раздел II	СК	8078	43,66
Баланс	Б	18512	100

Нормальная деятельность предприятия требует, чтобы эта величина превышала единицу, а лучше, если она имеет значение от 1,5 до 2,0.

4. Коэффициент абсолютной ликвидности (платежеспособности). Он показывает, какую часть краткосрочных обязательств предприятие может погасить немедленно, т. е. за счет имеющихся денежных средств:

$К_{ал}$ = ДС : ТО = 1238 : 10 434 = 0,119.

Рекомендуется, чтобы этот коэффициент был в границах от 0,05 до 0,20, так что АО «Самсон» имеет хороший показатель.

5. Доля собственного капитала в валюте баланса (его доля в активах). Этот показатель равен отношению СК : Б = 43,66%.

В мировой учетно-аналитической практике считается, что минимальное значение этого показателя должно быть не ниже 60%. Если доля собственного капитала меньше, предприятие рассматривается как рисковое для инвесторов и кредиторов.

11.4. ПОКАЗАТЕЛИ МНОГОМЕРНОЙ СТРУКТУРЫ С ПЕРЕСЕКАЮЩИМИСЯ ПРИЗНАКАМИ

Если общий объем признака подразделен по одному группировочному признаку, а затем каждый групповой и общий объем снова подразделены по другому группировочному признаку, то образуется **многомерная,** в простейшем случае – двумерная – **структура с пересекающимися признаками.**

Рассмотрим пример[1].

Кроме показателей структуры мирового производства электроэнергии по регионам мира и по типам станций, данные табл. 11.4 позволяют определить и другие, более аналитические характеристики структуры.

1. Доли каждого типа станций в каждом регионе в региональном итоге производства (технологическая структура регионального производства). Например, приняв за 100% итог производства электроэнергии в России, можно определить, что 71% выработали ТЭС, 17,1 – ГЭС, 11,9% – АЭС. В США доля АЭС достигала 20,1%, в Европе (без России) – 23,6%.

2. Доли каждого региона в общем производстве электроэнергии данным типом электростанций. Например, 49,6% электроэнергии,

[1] *Агапова Т. Н.* Методы статистического изучения структуры сложных систем и ее изменения. – М.: Финансы и статистика, 1996. – С. 54 – 55.

Таблица 11.4

Производство электроэнергии по регионам мира и типам электростанций в 1993 г.

Регионы	Произведено млрд. квт · ч					Доли региона %
	ТЭС	ГЭС	АЭС	ГеоТЭС и др.	Итого	
Европа*	2509	697	991	2	4200	35,0
США	2183	249	617	21	3070	25,6
Россия	667	161	112	...	940	7,8
Прочие	2322	1089	356	23	3790	31,6
Весь мир	7680	2196	2078	48	12000	100
В % к итогу по типу станций	64,0	18,3	17,3	0,4	100	

* без России.

выработанной ГЭС в мире, произведено прочими регионами, 47,7% энергии, выработанной АЭС, произведено в Европе и т. п.

3. Доли производства электроэнергии, выработанной в данном регионе данным типом станций в общем мировом производстве. Например, электроэнергия, произведенная АЭС США составляет 617 : 12000 = 5,14% всемирного производства. Эта же доля может быть вычислена, как произведение доли АЭС в производстве электроэнергии США (20,1%) на долю США в общем мировом производстве (25,6%).

Итак, двумерная пересекающаяся структура позволяет рассчитать пять видов структурных показателей (долей). При трех пересекающихся признаках группировки число разных видов структур достигает 19. В общем виде при n взаимопересекающихся признаках структура содержит $(n^3 - n^2 + 1)$ видов долей.

Конечно, вовсе необязательно при каждом конкретном исследовании вычислять все эти показатели. Исходить следует из поставленной задачи, и вычислять те виды показателей структуры, которые для данной задачи имеют существенное значение. В отличие от анализа балансовой структуры, где две стороны баланса взаимосвязаны, при анализе структуры с пересекающимися независимыми признаками соотношения между долями, образованными по равным группировочным признакам, смысла не имели, как например – со-

отношение доли АЭС в производстве электроэнергии в США с долей России в мировом производстве ГЭС.

Если же группировочные признаки, образующие многомерную пересекающуюся структуру, связаны друг с другом, то анализ такой структуры объекта позволяет измерить тесноту и направление связи (см. п. 8.15).

11.5. СРАВНИТЕЛЬНЫЙ АНАЛИЗ СТРУКТУР

Сравнение структурных показателей по разным признакам может служить важным аналитическим приемом исследования. Рассмотрим данные табл. 11.5.

Таблица 11.5

Структура естественного движения населения по субъектам Российской Федерации в 1995 г.

Субъекты Российской Федерации	Родилось за год		Умерло за год		Отношение доли умерших к доле родившихся
	тыс. чел.	% к РФ	тыс. чел.	% к РФ	
Российская Федерация	1367,3	100	2197,4	100	1,00
Москва	69,3	5,07	146,7	6,68	1,32
Московская область	47,5	3,47	115,8	5,27	1,52
Краснодарский край	50,5	3,69	76,8	3,50	0,95
Санкт-Петербург	33,8	2,47	76,7	3,49	1,41
Башкортостан	45,7	3,34	51,6	2,35	0,70
Татарстан	39,3	2,87	48,5	2,21	0,77
Нижегородская область	30,0	2,19	65,3	2,87	1,31
Челябинская область	33,2	2,43	54,5	2,48	1,02

Сопоставление абсолютных величин родившихся и умерших не раскрывает различия в естественном движении населения по субъектам Российской Федерации: во всех субъектах число умерших больше числа родившихся. Различие раскрывает сравнение структурных показателей: в Москве, Московской области, Петербурге доля умерших намного превышает долю родившихся, в Башкортостане и Татарстане, наоборот, больше доля родившихся. Построив показатель соотношения долей (последняя графа табл. 11.5), видим, что худшее положение занимает Московская область, на втором месте «снизу» – Санкт-Петербург. Краснодарский край и Челябинская область находятся примерно на среднероссийском уровне, а лучшее положение из перечисленных регионов занимает Башкортостан, чья доля среди умерших на 30% или на целый пункт ниже доли родившихся. Подчеркнем еще раз, что полученные новые показатели соотношения структур не тождественны ни по величине, ни по содержанию коэффициентам рождаемости и смертности – ведь и в Башкортостане смертность превышала рождаемость. Соотношения долей содержат новую, дополняющую информацию – в этом их значение.

Аналогично можно сравнить доли регионов в сумме средств, перечисляемых ими в федеральный бюджет с долей получаемых из него дотаций и субсидий, долю страны в территории суши с долей добываемых в стране полезных ископаемых. Так, Россия, занимая 10% площади суши Земли, добывает 11,6% нефти, 28,1 природного газа, 13% каменного угля. Это говорит о том, что Россия является мировым донором энергоносителей (а также – алмазов, апатита, калийных солей и других ископаемых). Другой яркий пример сравнения структурных показателей: Москва, имеющая 6,8% населения России, по сумме активов банков и обороту финансовых средств занимает в России около 50%, что говорит о ненормально высокой степени концентрации финансового капитала в столице.

При изучении распределения населения страны по душевому доходу (табл. 11.6) структурный анализ и сравнение структур позволяют раскрыть характер этого распределения, имеющий ключевое значение для понимания социальной структуры общества и социальной политики государства.

На основе данных табл. 11.6 можно вычислять разные структурные показатели, сравнивать распределение населения с распределением суммы денежных доходов. Выделим, например, 10% населения с наименьшими доходами: две первых группы сверху и три недостающих процента населения из третьей группы, т. е. 3 : 7,5 часть ее. Соответственно, возьмем сумму долей доходов d_{xj} первой, второй группы и 3 : 7,5 от доли доходов третьей группы. Имеем: доля доходов беднейших 10% населения составила:

$$0{,}30 + 1{,}19 + \frac{3}{7{,}5} \cdot 2{,}45 = 2{,}47\% \text{ всех доходов.}$$

Доля доходов 10% наиболее обеспеченных, в число которых входят полностью пять последних групп и еще 1,8 : 2,3 часть шестой с конца группы. Соответственно, объединяем доли доходов пяти

Таблица 11.6

Анализ* распределения населения России по среднедушевому среднемесячному доходу в 1995 г.**

Группы по доходу: тыс. руб. в мес. на 1 чел.	Млн чел. f_i	Доля населения d_{xj}	Середина интервала x'_j, тыс. руб. на 1 чел.	Общая сумма доходов по группе $x'_j \cdot f_j$ млрд руб.	Доля доходов d_{xj}, %	Модуль разности долей $\lvert d_{yj} - d_{xj} \rvert$, %	Нарастающая доля населения d'_{yj}, %	Нарастающая доля доходов d'_{xj}	
								факт.	$d'_{xj} = (d'_{yj})^B$
до 100	2,9	2,0	80	232	0,30	1,70	2,0	0,003	0,000
100,1–150	7,5	5,0	125	938	1,19	3,81	7,0	0,015	0,000
150,1–200	11,0	7,5	175	1925	2,45	5,05	14,5	0,039	0,003
200,1–250	12,6	8,5	225	2835	3,61	4,89	23,0	0,075	0,012
250,1–300	12,9	8,7	275	3548	4,54	4,18	31,7	0,120	0,032
300,1–350	12,2	8,2	325	3965	5,05	3,15	39,9	0,170	0,064
350,1–400	11,2	7,5	375	4200	5,35	2,15	47,4	0,224	0,107
400,1–450	10,0	6,8	425	4250	5,41	1,39	54,2	0,278	0,159
450,1–500	8,8	5,9	475	4180	5,32	0,58	60,1	0,331	0,217
500,1–600	14,3	9,7	550	7865	10,02	0,32	69,8	0,431	0,340
600,1–700	10,7	7,2	650	6955	8,86	1,64	77,0	0,520	0,457
700,1–800	8,0	5,4	750	6000	7,64	2,24	82,4	0,596	0,560
800,1–900	6,0	4,1	850	5100	6,49	2,39	86,5	0,661	0,647
900,1–1000	4,5	3,0	950	4275	5,44	2,44	89,5	0,715	0,717
1000,1–1100	3,4	2,3	1050	3570	4,55	2,25	91,8	0,761	0,774
1100,1–1200	2,6	1,7	1150	2990	3,81	2,11	93,5	0,799	0,817
1200,1–1300	2,0	1,3	1250	2500	3,18	1,88	94,8	0,831	0,852
1300,1–1400	1,5	1,0	1350	2025	2,58	1,58	95,8	0,857	0,879
1400,1–1500	1,2	0,8	1450	1740	2,22	1,42	96,6	0,880	0,901
Более 1500	4,9	3,4	1926**	9437	12,02	8,62	100	1,000	1
Итого	148,2	100		78530	100	53,81	–	–	–

* *Источник*: Россия в цифрах.1996: Крат. стат. ежегодник. – М., Финансы и статистика, 1996. – С. 54.
** Средний денежный доход был равен 530 тыс. руб. на 1 человека в месяц: сумма месячных доходов 78 530 млр руб. Исходя из этой суммы определен средний месячный доход на 1 человека в последней группе.

последних групп и 1,8 : 2,3 часть доли доходов шестой с конца группы. Имеем: доля доходов у 10% наиболее обеспеченного населения составляет:

$$12{,}02 + 2{,}22 + 2{,}58 + 3{,}18 + 3{,}81 + \frac{1{,}8}{2{,}3} \cdot 4{,}55 = 27{,}37\% \text{ всех доходов.}$$

Отношение дохода (или доли в доходе, что то же самое) 10% наиболее обеспеченного населения к доходу (доли) 10% наименее обеспеченных жителей страны является одной из основных характеристик степени неравномерности распределения доходов, важнейшего социально-экономического индикатора. В России в 1995 г. это отношение составило: 27,37 : 2,47 = 11,08 раз.

Аналогично можно рассчитать отношение доли доходов у 5% богатейших жителей к доле дохода у 5% беднейших жителей, по данным табл. 11.6 оно составляет: 16,3 : 1,014 = 16,1 раза.

Чем более мелкие доли населения «сверху» и «снизу» сравниваются, тем больше величина этого показателя. Если же взять доли доходов у половины беднейшего и более обеспеченного населения, то соотношение по данным табл. 11.6 составляет: 75,46% : 24,54% = 3,07 раза.

К сожалению, группировка Госкомстата РФ не позволяет вычислить соотношение доли доходов у 1% самых богатых к доле доходов у 1% наиболее бедных жителей. Последняя, высшая по доходам группа включает более 3% населения, вчетверо больше, чем предыдущая группа, что нарушает рекомендуемые для группировки, особенно – типологической, правила о постепенном уменьшении численности групп к началу и к концу группировки. Ясно, что социологические исследования требуют особенно подробного изучения доходов как раз у богатейшей части населения, может быть даже с выделением доли процента населения.

Одним из обобщающих показателей степени неравномерности распределения может служить *коэффициент Лоренца:*

$$K_L = \frac{\sum_{j=1}^{k} \lvert d_{yj} - d_{xj} \rvert}{2}. \qquad (11.3)$$

В нашем примере (K_L) равен полусумме модулей разности долей населения и долей доходов и составляет:

$$K_L = \frac{\sum_{j=1}^{k} \left| d_{nj} - d_{xj} \right|}{2} = \frac{53{,}81\%}{2} = 26{,}9\%,$$

где d_{xj} – доля j-й группы в объеме признака x;
d_{yj} – доля j-й группы в численности совокупности.

Знаменатель коэффициента – это максимально возможная величина суммы модулей разности долей.

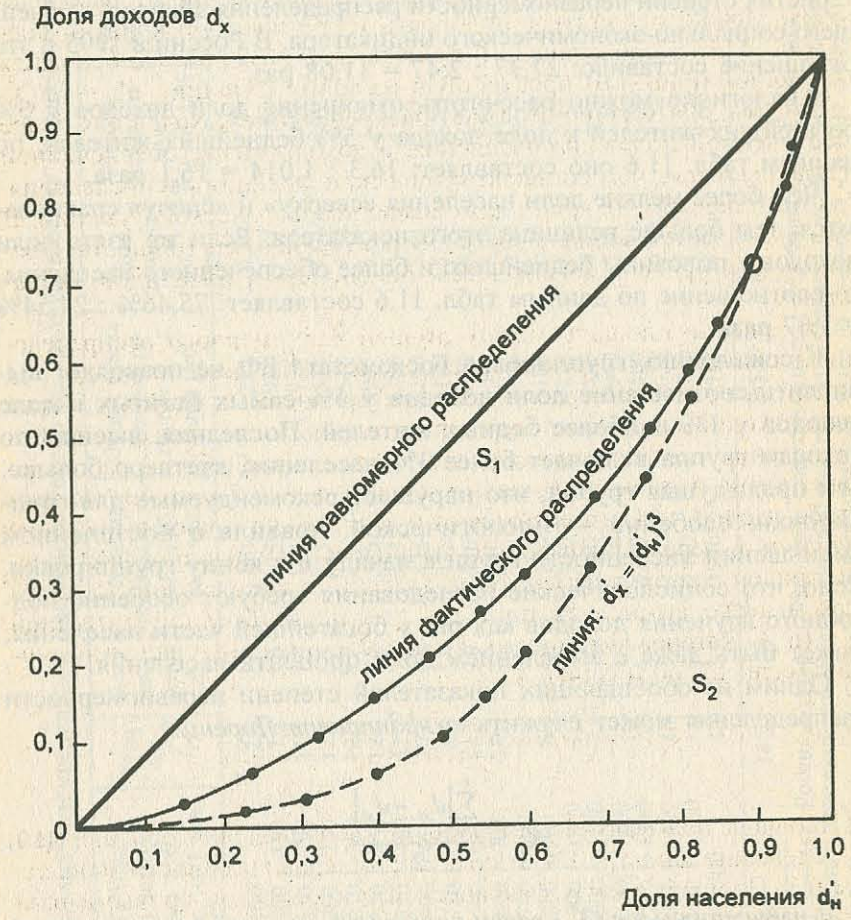


Рис. 11.2. Диаграмма Лоренца

По нарастающей доле населения и нарастающей доле доходов можно построить диаграмму Лоренца (рис. 11.1). В ней по оси абсцисс откладывается кумулятивная доля населения $d_н$, а по оси ординат – кумулятивная доля доходов d_x. Соединив точки ломаной линией, получим график, отражающий степень неравномерности распределения доходов. При строго равномерном распределении («абсолютной уравниловке») доли d_x и $d_н$ совпали бы, а ломаная линия обратилась в диагональ квадрата, которая и названа на графике «линией равномерного распределения». Чем дальше от диагонали в направлении к правому нижнему углу находится фактическая линия (ломаная), тем значительнее неравномерность распределения. Можно попытаться подобрать теоретическую кривую, достаточно близко проходящую к фактической ломаной. Не осложняя изложения, приводим простейшую из таких функций: $d'_x = (d'_н)^3$. Как видим, она хорошо отражает распределение доходов для более обеспеченной половины населения, а доходы менее обеспеченной половины выше, чем согласно кубической функции распределения. Для пяти беднейших групп даже выше квадратической функции. В этом сказывается социальная политика государства (общества), защищающая уровень жизни детей, пенсионеров, инвалидов, безработных.

Отношение площади между линией равномерного распределения и линией фактического распределения ко всей площади под линией равномерного распределения, принятой за единицу, также является одним из показателей неравномерности распределения. Без вычисления ясно, что площадь под линией равномерного распределения равна половине площади квадрата, т. е. 0,5 в принятом масштабе, где и $d_н$ и d_x изменяются от 0 до +1. Если точное выражение функции d_x через $d_н$ неизвестно вычислить площадь под фактической кривой нельзя, а измерить на диаграмме тоже сложно. Зато мы можем легко вычислить площадь под приближенной функцией $d'_x = (d'_н)^3$, которая выразится как определенный интеграл этой функции, если $d'_н$ обозначить X:

$$s_2 = \int_0^1 X^3 dX = \frac{X^4}{4}(1-10) = \frac{1^4}{4} - \frac{0^4}{4} = \frac{1}{4}.$$

Площадь под фактической кривой распределения больше, чем 1/4, следовательно $S_1 < 1/4$, и показатель неравномерности меньше, чем 0,5. Площадь между фактической и кубической кривыми раза в 3–4 меньше S_1, тогда коэффициент неравномерности составляет приблизительно от 0,5 – 0,5 : 3 до 0,5 – 0,5 : 4 или от 0,33 до 0,375.

11.6. ПОКАЗАТЕЛИ КОНЦЕНТРАЦИИ, СПЕЦИАЛИЗАЦИИ, МОНОПОЛИЗАЦИИ. МНОГОМЕРНАЯ СТРУКТУРА

Методы и показатели анализа структуры используются при изучении таких важных экономических процессов, как концентрация производства, специализация предприятий или отраслей, диверсификация капитала, степень монополизации рынка и др.

В гл. 5 рассмотрены показатели специализации предприятий региона, зоны, основанные на измерении вариации объемов производства или долей предприятий, а также отношения фактических мер вариации к предельно возможным при данной численности совокупности. В предыдущем параграфе данной главы рассмотрены показатели концентрации объема признака, основанные на неравномерности его распределения между единицами совокупности. Но эти характеристики не исчерпывают проблемы. Даже равномерное распределение производства, скажем, автомобилей в стране, где всего три предприятия, производящих по 33,3% всего выпуска автомашин, говорит о высокой степени концентрации в данной отрасли, и вероятности его монополизации, для устранения конкуренции и получения монопольной прибыли. Следовательно, показатель концентрации должен учитывать *две величины:* численность совокупности и степень неравномерности распределения признака между ее единицами. Рассмотрим методику конструирования показателя по заданным его свойствам.

Проще построить показатель, учитывающий численность совокупности и быстро убывающий, как убывает степень концентрации и вероятность монополизации, с ростом числа производителей n. Можно эту составляющую желаемого показателя представить например, как величину, обратную числу единиц совокупности, т. е. $1 : n$. При одном предприятии имеем абсолютный максимум, равный единице; при $n = 2$, $n = 3$, $n = 4$ доля все еще довольно значительна, но при большом n (большом числе производителей товара или услуг) эта составляющая уже становится несущественно малой и главное значение приобретает вторая составляющая – степень неравномерности распределения объема признака между единицами совокупности. Чтобы построить показатель, рассмотрим, как зависит от степени неравномерности распределения признака сумма накопленных долей объема признака $\sum_{i=1}^{n} d'_{xj}$ при условии, что изучаемая совокупность проранжирована в порядке нарастания долей объема признака.

При этом условии минимальная сумма накопленных долей будет в том случае, когда доли всех единиц совокупности кроме последней, равны нулю, а доля последней («монополиста») равна единице. Сумма накопленных долей тоже равна единице. Итак, $\sum_{i=1}^{n} d_{xj_{min}} = 1$.

Найдем теперь выражение для максимума этой суммы, которая, согласно условию ранжирования, образуется при строгом равенстве всех долей, каждая из которых будет равна $1 : n$.

Нарастающие доли будут $1 : n$, $2 : n$, $3 : n$, и т. д. до $n : n$, а их сумма, как сумма членов арифметической прогрессии, выражается как: $(1 : n)(1 + 2 + \ldots + n) = (1 : n)(n^2 + n) : 2 = (n + 1) : 2$. Чем дальше отстоит фактическая сумма накопленных долей от максимальной величины, тем сильнее неравномерность распределения. Следовательно, в числителе должны стоять величины: $(n+1):2-\sum_{j=1}^{n} d'_{xj}$. Чтобы измерить степень отклонения от равномерности распределения, нужно сравнить меру фактической неравномерности с максимально возможной, равной разности между максимально возможной суммой накопленных долей и минимальной их суммой, равной единице. Следовательно, знаменатель должен иметь вид: $[(n + 1) : 2] - 1 = (n + 1 - 2) : 2 = (n - 1) : 2$. Итак, показатель степени концентрации за счет неравномерности распределения имеет форму:

$$\frac{\frac{n+1}{2}-\sum_{j=1}^{n} d'_{xj}}{\frac{n-1}{2}} = \frac{n+2-2\sum_{j=1}^{n} d'_{xj}}{n-1}. \qquad (11.4)$$

Теперь, объединяем обе составляющие и получаем окончательный показатель степени концентрации объема признака в совокупности, состоящей из n единиц, проранжированных в порядке возрастания объема признака или доли его у данной единицы в общем объеме признака в совокупности. Обозначим его K:

$$K = \frac{1}{n} + \frac{n+1-2\sum_{j=1}^{n} d'_{xj}}{n-1}. \qquad (11.5)$$

Можно произвести преобразование этой формулы, но по нашему мнению, лучше сохранить выражения обеих составляющих частей,

чтобы их разная природа оставалась явной для пользователя. Остается выяснить свойства предлагаемого показателя концентрации и меры возможности монополизации рынка. При единственном монополисте: $n = 1$, первое слагаемое будет равно единице, второе – нулю. В итоге весь коэффициент равен единице. При $n = 2$ и равномерном распределении объема признака

$$K = \frac{1}{2} + \frac{2+1-2\cdot(0,5+1)}{2-1} = 0,5.$$

При сосредоточении всего объема признака у второго предприятия

$$K = \frac{1}{2} + \frac{2+1-2\cdot(0+1)}{2-1} = 1,5.$$

Эта величина показателя K максимальная из возможных. При росте n первое слагаемое уменьшается и при $n \to \infty$ стремится к нулю. Второе слагаемое при концентрации всего производства у одного предприятия всегда остается равной единице, значит, при абсолютной концентрации $K \to 1$ при $n \to \infty$. При полной равномерности, когда второе слагаемое равно нулю, имеем: $K \to 0$ при $n \to \infty$, как и должно быть логически. При реальных значениях распределений объема признака между единицами совокупности получаем промежуточные значения между $[(1 : n) + 1]$ и $1 : n$. Рассмотрим, например, распределение производителей грузовых автомашин в России в 1994 г. (табл. 11.7).

Показатель степени концентрации производства грузовых автомобилей в РФ в 1994 г. составил:

$$K = \frac{1}{5} + \frac{5+1-2\cdot 2,2176}{5-1} = 0,2 + 0,3912 = 0,5912.$$

Величина показателя указывает на значительную, хотя и не очень высокую, степень концентрации, большая часть которой следует из неравномерности распределения производства между фирмами. Опасность монополизации невелика. По-существу, она еще меньше, ибо грузовые автомашины разных фирм не вполне однородны, и нельзя рассматривать, как взаимозаменяющие товары. Из последнего замечания следует, что при экономической оценке величины концентрации и возможностей монополизации нельзя принимать в расчет только величину какого-то показателя, но надо проверить,

Таблица 11.7

Концентрация производства грузовых автомашин в РФ 1994 г.*

Производитель	Выпуск, штук	Доля к итогу d_{xj}	Накопленная доля, Σd_{xj}
Урал АЗ АО	14002	0,0885	0,0885
АО «Авто УАЗ»	19487	0,1232	0,2117
АО «КамАЗ»	25838	0,1633	0,3750
АМО «ЗИЛ»	26492	0,1674	0,5424
АО «ГАЗ»	72390	0,4576	1,0000
Итого	158209	1,000000	2,2176

* *Источник:* Теория статистики/Под. ред. Р. А. Шмойловой. – М.: Финансы и статистика, 1996. – С. 373.

насколько однородным является сам объемный признак, распределение которого изучается.

Даже хлеб, мясо и другие подобные товары на самом деле неоднородны по сортам, видам и т. п. Чем разнообразнее ассортимент товара, тем меньше, при прочих равных условиях, возможность монополизации производства. Качественная вариация, как и количественная вариация долей, облегчает конкуренцию.

Рассмотрим, измерение степени специализации привлекая методику *многомерной характеристики структуры*. Как известно, основным показателем специализации экономики считаются доли разных товаров, или их групп, или отраслей в общем объеме реализации продукции и услуг. Однако, нельзя игнорировать также структуру затрат труда, материалов, а в сельском хозяйстве – структуру использованных земельных ресурсов и т. д. Поэтому, более обоснованное суждение о специализации предприятия или региона можно получить с помощью многомерных показателей структуры, полученных тем или иным способом. В качестве примера рассмотрим показатели специализации в растениеводстве АПК Вологодской области (табл. 11.8).

Как видим, в стоимости продукции лен занимает первое место. При одномерной характеристике специализации только на основе стоимости продукции надо было бы считать область льноводческой и кормодобывающей. Но нельзя игнорировать, что зерновые занима-

Таблица 11.8

Показатели специализации растениеводства (АПК Вологодской обл., 1995 г.)*

Группы культур	Их доли в итогах по растениеводству, в %						
	в площади	в затратах труда	в материальных затратах	в стоимости продукции	простая средняя из всех долей	средняя взвешенная по экспертным баллам	средняя взвешенная по средним коэффициентам детерминации
Зерновые	35,8	48,7	51,9	23,7	40,02	37,04	41,56
Картофель	0,3	4,5	4,6	3,3	3,18	3,60	3,61
Овощи	0,1	1,5	1,3	1,7	1,150	1,41	1,28
Лен	1,5	9,4	3,2	37,5	12,90	19,76	13,26
Кормовые	62,3	35,9	39,0	33,8	42,75	38,19	40,29
Итого	100	100	100	100	100,00	100,00	100,00

**Источник:* *Агапова Т. Н.* Методы статистического изучения структуры сложных систем и ее улучшения. – М.: Финансы и статистика, 1996. – С. 48 – 51.

ют более 50% в материальных затратах и первое место (почти половину) в затратах труда, а также и то, что лен занимает лишь 1,5% площади посева.

Построим многомерные показатели структуры, тремя способами. *Первый* – с помощью простой арифметической средней величины долей. Очевидно, при этом полагаем все четыре признака структуры равноправными, что, конечно, является упрощением реальности. *Второй* способ состоит в вычислении взвешенных средних долей. При этом весами служат экспертные оценки сравнительной важности признаков специализации. Предположим, что наименее ценному признаку – доле в площади, присвоен балл 1, в затратах труда – балл 2, в материальных затратах –

балл 3, наиболее ценному признаку – стоимости продукции – балл 5. Тогда можно рассчитать взвешенные по баллам средние доли. Они приведены в предпоследней графе табл. 11.8. Теперь, растениеводство надо признать специализированным на кормодобывании и зерновом производстве, но существенную роль играет и льноводство.

Третий способ построения многомерных средних долей не требует привлечения каких-либо субъективных экспертных оценок – используется только информация, содержащаяся в исходных долях. Более информативным, а следовательно, весомым признается тот признак, который имеет более высокий коэффициент детерминации долей со всеми остающимися признаками. Вычислив попарные и средние коэффициенты детерминации, примем меньший из них за единицу (один балл) и получим баллы для других признаков, как отношения их средних коэффициентов детерминации к меньшему (см. табл. 11.9).

Таблица 11.9

Матрица коэффициентов детерминации долей

Признаки	Признаки			
	x_1	x_2	x_3	x_4
Посевная площадь, x_1	1	0,31	0,26	0,35
Затраты труда, x_2	0,31	1	0,98	0,36
Материальные затраты, x_3	0,26	0,98	1	0,87
Стоимость продукции, x_4	0,35	0,36	0,87	1
Средний коэффициент	0,307	0,55	0,703	0,527
Балл	1	1,79	2,29	1,72

Как видим, полученные баллы сильно расходятся с принятыми экспертными оценками по второму способу. Расходятся и многомерные доли, взвешенные по новым, информационным баллам, (на основе коэффициентов детерминации) приведенные в последней графе табл. 11.8. Резко снизился удельный вес льноводства, а на первое место вышло производство зерна.

11.7. АБСОЛЮТНЫЕ И ОТНОСИТЕЛЬНЫЕ ПОКАЗАТЕЛИ ИЗМЕНЕНИЯ СТРУКТУРЫ

Об особенностях измерения динамики относительных величин, в том числе и долей было сказано в п. 9.3. Здесь излагаются показатели, характеризующие не изменение отдельной доли, а изменение структуры в целом, т. е. «структурный сдвиг». Нередко под этим понятием понимают хорошо и давно известные индексы влияния изменения структуры на среднюю величину относительного показателя, например, показателей эффективности: производительности труда, себестоимости продукции, урожайности, рентабельности и т. п. Эти индексы измеряют не величину самого изменения структуры, а его влияние (они рассмотрены в гл. 10).

Обратимся к примеру (табл. 11.10).

Таблица 11.10

Изменение структуры ВВП России*

Направления использования ВВП	Доля в %		$\lvert d_{1j} - d_{0j} \rvert$	$\lvert d_{0j} - d_{1j} \rvert^2$	$\left\lvert \frac{d_{0j} - d_{1j}}{d_{0j}} \right\rvert$
	1992 d_{0j}	1995 d_{1j}			
Потребление домохозяйств	33,7	42,1	8,4	70,56	0,249
Потребление государственных учреждений и некоммерческих организаций	16,2	23,7	7,5	56,25	0,463
Валовое накопление	35,7	28,6	7,1	50,41	0,199
Сальдо экспорта-импорта	14,4	5,6	8,9	77,44	0,611
Итого	100	100	31,8	254,66	1,522

**Источник:* Россия в цифрах. 1996: Крат. стат. ежегодник. – М.: Финансы и статистика. – С. 171.

Эти данные свидетельствуют о существенном изменении долей ВВП, использованных на разные цели. Обобщающим абсолютным показателем изменения структуры может служить сумма модулей абсолютных изменений долей, выраженная в процентных пунктах:

$$Ad = \sum_{j=1}^{k} \left| d_{ij} - d_{0j} \right|. \qquad (11.6)$$

В 1995 г. по сравнению с 1992 г. это абсолютное изменение, обозначенное Ad, составило 31,8 процентных пункта.

Расчет среднего абсолютного изменения, приходящегося на одну долю (группу, единицу совокупности) не дает никакой добавочной информации, ибо отношение среднего изменения к величине средней доли тождественно суммарному изменению в отношении к сумме долей, равной единице. Зато очень важно определить, насколько сильно произошедшее изменение структуры в сравнении с предельно возможной величиной суммы модулей. Логически ясно, что максимальная сумма модулей изменения долей равна 2. Например, была одна доля в пределе равная 0, другая равная 1, а в следующем периоде наоборот. Сумма модулей разности долей равна 2. Теперь можно построить показатель степени интенсивности абсолютного структурного сдвига KAd:

$$KAd = \frac{\sum_{j=1}^{k} \left| d_{1j} - d_{0j} \right|}{2}; \qquad (11.7)$$

$$KAd = \frac{31{,}8\%}{2} = 15{,}9\%.$$

По данным табл. 11.10.

Изменение структуры использования ВВП страны на 16% всего за 3 года следует признать весьма быстрым. Чтобы избежать взаимопогашения разных по знаку изменений долей, вместо модулей можно применить квадраты и получить квадратическую меру абсолютного структурного сдвига, в форме среднего квадратического изменения долей:

$$\sigma d = \sqrt{\frac{\sum_{j=1}^{k} \left(d_{ij} - d_{0j} \right)^2}{k}}. \qquad (11.8)$$

По данным табл. 11.10.

$$\sigma_d = \sqrt{\frac{254{,}66}{4}} = 7{,}98 \text{ процентных пункта.}$$

В данном случае, все доли изменились почти на одинаковое число пунктов, поэтому средняя квадратическая величина почти равна арифметической средней: 31,8 : 4 = 7,95. При резко различных изменениях долей квадратическое изменение ближе к наибольшему из изменений, чем арифметическая средняя. Предельная величина суммы квадратов изменения долей также равна 2, как и сумма модулей изменений долей, так как $1^2 = 1$ $(-1)^2 = 1$. Для четырех долей максимальное значение $\sigma_d = \sqrt{2} : 4 = 0{,}71$. Фактическое значение составило 0,0798 : 0,71 = 0,112 или 11,2% максимального.

Абсолютные показатели изменения долей не учитывают величины долей базисного периода, т. е. считается, что изменение доли на 10 процентных пунктов равнозначно, была ли доля до этого равна 2% или 50%. Такой подход недостаточен. Ведь первая из долей при увеличении на 10 процентных пунктов возросла в 6 раз, а вторая только на одну пятую часть. Очевидно, изменение структуры следует охарактеризовать и относительным показателем, измеряющим среднее относительное изменение долей. Рассмотрим построение этого показателя. Средний темп изменения долей, взвешенный по величине базисных долей, тождественно равен 1:

$$\frac{\sum_{j=1}^{k} \frac{d_{1j}}{d_{0j}} \cdot d_{0j}}{\sum_{j=1}^{k} d_{0j}} = \frac{\sum_{j=1}^{k} d_{1j}}{\sum_{j=1}^{k} d_{0j}} = \frac{1}{1}.$$

Невзвешенный средний темп изменения при разных долях не обязательно равен 1, но из-за взаимопогашения темпов, больших 1 и темпов, меньших 1, близок к 1 и ничего не говорит о мере изменения структуры. Наиболее информативным оказывается среднее относительное линейное изменение (темп прироста) по модулю:

$$I_d = \frac{\sum_{j=1}^{k} \left| \frac{d_{ij} - d_{0j}}{d_{0j}} \right|}{k}. \qquad (11.9)$$

По данным табл. 11.10 эта величина составляет:

$$I_d = \frac{1{,}522}{4} = 0{,}381 \text{ или } 38{,}1\% \text{ (а не пункта).}$$

Этот показатель означает, что при изменении структуры использования ВВП России произошел в среднем 38-й процентный сдвиг

– изменение роли статей в итоге. Величина I_d предела не имеет, так как малая доля может возрасти в бесконечно большое число раз. Использовать необходимо лишь простую среднюю из относительных изменений долей, так как средняя величина, взвешенная по базисным долям, как легко можно убедиться, всегда равна ранее рассмотренному абсолютному изменению A_d.

К. Гатевым, С. В. Курышевой, Т. Н. Агаповой предложен еще ряд показателей относительного изменения структуры, о которых желающие расширить свои знания могут прочитать в указанной в конце главы литературе.

11.8. РАНГОВЫЕ И ИННОВАЦИОННЫЕ ПОКАЗАТЕЛИ ИЗМЕНЕНИЯ СТРУКТУРЫ

Изменения структуры не сводятся к возрастанию и уменьшению долей элементов этой структуры. В ряде практических задач особую роль играют ранги долей. Представим себе, что в каком-то комитете, на конференции, в Государственной Думе РФ, и т. д. обсуждался законопроект, и, по мере внесения в него поправок, проводилось три голосования, результаты которых представлены в табл. 11.11.

Таблица 11.11

Результаты голосования по законопроекту

Вид вотума	Результаты голосования, %			Ранги вотумов		
	I	II	III	I	II	III
За принятие	29	46	52	2	2	1
Против	54	51	46	1	1	2
Воздержались	11	3	2	3	3	3
Итого	100	100	100	–	–	–

При втором голосовании в сравнении с первым произошло существенное изменение структуры вотумов: абсолютное изменение (по модулю) $A_{d2/1} = 17 + 3 + 14 = 34$ процентных пункта, среднее изменение по 11,33 пункта на элемент. Абсолютный сдвиг при третьем голосовании в сравнении со вторым намного скромнее: $A_{d3/2} =$

= 6 + 5 + 1 = 12 или по 4 пункта на элемент структуры. Однако, качественное различие структур второго и третьего голосований принципиально, а различие второго и первого голосований не принципиально. И в первом и во втором голосовании законопроект не принят, а в третьем он одобрен. Это качественное различие проявилось в изменении рангов вотумов. Аналогичную ситуацию имеем в ряде других явлений. Так в результате экзаменационной сессии ранг («место», занятое группой) может быть гораздо важнее (скажем – группа, занявшая I и II места, награждаются путевкой, ценным призом) чем величина различия в долях отличников, «хорошистов», троечников и двоечников. Изменение рангов статей платежного баланса страны, рангов статей в структуре ВПП может иметь гораздо большее экономическое значение, чем даже значительный абсолютный структурный сдвиг без изменения рангов.

На основе изменения рангов долей можно построить **два показателя:**

1. *Линейный коэффициент изменения рангов долей.* Обозначим его KR. Он представляет собой отношение фактической суммы модулей изменения рангов к предельно возможной сумме модулей при n элементах структуры, равной $(n^2 : 2)$ для четного и $(n^2 - 1) : 2$ для нечетного n:

$$KR = \frac{\sum_{i=1}^{n} |R_{1i} - R_{0i}|}{n^2 : 2} \quad \text{или} \quad KR = \frac{\sum_{i=1}^{n} |R_{1i} - R_{0i}|}{(n^2 - 1) : 2}. \qquad (11.10)$$

По данным табл. 11.11 этот коэффициент составил

$$KR_{3/2} = \frac{1+1}{(3^2 - 1) : 2} = 0{,}5 \text{ или } 50\%.$$

Изменение рангов на 50% максимального, конечно, является существенным преобразованием структуры. Если подсчитать по ней ранги долей по данным табл. 11.10 получим: $KR_{1/0} = \frac{2}{4^2 : 2} = 0{,}25$, или 25% максимального, что также следует признать значительным изменением. О социально-экономическом значении этого изменения («хорошо» или «плохо») можно спорить, ибо сокращение доли накопления, да еще при абсолютном снижении всего объема ВВП, подрывает перспективы роста экономики в будущие годы.

2. *Квадратический коэффициент изменения рангов долей KRK.* Для его построения используем известный коэффициент корреляции рангов Спирмена (см. гл. 8).

При полном совпадении рангов долей в базисном и текущем периодах коэффициент Спирмена равен +1. При максимальном изменении рангов (первый становится последним, порядок рангов «переворачивается») коэффициент Спирмена составит –1, следовательно максимальное значение изменения коэффициента Спирмена равно 2. Чтобы построить показатель степени интенсивности изменения рангов элементов структуры, следует отклонение фактического коэффициента Спирмена от единицы разделить на 2. Получим формулу KRK:

$$KRK = \frac{1 - \left(1 - \frac{6\sum_{i=1}^{n} (R_{1i} - R_{0i})^2}{n^3 - n}\right)}{2} = \frac{3\sum_{i=1}^{n} (R_{1i} - R_{0i})^2}{n^3 - n}, \qquad (11.11)$$

где R_{1i} и R_{0i} – ранги долей элементов структуры в базисном и отчетном периодах.

Измерим с помощью этого показателя структурный сдвиг в распределении банков Санкт-Петербурга по сумме активов, рассматривая только банки, действовавшие и в 1994, и в 1995 гг. (табл. 11.12). Что касается измерения сдвига с обновлением состава элементов структуры, эта проблема рассмотрена ниже.

$KRK = \frac{374}{27^3 - 27} = 0{,}0113$ или 1,13%, что говорит об устойчивости иерархии петербургских банков, изменение их рангов за год было несущественным.

Рассмотрим, в заключение инновационные показатели изменения структуры, т. е. характеристики степени обновления ее качественного состава и элементов. Воспользуемся в качестве примера таблицей из уже упоминавшейся монографии Т. Н. Агаповой (табл. 11.13).

Линейный коэффициент интенсивности абсолютного структурного сдвига $KAd = \frac{0{,}90}{2} = 0{,}45$ или 45% максимального.

Таблица 11.12

Изменение рангов банков Санкт-Петербурга по сумме активов

Название банка	Ранги		$R_1 - R_0$	$(R_1 - R_0)^2$
	1994	1995		
Банк «Санкт-Петербург»	1	1	0	0
Промстройбанк	2	2	0	0
СПб Сбербанк	3	3	0	0
Петровский	4	4	0	0
Петроагропромбанк	5	5	1	1
Балтийский	6	6	–1	1
Леноблсбербанк	7	7	0	0
БНП Дрездер Банк, Россия	8	9	1	1
Лионский кредит, Россия	9	8	–1	1
Сибирский Торговый банк, филиал	10	11	1	1
Кредит Петербург	11	10	–1	1
Русский Торгово-Промышленный	12	18	6	36
Витабанк	13	13	0	0
Абт-Банк	14	12	–2	4
Токбанк, филиал	15	14	–1	1
Царскосельский банк	16	15	–1	1
Кредобанк	17	17	0	0
Энергомашбанк	18	16	–2	4
Петербургский лесопромышленный	19	19	0	0
Россия	20	21	1	1
Экспортно-импортный банк	21	20	–1	1
Викинг	22	22	0	0
Таврический	23	23	0	0
Порт Банк	24	27	3	9
Ипотена Банк	25	26	1	1
Технохимбанк	26	25	–1	1
Форбанк	27	24	–3	9
Итого	–	–	0	74

Таблица 11.13

Изменение и обновление структуры посевной площади сельскохозяйственного предприятия

Элементы структуры: виды сельхозструктур	Доли в итоге		Ранги		$d_{1j} - d_{0j}$	$(R_{1j} - R_{0j})^2$
	базисн. d_{0j}	текущ. d_{1j}	R_{0j}	R_{1j}		
Пшеница озимая	0,25	0,15	1	4	0,10	9
Пшеница яровая	0	0,30	8	1	0,30	49
Рожь	0,20	0	3	7	0,20	16
Овес	0,10	0,18	5	3	0,08	4
Картофель	0,05	0,04	6	6	0,01	0
Многолетние травы на сено	0,22	0,27	2	2	0,05	0
Лен	0,04	0,06	7	5	0,02	4
Однолетние травы на сено	0,14	0	4	8	0,14	16
Итого	1	1	38	38	0,90	98

Для построения рангового коэффициента логично будет условиться приписать нулевым значениям элементов последние по порядку ранги, если таких элементов несколько – в порядке их рангов в другом периоде. Тогда получим: $KR = \frac{22}{8^2 : 2} = \frac{22}{32} = 0{,}688$ или 68,8% максимального.

Квадратический коэффициент интенсивности изменения рангов:

$$KRK = \frac{3 \cdot 98}{8^3 - 8} = \frac{294}{504} = 0{,}583$$ или 58,3% максимального значения.

Все три показателя указывают на сильный количественный сдвиг в структуре. Но в отличие от ранее рассмотренных примеров, в данном примере нельзя этим ограничиться. Произошло качественное обновление структуры, состава сельскохозяйственных культур, и это качественное изменение отразится следующими показателями:

1. Показатель обновления по числу элементов структуры – отношение числа выбывших и числа новых элементов структуры к

общему числу имевшихся разных элементов за оба периода, его можно назвать «коэффициентом обновления состава»:

$$КОС = \frac{Ч_в + Ч_н}{П_0 + Ч_н} \text{ или } \frac{Ч_в + Ч_н}{П_1 + Ч_в}, \qquad (11.12)$$

где $Ч_в$, $Ч_н$ – число выбывших и число новых элементов:
$П_0$ и $П_1$ – число элементов базисной и текущей структуры.

$$КОС = \frac{2+1}{7+1} = 0{,}375 \text{ или } 37{,}5\% \text{ предельной величины.}$$

2. Принимая во внимание не только число обновившихся элементов структуры, но и их доли, т. е. значение в системе, получим отношение суммы обновившихся долей к максимальной сумме, как уже известно, равной двум целым. Этот показатель назовем «коэффициентом обновления долей»

$$КОД = \left(\sum_{j=1}^{k_1} d_в + \sum_{j=1}^{k_2} d_н \right) : 2,$$

где $d_в$, $d_н$ – выбывшие и новые доли;
$к_1$ и $к_2$ – их число.

В данном примере имеем:

КОД (0,14 + 0,20 + 0,30) : 2 = 0,32 или 32% максимального показателя.

При полном обновлении всех элементов структуры оба коэффициента обновления равны единице или 100%, так как числа выбывших и новых элементов равны в сумме числам прежних и новых элементов, а суммы выбывших долей и новых долей дают в числителе показателя КОД 2, и 2 в знаменателе. При отсутствии качественного обновления элементов структуры оба коэффициента, естественно, равны нулю, хотя количественный сдвиг может быть очень велик. Например, если при 20 элементах структуры 10 элементов имели по 0,01 и 10 элементов по 0,09, а в следующем периоде размеры их полностью поменяются, то абсолютный показатель интенсивности структурного сдвига достигнет (10 · 0,08 + + 10 · 0,08) : 2 = 0,8 или 80% максимального. Напротив, при сильном качественном обновлении, например, 18 элементов структуры из 20, если сумма долей этих обновившихся элементов составляет всего 0,18, а 2 доли, составляющие в сумме 0,82, остались неизменными, то количественные меры структурного сдвига окажутся низкими, хотя коэффициент обновления достигает по числу элементов: КОС = 18 : 20 = 0,9 или 90% максимального.

Приведенные примеры показывают, что при анализе изменения структуры следует применить не какой-то один показатель, а всю их систему, так как каждый показатель отражает, измеряет особый аспект структурного сдвига. Разные показатели изменения структуры связаны между собой не жесткой связью, а связью статистической, в среднем – прямой зависимостью, но в конкретных процессах изменения структуры разные показатели могут сильно расходиться и даже изменяться в разных направлениях.

Изменение структуры сложных систем включает не только изменение состава и долей материальных элементов структуры, но также изменение структуры связей между этими элементами. Об изучении структуры связей, в частности, коэффициента детерминации при многофакторной регрессии см. гл. 8.

Рекомендуемая литература к главе 11

1. *Агапова Т. Н.* Методы статистического изучения структуры сложных систем и ее изменения. – М.: Финансы и статистика, 1996.
2. *Казинец Л. С.* Измерение структурных сдвигов в экономике. – М.: Экономика, 1969.
3. *Казинец Л. С.* Темпы роста и структурные сдвиги в экономике. – М.: Экономика, 1981.
4. *Гатев К.* Статистическая оценка различий между структурами / Теоретические и методологические проблемы статистики / М., Статистика, 1979.
5. *Елисеева И. И., Рукавишников В. Н.* Группировка, корреляция, распознавание образов. – М.: Статистика, 1977.
6. *Миркин Б. Г.* Анализ качественных признаков и структур. – М.: Статистика, 1980.

ПРИЛОЖЕНИЕ.

Статистико-математические таблицы

1. Значение интеграла вероятностей $F(t) = \frac{1}{\sqrt{2\pi}} \int_{-t}^{+t} e^{-\frac{t^2}{2}} dt$

t	Сотые доли									
	0	1	2	3	4	5	6	7	8	9
0,0	0000	0080	0160	0239	0319	0399	0478	0558	0638	0718
0,1	0797	0876	0955	1034	1114	1192	1271	1350	1428	1507
0,2	1585	1663	1741	1819	1897	1974	2051	2128	2205	2282
0,3	2358	2434	2510	2586	2661	2737	2812	2886	2961	3035
0,4	3108	3182	3255	3328	3401	3473	3545	3616	3688	3752
0,5	3829	3899	3969	4039	4108	4177	4245	4313	4381	4448
0,6	4515	4581	4647	4713	4778	4843	4909	4971	5035	5098
0,7	5161	5223	5285	5346	5467	5497	5527	5587	5646	5705
0,8	5763	5821	5878	5935	5991	6047	6102	6157	6211	6265
0,9	6319	6372	6424	6476	6528	6579	6626	6679	6729	6778
1,0	6817	6875	6923	6970	7017	7063	7109	7154	7199	7243
1,1	7287	7330	7373	7415	7457	7499	7540	7580	7620	7660
1,2	7699	7737	7775	7813	7850	7887	7923	7959	7995	8030
1,3	8064	8098	8132	8165	8198	8230	8262	8293	8324	8355
1,4	8385	8415	8444	8473	8501	8529	8557	8584	8611	8638
1,5	8664	8690	8715	8740	8764	8788	8812	8836	8859	8882
1,6	8904	8926	8948	8969	8990	9011	9031	9051	9070	9089
1,7	9108	9127	9146	9164	9182	9199	9216	9233	9249	9265
1,8	9281	9297	9312	9327	9342	9357	9371	9385	9399	9412
1,9	9425	9438	9451	9464	9476	9488	9500	9512	9523	9534
2,0	9545	9556	9566	9576	9586	9596	9608	9615	9625	9634
2,1	9643	9652	9660	9669	9676	9684	9692	9700	9707	9715
2,2	9722	9729	9736	9743	9749	9755	9762	9768	9774	9780
2,3	9785	9791	9797	9802	9807	9812	9817	9822	9827	9832
2,4	9836	9840	9845	9849	9853	9857	9861	9866	9869	9872
2,5	9876	9879	9883	9886	9889	9892	9895	9898	9901	9904
2,6	9907	9909	9912	9915	9917	9920	9924	9926	9927	9929
2,7	9931	9933	9935	9937	9939	9940	9942	9944	9946	9947
2,8	9949	9950	9952	9953	9955	9956	9958	9959	9960	9961
2,9	9963	9964	9965	9966	9967	9968	9969	9970	9971	9972
3,0	99730	99739	99747	99755	99763	99771	99779	99786	99793	99800
3,1	99807	99813	99819	99825	99831	99837	99842	99847	99853	99858
3,2	99863	99867	99872	99876	99880	99884	99888	99892	99896	99900
3,3	99903	3,6	99911	3,9	999904	4,4	9999892	5,0		99999943
3,4	99933	3,7	99937	4,0	999937	4,6	9999957	5,0		99999996
3,5	99953	3,8	99957	4,2	999973	4,8	9999984	6,0		999999998

2. Значение *t*-критерия Стьюдента при уровне значимости 0,10, 0,05, 0,01

Число степеней свободы *d. f.*	*P*			*d. f.*	*P*		
	0,10	0,05	0,01		0,10	0,05	0,01
1	6,3138	12,706	63,657	18	1,7341	2,1009	2,8784
2	2,9200	4,3027	9,9248	19	1,7291	2,0930	2,8609
3	2,3534	3,1825	5,8409	20	1,7247	2,0860	2,8453
4	2,1318	2,7764	4,6041	21	1,7207	2,0796	2,8314
5	2,0150	2,5706	4,0321	22	1,7171	2,0739	2,8188
6	1,9432	2,4469	3,7074	23	1,7139	2,0687	2,8073
7	1,8946	2,3646	3,4995	24	1,7109	2,0639	2,7969
8	1,8595	2,3060	3,3554	25	1,7081	2,0595	2,7874
9	1,8331	2,2622	3,2498	26	1,7056	2,0555	2,7787
10	1,8125	2,2281	3,1693	27	1,7033	2,0518	2,7707
11	1,7959	2,2010	3,1058	28	1,7011	2,0484	2,7633
12	1,7823	2,1788	3,0545	29	1,6991	2,0452	2,7564
13	1,7709	2,1604	3,0123	30	1,6973	2,0423	2,7500
14	1,7613	2,1448	2,9768	40	1,6839	2,0211	2,7045
15	1,7530	2,1315	2,9467	60	1,6707	2,0003	2,6603
16	1,7459	2,1199	2,9208	120	1,6577	1,9799	2,6174
17	1,7396	2,1098	2,8982	∞	1,6449	1,9600	2,5758

3. Значение *F*-критерия Фишера

при уровне значимости 0,05

$d.f._2$	$d.f._1$																
	1	2	3	4	5	6	7	8	9	10	11	12	14	16	20	30	∞
1	161	200	216	225	230	234	237	239	241	242	243	244	245	246	248	250	254
2	18,51	19,00	19,16	19,25	19,30	19,33	19,36	19,37	19,38	19,39	19,40	19,41	19,42	19,43	19,44	19,46	19,50
3	10,13	9,55	9,28	9,19	9,01	8,94	8,88	8,84	8,81	8,78	8,76	8,74	8,71	8,69	8,66	8,62	8,53
4	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6,00	5,96	5,93	5,91	5,87	5,84	5,80	5,74	5,63
5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,78	4,74	4,70	4,68	4,64	4,60	4,56	4,50	4,36
6	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10	4,06	4,03	4,00	3,96	3,92	3,87	3,81	3,67
7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,63	3,60	3,57	3,52	3,49	3,44	3,38	3,2 3
8	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39	3,34	3,31	3,28	3,23	3,20	3,15	3,08	2,93
9	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,13	3,10	3,07	3,02	2,98	2,93	2,86	2,71
10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,97	2,94	2,91	2,86	2,82	2,77	2,7 0	2,54
11	4,84	3,98	3,59	3,36	3,20	3,09	3,01	2,95	2,90	2,86	2,82	2,79	2,74	2,70	2,65	2,57	2,40
12	4,75	3,88	3,49	3,26	3,11	3,00	2,92	2,85	2,80	2,76	2,72	2,69	2,64	2,60	2,54	2,46	2,3 0
13	4,67	3,80	3,41	3,18	3,02	2,92	2,84	2,77	2,72	2,67	2,63	2,60	2,55	2,51	2,46	2,38	2,21
14	4,60	3,74	3,34	3,11	2,96	2,85	2,77	2,70	2,65	2,60	2,56	2,53	2,48	2,44	2,39	2,31	2,13
15	4,54	3,68	3,29	3,06	2,90	2,79	2,70	2,64	2,59	2,55	2,51	2,48	2,43	2,39	2,33	2,25	2,07
16	4,49	3,63	3,24	3,01	2,85	2,74	2,66	2,59	2,54	2,49	2,45	2,42	2,37	2,33	2,28	2,20	2,01
17	4,45	3,59	3,20	2,96	2,81	2,70	2,62	2,55	2,50	2,45	2,41	2,38	2,33	2,29	2,23	2,15	1,96
18	4,41	3,55	3,16	2,93	2,77	2,66	2,58	2,51	2,46	2,41	2,37	2,34	2,29	2,25	2,19	2,11	1,92
19	4,38	3,52	3,13	2,90	2,74	2,63	2,55	2,48	2,43	2,38	2,34	2,31	2,26	2,21	2,15	2,07	1,88
20	4,35	3,49	3,10	2,87	2,71	2,60	2,52	2,45	2,40	2,35	2,31	2,28	2,23	2,18	2,12	2,04	1,84
21	4,32	3,47	3,07	2,84	2,68	2,57	2,49	2,42	2,37	2,32	2,28	2,25	2,20	2,15	2,09	2,00	1,81
22	4,30	3,44	3,05	2,82	2,66	2,55	2,47	2,40	2,35	2,30	2,26	2,23	2,18	2,13	2,07	1,98	1,78
23	4,28	3,42	3,03	2,80	2,64	2,53	2,45	2,38	2,32	2,28	2,24	2,20	2,14	2,10	2,04	1,96	1,7 6
24	4,26	3,40	3,01	2,78	2,62	2,51	2,43	2,36	2,30	2,26	2,22	2,18	2,13	2,09	2,02	1,94	1,73
25	4,24	3,88	2,99	2,76	2,60	2,49	2,41	2,34	2,26	2,24	2,20	2,16	2,11	2,06	2,00	1,92	1,71
26	4,22	3,37	2,98	2,74	2,59	2,47	2,39	2,32	2,27	2,22	2,18	2,15	2,10	2,05	1,99	1,90	1,69
27	4,21	3,35	2,96	2,73	2,57	2,46	2,37	2,30	2,25	2,20	2,16	2,13	2,08	2,03	1,97	1,88	1,67
28	4,20	3,34	2,95	2,71	2,56	2,44	2,36	2,29	2,24	2,19	2,15	2,12	2,06	2,02	1,96	1,87	1,65
29	4,18	3,33	2,93	2,70	2,54	2,43	2,35	2,28	2,22	2,18	2,14	2,10	2,05	2,00	1,94	1,85	1,64
30	4,17	3,32	2,92	2,69	2,53	2,42	2,34	2,27	2,21	2,16	2,12	2,09	2,04	1,99	1,93	1,84	1,62
40	4,08	3,23	2,84	2,61	2,45	2,34	2,25	2,18	2,12	2,07	2,04	2,00	1,95	1,90	1,84	1,74	1,51
50	4,03	3,18	2,79	2,56	2,40	2,29	2,20	2,13	2,07	2,02	1,98	1,95	1,90	1,85	1,78	1,69	1,44
60	4,00	3,15	2,76	2,52	2,37	2,25	2,17	2,10	2,04	1,99	1,95	1,92	1,86	1,81	1,75	1,65	1,39
100	3,94	3,09	2,70	2,46	2,30	2,19	2,10	2,03	1,97	1,92	1,88	1,85	1,79	1,75	1,68	1,57	1,28
∞	3,84	2,99	2,60	2,37	2,21	2,09	2,01	1,94	1,88	1,83	1,79	1,75	1,69	1,64	1,57	1,46	1,00

$d.f._1$ – число степеней свободы для большей дисперсии; $d.f._2$ – число степеней свободы для меньшей дисперсии.

4. Значение χ^2-критерия Пирсона при уровне значимости 0,10, 0,05, 0,01

d.f.	0,10	0,05	0,01	*d.f.*	0,10	0,05	0,01
1	2,71	3,84	6,63	21	29,62	32,67	38,93
2	4,61	5,99	9,21	22	30,81	33,92	40,29
3	6,25	7,81	11,34	23	32,01	35,17	41,64
4	7,78	9,49	13,28	24	33,20	36,42	42,98
5	9,24	11,07	15,09	25	34,38	37,65	44,31
6	10,64	12,59	16,81	26	35,56	38,89	45,64
7	12,02	14,07	18,48	27	36,74	40,11	46,96
8	13,36	15,51	20,09	28	37,92	41,34	48,28
9	14,68	16,92	21,67	29	39,09	42,56	49,59
10	50,99	18,31	23,21	30	40,26	43,77	50,89
11	17,28	19,68	24,72	40	51,80	55,76	63,69
12	18,55	21,03	26,22	50	63,17	67,50	76,15
13	19,81	22,36	27,69	60	74,40	79,08	88,38
14	21,06	23,68	29,14	70	85,53	90,53	100,42
15	22,31	25,00	30,58	80	96,58	101,88	112,33
16	23,54	26,30	32,00	90	107,56	113,14	124,12
17	24,77	27,59	33,41	100	118,50	124,34	135,81
18	25,99	28,87	34,81				
19	27,20	30,14	36,19				
20	28,41	31,14	37,57				

5. Критические значения коэффициентов корреляции для уровней значимости 0,05, 0,01

d.f.	$\alpha = 0,05$	$\alpha = 0,01$	*d.f.*	$\alpha = 0,05$	$\alpha = 0,01$
1	0,996917	0,9998766	17	0,4555	0,5751
2	0,995000	0,990000	18	0,4438	0,5614
3	0,8783	0,95873	19	0,4329	0,5487
4	0,8114	0,91720	20	0,4227	0,5368
5	0,7545	0,8745	25	0,3809	0,4869
6	0,07067	0,8343	30	0,3494	0,4487
7	0,6664	0,7977	35	0,3246	0,4182
8	0,6319	0,7646	40	0,3044	0,3932
9	0,6021	0,7348	45	0,2875	0,3721
10	0,5760	0,7079	50	0,2732	0,3541
11	0,5529	0,6835	60	0,2500	0,3248
12	0,5324	0,6614	70	0,2919	0,3017
13	0,5139	0,6411	80	0,2172	0,2830
14	0,4973	0,6226	90	0,2050	0,2673
15	0,4821	0,6055	100	0,1946	0,2540
16	0,4683	0,5897			

Для простой корреляции *d.f.* на 2 меньше, чем число пар вариантов; в случае частной корреляции необходимо также вычесть число исключаемых переменных.

6. *Z*-преобразование. Значения величины *z* для значений *r*

r	0	1	2	3	4	5	6	7	8	9
0,0	0,0000	0,0100	0,0200	0,0300	0,0400	0,0501	0,0601	0,0701	0,0802	0,0902
0,1	0,1003	0,1105	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	0,3206	0,3317	0,3428	0,3541	0,3654	0,3769	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	0,5763	0,5901	0,6042	0,6184	0,6328	0,6475	0,6625	0,6777
0,6	0,6931	0,7089	0,7250	0,7414	0,7582	0,7753	0,7928	0,8107	0,8291	0,8480
0,7	0,8673	0,8872	0,9076	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	1,7380	1,8318	1,9459	2,0923	2,2976	2,6467

7. Таблица случайных чисел

Ряд	Колонка							
	12345	67890	12345	67890	12345	67890	12345	67890
01	66194	28926	99547	16625	45515	67953	12108	57846
02	78240	43195	24837	32511	70880	22070	52622	61881
03	00833	88000	67299	68215	11274	55624	32991	17436
04	12111	86683	61270	58036	64192	90611	15145	01748
05	47189	99951	05755	03834	43782	90599	40282	51417
06	76396	72486	62423	27618	84184	78922	73561	52818
07	46409	17469	32483	09083	76175	19985	26309	91536
08	74626	22111	87286	46772	42243	68046	44250	42439
09	34450	81974	93723	49023	58432	67083	36876	93391
10	36327	72135	33005	28701	34710	49359	50693	89311
11	74185	77536	84825	09934	99103	09325	67389	45869
12	12296	41623	62873	37943	25584	09609	63360	47270
13	90822	60280	88925	99610	42772	60561	76873	04117
14	72121	79152	96591	90305	10189	79778	68016	13747
15	95268	41377	25684	08151	61816	58555	54305	86189
16	92603	09091	75884	93424	72586	88903	30061	14457
17	18813	90291	05275	01223	79607	95426	34900	09778
18	38840	26903	28624	67157	51986	42865	14508	49315
19	05959	33836	53758	16562	41081	38012	41230	20528
20	85141	21155	99212	32685	51403	31926	69813	58781
21	75047	59643	31074	38172	03718	32119	69506	67143
22	30752	95260	68032	62871	58781	34143	68790	69766
23	22986	82575	42187	62295	84295	30634	66562	31442
24	99439	86692	90348	66036	48399	73451	26698	39437
25	20389	93029	11881	71685	65452	89047	63669	02656
26	39249	05173	68256	36359	20250	68686	05947	09335
27	96777	33605	29481	20063	09398	01843	35139	61344
28	04860	32918	10798	50492	52655	33359	94713	28393
29	41613	42375	00403	03656	77580	87772	86877	57085
30	17930	00794	53836	53692	67135	98102	61912	11246
31	24649	31845	25736	75231	83808	98917	93829	99430
32	79899	34061	54308	59358	56462	58166	97302	86828
33	76801	49594	81002	30397	52728	15101	72070	33706

Продолжение

Ряд	Колонка							
	12345	67890	12345	67890	12345	67890	12345	67890
34	36239	63636	38140	65731	39788	06872	38971	53363
35	07392	64449	17886	63632	53995	17574	22247	62607
36	67133	04181	33874	98835	67453	59734	76381	63455
37	77759	31504	32832	70861	15152	29733	75371	39174
38	85992	72268	42920	20810	29361	51423	90306	73574
39	79553	75952	54116	65553	47139	60579	09165	85490
40	41101	17336	48951	53674	17880	45260	08575	49321
41	36191	17095	32123	91576	84221	78902	82010	30874
42	62329	63898	23268	74283	26091	68409	69704	82267
43	14751	13151	93115	01437	56945	89661	67680	79790
44	48462	59278	44185	29616	76537	19589	83139	28454
45	29435	88105	59651	44391	74588	55114	80834	85686
46	28340	29285	12965	14821	80425	16602	44653	70467
47	02167	58940	27149	80242	10587	79786	34959	75339
48	17864	00991	39557	54981	23588	81914	37609	13128
49	79675	80605	60059	35862	00254	36546	21545	78179
50	72335	82037	92003	34100	29879	46613	89720	13274
51	49280	88924	35779	00283	81163	07275	89863	02348
52	61870	41657	07468	08612	98083	97349	20775	45091
53	43898	65923	25078	86129	78496	97653	91550	08078
54	62993	93912	30454	84598	56095	20664	12872	64647
55	33850	58555	51438	85507	71865	79488	76783	31708
56	55336	71264	88472	04334	63919	36394	11095	92470
57	70543	29776	10087	10072	55980	64688	68239	20461
58	89382	93809	00796	95945	34101	81277	66090	88872
59	37818	72142	67140	50785	22380	16703	53362	44940
60	60430	22834	14130	96593	23298	56203	92671	15925
61	82975	66158	84731	19436	55790	69229	28661	13675
62	39087	71938	40355	54324	08401	26299	49420	59208
63	55700	24586	93247	32596	11865	63397	44251	43189
64	14756	23997	78643	75912	83832	32768	18928	57070
65	32166	53251	70654	92827	63491	04233	33825	69662
66	23236	73751	31888	81718	06546	83246	47651	04877

Продолжение

Ряд	Колонка							
	12345	67890	12345	67890	12345	67890	12345	67890
67	45794	26926	15130	82455	78305	55058	52551	47182
68	09893	20505	14225	68514	46427	56788	96297	78822
69	54382	74598	91499	14523	68479	27686	46162	83554
70	94750	89923	37089	20048	80336	94598	26940	36858
71	70297	34135	53140	33340	42050	82341	44104	82949
72	85157	47954	32979	26575	57600	40881	12250	73742
73	11100	02340	12860	74697	96644	89439	28707	25815
74	36871	50775	30592	57143	17381	68856	25853	35041
75	23913	48357	63308	16090	51690	54607	72407	55538
76	79348	36085	27973	65157	07456	22255	25626	57054
77	92074	54641	53673	54421	18130	60103	69593	49464
78	06873	21440	75593	41373	49502	17972	82578	16364
79	12478	37622	99659	31065	83613	69889	58869	29571
80	57175	55564	65411	42547	70457	03426	72937	83792
81	91616	11075	80103	07831	59309	13276	26710	73000
82	78025	73539	14621	39044	47450	03197	12787	47709
83	27587	67228	80145	10175	12822	86687	65530	49325
84	16690	20427	04251	64477	73709	73945	92396	68263
85	70183	58065	65489	31833	82093	16747	10386	59293
86	90730	35385	15679	99742	50866	78028	75573	67257
87	10934	93242	13431	24590	02770	48582	00906	58595
88	82462	30166	79613	47416	13389	80268	05085	96666
89	27463	10433	07606	16285	93699	60912	94532	95632
90	02979	52997	09079	92709	90110	47506	53693	49892
91	46888	69929	75233	52507	32097	37594	10067	67327
92	53638	83161	08289	12639	08141	12640	28437	09268
93	82433	61427	17239	89160	19666	08814	37841	12847
94	35766	31672	50082	22795	66948	65581	84393	15890
95	10853	42581	08792	13257	61973	24450	52351	16602
96	20341	27398	72906	63955	17276	10646	74692	48438
97	54458	90542	77563	51839	52901	53355	83281	19177
98	26337	66530	16687	35179	46560	00123	44546	79896
99	34314	23729	85264	05575	96855	23820	11091	79821
00	28603	10708	68933	34189	92166	15181	66628	58599

ОГЛАВЛЕНИЕ

Предисловие 3

Глава 1. **Понятие о статистике** 5
1.1. Что такое статистика 5
1.2. Статистическая закономерность. Статистические совокупности 7
1.3. Признаки и их классификация 12
1.4. Определение предмета статистики – основа статистической методологии 15

Глава 2. **Статистическое наблюдение** 19
2.1. Организация государственной статистики в Российской Федерации и международной статистики 19
2.2. Требования, предъявляемые к собираемым данным. Формы организации и виды статистического наблюдения 23
2.3. Подготовка статистического наблюдения 29
2.4. Статистическая отчетность 35
2.5. Ошибки статистического наблюдения. Методы контроля данных наблюдения 38

Глава 3. **Статистические показатели** 42
3.1. Сущность и значение статистических показателей. Показатель и его атрибуты 42
3.2. Классификация статистических показателей 44
3.3. Общие принципы построения относительных статистических показателей 50
3.4. Понятие о системах статистических показателей . . 52
3.5. Функции статистических показателей 54

Глава 4. **Представление статистических данных: таблицы и графики** 58
4.1. Статистические таблицы 58
4.2. Основные виды графиков 63
4.3. Картограммы и картодиаграммы 71
Глава 5. **Средние величины и изучение вариации** 74
5.1. Однородность и вариация массовых явлений 74
5.2. Средняя арифметическая величина 77
5.3. Другие формы средних величин 86
5.4. Средняя величина как выражение закономерности . . 90
5.5. Вариация массовых явлений 92
5.6. Построение вариационного ряда. Виды рядов. Ранжирование данных 94
5.7. Структурные характеристики вариационного ряда . . 100
5.8. Показатели размера и интенсивности вариации . . . 104
5.9. Моменты распределения и показатели его формы . . 109
5.10. Предельно возможные значения показателей вариации и их применение 113
Глава 6. **Группировка** 118
6.1. Значение и сущность группировки 118
6.2. Виды группировок 122
6.3. Многомерные группировки 133
Глава 7. **Выборочное наблюдение. Испытание статистических гипотез** 156
7.1. Причины применения выборочного наблюдения. Дескриптивная статистика и статистический вывод . . . 156
7.2. Способы отбора, обеспечивающие репрезентативность выборки. Виды выборки 160
7.3. Ошибка выборки 164
7.4. Влияние вида выборки на величину ошибки выборки 171
7.5. Задачи, решаемые при применении выборочного метода 181
7.6. Распространение данных выборочного наблюдения на генеральную совокупность. 187
7.7. Малая выборка 190
7.8. Статистическая проверка гипотез (общие понятия) . 193
7.9. Проверка гипотезы о законе распределения 196
7.10. Проверка гипотезы о связи на основе критерия χ^2 (хи-квадрат) 203
7.11. Проверка гипотезы о средних величинах 208
7.12. Основы дисперсионного анализа 211
7.13. Примеры применения выборочного метода и проверки статистических гипотез 218

Глава 8. **Корреляционно-регрессионный анализ и моделирование статистических связей** 226
8.1. Понятие о статистической и корреляционной связи . 226
8.2. Условия применения и ограничения корреляционно-регрессионного метода 230
8.3. Задачи корреляционно-регрессионного анализа и моделирования 232
8.4. Вычисление и интерпретация параметров парной линейной корреляции 238
8.5. Статистическая оценка надежности параметров парной корреляции 247
8.6. Применение парного линейного уравнения регрессии 250
8.7. Вычисление параметров парной линейной корреляции на основе аналитической группировки 253
8.8. Коэффициент корреляции рангов. 259
8.9. Параболическая корреляция 263
8.10. Гиперболическая корреляция 266
8.11. Множественное уравнение регрессии 268
8.12. Меры тесноты связей в многофакторной системе. . . 275
8.13. Вероятностные оценки параметров множественной регрессии и корреляции 283
8.14. Корреляционно-регрессионные модели (КРМ) и их применение в анализе и прогнозе 285
8.15. Измерение связи неколичественных признаков. . . . 290
Глава 9. **Статистическое изучение динамики** 301
9.1. Составляющие элементы динамики: основная тенденция и колебания 301
9.2. Показатели, характеризующие тенденцию динамики . 304
9.3. Особенности показателей динамики для рядов, состоящих из относительных уровней 309
9.4. Средние показатели тенденции динамики 313
9.5. Методы выявления типа тенденции динамики 321
9.6. Методика измерения параметров тренда 329
9.7. Методика изучения и показатели колеблемости . . . 338
9.8. Измерение устойчивости в динамике. 345
9.9. Сезонные колебания и полное разложение дисперсии уровней динамического ряда 347
9.10. Прогнозирование на основе тренда и колеблемости . 357
9.11. Корреляция рядов динамики 361
Глава 10. **Индексы** 368
10.1. Понятие индекса 368
10.2. Индекс как показатель центральной тенденции (индекс средний из индивидуальных) 370

Глава 10. Индексы
10.3. Агрегатные индексы. Система индексов 379
10.4. Свойства индексов 386
10.5. Индексный анализ взвешенной средней. Индекс структуры 388
10.6. Построение индексов при обобщении данных по единицам совокупности и по элементам 392
10.7. Границы и условия применения индексного метода . 402
10.8. Комплексное использование индексного и регрессионного методов анализа 407
10.9. Примеры использования индексов в экономико-статистических расчетах 424

Глава 11. **Статистическое изучение структуры совокупности и ее изменений** 433
11.1. Показатели простой (одномерной) структуры 433
11.2. Показатели иерархической «древовидной» структуры 435
11.3. Показатели балансовой структуры 439
11.4. Показатели многомерной структуры с пересекающимися признаками 444
11.5. Сравнительный анализ структур 446
11.6. Показатели концентрации, специализации, монополизации. Многомерная структура 452
11.7. Абсолютные и относительные показатели изменения структуры 458
11.8. Ранговые и инновационные показатели изменения структуры 461

Приложение. Статистико-математические таблицы 468

Е. К.

Учебное издание

Елисеева Ирина Ильинична
Юзбашев Михаил Михайлович

ОБЩАЯ ТЕОРИЯ СТАТИСТИКИ

Заведующая редакцией *Л.А. Табакова*
Редактор *Е.В. Стадниченко*
Художественный редактор *Ю.И. Артюхов*
Технический редактор *И.В. Завгородняя*
Корректоры *Е.В. Люминарская, Г.В. Хлопцева*
Обложка художника *Н.М. Биксентеева*

ИБ № 3853

Лицензия ЛР № 010156 от 29.01.97

Подписано в печать 01.10.2001.
Формат 60×88/16. Печать офсетная.
Гарнитура «Таймс».
Усл. п. л. 29,4. Уч.-изд. л. 28,7.
Доп. тираж 7000 экз. Заказ 3182. «С» 012

Издательство «Финансы и статистика»
101000, Москва, ул. Покровка, 7
Телефон (095) 925-35-02, факс (095) 925-09-57
E-mail: mail@finstat.ru, *http*://www.finstat.ru

ГУП «Великолукская городская типография»
Комитета по средствам массовой информации и связям
с общественностью администрации Псковской области,
182100, Великие Луки, ул. Полиграфистов, 78/12
Тел./факс: (811-53) 3-62-95
E-mail: VTL@MART.RU