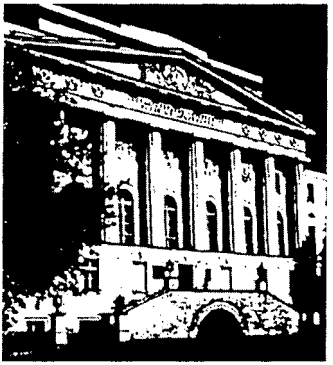
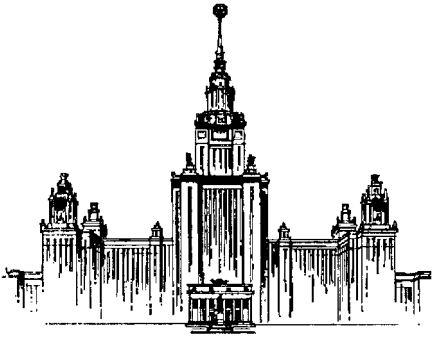
КЛАССИЧЕСКИЙ
УНИВЕРСИТЕТСКИЙ УЧЕБНИК

Серия основана в 2002 году по инициативе ректора
МГУ им. М.В. Ломоносова
академика РАН В.А. Садовничего
и посвящена

250-летию
Московского университета



КЛАССИЧЕСКИЙ
УНИВЕРСИТЕТСКИЙ УЧЕБНИК



Московский государственный университет имени М.В. Ломоносова

ТЕОРИЯ СТАТИСТИКИ

Учебник

Под редакцией профессора Г.Л. Громыко

2-е издание, переработанное и дополненное

Рекомендовано Министерством образования Российской Федерации в качестве учебника для студентов экономических специальностей высших учебных заведений

Москва
ИНФРА-М
2005

УДК 311(075.8)
ББК 65.051я73
Т11

Авторы: проф. Г.Л. Громыко (предисловие, главы 1, 2, 7 (кроме параграфа 7.8), 8 и 9 (кроме параграфа 9.8), приложения, предметный указатель);
доц. А.Н. Воробьев (главы 3 и 4);
доц. С.Е. Казаринова (глава 5 (кроме параграфов 5.4, 5.7 и 5.8)); доц. И.П. Мамий (параграфы 5.4, 5.7 и 5.8);
доц. Л.А. Карасева (глава 6);
доц. И.Н. Матюхина (параграф 7.8);
проф. Ю.Н. Иванов (параграф 9.8).

Рецензенты: кафедра статистики Московского государственного университета коммерции (зав. кафедрой — д-р экон. наук, проф. *О.Э. Башина*) и д-р экон. наук, проф. *М.В. Карманов*

Теория статистики: Учебник/Под ред. проф. Г.Л. Громыко. –
Т11 2-е изд., перераб. и доп. – М.: ИНФРА-М, 2005. – 476 с. – (Классический университетский учебник).

ISBN 5-16-002158-2

В учебнике рассмотрен широкий круг вопросов статистической методологии: организация статистического наблюдения, обработка данных и их анализ. Особое внимание уделено статистическим методам анализа вариационных рядов и рядов динамики, выборочному наблюдению, изучению корреляционных связей, индексному методу. В приложении содержатся основные формулы и математические таблицы, используемые при анализе статистических данных и проверке различных гипотез.

Для студентов и преподавателей экономических вузов и факультетов.

ББК 65.051я73

ISBN 5-16-002158-2

Уважаемый читатель!

Вы открыли одну из замечательных книг, изданных в серии «Классический университетский учебник», посвященной 250-летию Московского университета. Серия включает свыше 150 учебников и учебных пособий, рекомендованных к изданию Учеными советами факультетов, редакционным советом серии и издаваемых к юбилею по решению Ученого совета МГУ.

Московский университет всегда славился своими профессорами и преподавателями, воспитавшими не одно поколение студентов, впоследствии внесших заметный вклад в развитие нашей страны, составивших гордость отечественной и мировой науки, культуры и образования.

Высокий уровень образования, которое дает Московский университет, в первую очередь обеспечивается высоким уровнем написанных выдающимися учеными и педагогами учебников и учебных пособий, в которых сочетаются как глубина, так и доступность излагаемого материала. В этих книгах аккумулируется бесценный опыт методики и методологии преподавания, который становится достоянием не только Московского университета, но и других университетов России и всего мира.

Издание серии «Классический университетский учебник» наглядно демонстрирует тот вклад, который вносит Московский университет в классическое университетское образование в нашей стране и, несомненно, служит его развитию.

Решение этой благородной задачи было бы невозможным без активной помощи со стороны издательств, принявших участие в издании книг серии «Классический университетский учебник». Мы расцениваем это как поддержку ими позиции, которую занимает Московский университет в вопросах науки и образования. Это служит также свидетельством того, что 250-летний юбилей Московского университета — выдающееся событие в жизни всей нашей страны, мирового образовательного сообщества.

Ректор Московского университета
академик РАН, профессор *В. А. Садовничий*

Предисловие

Курс «Теория статистики» – одна из важных и обязательных дисциплин в учебных планах экономических вузов, поскольку высшее экономическое образование должно включать в себя как неотъемлемую составляющую статистическую грамотность.

Каждый экономист, работая с числами (статистическими данными), обязан владеть статистической методологией, т.е. знать основные приемы и методы изучения массовых данных, их сбора, обработки и анализа, понимать, как получены те или иные исходные данные, какова их природа, насколько они полны и достоверны; должен уметь обобщать их и оформлять в виде таблиц. Кроме того, экономист должен уметь использовать различные статистические методы анализа массовых данных при решении конкретных задач. Всему этому учит курс «Теория статистики».

Настоящий учебник написан коллективом кафедры статистики экономического факультета МГУ им. М.В. Ломоносова под руководством доктора экономических наук, профессора Г.Л. Громыко на основе обобщения многолетнего опыта преподавания данного курса в упомянутом вузе.

В учебнике рассмотрены основные приемы и методы сбора, обработки и анализа статистических данных, позволяющие: выявить особенности распределения единиц совокупностей по тем или иным признакам; определить средние значения отдельных показателей и их вариацию; выявить и измерить взаимосвязи между показателями; определить тенденцию развития отдельных показателей (явлений); оценить результаты выборочных данных и т.д.

Учебник соответствует программе данного курса, рассчитанного всего на 68 часов (аудиторных). Исходя из этого, некоторые темы (например, «Предмет и метод статистики», «Статистическое наблюдение», «Сводка и группировка данных») даны в кратком изложении. Так, в главе 3 дано самое общее представление о методе группировок в статистике и не рассмотрены более сложные методы многомерных классификаций, по которым существует специальная литература, в частности учебник А.М. Дуброва, В.С. Мхитаряна и Л.И. Трошина «Многомерные статистические

методы» (М.: Финансы и статистика, 1998), и имеются пакеты прикладных программ для ЭВМ. Также в главе 1 кратко, в общих чертах, упомянуты основные вехи развития статистики как науки и ее выдающиеся деятели. При этом учитывалось, что подробную информацию по этому вопросу читатель может получить в работе Б.Г. Плошко и И.И. Елисеевой «История статистики» (М.: Финансы и статистика, 1990).

Надо также отметить, что в предлагаемом учебнике «Теория статистики» отсутствует специальная глава (тема) «Графическое изображение статистических данных», но по ходу изложения в каждой главе, где это было необходимо, приведены графики с описанием их построения.

Основной акцент в данном издании учебника сделан на анализе вариационных рядов и рядов динамики (временны́х рядов), на методах изучения взаимосвязей в статистике, оценке результатов выборочных наблюдений и индексном методе.

Глава 1
ПРЕДМЕТ И МЕТОД СТАТИСТИКИ

1.1. Понятие статистики

Статистика – одна из древнейших отраслей знаний, возникшая на базе хозяйственного учета.

Первые учетные операции проводились еще в глубокой древности. Вначале они были довольно примитивны, нерегулярны и направлены главным образом на получение данных о численности населения, его составе и имущественном положении. Эти данные использовались прежде всего при налогообложении и в военных нуждах.

По мере развития производительных сил в обществе возрастал интерес к различного рода знаниям, расширялся круг учитываемых явлений и собираемых о них сведений; усложнялись сами учетные операции, они стали более регулярными. Постепенно накапливался опыт, появлялись рекомендации о том, каким образом организовать отдельные учетные операции и обработать собранные сведения, чтобы обобщить их и выявить различные закономерности.

Так постепенно сформировалась отрасль знаний, названная впоследствии «статистикой». Ее возникновение связано с потребностями общества в различного рода сведениях, информации, без которых невозможно управлять государством, изучать отдельные явления и процессы, происходящие в различных областях жизни, сферах деятельности.

Есть основания полагать, что термин «статистика» произошел от латинских слов *stato* (государство) и *status* (положение вещей, политическое состояние). До середины XVIII в. под статистикой подразумевалась совокупность сведений о государстве, о его достопримечательностях. И именно в таком понимании с 1666 г. в университетах Германии стал преподаваться курс «Государствоведение». Термин «статистика» в научный обиход ввел немецкий ученый Готфрид Ахенваль, представитель описательной школы государствоведения. В 1746 г. он предложил заменить название курса «Государствоведение», преподававшегося в университетах Германии, на «Статистику», положив тем самым начало развитию статистики как науки и учебной дисциплины.

В настоящее время термин «статистика» употребляется в нескольких значениях.

1. Статистикой часто называют ***совокупность сведений*** (фактов) о разных явлениях в той или иной стране или ее регионах, например:

сведения о численности и составе населения, о рождаемости, смертности, миграции и прочем (статистика населения);

сведения о доходах и расходах населения, о среднемесячной номинальной заработной плате, о размерах пенсий, потреблении различных продуктов питания на душу населения, величине прожиточного минимума и прочем (статистика уровня жизни);

сведения о числе промышленных предприятий, их отраслевой структуре и распределении по формам собственности, об объеме производимой продукции и прибыли, о численности занятых и прочем (статистика промышленности) и т.д.

Такие сведения обычно публикуются в специальных изданиях, справочниках. В России в настоящее время основным официальным изданием является «Российский статистический ежегодник» (в сокращенном варианте «Россия в цифрах»). Кроме того, издается целый ряд специализированных тематических статистических сборников: «Демографический ежегодник России», «Промышленность России», «Цены в России», «Социальное положение и уровень жизни населения России» и др.

Статистика как совокупность сведений о той или иной стране наиболее близка к понятию «государствоведение». Вместе с тем современная статистика существенно отличается от «государствоведения» прошлых лет как полнотой и разносторонностью содержащихся в ней сведений, так и характером последних.

В частности, к статистике относят сведения только о том, что имеет *количественное выражение*. Например, к статистике не относят сведения о государственном устройстве страны или о том, какой язык в ней признан государственным; но сведения о том, сколько в мире государств с тем или иным политическим устройством, формой правления (монархий, республик разного типа и пр.), относят к статистике, так же как и сведения о том, сколько жителей страны пользуется тем или иным языком в качестве разговорного.

Отметим еще одну особенность сведений, составляющих статистику в рассматриваемом значении.

Все они относятся не к единичному (индивидуальному) явлению, а к *совокупности*, состоящей из множества элементов. Другими словами, статистику интересуют только такие сведения об

индивидуальных явлениях, на основе которых можно получить *сводные* данные. Например, регистрируя каждый случай рождения или смерти, можно получить сводные данные по стране (или ее регионам) о числе родившихся и умерших, об уровне рождаемости и смертности.

2. Под статистикой понимают также процесс получения сведений с последующей их обработкой. В этом смысле статистика – ***практическая деятельность*** людей, направленная на сбор, обработку и анализ массовых данных, относящихся к тем или иным сферам общественной жизни.

Сбор сведений в целях обобщения можно вести в разных масштабах. Например, врач может вести статистику своих пациентов, чтобы обобщить опыт лечения; учитель может вести статистику своих учеников (их успеваемости), чтобы определить, какова результативность того или иного метода обучения, и т.д. Однако сбор сведений или статистической информации, касающейся всего государства или отдельных его регионов, возлагается на специальные органы – органы государственной статистики и органы ведомственной статистики, которые в совокупности можно назвать *статистической службой.*

В России формирование органов специальной государственной статистической службы относится к началу XIX в. (в 1811 г. создано статистическое отделение при Министерстве полиции), но окончательно сложившейся систему государственной статистики можно считать со второй половины XIX в., когда в 1858 г. был образован Центральный статистический комитет (ЦСК) при МВД. В последующие годы, и особенно в советский период, название и подчинение центрального статистического органа страны неоднократно менялось: с 1918 г. – ЦСУ (Центральное статистическое управление), с 1931 г. – ЦУНХУ (Центральное управление народнохозяйственного учета) при Госплане СССР, с 1940 г. – снова ЦСУ, но при Госплане СССР, с 1950 г. – ЦСУ при Совете Министров, с 1987 г. – Госкомстат СССР, после распада СССР – Госкомстат РФ.

С мая 2004 г. главным центральным органом *государственной статистики* в России является Федеральная служба государственной статистики (ФСГС). На нее, как и ранее на Госкомстат России, возлагается руководство всей статистической деятельностью в стране, осуществление сбора, обработки и анализа данных о развитии национальной экономики, представление Президенту, Правительству, Федеральному собранию и другим федеральным органам власти официальной статистической информации о социально-экономическом состоянии страны.

ФСГС России должна обеспечивать единство методологии учета и статистики в стране, разрабатывать формы отчетности, проводить (с помощью местных органов государственной статистики) переписи населения, осуществлять единовременный учет различных объектов исследования, публиковать статистические данные о социально-экономическом развитии страны и ее регионов.

Наряду с государственной существует и *ведомственная статистика*, которую ведут все министерства и ведомства на основе информации, поступающей к ним от подведомственных предприятий и организаций. Такая информация, ее обработка и анализ необходимы министерствам и ведомствам прежде всего для оперативного руководства и управления деятельностью соответствующих подведомственных производственных и других единиц, а также для ее планирования.

3. Под термином «статистика» понимают также некий ***параметр*** ряда случайных величин (x_1, x_2, ..., x_n), получаемый по определенному алгоритму из результатов индивидуальных наблюдений. Таким параметром – статистикой – являются средняя арифметическая значений x_1, x_2, ..., x_n, мода, среднее квадратическое отклонение и др. К числу статистик в этом смысле относятся разнообразные *статистические критерии* (критические статистики), которые применяют при проверке различных статистических гипотез (предположительных утверждений) относительно природы или значений отдельных показателей исследуемых данных, особенностей их распределения и пр.

Термин «статистика» как параметр, как статистический критерий употребляется преимущественно в математической статистике. Некоторые из таких статистик (χ^2, t и др.) рассмотрены в соответствующих главах данного учебника.

4. Наконец, под статистикой в широком смысле понимают ***науку***, изучающую с количественной стороны массовые явления и их закономерности.

Статистика как наука содержит теоретические положения о методах изучения массовых явлений.

1.2. Краткий обзор развития статистики как науки

Статистика как наука возникла в XVII в. почти одновременно **в Германии** и **Англии**. Ее зарождение произошло в недрах развившейся и расширившейся практики учетно-статистических работ. Несомненно, что говорить об оформлении статистики как науки стало возможным только тогда, когда появились первые научные труды, посвященные изучению массовых явлений, государства,

общества, когда в вузах было введено преподавание статистики как учебной дисциплины.

Развитие статистики как науки шло по двум направлениям.

Первое направление возникло в Германии и известно как ***государствоведение***, или ***описательная школа***. Представители этой школы основной своей задачей считали описание достопримечательностей государства: территории, населения, климата, политического устройства, вероисповедания, ведения хозяйства, торговли, благосостояния государства и граждан и т.п. – без анализа закономерностей и взаимосвязей между явлениями.

Основателем описательной школы был немецкий ученый **Герман Конринг** (1606–1681). Он же ввел и преподавание «Государствоведения» как учебной научной дисциплины в университетах Германии (1666 г.).

Много сделал для развития описательной школы и идей Конринга его последователь **Готфрид Ахенваль** (1719–1772), который, как уже указывалось, ввел термин «статистика», а также ученик последнего **Август Людвиг Шлецер** (1735–1809).

Второе направление развития статистики как науки возникло в Англии и известно под названием ***«политическая арифметика»***. Основателем школы этого направления был **Уильям Петти** (1623–1687), известный политэконом, которого К. Маркс называл отцом политической экономии и в некотором роде изобретателем статистики.

С деятельностью школы «политической арифметики» неразрывно связано имя **Джона Граунта** (1620–1674), друга и соратника У. Петти, а также имя **Эдмунда Галлея** (1656–1742), английского астронома, и др.

Представители данной школы в отличие от приверженцев государствоведения своей главной задачей считали выявление на основе большого числа наблюдений различных закономерностей и взаимосвязей в изучаемых явлениях. Так, Д. Граунт исследовал главным образом закономерности воспроизводства населения. В течение многих лет он изучал данные бюллетеней, в которых еженедельно публиковались сведения о числе родившихся и умерших в Лондоне, и сумел выявить ряд закономерностей. Например, он установил, что соотношение численности родившихся мальчиков и девочек составляло 14 : 13, что из числа родившихся до 6 лет доживало в то время 64% лондонцев, до 16 лет – 40%, что на 63 умерших приходилось 52 новорожденных и т.д.

Д. Граунт составил первую *таблицу смертности* и рассчитал *кривую дожития*. Результаты своих исследований он опубликовал

в 1662 г. в работе, название которой по традиции того времени отражало ее суть: «Естественные и политические наблюдения, перечисленные в прилагаемом оглавлении и сделанные над бюллетенями смертности, по отношению к управлению, религии, торговле, росту, болезням и пр.». Это был первый научный труд политических арифметиков.

Э. Галлей, как и Д. Граунт, интересовался естественным движением населения, составлением таблиц смертности с определением вероятности дожития до определенного возраста и использованием их в страховом деле.

У. Петти в отличие от Граунта и Галлея больше интересовался хозяйственными процессами, закономерностями в общественной и экономической жизни. Он первым, прибегнув к косвенным расчетам, попытался оценить национальное богатство и национальный доход страны. Круг его интересов отражен в работе, написанной в 1671–1676 гг., но опубликованной уже после его смерти в 1690 г. под названием «Политическая арифметика, или Рассуждения относительно размеров и стоимости земли, людей, сельского хозяйства, мануфактур, торговли, рыбной ловли, ремесленников, моряков, солдат; относительно государственных доходов, регистрации банков; относительно определения ценности людей, увеличения числа моряков; относительно портов, положения страны, кораблей, могущества на море и т.п.»*.

Школа «политической арифметики» имела немало последователей как в самой Англии, так и за ее пределами.

Заслугой политических арифметиков является то, что они понимали необходимость использования массовых данных для выявления тех или иных закономерностей, что при сводке и анализе использовали группировки, средние и относительные величины, старались рассматривать многие показатели взаимосвязанно, при отсутствии необходимых данных использовали косвенные расчеты и т.д.

Государствоведение и политическая арифметика развивались каждая своим путем, используя свои методы в исследованиях. Но предмет изучения у них был общий – государство, общество и, в частности, массовые явления и процессы, происходящие в нем.

Можно сказать, что статистика, родившись в связи с необходимостью решения практических государственных и хозяйственных проблем, сформировалась как наука в результате синтеза государствоведения и политической арифметики, причем от

* Название работы в различных источниках несколько варьирует из-за специфики перевода.

последней она взяла больше, поскольку статистика и в настоящее время призвана выявлять прежде всего различного рода закономерности в исследуемых явлениях.

Однако ни представители государствоведения, ни представители политической арифметики не дошли до теоретического обобщения практики учетно-статистических работ, до создания теории статистики. Эта задача была решена позднее, в XIX в., известным бельгийским ученым **Адольфом Кетле** (1796–1874), математиком по образованию, много лет возглавлявшим национальную статистику Бельгии. Именно он дал определение предмета статистики (массовые явления, связанные с жизнью общества, государства), увидел в ней орудие социального познания, раскрыл суть методов статистики.

Статистические исследования Кетле были посвящены в основном раскрытию различных закономерностей в общественной жизни, в частности в области преступности. По данным уголовной статистики Франции он установил постоянство в числе преступлений (и их орудий) за ряд лет и на этом основании сделал вывод о том, что общество содержит в себе зародыши всех преступлений, которые должны совершиться. Наличие устойчивых закономерностей в области социальных явлений он объяснял действием двух видов причин: «постоянных» (общих), определяющих закономерность, типичность, и «пертурбационных» (индивидуальных, случайных), вызывающих отклонения единичных явлений от типа.

Кетле считал, что основной задачей в выявлении закономерностей является взаимопогашение случайных причин. А последнее может быть достигнуто лишь при наличии массовых данных. Основным приемом статистического анализа он считал метод средних величин. Он даже ввел понятие «среднего человека».

Эти и другие идеи своего учения о предмете и методах статистического исследования Кетле изложил в работе «Социальная физика, или Опыт исследования о развитии человеческих способностей» (1836 г.), применив в ней термин «среднего человека».

Кетле считал, что теоретической основой статистики является теория вероятностей, что раскрыть (познать) различные закономерности в явлениях общественной жизни можно только на основе массовых статистических данных, при которых возможно погашение, исключение действия случайных причин, искажающих суть того или иного изучаемого явления (показателя).

Не все в учении Кетле было бесспорным, и имелось немало противников его взглядов, но в целом его идеи оказали большое

влияние на развитие статистики, статистическои методологии исследования.

Пожалуй, можно сказать, что под влиянием идей (и практики) Кетле в XIX в. возникло и успешно развивалось третье направление статистической науки — ***математико-статистическое.***

Развитие этого направления связано с именами таких ученых, как англичане **Фрэнсис Гальтон** (1822–1911), **Фрэнсис Эджворт** (1845–1926), **Карл Пирсон** (1857–1936), **Одни Дж. Юл** (1871–1951), **Артур Боули** (1869–1957), **Вильям Госсет** (1876–1937), **Рональд Фишер** (1890–1968), **Морис Дж. Кендэл** (р. 1907), немец **Вильгельм Лексис** (1837–1914), итальянец **Коррадо Джини** (1884–1965) и многие другие*.

Деятельность большинства из них протекала уже в первой половине XX в., для которого в целом характерно бурное развитие науки разных отраслей.

Работы перечисленных выше ученых обогатили статистическую науку такими новыми методами исследования, как: корреляционно-регрессионный анализ, дисперсионный анализ, анализ временных рядов, построение и анализ теоретических моделей распределения единиц совокупности по тем или иным признакам, оценивание гипотез и др.

Математико-статистическое направление статистической науки развивалось не только на Западе, но и в России. Но прежде чем называть имена русских ученых в этой области, вернемся несколько назад и рассмотрим основные этапы развития статистики как науки **в России.**

Хотя в России отсутствовало четкое деление статистических «школ», но характер научных работ и практической деятельности отдельных представителей статистики позволяет отнести их к сторонникам выделенных направлений статистики.

В России последователями «школы государствоведения» были И.К. Кирилов (1689–1737), В.Н. Татищев (1686–1750), М.В. Ломоносов (1711–1765), К.Ф. Герман (1767–1838), К.И. Арсеньев (1789–1865) и др.

Иван Кириллович Кирилов – яркая личность первой половины XVIII в. Более 20 лет он служил в Сенате и проявлял большой интерес к учетным данным, поступавшим в Сенат. В 1727 г. на материалах I петровской ревизии закончил работу под названием

---

* Подробнее о вкладе этих и других ученых в развитие статистики см.: *Плошко Б.Г., Елисеева И.И.* История статистики. – М.: Финансы и статистика, 1990.

«Цветущее состояние Всероссийского государства, в каковое начал, привел и оставил неизреченными трудами Петр Великий, отец отечествия, император и самодержец Всероссийский и проч.». Это было первое систематизированное статистическое и экономико-географическое описание России. Работа выдержана в классическом стиле описательного направления статистики (государствоведения).

Предметом описания служили города России. Работа содержала сведения не только о расположении городов, но и о их населении, строениях, фабриках и заводах, промысле, торговле, сельском хозяйстве, доходах и расходах, монастырях, церквях и пр. Отдельные данные приводились в виде «генеральных ведомостей и табелей» как сводные по губерниям и стране.

Такого детального и систематизированного описания государства не было прежде в Европе. Особо оригинальным и ценным было использование в этой работе *таблиц*. Кирилова по праву считают первооткрывателем табличного метода в статистике.

К сожалению, при жизни автора работа «Цветущее состояние Всероссийского государства...» не была опубликована. Издана она была лишь в 1831 г. историком М.П. Погодиным с рукописного экземпляра, обнаруженного в одной частной библиотеке. Но и в середине XIX в. этот труд Кирилова оценили как важное научное сочинение, дающее верное описание Петровской России.

Кирилову принадлежит и идея создания первого атласа России, среди его заслуг также частичное воплощение этой идеи в жизнь.

Представителем описательной школы был и русский историк, географ, государственный деятель Петровской эпохи **Василий Никитич Татищев**. На посту руководителя горного дела на Урале в 1720–1722 гг. и позднее, в 1734–1737 гг., он развил бурную деятельность: организовал строительство казенных заводов, дорог, поиск новых месторождений полезных ископаемых, геодезические съемки и составление картографических карт, содействовал открытию начальных и специальных горных школ и т.д. Будучи губернатором Астраханской губернии (1741–1745), В.Н. Татищев написал экономическую работу «Краткие экономические до деревни следующие записки», своеобразную инструкцию помещикам о том, как вести хозяйство.

Наряду с государственной деятельностью В.Н. Татищев занимался научными изысканиями. Проявляя интерес к различным наукам, он понял, что в первую очередь надо решить проблему *сбора* необходимой информации, т.е. проблему *источниковедения*.

Решению этой проблемы он уделял большое внимание до конца жизни.

Так, для сбора информации, необходимой, чтобы составить по заданию Петра I всестороннее экономико-географическое описание России, Татищев разработал специальную анкету (1737 г.) и представил ее на рассмотрение в Академию наук. Анкета содержала 198 вопросов, относящихся к истории, географии, этнографии, экономике и пр.

Не получив ответа от Академии наук, Татищев по собственной инициативе разослал анкету в канцелярии Сибири и Казанской губернии. На основе полученных сведений он разработал программу (план) «описания всей Сибири». Академия наук одобрила программу, и Татищеву поручили подготовить аналогичное описание всей России. К сожалению, работа не была завершена.

В.Н. Татищев понимал, как велико значение источников информации в любом научном исследовании. Он подверг критическому анализу результаты двух ревизий (переписей), проведенных в России, и высказал ряд идей, направленных на получение более качественных сведений. Это прежде всего: составление единого документа для сбора данных по сходным объектам в различных регионах страны, сокращение сроков сбора сведений и подготовка квалифицированных переписчиков. Вопросам о том, как организовать и усовершенствовать учет населения, посвящена его работа «Рассуждения о ревизии поголовной и касаюсчемся до оной» (1747 г.).

Продолжателем дела В.Н. Татищева в области сбора данных для всестороннего экономико-географического описания России стал **Михаил Васильевич Ломоносов**. Возглавив в 1758 г. Географический департамент Петербургской Академии наук, М.В. Ломоносов задумал создать новый «Российский атлас». Для сбора достоверных данных он на основе анкеты Татищева разработал «Академическую анкету», содержащую всего 30 вопросов, ответы на которые давали возможность получить подробное экономико-географическое описание городов и страны в целом.

Анкеты были разосланы в канцелярии всех губерний и провинций. Специальный правительственный указ обязывал администрацию обеспечить своевременное заполнение и представление анкет в Академию наук. Однако сведения были собраны не полностью и обработаны и изданы частично лишь в 1771–1774 гг. после смерти М.В. Ломоносова. Для организации сбора статистической информации Ломоносов сделал немало. Можно сказать, что под его руководством Географический департамент Академии наук превра-

тился в настоящий центр статистико-географического изучения хозяйства России, отдельных ее регионов.

М.В. Ломоносовым написан ряд работ экономического характера. Такова его работа «Слово похвальное императору Петру Великому» (1755 г.), в которой дается оценка I ревизии.

В трактате Ломоносова «О размножении и сохранении Российского народа» был выдвинут ряд условий, соблюдение которых могло бы привести к быстрому росту населения в России. Ломоносов отказался от термина «статистика» в смысле «описательного государствоведения» и впервые ввел термин «экономическая география», выделив ее из общей географии в самостоятельную науку.

Его работы не носят чисто описательный характер. Им присущ элемент анализа, использования числового метода, т.е. они значительно ближе к работам школы «политической арифметики».

Немало занималось описанием русских городов и изучением хозяйств наместничеств **«Вольное экономическое общество»** (ВЭО), созданное в Санкт-Петербурге в 1765 г. Это общество, созданное с целью распространения в государстве полезных сведений о земледелии и промышленности, ставило перед собой задачу найти в условиях крепостного строя пути увеличения производительности труда крестьян и доходов помещиков.

Оно собирало различного рода сведения статистического характера, на основании которых публиковались статистико-географические исследования. Созданное для определенных практических задач ВЭО сыграло большую роль в деле развития краевой географии и статистики.

Говоря о представителях описательной школы в России этого периода, нельзя не упомянуть имя **Карла Федоровича Германа**, первого руководителя Статистического комитета, созданного в 1811 г. при Министерстве полиции, автора таких работ, как «Статистическое описание Ярославской губернии» (1805 г.), «Статистическое исследование относительно Российской Империи, ч. 1. О народонаселении» (1818 г.) и др.

К.Ф. Герман преподавал статистику в учебных заведениях России*, написал первые учебные пособия «Краткое руководство ко всеобщей теории статистики для употребления в училищах Российской Империи» (1808 г.) и «Всеобщая теория статистики для обучающихся сей науке» (1809 г.).

Как представитель и последователь описательной школы К.Ф. Герман считал, что предметом статистики является государ-

* Сначала в педагогическом институте, а затем в Петербургском университете.

ство. Статистика, по его словам, есть «основательное познание государства в какое-либо известное время». Вместе с тем он не ограничивался лишь чистыми описаниями. В его работах уже присутствуют элементы анализа, группировки, динамические сопоставления и пр. Он уделял большое внимание истории статистики, критической оценке достоверности используемых статистических данных, объективности статистики.

Достойным продолжателем дела К.Ф. Германа был его ученик и соратник **Константин Иванович Арсеньев** – историк, географ, статистик. Он считал статистику наукой, призванной обобщать факты и давать им при анализе политическую и экономическую оценку. К.И. Арсеньев преподавал статистику и географию в Петербургском университете, а затем в военных учебных заведениях. Его «Статистические очерки России» (1848 г.) – серьезное экономико-географическое исследование, в котором дано обоснование экономического районирования России.

Еще раньше, в 1818–1819 гг., он написал работу «Начертание статистики Российского государства». Первая ее часть «О состоянии народа» включала разделы «О народонаселении», «О народном богатстве» и «О народном образовании» и содержала расчеты численности *всего* населения на основе данных ревизий (первые три ревизии не учитывали женщин). В работе приведены интересные группировки населения по национальности, по вероисповеданию, выделено городское и сельское население, производительное (земледельцы, мануфактуристы, ремесленники и купцы) и непроизводительное население (духовенство, дворянство, гражданские и военные чины, служители и пр.), для отдельных групп установлены соотношения. Так, соотношение численности непроизводительного населения к производительному составляло 1 : 9; соотношение фабрикантов, ремесленников и купцов к земледельцам – 1 : 20 и т.д. Это была одна из первых классовых группировок населения в России.

Обе работы Арсеньева сыграли большую роль в развитии экономической географии и статистики.

К.И. Арсеньев много сделал в организации и налаживании статистического дела в России. С 1835 по 1852 г. под его руководством создавались губернские статистические комитеты. И хотя К.И. Арсеньева и К.Ф. Германа считают представителями описательного направления, их работы содержат и анализ, что более свойственно второму направлению в развитии статистики – политической арифметике.

В правильной научной постановке статистического дела в России большая заслуга принадлежит **Русскому географическому обществу** (РГО), основанному в 1845 г. и имевшему в своем составе отделение статистики (впоследствии преобразованное в отделение экономической географии).

Статистическое отделение РГО привлекало к своей работе очень многих из тех, кто интересовался своим краем. В результате во второй половине XIX в. появились многочисленные описания отдельных губерний, уездов, городов, проведенные по заданию отделения статистики.

Среди таких работ особенно выделялось изданное в 1852 г. трехтомное «Статистическое описание Киевской губернии» Д.П. Журавского, в котором на богатом конкретном материале впервые в территориальном разрезе дан глубокий социальный анализ хозяйственных процессов, вскрыто имущественное неравенство отдельных слоев населения. По словам Н.Г. Чернышевского, который проявил живой интерес к статистике и экономической географии, «Статистическое описание Киевской губернии» было «одним из самых драгоценных приобретений, сделанных русскою наукою в течение всего настоящего столетия»*.

Говоря о **Дмитрии Петровиче Журавском** (1810–1856) как о крупнейшем русском статистике, нельзя не отметить, что немалая заслуга в его успехах принадлежит РГО, членом которого он был и по заданию которого работал.

Особое место в истории русской статистики занимает выдающийся русский географ, известный и как статистик, Петр Петрович Семенов, именуемый с 1906 г. за свои заслуги в области исследования Тян-Шаня **Семеновым Тян-Шанским** (1827–1914). Он является своего рода отцом русской государственной статистики. Семенов Тян-Шанский – автор многих ценнейших работ в области статистики. С 1873 по 1914 г. Семенов Тян-Шанский был председателем РГО. Одновременно с 1864 по 1875 г. он возглавлял ЦСК (Центральный статистический комитет) при МВД, инициатором создания которого он был, а затем с 1875 по 1897 г. был председателем Статистического совета, т.е. более 33 лет он находился в руководстве правительственной статистики.

П.П. Семенов Тян-Шанский упорядочил русскую статистику и исследование русского хозяйства. Он ввел новые для 70-х годов XIX в. подворные обследования. Это было очень ценно, так как только при подворном обследовании можно было изучить расслоение крестьянства. По инициативе Семенова Тян-Шанского

---

* *Чернышевский Н.Г.* Полн. собр. соч. Т III. – М., 1947. – С. 387.

была проведена Всероссийская перепись населения 1897 г. Ему же принадлежит заслуга обработки ее материалов.

П.П. Семенов Тян-Шанский прекрасно понимал, что без хорошо налаженного статистического дела в стране невозможно глубокое научное экономико-географическое исследование. Поэтому он старался поставить статистику на научные начала, пытался систематизировать и издавать различного рода справочные материалы по фабрично-заводской статистике; при нем начали собирать сведения по статистике урожаев; была проведена первая перепись всех паровых двигателей в России и т.д.

В 1870 г. П.П. Семенов Тян-Шанский провел I статистический съезд в России (единственный до революции). Как глава русской правительственной статистики он участвовал в V–IX международных статистических конгрессах, что способствовало изучению опыта зарубежной статистики и развитию статистики в России.

Вообще следует отметить, что вторая половина XIX – начало XX в. были для России периодом бурного развития статистической науки и практики. И большая заслуга в этом принадлежит представителям так называемой ***академической статистики***, к числу которых относили профессоров университетов, преподававших математику и статистику, авторов учебников по статистике и других научных трудов, посвященных различным проблемам статистической науки.

Это прежде всего представители математической школы Петербургского университета: **Пафнутий Львович Чебышёв** (1821–1894), сформулировавший закон больших чисел; **Андрей Андреевич Марков** (1856–1922) – создатель математической теории, способной описать сложные явления (так называемой схемы цепей Маркова), а также **Александр Михайлович Ляпунов** (1857–1918), обобщивший идеи Чебышёва и Маркова и заложивший теорстические основы в практику выборочного наблюдения в статистике.

Из статистиков к академической школе относится прежде всего **Юрий Эдуардович Янсон** (1835–1893), автор учебника «Теория статистики», придававший большое значение вопросам организации наблюдения и группировкам, а также выявлению причинно-следственных связей, но отрицавший роль закона больших чисел в изучении явлений общественной жизни.

К академической школе относится и **Александр Иванович Чупров** (1842–1908), читавший лекции по статистике в Московском университете, а также его сын **Александр Александрович Чупров** (1874–1926), профессор Петербургского политехнического инсти-

тута, заведующий кафедрой статистики, член-корреспондент Российской Академии наук. А.А. Чупров оставил глубокий след в развитии так называемой стохастической (вероятностной) статистики. Его заслуги в развитии статистики были признаны не только в России, но и за ее пределами. В 1924 г. он был избран почетным членом Лондонского королевского статистического общества.

Единомышленником А.А. Чупрова в вопросах стохастической статистики был профессор Московского университета **Николай Алексеевич Каблуков** (1849–1919), автор учебника «Курс статистики», принимавший активное участие в деятельности земской статистики.

В числе представителей академической статистики нельзя не упомянуть профессора Петербургского университета **Александра Аркадьевича Кауфмана** (1864–1919), автора весьма популярного в свое время учебника «Теория и методы статистики» (1916 г.) переиздававшегося и в советский период. Для А.А. Кауфмана характерно скептическое отношение ко многим разработкам стохастической статистики, в частности к теории корреляции. Саму статистику он рассматривал как методологическую науку.

Надо сказать, что представители академической статистики не были единомышленниками по всем вопросам статистики. У них были расхождения во взглядах, например, на предмет статистики, на использование тех или иных методов в изучении общественных явлений, на роль закона больших чисел и теории вероятностей в статистике. Но все они были преданы статистике, содействовали развитию статистической практики в России, создавали фундаментальные научные труды и учебники по статистике, по которым обучались студенты вузов разных специальностей, что обеспечивало подготовку статистически грамотных специалистов в разных областях.

Многие идеи представителей академической статистики получили дальнейшее развитие в трудах их учеников и последователей и были предметом спора, обсуждения и обобщения на протяжении всего XX в., в частности и в советский период.

1.3. Предмет и метод статистики

Статистика, как и любая другая наука, имеет свой предмет и метод исследования.

Как уже неоднократно подчеркивалось, статистика изучает, как правило, *массовые* явления, т.е. такие явления, которые состоят из *множества* отдельных элементов или фактов.

Каковы отличительные особенности массового явления?

1. Каждый элемент такого множества обладает как индивидуальными (отличающимися) признаками, так и общими (сходными). Например, изучая результат промышленного производства за тот или иной период, мы рассматриваем множество промышленных предприятий, каждое из которых имеет свои индивидуальные признаки, такие, как численность работников, стоимость основных фондов, ассортимент выпускаемой продукции, размер прибыли и т.п. В то же время все эти предприятия как единицы множества обладают общим признаком – все они являются промышленными (а не сельскохозяйственными или строительными).

Другими словами, все единицы определенного множества, изучаемого статистикой, как правило, о д н о к а ч е с т в е н н ы по сути.

2. Характеристики (показатели) одного из элементов массового явления не могут быть получены на основе характеристик других единиц (элементов), поскольку индивидуальные характеристики у разных элементов множества полностью или частично н е з а в и с и м ы.

Изучаемые статистикой массовые явления в виде множества однокачественных единиц с отличающимися индивидуальными признаками называют ***статистическими совокупностями.***

Это может быть, например, совокупность населения или отдельных его контингентов (трудоспособное население, пенсионеры, городское или сельское население и т.п.), совокупность промышленных предприятий (строительных, сельскохозяйственных, торговых и пр.), совокупность работников (на отдельном предприятии, в отрасли или секторе экономики), совокупность банков и т.д.

Другими словами, массовые явления находят свое оформление, выражение в статистических совокупностях. Причем каждая такая статистическая совокупность не абстрактна; она привязана к определенному месту и времени.

Исходя из этого можно сказать, что ***предметом*** статистики являются различные *статистические совокупности*, исследование которых связано с количественной характеристикой и выявлением присущих им закономерностей в конкретных условиях места и времени.

Каждая наука оперирует определенными понятиями, категориями.

Статистическая совокупность – одно из главных понятий статистической науки. С этим понятием непосредственно связаны и другие, такие, как *единица совокупности*, *признаки единиц совокупности*, *вариация признаков*, *статистическая закономерность* и т.д.

Элементы, множество которых образует изучаемую статистическую совокупность, называют ***единицами совокупности***. Так, при изучении совокупности фермерских хозяйств (в целом по стране или в отдельном регионе) единицей совокупности будет отдельное фермерское хозяйство; при переписи населения – отдельный человек; если изучается совокупность семей или домохозяйств, то каждая семья или домохозяйство является единицей совокупности (единицей наблюдения) и т.п.

Каждая единица совокупности может быть охарактеризована разного рода качественными (атрибутивными) и количественными ***признаками***. Например, каждое фермерское хозяйство можно охарактеризовать такими признаками, как площадь используемого земельного участка, поголовье скота (птицы), производство продукции (в натуральном и стоимостном выражении), рентабельность продукции и т.п., а каждого человека при переписи населения – такими признаками, как пол, возраст, национальность, семейное положение, место работы (или источник доходов), размер заработной платы (или дохода) и т.п.

Если определенный признак имеет разные значения у отдельных единиц совокупности, то говорят, что он варьирует или имеет некоторую ***вариацию***. Такие признаки, варьирующие от единицы к единице, составляют отличительную черту статистической совокупности, делающую ее предметом изучения статистики, и называются *статистическими*.

Любое статистическое исследование, т.е. исследование статистической совокупности, начинается с изучения отдельных единиц совокупности, с регистрации у них тех или иных признаков, знание которых необходимо для достижения цели исследования, для выявления различных закономерностей.

Статистика, как правило, оперирует числовыми данными, которые обусловлены влиянием множества различных факторов, причин, одни из которых являются главными, существенными, а другие – случайными.

Абстрагироваться от случайного и выявить типичное, закономерное – основная задача статистики, и эту задачу можно решить только на основе массовых данных. По единичному факту нельзя судить о закономерности, поскольку единичное явление несет на себе влияние случайного фактора. Только исследуя массу явлений, путем обобщения можно выявить, измерить и познать те или иные закономерности, например закономерность в распределении единиц совокупности по какому-то признаку (доля мальчиков и девочек в общей численности родившихся за год), или оп-

ределить средний уровень какого-либо количественного показателя (среднюю заработную плату в той или иной отрасли, среднюю производительность труда, среднюю рентабельность продукции в отдельных отраслях и т.п.).

Закономерность, выявленная на основе массового наблюдения, т.е. проявившаяся в большой массе данных через преодоление свойственной ее единичным элементам случайности, называется ***статистической закономерностью.***

В одних случаях существование определенной закономерности в изучаемых явлениях можно теоретически предположить, рассуждая логически и опираясь на знание сущности рассматриваемых явлений. Например, очевидно, что при внесении удобрений в почву урожайность сельскохозяйственных культур возрастает, при повышении цен на определенный товар спрос на него уменьшается и т.д. В этих случаях задача статистики – не только подтвердить существование предполагаемой зависимости, но и измерить ее, т.е. определить меру изменения одного показателя в зависимости от изменения другого.

В других случаях статистическую закономерность легко установить эмпирически при обработке массовых данных. Например, таким путем было выявлено, что при увеличении дохода семьи в ее бюджете снижается доля расходов на питание.

Статистические закономерности обнаруживаются при массовом наблюдении благодаря действию так называемого *закона больших чисел*, который выражает диалектику случайного и необходимого. Сущность закона больших чисел заключается в том, что по мере увеличения числа наблюдений влияние случайных факторов (причин), определяющих значение признака у единиц совокупности или соотношение между численностями единиц с определенными признаками, взаимопогашается в сводных (общих) характеристиках совокупности (например, в средних величинах) и на поверхность выступает действие основных факторов, которые и определяют закономерность.

Таким образом, ***массовое наблюдение*** – основа статистики и одна из составляющих ее метода.

Однако массовым наблюдением не заканчивается статистическое исследование. Результаты наблюдения подвергают обработке, сводке, что позволяет выделить во всей совокупности различные типы, группы единиц и затем для всей совокупности и отдельных ее частей рассчитать обобщающие показатели (характеристики).

Массовое наблюдение, *группировка* и *сводка его результатов*, *вычисление* и *анализ обобщающих показателей* – все это вместе составляет специфический ***метод*** статистики.

К какой бы области ни относился предмет исследования статистики (население, промышленность, торговля и т.п.), метод ее везде одинаков, т.е. везде используются массовое наблюдение, группировки и обобщающие показатели, в которых, благодаря действию закона больших чисел, взаимопогашается влияние случайных причин и выявляется типичное и закономерное. Иначе говоря, метод статистики обусловлен спецификой ее предмета.

1.4. Теория статистики как научная (учебная) дисциплина

В процессе развития статистики как науки возникли следующие самостоятельные научные дисциплины:

- *отраслевые статистики*, в которых освещаются сущность и методология расчета показателей, используемых при изучении соответствующей отрасли (статистика промышленности, сельского хозяйства, транспорта и т.п.);
- *экономическая статистика*, в которой раскрывается сущность и методология исчисления показателей, используемых при статистическом изучении экономики в целом;
- *общая теория статистики* (или *теория статистики*), в которой освещается статистическая методология, статистический метод, общий для всех отраслевых статистик.

Изучение статистических совокупностей связано с решением таких основных задач, как:

- получение итоговых данных по совокупности;
- определение структуры совокупности и соотношения отдельных ее частей;
- изучение особенности распределения единиц совокупности по отдельным признакам;
- определение средней величины того или иного количественного показателя и его вариации;
- выявление взаимосвязи между отдельными показателями (признаками);
- изучение динамики отдельных показателей (как единичных, так и агрегированных);
- оценка выборочных данных и пр.

Для решения всех этих и некоторых других задач в статистике разработаны различные приемы, методы, показатели. И именно в

курсе «Теория статистики» рассматриваются все эти приемы, методы и показатели, т.е. методы сбора, обработки и анализа массовых статистических данных.

В этом плане можно сказать, что «Теория статистики» – методологическая наука, т.е. наука о методах, применяемых для количественных характеристик и выявления закономерностей в изучаемых явлениях, где выводы строятся на основе массового наблюдения, где имеет место вариация признака у единичных элементов совокупности, где общие закономерности могут проявиться только через взаимопогашение случайностей у отдельных единиц при расчете обобщающих показателей.

В статистике, имеющей дело с количественными показателями, естественно применение математики. Так как выводы статистики основаны на большом числе единичных случайных явлений (событий), она неизбежно соприкасается с теорией вероятностей, которая изучает законы случайных величин.

В результате взаимодействия теории вероятностей с другими приемами математики возникла самостоятельная научная дисциплина – *математическая статистика*, в которой на основе математической логики и строгих доказательств рассматриваются различные вероятностные модели распределения случайных величин, их взаимосвязи, дается оценка выборочных данных и т.п.

Естественно, что курс «Теория статистики» не может не включать отдельные приемы математической статистики, помогающие раскрыть статистические закономерности. Однако эти приемы рассматриваются в данном курсе в основном в прикладном плане (без строго математического доказательства).

В статистике, как и в любой другой науке, математика служит средством, инструментом исследования, о чем в свое время сказал известный русский статистик А.А. Чупров. По его словам, математика «предоставляет в распоряжение исследователя богатый набор усовершенствованных рабочих инструментов»*, с помощью которых он может более глубоко проникнуть в сущность изучаемых явлений и обнаружить присущие им закономерности.

---

* *Чупров А.А.* Основные проблемы теории корреляции. — М., 1960. — С. 137.

Глава 2
ОБЩИЕ СВЕДЕНИЯ О СТАТИСТИЧЕСКОМ НАБЛЮДЕНИИ

2.1. Статистическое наблюдение как первый этап статистического исследования. Организационные формы статистического наблюдения

Этапы статистического наблюдения

В любом статистическом исследовании, статистической работе можно выделить несколько этапов.

Статистическое изучение тех или иных явлений предполагает как обязательное условие наличие информации, сведений об этих явлениях. Поэтому первый этап, начало статистического исследования сводится к **сбору необходимой информации**.

Научно организованный сбор сведений, заключающийся в регистрации тех или иных фактов, признаков, относящихся к каждой единице изучаемой совокупности, именуют ***статистическим наблюдением***.

В результате статистического наблюдения образуется масса первичной информации (сведений) о каждой единице совокупности. Чтобы получить характеристику всей исследуемой совокупности в целом, первичные данные должны быть подвергнуты обработке, обобщению. Обработка собранных первичных данных, включающая их группировку, обобщение и оформление в таблицах, составляет второй этап статистического исследования, который именуют **сводкой**.

И наконец, на основе итоговых данных сводки осуществляется **научный анализ исследуемых явлений**: рассчитываются различные обобщающие показатели в виде средних и относительных величин, выявляются определенные закономерности в распределениях, динамике показателей и т.п. Это третий этап статистического исследования.

Таким образом, любое законченное статистическое исследование проходит три этапа, между которыми, разумеется, могут быть разрывы во времени.

Все этапы работы, несомненно, важны. Но поскольку статистическое наблюдение является начальным, то оно во многом определяет успех всей работы. От того, насколько полными и качественными окажутся собранные первичные данные, зависят в значительной степени и конечные результаты работы, и выводы исследователей. Поэтому статистическому наблюдению всегда уделялось и уделяется большое внимание в статистических исследованиях.

В статистической практике используются разные формы, виды и способы наблюдения.

Формы организации статистического наблюдения

По своей организации статистическое наблюдение может быть осуществлено по-разному. Различают следующие формы его организации:

- статистическая отчетность;
- специально организованные статистические обследования (наблюдения);
- регистры.

1. ***Статистическая отчетность*** – это особая форма организации сбора данных государственной статистикой о деятельности хозяйствующих субъектов через специально заполняемые последними документы-бланки, именуемые формами статистической отчетности. *Форма статистической отчетности* – это специальный документ, бланк, содержащий перечень определенных показателей, сведений, характеризующих ту или иную хозяйственную единицу и результаты ее деятельности, представляемый в государственные статистические органы для дальнейшего обобщения. (В свою очередь статистическая отчетность заполняется каждым предприятием, организацией на основе данных оперативного или бухгалтерского учета, ведущегося на этих предприятиях, т.е. является обобщением документов первичного учета.)

Отчетность (перечень и содержание ее форм) утверждается органами государственной статистики и является обязательной для установленного круга предприятий и организаций. Каждая форма отчетности имеет свой шифр и название.

Перечень и содержание форм отчетности не остаются неизменными, они меняются со временем с учетом требований меняющейся экономики и международной практики учета и статистики.

В настоящее время для средних и крупных предприятий всех отраслей экономики разработана унифицированная отчетность, включающая следующие формы: № П-1 «Сведения о производ-

стве и отгрузке товаров и услуг»; № П-2 «Сведения об инвестициях»; № П-3 «Сведения о финансовом состоянии предприятия»; № П-4 «Сведения о численности, заработной плате и движении работников».

Малые предприятия с 1999 г. отчитываются ежеквартально лишь по форме ПМ «Сведения об основных показателях деятельности малого предприятия».

В соответствии со сроками представления отчетность бывает *суточная* (ежедневная), *недельная*, *месячная*, *квартальная*, *полугодовая*, *годовая*. Все упомянутые выше виды отчетности, кроме годовой, объединяют одним названием – *текущая отчетность*. Как правило, чем больше период, за который представляется отчетность, тем больше последняя содержит показателей.

Каждая форма отчетности должна представляться в установленные для нее сроки.

По способу передачи (представления) сведений отчетность делится на *электронную*, *телеграфную*, *телетайпную*, *почтовую*.

2. Круг явлений общественной жизни настолько велик, что полный охват их отчетностью не только затруднителен, но и невозможен. Например, при изучении бюджетов населения невозможно заставить каждую семью ежемесячно отчитываться в своих доходах и расходах; нет необходимости (и возможности) через отчетность изучать обеспеченность населения жильем, определять численность и состав населения и т.д. Трудно использовать отчетность при изучении многих явлений в условиях рыночной экономики.

Во всех случаях, когда необходимо получить сведения, по которым отсутствует отчетность, когда требуется уточнить или дополнить данные той или иной отчетности либо провести разовое детальное, всестороннее обследование каких-либо объектов (например, состояния малых рек, экологической ситуации в отдельных промышленных регионах, состояния сельских библиотек или клубов и т.п.), – применяют ***специально организованные статистические наблюдения***, проводимые в виде переписей или специальных обследований (выборочных или сплошных).

Наиболее типичным примером специально организованного наблюдения являются различного рода *переписи*. Это прежде всего переписи населения. Последняя перепись населения в России проведена в 2002 г. В практике советской и российской статистики, кроме переписей населения, известны переписи промышленности, материальных ресурсов, неустановленного оборудования, библиотек, школ и др. Любая перепись дает представление о чис-

ленности, размещении по территории, составе и состоянии объекта наблюдения на определенный момент времени.

Кроме переписей к специально организованным наблюдениям относятся проводимые в России ежегодные выборочные *бюджетные обследования населения*, периодические выборочные *обследования занятости населения*. Примером специально организованного статистического наблюдения может служить проведенное в январе–апреле 2001 г. сплошное единовременное *обследование малых предприятий* (по результатам работы за 2000 г.).

Специально организованное статистическое наблюдение используется как органами статистики, так и отдельными учреждениями, предприятиями, организациями.

3. Наблюдение через регистры – сравнительно новая форма организации статистического наблюдения, появление которой стало возможным только при соответствующем уровне развития вычислительной электронной техники, компьютеризации. Регистры могут разрабатываться для разных объектов статистического наблюдения (*регистр населения*, *регистр промышленных предприятий*, *регистр подрядных строительных организаций* и пр.).

Регистр – это поименованный и постоянно уточняемый перечень тех или иных единиц наблюдения, созданный для непрерывного длительного статистического наблюдения за определенной совокупностью. В регистре содержится информация о каждой единице совокупности. При этом одни сведения остаются неизменными в течение всего периода наблюдения (например, в регистре предприятий это название и адрес предприятия, его организационно-правовая форма, вид экономической деятельности и т.п.), другие – обновляются по мере их изменения.

В России завершена работа по созданию *Единого государственного регистра предприятий и организаций всех форм собственности* (ЕГРПО). В нем кроме адресной части содержатся такие сведения о каждом предприятии, как: среднесписочная численность работников, остаточная стоимость основных средств, уставный фонд, балансовая прибыль и др. Поскольку регистр ведется по отдельным территориям, то у региональной статистической службы есть возможность расширить круг экономических показателей с учетом своих потребностей.

ЕГРПО является надежной информационной базой для ведения системы национальных счетов в статистической практике России. Также ЕГРПО является хорошей базой (как генеральная совокупность) для проведения выборочных обследований предприятий с различной целью.

Важно отметить, что все три организационные формы статистического наблюдения не противостоят друг другу, а дополняют друг друга, позволяют более глубоко, всесторонне изучать отдельные явления и процессы общественной жизни.

2.2. Виды и способы статистического наблюдения

Виды статистического наблюдения

Выше отмечалось, что статистическое исследование сводится к сбору информации о той или иной статистической совокупности, к регистрации тех или иных фактов, признаков, относящихся к каждой единице исследуемой совокупности.

Об одних явлениях факты регистрируются по мере их возникновения (например, регистрация рождения и смерти каждого человека, заключения и расторжения браков и т.п.).

В других случаях факты регистрируются независимо от того, когда они возникли. Например, при проведении переписи населения регистрируются такие признаки, факты, как семейное положение, образование, род занятий и прочее независимо от того, когда они были получены или оформлены.

Иными словами, одни факты учитываются (регистрируются) постоянно, непрерывно, по мере их возникновения; другие – с разрывом во времени между моментом возникновения факта и его регистрацией.

В соответствии с этим по *времени регистрации фактов* в статистике различают ***текущее (непрерывное) наблюдение*** и ***прерывное***. Последнее, в свою очередь, может быть ***единовременным***, если наблюдение происходит от случая к случаю, по мере необходимости, и ***периодическим***, если оно повторяется через определенные равные интервалы времени (год, 5 лет, 10 лет и т.д.).

Прерывное наблюдение, как единовременное, так и периодическое, играет роль моментальной фотографии, отражающей состояние изучаемого явления на определенный момент времени, определенную дату.

Различного рода переписи, в частности переписи населения, представляют собой пример прерывного наблюдения.

В дореволюционной России в 1897 г. была проведена единственная Всеобщая перепись населения. В бывшем СССР переписи населения проводились в 1926, 1939, 1959, 1970, 1979 и 1989 гг. В постсоветской России – в 2002 г. (до этого, в 1994 г., была проведена выборочная микроперепись населения). Судя по

датам проведения, переписи в России (и СССР) скорее можно отнести к единовременным наблюдениям, чем к периодическим.

Примерами текущего наблюдения являются: наблюдение за деятельностью предприятий, включенных в Реестр хозяйствующих субъектов, имеющих на рынке определенного товара долю более 35%; статистическое наблюдение за производством продукции в отдельных отраслях и секторах экономики; наблюдение за естественным движением населения (рождаемостью, смертностью) и т.д. Можно сказать, что практически все публикуемые органами статистики итоговые показатели, достигнутые за определенный период, получены на основе текущего наблюдения.

Статистическое наблюдение может охватывать все единицы изучаемой совокупности, но в отдельных случаях представление о всей совокупности можно получить, исследуя лишь ее часть.

По *охвату единиц наблюдаемого объекта* в статистике различают ***сплошное*** и ***несплошное наблюдение.***

При сплошном наблюдении ставится задача получить сведения о всех единицах изучаемой совокупности. Примером сплошного наблюдения может быть перепись населения, при проведении которой ставится задача получить сведения по определенному перечню вопросов о каждом человеке (или о каждой семье, домохозяйстве), а также любая другая перепись (перепись промышленности, строительных подрядных организаций и др.), имеющая своей целью полный охват (учет) всех единиц исследуемого объекта. В виде сплошного наблюдения организована регистрация родившихся, умерших и др.

Сплошное наблюдение при всей своей привлекательности и кажущейся большей достоверности не лишено недостатков. В частности, оно требует бóльших затрат (стоимостных и трудовых) как при сборе информации, так и при последующей ее обработке. Кроме того, в ряде случаев просто невозможно осуществление сплошного наблюдения (например, изучение флоры или фауны океана, изучение бюджетов населения и др.). Поэтому в статистической практике наряду со сплошным широко используется и несплошное наблюдение.

К несплошному наблюдению прибегают в тех случаях, когда физически невозможно, трудно или нецелесообразно осуществить сплошное наблюдение, в частности, когда наблюдение влечет за собой порчу или уничтожение наблюдаемой единицы (например, при исследовании качества пищевых продуктов и иных товаров, при проверке длительности срока службы тех или

иных машин, деталей и пр.), а также когда имеется ограничение во времени или средствах, не позволяющее осуществить сплошное наблюдение большого числа единиц изучаемой совокупности, и в других случаях.

Основным преимуществом несплошного наблюдения является то, что оно требует меньших затрат времени на сбор и обработку информации и, как следствие этого, сокращает стоимостные затраты, связанные с исследованием изучаемой совокупности.

Несплошное наблюдение может быть осуществлено по-разному. Различают следующие его виды:

1) наблюдение основного массива;
2) анкетное;
3) выборочное;
4) монографическое.

Наблюдение основного массива предполагает исключение из состава совокупности малозначимых единиц и исследование основной ее части. Примером такого наблюдения может служить ведение Реестра хозяйствующих субъектов, сформированного в 1996 г. во исполнение Закона РСФСР от 22 марта 1991 г. «О конкуренции и ограничении монополистической деятельности на товарных рынках». Статистическое наблюдение за деятельностью предприятий, включенных в этот Реестр, осуществляет Госкомстат РФ. Изучение работы городских рынков также построено по принципу исследования (наблюдения) основного массива, т.е. по определенному кругу городов, а не по всем.

При использовании наблюдения основного массива исходят из соображения, что исключение определенной части «малозначимых» единиц не отразится существенно на результатах наблюдения, в то время как включение этих единиц значительно увеличит объем работы и, соответственно, затраты.

Из сказанного очевидно, что применение наблюдения основного массива возможно лишь в тех случаях, когда известен состав всей совокупности и можно заранее решать, какие единицы малозначимы, а какие нет.

Анкетное наблюдение проводится следующим образом. Организации (или учреждения), поставившие перед собой задачу изучить тот или иной вопрос, составляют и рассылают (или раздают) определенному кругу лиц (респондентам) особые *анкеты* с вопросами. При этом заполнение и возврат анкет является делом добровольным. Как правило, число заполненных и возвращенных анкет бывает значительно меньше розданных. Таким образом, сама организация этого наблюдения, предусматрива-

ющая неполный возврат разосланных (или розданных) анкет, дает основание относить этот вид наблюдения (анкетный) к несплошному.

Выборочное наблюдение – это такой вид несплошного наблюдения, при котором из всей изучаемой совокупности случайно, наудачу (путем жеребьевки или другим методом) отбирается определенное число единиц (выборочная совокупность), для них регистрируются интересующие исследователя признаки, на основании которых исчисляются искомые выборочные показатели (средние величины, относительные и пр.), распространяемые затем на исходную генеральную совокупность.

Выборочный метод наиболее научно обоснован. При использовании этого метода несплошного наблюдения всегда можно оценить возможные расхождения между показателями выборочного и сплошного наблюдения, т.е. ошибки выборки (об этом подробнее будет сказано в главе 6).

Наконец, ***монографическое наблюдение*** представляет собой детальное, тщательное изучение (описание) какой-то одной единицы. Это может быть один рабочий, одна бригада, одно предприятие, один район и т.д.

Иногда эта единица рассматривается как типичная, и ее детальное изучение дает более широкое представление о единицах совокупности; иногда она представляет собой что-то новое, зарождающееся и изучается в целях распространения передового, прогрессивного опыта. Детальное изучение отдельной единицы помогает обнаружить и отрицательные моменты, нуждающиеся в устранении.

Монографическое наблюдение, как правило, – область деятельности отдельных научных учреждений. Монографическое наблюдение не противостоит массовому наблюдению в статистике, а дополняет его, углубляя, когда это необходимо, познание отдельных единиц.

Способы статистического наблюдения

Регистрация необходимых сведений при статистическом наблюдении может проводиться на основе разных источников.

В одних случаях необходимые сведения регистраторы получают путем непосредственного осмотра, измерения, взвешивания и т.п. (например, личный осмотр и подсчет поголовья скота в хозяйстве, непосредственное взвешивание остатков товаров при учете в магазинах, на складах и т.п.). Это так называемое ***непосредственное наблюдение***.

В других случаях регистрация сведений проводится только на основе данных, зафиксированных в документе. Например, рождение ребенка регистрируется в ЗАГСе только на основе представления справки из роддома, где зафиксировано рождение ребенка. Аналогично смерть человека фиксируется в ЗАГСе только на основании справки врача, установившего факт смерти. Регистрация данных о расходе на предприятии сырья или производстве продукции за определенный период также проводится на основе соответствующих первичных документов. Такой способ регистрации сведений называют ***документальным***. Он лежит в основе заполнения отчетности.

В третьих случаях регистрация сведений (фактов) проводится на основе ***опроса***. Этот способ применяется при переписях населения, где все сведения о каждом человеке (пол, возраст, национальность, образование и т.д.) записываются со слов опрашиваемого, т.е. путем опроса.

Таким образом, непосредственное наблюдение, документальное и опрос – это важнейшие (основные) способы статистического наблюдения по *источникам информации*.

Говоря о способах статистического наблюдения, следует отметить, что последние различаются не только по источникам собираемых сведений, но и по *организации сбора*. Особенно это касается сбора сведений путем опроса. Так, опрос может быть осуществлен следующими способами:

- экспедиционным;
- способом саморегистрации;
- корреспондентским;
- явочным.

При ***экспедиционном способе*** специально подготовленные работники (регистраторы) отправляются к лицам (единицам), о которых должны быть получены сведения, и на месте опрашивают их устно, регистрируя ответы в специальных бланках. Например, при переписи населения счетчики-регистраторы являются по месту жительства отдельных лиц, где их и опрашивают (переписывают).

При ***саморегистрации*** переписные листы (бланки) заполняют сами респонденты (опрашиваемые), а счетчики-регистраторы раздают респондентам эти бланки, объясняют правила их заполнения и собирают заполненные опросные листы.

При ***корреспондентском способе*** сведения по заранее определенному кругу показателей поступают в органы, ведущие наблюдение, от специального штата добровольных корреспондентов, работающих на местах, откуда запрашиваются данные.

Наконец, при ***явочном способе*** наблюдения предполагается, что сведения по определенному кругу показателей должны сообщаться в органы, ведущие наблюдение за этими явлениями, в явочном порядке. Например, сведения о родившихся и умерших, заключенных браках и разводах собираются явочным способом.

2.3. Программно-методологические и организационные вопросы статистического наблюдения

Проведение любого статистического наблюдения и в особенности специально организованного – дело весьма непростое, требующее большой подготовительной работы, которая должна обеспечить полноту получения сведений, их достоверность, единообразие, сравнимость. Поэтому многие из вопросов, относящихся к наблюдению, решаются задолго до проведения самого наблюдения. Эти вопросы можно разделить на так называемые программно-методологические и организационные.

Программно-методологические вопросы включают в себя:

- определение цели, объекта и единицы наблюдения;
- разработку программы наблюдения и статистического формуляра, содержащего ее.

При планировании любого статистического наблюдения прежде всего необходимо точно сформулировать его ***цель***, так как только конкретная цель, конкретные задачи исследования определят те сведения, которые должны быть получены в процессе наблюдения.

Так, цель Всесоюзной переписи населения 1989 г., как и предыдущих (1939, 1959, 1970, 1979 гг.), состояла в том, чтобы установить численность наличного и постоянного населения по отдельным республикам, населенным пунктам, городам, районам и т.д., определить состав населения по полу, возрасту, семейному положению, национальности, языку, уровню образования и т.д. Цель Всероссийской переписи населения 2002 г. несколько шире. Кроме перечисленных выше задач в этой переписи ставилась задача получения сведений и о жилищных условиях населения, и о занятости (за неделю до переписи), и о миграции (указание места жительства в 1989 г., а также места рождения).

Целью проведенного в России в начале 2001 г. статистического обследования малых предприятий по итогам работы за 2000 г. было получение полного перечня малых предприятий и сведений о производимых ими товарах и услугах.

Точное определение цели каждого конкретного статистического наблюдения позволяет наметить не только те признаки, которыми должна быть охарактеризована каждая исследуемая единица для решения поставленной задачи, но и форму организации данного наблюдения, его вид и способ проведения.

Параллельно с установлением цели определяется ***объект наблюдения***, т.е. совокупность единиц, сведения о которых должны быть получены. *Определить объект наблюдения – значит точно установить границы изучаемой совокупности*, т.е. решить, что должно быть обследовано или кто должен быть обследован в процессе наблюдения. Так, в переписи населения 1989 г. таким объектом наблюдения было наличное и постоянное население. При переписи 2002 г. объектом наблюдения было постоянное население. Точное определение объекта наблюдения во многом гарантирует полноту и достоверность получаемых при наблюдении сведений.

При определении объекта наблюдения должны быть строго проведены его границы во времени, пространстве и материальной сущности. Так, если взять пример из области наблюдения в промышленности, то при переписи промышленности важно определить, к какой территории должны относиться сведения (ко всей ли стране, к республике, области, городу и т.п.), за какой период или на какой момент должны быть получены сведения (за год, за два или иной период) и, наконец, какой вид деятельности должен быть отнесен к промышленности. При обследовании малых предприятий надо четко определить, какой вид деятельности относится к малому предпринимательству.

Объект наблюдения – это всегда определенная совокупность, состоящая из отдельных элементов, единиц. Характеристика отдельных единиц (частей) объекта и позволяет изучить объект в целом. Поэтому определение объекта наблюдения непосредственно связано с определением единиц наблюдения, которые подлежат статистической характеристике.

Единицей наблюдения в статистике называют ту единицу, тот элемент объекта наблюдения, который характеризуется рядом признаков и относительно которого ведется регистрация этих признаков. Так, при переписи населения 1989 г., как и в предыдущие годы, единицей наблюдения являлся человек. У каждого человека регистрировались такие признаки, как пол, возраст, национальность, образование и т.п. При переписи 2002 г. единицей наблюдения была установлена учетная единица – домохозяйства (домашние хозяйства), т.е., хотя переписывался каждый

человек, регистрация отдельных лиц велась по домохозяйствам, и часть вопросов программы наблюдения (в частности, касающихся жилищных условий) относилась в целом к домохозяйствам.

Правильное определение единицы наблюдения (например, малого предприятия) – залог полноты учета объекта наблюдения.

Следует отметить, что иногда сведения о той или иной единице наблюдения могут быть получены не от нее самой, а от какой-то организационной ячейки. Например, данные о производительности труда рабочих, их заработной плате можно получить не от самого рабочего, а от предприятия, где эти сведения имеются.

Тот орган или ту ячейку, к которой обращаются за получением сведений о каждой единице наблюдения, в статистической литературе именуют *отчетной* либо *учетной единицей.*

После определения единицы наблюдения разрабатывается программа наблюдения. Под ***программой статистического наблюдения*** понимается перечень тех признаков, которыми каждая единица наблюдения должна быть охарактеризована. Другими словами, это перечень вопросов, на которые в процессе наблюдения должны быть получены ответы.

Признаки, которыми отдельные единицы совокупности могут отличаться одна от другой, носят разный характер. Они могут быть *количественными* (возраст, стаж работы, рост, вес и т.п.), и тогда отдельные единицы отличаются друг от друга по величине данного признака. Признаки могут быть *качественными* (пол, семейное положение, занятия и т.п.), и тогда отдельные единицы наблюдения отличаются друг от друга наличием или отсутствием того или иного качества.

При статистическом наблюдении у отдельных единиц регистрируется наличие или отсутствие тех или иных качественных признаков и определяется величина количественного признака. Но признаков, которыми можно охарактеризовать каждую единицу наблюдения, может быть множество.

Составить программу статистического наблюдения – значит выбрать те признаки, которые помогут решить намеченную наблюдением цель, т.е. программа должна определяться целью наблюдения.

Вопросы программы и ответы на них фиксируются в особых статистических формулярах, которые могут именоваться по-разному: переписной лист, бланк, форма и др.

Программа переписи населения 1989 г. содержала такие вопросы, как пол, возраст, национальность, образование, источник средств существования, и ряд других. Всего было 16 вопросов,

на 11 из них ответы получались в виде сплошного наблюдения (переписи), а на 5 – выборочно, от 25% постоянного населения.

Программа Всероссийской переписи населения 2002 г. также состояла из программы сплошного наблюдения и программы выборочного наблюдения.

По программе сплошного наблюдения было опрошено все постоянно проживающее население России. Эта программа, содержавшаяся в переписных листах формы К и формы Д, включала такие вопросы, как пол, возраст (дата рождения), семейное положение, место рождения, гражданство, национальная принадлежность, образование, владение языками, источники средств существования, занятость.

По программе выборочного наблюдения было опрошено 25% постоянного населения. В программу выборочного наблюдения, содержащуюся в переписном листе формы Д (длинной), кроме перечисленных выше вопросов сплошного опроса были включены дополнительные вопросы:

а) о занятости (в какой отрасли экономики занят респондент, основной вид производимой предприятием продукции или услуги, место нахождения работы, характер выполняемой работы или занятия; для не имеющих работу требовалось указать, занимался ли опрашиваемый ее поисками в течение последнего месяца);

б) о том, проживает ли респондент в данном населенном пункте с рождения, и если нет, то с какого года проживает по указанному адресу и где проживал во время предыдущей переписи (в январе 1989 г.);

в) о числе рожденных детей (вопрос только для женщин старше 15 лет).

Третья форма бланка переписного листа (форма П) предназначалась для сбора сведений о жилищных условиях жителей Российской Федерации. Она содержала вопросы о типе жилого помещения, времени его постройки и материале наружных стен жилого помещения, в котором проживал респондент, о размере общей и жилой занимаемой площади индивидуального дома или квартиры и числе жилых комнат, о видах благоустройства и др.

Как видно из вышесказанного, программа Всероссийской переписи населения 2002 г. была очень широкой.

Кроме постоянно проживающего населения переписью были охвачены и лица, временно находящиеся на территории России, но постоянно проживающие за рубежом. Они были опрошены по отдельной сокращенной программе (форма В), которая содержала, в частности, вопрос о цели приезда в Россию.

Составление программы наблюдения – важный и ответственный момент в подготовке самого наблюдения. При ее составлении очень важно обращать внимание на формулировку вопросов. Необходимо добиваться того, чтобы вопросы допускали единое толкование для всех, кто на них отвечает. Для этого к вопросам обычно дается подсказ возможных ответов. Например, вопрос «Родственное отношение к тому, кто записан первым» (в бланке) сопровождается подсказом: жена, муж, сын, дочь, мать, отец, сестра, племянник, зять, свекровь, теща и т.п. Вопрос об источниках средств существования сопровождается подсказом: доходы от трудовой деятельности, личное подсобное хозяйство, на иждивении, сбережения, пенсия, стипендия, доход от сдачи в наем или аренду имущества, иной источник.

Толкование отдельных вопросов и разъяснение того, как на них следует отвечать, обычно даются в инструкции, составляемой для каждого статистического наблюдения.

Статистические формуляры, содержащие программу и результаты регистрации, встречаются двух видов: *индивидуальные* (карточные) и *списочные*. В первом случае формуляр (индивидуальный) заводится на каждую единицу наблюдения отдельно, т.е. в каждом формуляре содержатся сведения лишь об одной единице наблюдения. Во втором случае один формуляр (список) составляется на несколько единиц.

Статистический формуляр должен быть удобен для чтения и заполнения, для шифровки и обработки данных.

В связи с широким внедрением в статистической практике компьютеров в последние годы появились новые носители информации: магнитные ленты, диски и пр.

Организационные вопросы статистического наблюдения включают в себя решение таких важных моментов, как определение:

- субъекта наблюдения;
- места и времени наблюдения;
- организационной формы, вида и способа наблюдения.

Определение ***субъекта наблюдения*** сводится к решению вопроса о том, к т о будет осуществлять статистическое наблюдение. В одних случаях это могут быть органы статистики со своими кадровыми работниками статистических аппаратов. В других случаях (если речь идет о наблюдении в больши́х масштабах, как при переписи населения) для статистического наблюдения наряду со специалистами-статистиками могут привлекаться широкие круги общественности. Наконец, к наблюдению может быть привлечено само население, которому счетчики раздают опрос-

ные листы с просьбой заполнить их самостоятельно, а затем собирают их уже заполненными (это так называемая саморегистрация).

При планировании статистического наблюдения необходимо решить вопрос о ***времени наблюдения***, т.е. определить, к о г д а оно будет проведено. Например, перепись населения следует проводить в такой период года, когда население наименее подвижно (в смысле переездов). В практике российской статистики перепись населения проводится, как правило, в зимний период.

В выбранном периоде важно наметить определенный момент, именуемый *критическим*, по состоянию на который должны регистрироваться все сведения. От критического момента следует отличать *срок наблюдения*, или *время производства наблюдения*, – отрезок времени, в течение которого должны быть собраны сведения об изучаемом явлении. Длительность срока, или времени производства наблюдения, устанавливается исходя из того, какова численность лиц, проводящих наблюдение, какова численность единиц наблюдения.

В переписи населения 2002 г. срок проведения переписи равнялся 8 дням: с 9 по 16 октября включительно. Критическим моментом служили 0 часов 9 октября. Это значит, что независимо от того, когда счетчик явился в ту или иную семью (9, 10 октября или позже), он должен был включить в переписные листы всех проживающих по состоянию на критический момент, т.е. если кто-то умер после 0 часов 9 октября – он должен был быть переписан; все же родившиеся после этого момента в перепись не включались.

Наряду с вопросами о времени наблюдения решается вопрос о ***месте наблюдения***, т.е. о том, г д е проводить наблюдение. Так, при переписях населения в практике советской и российской статистики местом наблюдения принято считать место жительства, отдельные квартиры. Для других наблюдений вопрос о месте наблюдения может решаться по-разному.

До проведения статистического наблюдения должно быть решено, какова будет ***организационная форма наблюдения***, какое наблюдение будет применено по охвату единиц (сплошное или несплошное; если несплошное, то какое по виду: выборочное, основного массива, анкетное), на основании каких источников будет проводиться регистрация сведений (путем непосредственного наблюдения на основе документов, путем опроса), а также каким ***способом*** будет организован сбор информации (экспедиционным, явочным или иным).

При проведении переписи населения в России в 2002 г., как уже отмечалось ранее, одни сведения собирались на основе сплошного наблюдения, а другие выборочно. Заполнение переписных листов осуществлялось экспедиционным способом (в порядке обхода счетчиками-регистраторами жилых помещений), методом опроса населения.

Для того чтобы обеспечить успех такой большой учетной работы, как перепись населения, и опробовать проект программы (с точки зрения ее содержания и удачности формулировок вопроса), а также возможную нагрузку счетчиков, обычно проводятся *пробные переписи* по отдельным районам страны, различающимся по природно-географическим условиям, экономической структуре хозяйства, социальному и национальному составу населения. Поскольку перепись населения в России первоначально намечалась на 1999 г., а затем была перенесена на 2002 г., то фактически были проведены две пробные переписи – в 1997 и 2001 гг. Пробные переписи помогают учесть и скорректировать многие вопросы статистического наблюдения.

К числу подготовительных работ, предшествующих любому большому наблюдению, в частности переписи населения, относятся также: уточнение границ городских поселений, названий улиц; составление списков домовладений населенных пунктов; разбивка территории каждого района и города на переписные отделы, инструкторские и счетные участки; определение необходимой численности переписных кадров, их подбор и подготовка; подготовка документации переписи; проведение разъяснительной работы среди населения.

2.4. Ошибки статистического наблюдения и контроль данных наблюдения

Тщательно разработанные и продуманные вопросы статистического наблюдения – залог успеха в получении достоверных данных об изучаемом явлении.

Однако как бы тщательно ни были предусмотрены отдельные моменты статистического наблюдения и программы наблюдения и как бы точно ни руководствовались всеми указаниями инструкции лица, осуществляющие сбор сведений, при любом статистическом наблюдении могут возникнуть ошибки (погрешности). Эти **ошибки наблюдения** могут возникнуть по разным причинам: за счет описок, оговорок, округлений, неправильного заполнения формуляра, запамятования тех, кто отвечает, или стремления

скрыть, исказить факты; при непосредственном наблюдении (при взвешивании, измерении) ошибки могут возникать и из-за неточности измерительных приборов и т.п.

Все ошибки можно разделить на *преднамеренные* и *непреднамеренные*. Непреднамеренные, в свою очередь, могут носить *случайный* и *систематический* характер.

Случайные ошибки наблюдения, возникающие и по вине отвечающего, и по вине регистраторов в результате описок, оговорок, незнания и тому подобного, не столь опасны для результатов наблюдения, так как такие ошибки одинаково часто могут встретиться и в сторону преуменьшения, и в сторону преувеличения, а при большом числе наблюдений они взаимопогашаются, нейтрализуются.

Непреднамеренные систематические ошибки возникают главным образом при опросе за счет округлений количественных показателей (округление возраста, стажа работы, дохода и т.п.) или за счет неточностей измерительных приборов при непосредственном наблюдении. Так, замечено, что при регистрации возраста путем опроса наиболее часто возраст округляется до чисел, оканчивающихся на 0 и 5. В результате получается, что 40-летних оказывается по записям значительно больше, чем 39- и 41-летних. Это явление получило в статистике название аккумуляции возрастов. Такие погрешности приходится исправлять уже при обработке собранного статистического материала.

Преднамеренные ошибки, как говорит само название, возникают в силу сознательного стремления лиц, дающих сведения, исказить истину: уменьшить или увеличить величину того или иного показателя. Ясно, что преднамеренные ошибки искажают сведения в одном направлении (либо преуменьшают, либо преувеличивают). Этот род ошибок наиболее опасен для статистического исследования, и надо всегда приложить максимум усилий, чтобы выявить эти ошибки и устранить.

Все указанные ошибки могут возникнуть как при сплошном, так и при несплошном статистическом наблюдении в процессе регистрации самих фактов. Отсюда и их название — **ошибки регистрации**.

При несплошном наблюдении наряду с ошибками регистрации могут возникнуть расхождения между показателями несплошного наблюдения и показателями для всей совокупности при условии сплошного наблюдения. Расхождения между показателями несплошного и сплошного наблюдения в статистике именуют **ошибками репрезентативности**. Эти ошибки тоже могут но-

сить *случайный характер* (в силу несплошного наблюдения) и *систематический.*

Случайные ошибки, в частности, при выборочном наблюдении неизбежны, но они легко поддаются учету, и при правильно организованном случайном отборе всегда можно определить величину таких ошибок и пределы, в которых может заключаться величина изучаемого показателя во всей совокупности (см. главу 6).

Систематические ошибки репрезентативности, как правило, возникают при неправильной организации выборки, т.е. в том случае, когда нарушен принцип случайности отбора единиц из так называемой генеральной совокупности (например, если специально отбираются единицы с заведомо заниженными или завышенными значениями показателей).

Так как в процессе наблюдения всегда могут возникнуть ошибки, то, естественно, весь собранный материал должен быть подвергнут **контролю** в целях устранения обнаруженных ошибок. Проверка правильности зафиксированных в статистических формулярах сведений должна проводиться с точки зрения *логического* и *арифметического контроля.*

Логический контроль ставит своей целью определить соответствие ответа поставленному вопросу или соответствие между ответами на разные вопросы программы. Например, если на вопрос «возраст» обнаружен ответ «русский», то ясно, что ответ в данном случае не соответствует вопросу, что это ошибка, вызванная записью ответа не в той строке или графе.

Если же на вопрос «возраст» получен ответ «3 года», а на вопрос «состоит ли в браке» – ответ «да», то каждый ответ здесь соответствует вопросу, однако ответы не соответствуют друг другу. Чтобы установить, в каком же ответе содержится ошибка, следует рассмотреть ответы на другие взаимно контролируемые вопросы. Так, если в рассматриваемом случае в графе «место работы» записано наименование определенного предприятия и в графе «образование» указано «среднее» или «высшее», то ясно, что допущена ошибка в возрасте.

Можно установить логическую неточность в ответах, сопоставляя фактические показатели с плановыми, с показателями за предшествующие периоды, сопоставляя показатели по районам, находящимся в одинаковых природных условиях, показатели, относящиеся к одному и тому же явлению, полученные по нескольким источникам, и т.п.

К количественным ответам, полученным как сумма, разность, произведение или часть других показателей, всегда следует при-

менять наряду с логическим и арифметический контроль. Целью ***арифметического контроля*** является проверка правильности вычислений.

Все обнаруженные ошибки по возможности должны быть исправлены. Для этого порой приходится проводить контрольные опросы, запросы почтой, по телефону, факсу и пр.

Глава 3
СВОДКА И ГРУППИРОВКА СТАТИСТИЧЕСКИХ ДАННЫХ

В результате проведения статистического наблюдения получают данные о признаках каждой обследованной единицы статистической совокупности. Однако эти массивы данных, содержащие подробные сведения о каждой единице совокупности, собирают не для того, чтобы получить характеристики каждой из них, а с целью изучить совокупность в целом, выявить ее характерные группы и закономерности.

Для этого необходимо обобщить и систематизировать сведения, полученные в ходе статистического наблюдения.

Обобщение и систематизация первичных статистических данных – это самостоятельный этап статистического исследования, основная задача которого получить полную и всестороннюю характеристику как совокупности в целом, так и отдельных ее частей и представить полученную информацию об изучаемой совокупности в наиболее удобной для пользователей форме.

В статистической практике этот этап статистического исследования именуют *этапом сводки и группировки статистических данных.*

Сводка данных, полученных в результате статистического наблюдения, состоит в систематизации, обработке и получении общих и групповых итогов, а также расчете производных показателей (средних и относительных величин).

По способу организации различают централизованную и децентрализованную сводку. При *централизованной сводке* все данные наблюдения сосредоточиваются в одном центре, где они обрабатываются. Примером применения централизованной сводки является обработка результатов федерального статистического наблюдения за составом затрат на рабочую силу, статистического наблюдения за внешней трудовой миграцией. При *децентрализованной сводке* первичные статистические материалы разрабатываются на уровне административных районов, итоги сводятся на уровне субъектов Российской Федерации, а затем на уровне государства в целом. Такой вид сводки используется при обработке данных, получаемых от предприятий и организаций по установленным формам статистической отчетности.

На практике имеет место *сочетание децентрализованной* и *централизованной сводки*. При разработке материалов переписи населения часть итогов получают в порядке децентрализованной сводки (о численности населения городов и других населенных пунктов, численности мужчин и женщин), но полные итоги по всем признакам получают в результате централизованной обработки данных.

Единицы статистической совокупности отличаются друг от друга как качественными, так и количественными признаками. В связи с этим отдельные единицы совокупности, сходные по своему виду, размеру, отношению к другим частям совокупности и т.д., необходимо объединить в обособленные группы. Разбиение совокупности на однородные виды, классы выполняют в ходе группировки.

Группировкой называется расчленение единиц статистической совокупности на группы, однородные по какому-либо одному или нескольким признакам. Группировка позволяет систематизировать данные статистического наблюдения. В результате группировки они превращаются в упорядоченную статистическую информацию, пригодную для дальнейшего статистического анализа.

3.1. Виды группировок

Каждая единица исследуемой совокупности обладает рядом свойств, или признаков. Отдельные значения, которые может принимать тот или иной варьирующий признак, называются его *вариантами*.

По характеру вариантов признаки делятся на атрибутивные и количественные. Признак называется *атрибутивным* в том случае, если его варианты не выражаются числами, и *количественным*, если его варианты выражаются в виде чисел.

Признаки, на основе которых получена группировка, называются *группировочными*.

Например, население может быть сгруппировано на основе таких признаков, как пол, национальность, статус в занятости. Эти признаки являются атрибутивными. Группировки, полученные по этим признакам, называются ***атрибутивными*** или ***качественными.***

Если группировка получена по количественному признаку, она называется ***количественной.*** Примерами таких группировок служат распределение населения по возрасту, по размеру дохода, группировка предприятий по численности работников и др.

Выбор группировочных признаков имеет огромное значение. В основу группировки должны быть положены наиболее важные,

существенные признаки. Их выбор определяется как качественной особенностью изучаемых процессов и явлений, так и целями исследования. Выделение наиболее типичных черт, которые присущи некоторым единицам совокупности, позволяет получить качественно однородные группы. В таких группах легче обнаружить закономерности изменения и развития явления, более наглядна реакция на те факторы, которые влияют на изменение его состояния.

Анализируя экономическую и социальную жизнь общества, выделяют и изучают отдельные типы явлений. Такого рода группировки называются ***типологическими.*** Довольно часто между типологическими и качественными группировками ставят знак равенства. Это не совсем верно, поскольку некоторые типы явлений могут быть выделены и по количественному признаку. Например, группировка предприятий на малые, средние и крупные проводится по таким количественным признакам, как численность персонала, объем продукции, стоимость основных фондов, причем для разных видов деятельности значение этих признаков различно.

Примером типологических группировок служит деление населения на такие группы, как молодежь, лица среднего возраста и др. Следует отметить, что пороговые значения количественных признаков, отделяющие одну группу от другой, изменяются во времени и пространстве.

При анализе явлений часто используют ***пространственные*** группировки, созданные по географическому признаку, при этом в основу группировок могут быть положены существующее административно-территориальное деление, природно-климатические зоны, части света и т.д. Данные, сгруппированные по территориальному признаку, представляют важный информационный массив как для анализа явлений в пределах отдельных территорий, так и для сопоставления одних и тех же явлений (например, уровни цен и доходов, показатели рождаемости и смертности и др.) на различных территориях.

Данные любой группировки соответствуют определенному моменту времени или периоду.

С течением времени изменяется как численность совокупности, так и численность и соотношение отдельных ее групп. В табл. 3.1 приведены данные о том, как изменилась численность населения России за столетие, а также как изменилась численность и соотношение между двумя группами населения – городским и сельским.

Таблица 3.1

Изменение численности населения России (в современных границах) в 1897—2002 гг.

Год	Численность населения, млн чел.			Доля населения в общей численности, %	
	Всего	городского	сельского	городского	сельского
1897	67,5	9,9	57,6	15	85
1926	92,7	16,4	76,3	18	82
1939	108,4	36,3	72,1	33	67
1959	107,5	61,6	55,9	52	48
1970	130,0	81,0	49,0	62	38
1979	137,6	95,4	42,2	69	38
1989	147,4	108,4	39,0	74	26
2002	145,2	106,4	38,8	73	27

Примечание. По данным переписей населения на даты их проведения.

Для исследования зависимости между явлениями используют ***аналитические*** группировки. При их построении можно установить взаимосвязь между двумя признаками и более. При этом один признак будет результативным, а другой (другие) – факторным. *Факторными* называются признаки, под воздействием которых изменяются результативные признаки.

Для того чтобы установить взаимосвязь между признаками, данные следует сгруппировать по признаку-фактору и затем вычислить среднее значение результативного признака в каждой группе. Сопоставляя изменения значений факторного и результативного признаков, определяют характер связи между ними. Если с увеличением значения факторного признака возрастает и значение результативного признака, то между ними существует прямая связь. Изменение их значений в противоположных направлениях свидетельствует об обратной связи между признаками.

В качестве примера взаимосвязи между признаками рассмотрим табл. 3.2.

Данные, приведенные в табл. 3.2, показывают, что чем меньше предприятие (по численности работников), тем продолжительнее рабочая неделя. Иной характер связи прослеживается при сопоставлении размера торгового предприятия и среднемесячной заработной платы работников. Однако заработная плата зависит не только от размера предприятия, но и от продолжительности рабочей недели. Если сопоставить отработанное время по различ-

Таблица 3.2

Группировка магазинов по численности работников
(данные условные)

Численность работников, чел.	Количество магазинов	Фактическая продолжительность рабочей недели, ч	Среднемесячная заработная плата работников, руб.
До 5	83	42,0	4750
5–10	49	39,5	4940
11–19	52	38,1	5670
20–49	29	37,4	5420
50–99	12	37,6	5560
100 и более	11	37,2	5490

ным группам предприятий со среднемесячной заработной платой, можно говорить о наличии прямой связи между этими двумя признаками. Отметим, что в связке «численность работников – продолжительность рабочей недели» последняя является результативным признаком, а в паре «продолжительность рабочей недели – среднемесячная заработная плата» этот же признак становится факторным.

На размер заработной платы влияют оба фактора (размер предприятия и продолжительность работы). В случае если изучается влияние на результат нескольких факторов, используют *многофакторную* аналитическую группировку.

От выбора группировочного признака часто зависит и **число образуемых групп**. Так, при группировке населения по полу возможны только две группы, а при группировке по национальности может быть образовано столько групп, сколько разнообразных национальностей и народностей зафиксировано на данной территории на момент обследования.

Следует иметь в виду, что многие экономические и социальные явления и процессы хорошо изучены, поэтому для качественных группировок предусмотрено устойчивое разбиение совокупности на группы достаточно однородных явлений. Такое устойчивое разбиение на группы проводится на основе свойств и различий элементов совокупности и называется *классификацией*.

Классификации играют большую роль при систематизации статистических данных. Значение классификаций все время возрастает. Они служат статистическим нормативом, в соответствии с которым группируется статистическая информация. Классификации не остаются неизменными – появляются новые типы, изменяются условия и принципы, на которых базируются те или иные

классификаторы. Например, классификация форм собственности, разработанная для централизованно планируемой экономики, не соответствовала условиям рыночной экономики, поэтому действовавший в России классификатор форм собственности пришлось пересмотреть. Однако глобальные, коренные изменения происходят в экономической и социальной жизни не так уж часто. Кроме того, далеко не все изменения требуют пересмотра классификации в целом. Накопленный опыт позволяет в рамки действующих классификаций встроить новые блоки, если сформировались новые группы либо значение тех или иных типов явлений возросло настолько, что их нужно выделить в самостоятельную группу.

В отличие от классификации группировка проводится обычно для целей конкретного обследования. Такие группировки можно использовать и в последующих обследованиях. Это даже желательно, поскольку обеспечивается сопоставимость их данных. При необходимости можно применять другие группировки.

Число групп при использовании количественного признака зависит от числа единиц изучаемого явления, степени колеблемости группировочного признака, а также от того, является ли признак дискретной величиной (т.е. характеризуется только целыми значениями) или непрерывной (т.е. в пределах вариации может принимать любые значения, отличающиеся друг от друга на сколь угодно малую величину).

В совокупности, где варьирующий признак носит дискретный характер и может принимать ограниченное число значений, количество групп, как правило, равно количеству возможных значений. Примером такой группировки служит распределение семей одного из городов по числу детей, приведенное в табл. 3.3.

Таблица 3.3

Распределение семей города *N* по числу детей на 1 января 2004 г. (данные условные)

№ п/п	Количество детей в семье	Количество семей
1	0	1830
2	1	3953
3	2	2780
4	3	801
5	4	24
6	5	11
7	6 и более	4
Всего		9403

Следует обратить внимание на последнюю группу «6 и более», в ней нарушен принцип, по которому образованы группы для данной совокупности, – указывалось точное число детей в семье. Обычно это делается из практических соображений – чтобы не увеличивать число групп, вводя значения признака, которые редко встречаются в совокупности.

Группировки, образованные на основе точных значений варьирующего группировочного признака, применяют тогда, когда количество возможных дискретных значений невелико (например, если речь идет о группировке семей по численности членов семьи, о распределении жилых помещений, занимаемых одной семьей, по числу комнат и др.).

Если варьирующий признак является непрерывной величиной или дискретной величиной, которая может принимать очень большое число значений (например, численность работников на предприятии может изменяться от одного до нескольких тысяч), то в этом случае число групп зависит от степени колеблемости данного признака, а также от объема изучаемой совокупности.

При группировке данных возникает вопрос о том, на сколько групп будет разбита изучаемая совокупность. На этот вопрос нет стандартного, однозначного ответа.

Если распределение признака в границах его вариации достаточно равномерно или близко к нормальному, диапазон колебаний признака разбивают на р а в н ы е интервалы, длину которых определяют по формуле

$$h = \frac{x_{max} - x_{min}}{k},$$

где x_{max} – максимальное значение признака в совокупности;
x_{min} – минимальное значение признака в совокупности;
k – число групп.

Число групп может быть задано (на основе опыта предыдущих обследований). В том случае, если вопрос о числе групп приходится решать самостоятельно, можно использовать **формулу Стерджесса для определения оптимального числа групп**:

$$k = 1 + 3{,}322 \lg N,$$

где N – число единиц в совокупности.

Например, необходимо осуществить группировку работников предприятия по размеру месячной заработной платы, при условии, что ее минимальный размер составил 1359 руб., а максимальный – 6449 руб. при среднесписочной численности работни-

ков предприятия 645 человек. Находим длину интервала, используя формулу Стерджесса для определения оптимального числа групп:

$$h = \frac{6449 - 1359}{1 + 3{,}322 \lg 645} = 492{,}6 \text{ руб.}$$

Полученное значение следует округлить для облегчения расчетов до 500 руб. Процедуру округления при расчете интервала проводят всегда. Трехзначное, четырехзначное или большее число округляют до ближайшего числа, кратного 50 или 100. Если число имеет два знака до запятой и несколько знаков после запятой, его округляют до целого, если один знак до запятой и несколько знаков после запятой – до десятых и т.д.

В нашем примере диапазон колебаний заработной платы будет разбит на следующие интервалы: 1) 1000–1500 руб.; 2) 1500–2000 руб.; 3) 2000–2500 руб. и т.д. Последним интервалом будет 6000 руб. и более.

Часто значения варьирующего признака распределены таким образом, что при использовании равного интервала для образования групп излишне увеличивается их количество, при этом многие группы будут малочисленными. В этих условиях совокупность разбивают на группы с неравными интервалами. Примером такой группировки может служить распределение населения по размеру среднедушевого дохода, приведенное в табл. 3.4.

Таблица 3.4

Распределение населения России по размеру среднедушевых месячных доходов в 2000 и 2002 гг. (в %)

	2000 г.	2002 г.
Все население	100	100
В том числе со среднедушевыми денежными доходами в месяц, руб.:		
до 500	3,4	0,8
500,1–750	7,3	2,3
750,1–1000	9,6	3,9
1000,1–1500	19,8	10,7
1500,1–2000	16,3	11,9
2000,1–3000	20,6	21,0
3000,1–4000	10,5	15,2
свыше 4000	12,5	34,2

Источник: Российский статистический ежегодник. 2003. – М.: Госкомстат России, 2003. – С. 185.

Различия в длине интервала могут быть обусловлены не только характером изменения варьирующего признака, но и особенностями изучаемых экономических и социальных явлений. При этом не наблюдается какой-либо определенной тенденции увеличения или уменьшения интервала при образовании групп. Рассмотрим это на примере группировки экономически активного и занятого населения по возрастным группам, приведенной в табл. 3.5.

Таблица 3.5

Распределение численности экономически активного населения и населения, занятого в экономике, по возрастным группам в 2003 г. (на конец августа, в %)

	Всего	В том числе в возрасте, лет						
		15–19	20–24	25–29	30–49	50–54	55–59	60–72
Экономически активное население:								
всего	100	3,0	10,2	12,4	54,5	11,2	3,9	4,8
мужчины	100	3,3	10,8	13,1	53,3	10,4	4,3	4,9
женщины	100	2,7	9,6	11,7	55,7	12,0	3,5	4,8
Занятое население:								
всего	100	2,4	9,4	12,4	55,1	11,5	4,1	5,0
мужчины	100	2,8	10,1	13,0	53,9	10,7	4,5	5,1
женщины	100	1,9	8,7	11,7	56,5	12,3	3,7	5,0

Источник: Обследование населения по проблемам занятости, август 2003 г. — М.: Госкомстат России, 2003. — С. 39, 49.

В табл. 3.5 не наблюдается прогрессивного увеличения или уменьшения интервала. Специфика анализируемых явлений – экономически активного и занятого населения – требует более детальной информации о молодежи (до 30 лет) и лицах предпенсионного возраста (50–54 года у женщин и 55–59 лет у мужчин).

В приведенных выше примерах используются два вида интервалов: закрытые и открытые. *Закрытыми* называются интервалы, у которых указаны обе границы, *открытыми* – интервалы с одной границей (верхней у первого интервала и нижней у последнего интервала).

Для расчета показателей статистической совокупности необходимо «закрыть» открытые интервалы. Для этой цели используют интервал, соседний с открытым.

Если обратиться к данным табл. 3.4, то в результате операции закрытия интервалов первый интервал будет «250,1–500», последний – «4000,1–5000».

Однако следует помнить, что существуют логические и установленные границы совокупностей. Например, в группировке населения по возрасту: до 3 лет; 3–7 лет и т.д. – для первой группы логической нижней границей интервала будет 0, т.е. целесообразно рассматривать интервал от 0 до 3 лет.

Если речь идет о группировке населения в трудоспособном возрасте, то для открытых интервалов следует использовать установленные законодательством, т.е. юридические, границы совокупности: от 16 до 55 лет у женщин и от 16 до 60 лет у мужчин.

В целях статистического исследования часто приходится пользоваться данными, относящимися к различным периодам, сопоставлять информацию по отдельным отраслям, регионам, странам, опираясь на уже сгруппированные данные, причем сгруппированными, как правило, на разной основе. В этих условиях требуется перегруппировка уже сгруппированных данных.

Операция перегруппировки, т.е. образование новых групп на базе ранее созданных группировок, называется ***вторичной группировкой***.

При анализе явления необходимо из большого количества первоначально созданных групп образовать более крупные группы. Например, при переписи населения базисными являются погодовые группы населения, на основе которых можно образовать любые группы. Вторичная группировка в данном случае не вызывает проблем.

Другой пример. Имеются группировки предприятий различных отраслей экономики по численности работников. В силу специфики отраслей группировки по этому признаку довольно значительно отличаются. Причина таких различий в том, что максимальная численность персонала малого предприятия в промышленности в несколько раз превышает аналогичный показатель в торговле, науке и других непроизводственных отраслях. В этом случае за базисную может быть выбрана группировка, используемая в промышленности, либо другая стандартная группировка, которая учитывает специфику не одной отрасли, а широкого круга отраслей. При этом в ходе укрупнения интервалов некоторые из них целиком войдут во вновь образованные интервалы. Другие интервалы придется разбивать на части согласно новым границам. При этом в новом интервале число единиц признака будет пропорционально части старого интервала, которая попадает в соответствующий новый интервал.

Вторичные группировки проводятся для совокупностей, сгруппированных не только по количественным, но и по качественным

признакам. Наиболее часто это приходится делать при сопоставлении данных, полученных в разных странах. В этом случае показатели, рассчитанные на базе отличающихся друг от друга национальных классификаций, перегруппировывают. За основу, как правило, принимают действующую международную классификацию, которая служит международной статистической нормой. Эту работу осуществляют в два этапа: сначала разрабатывают ключи перехода от национальной классификации к международной, затем на этой основе проводят перегруппировку данных национальной статистики.

Метод группировок – один из важнейших методов статистики, без которого немыслимо изучение массовых явлений. Данная глава содержит самые общие сведения о группировках как обязательном этапе статистического исследования, элементе сводки, приеме систематизации и обобщения массовых данных.

На практике при обработке массовых данных задача расчленения множества единиц изучаемой совокупности на группы по определенным признакам решается порой более сложными приемами, разработанными в последние годы и требующими использования компьютеров. Особенно это относится к группировкам по нескольким признакам, т.е. на основе множества признаков. Для этой цели разработаны так называемые методы многомерной классификации: классификация на основе многомерной средней, кластерный анализ, метод главных компонент*.

3.2. Статистические таблицы

Результаты сводки и группировки статистических данных оформляются в статистические таблицы. *Статистическая таблица* – это форма наиболее краткого и рационального изложения цифровых данных об изучаемой статистической совокупности.

Незаполненная цифрами статистическая таблица называется *макетом*. Макеты статистических таблиц разрабатывают на стадии подготовки к этапу сводки и группировки данных статистического наблюдения.

Макет таблицы – это сетка, состоящая из горизонтальных строк и вертикальных колонок (граф), каждая из которых имеет название. Клетки, образуемые на пересечении строк и колонок, заполняют статистическими данными.

---

* Подробнее см., например: *Дубров А.М., Мхитарян В.С., Трошин Л.И.* Многомерные статистические методы. — М.: Финансы и статистика, 1998; *Енюков И.С.* Методы — алгоритмы — программы многомерного статистического анализа. — М.: Финансы и статистика, 1986; *Елисеева И.И., Рукавишников В.О.* Группировка, корреляция, распознавание образов. — М.: Статистика, 1977; *Мандель Н.Д.* Кластерный анализ. — М.: Финансы и статистика, 1997.

Каждая статистическая таблица содержит подлежащее и сказуемое. *Подлежащим* таблицы называется объект, отдельные единицы или его части (группы), которые характеризуются соответствующими показателями. *Сказуемым* называются показатели, которые характеризуют подлежащее. Подлежащее таблицы обычно составляет название ее строк, сказуемое – название колонок. Иногда в целях получения более компактной таблицы подлежащее и сказуемое меняют местами, т.е. подлежащее указывают по графам, а сказуемое – по строкам.

Подлежащее статистической таблицы может быть простым и сложным. По характеру подлежащего различают простые, групповые и комбинационные таблицы.

В ***простой*** таблице подлежащее представляет перечень отдельных единиц изучаемого объекта. Например, перечень предприятий и организаций или перечень отдельных дат (годы, кварталы и т.д.). Примером простой таблицы является табл. 3.6, в которой подлежащим выступают иностранные инвестиции в экономику России (перечень наиболее крупных стран-инвесторов), а сказуемым – объем инвестиций отдельной страны как по абсолютной величине, так и в процентах.

Таблица 3.6

Страны с наиболее значительными инвестициями в экономику России в 2002 г.

	Млн долл. США	% к итогу
Иностранные инвестиции:		
всего	19 780	100
в том числе:		
Германия	4001	20,2
Кипр	2327	11,8
Великобритания	2271	11,5
Швейцария	1349	6,8
Виргинские острова (Британские)	1307	6,6
Люксембург	1258	6,4
Франция	1184	6,0
Нидерланды	1168	5,9
США	1133	5,7
Финляндия	592	3,8
Япония	441	2,2
прочие страны	2749	13,9

Источник: Инвестиции в России. 2003: Стат. сб. — М.: Госкомстат России, 2003. — С. 98.

В ***групповых*** таблицах статистическая совокупность разбивается на отдельные группы по какому-то одному признаку. При этом каждую группу можно охарактеризовать одним или несколькими показателями. Примером групповой таблицы служит табл. 3.4, в которой подлежащим выступают группы населения с различным среднедушевым доходом в месяц, а сказуемым – удельный вес каждой группы, а также табл. 3.2, в которой подлежащим являются группы магазинов с разной численностью работников, а сказуемым – показатели количества магазинов, продолжительности рабочей недели и среднемесячной заработной платы работников.

В ***комбинационных*** таблицах объект исследования, т.е. подлежащее, разбивается на группы не по одному, а по нескольким признакам. Примером может служить макет табл. 3.7.

Подлежащее в этом макете таблицы – группировка предприятий обрабатывающей промышленности по двум признакам: отраслевой принадлежности и формам собственности.

Разработка сказуемого таблицы также может быть простой и сложной.

При *простой разработке* показатели, характеризующие подлежащее таблицы, располагаются параллельно друг другу. Примером простой разработки сказуемого служат табл. 3.2 и 3.7.

При *сложной разработке* сказуемого один признак комбинируется с другим. В качестве примера сложной разработки сказуемого приведем табл. 3.8. Она содержит данные о затратах на рабочую силу по отраслям экономики России, полученные из материалов выборочного обследования о составе затрат на рабочую силу за 2002 г.

При построении статистических таблиц следует соблюдать ряд условий.

Каждая таблица должна иметь краткий заголовок, который в то же время должен достаточно полно и четко отражать содержание анализируемого объекта.

Все строки и графы таблицы должны иметь названия, при этом повторяющиеся термины следует выносить в общие заголовки.

В строках и графах таблицы должны быть указаны единицы измерения, соответствующие показателям, содержащимся в подлежащем и сказуемом, при этом следует использовать общепринятые сокращения единиц измерения (руб., км^2 и т.д.). Единая единица измерения для всех строк и столбцов должна быть вынесена за пределы таблицы и размещена с правой стороны над таблицей.

Таблица 3.7

Группировка предприятий обрабатывающей промышленности по отраслям и формам собственности

Отрасль обрабатывающей промышленности	Форма собственности предприятия	Число предприятий	Среднесписочная численность работающих	Основной капитал, млн руб.	Произведенная продукция, млн руб.
Пищевая промышленность	Государственная				
	Муниципальная				
	Общественных организаций				
	Частная				
	Иностранная				
	Прочие				
	Всего				
Легкая промышленность	Государственная				
	Муниципальная				
	Общественных организаций				
	Частная				
	Иностранная				
	Прочие				
	Всего				
.....	Государственная				
	Муниципальная				
	Общественных организаций				
	Частная				
	Иностранная				
	Прочие				
	Всего				
Всего по обрабатывающей промышленности	Государственная				
	Муниципальная				
	Общественных организаций				
	Частная				
	Иностранная				
	Прочие				
	Всего				

Таблица 3.8

Среднечасовые и среднемесячные затраты на рабочую силу по отраслям экономики России в зависимости от форм собственности

(руб.)

	Затраты на рабочую силу в расчете на					
	отработанный человекочас			одного работника в месяц		
	Всего	В том числе по организациям		Всего	В том числе по организациям	
		государственным и муниципальным	негосударственным		государственным и муниципальным	негосударственным
Всего по отраслям экономики	53,9	46,8	56,9	7644,0	6649,0	8064,5
В том числе:						
промышленность	55,7	45,8	57,2	7766,6	6388,0	7973,8
строительство	56,3	52,5	56,8	8037,9	7517,6	8112,2
транспорт	58,2	56,1	62,9	8220,2	7878,0	8987,9
связь	51,6	36,7	64,5	7470,2	5316,7	9311,7
оптовая и розничная торговля, общественное питание	37,1	44,1	35,6	5504,7	6451,8	5300,9
финансы, кредит, страхование, пенсионное обеспечение	107,3	83,7	118,2	15 613,2	11 996,9	17 330,3

Источник: Труд и занятость в России: Стат. сб. – М.: Госкомстат России, 2003. – С. 354.

Все данные одной строки (графы) следует представлять с одинаковой степенью точности.

Итоговые строки (столбцы) могут располагаться как в первых, так и в последних строках (столбцах) таблицы.

Желательно нумеровать строки и столбцы таблицы. Для больших таблиц это обязательное требование.

Все клетки таблицы должны быть заполнены. Причины отсутствия данных в той или иной клетке могут быть различны, поэто-

му как в отечественной, так и в зарубежной статистике при заполнении таблиц используют следующие условные обозначения:

- «...» (многоточие) – явление существует, но сведений о нем нет;
- «0» (нуль) – явление существует, но значение его показателя меньше половины единицы, принятой при округлении (например, меньше 0,5 при записи данных целыми числами либо меньше 0,05, если данные выражены с точностью до одного знака после запятой, и т.д.);
- «–» (тире) – явление отсутствует;
- «х» (крестик) – клетка не подлежит заполнению.

В таблице должны быть отмечены (цифрами, буквами либо другими условными обозначениями) предварительные данные, а также данные, которые рассчитаны по методологии, отличной от методологии расчета остальных данных. Таблица, содержащая подобные сведения, должна быть снабжена примечаниями, сносками, где даны необходимые разъяснения. Обычно их располагают ниже таблицы, иногда в конце текста (материала).

Необходимо указывать источники данных, приведенных в таблице (название обследования с указанием организации, которая его проводила, название публикации или указание на условность данных).

Составляя таблицы на этапе обобщения и систематизации статистических данных, необходимо соблюдать указанные правила.

Глава 4
ОБОБЩАЮЩИЕ СТАТИСТИЧЕСКИЕ ПОКАЗАТЕЛИ

В результате сводки данных статистического наблюдения получают различные показатели, одни из которых характеризуют совокупность в целом, другие – отдельные ее части.

Статистика, имея дело с массовыми явлениями и процессами, давая им количественную оценку, оперирует не просто числами, а статистическими показателями.

Под ***статистическим показателем*** понимается обобщающая количественная характеристика изучаемого объекта или его свойства.

Именно на этапе статистической сводки от индивидуальных значений признаков у отдельных единиц совокупности путем суммирования переходят к показателям совокупности, которые называются ***обобщающими***.

Естественно, что переход к обобщающим показателям операцией суммирования не ограничивается. Вычисления на основе итоговых значений, которые уже являются показателями, можно продолжить. Например, если известны суммарный объем выпуска продукции за период и количество человекочасов, затраченных на ее производство, можно рассчитать среднюю часовую выработку работников. Это тоже обобщающий показатель, он получен как частное от деления двух итоговых показателей и отражает уровень производительности труда данной совокупности работников.

В других случаях для получения обобщающих показателей приходится сопоставлять индивидуальные или итоговые значения одних и тех же или различных признаков; таким образом вычисляют относительные показатели.

В зависимости от методов расчета обобщающие показатели могут быть *абсолютными*, *относительными* или *средними* величинами.

Каждому показателю соответствует конкретная методология расчета или способ вычисления. При этом любой статистический показатель должен быть точно определенным, что выдвигает ряд требований к его наименованию. В нем должны быть указаны:

- статистическая структура показателя (среднее значение, сумма, процент к итогу и т.д.);

- его содержание (население, инвестиции, объем добычи и т.д.);
- позиция в классификации, совокупность объектов (обрабатывающая промышленность России, предприятия угольной промышленности Кузнецкого бассейна и т.п.);
- единица измерения (человек, тонна, километр и др.);
- время (на начало года, за 1990–2002 гг. и т.п.);
- специальные уточнения (в рыночных ценах 2000 г. и пр.).

4.1. Абсолютные величины

Абсолютные обобщающие показатели – это число единиц по совокупности в целом или по ее отдельным группам, которое получают в результате суммирования зарегистрированных значений признаков первичного статистического материала. Данные показатели могут быть получены и расчетным путем на основе других показателей (например, прирост банковских вкладов населения за период определяется как разность вкладов на конец и начало периода).

Абсолютные величины как обобщающие показатели характеризуют либо численность совокупности (численность экономически активного населения, количество предприятий различных форм собственности и т.д.), либо объем признаков совокупности (размер инвестиций, затраты на рабочую силу и т.д.).

Любая абсолютная величина всегда имеет свою единицу измерения, присущую тем или иным явлениям.

Широкое применение находят *натуральные* единицы измерения, как простые (тонна, штука, квадратный и кубический метр, километр и т.д.), так и сложные, представляющие собой комбинацию двух величин (тонно-километр, киловатт-час и др.).

Разновидностью натуральных показателей являются *условно-натуральные* показатели. Их применяют для получения абсолютных обобщающих показателей, когда отдельные группы слагаемых, входящие в совокупность, не поддаются непосредственному суммированию. Предварительно все слагаемые необходимо привести к сопоставимому виду. С помощью специальных коэффициентов пересчета слагаемые выражают в единой стандартной единице измерения, что позволяет получить обобщающий показатель. Например, различные виды топлива соизмеряют по условному топливу с теплотворной способностью 7000 ккал/кг, продукты химической промышленности, руды металлов – по содержанию полезного вещества.

В качестве абсолютных обобщающих показателей используют *стоимостные* показатели, они позволяют соизмерить в денежной

форме величины, которые нельзя соизмерить в натуральной форме (например, затраты на производство и расходы населения).

Кроме того, в качестве абсолютных обобщающих показателей используют и показатели, измеренные в *единицах труда*. В человекоднях или человекочасах измеряются различные фонды времени, которыми располагают отдельные производственные единицы, а также затраты времени на каждую технологическую операцию, на производство того или иного вида продукции.

4.2. Относительные величины

Анализируя статистические данные, необходимо сопоставлять явления во времени и пространстве, исследовать закономерности их изменения и развития, изучать структуру совокупностей. С помощью абсолютных величин эти задачи невыполнимы, в этом случае необходимо использовать относительные величины.

Относительная величина представляет собой результат деления (сравнения) двух величин. В числителе дроби стоит величина, которую сравнивают, в знаменателе – величина, с которой сравнивают. Последняя называется *базой* (или *основанием*) сравнения. Так, если сопоставить численность населения двух крупнейших городов России – Москвы (10,383 млн чел.) и Санкт-Петербурга (4,661 млн чел.), полученную в результате Всероссийской переписи населения 2002 г., то относительная величина покажет, что численность населения Москвы больше численности населения Санкт-Петербурга в 2,23 раза (10,383/4,661 = 2,23). При этом базой сравнения является численность населения Санкт-Петербурга. Полученная относительная величина выражена в виде коэффициента, который показывает, во сколько раз сравниваемый абсолютный показатель больше базисного. В данном примере база сравнения принята за единицу.

В случае если основание принимается за 100, относительная величина выражается в *процентах* (%), если за 1000 – в *промилле* (‰), если за 10 000 – в *продецимилле* (‱).

Выбор той или иной формы относительной величины зависит от ее абсолютного значения. Если сравниваемая величина больше базы сравнения в 2 раза и более, то обычно выбирают форму коэффициента (как в вышеприведенном примере). Если относительная величина близка к единице, как правило, ее выражают в процентах. Так, сравнив численность безработных в России по состоянию на конец августа 2003 г. (5,68 млн чел.) с их численностью на конец августа 2002 г. (5,203 млн чел.), можно сказать, что численность безработных в августе 2003 г. составила 109,2%

по сравнению с численностью безработных за аналогичный период 2002 г. Если сопоставить численность родившихся в России за 2002 г. со средней численностью населения в 2002 г., получим значение 0,0104. Это число значительно меньше единицы, и его рационально выразить в промилле. Показатель рождаемости 10,4‰ означает, что на 1000 человек населения в 2002 г. рождалось в среднем 10,4 ребенка.

Различают относительные величины структуры, динамики, сравнения, интенсивности и координации.

Относительные величины структуры показывают удельный вес каждой группы в общей численности совокупности. Их получают путем деления численности каждой группы, входящей в совокупность, на численность всей совокупности. Примером таких показателей служат данные последних двух граф в табл. 3.4.

Относительные величины структуры дают возможность сопоставлять структуры одной и той же совокупности в различные моменты времени. Такое сопоставление позволяет делать выводы о тенденциях и закономерностях структурных изменений во времени.

Для примера обратимся к табл. 3.1. Ее данные свидетельствуют, что удельный вес городского населения России увеличился с 15% в 1897 г. до 73% в 2002 г. В то же время данные о структуре городского и сельского населения в последнем десятилетии позволяют сделать вывод о том, что рост удельного веса городского населения в общем населении страны прекратился, более того, наметилась тенденция некоторого роста удельного веса сельского населения.

Весьма часто приходится сопоставлять структуры совокупностей, объем которых различен. Допустим, требуется сравнить структуру населения, занятого в сельском хозяйстве Татарстана и Чувашии в 2000 г. Сопоставление абсолютных показателей – 253,0 тыс. чел. в Татарстане и 154,5 тыс. чел. в Чувашии – не позволяет получить правильный ответ. Для этого нужно сопоставить удельный вес сельскохозяйственного населения в общем объеме занятого населения каждой республики. В Татарстане его удельный вес равен 14,9%, в Чувашии – 25,3%. Для того чтобы представить структурные различия в отдельных частях совокупности, приходится сравнивать структуру совокупности в целом со структурой отдельных ее частей. Например, удельный вес населения, занятого в сельском хозяйстве Приволжского федерального округа, в состав которого входят Татарстан и Чувашия, равен 15,5%. Таким образом, сопоставляя относительные пока-

затели, получаем, что для Татарстана удельный вес указанного населения несколько меньше, чем по региону в целом, а для Чувашии на 63% больше, чем по региону в целом.

Относительные величины динамики – это результат сопоставления уровней одного и того же явления, относящихся к различным периодам или моментам времени. Например, сопоставляя объем добычи нефти в России в 2002 г. и 2000 г., получаем относительную величину динамики:

$$\frac{379{,}6 \text{ млн т}}{323{,}5 \text{ млн т}} 100\% = 117{,}3\%.$$

Она показывает, что добыча нефти в России в 2002 г. была на 17,3% больше, чем в 2000 г.

При определении относительных показателей динамики важно обеспечить сопоставимость показателей, которые участвуют в расчете. Несопоставимость может возникнуть по многим причинам: меняется методология расчета показателей или степень охвата совокупности, показатели относятся к периодам разной продолжительности и т.д.

Например, при исчислении затрат на рабочую силу в одном периоде командировочные расходы не учитывались, а в другом – учитывались. В данном случае методология расчета величин различна, поэтому без соответствующей корректировки они несопоставимы.

Другой пример. Допустим, требуется сравнить объем выпуска промышленной продукции за два периода. Однако если в первом периоде учитывают продукцию только промышленных предприятий, а во втором – также и продукцию, созданную на предприятиях других отраслей и в домашних хозяйствах, то эти величины несопоставимы, так как границы двух совокупностей различны.

Часто возникает необходимость сопоставлять данные, относящиеся к периодам различной продолжительности. Так, для того чтобы сравнить данные февраля с данными января или марта, надо учесть различную продолжительность этих периодов, т.е. прежде чем рассчитывать соответствующие показатели динамики, необходимо абсолютные величины привести к сопоставимому виду – рассчитать среднесуточные показатели января и февраля.

При исчислении относительных показателей динамики следует иметь в виду, что многие явления изменяются под влиянием «сезонной волны». Так, число отдыхающих на курортах Черноморского побережья всегда будет больше в июле, чем в январе, а потребление топлива будет больше в январе, чем в июле. Для

такого круга явлений интерпретировать соответствующие показатели динамики нужно весьма осторожно (подробнее об этом см. в главе 8).

Примыкают к относительным показателям динамики и показатели выполнения плана, по которым судят о ходе реализации различных программ как на национальном и региональном уровнях, так и на уровне фирмы. В этом случае относительная величина получается как результат сопоставления фактической и плановой абсолютных величин, относящихся к одному и тому же периоду.

Относительные величины сравнения получают в результате сопоставления одноименных абсолютных показателей, относящихся к разным совокупностям. Например, сравниваем размер основных фондов пищевой промышленности двух регионов по состоянию на 1 января 2003 г. или уровень потребления в расчете на душу населения жителями Кировской и Ростовской областей в третьем квартале текущего года и т.д. При определении относительных величин сравнения необходимо обеспечить единство методологии исчисления абсолютных показателей, подлежащих сопоставлению.

Относительные величины интенсивности получают, сопоставляя разноименные признаки одной совокупности, а также объекты двух связанных между собой совокупностей.

Примерами такого рода показателей могут служить коэффициент рождаемости (число родившихся в расчете на 1000 человек населения), уровень занятости (отношение числа занятых к численности экономически активного населения). Здесь показатели интенсивности получены как отношение значений различных признаков одной совокупности. Эти показатели обычно выражаются в процентах, промилле и т.д.

К показателям интенсивности, полученным на основе разных совокупностей, относятся плотность населения (число людей, приходящихся на 1 км^2 территории), фондоотдача (стоимость продукции, произведенной на 1 руб. основных фондов) и т.п. В этом случае единицы измерения относительных величин интенсивности определяются показателями, на основе которых они рассчитаны.

Относительные величины координации получают как соотношение между частями одного целого. Примерами такого рода показателей являются соотношение числа мужчин и женщин, отношение численности неработающих лиц к численности занятого населения, отношение стоимости импортных продуктов питания к стоимости отечественного продовольствия и др.

4.3. Средние величины

Среди обобщающих показателей, характеризующих статистические совокупности, большое значение имеют средние величины.

Средняя величина – это обобщающая характеристика множества индивидуальных значений некоторого количественного признака.

Совокупность, изучаемая по количественному признаку, состоит из индивидуальных значений; на них оказывают влияние как общие причины, так и индивидуальные условия. В среднем значении отклонения, характерные для индивидуальных значений, погашаются. Средняя, являясь функцией множества индивидуальных значений, представляет одним значением всю совокупность и отражает то общее, что присуще всем ее единицам.

Практическое применение средних величин как обобщающих характеристик явлений и процессов в природе и обществе чрезвычайно широко.

Можно рассчитать среднемесячную заработную плату работника той или иной профессиональной группы (шахтера, библиотекаря, врача) и среднемесячный денежный доход, который приходится на одного жителя страны, среднюю себестоимость продукции по группе предприятий, выпускающих данный вид продукции, и среднегодовую температуру воздуха в 2003 г. в Москве и т.д.

В приведенных примерах средние величины характеризуют качественно однородные группы изучаемого явления. Разумеется, уровни месячной заработной платы шахтеров в силу различия их квалификации, стажа работы, отработанного за месяц времени и многих других факторов отличаются как друг от друга, так и от уровня средней заработной платы. Однако в среднем уровне отражены основные факторы, которые влияют на уровень заработной платы, и взаимно погашаются различия, которые возникают вследствие индивидуальных особенностей работника. Средняя заработная плата отражает типичный уровень оплаты труда для данного вида работников. Средняя величина в этом случае является не просто обобщающей, но и типической характеристикой совокупности. Получению *типической средней* должен предшествовать анализ того, насколько данная совокупность качественно однородна. Если совокупность состоит из разнокачественных частей, следует разбить ее на типические группы.

Например, если доходы 70% населения сокращаются в несколько раз, доходы 20% увеличиваются в несколько десятков раз и только у 10% населения остаются на прежнем уровне, то, опираясь на такую обобщающую характеристику, как среднедушевой

доход, можно сделать вывод о том, что доходы населения неизменны. Однако полученные средние обобщающие показатели не являются типичными. В этой совокупности ярко выражены противоположные тенденции изменения уровня доходов, поэтому обобщающие показатели следует рассчитать для отдельных ее однородных частей.

Средние величины используются в качестве типических характеристик не только для однородных, но и для неоднородных совокупностей. Например, если рассчитывается потребление крепких спиртных напитков на душу населения, то из всей совокупности населения можно исключить детей в возрасте до 10 лет, не говоря уже о том, что и довольно значительная часть других возрастных групп не потребляет этот продукт.

Средняя величина ВВП на душу населения, средняя величина потребления различных групп товаров на человека и другие подобные величины представляют обобщающие характеристики государства как единой экономической системы и носят название *системных средних.*

Системные средние могут служить обобщающими характеристиками не только разнородных пространственных, но и динамических систем. Пример такой средней – среднегодовая или среднемесячная температура воздуха той или иной местности. Полученную среднюю нельзя рассматривать как типическую величину для года или какого-то месяца. Среднемесячная температура сентября данного года может существенно отличаться от среднемесячной температуры сентября прошлого года. Для вычисления типической средней нужно рассчитать среднюю температуру сентября на основе многолетних наблюдений. В этом случае типическая средняя получается при расчете средней величины из системных средних.

В других случаях из типических средних можно получить системную среднюю. Например, если известны средние значения доходов на душу населения для типических групп, то общая средняя, рассчитанная на основе этих групповых средних, представляет собой системную среднюю.

В статистике используются различные виды (формы) средних величин. Наиболее часто применяются следующие средние величины:

- средняя арифметическая;
- средняя гармоническая;
- средняя геометрическая;
- средняя квадратическая.

Выбор той или иной формы средней зависит от содержания осредняемого признака и конкретных данных, по которым ее приходится вычислять.

Указанные средние величины могут быть вычислены, либо когда каждый вариант в данной совокупности встречается только один раз, при этом средняя называется *простой* или *невзвешенной*, либо когда варианты повторяются различное число раз, при этом число повторений вариантов называется *частотой* или *статистическим весом*, а средняя, вычисленная с учетом весов, – *средней взвешенной*.

Все указанные средние величины можно рассчитать по *формулам средней степенной*:

а) если имеются только варианты $x_1^z, x_2^z, \ldots, x_n^z$ – по формуле средней степенной порядка z

$$\bar{x} = \sqrt[z]{\frac{\sum\limits_i x_i^z}{n}}; \tag{4.1}$$

б) если имеются варианты и частоты $f_1, f_2, \ldots, f_n$ – по формуле средней степенной взвешенной

$$\bar{x} = \sqrt[z]{\frac{\sum\limits_i x_i^z f_i}{\sum\limits_i f_i}}, \tag{4.2}$$

где $\bar{x}$ – средняя степенная;

z – показатель степени, позволяющий определить вид средней;

x_i – вариант;

f_i – частота, или статистический вес, варианта.

Средняя арифметическая – самый распространенный вид средней величины. Следует отметить, что если вид средней величины не указывается, подразумевается средняя арифметическая.

Средняя арифметическая получается при подстановке в формулу степенной средней значения $z = 1$.

Средняя арифметическая простая рассчитывается по формуле

$$\bar{x}_{\text{арифм}} = \frac{\sum\limits_i x_i}{n}, \tag{4.3}$$

а *средняя арифметическая взвешенная* – по формуле

$$\bar{x}_{\text{арифм}} = \frac{\sum\limits_i x_i f_i}{\sum\limits_i f_i}. \qquad (4.4)$$

Пример. Обследование пяти квартир первого этажа жилого дома показало, что в них проживает соответственно 1, 2, 3, 4 и 5 человек.

Определить среднюю арифметическую.

Средняя арифметическая из этих чисел

$$\bar{x}_{\text{арифм}} = \frac{\sum\limits_i x_i}{n} = \frac{1 + 2 + 3 + 4 + 5}{5} = 3 \text{ чел.},$$

т.е. в среднем на одну квартиру первого этажа приходится 3 человека.

Пример. Результаты обследования всех квартир одного подъезда жилого дома приведены в табл. 4.1.

Таблица 4.1

Количество проживающих в квартире, чел. x_i	Количество квартир f_i	$x_i f_i$
1	2	3
1	6	6
2	9	18
3	10	30
4	20	80
5	15	75
Итого	60	209

Вычислить среднее число жителей, проживающих в одной квартире.

Находим среднюю арифметическую взвешенную, предварительно заполнив графу 3 в табл. 4.1:

$$\bar{x}_{\text{арифм}} = \frac{\sum\limits_i x_i f_i}{\sum\limits_i f_i} = \frac{209}{60} = 3{,}48 \text{ чел.},$$

т.е. в среднем на одну квартиру в этом подъезде приходится 3,48 человека.

Средняя арифметическая обладает рядом свойств.

1. Средняя арифметическая постоянной величины a равна этой же постоянной величине:

$$\bar{a} = a.$$

2. Сумма отклонений значений вариантов от средней равна нулю:

$$\sum_i (x_i - \bar{x}) = 0 \text{ (если частоты равны единице);}$$

$$\sum_i (x_i - \bar{x}) f_i = 0 \text{ (если частоты различны).}$$

3. Если из всех вариантов x_i вычесть постоянную величину x_0 и на основе разностей ($x_i - x_0 = x_i'$) вычислить среднюю $\bar{x}'$, то она будет меньше средней исходного ряда на эту постоянную величину. Поэтому, чтобы получить среднюю из исходных вариантов, необходимо к средней $\bar{x}'$ прибавить ту же постоянную величину x_0:

$$\bar{x} = \bar{x}' + x_0.$$

4. Если все варианты x_i разделить на постоянную величину h и из частных ($x_i/h = x_i'$) вычислить среднюю, то она будет меньше средней исходного ряда в h раз. Для того чтобы получить среднюю из исходных вариантов, нужно среднюю $\bar{x}'$ умножить на эту постоянную величину h:

$$\bar{x} = \bar{x}' h.$$

5. Если у всех вариантов x_i частоты f_i равны друг другу ($f_1 = f_2 = \ldots = f_n = k$), то средняя арифметическая взвешенная равна средней арифметической простой:

$$\bar{x} = \frac{\sum_i x_i f_i}{\sum_i f_i} = \frac{\sum_i x_i k}{\sum k} = \frac{k \sum_i x_i}{kn} = \frac{\sum_i x_i}{n}.$$

Использование свойств средней позволяет упростить вычисление средней арифметической.

Упрощенная формула расчета средней арифметической:

$$\bar{x} = \bar{x}' h + x_0. \qquad (4.5)$$

При этом $\bar{x}'$ получено из $x_i' = (x_i - x_0)/h$.

Формула (4.5) применяется в том случае, когда данные сгруппированы и варианты каждой группы представлены интервалом ($x_{i\,\min}$–$x_{i\,\max}$). При этом значение в центре интервала принимается за значение признака у всех единиц в этом интервале.

Упрощенный способ вычисления средней арифметической называется *методом отсчета от условного нуля.*

Пример. Имеется распределение работников цеха предприятия по стажу работы (графы 1 и 2 в табл. 4.2).

Определить среднюю арифметическую по упрощенной формуле.

Сначала по упрощенной формуле находим середину интервалов, т.е. значения центра интервала (графа 3), затем отнимаем от них постоянную величину x_0. В качестве x_0 целесообразно брать значение середины интервала, находящегося в середине ряда. Если в ряду четное число групп (как в нашем примере), из двух срединных интервалов выбирают интервал с бóльшим весом. В данном случае $x_0 = 17{,}5$.

В качестве постоянной величины h при равенстве интервалов берется длина интервала (в нашем случае $h = 5$).

Если интервалы не равны, в качестве h следует брать наибольший общий делитель ряда $x_{i\,\text{ср}}$ либо любое другое число, которое позволяет упростить расчеты.

Таблица 4.2

Стаж работников, лет x_i	Количество работников f_i	Середина интервала $x_{i\,\text{ср}}$	$x_i - x_0$	$\frac{x_i - x_0}{h}$	$\frac{x_i - x_0}{h} f_i$
1	2	3	4	5	6
0–5	12	2,5	–15	–3	–36
5,1–10	16	7,5	–10	–2	–32
10,1–15	23	12,7	–5	–1	–23
15,1–20	28	17,5	0	0	0
20,1–25	17	22,5	5	1	17
25,1–30	14	27,5	10	2	28
Итого	110	–	–	–	–46

Средняя арифметическая рассчитывается по формуле

$$\bar{x} = \frac{\sum_i \left(\frac{x_i - x_0}{h}\right) f_i}{\sum_i f_i} h + x_0 = \frac{-46}{110} 5 + 17{,}5 = -2{,}09 + 17{,}5 = 15{,}41.$$

Средний стаж работников цеха будет равен 15,41 года.

В случае если совокупность разбита на группы и для каждой группы исчислена средняя арифметическая, *общая средняя* для всей совокупности рассчитывается по формуле

$$\bar{x}_{\text{общ}} = \frac{\sum_i \bar{x}_i n_i}{\sum_i n_i}, \qquad (4.6)$$

где $\bar{x}_i$ – средняя арифметическая i-й группы;

n_i – численность i-й группы.

Как видим, переход от частных средних к средней по всей совокупности осуществляется по формуле средней арифметической взвешенной, при этом в качестве вариантов выступают групповые средние, а весом служит численность каждой группы.

Например, если известно, что среднемесячная заработная плата руководителей предприятия составляет 10000 руб., специалистов – 8000 руб., служащих – 5000 руб., рабочих – 7000 руб., а численность этих групп персонала на предприятии соответственно равна 12, 75, 60 и 653 человека, то среднемесячная заработная плата всех работников этого предприятия

$$\bar{x}_{\text{общ}} = \frac{\sum_i \bar{x}_i n_i}{\sum_i n_i} = \frac{10000 \cdot 12 + 8000 \cdot 75 + 5000 \cdot 60 + 7000 \cdot 653}{12 + 75 + 60 + 653} =$$

$$= \frac{5591000}{800} = 6988{,}75 \text{ руб.}$$

Средняя арифметическая – это всегда обобщающая количественная характеристика варьирующего признака совокупности. Отметим, что при вычислении средней арифметической не обязательно знать значение каждого варианта. Часто она вычисляется на основе итоговых данных осредняемого признака для всей совокупности. При этом итоговые данные могут быть получены не только путем суммирования индивидуальных значений признака, возможен также подсчет суммарного значения без фиксации значений вариантов по тем или иным признакам. Так, при подсчете средней себестоимости изделия общий уровень затрат на его производство делится на количество продукции данного вида. При определении средней урожайности, естественно, не фиксируется урожайность каждого гектара посевных площадей, а подсчитывается валовой сбор урожая тех или иных культур, и эта величина делится на площадь, отведенную под эти культуры.

В ряде случаев даже теоретически невозможно определить значение вариантов. Например, если рассчитывается уровень среднедушевого производства ВВП для страны в целом, то невозможно получить объем производства продукции каждым жителем в силу того, что ВВП – это результат деятельности не всего населения страны, а только лиц, занятых в экономике.

Такая средняя величина, по существу, не отличается от относительных величин интенсивности ни по способу расчета, ни по аналитическому значению. Следует подчеркнуть, что между относительными и средними величинами порой трудно провести четкую границу, поскольку средняя величина – это отношение двух абсолютных величин, т.е. относительная величина.

Средняя гармоническая величина получается при подстановке в формулу степенной средней значения $z = -1$.

Формула *средней гармонической простой* такова:

$$\bar{x}_{\text{гарм}} = \sqrt[-1]{\frac{\sum_i x_i^{-1}}{n}} = \frac{n}{\sum_i \frac{1}{x_i}}. \qquad (4.7)$$

Средняя гармоническая взвешенная определяется по формуле

$$\bar{x}_{\text{гарм}} = \sqrt[-1]{\frac{\sum_i x_i^{-1} V_i}{\sum_i V_i}} = \frac{\sum_i V_i}{\sum_i \frac{V_i}{x_i}}, \qquad (4.8)$$

где V_i – веса для обратных значений x_i.

Средняя гармоническая вычисляется в тех случаях, когда приходится суммировать не сами варианты, а обратные им величины:

$$1/x_1,\ 1/x_2,\ \ldots,\ 1/x_n.$$

Пример. В ходе торгов на валютной бирже за первый час работы заключено пять сделок. Данные о сумме продажи рублей и курсе доллара США приведены в табл. 4.3 (графы 2 и 3).

Определить средний курс доллара США на первый час торгов.

Таблица 4.3

Номер сделки	Сумма продажи V_i, млн руб.	Курс доллара x_i, руб. за 1 долл.	$\frac{V_i}{x_i}$, млн долл.
1	2	3	4
1	197,4	28,2	7,0
2	142,0	28,4	5,0
3	228,0	28,5	8,0
4	114,8	28,7	4,0
5	144,0	28,8	5,0
Итого	826,2	—	29,0

Для того чтобы определить средний курс доллара, нужно найти соотношение между суммой продажи рублей, которые затрачены на покупку долларов в ходе всех сделок, и суммой приобретенных в результате этих сделок долларов.

Если итоговую сумму продажи можно получить на основе данных, приведенных в графе 2 (см. табл. 4.3), то данные о количестве купленных долларов в таблице отсутствуют. Однако их можно получить, рассчитав для каждой сделки отношение суммы продажи к курсу доллара (графа 4).

Следовательно, в ходе пяти сделок куплено 29 млн долл. При этом

$$\bar{x}_{\text{гарм}} = \frac{\sum_i V_i}{\sum_i \frac{V_i}{x_i}} = \frac{826,2}{29,0} = 28,49 \text{ руб.},$$

т.е. средний курс доллара составил 28,49 руб.

В данном случае для расчетов использована формула средней гармонической взвешенной.

В практике расчетов довольно часто встречаются ситуации, когда данные о весах признака отсутствуют, но известны варианты осредняемого признака и произведение значений этих вариантов на количество единиц, обладающих этим значением (например, стоимость товарооборота по отдельным товарным группам и индексы цен по этим группам, валовые сборы зерновых по регионам и средняя урожайность по этим регионам и т.д.). В этих случаях средние значения необходимо рассчитывать по формуле средней гармонической.

Средняя геометрическая получается при подстановке в формулу степенной средней значения $z = 0$:

$$\bar{x}_{\text{геом}} = \sqrt[0]{\frac{\sum x^0}{n}} = \left(\sum \frac{1}{n}\right)^{1/0} = \left(\frac{n}{n}\right)^{\infty} = 1^{\infty}.$$

Для раскрытия неопределенностей этого вида прологарифмируем обе части формулы степенной средней [см. формулу (4.1)]:

$$\ln \bar{x} = \frac{1}{z}(\ln \sum x^z - \ln n) = (\ln \sum x^z - \ln n)/z.$$

Подставляя в правую часть равенства $z = 0$, получаем неопределенность вида $\frac{0}{0}$. Используя правило Лопиталя и дифференцируя отдельно числитель и знаменатель по переменной z, получаем

$$\lim_{z \to 0}(\ln \bar{x}) = \lim_{z \to 0} \frac{\sum x^z \ln x}{1 \cdot \sum x^z} = \frac{\sum \ln x}{n}.$$

Следовательно, при $z = 0$

$$\ln \bar{x} = \frac{\sum \ln x}{n}.$$

Потенцируя, находим

$$\bar{x}_{\text{геом}} = \sqrt[n]{x_1 x_2 \cdots x_n} = \sqrt[n]{\prod_{i=1}^{n} x_i}. \qquad (4.9)$$

Формула (4.9) является формулой *средней геометрической простой*. Если использовать частоты f, получим формулу *средней геометрической взвешенной*:

$$\bar{x}_{\text{геом}} = \sqrt[\sum f]{x_1^{f_1} x_2^{f_2} \cdots x_n^{f_n}} = \sqrt[\sum f]{\prod_{i=1}^{n} x_i^{f_i}}. \qquad (4.10)$$

Средняя геометрическая используется для анализа динамики явлений и позволяет определить средний коэффициент роста. При расчете средней геометрической индивидуальные значения признака обычно представляют собой относительные показатели динамики, построенные в виде цепных величин, как отношения каждого уровня ряда к предыдущему уровню (см. главу 8).

Средняя квадратическая получается при подстановке в формулу степенной средней $z = 2$.

Для несгруппированных данных используют формулу *средней квадратической простой*:

$$\bar{x}_{\text{кв}} = \sqrt{\frac{\sum\limits_i x_i^2}{n}} = \sqrt{\frac{x_1^2 + x_2^2 + \ldots + x_n^2}{n}}, \qquad (4.11)$$

для сгруппированных данных – формулу *средней квадратической взвешенной*:

$$\bar{x}_{\text{кв}} = \sqrt{\frac{\sum\limits_i x_i^2 f}{\sum\limits_i f}} = \sqrt{\frac{x_1^2 f_1 + x_2^2 f_2 + \ldots + x_n^2 f_n}{f_1 + f_2 + \ldots + f_n}}. \qquad (4.12)$$

Средняя квадратическая применяется, когда изучается вариация признака. В качестве вариантов используются отклонения фактических значений признака либо от средней арифметической, либо от заданной нормы (см. главу 5).

Средняя арифметическая, средняя гармоническая, средняя геометрическая и средняя квадратическая, рассчитанные для одного и того же ряда вариантов, отличаются друг от друга: их численные значения возрастают с ростом показателя степени z в формуле степенной средней, т.е.

$$\bar{x}_{\text{гарм}} < \bar{x}_{\text{геом}} < \bar{x}_{\text{арифм}} < \bar{x}_{\text{кв}}.$$

Это так называемое *правило мажорантности средних*, которое впервые сформулировал профессор А.Я. Боярский.

Глава 5
АНАЛИЗ ВАРИАЦИОННЫХ РЯДОВ

Как уже отмечалось, единицы изучаемой совокупности обладают различными признаками. Для каждой единицы совокупности определенный признак принимает различные значения, т.е. имеет некоторую вариацию.

Вариацией признака называется наличие различий в численных его значениях у отдельных единиц совокупности.

Чтобы выявить характер распределения единиц совокупности по варьирующим признакам, определить закономерности в этом распределении, строят ряды распределения единиц совокупностей по какому-либо варьирующему признаку.

Ряды распределения, построенные по количественному признаку, называются ***вариационными.***

Форма статистических распределений может быть разнообразной. В одних случаях значения признака концентрируются возле некоторого центра распределения очень тесно, в других случаях наблюдается значительное рассеяние, хотя средние величины могут быть одинаковыми. В связи с этим необходимо определить характер рассеяния признака.

С этой целью решают следующие задачи. Во-первых, определяют меру вариации, т.е. количественно измеряют степень колеблемости признака. Это позволяет сравнивать различные совокупности между собой по степени рассеяния и отслеживать уровень вариации признака одной и той же совокупности в различные периоды.

Во-вторых, для изучения изменчивости признаков выясняют причины, вызывающие вариацию, что предполагает исследование закономерностей вариации в статистических совокупностях.

Для описания статистических распределений обычно используются следующие четыре вида характеристик (показателей):

1) средние, или характеристики центральной тенденции;
2) характеристики вариации (рассеяния);
3) характеристики дифференциации и концентрации;
4) характеристики формы распределения.

Для каждой характеристики существуют различные способы исчисления, выбор которых диктуется условиями задачи.

5.1. Построение и графическое изображение вариационных рядов

Построение вариационных рядов

По своей конструкции вариационный ряд состоит из двух столбцов (граф): один столбец – значения варьирующего признака (x – варианты), другой – частоты (m – абсолютное число случаев данного варианта) или частости (w – относительная доля каждой частоты в общей сумме частот)*.

Вариационные ряды по способу построения бывают двух видов: дискретные и интервальные.

Дискретный ряд распределения можно рассматривать как такое преобразование ранжированного (упорядоченного) ряда, при котором перечисляются отдельные значения признака и указывается их частота. Примером дискретного ряда может служить распределение домашних хозяйств по числу их членов, представленное в табл. 5.1.

Таблица 5.1

Распределение домашних хозяйств России по числу совместно проживающих их членов в 2002 г. (на 1000 домашних хозяйств)

Число членов домашних хозяйств, чел. x_i	Число домашних хозяйств m_i
1	223
2	276
3	238
4	170
5	58
6 и более	35
Итого	1000

Общая схема ряда распределения такова: в совокупности, состоящей из N единиц, некоторая переменная величина x_i (т.е. какой-либо варьирующий признак) принимает различные значения, а каж-

* В предыдущей главе частоты (m) и частости (w) при расчете взвешенных показателей обозначены одним символом f.

дое из этих значений имеет частоту m_i. Исходя из этого, дискретный ряд распределения можно представить следующим образом:

Вариант x_i	Частота m_i
x_1 x_2 ⋮ x_n	m_1 m_2 ⋮ m_n
Итого	$\sum_i m_i$ (или N)

Однако приведенная схема вариационного ряда применима лишь для тех случаев, когда варьирующий признак может принимать небольшое количество значений, т.е. когда число вариантов невелико. Если же вариантов много, невозможно образовать группы для каждого из них. Число групп не должно превышать 12–15 (при достаточно большом числе наблюдений, например свыше 500), в противном случае вариационный ряд становится слишком громоздким.

Если число вариантов велико или признак имеет непрерывную вариацию, то объединение отдельных наблюдений в группы возможно лишь на базе интервала, т.е. такой группы, которая имеет определенные пределы значений варьирующего признака. Эти пределы обозначаются двумя числами, они указывают верхнюю и нижнюю границы, т.е. значение, с которого начинается данная группа, и значение, на котором она заканчивается. При использовании интервалов образуются ***интервальные ряды распределения***. Строя интервальный вариационный ряд, определяют прежде всего число групп, на которые следует разбить всю совокупность. Чем больше групп, тем у́же будет интервал и тем точнее описание распределения. Однако слишком большое число групп затрудняет понимание характера вариации. Вопрос о числе групп следует решать в каждом случае особо в зависимости от изучаемого объекта, объема совокупности. Чаще всего строят вариационные ряды из 7–10 групп.

Как уже отмечалось, интервалы могут быть закрытые и открытые. *Закрытые* интервалы ограничены с обеих сторон, т.е. имеют границу как нижнюю («от»), так и верхнюю («до»). *Открытые* интервалы имеют какую-либо одну границу: либо верхнюю, либо нижнюю. Наличие открытых интервалов хотя и нежелательно, но тем не менее почти неизбежно, так как ради компактности ряда все крайние случаи необходимо сводить в одну группу. Однако,

признавая неизбежность образования открытых интервалов, следует подчеркнуть, что они не должны включать в себя значительную часть общего числа наблюдений, иначе описание всего распределения будет недостаточно точным.

Как обозначают границы интервалов? Строго говоря, требуется, чтобы верхняя граница данного интервала несколько отличалась от нижней границы следующего за ним интервала, как, например, в табл. 4.2. Однако это правило часто не соблюдается, более того, иногда его даже и не следует соблюдать, чтобы не создавать трудности в понимании границ интервалов. В таких случаях, особенно при исследовании непрерывно варьирующего признака, можно использовать интервалы, в которых как нижние, так и верхние границы выражены круглыми числами. Правда, если верхняя граница одного интервала совпадает с нижней границей следующего интервала, остается неясным, в какой интервал попали пограничные случаи. Поэтому всегда необходимо уточнять, как понимаются границы интервалов: включительно или исключительно. Сравним два варианта записи интервалов:

Вариант 1	Вариант 2
От 20 до 30	От 20 до 30
От 30 до 40	От 30 до 40
40 и выше	Свыше 40

Здесь верхняя граница первого интервала совпадает с нижней границей второго интервала и т.д. К какому интервалу отнести пограничные значения? Вопрос может быть решен двояко: во-первых, по принципу «включительно», когда единицы совокупности, имеющие значение 30, относятся к первому (предыдущему) интервалу; во-вторых, по принципу «исключительно», когда единицы относятся ко второму (последующему) интервалу. Судя по последнему интервалу, в варианте 1 принят принцип «исключительно», так как единицы совокупности, имеющие значение 40, попали в последнюю группу, а не в предшествующую, а в варианте 2 принят принцип «включительно», так как в последней группе представлены единицы совокупности со значениями, превышающими 40. Принцип «включительно» встречается чаще.

Построим вариационный ряд и рассчитаем основные его характеристики на основе данных об активах (в млрд руб.) 50 круп-

нейших коммерческих банков России по состоянию на 1 июля 2003 г.

Для построения интервального вариационного ряда ранжируем значения признака в порядке убывания (первые 10 банков (после Сбербанка России) приведены с названиями, остальные только пронумерованы):

Сбербанк России*	1322,7								
1. Внешторгбанк	228,7	11	54,3	21	23,0	31	17,2	41	13,3
2. Альфа-банк	187,3	12	51,8	22	22,2	32	17,1	42	13,3
3. Газпромбанк	180,7	13	45,5	23	21,8	33	17,0	43	13,0
4. Международный промбанк	140,1	14	38,6	24	21,6	34	15,8	44	12,3
5. Банк Москвы	110,9	15	32,4	25	21,5	35	15,7	45	11,6
6. МДМ-банк	108,0	16	32,1	26	20,9	36	15,1	46	11,5
7. Росбанк	81,3	17	30,3	27	18,8	37	15,0	47	11,4
8. Международный Московский банк	73,0	18	30,1	28	18,8	38	14,7	48	11,3
9. Уралсиб	62,1	19	27,2	29	17,5	39	14,1	49	11,2
10. Промышленно-строительный банк	61,0	20	26,9	30	17,4	40	14,0	50	10,9

* В связи с тем, что размер активов Сбербанка сильно отличается от остальных значений и может существенно исказить рассчитываемые средние величины, данную единицу совокупности мы не будем использовать при построении вариационного ряда.

Найдем максимальное и минимальное значения признака в ряду. В рассматриваемой совокупности $x_{min} = 10{,}9$, $x_{max} = 1322{,}7$. Численность совокупности невелика: $N = 51$ единица. Для определения числа групп, на которые будем делить совокупность, воспользуемся формулой Стерджесса:

$$k = 1 + 3{,}322 \lg N = 1 + 3{,}322 \lg 51 \cong 7.$$

По формуле Стерджесса можно определить и длину интервала h, если отбросить «аномальное» значение активов у Сбербанка России ($x_{max} = 1322{,}7$) и построить ряд с равными интервалами. Тогда

$$h = \frac{x_{max} - x_{min}}{k} = \frac{228{,}7 - 10{,}9}{7} \cong 30.$$

Примечание. При формировании первого интервала от минимального значения следует отступить на половину длины интервала, а не рассчитывать данный интервал как $x_{min} + h$.

В соответствии с формулой Стерджесса получим распределение 50 банков по величине активов, приведенное в табл. 5.2.

Таблица 5.2

Распределение 50 банков по величине активов

Величина активов, млрд руб. $x_{i-1}-x_i$	Число банков m_i
До 30	32
30,1–60	8
60,1–90	4
90,1–120	2
120,1–150	1
150,1–180	0
Свыше 180	3
Итого	50

Применение формулы Стерджесса не всегда дает хорошие результаты, что видно из приведенного в табл. 5.2 распределения, где почти половина единиц совокупности оказалась в первом интервале. При значительном разбросе значений можно получить приемлемое распределение, если брать не равные интервалы, а последовательно возрастающие. При этом сохраняется информация о единицах совокупности с «аномальными» значениями. В соответствии с вышесказанным образуем новые интервалы и подсчитаем численность объектов в каждом интервале абсолютно (m) и относительно (w). Полученный интервальный вариационный ряд запишем в виде таблицы (графы А и 1, 2 табл. 5.3).

Для анализа структуры совокупности и расчета обобщающих характеристик дополним табл. 5.3 несколькими колонками (графами), в которых покажем такие элементы вариационного ряда, как середина интервала, накопленная частота и накопленная частость, плотность распределения.

Середину (центр) каждого интервала находят как полусумму нижнего и верхнего значений интервала (см. графу 3 табл. 5.3). В нашем примере центральные варианты будут такими: $\frac{10 + 12}{2} = 11$; $\frac{12 + 15}{2} = 13{,}5$; $\frac{15 + 20}{2} = 17{,}5$ и т.д. (Одна десятая в начале каждого интервала не учитывается, она указывает лишь на то, что интервал читается следующим образом: свыше нижнего значения интервала до верхнего включительно.)

Таблица 5.3

Группировка 50 крупнейших коммерческих банков России по величине активов на 1 июля 2003 г.

Величина активов, млрд руб. $x_{k-1}-x_k$	Количество банков		Середина интервала x_i	$x_i m_i$	Накопленные		Плотность распределения $y_i = \frac{w_i}{h_i}$	Доля активов групп банков в общей сумме активов		$\left(\frac{x_i m_i}{\sum_i x_i m_i}\right)^2$
	единиц m_i	% к итогу w_i			частоты F_i	частости p_i		$\frac{x_i m_i}{\sum_i x_i m_i}$	нарастающим итогом q_i	
А	1	2	3	4	5	6	7	8	9	10
10,1–12	6	12	11	66	6	12	6	0,029	0,029	0,001
12,1–15	8	16	13,5	108	14	28	5,3	0,047	0,076	0,002
15,1–20	10	20	17,5	175	24	48	4	0,076	0,152	0,006
20,1–30	8	16	25	200	32	64	1,6	0,087	0,239	0,008
30,1–50	6	12	40	240	38	76	0,6	0,105	0,344	0,011
50,1–100	6	12	75	450	44	88	0,24	0,197	0,541	0,039
100,1–250	6	12	175	1050	50	100	0,08	0,459	1,000	0,210
Итого	50	100		2289				1,000		0,277

Что касается открытых интервалов, то длина первого интервала приравнивается условно к длине второго, а центральным вариантом последнего интервала обычно служит сумма его нижнего значения и половины предпоследнего интервала.

Любое распределение можно охарактеризовать с помощью накопленных частот. ***Накопленная частота*** показывает число единиц совокупности, у которых значение варианта не больше данного. Накопленная частота для данного варианта или для верхней границы данного интервала получается суммированием (накапливанием) частот всех предшествующих интервалов, включая данный (см. графу 5 табл. 5.3).

Если вместо абсолютных частот использовать частости, то аналогично получим ***накопленные частости*** (см. графу 6 табл. 5.3). Ряд частостей обычно применяют, когда совокупность очень велика. Кроме того, они позволяют сравнивать распределения по одному и тому же признаку в разных по численности совокупностях. В графе 7 табл. 5.3 найдена относительная плотность распределения, которую используют для приведения частостей, относящихся к интервалам разной длины, к сопоставимому виду. Можно рассчитать как абсолютную, так и относительную плот-

ность распределения. ***Абсолютная плотность распределения*** – это частота, приходящаяся на единицу длины интервала, т.е. $\frac{m_i}{h_i}$, а ***относительная плотность распределения*** – частость, приходящаяся на единицу длины интервала, т.е. $\frac{w_i}{h_i}$, где h_i – длина i-го интервала.

Плотность распределения используется в рядах с неравными интервалами для расчета такой характеристики, как мода (см. параграф 5.2), или для графического изображения вариационного ряда в виде гистограммы.

Графическое изображение вариационных рядов

Графически вариационный ряд можно изобразить, как и любой ряд значений аргумента и функции, используя прямоугольную систему координат и строя точки с координатами (x_1, m_1), (x_2, m_2), ..., (x_n, m_n). Если затем последовательно соединить полученные точки отрезками прямой, а из первой и последней точки опустить перпендикуляры на ось x, получим замкнутую фигуру в виде многоугольника, которая называется ***полигоном*** и графически представляет распределение совокупности по признаку x. Полигон чаще используется для дискретных вариационных рядов. На рис. 5.1 представлен полигон распределения домашних хозяйств по числу их членов (см. табл. 5.1).

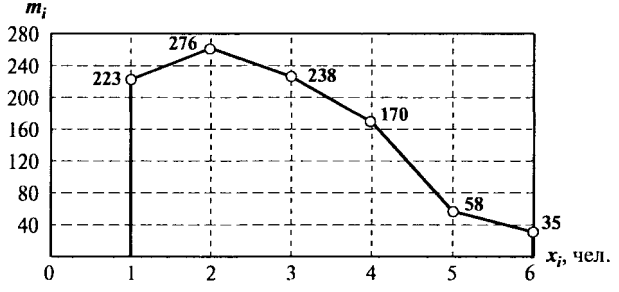


Рис. 5.1. Полигон распределения

Интервальный вариационный ряд изображают в виде ***гистограммы***. Для интервального ряда с равными интервалами на оси x откладывают отрезки, равные длине интервала. На этих отрезках, как на основаниях, строят прямоугольники, высота которых пропорциональна частоте или частости. Для интервального ряда с неравными интервалами на оси ординат отклады-

вают плотности распределения, так как в этом случае именно плотность дает представление о заполненности каждого интервала. На рис. 5.2 изображена гистограмма распределения банков по величине активов (см. табл. 5.3), построенная по относительной плотности распределения.

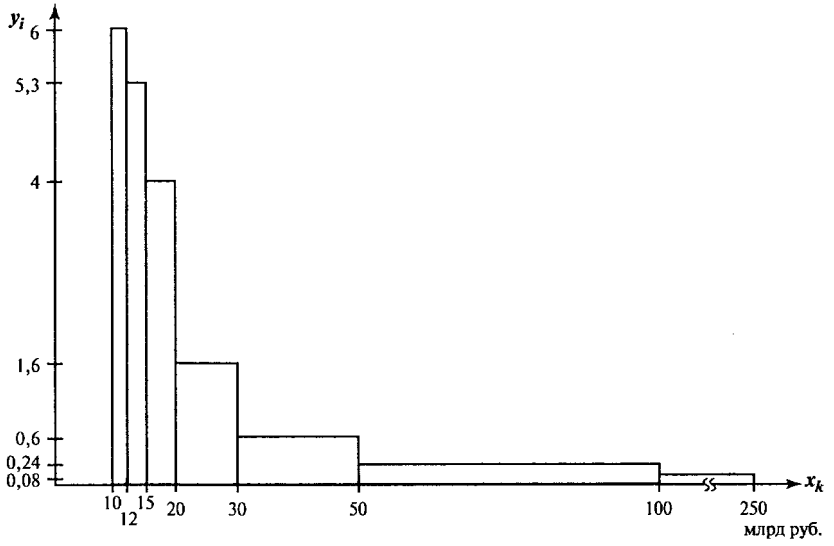


Рис. 5.2. Гистограмма

Площадь всей гистограммы численно равна сумме частот, или численности единиц в совокупности (если на оси ординат отложить частоты).

Любой вариационный ряд можно представить графически в виде кривой накопленных частот (или частостей). При этом на оси x откладывают варианты или верхние границы интервалов, а на оси y – соответствующие накопленные частоты (или частости). Полученные точки соединяют для непрерывного признака плавной кривой, которая называется ***кумулятивной кривой***, или ***кумулятой***. Если значения x (варианты) откладывать на оси y, а накопленные частоты (или частости) на оси x, то построенная на них кумулятивная кривая называется ***огивой***.

На рис. 5.3 представлена кумулята распределения банков по величине активов (см. табл. 5.3). Кумулята имеет начальную точку на оси x с координатами $(x_0, 0)$, где x_0 – нижняя граница первого интервала. Это означает, что в нашей совокупности нет ни одного банка с активами 10 млрд руб. и менее.

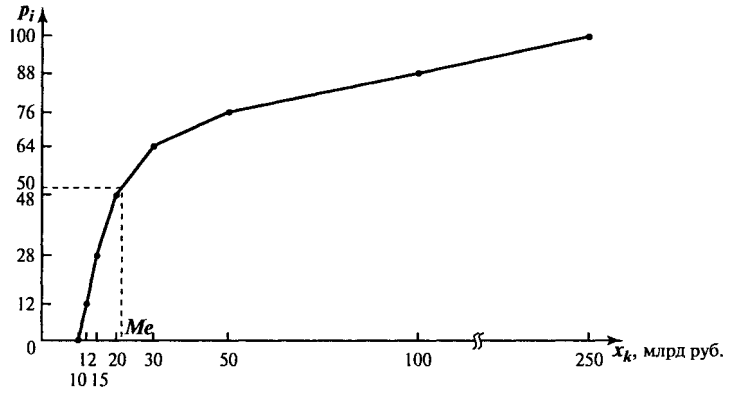


Рис. 5.3. Кумулятивная кривая

Ряд накопленных частот по сравнению с первоначальным рядом распределения обладает некоторыми преимуществами. Например, длина интервала для такого ряда имеет уже второстепенное значение.

Иногда при исследовании вариационных рядов нас интересует параллельное изменение нарастающих долей единиц совокупности и нарастающих долей значений признака в общем объеме. Такая задача возникает при изучении концентрации какого-либо признака в тех или иных группах совокупности. В этих случаях для анализа концентрации строят так называемую ***кривую Лоренца*** (рис. 5.4).

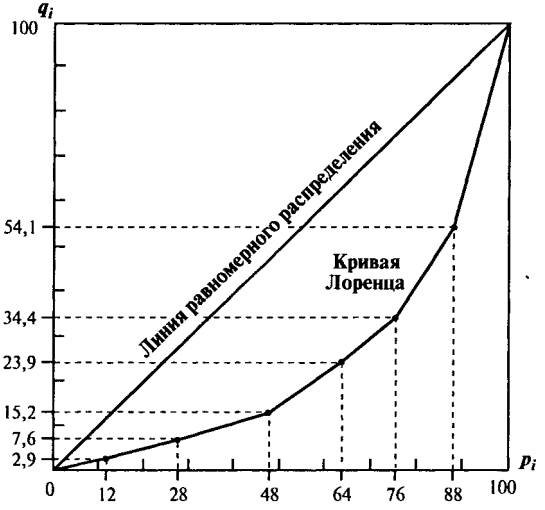


Рис. 5.4. График Лоренца

По оси абсцисс откладывают накопленные частости, характеризующие распределение единиц совокупности (p_i), по оси ординат – кумулятивные доли значений признака в общем объеме (q_i). Так, на рис. 5.4 представлена кривая Лоренца распределения активов по крупнейшим банкам России (см. графы 6 и 9 табл. 5.3).

5.2. Основные показатели среднего уровня вариационного ряда

Вычисление средней арифметической

При изучении особенностей статистического распределения прежде всего следует найти его центральное значение, т.е. средний уровень. Для характеристики центра распределения применяются показатели, получившие название *средних величин.*

Как уже отмечалось, в статистике применяются различные виды (формы) средних величин. Форму средней выбирают исходя из экономической сущности осредняемого признака. Самый распространенный вид средних – средняя арифметическая: простая или взвешенная.

Средняя арифметическая простая

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{1}{n}\sum_i x_i \qquad (5.1)$$

применяется, когда объем совокупности варьирующего признака представляет сумму всех индивидуальных значений.

Расчет средней арифметической простой возможен, например, на основе несгруппированных данных об активах 50 банков (см. с. 84):

$$\bar{x} = \frac{228{,}7 + 187{,}3 + \ldots + 10{,}9}{50} = \frac{2081{,}3}{50} = 41{,}6 \text{ млрд руб.}$$

Для построенного интервального вариационного ряда (см. табл. 5.3) расчет средней арифметической должен быть выполнен по формуле средней арифметической взвешенной, где совокупный объем активов 50 банков находится не путем суммирования всех значений признака, а путем перемножения (взвешивания) вариантов признака на их частоты и последующего сложения произведений $x_i m_i$, число которых (произведений) равно количеству интервалов. Следовательно, взвешивание – это лишь технический прием, посредством которого суммирование

одинаковых значений заменяется умножением этих значений на их частоты. В основе взвешивания лежит равенство

$$x_1 m_1 + x_2 m_2 + \ldots + x_n m_n = \bar{x} m_1 + \bar{x} m_2 + \ldots + \bar{x} m_n = \\ = \bar{x}(m_1 + m_2 + \ldots + m_n).$$

Отсюда ***средняя арифметическая взвешенная***

$$\bar{x} = \frac{\sum\limits_i x_i m_i}{\sum\limits_i m_i}. \qquad (5.2)$$

В отдельных случаях веса могут быть представлены не абсолютными, а относительными величинами (в процентах или долях единицы). При этом упрощаются расчеты, так как $\sum\limits_i w_i$ составляет единицу, или 100%. При замене частот на частости средняя величина характеристики не изменится, а формула $\bar{x}$ примет следующий вид:

$$\bar{x} = \frac{\sum\limits_i x_i w_i}{\sum\limits_i w_i}.$$

Обычно формулу средней арифметической взвешенной записывают в виде

$$\bar{x} = \frac{\sum\limits_i x_i f_i}{\sum\limits_i f_i}, \qquad (5.3)$$

где f_i – веса, в роли которых могут выступать и частоты, и частости.

В формулах средней арифметической взвешенной, рассчитываемой для интервального вариационного ряда, в качестве x_i принято брать середину интервала, исходя из предположения о равномерном распределении единиц совокупности на данном интервале. Середину интервала найдем как полусумму значений его нижней и верхней границ (при условии, что верхняя граница данного интервала совпадает с нижней границей следующего интервала). Тогда на основе данных графы 4 табл. 5.3 средняя арифметическая получит значение

$$\bar{x} = \frac{\sum\limits_i x_i m_i}{\sum\limits_i m_i} = \frac{2289}{50} = 45{,}78 \text{ млрд руб.}$$

Несовпадение средних арифметических, вычисленных на основе исходных данных, приведенных на с. 84, и на основе вариационного ряда (41,63 < 45,78), вызвано тем, что сделанное нами допущение о равномерном распределении значений на интервале не всегда выполняется. Средние, исчисленные на основе интервального ряда, являются приближенными. Степень точности зависит от того, в какой мере распределение единиц внутри интервала приближается к такому распределению, для которого средняя арифметическая взвешенная совпадает с серединой интервала. Точность средней зависит также от длины интервала. Чем у́же интервал, тем меньше ошибка, вызванная тем, что середина интервала принимается в качестве среднего его значения (см. графу 3 табл. 5.3). При неравных интервалах точность средней меньше, чем при равных.

Средняя арифметическая обладает рядом свойств (см. с. 73).

Поскольку средняя арифметическая вычисляется как отношение суммы значений x_i к их общей численности, то она никогда не выходит за пределы этих значений, а находится между минимальным и максимальным значениями x_i. При увеличении или уменьшении каждого значения x_i средняя арифметическая также увеличивается или уменьшается.

Частоты (веса) вариационного ряда показывают, сколько раз повторяется каждое значение осредняемого признака, положенного в основу группировки. Однако в случае, когда единица измерения признака, положенного в основу группировки, не совпадает с единицами измерения элемента совокупности, частоты вариационного ряда не могут служить весами для определения средней. Примером такого вариационного ряда может служить приведенное в табл. 5.4 распределение 20 фермерских хозяйств по урожайности зерновых (см. графы 1 и 2).

Таблица 5.4

Распределение 20 хозяйств по урожайности зерновых культур

Урожайность, ц/га x_i	Число хозяйств m_i	Посевная площадь под зерновыми по группам хозяйств, га S_i	Валовой сбор зерновых, ц $x_i S_i$
1	2	3	4
25	4	50	1250
27	5	100	2700
28	8	220	6160
30	3	130	3900
Итого	20	500	14010

При расчете средней урожайности зерновых по всем 20 хозяйствам в качестве весов для средней арифметической взвешенной надо брать не число хозяйств в каждой группе как частоты, а их посевную площадь под зерновыми S_i (см. графу 3 табл. 5.4). Тогда, умножая урожайность на посевную площадь, получим валовой сбор зерновых по группам хозяйств и в целом $\left(\sum_i x_i S_i\right)$ (см. графу 4 табл. 5.4), который необходим для расчета средней урожайности зерновых:

$$\bar{x} = \frac{\sum_i x_i f_i}{\sum_i f_i} = \frac{\sum_i x_i S_i}{\sum_i S_i} = \frac{14010}{500} = 28{,}02 \text{ ц/га}.$$

Рассмотрим еще один пример: распределение предприятий оптовой торговли по объему товарооборота, где одновременно дано и распределение численности работников по этим группам предприятий (табл. 5.5).

Таблица 5.5

Группировка предприятий оптовой торговли по объему оптового товарооборота по состоянию на 1 мая 1995 г.

Объем оптового товарооборота за апрель 1995 г., тыс. руб.	Число предприятий		Среднесписочная численность работников		
	тыс. единиц	% от общего числа предприятий	тыс. чел.	% от общей численности работников	на одно предприятие, чел.
	m_i	w_i	T_i	$\frac{T_i}{\sum_i T_i}100\%$	T_i/m_i
А	1	2	3	4	5
Менее 1	15,620	37,5	69,6	10,5	4
1–25	9,362	22,5	71,4	10,8	8
25–50	3,633	8,7	37,3	5,6	10
50–100	3,618	8,7	52,2	7,9	14
100–200	3,261	7,8	61,9	9,3	19
200–500	3,034	7,3	88,8	13,4	29
Свыше 500	3,100	7,5	283,1	42,5	91
Итого	41,628	100,0	664,3	100,0	16

Для расчета среднего товарооборота, приходящегося на одно предприятие, весами будут служить частоты вариационного

ряда – число предприятии, а вариантами – средние значения интервалов:

$$\bar{x} = \frac{\sum_i x_i m_i}{\sum_i m_i} = \frac{0{,}5 \cdot 15{,}62 + 13 \cdot 9{,}362 + 37{,}5 \cdot 3{,}633}{41{,}628} +$$

$$+ \frac{75 \cdot 3{,}618 + 150 \cdot 3{,}261 + 350 \cdot 3{,}034 + 650 \cdot 3{,}1}{41{,}628} =$$

$$= \frac{4238{,}8285}{41{,}628} = 101{,}8 \text{ тыс. руб. (деноминированных).}$$

По данным табл. 5.5 можно рассчитать и средний товарооборот, приходящийся на одного работника, для чего найдем соотношение объема товарооборота и числа работающих:

$$\bar{x} = \frac{\sum_i x_i m_i}{\sum_i T_i} = \frac{4238{,}8285}{664{,}3} = 6380{,}9 \text{ руб./чел.}$$

Наряду с этой общей средней для аналитических целей можно найти в каждой группе предприятий объем товарооборота, приходящийся на одного работника, т.е. частные или групповые средние как показатели эффективности для мелких, средних и крупных предприятий.

Для оценки размера предприятий оптовой торговли найдем среднесписочное число работников, приходящееся на одно предприятие:

$$\bar{T} = \frac{\sum_i T_i}{\sum_i m_i} = \frac{664{,}3}{41{,}628} \cong 16 \text{ чел./предприятие.}$$

Эту общую среднюю можно дополнить групповыми средними (см. графу 5 табл. 5.5).

С помощью средних обобщаются не только абсолютные, но и относительные значения варьирующего признака. Способ расчета *средних из относительных величин* зависит от того, какие относительные величины обобщаются и какие данные известны. Однако общее определение средней арифметической сохраняет силу и в этом случае. При вычислении таких средних необходи-

мо, чтобы сохранялось суммарное значение объемного признака $\left(\sum_i x_i m_i\right)$, т.е. чтобы числитель в формуле средней арифметической был неизменным. В качестве весов при расчете средней арифметической относительного показателя необходимо принять значения того признака, который является знаменателем при определении относительного показателя.

Рассмотрим пример расчета среднего удельного веса женщин в численности экономически активного населения (табл. 5.6).

Таблица 5.6

Распределение по группам численности экономически активного населения России в 2002 г.

	Млн чел. m_i	Доля женщин в группе x_i	Численность женщин в группе, млн чел. $x_i m_i$
А	1	2	3
Экономически активное население:			
занятые в экономике	65,766	0,489	32,151
безработные	6,153	0,460	2,831
Всего	71,919	0,486	34,982

Средний удельный вес (доля) женщин в экономически активном населении

$$\bar{x} = \frac{0{,}489 \cdot 65{,}766 + 0{,}46 \cdot 6{,}153}{65{,}766 + 6{,}153} = \frac{34{,}982}{71{,}919} = 0{,}486 \text{ (или } 48{,}6\%).$$

Средняя доля ближе по значению к доле занятых в экономике, так как они преобладают в экономически активном населении. Числитель средней величины $\sum_{i=1}^{2} x_i m_i$ – это общая численность женщин в обеих группах, т.е. техника расчета средней из относительных величин аналогична технике расчета средней из групповых средних. Отметим, что средняя из относительных величин остается относительной величиной.

Вычисление средней гармонической

Кроме средней арифметической в статистике используется и средняя гармоническая, как простая $\bar{x} = \dfrac{n}{\sum\limits_i \dfrac{1}{x_i}}$, так и взвешенная $\bar{x} = \dfrac{\sum\limits_i V_i}{\sum\limits_i \dfrac{V_i}{x_i}}$, где V_i – веса для обратных значений x_i.

Рассмотрим примеры вычисления средней гармонической.

Пример. Найти средний процент изменения объема производства продукции за восемь месяцев 1997 г. на основе представленных в табл. 5.7 данных (графа 3).

Таблица 5.7

Производство минеральных удобрений предприятиями химической промышленности России в 1997 г.

Годовая мощность выпуска минеральных удобрений, тыс. т	Количество предприятий	Произведено продукции в январе–августе 1997 г.	
		тыс. т V_i	% к январю–августу 1996 г. x_i
А	1	2	3
Менее 100	6	27	133,7
100–500	11	957	103,5
500–1000	8	1883	92,0
Свыше 1000	2	2020	147,1
Итого	27	4887	?

Рассчитаем средний процент изменения объема выпуска минеральных удобрений по всем предприятиям в январе–августе 1997 г. по сравнению с объемом за аналогичный период предыдущего года. Для этого объем производства в текущем периоде (числитель) нужно сравнить с объемом производства за аналогичный период предыдущего года (знаменатель):

$$\bar{x} = \frac{27 + 957 + 1883 + 2020}{\dfrac{27}{1{,}337} + \dfrac{957}{1{,}035} + \dfrac{1883}{0{,}92} + \dfrac{2020}{1{,}471}} =$$

$$= \frac{4887}{20{,}2 + 924{,}6 + 2046{,}7 + 1373{,}2} = \frac{4887}{4364{,}7}100 = 112\%.$$

Расчет средней величины при имеющихся исходных данных осуществлен по формуле средней гармонической взвешенной.

Если же в табл. 5.7 заменить исходные данные о производстве продукции в январе–августе 1997 г. на данные 1996 г. (табл. 5.8), расчет среднего процента изменения объема выпуска следует осуществлять по формуле средней арифметической.

Таблица 5.8

Производство минеральных удобрений предприятиями химической промышленности России в 1996 г.

Годовая мощность выпуска минеральных удобрений, тыс. т	Количество предприятий	Произведено продукции в январе–августе	
		1996 г., тыс. т V_i	1997 г., % к январю–августу 1996 г. x_i
А	1	2	3
Менее 100	6	20,2	133,7
100–500	11	924,6	103,5
500–1000	8	2046,7	92,0
Свыше 1000	2	1373,2	147,1
Итого	27	4364,7	?

Смысл вычислений не изменился: объем выпуска 1997 г. сравнивается с объемом выпуска 1996 г., но для этого нужно найти числитель – объем выпуска продукции за 1997 г. Здесь при осреднении относительных величин весами служат величины из знаменателя, принятые в качестве базы сравнения:

$$\bar{x} = \frac{1{,}337 \cdot 20{,}2 + 1{,}035 \cdot 924{,}6 + 0{,}92 \cdot 2046{,}7 + 1{,}471 \cdot 1373{,}2}{20{,}2 + 924{,}6 + 2046{,}7 + 1373{,}2} =$$

$$= \frac{27 + 957 + 1883 + 2020}{4364{,}7} = \frac{4887}{4364{,}7} 100 = 112\%.$$

Выбор вида средней арифметической или гармонической обусловлен наличием исходных данных (какие данные имеются: для числителя или для знаменателя), так как, вычисляя процент выполнения бизнес-плана или процент изменения объема выпуска при сравнении отчетного и базисного периодов, следует сопоставлять фактический выпуск с плановым или с выпуском за прошлый период.

Если известны данные за прошлый (или планируемый) период и процент изменения объема выпуска в отчетном периоде, при-

меняют *среднюю арифметическую*, для чего следует найти числитель – фактический выпуск.

Если известны фактический выпуск и процент изменения его по сравнению с выпуском за прошлый (базисный) период, применяют *среднюю гармоническую*, для чего следует найти знаменатель – выпуск за предыдущий период или планируемый.

В общем виде формула ***средней гармонической взвешенной*** (обозначаемой как $\bar{x}_{\text{гарм}}$) имеет вид

$$\bar{x}_{\text{гарм}} = \frac{V_1 + V_2 + \ldots + V_n}{\frac{V_1}{x_1} + \frac{V_2}{x_2} + \ldots + \frac{V_n}{x_n}} = \frac{\sum_i V_i}{\sum_i \frac{V_i}{x_i}}, \qquad (5.4)$$

а формула ***средней гармонической простой*** –

$$\bar{x}_{\text{гарм}} = \frac{n}{\sum_i \frac{1}{x_i}}. \qquad (5.5)$$

В формуле средней гармонической веса обозначены другой буквой (не m_i). В данном случае также необходимо, чтобы средние не только из прямых, но и из обратных значений признака определялись на основе соотношения

$$\frac{\text{Объем варьирующего признака}}{\text{Объем совокупности}}.$$

Область применения средней гармонической довольно узкая. Ее применяют в тех случаях, когда изучаемые показатели связаны между собой как x и $\frac{1}{x}$ (например, показатели выхода продукции на единицу сырья и соответствующие им обратные показатели удельного расхода сырья, материалов, электроэнергии или затраты времени на единицу продукции и выработка продукции в единицу времени).

Пример. Допустим, что шесть рабочих заняты обработкой деталей:

Рабочий	1	2	3	4	5	6
Затраты времени на обработку одной детали, мин (ч)	5 ($^1/_{12}$)	6 ($^1/_{10}$)	10 ($^1/_6$)	6 ($^1/_{10}$)	5 ($^1/_{12}$)	6 ($^1/_{10}$)

Определить средние затраты времени на одну деталь. Для этого нужно общие затраты времени всех рабочих разделить на количество деталей. Общие затраты времени шести рабочих (если их время работы один час или одинаковое) можно принять за 6 человекочасов. Количество деталей, обработанных ими за один час, найдем в знаменателе. Средняя гармоническая

$$\bar{x}_{\text{гарм}} = \frac{n}{\sum_i \frac{1}{x_i}} = \frac{1+1+1+1+1+1}{\frac{1}{1/12}+\frac{1}{1/10}+\frac{1}{1/6}+\frac{1}{1/10}+\frac{1}{1/12}+\frac{1}{1/10}} =$$

$$= \frac{6}{12+10+6+10+12+10} = \frac{6}{60} = 0{,}1 \text{ ч},$$

т.е. на одну деталь затрачивалось в среднем 0,1 ч, или 6 мин.

Среднюю гармоническую взвешенную будем применять при наличии данных о фактически отработанном времени каждым рабочим или группой рабочих.

Такие данные приведены в табл. 5.9. Эта группировка, как и при вычислении средней арифметической, позволяет заменить суммирование одинаковых слагаемых умножением их на веса V_i.

В нашем примере некоторые рабочие имеют одинаковые затраты времени: трое рабочих по 6 мин, а двое – по 5 мин. Поэтому группировка рабочих по затратам времени получит вид, приведенный в табл. 5.9 (графы А и 1).

Таблица 5.9

Затраты времени на одну деталь, ч x_i	Число рабочих	Количество деталей, обработанных в течение часа	Отработанные в течение 8-часового рабочего дня человекочасы V_i
А	1	2	3
1/6	1	6	8
1/10	3	10	24
1/12	2	12	16
Итого	6		48

В графе 2 табл. 5.9 показана выработка деталей в час одним рабочим. Определим средние затраты времени на одну деталь при условии, что рабочие заняты обработкой деталей в течение 8-часового рабочего дня. В графе 3 табл. 5.9 приведены данные о затратах времени группой рабочих в течение рабочего дня, которые

и выступают в качестве весов при расчете средней гармонической взвешенной.

На основе данных, приведенных в табл. 5.9, получим

$$\bar{x}_{\text{гарм}} = \frac{\sum_i V_i}{\sum_i \frac{V_i}{x_i}} = \frac{8 + 24 + 16}{\frac{8}{1/6} + \frac{24}{1/10} + \frac{16}{1/12}} = \frac{48}{480} = 0{,}1 \text{ ч, или 6 мин.}$$

Эта же формула применима для случая, когда продолжительность рабочего дня различна у отдельных категорий рабочих.

Мода

Важнейшей характеристикой центра распределения, кроме средней арифметической, является мода.

Мода – это значение признака, которое чаще всего встречается в вариационном ряду. Во многих случаях эта величина наиболее характерна для ряда распределения и вокруг нее концентрируется бóльшая часть вариантов. При изменении распределения в его концах мода не меняется, т.е. она обладает определенной устойчивостью к вариации признака. Поэтому моду наиболее удобно применять при изучении рядов с неопределенными границами.

Для дискретного ряда мода находится непосредственно по определению. Так, по данным табл. 5.1 исходя из наибольшего значения частоты определяем, что типичное число членов домашних хозяйств – 2 человека. Из 1000 домашних хозяйств 276 состоят всего из 2 человек (276 – максимальная частота ряда, а 2 – значение признака, которое встречается чаще всего).

Для интервального ряда с равными интервалами сначала определяется модальный интервал $x_{k-1} - x_k$, которому соответствует максимальная частота m_k или частость w_k. Значение моды внутри модального интервала определяется по интерполяционной формуле*

$$Mo = x_{k-1} + h_k \frac{m_k - m_{k-1}}{(m_k - m_{k-1}) + (m_k - m_{k+1})}, \qquad (5.6)$$

где x_{k-1} – нижняя граница модального интервала;

h_k – длина модального интервала;

---

* Интерполяционная формула моды была предложена известным статистиком Р.М. Орженцким (1863–1923). Формула (5.6) выведена математически исходя из допущения, что в модальном и двух соседних с ним интервалах кривая распределения представляет собой параболу 2-го порядка. Тогда *Mo* находится как абсцисса точки максимума кривой распределения.

m_{k-1}, m_k, m_{k+1} – частота интервала, соответственно предшествующего модальному, модального и следующего за модальным.

Для ряда с неравными интервалами модальный интервал определяется по наибольшей плотности распределения. Строго говоря, мода – это значение признака, которому соответствует максимальная плотность распределения. Поэтому в формуле моды вместо частот m_{k-1}, m_k, m_{k+1} следует взять плотности распределения y_{k-1}, y_k, y_{k+1}.

В табл. 5.3 плотности распределения представлены в графе 7. Наибольшая плотность, равная 6, соответствует интервалу 10–12. Это означает, что чаще всего встречаются банки с активами от 10 до 12 млрд руб. В этом случае значение моды

$$Mo = x_{k-1} + h_k \frac{y_k - y_{k-1}}{(y_k - y_{k-1}) + (y_k - y_{k+1})} =$$

$$= 10 + 2\frac{6 - 0}{(6 - 0) + (6 - 5{,}3)} = 11{,}79 \text{ млрд руб.}$$

Таким образом, наиболее типичный объем активов банков (11,79 млрд руб.) оказался гораздо меньше средней арифметической (45,78 млрд руб.). Это обстоятельство следует учитывать, принимая решение, например, об изменении нормативной (минимальной) величины уставного фонда. Задачи, связанные с отысканием моды, обычно решаются применительно к одновершинным (одномодальным) распределениям.

Графически моду определяют по гистограмме распределения. Для этого выбирают самый высокий прямоугольник, который и является модальным, далее верхнюю правую вершину модального прямоугольника соединяют с верхней правой вершиной предшествующего прямоугольника, а верхнюю левую вершину модального прямоугольника с верхней левой вершиной последующего прямоугольника. Абсцисса точки пересечения этих отрезков и будет модой распределения.

Медиана

В статистическом анализе часто применяют структурные, или порядковые, средние, например медиану.

В отличие от средней арифметической, на которую оказывают влияние все значения x_i, структурные средние не зависят от крайних значений признака.

Медианой называют такое значение признака, которое приходится на середину ранжированного ряда. Таким образом, в ранжированном ряду распределения одна половина ряда имеет значения признака больше медианы, другая – меньше медианы.

В дискретном ряду медиана находится непосредственно по определению на основе накопленных частот. Для распределения домашних хозяйств (см. табл. 5.1) номер медианы 1000 : 2 = = 500. Накапливаем частоты до тех пор, пока не будет превзойден номер медианы. Так, 223 домашних хозяйства имеют не более одного человека, 223 + 276 = 499 домашних хозяйств – не более 2 человек, а 499 + 238 = 787 домашних хозяйств – не более 3, т.е. 500-е и 501-е домашние хозяйства состоят из 3 человек. Таким образом, медиана данного ряда равна 3.

Ряд с четным числом единиц делит пополам не одна, а две единицы совокупности. Так, в распределении 50 коммерческих банков в середине ряда расположены единицы совокупности под номерами 25 и 26.

Тогда $Me = \frac{x_{25} + x_{26}}{2} = \frac{21{,}5 + 20{,}9}{2} = 21{,}2$ млрд руб. (см. ранжированные исходные данные на с. 84).

Однако на практике для простоты счета номер медианы при четном числе членов ряда определяется как $\frac{1}{2}\sum_i m_i$ либо $\frac{1}{2}\sum_i w_i$. Номер медианы для ряда с нечетным числом членов равен $\frac{N+1}{2}$.

В случае интервального вариационного ряда медиану определяют в такой последовательности. Прежде всего находят медианный интервал. Для этой цели используются накопленные частоты (или частости). Соответственно номер медианы равен $\frac{1}{2}\sum_i m_i$ или $\frac{1}{2}\sum_i w_i$.

Так, по данным табл. 5.3 номер медианы, рассчитанный на основе накопленных частот, равен 25 $\left(\frac{50}{2}\right)$ или 50 $\left(\frac{100}{2}\right)$, если исходить из частостей.

Далее на основе накопленных частот (см. графу 5 табл. 5.3) определяют, что 25-й банк находится в интервале 20–30. Точное

нахождение медианы на данном интервале осуществляется по следующей интерполяционной формуле:

$$Me = x_{k-1} + h_k \frac{\frac{1}{2}\sum_i m_i - F_{k-1}}{m_k}, \qquad (5.7)$$

где x_{k-1} – нижняя граница медианного интервала;
h_k – длина медианного интервала;
F_{k-1} – накопленная частота интервала, предшествующего медианному;
m_k – частота медианного интервала.

Таким образом,

$$Me = 20 + 10 \frac{25 - 24}{8} = 21{,}25 \text{ млрд руб.}$$

Аналогичный ответ получится, если вместо частот использовать частости.

Номер медианы $N_{Me} = \frac{100}{2} = 50.$

Накапливаем частости в графе 6 табл. 5.3 до тех пор, пока не будет превзойден номер медианы, равный 50. Итак, активы 48% банков не превышают 20 млрд руб. Следовательно, 50-я единица, т.е. медиана, находится в интервале 20–30:

$$Me = x_{k-1} + h_k \frac{\frac{1}{2}\sum_i w_i - p_{k-1}}{w_k} = 20 + 10\frac{50 - 48}{16} = 21{,}25 \text{ млрд руб.},$$

где p_{k-1} – накопленная частость интервала, предшествующего медианному;
w_k – частость медианного интервала.

Из определения медианы следует, что она не зависит от тех значений признака, которые расположены по обе стороны от нее. В связи с этим медиана является лучшей характеристикой центральной тенденции в тех случаях, когда концы распределений расплывчаты (например, границы крайних интервалов открыты) или в ряду распределения имеются чрезмерно большие или малые значения.

Значение медианы можно использовать, например, для установления официального прожиточного минимума или уровня бедности. В разных странах за прожиточный минимум принимают 40, 50 или 60% медианного дохода.

В интервальном ряду медиану можно определить графически. Медиана рассчитывается по кумуляте (см. рис. 5.3). Для этого из точки на шкале накопленных частот (частостей), соответствующей $\frac{1}{2}\sum_i m_i$ (или 50%), проводится прямая, параллельная оси абсцисс, до пересечения с кумулятой. Затем из точки пересечения указанной прямой с кумулятой опускается перпендикуляр на ось абсцисс. Абсцисса точки пересечения и является медианой.

5.3. Показатели вариации и способы их расчета

В практическом анализе оценка рассеяния значений признака может оказаться не менее важной, чем определение средней.

Самая грубая оценка рассеяния, легко определяемая по данным вариационного ряда, может быть дана с помощью ***размаха вариации***

$$R = x_{max} - x_{min},$$

где x_{max} и x_{min} – наибольшее и наименьшее значения варьирующего признака.

Этот показатель представляет интерес в тех случаях, когда важно знать, какова амплитуда колебаний значений признака, например, каковы колебания цены на данный товар в течение недели или по разным регионам в данный отрезок времени.

Однако этот показатель не дает представления о характере вариационного ряда, расположении вариантов вокруг средней и может сильно меняться, если добавить или исключить крайние варианты (когда эти значения аномальны для данной совокупности). В этих случаях размах вариации дает искаженную амплитуду колебания против нормальных ее размеров. Поэтому следует очистить совокупность от аномальных наблюдений, прежде чем определять размах вариации. Так, для совокупности банков (см. с. 84), если отбросить так называемое аномальное значение (1322,7), размах вариации составит 228,7 – 10,9 = 217,8 млрд руб.

Для оценки колеблемости значений признака относительно средней используются характеристики *рассеяния*. Они различаются выбранной формой средней и способами оценки отклонений от нее отдельных вариантов. К таким показателям относятся:

- среднее линейное отклонение;
- дисперсия;
- среднее квадратическое отклонение;
- коэффициент вариации.

Среднее линейное отклонение есть средняя арифметическая из абсолютных значений отклонений отдельных вариантов от их средней величины:

для несгруппированных данных

$$\bar{d} = \frac{\sum_i |x_i - \bar{x}|}{n}, \tag{5.8}$$

для сгруппированных данных

$$\bar{d} = \frac{\sum_i |x_i - \bar{x}| m_i}{\sum_i m_i}, \tag{5.9}$$

где x_i – значение признака в дискретном ряду или середина интервала в интервальном распределении;

m_i – частота признака.

Поскольку $\frac{\sum_i |x_i - \bar{x}| m_i}{\sum_i m_i} = 0$ согласно свойству средней арифметической, то мерой вариации выступает не алгебраическая средняя из отклонений, а средний модуль отклонений. Он не зависит от случайных колебаний и учитывает всю сумму отклонений конкретных вариантов от средней.

Среднее линейное отклонение выражено в тех же единицах измерения, что и варианты или их средняя. Оно дает абсолютную меру вариации.

Чтобы избежать равенства нулю суммы отклонений от средней, используют либо абсолютные значения отклонений, либо их четные степени, например квадраты. В последнем случае мера вариации называется ***дисперсией*** и обозначается D или σ^2:

для несгруппированных данных

$$D = \frac{\sum_i (x_i - \bar{x})^2}{n}, \tag{5.10}$$

для сгруппированных данных

$$D = \frac{\sum_i (x_i - \bar{x})^2 m_i}{\sum_i m_i}. \tag{5.11}$$

Исчисление дисперсии сопряжено с громоздкими расчетами, особенно если средняя величина выражена числом с несколькими десятичными знаками. Расчеты можно упростить, если использовать следующую модификацию формулы дисперсии:

$$\sigma^2 = \frac{\sum_i (x_i - \bar{x})^2 m_i}{\sum_i m_i} = \frac{\sum_i x_i^2 m_i}{\sum_i m_i} - \frac{2\sum_i x_i \bar{x} m_i}{\sum_i m_i} + \frac{\sum_i \bar{x}^2 m_i}{\sum_i m_i} =$$

$$= \frac{\sum_i x_i^2 m_i}{\sum_i m_i} - 2\bar{x}\frac{\sum_i x_i m_i}{\sum_i m_i} + \bar{x}^2 = \overline{x^2} - 2\bar{x}^2 + \bar{x}^2 = \overline{x^2} - \bar{x}^2. \quad (5.12)$$

Существуют и другие способы для упрощения исчисления дисперсии.

Однако вследствие суммирования квадратов отклонений дисперсия дает искаженное представление об отклонениях, измеряя их в квадратных единицах. Поэтому на основе дисперсии вводятся еще две характеристики: среднее квадратическое отклонение и коэффициент вариации.

Среднее квадратическое отклонение измеряется в тех же единицах, что и варьирующий признак, и исчисляется путем извлечения квадратного корня из дисперсии:

для несгруппированных данных

$$\sigma = \sqrt{\frac{\sum_i (x_i - \bar{x})^2}{n}}, \quad (5.13)$$

для сгруппированных данных

$$\sigma = \sqrt{\frac{\sum_i (x_i - \bar{x})^2 m_i}{\sum_i m_i}}. \quad (5.14)$$

Среднее квадратическое отклонение, как и среднее линейное отклонение, показывает, на сколько в среднем отклоняются конкретные варианты признака от его среднего значения. Величина σ часто используется в качестве единицы измерения отклонений от средней арифметической. Отклонение, выраженное в σ, называется *нормированным* или *стандартизированным*.

Для оценки меры вариации и ее значимости пользуются также ***коэффициентом вариации*** V, который дает относительную оценку вариации и получается путем сопоставления среднего линейного или среднего квадратического отклонения со средним уровнем явления, а результат выражается в процентах:

$$V = \frac{\bar{d}}{\bar{x}} 100\% \quad \text{либо} \quad V = \frac{\sigma}{\bar{x}} 100\% \quad (\bar{x} \neq 0). \tag{5.15}$$

Так как коэффициенты вариации дают относительную характеристику однородности явлений и процессов, они позволяют сравнивать степень вариации разных признаков.

Рассчитаем показатели вариации на основе распределения банков (см. табл. 5.3). Дополним табл. 5.3 графами 11–13, где приведены величины, которые нужны для расчета среднего линейного и среднего квадратического отклонений:

$x_i - \bar{x}$ ($\bar{x} = 45{,}78$)	$\lvert x_i - \bar{x} \rvert m_i$	$(x_i - \bar{x})^2 m_i$
11	12	13
−34,78	208,68	7257,89
−32,28	258,24	8335,99
−28,28	282,80	7997,58
−20,78	166,24	3454,47
−5,78	34,68	200,45
29,22	175,32	5122,85
129,22	775,32	100186,85
Σ	1901,28	132556,08

Итак, среднее линейное отклонение

$$\bar{d} = \frac{\sum_i \lvert x_i - \bar{x} \rvert m_i}{\sum_i m_i} = \frac{1901{,}28}{50} = 38{,}03 \text{ млрд руб.},$$

дисперсия

$$\sigma^2 = \frac{\sum_i (x_i - \bar{x})^2 m_i}{\sum_i m_i} = \frac{132556{,}08}{50} = 2651{,}12,$$

среднее квадратическое отклонение $\sigma = 51{,}49$ млрд руб.,

коэффициент вариации

$$V = \frac{\sigma}{\bar{x}} 100\% = \frac{51,49}{45,78} 100 = 112,5\%,$$

или, если пользоваться линейным отклонением,

$$V = \frac{\bar{d}}{\bar{x}} 100\% = \frac{38,03}{45,78} 100 = 83,1\%.$$

Однако надо отметить, что если коэффициент вариации рассчитывается на основе среднего линейного отклонения ($\bar{d}$), то в знаменателе чаще используют медиану, а не $\bar{x}$, т.е. в этом случае

$$V = \frac{\bar{d}}{Me} 100\% = \frac{38,03}{21,2} 100 = 179\%.$$

Среднее квадратическое отклонение σ всегда будет больше среднего линейного отклонения $\bar{d}$, что обусловлено разными способами их вычисления.

Среди признаков, изучаемых статистикой, есть и такие, которым свойственны лишь два взаимоисключающих значения. Такие признаки называются *альтернативными*. Им придается соответственно два количественных значения: 1 и 0. Частостью варианта 1 (она обозначается p) является доля единиц, обладающих данным признаком, в общей численности совокупности. Разность $1 - p = q$ является частостью варианта 0. Таким образом,

x_i	w_i
1	p
0	q

Средняя арифметическая альтернативного признака

$$\bar{x} = \frac{1 \cdot p + 0 \cdot q}{p + q} = p. \qquad (5.16)$$

Дисперсия альтернативного признака

$$\sigma^2 = \frac{(1 - p)^2 p + (0 - p)^2 q}{p + q} = \frac{q^2 p + p^2 q}{p + q} = pq, \qquad (5.17)$$

т.е. дисперсия альтернативного признака равна произведению доли единиц, обладающих данным признаком, и доли единиц, не обладающих этим признаком.

Если значения 1 и 0 встречаются одинаково часто, т.е. $p = q$, то дисперсия достигает своего максимума $pq = 0{,}25$.

Дисперсия альтернативного признака используется в выборочных обследованиях, например, качества продукции.

5.4. Виды дисперсий в совокупности, разделенной на части. Правило сложения дисперсий

Межгрупповая и внутригрупповая дисперсии

Если статистическая совокупность разбита на группы по какому-либо признаку и для этих групп известны (или могут быть найдены) средний уровень и дисперсия, то нередко при объединении частных групп в совокупность требуется оценить вариации показателей объединенной совокупности на основе показателей отдельных частных групп. При этом необходимо учитывать, что вариация признака в целом по совокупности зависит как от вариации признака внутри каждой группы, так и от вариации групповых средних, т.е. от межгрупповой вариации признака. Другими словами, общую дисперсию $\sigma^2_{\text{общ}}$, характеризующую вариацию признака под влиянием всех факторов, можно получить на основе ее составляющих – межгрупповой и внутригрупповой дисперсий.

Рассмотрим простейший случай, когда исходная совокупность делится на m однородных групп по одному признаку-фактору.

Допустим, имеется распределение исходной совокупности, представленное в табл. 5.10.

Таблица 5.10

Распределение исходной совокупности по группам

Значение признака x_i	Число единиц в j-й группе				*Итого*
	1	2	...	m	
x_1	f_1	s_1	...	t_1	$f_1 + s_1 + ... + t_1 = n_1$
x_2	f_2	s_2	...	t_2	$f_2 + s_2 + ... + t_2 = n_2$
...	...	...	...	...	...
x_k	f_k	s_k	...	t_k	$f_k + s_k + ... + t_k = n_k$
Итого	N_1	N_2	...	N_m	N

Сначала вычисляем m частных средних, т.е. среднее значение признака в каждой группе:

$$\bar{x}_1 = \frac{\sum_{i=1}^{k} x_i f_i}{N_1}, \quad \bar{x}_2 = \frac{\sum_{i=1}^{k} x_i s_i}{N_2}, \quad \ldots, \quad \bar{x}_m = \frac{\sum_{i=1}^{k} x_i t_i}{N_m}.$$

На основе частных средних $\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_m$ определяем общую среднюю по формуле

$$\bar{x}_{\text{общ}} = \frac{\sum_{j=1}^{m} \bar{x}_j N_j}{N}, \tag{5.18}$$

где $N = \sum_{j=1}^{m} N_j = \sum_{i=1}^{k} n_i$.

Общая дисперсия совокупности

$$\sigma^2_{\text{общ}} = \frac{\sum_{i=1}^{k} \left(x_i - \bar{x}_{\text{общ}}\right)^2 n_i}{N}. \tag{5.19}$$

Общая дисперсия отражает вариацию признака за счет всех условий (факторов), действующих в данной совокупности.

Вариацию между группами за счет признака-фактора, положенного в основу группировки, отражает ***межгрупповая дисперсия***, которая исчисляется по отклонениям групповых средних от общей средней:

$$\delta^2 = \frac{\sum_{j=1}^{m} \left(\bar{x}_j - \bar{x}_{\text{общ}}\right)^2 N_j}{N}. \tag{5.20}$$

Вариацию внутри каждой группы изучаемой совокупности отражает ***частная групповая дисперсия***, которая исчисляется как средний квадрат отклонений значений признака x от частной средней $\bar{x}_j$:

$$\sigma_1^2 = \frac{\sum_{i=1}^{k} x_i^2 f_i}{N_1} - (\bar{x}_1)^2, \quad \text{или} \quad \sigma_1^2 = \frac{\sum_{i=1}^{k} \left(x_i - \bar{x}_1\right)^2 f_i}{N_1},$$

$$\sigma_2^2 = \frac{\sum_{i=1}^{k} x_i^2 s_i}{N_2} - (\bar{x}_2)^2, \quad \text{или} \quad \sigma_2^2 = \frac{\sum_{i=1}^{k} \left(x_i - \bar{x}_2\right)^2 s_i}{N_2},$$

$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad \ldots\ldots \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots$$

$$\sigma_m^2 = \frac{\sum_{i=1}^{k} x_i^2 t_i}{N_m} - (\bar{x}_m)^2, \quad \text{или} \quad \sigma_m^2 = \frac{\sum_{i=1}^{k} \left(x_i - \bar{x}_m\right)^2 t_i}{N_m}.$$

В общем виде частную дисперсию запишем так:

$$\sigma_j^2 = \frac{\sum_{i=1}^{k} x_i^2 N_{ij}}{N_j} - (\bar{x}_j)^2,$$

где N_{ij} – частоты от $i = 1 \div k$ в каждой j-й группе.

Так как изучаемая совокупность разбита на несколько групп, то для всей совокупности внутригрупповую вариацию будет выражать ***внутригрупповая дисперсия***, которая рассчитывается как средняя арифметическая из групповых дисперсий:

$$\overline{\sigma^2} = \frac{\sum_{j=1}^{m} \sigma_j^2 N_j}{N}. \qquad (5.21)$$

Между представленными видами дисперсий существует определенное соотношение: общая дисперсия равна сумме дисперсий внутригрупповой (средней из групповых дисперсий) и межгрупповой (дисперсии частных средних), т.е.

$$\sigma_{\text{общ}}^2 = \overline{\sigma^2} + \delta^2. \qquad (5.22)$$

Это равенство известно как *правило сложения дисперсий*, его автором является Вильгельм Лексис (1837–1914), немецкий статистик и экономист.

Докажем равенство (5.22), для чего формулу частной дисперсии

$$\sigma_j^2 = \frac{\sum_{i=1}^{k} x_i^2 N_{ij}}{N_j} - (\bar{x}_j)^2$$

перепишем в виде

$$\sigma_j^2 N_j = \sum_{i=1}^{k} x_i^2 N_{ij} - (\bar{x}_j)^2 N_j,$$

откуда

$$\sum_{i=1}^{k} x_i^2 N_{ij} = \sigma_j^2 N_j + (\bar{x}_j)^2 N_j.$$

Составив для каждой группы аналогичные уравнения и просуммировав их, получим

$$\sum_{j=1}^{m} \sum_{i=1}^{k} x_i^2 N_{ij} = \sum_{j=1}^{m} \sigma_j^2 N_j + \sum_{j=1}^{m} (\bar{x}_j)^2 N_j. \qquad (5.23)$$

Однако

$$\sum_{j=1}^{m}\sum_{i=1}^{k} x_i^2 N_{ij} = \sum_{i=1}^{k} x_i^2 n_i,$$

т.е. мы получили не что иное, как сумму взвешенных квадратов значений x_i по совокупности в целом.

Разделим равенство (5.23) на общую численность совокупности N:

$$\frac{\sum_{i=1}^{k} x_i^2 n_i}{N} = \frac{\sum_{j=1}^{m} \sigma_j^2 N_j}{N} + \frac{\sum_{j=1}^{m} \left(\bar{x}_j\right)^2 N_j}{N},$$

затем вычтем квадрат общей средней из обеих частей уравнения:

$$\frac{\sum_{i=1}^{k} x_i^2 n_i}{N} - \left(\bar{x}_{\text{общ}}\right)^2 = \frac{\sum_{j=1}^{m} \sigma_j^2 N_j}{N} + \frac{\sum_{j=1}^{m} \left(\bar{x}_j\right)^2 N_j}{N} - \left(\bar{x}_{\text{общ}}\right)^2. \quad (5.24)$$

В левой части (5.24) представлена общая дисперсия, а в правой – сумма внутригрупповой и межгрупповой дисперсий:

$$\sigma^2_{\text{общ}} = \overline{\sigma^2} + \delta^2.$$

Таким образом, общая дисперсия складывается из двух слагаемых: первое измеряет вариацию внутри частей совокупности, а второе – вариацию между средними этих частей.

Эмпирический коэффициент детерминации и эмпирическое корреляционное отношение

Правило сложения дисперсий позволяет выявить зависимость результатов от определяющих факторов с помощью соотношения межгрупповой и общей дисперсий. Это соотношение называется ***эмпирическим коэффициентом детерминации*** $\eta^2_{\text{эмп}}$ и показывает, какая доля в общей дисперсии приходится на дисперсию, обусловленную вариацией признака, положенного в основу группировки:

$$\eta^2_{\text{эмп}} = \frac{\delta^2}{\sigma^2_{\text{общ}}}. \quad (5.25)$$

Используется правило сложения дисперсий и для определения степени связи между изучаемыми признаками. Для этого необходимо найти ***эмпирическое корреляционное отношение*** $\eta_{\text{эмп}}$, которое показывает, насколько тесно связаны исследуемое явление и группировочный признак:

$$\eta_{\text{эмп}} = \sqrt{\frac{\delta^2}{\sigma^2_{\text{общ}}}}. \quad (5.26)$$

Эмпирическое корреляционное отношение изменяется от 0 до 1. Если связь отсутствует, то $\eta_{эмп} = 0$. В этом случае дисперсия групповых средних равна нулю ($\delta^2 = 0$), т.е. все групповые средние равны между собой и межгрупповой вариации нет. Это означает, что группировочный признак не влияет на вариацию исследуемого признака x.

Если связь функциональная, то $\eta_{эмп} = 1$. В этом случае дисперсия групповых средних равна общей дисперсии ($\delta^2 = \sigma^2_{общ}$), т.е. не будет внутригрупповой вариации. Это означает, что группировочный признак полностью определяет вариацию изучаемого признака.

Чем больше значение корреляционного отношения приближается к единице, тем полнее (сильнее) корреляционная связь между признаками (табл. 5.11).

Таблица 5.11

Качественная оценка связи между признаками

$\eta_{эмп}$	Связь	$\eta_{эмп}$	Связь
0	Отсутствует	0,5–0,7	Заметная
0–0,2	Очень слабая	0,7–0,9	Тесная
0,2–0,3	Слабая	0,9–0,99	Весьма тесная
0,3–0,5	Умеренная	1	Функциональная

Пример. Рассчитать дисперсию, эмпирический коэффициент детерминации и эмпирическое корреляционное отношение по данным, приведенным в табл. 5.12.

Таблица 5.12

Среднемесячная номинальная заработная плата работников предприятий и организаций России по федеральным округам в 2002 г.

Федеральный округ	Средний размер заработной платы, тыс. руб. $\bar{x}_j$	Численность занятых, млн чел. N_j	Дисперсия заработной платы σ_j^2
Центральный	4,433	17,508	0,58
Северо-Западный	5,068	7,091	3,769
Южный	2,974	8,505	0,115
Приволжский	3,142	14,624	0,128
Уральский	6,589	5,795	6,743
Сибирский	4,310	9,147	3,299
Дальневосточный	5,979	3,401	3,458
Итого		66,071	

Источник: Труд и занятость в России: Стат. сб. – М.: Госкомстат России, 2003. – С. 109–126.

Сначала найдем средний размер заработной платы по стране:

$$\bar{x}_{\text{общ}} = \frac{\sum_j \bar{x}_j N_j}{\sum_j N_j} =$$

$$= [4{,}433 \cdot 17{,}508 + 5{,}068 \cdot 7{,}091 + 2{,}974 \cdot 8{,}505 + 3{,}142 \cdot 14{,}624 + 6{,}589 \cdot 5{,}795 + 4{,}310 \cdot 9{,}147 + 5{,}979 \cdot 3{,}401] / 66{,}071 = 4{,}339 \text{ тыс. руб.}$$

Вариация средней заработной платы по федеральным округам, обусловленная различием в местах проживания занятого населения, характеризуется межгрупповой дисперсией:

$$\delta^2 = \frac{\sum_j \left(\bar{x}_j - \bar{x}_{\text{общ}}\right)^2 N_j}{\sum_j N_j} =$$

$$= [(4{,}433 - 4{,}339)^2 \cdot 17{,}508 + (5{,}068 - 4{,}339)^2 \cdot 7{,}091 + (2{,}974 - 4{,}339)^2 \cdot 8{,}505 + (3{,}142 - 4{,}339)^2 \cdot 14{,}624 + (6{,}589 - 4{,}339)^2 \cdot 5{,}795 + (4{,}310 - 4{,}339)^2 \cdot 9{,}147 + (5{,}979 - 4{,}339)^2 \cdot 3{,}401] / 66{,}071 = 1{,}072.$$

Средняя из групповых дисперсий дает обобщающую характеристику случайной вариации, обусловленную всеми отдельными факторами, кроме места проживания работающего населения (например, характером занятости, стажем работы и т.п.):

$$\overline{\sigma^2} = \frac{\sum_j \sigma_j^2 N_j}{\sum_j N_j} =$$

$$= [0{,}58 \cdot 17{,}508 + 3{,}769 \cdot 7{,}091 + 0{,}115 \cdot 8{,}505 + 0{,}128 \cdot 14{,}624 + 6{,}743 \cdot 5{,}795 + 3{,}299 \cdot 9{,}147 + 3{,}458 \cdot 3{,}401] / 66{,}071 = 1{,}827.$$

Вариация средней заработной платы в регионах России, обусловленная влиянием всех факторов, вместе взятых, определяется общей дисперсией:

$$\sigma^2_{\text{общ}} = \delta^2 + \overline{\sigma^2} = 1{,}072 + 1{,}827 = 2{,}899.$$

Сопоставляя межгрупповую дисперсию с общей, рассчитаем эмпирический коэффициент детерминации:

$$\eta^2_{\text{эмп}} = \frac{\delta^2}{\sigma^2_{\text{общ}}} = \frac{1{,}072}{2{,}899} = 0{,}369.$$

Полученный эмпирический коэффициент детерминации показывает, что дисперсия заработной платы зависит от места проживания работающего населения на 36,9%. Остальные 63,1% определяются множеством других неучтенных факторов. Извлекая квадратный корень из эмпирического коэффициента детерминации, определяем эмпирическое корреляционное отношение:

$$\eta_{\text{эмп}} = \sqrt{0{,}369} = 0{,}607.$$

Полученное значение эмпирического корреляционного отношения позволяет утверждать, что существует заметная связь между местом проживания работающего населения и размером заработной платы (см. табл. 5.11).

Для проверки существенности связи между группировочным признаком и вариацией исследуемого показателя часто используется дисперсионное отношение F (критерий Фишера):

$$F = \frac{\delta^2}{\nu_1} : \frac{\sigma^2}{\nu_2}, \qquad (5.27)$$

где ν_1 и ν_2 – число степеней свободы для сравниваемых дисперсий.

При этом

$$\nu_1 = m - 1, \qquad \nu_2 = N - m,$$

где m – число групп;

N – число наблюдений.

Расчетное значение критерия Фишера ($F_{\text{расч}}$) сравнивается с критическим ($F_{\text{кр}}$), определяемым по таблице Приложения 8 в зависимости от числа степеней свободы и уровня значимости α. Если $F_{\text{расч}} > F_{\text{кр}}$, наличие связи доказано, так как проверяется нулевая гипотеза об отсутствии взаимосвязи признаков, т.е. об отсутствии влияния группировочного признака на исследуемый признак.

Правило сложения дисперсий для доли признака

Рассмотренное правило сложения дисперсий распространяется и на дисперсии доли признака, т.е. доли единиц с определенным признаком в совокупности, разбитой на части (группы). При

этом изучение вариации происходит непосредственно при вычислении и анализе следующих видов дисперсий доли признака.

Групповая дисперсия доли признака

$$\sigma^2_{p_i} = p_i(1 - p_i), \tag{5.28}$$

где p_i – доля изучаемого признака в отдельных группах.

Внутригрупповая дисперсия, т.е. средняя из групповых дисперсий,

$$\overline{\sigma^2_{p_i}} = \frac{\sum_i p_i(1 - p_i)n_i}{\sum_i n_i} = \overline{p_i(1 - p_i)}, \tag{5.29}$$

где n_i – численность единиц в отдельных группах.

Межгрупповая дисперсия

$$\delta^2_{p_i} = \frac{\sum_i (p_i - \bar{p})^2 n_i}{\sum_i n_i}. \tag{5.30}$$

При этом $\bar{p}$ – доля изучаемого признака во всей совокупности – определяется по формуле средней арифметической взвешенной:

$$\bar{p} = \frac{\sum_i p_i n_i}{\sum_i n_i}.$$

Общая дисперсия

$$\sigma^2_{\bar{p}} = \bar{p}(1 - \bar{p}). \tag{5.31}$$

Кроме того, общую дисперсию можно определить как сумму средней из групповых дисперсий и межгрупповой дисперсии, т.е. по *правилу сложения дисперсий доли признака:*

$$\sigma^2_{\bar{p}} = \overline{\sigma^2_{p_i}} + \delta^2_{p_i}. \tag{5.32}$$

Зная любые два вида дисперсий из трех, входящих в формулу (5.32), можно определить дисперсию третьего вида или проверить правильность ее расчета.

Пример. Определить дисперсию доли безработных с высшим образованием по данным, представленным в табл. 5.13.

Таблица 5.13

Доля безработных с высшим образованием по федеральным округам России в 2002 г.

Федеральный округ	Доля безработных с высшим образованием		Численность безработных, млн чел. n_i
	%	p_i	
Центральный	12,2	0,122	0,995
Северо-Западный	11,5	0,115	0,478
Южный	10,4	0,104	1,162
Приволжский	8,3	0,083	1,217
Уральский	7,6	0,076	0,515
Сибирский	9,2	0,092	1,024
Дальневосточный	10,6	0,106	0,321
Σ			5,712

Источник: Труд и занятость в России: Стат. сб. – М.: Госкомстат России, 2003. – С. 109–111, 134–136.

Вначале определяем среднюю долю безработных с высшим образованием по России:

$$\bar{p} = [0{,}122 \cdot 0{,}995 + 0{,}115 \cdot 0{,}478 + 0{,}104 \cdot 1{,}162 + 0{,}083 \cdot 1{,}217 + 0{,}076 \cdot 0{,}515 + 0{,}092 \cdot 1{,}024 + 0{,}106 \cdot 0{,}321] / 5{,}712 \approx \approx 0{,}099 \text{ (или } 9{,}9\%).$$

Затем находим общую дисперсию этой доли по формуле (5.31):

$$\sigma^2_{\bar{p}} = \bar{p}(1 - \bar{p}) = 0{,}099(1 - 0{,}099) = 0{,}0892.$$

Для расчета общей дисперсии по формуле (5.32) определяем групповые дисперсии по федеральным округам, используя формулу (5.28):

$$\sigma^2_{\text{Центр}} = 0{,}122(1 - 0{,}122) = 0{,}107,$$
$$\sigma^2_{\text{Сев-Зап}} = 0{,}115(1 - 0{,}115) = 0{,}102,$$
$$\sigma^2_{\text{Юж}} = 0{,}104(1 - 0{,}104) = 0{,}093,$$
$$\sigma^2_{\text{Прив}} = 0{,}083(1 - 0{,}083) = 0{,}076,$$
$$\sigma^2_{\text{Ур}} = 0{,}076(1 - 0{,}076) = 0{,}070,$$
$$\sigma^2_{\text{Сиб}} = 0{,}092(1 - 0{,}092) = 0{,}084,$$
$$\sigma^2_{\text{Д}} = 0{,}106(1 - 0{,}106) = 0{,}095.$$

Определив групповые дисперсии, можно перейти к расчету средней дисперсии из групповых по формуле (5.29):

$$\overline{\sigma^2_{p_i}} = [0{,}107 \cdot 0{,}995 + 0{,}102 \cdot 0{,}478 + 0{,}093 \cdot 1{,}162 + 0{,}076 \cdot 1{,}217 + 0{,}070 \cdot 0{,}515 + 0{,}084 \cdot 1{,}024 + 0{,}095 \cdot 0{,}321] / 5{,}712 = 0{,}088995 \approx 0{,}0890.$$

Далее, зная долю безработных с высшим образованием в каждом округе и по стране в целом, а также численность безработного населения в каждом округе, по формуле (5.30) рассчитаем межгрупповую дисперсию:

$$\delta^2_{p_i} = [(0{,}122 - 0{,}099)^2 \cdot 0{,}995 + (0{,}115 - 0{,}099)^2 \cdot 0{,}478 + (0{,}104 - 0{,}099)^2 \cdot 1{,}162 + (0{,}083 - 0{,}099)^2 \cdot 1{,}217 + (0{,}076 - 0{,}099)^2 \cdot 0{,}515 + (0{,}092 - 0{,}099)^2 \cdot 1{,}024 + (0{,}106 - 0{,}099)^2 \cdot 0{,}321] / 5{,}712 = 0{,}0002.$$

По правилу сложения дисперсий общая дисперсия равна

$$\sigma^2_{\bar{p}} = 0{,}089 + 0{,}0002 = 0{,}0892.$$

Оба метода дали аналогичный результат, что подтверждает правильность расчета.

5.5. Показатели дифференциации и концентрации

Показатели дифференциации

Если возникает необходимость изучить структуру вариационного ряда более подробно, вычисляют значения признака, аналогичные медиане. Такие значения признака, которые делят все единицы распределения на равные численности, получили название ***квантилей***, или ***градиентов***. Квартили, квинтили, децили – частные случаи квантилей.

Квартилями называются такие значения признака, которые делят распределение на четыре равные части.

Общая идея построения квантилей довольно проста – расширить понятие медианы. С этой точки зрения медиана представляет собой центральный квартиль.

Обозначим значения x_i, делящие вариационный ряд на четыре равные части, через Q_1, Q_2, Q_3. Ниже первого квартиля лежит $1/4$ значений x_i, $3/4$ элементов совокупности имеют значения x_i, превышающие Q_1. Второй квартиль делит распределение пополам и совпадает с медианой. Между медианой и третьим кварти-

лем Q_3 располагается $1/4$ всей совокупности, и, наконец, $1/4$ значений лежит выше Q_3. При этом Q_3 называется *верхним квартилем*, Q_1 – *нижним квартилем.*

Квинтили делят распределение на пять равных частей.

Дециль – такое значение признака в ряду распределения, которому соответствуют десятые доли численности совокупности.

При изучении дифференциации доходов широко применяется *децильный коэффициент* $K_д$ – отношение девятого дециля к первому децилю. Сравнивая девятый и первый децили, измеряют соотношение уровней доходов 10% наиболее обеспеченного и 10% наименее обеспеченного населения (в разах).

Интерполяционные формулы для определения децилей в интервальном ряду распределения имеют следующий вид:

$$\text{Первый дециль} = x_{k-1} + h_k \frac{\frac{1}{10}\sum_i m_i - F_{k-1}}{m_k} \qquad (5.33)$$

или

$$\text{Первый дециль} = x_{k-1} + h_k \frac{10\% - p_{k-1}}{w_k}, \qquad (5.34)$$

где x_{k-1} – нижняя граница интервала, содержащего первый дециль;

$h_k = x_k - x_{k-1}$ – длина интервала, содержащего первый дециль;

F_{k-1} и p_{k-1} – соответственно накопленные частоты и накопленные частости предшествующего интервала;

m_k и w_k – соответственно частота и частость интервала, содержащего первый дециль.

Номер первого дециля определяется как

$$\frac{1}{10}\sum_i m_i \quad \text{или} \quad \frac{1}{10}\sum_i w_i,$$

где $\sum_i w_i = 100\%$ (или 1).

Для нахождения интервала, содержащего первый дециль, накапливают частоты или частости до тех пор, пока они не превзойдут номер единицы совокупности, соответствующей первому децилю.

Девятый дециль находится аналогично:

$$\text{Девятый дециль} = x_{k-1} + h_k \frac{\frac{9}{10}\sum_i m_i - F_{k-1}}{m_k} \qquad (5.35)$$

или

$$\text{Девятый дециль} = x_{k-1} + h_k \frac{90\% - p_{k-1}}{w_k}. \qquad (5.36)$$

Для нахождения интервала, содержащего девятый дециль, частости накапливают до тех пор, пока они не превзойдут номер единицы совокупности, соответствующей девятому децилю, т.е. 90%.

Рассмотрим пример с распределением банков по величине активов (см. табл. 5.3). Первый дециль найдем в интервале 10–12 (так как в этом интервале находится единица совокупности, которая делит ее в соотношении 10 к 90):

$$\text{Первый дециль} = 10 + 2\frac{10 - 0}{12} = 11{,}67 \text{ млрд руб.},$$

что характеризует максимальную величину активов 10% самых мелких банков.

Чтобы найти девятый дециль, определим интервал, содержащий единицу совокупности, которая делит ряд в соотношении 90 к 10. Это интервал 100–250. Таким образом,

$$\text{Девятый дециль} = 100 + 150\frac{90 - 88}{12} = 125 \text{ млрд руб.},$$

что характеризует минимальную величину активов 10% самых крупных банков.

Сопоставляя девятый и первый децили, находим децильный коэффициент дифференциации:

$$K_д = \frac{125}{11{,}67} \cong 11 \text{ раз}.$$

Данный показатель не совсем точно измеряет уровень дифференциации, так как сопоставляются минимальная величина активов 10% самых крупных банков с максимальной величиной активов 10% самых мелких банков.

Более точно уровень дифференциации можно измерить, сопоставив средние уровни активов 10% самых крупных и 10% самых мелких банков. Такой показатель называют *фондовым коэффициентом* $K_ф$.

Так как в нашем примере 10% самых крупных и 10% самых мелких банков составляют одну и ту же величину ($\frac{1}{10}$ 50 = 5 ед.), то фондовый коэффициент

$$K_{\text{ф}} = \frac{\frac{1}{5}\sum_j x_j}{\frac{1}{5}\sum_s x_s} = \frac{\sum_j x_j}{\sum_s x_s}, \qquad (5.37)$$

где $\sum_j x_j$ – сумма активов 10% самых крупных банков;

$\sum_s x_s$ – сумма активов 10% самых мелких банков.

Для расчета фондового коэффициента обратимся к исходным данным (см. с. 84). Подставив в формулу (5.37) соответствующие значения, получим

$$K_{\text{ф}} = \frac{(228{,}7 + 187{,}3 + 180{,}7 + 140{,}1 + 110{,}9)/5}{(10{,}9 + 11{,}2 + 11{,}3 + 11{,}4 + 11{,}5)/5} = \frac{169{,}54}{11{,}26} = 15 \text{ раз.}$$

Полученный коэффициент показывает, что уровень дифференциации довольно высок, средняя (а также суммарная) величина активов 10% самых крупных банков в 15 раз превышает среднюю (или совокупную) величину активов 10% самых мелких банков.

Показатели концентрации

К показателям дифференциации близки по значению показатели концентрации:

- коэффициент концентрации Джини;
- коэффициент Герфиндаля;
- коэффициент Лоренца и др.

Для описания концентрации можно использовать данные, приведенные в табл. 5.3 (см. графы 2 и 8). В графе 8 найдены доли активов групп банков – мелких, средних и крупных – в совокупном объеме активов. Так, только у шести банков (или 12% от общего числа банков) активы меньше 12 млрд руб., в совокупном объеме активов эти шесть банков занимают всего 2,9%. В то же время один Сбербанк (или 2% от общего числа банков) владеет 39% активов.

Сравним графы 6 и 9 табл. 5.3. В графе 6 показано нарастающим итогом распределение банков, а в графе 9 – нарастающим

итогом доли активов групп банков, распределенных в порядке возрастания активов. Например, 28% мелких банков с активами меньше 15 млрд руб. располагают лишь 7,6% активов, а банки с активами свыше 100 млрд руб. (их всего 12% от общего числа) располагают 45,9% активов.

Для оценки концентрации нужно рассчитать обобщающий показатель. Таким показателем является ***коэффициент концентрации Джини***

$$G = \sum_{i=1}^{n-1} p_i q_{i+1} - \sum_{i=1}^{n-1} p_{i+1} q_i, \qquad (5.38)$$

где p_i – накопленная доля (частость) численности единиц совокупности;

q_i – накопленная доля активов, приходящихся на все единицы совокупности, с активами не более x_i.

Рассчитаем коэффициент Джини по данным табл. 5.3 (см. графы 6 и 9), дополнив указанную таблицу графами 14 и 15:

p_i	...	q_i	...	$p_i q_{i+1}$	$p_{i+1} q_i$
6	...	9	...	14	15
12	...	0,029	...	0,912	–
28	...	0,076	...	4,256	0,812
48	...	0,152	...	11,472	3,648
64	...	0,239	...	22,016	9,728
76	...	0,344	...	41,116	18,164
88	...	0,541	...	88	30,272
100	...	1,000	...	–	54,1
				167,772	116,724

Коэффициент Джини $G = \dfrac{167{,}772 - 116{,}724}{100} = 0{,}51.$

Коэффициент Джини может принимать значения от 0 до 1, поэтому результат следует разделить либо на 100, если p_i или q_i выражен в процентах, либо на 10 000, если оба показателя выражены в процентах.

Коэффициент Джини по существу строится на основе кривой Лоренца, характеризующей накопление значения изучаемого признака в зависимости от накопления элементов совокупности (в данном случае – накопление доли активов в зависимости от

накопления доли владельцев активов). В прямоугольной системе координат кривая Лоренца является вогнутой и проходит под диагональю квадрата (рис. 5.5, *а*). При таком положении кривой доли значений исследуемого признака концентрируются в последних группах. При концентрации исследуемого признака в первых группах кривая Лоренца становится выпуклой (рис. 5.5, *б*).

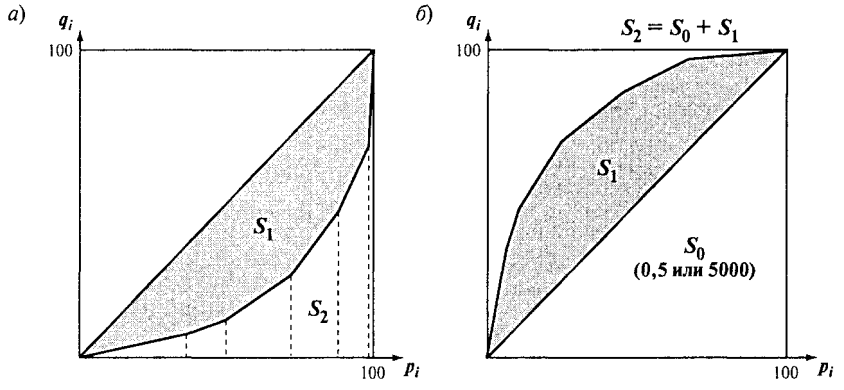


Рис. 5.5. График Лоренца

Как известно, чем больше кривая Лоренца отклоняется от диагонали, тем выше степень неравномерности распределения признака в совокупности. Эту степень неравномерности распределения, т.е. меру дифференциации признака в совокупности, можно выразить через площадь S_1, заключенную между диагональю квадрата и кривой Лоренца (см. рис. 5.5, *а*). Для того чтобы мера дифференциации находилась в стандартных границах (от 0 до 1), необходимо определить ее как отношение указанной площади S_1 к площади треугольника, которая включает в себя S_1 и S_2 и равна 0,5 или $\frac{100 \cdot 100}{2} = 5000$.

Для удобства вычисления выразим коэффициент Джини через отношение площадей:

$$G = \frac{S_1}{S_1 + S_2} = \frac{5000 - S_2}{5000}. \qquad (5.39)$$

Площадь S_2 приближенно равна сумме площадей одного треугольника и нескольких прямоугольных трапеций, образованных кривой Лоренца (см. рис. 5.5, *а*).

Итак, по данным табл. 5.3 (см. графы 2 и 9)

$$S_2 = \frac{1}{2}2{,}9 \cdot 12 + \frac{2{,}9 + 7{,}6}{2}16 + \frac{7{,}6 + 15{,}2}{2}20 + \frac{15{,}2 + 23{,}9}{2}16 +$$

$$+ \frac{23{,}9 + 34{,}4}{2}12 + \frac{34{,}4 + 54{,}1}{2}12 + \frac{54{,}1 + 100}{2}12 = 2447{,}6,$$

тогда

$$G = \frac{5000 - 2447{,}6}{5000} = 0{,}51.$$

Полученное значение показателя отражает высокую степень неравномерности распределения активов.

Площадь фигуры S_2 приближенно можно представить и следующим образом:

$$S_2 = \frac{1}{2}\sum_{i=1}^{n} w_i(q_{i-1} + q_i), \qquad (5.40)$$

где $q_0 = 0$, а $q_n = 1$ (n – число групп).

Тогда, подставив выражение для S_2 в формулу (5.39), получим модификацию формулы коэффициента Джини:

$$G = \frac{0{,}5 - 0{,}5\sum_{i=1}^{n} w_i(q_{i-1} + q_i)}{0{,}5} = 1 - \sum_{i=1}^{n} w_i(q_{i-1} + q_i). \qquad (5.41)$$

Для распределения, имеющего вид, представленный на рис. 5.5, *б*, площадь S_2, которая приближенно вычисляется по той же формуле (5.40), включает в себя площадь треугольника S_0, равную 0,5, и искомую площадь S_1. Поэтому для распределений, в которых доли значений исследуемого признака концентрируются в первых группах, коэффициент Джини

$$G = \frac{S_2 - 0{,}5}{0{,}5} = \frac{0{,}5\sum_{i=1}^{n} w_i(q_{i-1} + q_i) - 0{,}5}{0{,}5} =$$

$$= \sum_{i=1}^{n} w_i(q_{i-1} + q_i) - 1. \qquad (5.42)$$

Коэффициент Джини используют для характеристики степени неравномерности распределения населения по уровню доходов. В случае уравнительного распределения каждая группа получает доход пропорционально своей численности; при значительной

неравномерности преобладающая часть доходов сосредоточена у небольшой по удельному весу (численности) группы.

Коэффициенты концентрации рассчитывают для вариационных рядов, характеризующих распределение продукции по группам предприятий, а также распределение доходов. Кроме того, с помощью коэффициента Джини можно оценить концентрацию каких-либо явлений в различных регионах. Тогда его уместнее назвать *коэффициентом локализации.*

Для оценки концентрации производства можно использовать и более простой показатель – ***коэффициент Герфиндаля.*** Он вычисляется на основе данных о доле производства (или доходов) отдельных групп в совокупном объеме производства (или доходов). Коэффициент Герфиндаля

$$H = \sum_i \left(\frac{x_i m_i}{\sum_i x_i m_i} \right)^2 \quad \text{или} \quad H = \sum_i \left(\frac{Q_i}{\sum_i Q_i} \right)^2, \qquad (5.43)$$

где $\frac{x_i m_i}{\sum_i x_i m_i}$ – доля производства (доходов) i-й группы в общем объеме производства (доходов);

Q_i – объем производства в i-й группе.

Группами с незначительной долей производства можно пренебречь, так как, будучи возведенной в квадрат, такая доля выражается незначащим числом. Таким образом, значение коэффициента Герфиндаля определяется влиянием лишь доминирующих групп. В нашей задаче $H = 0{,}277$ (см. графу 10 табл. 5.3), что подтверждает доминирующую роль нескольких крупнейших банков.

Вторая формула (5.43) применяется тогда, когда данные об объемах производства продукции по группам уже известны и нет необходимости находить их менее точным способом (как $x_i m_i$).

Показатель H зависит от числа предприятий в группах.

В некоторых странах в качестве *обобщающего показателя уровня концентрации производства* принимается доля фиксированного числа предприятий с наибольшим удельным весом в общем объеме производства продукции (услуг) в данной отрасли по стоимости (*CR*). Все предприятия ранжируются по показателю стоимости продукции и объединяются в группы на основе принятого фиксированного числа предприятий. В США принято рассчитывать показатели *CR* как доли 4, 8 и 20 крупнейших фирм; в Англии, Канаде, Германии – как доли 3, 6 и 10 предприятий. В России расчет этого показателя проводится с использованием нескольких вариантов фиксированного числа предприятий.

Однако показатель уровня концентрации производства CR не лишен недостатков. Например, с его помощью трудно однозначно оценить уровень концентрации производства.

Так, CR_3, рассчитанный на основе ранжированных данных о крупнейших банках России (см. с. 84), покажет господствующее положение трех банков:

$$CR_3 = \frac{\sum_{i=1}^{3} x_i}{\sum_{i=1}^{n} x_i} = \frac{1322{,}7 + 228{,}7 + 187{,}3}{3404} = \frac{1738{,}7}{3404} = 0{,}51.$$

Господствующее положение понимается как преобладание на рынке одного предприятия или группы предприятий. Количественным выражением такого преобладания может быть 50% (и более) объема производства, объема продаж или активов.

Рассчитаем показатели концентрации по данным табл. 5.5.

Коэффициенты Джини и Герфиндаля на основе имеющихся данных могут быть исчислены как для оптового товарооборота, так и для среднесписочной численности работников. Поскольку для оптового товарооборота показатели исчисляются аналогично показателям концентрации активов банка, найдем коэффициенты Джини и Герфиндаля лишь для измерения концентрации работников на предприятиях оптовой торговли, распределенных по мере возрастания товарооборота.

Из табл. 5.5 следует, что преобладают мелкие предприятия оптовой торговли. Так, с товарооборотом за апрель менее 1 тыс. руб. обнаружено 37,5% предприятий, на которых сосредоточено 10,5% общей численности работников $\left(\frac{T_i}{\sum_i T_i} 100\%\right)$ (см. графу 4 табл. 5.5). В то же время 42,5% общей численности работников занято на крупнейших предприятиях, которые составляют от общего числа 7,5%.

Используя данные граф 2 и 4 табл. 5.5, получаем два ряда накопленных частостей (в %): количество предприятий и численность работников на них (табл. 5.14).

Коэффициент Джини

$$G = \frac{22\,365{,}85 - 16\,969{,}58}{100 \cdot 100} = 0{,}54$$

Таблица 5.14

Расчет показателей концентрации

Объем оптового товарооборота за апрель 1995 г., тыс. руб.	Накопленная частость, %		$p_i q_{i+1}$	$p_{i+1} q_i$	$\left(\frac{T_i}{\sum_i T_i}\right)^2$
	Количество предприятий p_i	Численность работников q_i			
А	1	2	3	4	5
Менее 1	37,5	10,5	798,75	—	0,011
1–25	60,0	21,3	1614	630	0,012
25–50	68,7	26,9	2390,76	1463,31	0,003
50–100	77,4	34,8	3413,34	2082,06	0,006
100–200	85,2	44,1	4899	2964,96	0,009
200–500	92,5	57,5	9250	4079,25	0,018
Свыше 500	100,0	100,0	—	5750	0,181
Σ			22365,85	16969,58	0,240

свидетельствует о высоком уровне концентрации работников, что также подтверждает и коэффициент Герфиндаля

$$H = \sum_i \left(\frac{T_i}{\sum_i T_i}\right)^2 = 0{,}24.$$

Расчет коэффициента Герфиндаля приведен в табл. 5.14 (см. графу 5) на основе исходных данных табл. 5.5 (см. графу 4) (будучи возведенным в квадрат, удельный вес доминирующей группы остался значащим числом).

Основное достоинство коэффициента Герфиндаля – его высокая чувствительность к изменению в суммарном обороте долей крупнейших участников, что позволяет отслеживать концентрацию рыночного оборота. Другое достоинство данного коэффициента заключается в том, что он реагирует на число участников рынка. Однако его крупнейшим участникам придается наибольший вес. Вследствие этого существует опасность преувеличения уровня концентрации.

Наряду с коэффициентом Герфиндаля целесообразно применять ***коэффициент Лоренца***, который также характеризует концентрацию, степень неравномерности распределения доходов

путем сравнения долей численности единиц в группах (w_i) и долей значений признака в общем объеме $\left(\frac{x_i w_i}{\sum\limits_i x_i w_i} \text{ или } \frac{T_i}{\sum\limits_i T_i}\right)$.

Коэффициент Лоренца (L) исчисляется по формуле

$$L = \frac{1}{2}\sum_{i=1}^{n}\left|w_i - \frac{x_i w_i}{\sum\limits_i x_i w_i}\right|. \qquad (5.44)$$

Рассчитаем и сравним коэффициенты Лоренца и Герфиндаля на основе данных о распределении доходов населения России, представленных в табл. 5.15.

Таблица 5.15

Распределение общего объема денежных доходов населения России в январе–сентябре 2003 г.

20-процентные группы населения	Денежные доходы групп, % от общего числа
А	1
Первая (с наименьшими доходами)	5,6
Вторая	10,3
Третья	15,3
Четвертая	22,7
Пятая (с наивысшими доходами)	46,1
Итого	100,0

В каждой группе по 20% населения, следовательно, известны частости w_1, w_2, w_3, w_4, w_5 и все они равны 20%, или 0,2. В графе 1 табл. 5.15 даны удельные веса доходов групп от общего объема доходов, что можно обозначить как $\frac{x_i w_i}{\sum\limits_i x_i w_i}$. Следовательно, коэффициент Лоренца получит значение

$$L = \frac{1}{2}\left[\left|0{,}2 - 0{,}056\right| + \left|0{,}2 - 0{,}103\right| + \left|0{,}2 - 0{,}153\right| + \right.$$
$$\left. + \left|0{,}2 - 0{,}227\right| + \left|0{,}2 - 0{,}461\right|\right] = 0{,}228,$$

что свидетельствует о значительной степени социально-экономического расслоения населения. Коэффициент Лоренца прибли-

жался бы к нулю в случае совпадения долей численности населения в группах и долей доходов групп.

Коэффициент Герфиндаля получит следующее значение:

$$H = 0{,}056^2 + 0{,}103^2 + 0{,}153^2 + 0{,}227^2 + 0{,}461^2 = 0{,}30,$$

что свидетельствует о концентрации доходов в одной из групп, а именно в пятой группе, где сосредоточились 46,1% всех доходов.

5.6. Моменты распределения. Показатели формы распределения

Моменты распределения

Для подробного описания особенностей распределения используют дополнительные характеристики – моменты распределения, предложенные русским математиком П.Л. Чебышёвым.

Моментом k-го порядка называют среднюю из k-х степеней отклонений вариантов x от некоторой постоянной величины A:

$$M_k = \frac{\sum_i (x_i - A)^k m_i}{\sum_i m_i}. \qquad (5.45)$$

Порядок момента определяется величиной k. При исчислении моментов в качестве весов можно использовать частоты или частости. В зависимости от выбора постоянной величины A различают начальные, условные и центральные моменты.

1. Если $A = 0$, то моменты называются ***начальными***:

$$M_k = \frac{\sum_i x_i^k m_i}{\sum_i m_i}. \qquad (5.46)$$

При $k = 0$ получаем начальный момент нулевого порядка

$$M_0 = \frac{\sum_i x_i^0 m_i}{\sum_i m_i} = 1;$$

при $k = 1$ – начальный момент первого порядка

$$M_1 = \frac{\sum_i x_i m_i}{\sum_i m_i} = \bar{x}.$$

Примечание. Согласно определению начальный момент первого порядка есть средняя арифметическая;

при $k = 2$ – начальный момент второго порядка

$$M_2 = \frac{\sum_i x_i^2 m_i}{\sum_i m_i} = \overline{x^2}$$

и т.д.

Практически используют моменты первых четырех порядков.

2. Если A равно не нулю, а некоторой произвольной величине x_0 (начало отсчета), то моменты называются ***начальными относительно*** x_0 или ***условными***:

$$M'_k = \frac{\sum_i (x_i - x_0)^k m_i}{\sum_i m_i}. \qquad (5.47)$$

С их помощью упрощают вычисления основных характеристик.

При $k = 0$ получаем начальный момент относительно x_0 нулевого порядка

$$M'_0 = \frac{\sum_i (x_i - x_0)^0 m_i}{\sum_i m_i} = 1;$$

при $k = 1$ – первого порядка

$$M'_1 = \frac{\sum_i (x_i - x_0) m_i}{\sum_i m_i} = \frac{\sum_i x_i m_i}{\sum_i m_i} - \frac{\sum_i x_0 m_i}{\sum_i m_i} = \bar{x} - \frac{x_0 \sum_i m_i}{\sum_i m_i} = \bar{x} - x_0$$

и т.д.

Из последней формулы вытекает, что $\bar{x} = M'_1 + x_0$, т.е. средняя арифметическая равна условному моменту первого порядка плюс начало отсчета.

Если отклонения $(x_i - x_0)$ имеют общий множитель c, то на него можно разделить отклонения, а по окончании вычислений полученный момент умножить на этот множитель в соответствующей степени. Таким образом,

$$M'_1 = \frac{\sum_i \left(\frac{x_i - x_0}{c} \right)^1 m_i}{\sum_i m_i}.$$

Отсюда следует, что $\bar{x} = M'_1 c + x_0$.

Вычисление средней методом отсчета от условного нуля иногда называют *методом моментов.*

Практически начальные моменты относительно x_0 определяют следующим образом:

- из всех вариантов вычитают начало отсчета и находят отклонения $(x_i - x_0)$;
- делят эти отклонения на общий множитель: $\dfrac{x_i - x_0}{c} = x'$;
- вычисляют начальные моменты относительно x';
- умножают найденные начальные моменты на c.

В результате такого умножения получают искомые начальные моменты относительно x_0.

3. Если за постоянную величину A принять среднюю (т.е. $A = \bar{x}$), то моменты называются ***центральными*** и обозначаются μ_k:

$$\mu_k = \frac{\sum_i (x_i - \bar{x})^k m_i}{\sum_i m_i} = \overline{(x_i - \bar{x})^k}. \qquad (5.48)$$

Центральный момент нулевого порядка равен 1:

$$\mu_0 = \frac{\sum_i (x_i - \bar{x})^0 m_i}{\sum_i m_i} = 1;$$

центральный момент первого порядка равен 0:

$$\mu_1 = \frac{\sum_i (x_i - \bar{x}) m_i}{\sum_i m_i} = \frac{\sum_i x_i m_i}{\sum_i m_i} - \frac{\sum_i \bar{x} m_i}{\sum_i m_i} = \bar{x} - \bar{x} = 0;$$

центральный момент второго порядка равен дисперсии:

$$\mu_2 = \frac{\sum_i (x_i - \bar{x})^2 m_i}{\sum_i m_i} = \sigma^2;$$

центральный момент третьего порядка

$$\mu_3 = \frac{\sum_i (x_i - \bar{x})^3 m_i}{\sum_i m_i} = \overline{(x_i - \bar{x})^3}.$$

Показатели формы распределения

Центральный момент третьего порядка используется при исчислении показателя асимметрии распределения. Для того чтобы показатель асимметрии не зависел от масштаба, выбранного при измерении варианта, вводят безразмерную характеристику – ***коэффициент асимметрии*** (нормированный момент третьего порядка):

$$r_3 = \frac{\mu_3}{\sigma^3}. \qquad (5.49)$$

При симметричном распределении варианты, равноудаленные от $\bar{x}$, имеют одинаковую частоту, поэтому $\mu_3 = 0$, а следовательно, и $r_3 = 0$. Если $r_3 < 0$, то в вариационном ряду преобладают (имеют бóльшую частоту) варианты, которые меньше, чем средняя, т.е. ряд отрицательно асимметричен (или с левосторонней скошенностью – более длинная ветвь влево). Положительная асимметрия (правосторонняя скошенность – более длинная ветвь вправо) характеризуется значением $r_3 > 0$ (рис. 5.6).

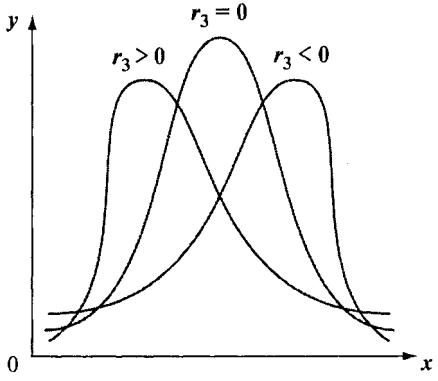


Рис. 5.6. Асимметрия распределения

В качестве показателя асимметрии применяется и ***коэффициент асимметрии Пирсона***, представляющий собой отношение разности между средней арифметической и модой к среднему квадратическому отклонению:

$$As = \frac{\bar{x} - Mo}{\sigma}. \qquad (5.50)$$

Если $As > 0$, скошенность правосторонняя (как и для r_3); если $As < 0$, скошенность левосторонняя; если $As = 0$, вариационный ряд симметричен.

Для характеристики крутизны распределения используется центральный момент четвертого порядка

$$\mu_4 = \frac{\sum_i (x_i - \bar{x})^4 m_i}{\sum_i m_i} = \overline{(x_i - \bar{x})^4}.$$

Для образования безразмерной характеристики определяется отношение $\frac{\mu_4}{\sigma^4} = r_4$ (нормированный момент четвертого порядка). Данное отношение и характеризует крутизну (заостренность) графика распределения.

При измерении асимметрии эталоном служит симметричное распределение, для которого $r_3 = 0$. Аналогично при оценке крутизны в качестве эталонного выбирается так называемое нормальное (симметричное) распределение, которое будет подробно рассмотрено в подпараграфе 5.7.1.

Для нормального распределения $\frac{\mu_4}{\sigma^4} = 3$, поэтому для оценки крутизны данного распределения в сравнении с нормальным вычисляется ***эксцесс распределения*** (рис. 5.7):

$$Ex = \frac{\mu_4}{\sigma^4} - 3. \qquad (5.51)$$

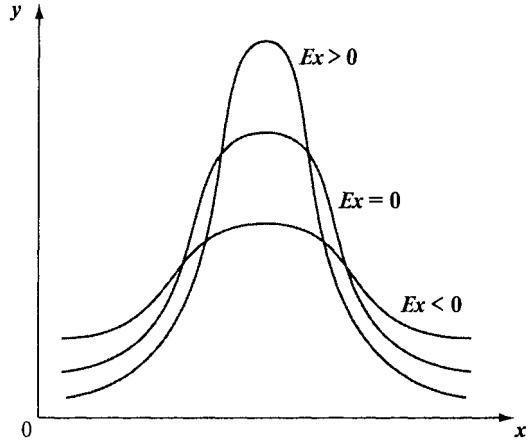


Рис. 5.7. Эксцесс распределения

Пример. По данным табл. 5.1 найти коэффициент асимметрии и нормированные моменты третьего и четвертого порядков.

Необходимые для расчета нормированных моментов величины приведены в табл. 5.16.

Таблица 5.16

Расчет нормированных моментов

x_i	m_i	$x_i m_i$	$x_i - \bar{x}$	$(x_i - \bar{x})^2 m_i$	$(x_i - \bar{x})^3 m_i$	$(x_i - \bar{x})^4 m_i$
A	1	2	3	4	5	6
1	223	223	−1,669	621,180	−1036,7	1730,3
2	276	552	−0,669	123,527	−82,6	55,3
3	238	714	0,331	26,076	8,6	2,8
4	170	680	1,331	301,165	400,8	533,5
5	58	290	2,331	315,146	734,6	1712,4
6	35	210	3,331	388,345	1293,6	4309,0
Σ	1000	2669	–	1775,439	1318,3	8343,3

Сначала найдем среднюю арифметическую:

$$\bar{x} = \frac{\sum\limits_i x_i m_i}{\sum\limits_i m_i} = \frac{2669}{1000} = 2{,}669 \text{ чел.}$$

Далее найдем среднее квадратическое отклонение (стандарт) (см. графу 4 табл. 5.16):

$$\sigma = \sqrt{\frac{\sum\limits_i (x_i - \bar{x})^2 m_i}{\sum\limits_i m_i}} = \sqrt{\frac{1775{,}439}{1000}} = \sqrt{1{,}775439} = 1{,}33.$$

Центральный момент третьего порядка (см. графу 5 табл. 5.16)

$$\mu_3 = \frac{\sum\limits_i (x_i - \bar{x})^3 m_i}{\sum\limits_i m_i} = \frac{1318{,}3}{1000} = 1{,}3183,$$

а нормированный момент третьего порядка

$$r_3 = \frac{\mu_3}{\sigma^3} = \frac{1{,}3183}{1{,}33^3} = 0{,}56.$$

Коэффициент асимметрии Пирсона

$$K_a = \frac{\bar{x} - Mo}{\sigma} = \frac{2{,}669 - 2}{1{,}33} = 0{,}50$$

(мода была найдена ранее в параграфе 5.2).

В данном случае асимметрия заметная и скошенность правосторонняя.

Найдем центральный момент четвертого порядка (см. графу 6 табл. 5.16):

$$\mu_4 = \frac{\sum_i (x_i - \bar{x})^4 m_i}{\sum_i m_i} = \frac{8343{,}3}{1000} \approx 8{,}343,$$

а также нормированный момент четвертого порядка

$$r_4 = \frac{\mu_4}{\sigma^4} = \frac{8{,}343}{1{,}33^4} = 2{,}666.$$

Эксцесс распределения

$$Ex = r_4 - 3 = 2{,}666 - 3 = -0{,}334.$$

Так как $Ex < 0$, распределение низковершинное.

5.7. Теоретические кривые распределения

Анализ вариационных рядов предполагает выявление закономерностей распределения, определение и построение (получение) некой теоретической (вероятностной) формы распределения. Характер распределения лучше всего проявляется при большом числе наблюдений и малых интервалах. В этом случае графическое изображение эмпирического вариационного ряда принимает вид плавной кривой, именуемой *кривой распределения.* Кривая распределения может рассматриваться как некая теоретическая (вероятностная) форма распределения, свойственная определенной совокупности в конкретных условиях.

Таким образом, анализируя частоты в эмпирическом распределении, можно описать его с помощью математической модели – закона распределения, установить по исходным данным параметры теоретической кривой и проверить правильность выдвинутой гипотезы о типе распределения данного ряда.

При исследовании закономерностей распределения очень важно выдвинуть верную гипотезу о типе кривой распределения, так как, если кривая описана математически (с помощью уравнения) верно, она более точно отражает закономерности данного рас-

пределения и может быть использована в различных практических расчетах и прогнозах. Кроме того, в этом случае можно сформулировать рекомендации для принятия практических решений.

Что понимается под теоретическим распределением? Это гипотетическое распределение вероятностей, которое предполагается для наблюдаемых частот вариационного ряда.

В практике статистического исследования встречаются различные распределения: нормальное, логарифмически нормальное, биномиальное, Пуассона, Шарлье и др. Каждое распределение имеет свою специфику и область применения. Далее будут рассмотрены только нормальное распределение и распределение Пуассона.

5.7.1. Нормальное распределение

При построении статистических моделей весьма широко применяется нормальное распределение.

В 1727 г. английский математик Абрахам де Муавр (1667–1754) открыл закон распределения вероятностей, названный законом нормального распределения. Позднее, в начале XIX в., разработкой вопросов, относящихся к данному закону, занимались Пьер Лаплас (1749–1827) и Карл Гаусс (1777–1855). Общие условия возникновения закона нормального распределения установил А.М. Ляпунов (1857–1918).

Распределение непрерывной случайной величины x называют ***нормальным*** $N(x, \sigma)$, если соответствующая ей плотность распределения выражается формулой

$$f(x) = \varphi(x, \bar{x}, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\bar{x})^2}{2\sigma^2}}$$

$$\text{или} \quad \varphi(t) = \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}, \tag{5.52}$$

где x – значение изучаемого признака;

$\bar{x}$ – средняя арифметическая ряда;

σ^2 – дисперсия значений изучаемого признака;

σ – среднее квадратическое отклонение изучаемого признака;

$\pi = 3{,}1415$ – постоянное число (отношение длины окружности к ее диаметру);

$e = 2{,}7182$ – основание натурального логарифма;

$t = \dfrac{x - \bar{x}}{\sigma}$ – нормированное отклонение.

При графическом изображении плотности распределения $f(x)$ получим кривую нормального распределения, симметричную относительно вертикальной прямой $x = \bar{x}$ (рис. 5.8), поэтому величину $\bar{x}$ называют *центром распределения.*

Случайные величины, распределенные по нормальному закону, различаются значениями параметров $\bar{x}$ и σ, поэтому очень важно выяснить, как эти параметры влияют на вид нормальной кривой.

Если $\bar{x}$ не меняется, а изменяется только σ, то:

1) чем меньше σ, тем более вытянута вверх кривая (см. рис. 5.8, *а*), а так как площадь, ограниченная осью x и данной кривой, равна 1, то вытягивание вверх компенсируется сжатием около центра распределения $\bar{x}$ и более быстрым приближением кривой к оси абсцисс;
2) чем больше σ, тем более плоской и растянутой вдоль оси абсцисс становится кривая.

Если σ остается неизменной, а $\bar{x}$ изменяется, то кривые нормального распределения имеют одинаковую форму, но отличаются друг от друга положением максимальной ординаты (см. рис. 5.8, *б*).

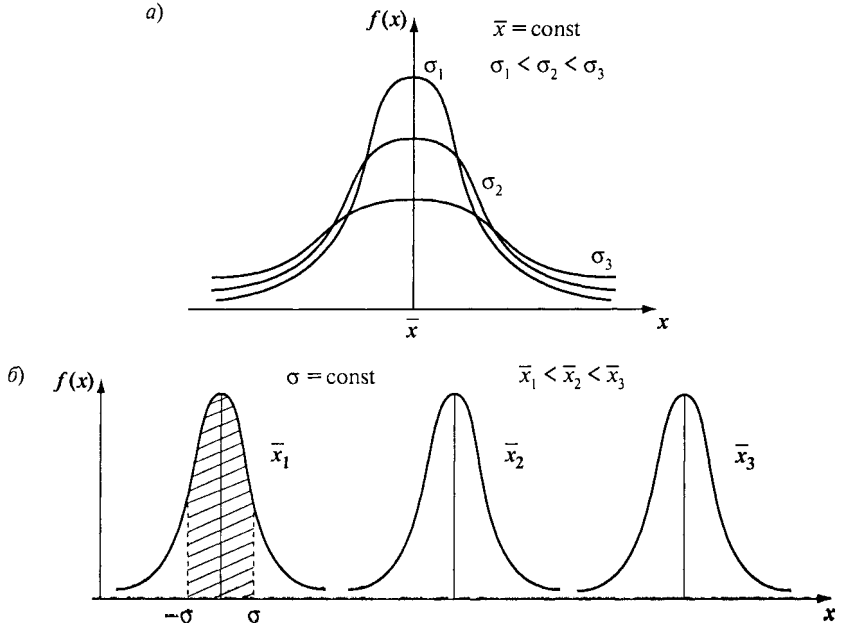


Рис. 5.8. Кривые нормального распределения

Итак, выделим **особенности кривой нормального распределения.**

1. Кривая симметрична и имеет максимум в точке, соответствующей значению $\bar{x} = Mo = Me$.
2. Кривая асимптотически приближается к оси абсцисс, продолжаясь в обе стороны до бесконечности. Чем больше отдельные значения x отклоняются от $\bar{x}$, тем реже они встречаются.
3. Кривая имеет две точки перегиба на расстоянии $\pm\sigma$ от $\bar{x}$.
4. Площадь между ординатами, проведенными на расстоянии $\bar{x} \pm \sigma$ (заштрихованная область на рис. 5.8, *б*), составляет 0,683. Это означает, что 68,3% всех исследуемых единиц (частот) отклоняется от средней арифметической не более чем на σ, т.е. находится в пределах $\bar{x} \pm \sigma$. В промежутке $\bar{x} \pm 2\sigma$ находится 95,4%, а в промежутке $\bar{x} \pm 3\sigma$, соответственно, 99,7% всех единиц исследуемой совокупности.
5. Коэффициенты асимметрии и эксцесса равны нулю.

Порядок расчета теоретических частот кривой нормального распределения таков:

- по эмпирическим данным рассчитывают среднюю арифметическую ряда $\bar{x}$ и среднее квадратическое отклонение σ;
- находят нормированное отклонение каждого варианта от средней арифметической:

$$t = \frac{x - \bar{x}}{\sigma};$$

- по таблице распределения функции $\varphi(t)$ (см. Приложение 1) определяют ее значения;
- вычисляют теоретические частоты m' по формуле

$$m' = \frac{Nh_k}{\sigma}\varphi(t),$$

где N – объем совокупности;

h_k – длина интервала.

В случае если вариационный ряд имеет равные интервалы,

$$\frac{Nh_k}{\sigma} = \text{const}.$$

Пример. Рассчитать теоретические частоты ряда распределения на основе данных, представленных в табл. 5.17.

Выдвинув гипотезу о нормальном распределении, определим по эмпирическим данным параметры этой кривой.

Таблица 5.17

Распределение призывников района по росту
(данные условные)

Рост призывников, см $x_{k-1}-x_k$	Количество человек m_i	x_i	$x_i m_i$	$(x_i-\bar{x})^2 m_i$	$t=\frac{x_i-\bar{x}}{\sigma}$	$\varphi(t)$	m_i'
А	1	2	3	4	5	6	7
156–160	8	157,5	1260,0	2918,48	2,34	0,0258	5
161–165	17	162,5	2762,5	3379,77	1,73	0,0893	16
166–170	42	167,5	7035,0	3478,02	1,11	0,2155	40
171–175	54	172,5	9315,0	907,74	0,50	0,3521	65
176–180	73	177,5	12957,5	59,13	0,11	0,3965	73
181–185	57	182,5	10402,5	1984,17	0,72	0,3079	57
186–190	38	187,5	7125,0	4514,78	1,33	0,1647	30
191–195	11	192,5	2117,5	2780,91	1,95	0,0596	11
Σ	300		52975,0	20023,00			297

Сначала рассчитаем средний уровень ряда:

$$\bar{x} = \frac{\sum\limits_i x_i m_i}{\sum\limits_i m_i} = \frac{52975}{300} = 176{,}6 \text{ см}.$$

Затем определим еще один параметр – среднее квадратическое отклонение, для чего предварительно вычислим дисперсию (см. графу 4 табл. 5.17):

$$\sigma^2 = \frac{\sum\limits_i (x_i - \bar{x})^2 m_i}{\sum\limits_i m_i} = \frac{20023}{300} = 66{,}74,$$

отсюда $\sigma = 8{,}17$ см.

Далее определим нормированное отклонение t для каждого варианта (см. графу 5 табл. 5.17) с точностью до сотых, после чего по таблице распределения функции $\varphi(t)$ (см. Приложение 1) найдем значения функции при значениях аргумента, полученных в графе 5. При этом необходимо учитывать, что функция $\varphi(t)$ четная, т.е. $\varphi(-t) = \varphi(t)$.

Анализируемый вариационный ряд имеет равные интервалы, следовательно, можно определить

$$\text{const} = \frac{Nh_k}{\sigma} = \frac{300 \cdot 5}{8{,}17} = 183{,}6 \cong 184.$$

Последовательно умножив const на величину $\varphi(t)$ для каждого варианта, получим теоретические частоты m' (см. графу 7 табл. 5.17).

Иногда за счет округлений при расчетах может быть нарушено равенство сумм эмпирических и теоретических частот, что и произошло в данном случае ($\sum_i m_i = 300$, $\sum_i m'_i = 297$).

Сравним на графике эмпирические m и теоретические m' частоты, полученные на основе данных табл. 5.17 (рис. 5.9). Близость этих частот очевидна, но объективная оценка их соответствия может быть получена только с помощью критериев согласия (см. параграф 5.8).

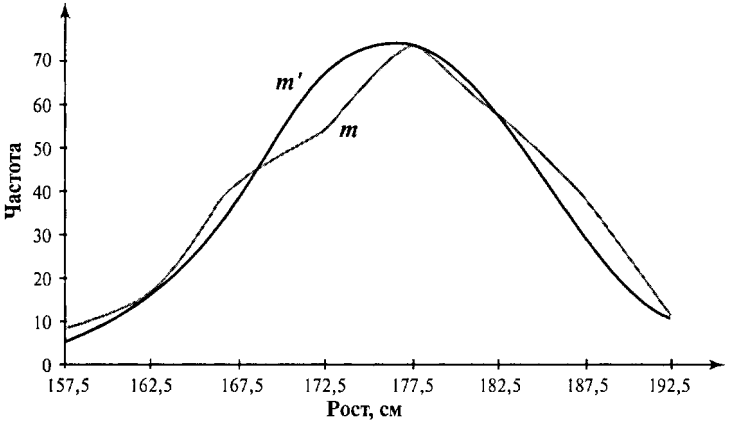


Рис. 5.9. Распределение призывников по росту

5.7.2. Распределение Пуассона

К числу важнейших теоретических распределений, имеющих практическое применение, относится пуассоновское распределение, названное по фамилии французского математика Симеона Пуассона (1781–1840).

Классическую форму распределение Пуассона принимает в том случае, если значения признака носят дискретный характер $x = 0, 1, 2, 3, \ldots$ и являются результатом какого-либо редко возникающего события среди наблюдаемых единиц. Причем с увеличением значений признака вероятность наступления события падает. Природа распределения Пуассона наиболее полно раскрывается в теории случайных процессов, поэтому его еще называют законом

распределения редких явлений. Распределение Пуассона наблюдается в совокупностях, число единиц которых достаточно велико ($N \geqslant 100$), а доля единиц, обладающих большими значениями признака, мала.

Средняя арифметическая ряда и дисперсия, вычисленные по эмпирическим данным, как правило, совпадают или мало отличаются друг от друга.

Графически распределение Пуассона представлено на рис. 5.10.

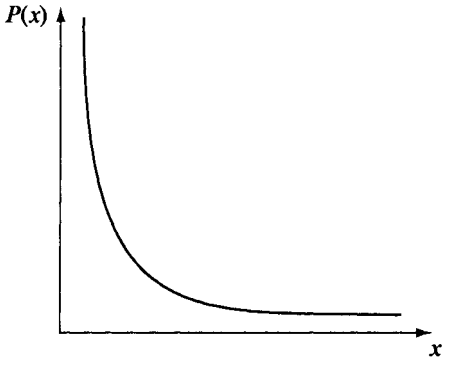


Рис. 5.10. Кривая распределения Пуассона

Аналитически распределение Пуассона можно выразить формулой

$$P(x) = \frac{a^x e^{-a}}{x!}, \qquad (5.53)$$

где $P(x)$ – вероятность того, что признак примет то или иное значение;

$a = \bar{x}$ – средняя арифметическая ряда.

Из формулы (5.53) видно, что единственным параметром распределения является средняя арифметическая.

Порядок расчета теоретических частот кривой распределения Пуассона таков:

- находят среднюю арифметическую ряда, т.е. $\bar{x} = a$;
- по таблицам определяют e^{-a};
- для каждого значения x вычисляют теоретическую частоту по формуле

$$m' = N\frac{a^x e^{-a}}{x!} = NP(x), \qquad (5.54)$$

где N – число единиц в изучаемой совокупности.

Пример. Государственная инспекция безопасности дорожного движения (ГИБДД) провела проверку 400 автомобилей. Результаты этой проверки представлены в табл. 5.18. Выровнять полученный ряд по кривой Пуассона.

Таблица 5.18

Распределение автомобилей по числу неисправностей

Число неисправностей в автомобиле x_i	Количество автомобилей m_i	$x_i m_i$	$(x_i - \bar{x})^2 m_i$	m_i'
0	215	0	77,40	213
1	135	135	21,60	134
2	38	76	74,48	42
3	8	24	46,08	9
4	3	12	34,68	1
5	1	5	19,36	1
Σ	400	252	273,60	400

Рассчитаем среднюю арифметическую ряда и дисперсию:

$$\bar{x} = \frac{\sum_i x_i m_i}{\sum_i m_i} = \frac{252}{400} = 0{,}63; \qquad \sigma^2 = \frac{\sum_i (x_i - \bar{x})^2 m_i}{\sum_i m_i} = \frac{273{,}6}{400} = 0{,}68.$$

Так как $\bar{x} \cong \sigma^2$, есть основание полагать, что данное распределение подчиняется закону Пуассона. Для определения теоретических частот выполним следующие расчеты.

Поскольку $a = \bar{x} = 0{,}63$, по таблице Приложения 11 найдем значение $e^{-0{,}63} = 0{,}5326$. Затем, подставив в формулу (5.54) значения $\bar{x}$ от 0 до 5, вычислим теоретические частоты:

$$m_0' = 400\frac{0{,}63^0 \cdot 0{,}5326}{0!} = 213{,}0 \text{ (так как } 0! = 1);$$

$$m_1' = 400\frac{0{,}63^1 \cdot 0{,}5326}{1} = 134{,}2;$$

$$m_2' = 400\frac{0{,}63^2 \cdot 0{,}5326}{1 \cdot 2} = 42{,}3;$$

$$m_3' = 400\frac{0{,}63^3 \cdot 0{,}5326}{1 \cdot 2 \cdot 3} = 8{,}9;$$

$$m_4' = 400\frac{0{,}63^4 \cdot 0{,}5326}{1 \cdot 2 \cdot 3 \cdot 4} = 1{,}4;$$

$$m_5' = 400\frac{0{,}63^5 \cdot 0{,}5326}{1 \cdot 2 \cdot 3 \cdot 4 \cdot 5} = 0{,}2.$$

Полученные теоретические частоты округлим до целых чисел (см. последнюю графу табл. 5.18).

Распределение Пуассона используется в теории надежности, теории массового обслуживания для описания числа заявок, поступающих в систему массового обслуживания (например, такую, как телефонная станция, ремонтная мастерская, торговое предприятие, билетная касса) в единицу времени, а также для описания числа отказов технологического оборудования в единицу времени.

5.8. Критерии согласия

Так как все предположения о характере того или иного распределения – это гипотезы, а не категорические утверждения, то они, естественно, должны быть подвергнуты статистической проверке с помощью так называемых *критериев согласия*. Критерии согласия, опираясь на установленный закон распределения, дают возможность установить, когда расхождения между теоретическими и эмпирическими частотами следует признать несущественными (случайными), а когда – существенными (неслучайными). Таким образом, критерии согласия позволяют отвергнуть или подтвердить правильность выдвинутой при выравнивании ряда гипотезы о характере распределения в эмпирическом ряду и дать ответ, можно ли принять для данного эмпирического распределения модель, выраженную некоторым теоретическим законом распределения.

Существует ряд критериев согласия. Чаще других применяют критерии Пирсона, Романовского и Колмогорова. Рассмотрим их.

Критерий согласия Пирсона χ^2 (хи-квадрат) – один из основных критериев согласия. Критерий предложен английским математиком Карлом Пирсоном (1857–1936) для оценки случайности (существенности) расхождений между частотами эмпирического и теоретического распределений. Критерий Пирсона

$$\chi^2 = \sum_{i=1}^{k} \frac{\left(m_i - m_i'\right)^2}{m_i'},$$

где k – число групп, на которые разбито эмпирическое распределение;

m_i – наблюдаемая частота признака в i-й группе;

m_i' – теоретическая частота, рассчитанная по предполагаемому распределению.

Для распределения χ^2 составлены таблицы, где указано критическое значение критерия согласия χ^2 для выбранного уровня значимости α и данного числа степеней свободы ν (см. Приложение 4).

Уровень значимости α – вероятность ошибочного отклонения выдвинутой гипотезы, т.е. вероятность того, что будет отвергнута правильная гипотеза. В статистических исследованиях в зависимости от важности и ответственности решаемых задач пользуются следующими тремя уровнями значимости:

1) $\alpha = 0{,}10$, тогда $P = 0{,}90$;
2) $\alpha = 0{,}05$, тогда $P = 0{,}95$;
3) $\alpha = 0{,}01$, тогда $P = 0{,}99$.

Например, вероятность 0,01 означает, что в одном случае из 100 может быть отвергнута правильная гипотеза. В экономических исследованиях считается практически приемлемой вероятность ошибки 0,05, т.е. в 5 случаях из 100 может быть отвергнута правильная гипотеза.

Кроме того, χ^2-критерий, определяемый по таблице, зависит и от числа степеней свободы. *Число степеней свободы* ν определяется как число групп в ряду распределения k минус число связей z:

$$\nu = k - z.$$

Под *числом связей* понимается число показателей эмпирического ряда, использованных при исчислении теоретических частот, т.е. показателей, связывающих эмпирические и теоретические частоты $\left(\bar{x},\ \sigma,\ \sum_i m_i\right)$.

Так, в случае выравнивания по кривой нормального распределения имеется т р и связи:

$$\bar{x}_{\text{эмп}} = \bar{x}'_{\text{теор}};\qquad \sigma_{\text{эмп}} = \sigma'_{\text{теор}};\qquad \sum_i m_{i\,\text{эмп}} = \sum_i m'_{i\,\text{теор}}.$$

Поэтому при выравнивании по кривой нормального распределения число степеней свободы определяется как $\nu = k - 3$, где k – число групп в ряду.

В случае выравнивания по кривой Пуассона $\nu = k - 2$, так как при построении частот используются д в е ограничивающие связи: $\bar{x}$, $\sum_i m_i$.

Для оценки существенности расчетное значение $\chi^2_{\text{расч}}$ сравнивается с табличным $\chi^2_{\text{табл}}$.

При полном совпадении теоретического и эмпирического распределений $\chi^2 = 0$, в противном случае $\chi^2 > 0$.

Если $\chi^2_{\text{расч}} > \chi^2_{\text{табл}}$, то при заданном уровне значимости α и числе степеней свободы ν гипотезу о несущественности (случайности) расхождений отклоняем.

В случае если $\chi^2_{\text{расч}} \leqslant \chi^2_{\text{табл}}$, заключаем, что эмпирический ряд хорошо согласуется с гипотезой о предполагаемом распределении и с вероятностью $(1 - \alpha)$ можно утверждать, что расхождение между теоретическими и эмпирическими частотами случайно.

Используя критерий согласия χ^2, необходимо соблюдать следующие условия:

1) объем исследуемой совокупности должен быть достаточно большим ($N \geqslant 50$), при этом частота или численность каждой группы должна быть не менее 5. Если это условие нарушается, необходимо предварительно объединить маленькие частоты;
2) эмпирическое распределение должно состоять из данных, полученных в результате случайного отбора, т.е. они должны быть независимыми.

Если в эмпирическом ряду распределение задано частостями $\left(w_i = \dfrac{m_i}{\sum\limits_i m_i} \right)$, то χ^2 следует исчислять по формуле

$$\chi^2 = N \sum \frac{(w - w')^2}{w'}.$$

Критерий Романовского $K_{\text{Р}}$ основан на использовании критерия Пирсона χ^2, т.е. уже найденных значений χ^2, и числа степеней свободы ν:

$$K_{\text{Р}} = \frac{\left| \chi^2 - \nu \right|}{\sqrt{2\nu}}.$$

Он весьма удобен при отсутствии таблиц для χ^2.

Если $K_{\text{Р}} < 3$, то расхождения между теоретическим и эмпирическим распределением случайны, если же $K_{\text{Р}} > 3$, то не случайны и, соответственно, теоретическое распределение не может служить моделью для изучаемого эмпирического распределения.

Критерий Колмогорова λ основан на определении максимального расхождения между накопленными частотами или частостями эмпирических и теоретических распределений:

$$\lambda = \frac{D}{\sqrt{N}} \quad \text{или} \quad \lambda = d\sqrt{N},$$

где D и d — соответственно максимальная разность между накопленными частотами $(F - F')$ и между накоплен-

ными частостями ($p - p'$) эмпирического и теоретического рядов распределений;

N – число единиц в совокупности.

Рассчитав значение λ, по таблице $P(\lambda)$ (см. Приложение 6) определяют вероятность, с которой можно утверждать, что отклонения эмпирических частот от теоретических случайны. Вероятность $P(\lambda)$ может изменяться от 0 до 1. При $P(\lambda) = 1$ происходит полное совпадение частот, при $P(\lambda) = 0$ – полное расхождение. Если λ принимает значения до 0,3, то $P(\lambda) = 1$.

Основное условие для использования критерия Колмогорова – достаточно большое число наблюдений.

Пример. Используя данные табл. 5.17, проверить правильность выдвинутой гипотезы о распределении призывников района по закону нормального распределения. Величины, необходимые для расчета критериев согласия, приведены в табл. 5.19.

Таблица 5.19

Расчет величин для определения критериев согласия Пирсона χ^2 и Колмогорова λ

Рост, см	Частоты ряда распределения		$\frac{(m - m')^2}{m'}$	F	F'	$\lvert F - F' \rvert$
	m	m'				
А	1	2	3	4	5	6
156–160	8	5	1,8	8	5	3
161–165	17	16	0,1	25	21	4
166–170	42	40	0,1	67	61	6
171–175	54	65	1,9	121	126	5
176–180	73	73	0	194	199	5
181–185	57	57	0	251	256	5
186–190	38	30	2,1	289	286	3
191–195	11	11	0	300	297	3
Σ	300	297	6,0			

Сначала рассчитаем критерий Пирсона

$$\chi^2_{\text{расч}} = \sum \frac{(m - m')^2}{m'} = 6{,}0.$$

Затем выберем уровень значимости $\alpha = 0{,}05$ и определим число степеней свободы ν. В данном распределении 8 групп и число связей (параметров) равно 3, следовательно, $\nu = 8 - 3 = 5$. По

таблице Приложения 4 найдем χ^2: при $\alpha = 0{,}05$ и $\nu = 5$ критерий Пирсона $\chi^2 = 11{,}07$.

Так как $\chi^2_{\text{расч}} < \chi^2_{\text{табл}}$, с вероятностью 0,95 можно утверждать, что в основе эмпирического распределения призывников по росту лежит закон нормального распределения, т.е. выдвинутая гипотеза не отвергается, а расхождения объясняются случайными факторами.

Проверим выдвинутую гипотезу, используя критерий Романовского:

$$K_{\text{Р}} = \frac{\left|\chi^2 - \nu\right|}{\sqrt{2\nu}} = \frac{\left|6{,}0 - 5\right|}{\sqrt{2 \cdot 5}} = \frac{1}{3{,}16} = 0{,}3.$$

Так как $K_{\text{Р}} < 3$, гипотеза не отвергается.

Критерий Романовского также подтверждает, что расхождения между эмпирическими и теоретическими частотами несущественны.

Рассмотрим теперь применение критерия Колмогорова λ. Как видно из табл. 5.19, максимальная разность между кумулятивными частотами равна 6, т.е. $D = \max\left|F - F'\right| = 6$. Следовательно, критерий Колмогорова

$$\lambda = \frac{D}{\sqrt{N}} = \frac{6}{\sqrt{300}} = 0{,}35.$$

По таблице Приложения 6 находим значение вероятности при $\lambda = 0{,}35$: $P(\lambda) = 0{,}9997$. Это означает, что с вероятностью, близкой к единице, можно утверждать, что гипотеза о нормальном распределении не отвергается, а расхождения эмпирического и теоретического распределений носят случайный характер.

Теперь, подтвердив правильность выдвинутой гипотезы с помощью известных критериев согласия, можно использовать результаты распределения для практической деятельности.

Пример. Используя данные табл. 5.18, проверить гипотезу о подчинении распределения числа неисправностей в автомобилях закону Пуассона.

Исходные данные и расчет величин, необходимых для определения критериев согласия, приведены в табл. 5.20.

Подсчитаем величину χ^2:

$$\chi^2_{\text{расч}} = \sum \frac{(m - m')^2}{m'} = 4{,}21$$

(см. табл. 5.20).

Таблица 5.20

x_i	m	m'	$\frac{(m-m')^2}{m'}$	F	F'	$\lvert F-F'\rvert$
А	1	2	3	4	5	6
0	215	220	0,11	215	220	5
1	135	132	0,07	350	352	2
2	38	39	0,03	388	391	3
3	8	8	0	396	399	3
4	3	1	4	399	400	1
5	1	0	0	400	400	0
Σ	400	400	4,21			

При уровне значимости $\alpha = 0{,}05$ и числе степеней свободы $\nu = 6 - 2 = 4$

$$\chi^2_{\text{табл}} = 9{,}49$$

(см. Приложение 4).

Поскольку $\chi^2_{\text{расч}} < \chi^2_{\text{табл}}$, можно сделать вывод о том, что расхождения эмпирических и теоретических частот случайны.

Таким образом, выдвинутая гипотеза о распределении числа неисправностей в автомобилях по закону Пуассона не отвергается.

Критерий Романовского также подтверждает выдвинутую гипотезу:

$$K_{\text{Р}} = \frac{\lvert \chi^2 - \nu \rvert}{\sqrt{2\nu}} = \frac{\lvert 4{,}21 - 4 \rvert}{\sqrt{2 \cdot 4}} = 0{,}05 < 3.$$

Для определения критерия Колмогорова λ необходимо вычислить F и F' (соответственно накопленные эмпирические и теоретические частоты) и найти максимальное расхождение между ними (см. графы 4–6 табл. 5.20):

$$D = \max \lvert F - F' \rvert = 5.$$

Так как $N = \sum m = 400$, то

$$\lambda = \frac{5}{\sqrt{400}} = 0{,}25.$$

Можно, не обращаясь к таблице Приложения 6, согласно сформулированному выше принципу, сделать вывод о том, что расхождения между эмпирическим и теоретическим распределениями несущественны. На этом основании выдвинутая гипотеза о распределении неисправностей в автомобилях по закону Пуассона не отвергается.

Глава 6
ВЫБОРОЧНОЕ НАБЛЮДЕНИЕ

6.1. Общая характеристика выборочного наблюдения

Наиболее широко распространенным видом несплошного наблюдения является выборочное наблюдение, при котором обследуются не все единицы изучаемой совокупности, а лишь определенным образом отобранная их часть. Вся совокупность единиц, из которой осуществляется отбор, называется ***генеральной совокупностью***, а единицы, отобранные для непосредственного наблюдения, представляют собой ***выборочную совокупность***, или просто ***выборку***. Отбор из генеральной совокупности проводится таким образом, чтобы на основе выборки можно было получить достаточно точное представление об основных параметрах совокупности в целом. При этом речь идет как о точечной оценке, в качестве которой принимается соответствующее значение средней, доли и т.д., полученное в результате выборки, так и об интервальной оценке, т.е. о тех пределах, в которых с определенной вероятностью может находиться значение искомого параметра в генеральной совокупности. Главное требование, которому должна отвечать выборочная совокупность, – это требование ее р е п р е з е н т а т и в н о с т и, т.е. представительности.

В статистике результаты сплошного наблюдения иногда оцениваются как выборочные характеристики. Такая трактовка полученных данных имеет место в тех случаях, когда число обследованных единиц невелико и нет твердой уверенности в том, что изучаемые характеристики не могут принимать иных значений, кроме выявленных в результате наблюдения. При проведении экспериментов число значений может быть бесконечно большим, поэтому, формулируя выводы на основе ограниченного их числа, необходимо рассматривать полученные данные как выборочные характеристики.

При организации выборочного обследования нужно соблюдать принцип с л у ч а й н о с т и отбора. Каждая единица совокупности должна иметь равную вероятность попасть в выборку. На практике не всегда удается обеспечить соблюдение данного принципа. Для этого необходимо учесть все элементы генеральной совокупности. Например, невозможно пронумеровать все домаш-

ние хозяйства или все население страны, так как это очень большая совокупность и состав ее постоянно меняется. В таких случаях прибегают к методике неслучайного отбора, стараясь, чтобы элементы случайности присутствовали. Примером такого отбора служит механическая выборка, при которой вся исследуемая совокупность предварительно упорядочивается и правило выбора из нее отдельных единиц устанавливает исследователь.

Выборочный метод наблюдения широко используется на практике как в области естественных наук для оценки результатов экспериментов, так и в экономике. Госкомстат России проводит выборочные обследования бюджетов домашних хозяйств, потребительских ожиданий населения, обследования населения по проблемам занятости и др. На выборочной основе организовано статистическое наблюдение за деятельностью малых предприятий, за их деловой активностью, наличием и движением основных фондов. Выборочный метод используется также при изучении объема и состава затрат организаций на рабочую силу. Сфера применения этого метода постоянно расширяется, что связано с рядом его преимуществ.

Во-первых, выборочный метод обеспечивает значительную экономию материальных и финансовых ресурсов при проведении статистического наблюдения, что позволяет расширить программу обследования и повысить его оперативность. Второе преимущество – высокая достоверность получаемых данных, так как при относительно небольшом объеме выборки можно организовать эффективный контроль за качеством собираемой информации. Таким образом, при использовании выборочного метода снижается вероятность появления ошибок регистрации и необнаружения их на стадии проверки первичной информации. И наконец, в ряде случаев, когда сплошное наблюдение связано с уничтожением или порчей обследуемых единиц (например, при проверке качества поступающих в продажу продуктов питания), возможно только выборочное обследование.

Точность оценок, полученных на основе выборочного метода, зависит не от доли обследованных единиц, а от их числа. Если объем генеральной совокупности достаточно велик, то доля отобранных для наблюдения единиц может быть очень небольшой, а точность оценок – высокой. Например, выборочное обследование по проблемам занятости в России охватывает около 0,2% населения в возрасте от 15 до 72 лет, но обеспечивает высокую точность оценок параметров генеральной совокупности. Если же объем такой совокупности невелик, то эффект от применения

выборочного наблюдения может выражаться не столько в экономии материальных ресурсов, сколько в повышении качества собираемой исходной информации. Для получения несмещенной оценки в этом случае процент отбора должен быть значительно больше. Под *несмещенной оценкой* подразумевается такая характеристика выборочной совокупности, математическое ожидание которой совпадает с ее значением в генеральной совокупности.

Методологически выборочное наблюдение сложнее, так как требует глубокой предварительной проработки программы, а в ряде случаев и организации пробного обследования. Если такое наблюдение проводится на постоянной основе, необходимо периодически обновлять совокупность обследуемых единиц, т.е. требуется ротация выборки.

Распространяя результаты выборочного обследования на генеральную совокупность, следует иметь в виду, что между характеристиками генеральной и выборочной совокупности возможно расхождение, обусловленное тем, что обследуется не вся совокупность, а лишь ее часть. Такого рода несовпадения называются ***ошибками репрезентативности***, которые подразделяются на систематические и случайные. *Систематические ошибки* возникают в связи с принятым способом отбора или нарушением его правил. Например, результаты проводимых в России обследований бюджетов домашних хозяйств содержат значительную систематическую ошибку, так как в выборочной совокупности фактически не представлены наиболее богатые слои населения.

Случайные ошибки репрезентативности неизбежно возникают при проведении выборочных обследований, так как обеспечить абсолютную адекватность характеристик выборочной и генеральной совокупности даже при тщательно спланированном наблюдении практически невозможно. Оценка таких ошибок – одна из задач статистики. Важно определить не только абсолютную величину ошибки, но и ее допустимый уровень. Стремление максимально уменьшить случайную ошибку выборки приводит к росту ее объема, а большая ошибка ставит под сомнение возможность практического использования полученных результатов. Допустимый уровень ошибки должен быть установлен при разработке программы обследования.

Основные этапы выборочного наблюдения:

1) определение цели, задач и составление программы наблюдения;
2) анализ информационных источников, используемых для выделения генеральной совокупности объектов наблюдения (основы выборки);

3) формирование генеральной совокупности для проведения выборочного обследования;
4) разработка методологии формирования выборочной совокупности, включающей выбор способа отбора, определение необходимого объема выборки, этапов отбора единиц из генеральной совокупности, планирование и проведение пробной выборки;
5) формирование выборки;
6) сбор данных на основе разработанной программы;
7) анализ полученных результатов и расчет основных характеристик выборочной совокупности;
8) расчет ошибок выборки и распространение ее результатов на генеральную совокупность.

Условные обозначения, использованные при изложении материала в данной главе, приведены в табл. 6.1.

Таблица 6.1

Условные обозначения

Показатель	Совокупность	
	генеральная	выборочная
Объем (число единиц) совокупности	N	n
Среднее значение признака	$\bar{x}$	$\tilde{x}$
Доля единиц, обладающих изучаемым признаком	p	w
Доля единиц, не обладающих изучаемым признаком	q	$1 - w$
Дисперсия	$\sigma_г^2$	σ^2
Среднее квадратическое отклонение	$\sigma_г$	σ

6.2. Ошибки выборки при собственно случайном отборе

Виды случайного отбора

Теоретические основы выборочного метода, первоначально разработанные применительно к собственно случайному отбору, используют и для определения ошибок выборки при других способах наблюдения.

Рассмотрим наиболее простой способ формирования выборочной совокупности – ***собственно случайный отбор.***

Собственно случайный отбор может быть повторным и бесповторным. При *повторном* отборе каждая единица, отобранная в случайном порядке из генеральной совокупности, после проведе-

ния наблюдения возвращается в эту совокупность и может быть вновь подвергнута обследованию. На практике такой способ отбора встречается редко. Гораздо более распространен собственно случайный *бесповторный* отбор, при котором обследованные единицы в генеральную совокупность не возвращаются и не могут быть обследованы повторно. При повторном отборе вероятность попадания в выборку для каждой единицы генеральной совокупности остается неизменной. При бесповторном отборе она меняется, но для всех единиц, оставшихся в генеральной совокупности после отбора из нее нескольких единиц, вероятность попадания в выборку одинакова.

Для обеспечения случайности отбора используются разные способы. Если параметры генеральной совокупности известны и все ее единицы могут быть пронумерованы, то случайный отбор обеспечивается с помощью жребия. При большом объеме совокупности выборка может осуществляться с использованием таблиц случайных чисел. Такие таблицы представляют собой набор четырех- или пятизначных чисел. Если число единиц в генеральной совокупности трехзначное, то из любого столбца или строки таблицы последовательно выписывают столько чисел, сколько единиц в выборочной совокупности. От каждого числа отбрасывают первую или последнюю цифру (или две цифры, если таблицы состоят из пятизначных чисел). Затем отбирают числа, не превышающие число единиц в генеральной совокупности.

Пример. В первом столбце таблицы случайных чисел содержатся числа: 5489, 3522, 7555, 5759, 6303 и т.д. Предположим, что генеральная совокупность состоит из 600 единиц. При этом в соответствии с программой выборки должно быть обследовано 30 единиц. Номера единиц, попавших в выборку: 489, 522, 555, 303 и т.д. Единицы с номером 759 в генеральной совокупности нет, поэтому в выписанные порядковые номера единиц наблюдения это число не попадает.

Ошибки выборки при случайном повторном отборе

Ошибка выборки для средней. Основные свойства выборочной совокупности, сформированной методом собственно случайного повторного отбора, рассмотрим на следующем примере.

Пример. Из генеральной совокупности (например, студенты I курса, данные о возрасте которых приведены в табл. 6.2) с числом единиц $N = 4$ методом собственно случайного повторного отбора осуществлена выборка, объем которой равен 2 единицам, т.е. $n = 2$.

Таблица 6.2

Порядковый номер студента	1	2	3	4
Возраст x_i, лет	16	17	17	18

Результаты всех возможных испытаний представлены в табл. 6.3.

Таблица 6.3

Номера отобранных единиц	Выборочная средняя $\bar{x}_i$	Номера отобранных единиц	Выборочная средняя $\bar{x}_i$
1 и 1	16,0	3 и 1	16,5
1 и 2	16,5	3 и 2	17,0
1 и 3	16,5	3 и 3	17,0
1 и 4	17,0	3 и 4	17,5
2 и 1	16,5	4 и 1	17,0
2 и 2	17,0	4 и 2	17,5
2 и 3	17,0	4 и 3	17,5
2 и 4	17,5	4 и 4	18,0

В генеральной совокупности средний возраст студентов

$$\bar{x} = \frac{\sum x_i}{N} = \frac{16 + 17 + 17 + 18}{4} = 17 \text{ лет},$$

дисперсия изучаемого признака

$$\sigma_г^2 = \frac{\sum (x_i - \bar{x})^2}{N} =$$

$$= \frac{(16 - 17)^2 + (17 - 17)^2 + (17 - 17)^2 + (18 - 17)^2}{4} = 0{,}5.$$

На основе результатов расчета $\bar{x}$ и $\sigma_г^2$ можно построить распределение полученных значений выборочных средних (табл. 6.4).

Таблица 6.4

i	Средний возраст студентов в выборке, лет $\bar{x}_i$	Отклонение выборочной средней от генеральной средней $\bar{x}_i - \bar{x}$	Частота появления i-го значения выборочной средней f_i	Вероятность появления i-го значения выборочной средней p_i
1	2	3	4	5
1	16,0	−1,0	1	0,0625
2	16,5	−0,5	4	0,2500
3	17,0	0,0	6	0,3750
4	17,5	0,5	4	0,2500
5	18,0	1,0	1	0,0625
Итого			16	1,0000

Вероятности появления различных значений выборочной средней, равные вероятностям соответствующего отклонения выборочной средней от генеральной средней, неодинаковы. Чем больше отклонение выборочной характеристики от генеральной, тем меньше вероятность его появления. Наиболее часто оценка, полученная на основе выборки, совпадает с соответствующей характеристикой генеральной совокупности. В приведенном примере вероятность появления в выборке среднего возраста студентов, равного 17 годам, наиболее велика ($p_3 = 0{,}3750$).

Рассчитаем математическое ожидание выборочной средней:

$$M(\tilde{x}) = \sum \tilde{x}_i p_i = 16{,}0 \cdot 0{,}0625 + 16{,}5 \cdot 0{,}25 + 17{,}0 \cdot 0{,}375 + \\ + 17{,}5 \cdot 0{,}25 + 18{,}0 \cdot 0{,}0625 = 17 \text{ лет.}$$

Таким образом, $\bar{x} = M(\tilde{x})$, т.е. выборочная средняя является несмещенной оценкой генеральной средней. Аналогичный результат можно получить, используя вместо вероятности p_i частоту появления соответствующих значений выборочных средних:

$$\tilde{x} = \frac{\sum \tilde{x}_i f_i}{\sum f_i} = \frac{16{,}0 \cdot 1 + 16{,}5 \cdot 4 + 17{,}0 \cdot 6 + 17{,}5 \cdot 4 + 18{,}0 \cdot 1}{16} = 17 \text{ лет.}$$

Отклонение выборочной средней от генеральной равно нулю лишь в 6 выборках из 16. В остальных случаях значения выборочной и генеральной средней не совпадают, при этом вероятность появления наибольшего по абсолютной величине отклонения, равного единице, минимальна. Таким образом, существует предел, к которому стремится отклонение выборочной средней от генеральной.

Рассчитаем среднюю величину этих отклонений. Учитывая, что сумма отклонений, взятая в абсолютном выражении, равна нулю, указанную среднюю рассчитаем как среднее квадратическое отклонение:

$$\mu = \sqrt{\frac{\sum \left(\tilde{x}_i - \bar{x}\right)^2 f_i}{\sum f_i}}.$$

Так как $\frac{f_i}{\sum f_i} = p_i$, то

$$\mu = \sqrt{\sum \left(\tilde{x}_i - \bar{x}\right)^2 p_i} = \sqrt{M\left(\tilde{x}_i - \bar{x}\right)^2} =$$

$$= \sqrt{(-1)^2 \cdot 0{,}0625 + (-0{,}5)^2 \cdot 0{,}25 + 0{,}5^2 \cdot 0{,}25 + 1^2 \cdot 0{,}25} = 0{,}5.$$

Полученная величина μ называется средней ошибкой выборки. ***Средняя ошибка выборки*** – это среднее квадратическое отклонение всех возможных значений выборочной средней от генеральной средней, т.е. от своего математического ожидания.

Дисперсия возможных значений выборочной средней

$$\mu^2 = \frac{\sum(\tilde{x}_i - \bar{x})^2 f_i}{\sum f_i} = M(\tilde{x}_i - \bar{x})^2 = 0{,}25.$$

В математической статистике доказано, что эта величина в n раз меньше дисперсии в генеральной совокупности. В данном примере дисперсия в генеральной совокупности $\sigma_г^2 = 0{,}5$, а объем выборки $n = 2$, тогда

$$\mu^2 = \frac{\sigma_г^2}{n} = \frac{0{,}5}{2} = 0{,}25.$$

Следовательно, средняя ошибка выборки может быть определена по формуле

$$\mu = \sqrt{\frac{\sigma_г^2}{n}} = 0{,}5.$$

При собственно случайном повторном отборе средняя ошибка выборки зависит от:

- вариации изучаемого признака в генеральной совокупности;
- объема выборки.

Чем больше вариация признака, тем больше ошибка выборки. Для ее уменьшения необходимо увеличить объем выборочной совокупности.

В действительности решается обратная задача: на основе выборочных данных делается вывод о некоторых характеристиках генеральной совокупности. Согласно правилу сложения дисперсий дисперсия в генеральной совокупности $M(x - \bar{x})^2$ может быть представлена как сумма двух слагаемых: средней величины из отклонений отдельных значений от выборочных средних $M(x - \tilde{x})^2$ и средней величины из отклонений выборочных средних от генеральной средней $M(\tilde{x} - \bar{x})^2$, т.е.

$$M(x - \bar{x})^2 = M(x - \tilde{x})^2 + M(\tilde{x} - \bar{x})^2.$$

Учитывая, что $M(x - \bar{x})^2 = \sigma_г^2$, $M(x - \tilde{x})^2 = \sigma^2$, а $M(\tilde{x} - \bar{x})^2 = \frac{\sigma_г^2}{n}$, получаем

$$\sigma_г^2 = \overline{\sigma^2} + \frac{\sigma_г^2}{n},$$

или

$$\overline{\sigma^2} = \sigma_г^2 - \frac{\sigma_г^2}{n} = \sigma_г^2 \frac{n-1}{n},$$

где $\overline{\sigma^2}$ – средняя дисперсия выборочных совокупностей.

Следовательно,

$$\sigma_г^2 = \frac{n\overline{\sigma^2}}{n-1}.$$

В таком случае средняя ошибка выборки

$$\mu = \sqrt{\frac{\overline{\sigma^2}}{n-1}}. \qquad (6.1)$$

Так как все возможные значения дисперсии в выборочной совокупности неизвестны, при нахождении средней ошибки выборки вместо $\overline{\sigma^2}$ в формуле (6.1) используют дисперсию конкретной выборки σ^2. При такой замене велика вероятность малой погрешности. При достаточно большом объеме выборочной совокупности в формуле (6.1) вместо $(n-1)$ можно использовать величину n. Таким образом, средняя ошибка выборки при собственно случайном повторном отборе будет рассчитываться по формуле

$$\mu = \sqrt{\frac{\sigma^2}{n}}. \qquad (6.2)$$

Учитывая, что на основе выборочного обследования нельзя точно оценить изучаемый параметр генеральной совокупности, необходимо найти пределы, в которых он находится. В конкретной выборке разность $|\tilde{x}_i - \bar{x}|$ может быть больше, меньше или равна μ. Каждое из отклонений $|\tilde{x}_i - \bar{x}|$ от μ имеет определенную вероятность. При выборочном обследовании реальное значение $\bar{x}$

в генеральной совокупности неизвестно. Зная среднюю ошибку выборки, с определенной вероятностью можно оценить отклонение выборочной средней от генеральной и установить пределы, в которых находится изучаемый параметр (в данном случае средняя) в генеральной совокупности. Отклонение выборочной характеристики от генеральной называется ***предельной ошибкой выборки*** Δ. Она определяется в долях средней ошибки с заданной вероятностью, т.е.

$$\Delta = t\mu, \tag{6.3}$$

где t — коэффициент доверия, зависящий от вероятности, с которой определяется предельная ошибка выборки.

Вероятность появления определенной ошибки выборки находят с помощью теорем теории вероятностей. Согласно теореме П.Л. Чебышёва, *при достаточно большом объеме выборки и ограниченной дисперсии генеральной совокупности вероятность того, что разность между выборочной средней и генеральной средней будет сколь угодно мала, близка к единице:*

$$P\left(\left|\tilde{x} - \bar{x}\right| \leqslant \xi\right) \to 1 \text{ при } n \to \infty.$$

А.М. Ляпунов доказал, что *независимо от характера распределения генеральной совокупности при увеличении объема выборки распределение вероятностей появления того или иного значения выборочной средней приближается к нормальному распределению.* (Это так называемая центральная предельная теорема.) Следовательно, вероятность отклонения выборочной средней от генеральной средней, т.е. вероятность появления заданной предельной ошибки, также подчиняется указанному закону и может быть найдена как функция от t с помощью интеграла вероятностей Лапласа:

$$P\left(\left|\tilde{x} - \bar{x}\right| \leqslant t\mu\right) = \frac{1}{\sqrt{2\pi}} \int_{-t}^{+t} e^{-\frac{t^2}{2}} dt,$$

где $t = \dfrac{\tilde{x} - \bar{x}}{\mu}$ — нормированное отклонение выборочной средней от генеральной средней.

Значения интеграла Лапласа для разных t рассчитаны и приводятся в специальных таблицах (см. Приложение 2).

Поясним графически процедуру нахождения вероятности t-кратного отклонения генеральной средней $\bar{x}$ от выборочной $\tilde{x}$ (рис. 6.1).

Вероятность t-кратного отклонения

$$\varphi(t) = P(t) = \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}.$$

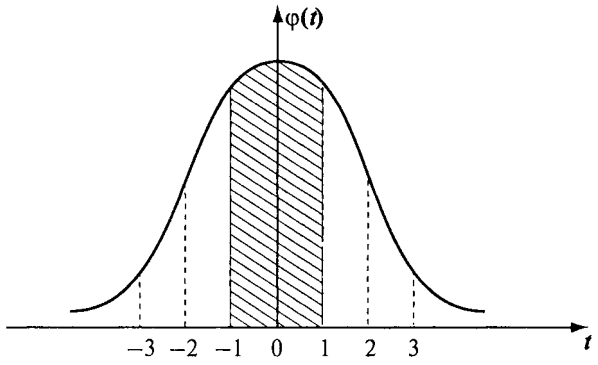


Рис. 6.1. Кривая нормального распределения

Площадь, ограниченная кривой нормального распределения и осью абсцисс, равна суммарной вероятности возникновения различных отклонений $\bar{x}$ от $\tilde{x}$, т.е. равна 1. Заштрихованная часть (см. рис. 6.1), которая находится в пределах от -1 до $+1$, равна 0,683, т.е. с вероятностью 68,3% можно гарантировать, что отклонение генеральной средней от выборочной не превысит однократной средней ошибки выборки. С этой вероятностью можно утверждать, что среднее значение признака в генеральной совокупности находится в пределах $\tilde{x} - \mu \leqslant \bar{x} \leqslant \tilde{x} + \mu$.

Вероятность того, что отклонение средней в генеральной совокупности от выборочной средней не выйдет за пределы 2μ (т.е. $t = 2$), равна 0,954, а вероятность того, что оно не выйдет за пределы 3μ, – соответственно 0,997. Таким образом, зная среднее значение признака в выборке, можно почти достоверно утверждать, что в генеральной совокупности соответствующее значение будет находиться в пределах $\tilde{x} - 3\mu \leqslant \bar{x} \leqslant \tilde{x} + 3\mu$. На практике доверительная вероятность принимается чаще всего на уровне 0,95 или 0,99. Соответствующие значения коэффициента доверия равны 1,96 и 2,58 (см. Приложение 2).

Пользуясь приведенными рассуждениями, можно определить вероятность только верхнего или нижнего предела для искомой характеристики генеральной совокупности. Например, вероятность того, что средняя в генеральной совокупности не превысит $\tilde{x} + 2\mu$, будет равна $0{,}5 + \dfrac{0{,}954}{2} = 0{,}977$.

Ошибка выборки для доли. Для того чтобы на основе результатов выборочного наблюдения найти долю единиц, обладающих

изучаемым признаком в генеральной совокупности, используют формулы, аналогичные приведенным ранее. Дисперсия для доли в генеральной совокупности $\sigma_г^2$ равна произведению pq, где p – доля единиц, обладающих изучаемым признаком в генеральной совокупности, а $q = 1 - p$ – доля единиц, не обладающих изучаемым признаком. Так как наблюдение выборочное и величины p и q неизвестны, в формуле средней ошибки выборки используются соответствующие значения, полученные на основе выборочного обследования. Средняя ошибка выборки для доли при собственно случайном повторном отборе рассчитывается по формуле

$$\mu = \sqrt{\frac{w(1 - w)}{n}}, \qquad (6.4)$$

где w – доля единиц, обладающих изучаемым признаком в выборочной совокупности;

$(1 - w)$ – доля единиц в выборке, не обладающих изучаемым признаком.

Предельная ошибка выборки в этом случае определяется так же, как и для средней: $\Delta = t\mu$.

Частный случай теоремы П.Л. Чебышёва для доли доказан Я. Бернулли: *при достаточно большом объеме выборки вероятность того, что расхождение между долями единиц, обладающих изучаемым признаком, в выборочной и генеральной совокупности будет сколь угодно малым, стремится к единице.* При этом распределение вероятностей различных отклонений доли в выборочной совокупности от доли в генеральной также подчиняется нормальному закону. Зная долю в выборочной совокупности, с соответствующей вероятностью можно гарантировать, что доля в генеральной совокупности не выйдет за пределы $w - t\mu \leqslant p \leqslant w + t\mu$.

Ошибки выборки при случайном бесповторном отборе

Приведенные ранее формулы средней ошибки выборки справедливы только при повторном отборе. Однако на практике чаще используется бесповторный отбор: обследованная единица не возвращается в генеральную совокупность и не может быть обследована повторно. При этом принцип независимости испытаний нарушается. Очевидно, что при бесповторном отборе из четырех студентов двоих (см. табл. 6.2) средний возраст студентов I курса не может быть равен 16 и 18 годам. Следовательно, средняя ошибка выборки будет меньше.

Средняя ошибка выборки при собственно случайном бесповторном отборе рассчитывается по формуле

$$\mu = \sqrt{\frac{\sigma^2}{n}\frac{N-n}{N-1}}. \qquad (6.5)$$

При больших значениях N величину $(N-1)$ в формуле (6.5) можно заменить на N, тогда упрощенная формула средней ошибки выборки запишется следующим образом:

для средней

$$\mu = \sqrt{\frac{\sigma^2}{n}\left(1-\frac{n}{N}\right)}, \qquad (6.6)$$

для доли

$$\mu = \sqrt{\frac{w(1-w)}{n}\left(1-\frac{n}{N}\right)}, \qquad (6.7)$$

где $\frac{n}{N}$ – доля обследованных единиц совокупности.

При собственно случайном бесповторном отборе средняя ошибка выборки зависит от:

- вариации изучаемого признака;
- объема выборки;
- доли обследованных единиц.

Чем больше объем выборки и доля обследованных единиц, тем меньше ошибка выборки; вариация признака связана с ней прямо пропорционально.

Если доля обследованных единиц невелика, то дополнительный множитель под знаком радикала практически не влияет на ошибку выборки. В этом случае ошибку выборки при бесповторном отборе можно найти по формулам, которые применяются при повторном отборе.

Наряду с абсолютной величиной средней и предельной ошибок выборки в статистической практике используется относительная их величина, рассчитываемая как отношение ошибки к исследуемому параметру: $\Delta_{\text{отн}} = \frac{\Delta}{\bar{x}}$ или $\Delta_{\text{отн}} = \frac{\Delta}{p}$. Теоретически в знаменателе должно быть значение исследуемого параметра генеральной совокупности. Однако оно неизвестно, поэтому относи-

тельная ошибка рассчитывается через соответствующий параметр выборки: $\Delta_{\text{отн}} = \frac{\Delta}{\bar{x}}$ или $\Delta_{\text{отн}} = \frac{\Delta}{w}$. Относительная ошибка выражается в процентах. Выборка считается репрезентативной, если $\Delta_{\text{отн}} \leqslant 5\%$.

Пример. По данным выборочного обследования, проведенного Госкомстатом России по состоянию на конец марта 1996 г., средний возраст безработных в России составил $\bar{x} = 34{,}4$ года при среднем квадратическом отклонении $\sigma = 13{,}8$ года. С вероятностью $P = 0{,}997$ определить пределы, в которых находится средний возраст безработных в генеральной совокупности, если известно, что в ходе обследования опрошено $n = 155$ тыс. человек в возрасте 15–72 лет, что составляет $\frac{n}{N} = 0{,}15\%$ от общей численности населения в этом возрасте.

Средняя ошибка выборки при собственно случайном бесповторном отборе составит

$$\mu = \sqrt{\frac{\sigma^2}{n}\left(1 - \frac{n}{N}\right)} = \sigma\sqrt{\frac{1}{n}\left(1 - \frac{n}{N}\right)} =$$

$$= 13{,}8\sqrt{\frac{1}{155000}(1 - 0{,}0015)} = 0{,}035 \text{ года}.$$

При $P\left(|\bar{x} - \bar{x}| \leqslant t\mu\right) = 0{,}997$ коэффициент доверия $t = 3$ (см. Приложение 2), а предельная ошибка выборки

$$\Delta = t\mu = 3 \cdot 0{,}035 \cong 0{,}1 \text{ года}.$$

Таким образом, средний возраст безработных в России с вероятностью 0,997 находится в пределах

$$\bar{x} - \Delta \leqslant \bar{x} \leqslant \bar{x} + \Delta,$$

т.е.

$$34{,}4 - 0{,}1 \leqslant \bar{x} \leqslant 34{,}4 + 0{,}1,$$

или

$$34{,}3 \text{ года} \leqslant \bar{x} \leqslant 34{,}5 \text{ года}.$$

При решении данной задачи среднюю ошибку выборки можно было рассчитать по формуле для повторного отбора, поскольку величина $\frac{n}{N}$ мала.

Пример. С вероятностью 0,954 определить предельную ошибку выборки для доли мужчин среди безработных в России в конце марта 1996 г., если известно, что в выборке (n = 155 тыс. человек) их доля составила 54,9%.

При $P\left(\left|w - p\right| \leqslant t\mu\right) = 0{,}954$ коэффициент доверия $t = 2$ (см. Приложение 2).

Предельная ошибка выборки

$$\Delta = t\mu = t\sqrt{\frac{w(1 - w)}{n}} = 2\sqrt{\frac{0{,}549(1 - 0{,}549)}{155000}} = 0{,}0025,$$

т.е. $\Delta = 0{,}25\%$.

Следовательно, с вероятностью 0,954 можно утверждать, что в генеральной совокупности предельная ошибка выборки для доли безработных мужчин не превысит 0,25%.

Можно решить и обратную задачу: задав предельную ошибку выборки, определить вероятность, с которой она может быть гарантирована. При этом, зная Δ и μ, сначала находят коэффициент доверия $t = \Delta/\mu$, а затем по таблице (см. Приложение 2) искомое значение вероятности.

6.3. Основные способы формирования выборочной совокупности

Рассмотренный в параграфе 6.2 собственно случайный способ формирования выборочной совокупности теоретически наиболее простой, но он предполагает, что в распоряжении исследователя имеется полный перечень единиц. На практике более широко используется ***механический отбор***, основанный на предварительном упорядочении генеральной совокупности. Например, при проведении выборочных социально-демографических обследований составляются списки жилых помещений, при обследовании предприятий – их регистры и т.д. Устанавливается процент отбора, исходя из которого определяется число отбираемых единиц. Например, при формировании 5-процентной выборки из 1 млн единиц необходимо обследовать 50 тыс., т.е. из каждых 20 единиц одну. Затем определяется начало отбора, т.е. номер первой обследуемой единицы. Каждая следующая единица включается в выборку в соответствии с установленным шагом отбора. В приведенном примере из списка следует отбирать каждую 20-ю единицу.

Механический способ отбора часто используется на практике, так как он позволяет проводить оперативную замену одной единицы наблюдения на другую, стоящую в списке непосредственно перед или за единицей, первоначально включенной в выборку. Необходимость замены единиц наблюдения довольно часто возникает в связи с отказом респондентов от участия в обследовании или отсутствием их по соответствующему адресу. Например, при проведении выборочных бюджетных обследований населения в России ротация выборки, обусловленная указанной причиной, ежегодно составляет 15–20%.

Начало отсчета определяется разными способами. Если предварительное упорядочение единиц генеральной совокупности по какому-либо признаку не проводится, а список единиц (или регистр) составляется в порядке их поступления (например, регистр предприятий), то начало отсчета устанавливается в случайном порядке, а каждая следующая единица отбирается из списка через установленный интервал отсчета. Например, если началом отсчета при 5-процентном отборе является 20-я единица, то следующая – 40-я, затем 60-я и т.д.

При механическом отборе из совокупности, упорядоченной по какому-либо признаку, очень важно правильно выбрать начало отсчета. Например, при изучении бюджетов домашних хозяйств списки можно составлять в зависимости от заработной платы работников, поэтому выбор в качестве начала отсчета первой или последней единицы из каждой группы приведет к появлению систематической ошибки выборки. Чтобы этого не происходило, необходимо из каждой группы отобрать единицу, которая находится в ее середине. При проведении выборочных социально-демографических обследований необходимо включить в выборку семьи разного состава. Поэтому из списка жилых помещений недостаточно механически отобрать каждое следующее в соответствии с установленным номером, требуется менять начало отсчета. Это позволяет исключить в выборке возможный перекос в сторону семей того или иного состава.

Если предварительное упорядочение проводится, то теоретически из общей ошибки выборки необходимо выделить ее случайную и систематическую компоненты. Однако практически это невозможно, поскольку в каждой группе обследуется одна единица. Поэтому при механическом отборе используются те же формулы для нахождения ошибки выборки, что и при собственно случайном отборе (табл. 6.5).

Таблица 6.5

Формулы средней ошибки выборки при собственно случайном и механическом отборе

Оцениваемый параметр	Повторный отбор	Бесповторный отбор
Средняя	$\mu = \sqrt{\frac{\sigma^2}{n}}$	$\mu = \sqrt{\frac{\sigma^2}{n}\left(1 - \frac{n}{N}\right)}$
Доля	$\mu = \sqrt{\frac{w(1-w)}{n}}$	$\mu = \sqrt{\frac{w(1-w)}{n}\left(1 - \frac{n}{N}\right)}$

В последние годы более широкое практическое применение получил ***типический (стратифицированный, расслоенный) отбор***, при котором обследуемая совокупность предварительно разбивается на типически однородные группы и выбор осуществляется из каждой такой группы механическим или собственно случайным способом.

Типический способ отбора используется в нашей стране при проведении выборочных бюджетных обследований домашних хозяйств, изучении потребительских ожиданий населения, при организации выборочных обследований по проблемам занятости, анализе результатов деятельности малых предприятий и их деловой активности.

До 1996 г. при формировании выборочной совокупности домашних хозяйств для проведения бюджетных обследований применялась типическая выборка с механическим отбором единиц внутри групп. На первом этапе отбора в качестве типических групп использовалась территориально-отраслевая группировка рабочих и служащих, а на втором – группировка по средней месячной оплате труда. В настоящее время при проведении таких обследований применяется более сложная процедура отбора, также основанная на выделении типически однородных групп.

При проведении ежеквартальных выборочных обследований малых предприятий используется многомерная типическая выборка, объем которой составляет примерно 20% от общей численности таких предприятий. Расслоение объектов генеральной совокупности по типически однородным группам проводится в соответствии со следующими признаками: территория, отрасль, форма собственности, выручка от реализации продукции (работ, услуг).

Из каждой выделенной группы при проведении типического отбора в выборочную совокупность отбирается определенное число единиц. Обозначим число единиц, попавших в выборку из i-й группы, через n_i, а общее число образованных групп через m ($i = 1, 2, ..., m$). Величину n_i можно задать одним из трех способов:

- отбор из каждой группы равного числа единиц, т.е. $n_i = \frac{n}{m}$.

  Использование такого принципа отбора позволяет получить достаточно надежные результаты лишь при равных размерах выделенных типических групп. Если же их численность существенно различается между собой, то использование равномерного отбора может привести к смещению оценок, полученных по результатам выборочного обследования;
- отбор единиц пропорционально их численности в соответствующих группах генеральной совокупности, т.е. $n_i = n\frac{N_i}{N}$, где N_i – число единиц в i-й типической группе генеральной совокупности. Использование этого принципа формирования выборочной совокупности обеспечивает достаточно надежные результаты, если колеблемость признака несущественно различается в разных группах генеральной совокупности. Если же коэффициенты вариации в них различаются существенно, то репрезентативность выборки при таком способе ее формирования может оказаться невысокой;
- оптимальное размещение, учитывающее не только численность групп, но и степень вариации в них изучаемого признака, т.е. $n_i = n\frac{N_i\sigma_{ir}}{\sum N_i\sigma_{ir}}$, где σ_{ir} – среднее квадратическое отклонение признака в i-й группе генеральной совокупности. Данная формула получена следующим образом:

$$n_i = n\frac{N_i}{\sum N_i}\frac{\sigma_{ir}}{\overline{\sigma_{ir}}} = n\frac{N_i\sigma_{ir}}{\sum N_i\sigma_{ir}}, \qquad (6.8)$$

где $\sum N_i = N$, а $\overline{\sigma_{ir}} = \frac{\sum\sigma_{ir}N_i}{\sum N_i}$ – средняя из групповых средних квадратических отклонений.

Оптимальное размещение позволяет минимизировать среднюю ошибку выборки. Впервые ответ на вопрос о наиболее эффектив-

ной организации типической (расслоенной) выборки был получен в 1920 г. А.А. Чупровым и независимо от него в 1934 г. Е. Нейманом. В статистике такое размещение называется также *неймановым*. Хотя оно позволяет получить более точные результаты, на практике осуществить его сложно, поскольку необходимо знать вариацию признака в генеральной совокупности еще до проведения обследования.

В статистической практике нашей страны пропорциональный способ отбора используется при формировании выборочной совокупности для проведения наблюдения за деловой активностью малых предприятий в промышленности.

Для ежеквартального наблюдения за основными экономическими показателями, характеризующими деятельность малых предприятий, выборка строится с использованием принципа оптимального размещения.

Опыт практического применения принципа оптимального размещения показал, что использование в качестве показателя вариации в формуле (6.8) среднего квадратического отклонения по i-й типической группе генеральной совокупности ($\sigma_{i\Gamma}$) позволяет обеспечить оптимальное размещение единиц выборочной совокупности по типическим группам лишь в том случае, если нет существенных различий в значениях коэффициентов вариации по этим группам. При проведении выборочного обследования всех предприятий региона, а не только малых, высокое значение среднего квадратического отклонения может иметь место при низком значении коэффициента вариации (например, на крупных предприятиях) и наоборот. В результате крайне неоднородные группы объектов, например малые предприятия, будут недостаточно представлены в выборке. В таком случае в формуле (6.8) целесообразно использовать не среднее квадратическое отклонение, а коэффициент вариации.

Общая дисперсия изучаемого признака, согласно правилу сложения дисперсий, может быть представлена как сумма $\sigma^2 = \overline{\sigma_i^2} + \delta^2$, где $\overline{\sigma_i^2}$ – средняя из групповых дисперсий, а δ^2 – межгрупповая дисперсия. При проведении типического отбора межгрупповая дисперсия не носит характера случайной вариации, так как группы образованы еще до начала выборочного обследования. Следовательно, при нахождении ошибки выборки необходимо из двух указанных слагаемых общей дисперсии учесть лишь то, которое связано со случайной вариацией, а именно среднюю из групповых дисперсий в выборке $\overline{\sigma_i^2}$.

При отборе пропорционально численности групп средняя из групповых дисперсий $\overline{\sigma_i^2} = \frac{\sum \sigma_i^2 N_i}{\sum N_i} = \frac{\sum \sigma_i^2 n_i}{\sum n_i}$. В этом случае средняя ошибка выборки при бесповторном отборе:

для средней

$$\mu = \sqrt{\frac{\overline{\sigma_i^2}}{n}\left(1 - \frac{n}{N}\right)},$$

для доли

$$\mu = \sqrt{\frac{\overline{w_i(1 - w_i)}}{n}\left(1 - \frac{n}{N}\right)},$$

где w_i — доля единиц совокупности, обладающих изучаемым признаком в i-й типической группе;

$\overline{w_i(1 - w_i)}$ — средняя из групповых дисперсий для доли.

В табл. 6.6 представлены формулы для исчисления средней ошибки выборки при типическом отборе.

Таблица 6.6

Формулы средней ошибки выборки при типическом отборе

Оцениваемый параметр	Повторный отбор	Бесповторный отбор
Пропорциональное распределение единиц выборочной совокупности по группам		
Средняя	$\mu = \sqrt{\frac{\overline{\sigma_i^2}}{n}}$	$\mu = \sqrt{\frac{\overline{\sigma_i^2}}{n}\left(1 - \frac{n}{N}\right)}$
Доля	$\mu = \sqrt{\frac{\overline{w_i(1 - w_i)}}{n}}$	$\mu = \sqrt{\frac{\overline{w_i(1 - w_i)}}{n}\left(1 - \frac{n}{N}\right)}$
Оптимальное распределение единиц выборочной совокупности по группам		
Средняя	$\mu = \frac{1}{N}\sqrt{\sum \frac{\sigma_i^2 N_i^2}{n_i}}$	$\mu = \frac{1}{N}\sqrt{\sum\left[\frac{\sigma_i^2 N_i^2}{n_i}\left(1 - \frac{n_i}{N_i}\right)\right]}$
Доля	$\mu = \frac{1}{N}\sqrt{\sum \frac{w_i(1 - w_i)N_i^2}{n_i}}$	$\mu = \frac{1}{N}\sqrt{\sum\left[\frac{w_i(1 - w_i)N_i^2}{n_i}\left(1 - \frac{n_i}{N_i}\right)\right]}$

Разновидностью типической является ***районированная выборка***, при которой отбор единиц для наблюдения проводится из групп, представленных административно-территориальными образованиями. В этом случае преимущества типической выборки проявляются лишь при заметном расхождении среднего значения изучаемого признака по отдельным регионам.

Применение типического отбора позволяет уменьшить среднюю ошибку выборки, но его преимущества проявляются только при условии, что различия в средних значениях изучаемого признака между группами достаточно ощутимы, а вариация признака внутри каждой группы невелика.

Если при построении типической выборки предполагается получить значение нескольких показателей, то расслоение проводится, как правило, не по наблюдаемым, а по вспомогательному признаку. При этом вспомогательный признак должен коррелировать с наблюдаемыми. Например, численность домашних хозяйств определенного типа (одиночки; семьи, состоящие из двоих взрослых; семьи с разным числом детей) коррелирует с показателями доходов и расходов.

Пример. Для изучения объема и структуры доходов работников городских торговых предприятий, относящихся к разным формам собственности, проведен 2-процентный бесповторный типический отбор, результаты которого по одному из обследованных показателей приведены в табл. 6.7.

Таблица 6.7

Форма собственности	Численность занятых, чел.	Обследовано человек	Доход от участия в собственности предприятия на одного работника в год, тыс. руб.	
			средний	среднее квадратическое отклонение
	N_i	n_i	x_i	σ_i
Государственная	5000	100	270	90
Негосударственная	25000	500	880	260
Всего	30000	600		

В данном случае отбор проведен пропорционально численности работников, занятых на предприятиях каждой выделенной группы. Для того чтобы найти пределы, в которых указанный вид дохода работников торговли находится в генеральной совокупно-

сти, зададим доверительную вероятность $P = 0{,}95$. Следовательно, коэффициент доверия $t = 1{,}96$ (см. Приложение 2).

В выборочной совокупности средняя сумма дохода от участия в собственности предприятия

$$\tilde{x} = \frac{\sum x_i n_i}{\sum n_i} = \frac{270 \cdot 100 + 880 \cdot 500}{100 + 500} = 778{,}3 \text{ тыс. руб.}$$

Для нахождения средней ошибки выборки необходимо знать среднюю из групповых дисперсий:

$$\overline{\sigma_i^2} = \frac{\sum \sigma_i^2 n_i}{\sum n_i} = \frac{90^2 \cdot 100 + 260^2 \cdot 500}{100 + 500} = 57\,683.$$

Предельная ошибка выборки

$$\Delta = t\mu = t\sqrt{\frac{\overline{\sigma_i^2}}{n}\left(1 - \frac{n}{N}\right)} = 1{,}96\sqrt{\frac{57683}{600}(1 - 0{,}02)} \cong 19{,}0 \text{ тыс. руб.}$$

Следовательно, с вероятностью 0,95 можно сделать вывод о том, что среди торговых работников города средний годовой доход от участия в собственности предприятия находится в пределах

$$\tilde{x} - \Delta \leqslant \bar{x} \leqslant \tilde{x} + \Delta,$$

т.е.

$$778{,}3 - 19{,}0 \leqslant \bar{x} \leqslant 778{,}3 + 19{,}0,$$

или

$$759{,}3 \text{ тыс. руб.} \leqslant \bar{x} \leqslant 797{,}3 \text{ тыс. руб.}$$

Для того чтобы на основе приведенных данных сформировать выборочную совокупность с учетом не только численности выделенных групп в генеральной совокупности, но и степени вариации в них изучаемого признака, надо исходить из предположения, что групповые дисперсии в выборочной и генеральной совокупности равны, т.е. $\sigma_{i\text{г}}^2 \cong \sigma_i^2$. Общий объем выборки $n = 600$ чел. с учетом степени вариации признака по формам собственности должен быть распределен следующим образом:

$$n_1 = n\frac{N_1\sigma_1}{\sum_{i=1}^{2} N_i\sigma_i} = 600\frac{5000 \cdot 90}{5000 \cdot 90 + 25000 \cdot 260} = 39 \text{ чел.};$$

$$n_2 = n\frac{N_2\sigma_2}{\sum_{i=1}^{2} N_i\sigma_i} = 600\frac{25000 \cdot 260}{5000 \cdot 90 + 25000 \cdot 260} = 561 \text{ чел.}$$

Таким образом, при оптимальном размещении необходимо обследовать 39 работников государственных торговых предприятий и 561 – негосударственных. В этом случае средняя ошибка выборки

$$\mu = \frac{1}{N}\sqrt{\sum\left[\frac{\sigma_i^2 N_i^2}{n_i}\left(1 - \frac{n_i}{N_i}\right)\right]} =$$

$$= \frac{1}{30000}\sqrt{\frac{90^2 \cdot 5000^2}{39}\left(1 - \frac{39}{5000}\right) + \frac{260^2 \cdot 25000^2}{561}\left(1 - \frac{561}{25000}\right)} =$$

$$= 9{,}4 \text{ тыс. руб.}$$

При $P = 0{,}95$ предельная ошибка выборки равна 18,4 тыс. руб., т.е. немного меньше, чем при пропорциональном отборе.

Серийный (гнездовой) отбор применяется в том случае, если генеральная совокупность разбита на группы еще до начала выборочного обследования. При проведении выборки исследователь может из генеральной совокупности отбирать не отдельные единицы, а целые их серии и обследовать в рамках каждой серии все попавшие в нее единицы. Такой способ отбора широко применяется при контроле качества продукции, когда для проведения наблюдения вскрывается упаковка, содержащая определенное количество изделий, и все они проверяются. Если бы в этом случае из каждой упаковки обследовалась лишь одна единица, потребовалось бы повторно упаковывать всю партию товара, что привело бы к дополнительному увеличению затрат, связанных с обследованием.

Исследователь отбирает из генеральной совокупности в случайном порядке более крупные единицы (серии), поэтому возникновение случайной ошибки связано с отклонением серийных средних от общей средней. Поскольку внутри отобранных серий обследуются все единицы, то вариация признака в рамках каждой серии носит характер не случайной, а систематической составляющей и, следовательно, не должна учитываться при расчете средней ошибки выборки. Таким образом, в формуле средней ошибки выборки вместо общей дисперсии необходимо использовать межсерийную дисперсию.

Обозначим число серий в генеральной совокупности через S, а в выборке – через s.

Межсерийная дисперсия при равновеликих сериях $\delta_{\bar{x}}^2$ рассчитывается по формуле

$$\delta_{\bar{x}}^2 = \frac{\sum_{i=1}^{s}\left(\bar{x}_i - \bar{x}\right)^2}{s},$$

где $\bar{x}_i$ – среднее значение признака в i-й серии;

$\bar{x} = \frac{\sum \bar{x}_i}{s}$ – общая средняя в выборочной совокупности.

При бесповторном серийном отборе средняя ошибка выборки

$$\mu = \sqrt{\frac{\delta_{\bar{x}}^2}{s}\frac{S-s}{S-1}}.$$

Если число серий в генеральной совокупности велико, то вместо $(S-1)$ в последней формуле можно использовать величину S. В знаменателе первой дроби величина s берется лишь при большом объеме выборки ($s > 30$). Если число отобранных серий невелико, вместо s должна быть величина $(s-1)$.

При нахождении межсерийной дисперсии для доли необходимо учесть, что среднее значение альтернативного признака равно p, т.е. доле единиц, обладающих этим признаком. Соответственно, оно будет равно $\bar{w}$ в выборке и w_i в каждой отобранной серии. В таком случае межсерийная дисперсия для доли

$$\delta_w^2 = \frac{\sum_{i=1}^{s}\left(w_i - \bar{w}\right)^2}{s}.$$

При равновеликих сериях $\bar{w} = \frac{\sum_{i=1}^{s} w_i}{s}$.

При серийном отборе средняя ошибка выборки рассчитывается по формулам, приведенным в табл. 6.8.

При рассмотренных способах формирования выборочной совокупности отбор единиц для наблюдения осуществляется уже на первом этапе. Такой отбор называется ***одноступенчатым***. Однако на практике часто используется ***многоступенчатый отбор***, при котором на первом этапе из совокупности отбираются укрупненные единицы (серии), а затем без проведения наблюдения за всеми единицами в рамках серии осуществляется собственно случайный или механический отбор единиц из каждой отобранной серии.

Таблица 6.8

Формулы средней ошибки выборки при серийном отборе

Оцениваемый параметр	Повторный отбор*	Бесповторный отбор
Средняя	$\mu = \sqrt{\frac{\delta_{\bar{x}}^2}{s}}$	$\mu = \sqrt{\frac{\delta_{\bar{x}}^2}{s}\frac{S-s}{S-1}}$
Доля	$\mu = \sqrt{\frac{\delta_w^2}{s}}$	$\mu = \sqrt{\frac{\delta_w^2}{s}\frac{S-s}{S-1}}$

* При серийной выборке повторный отбор практически не применим, поэтому в основном используются формулы ошибок для бесповторного отбора.

Ошибка выборки при двухступенчатом отборе складывается из ошибок, возникающих на каждой ступени. В данном случае

$$\mu = \sqrt{\frac{\delta_{\bar{x}}^2}{s}\frac{S-s}{S-1} + \frac{\overline{\sigma_i^2}}{n}\left(1 - \frac{n}{N_s}\right)}, \qquad (6.9)$$

где $\overline{\sigma_i^2}$ – средняя из серийных дисперсий;

N_s – общее число единиц совокупности в отобранных сериях.

По схеме двухступенчатой выборки организовано обследование населения по проблемам занятости, а также выборочные бюджетные обследования. На основе ежеквартальных выборочных бюджетных обследований населения изучаются источники средств существования разных категорий домашних хозяйств, структура их расходов, объем потребления продуктов питания, наличие домашнего имущества, обеспеченность жильем. В настоящее время такие обследования ориентированы в основном на сбор информации о расходах и объеме личного потребления населения. На основе полученных таким образом данных рассчитываются весовые коэффициенты, необходимые для исчисления сводного индекса потребительских цен.

Выборка строится по территориальному принципу с учетом необходимости представительства в ней разных типов домашних хозяйств. Базой для таких обследований служат материалы переписи населения. На первой ступени единицей отбора в рамках каждой территории является счетный участок переписи населения. На второй ступени на каждом отобранном участке отбирает-

ся 25 (30) домохозяйств, имеющих разные социально-экономические характеристики. В качестве группировочных признаков используются такие показатели, как размер домашнего хозяйства, принадлежность и тип жилого помещения, источники средств существования, возраст, пол и уровень образования обследованных лиц. В настоящее время выборочными обследованиями охвачено 49,2 тыс. домашних хозяйств всех субъектов Российской Федерации.

Эта выборочная совокупность служит также основой для проведения ежеквартальных выборочных обследований потребительских ожиданий населения. Такие обследования направлены на изучение мнения населения об общей экономической ситуации в стране и личном материальном положении, а также о ситуации на рынке товаров и услуг. Объем выборки при проведении обследований потребительских ожиданий составляет 5 тыс. человек в возрасте 16 лет и старше.

Аналогичным образом формируется выборочная совокупность домашних хозяйств для проведения обследований по проблемам занятости населения. Такие обследования проводятся в России с 1992 г. в целях получения информации о численности и составе экономически активного населения, занятых и безработных. В настоящее время выборочные обследования по проблемам занятости проводятся ежеквартально. Единицей наблюдения являются лица в возрасте от 15 до 72 лет, входящие в состав отобранных домохозяйств. Ежеквартальный объем выборки составляет около 63,8 тыс. человек (31 тыс. домашних хозяйств), а годовой ее объем – 255 тыс. человек.

При построении многоступенчатой выборки используется комбинация разных способов отбора, поэтому такой способ отбора иногда называют ***комбинированной выборкой.***

От многоступенчатого следует отличать ***многофазный отбор***, при котором из единиц совокупности, отобранных на первом этапе, осуществляется подвыборка в целях изучения дополнительных характеристик обследуемой совокупности. При многофазном отборе единица отбора на каждом этапе одна и та же, а при многоступенчатом она меняется: на первой ступени отбираются единицы более высокого порядка (например, серии), чем на второй. Многофазная выборка используется для расширения программы обследования. На второй фазе целесообразно изучать такие признаки, которые обладают меньшей вариацией в генеральной совокупности. Это позволяет сэкономить средства, необходимые для выборочного обследования. Отметим, что на каждой фазе

многофазных выборок рассчитывается особое значение ошибки выборки.

Взаимопроникающие выборки – форма выборочного обследования, при которой из одной генеральной совокупности одним и тем же способом формируются две (или более) выборочные совокупности. При этом происходит взаимное уточнение результатов обследования.

Выборочные обследования широко применяются в отечественной статистической практике. Этот метод используется, например, при организации специальных статистических обследований, изучении общественного мнения и т.д.

6.4. Определение необходимой численности выборки

При разработке программы выборочного обследования одним из наиболее сложных является вопрос о том, сколько единиц изучаемой совокупности необходимо обследовать, т.е. об объеме выборки. В параграфах 6.2 и 6.3 показано, что при любом способе отбора предельная ошибка выборки обратно пропорциональна числу обследованных единиц. Чтобы уменьшить ошибку выборки, необходимо увеличить ее объем, но при этом возрастут и затраты на проведение обследования. Определяя необходимую численность выборочной совокупности, приходится прежде всего оценивать допустимую ошибку.

Как определить необходимую численность выборки при собственно случайном или механическом повторном отборе? В этом случае предельная ошибка выборки для средней

$$\Delta = t\mu = t\sqrt{\frac{\sigma^2}{n}} = \frac{t\sigma}{\sqrt{n}},$$

а необходимая ее численность

$$n = \frac{t^2\sigma^2}{\Delta^2}. \qquad (6.10)$$

Для определения необходимой численности выборки должны быть заданы предельная ее ошибка и вероятность того, что эта ошибка не превысит заданного предела. В соответствии с этой вероятностью по таблице Приложения 2 находят коэффициент доверия t.

Наиболее сложно определить дисперсию изучаемого признака в генеральной совокупности. До проведения обследования при-

ближенно оценить дисперсию или среднее квадратическое отклонение можно на следующей основе:

- исходя из результатов специально организованного пробного обследования;
- опираясь на данные предыдущих обследований, как выборочных, так и сплошных. В последние годы в статистической практике все чаще вместо сплошного наблюдения применяют выборочный метод. Например, с 1996 г. проводят выборочное наблюдение за деятельностью малых предприятий. Таким образом, дисперсию изучаемого признака в выборке можно оценить, зная коэффициент вариации, значение которого получено по итогам предшествующего сплошного наблюдения или предшествующей выборки. Коэффициент вариации $V = \frac{\sigma}{\bar{x}} 100\%$. Следовательно, дисперсия $\sigma^2 = \frac{V^2(\bar{x})^2}{100^2}$;
- исходя из закона распределения изучаемого признака в генеральной совокупности. Если распределение близко к нормальному, то размах вариации R в 6 раз больше среднего квадратического отклонения: $R = 6\sigma$, где $R = x_{max} - x_{min}$. В таком случае, зная максимальное и минимальное значения признака, можно оценить σ:

$$\sigma = \frac{R}{6} = \frac{x_{max} - x_{min}}{6}.$$

Если в результате выборочного обследования необходимо установить долю единиц, обладающих определенным значением альтернативного признака, то дисперсия для доли будет равна pq. В этом случае формула необходимой численности выборки примет вид

$$n = \frac{t^2 pq}{\Delta^2}. \qquad (6.11)$$

Максимальное значение дисперсии альтернативного признака равно 0,25, т.е. $\max(pq) = 0{,}25$ (при $p = q = 0{,}5$). Если доля единиц, обладающих изучаемым признаком, т.е. p, неизвестна, в расчете необходимой численности выборки можно использовать указанное максимальное значение для дисперсии альтернативного признака.

На практике величина допустимой ошибки выборки, как правило, устанавливается не в абсолютном, а в относительном выра-

жении: $\Delta_{\text{отн}} = \frac{\Delta}{\bar{x}}100\%$. Так как $\sigma^2 = \frac{V^2(\bar{x})^2}{100^2}$, формулу для определения необходимой численности выборки при собственно случайном или механическом повторном отборе можно представить следующим образом:

$$n = \frac{t^2\sigma^2}{\Delta^2} = \frac{t^2V^2(\bar{x})^2}{\Delta^2_{\text{отн}}(\bar{x})^2} = \frac{t^2V^2}{\Delta^2_{\text{отн}}}. \qquad (6.12)$$

Пример. Для изучения товарооборота по выделенной товарной группе планируется провести выборочное обследование торговых предприятий региона. Сколько предприятий розничной торговли необходимо обследовать, если по данным предшествующего обследования известно, что коэффициент вариации товарооборота по данной группе товаров составляет 90%, а предельная относительная ошибка выборки с вероятностью 0,95 не должна превысить 5%?

При $P = 0{,}95$ коэффициент доверия $t = 1{,}96$ (см. Приложение 2). Следовательно, $n = \frac{1{,}96^2 \cdot 90^2}{5^2} = 1245$, т.е. при повторном отборе необходимо обследовать 1245 торговых предприятий.

Рассмотрим формулы для нахождения необходимой численности выборки при бесповторном отборе.

Предельная ошибка выборки при собственно случайном или механическом бесповторном отборе рассчитывается по формуле $\Delta = t\sqrt{\frac{\sigma^2}{n}\left(1 - \frac{n}{N}\right)}$, поэтому необходимая для достижения заданной ошибки численность выборки

$$n = \frac{Nt^2\sigma^2}{N\Delta^2 + t^2\sigma^2}. \qquad (6.13)$$

Если задана предельная относительная ошибка выборки и известен коэффициент вариации, то численность выборки определяется по формуле

$$n = \frac{Nt^2V^2(\bar{x})^2}{N\Delta^2_{\text{отн}}(\bar{x})^2 + t^2V^2(\bar{x})^2} = \frac{Nt^2V^2}{N\Delta^2_{\text{отн}} + t^2V^2}. \qquad (6.14)$$

При бесповторном отборе для нахождения доли альтернативного признака необходимая численность выборки

$$n = \frac{Nt^2pq}{N\Delta^2 + t^2pq}. \qquad (6.15)$$

Пример. Исходя из условия предыдущего примера, но зная, что общее число торговых предприятий, осуществляющих продажу товаров изучаемой группы, составляет 15 тыс. единиц, объем выборки при бесповторном отборе рассчитывают следующим образом:

$$n = \frac{15000 \cdot 1{,}96^2 \cdot 90^2}{15000 \cdot 5^2 + 1{,}96^2 \cdot 90^2} = \frac{466754400}{406116{,}96} = 1150 \text{ единиц},$$

т.е. выборка должна быть 8-процентной.

Как правило, цель выборочного обследования – определить пределы, в которых находится в генеральной совокупности не один, а несколько показателей. В таком случае дисперсия для каждого из них будет различна, соответственно будет различаться и необходимая численность выборки. Число обследуемых единиц будет максимальным при изучении показателя с максимальной дисперсией. Соответственно, и необходимая численность выборки должна быть принята на максимальном уровне из всех рассчитанных.

В табл. 6.9 приведены формулы для нахождения необходимой численности выборки при разных способах отбора, где $\overline{p_i q_i}$ – средняя из групповых дисперсий в генеральной совокупности, $\delta_{\bar{x}}^2$ – межсерийная дисперсия в генеральной совокупности, δ_p^2 – межсерийная дисперсия для доли в генеральной совокупности.

Значения величин $\overline{p_i q_i}$, $\delta_{\bar{x}}^2$ и δ_p^2 устанавливаются так же, как и дисперсия при собственно случайном отборе. Если их оценка основана на данных пробных выборок, то в соответствующих формулах вместо генеральных характеристик необходимо использовать выборочные.

Таким образом, формула предельной ошибки выборки используется не только для оценки пределов, в которых находится изучаемый признак в генеральной совокупности, но и для определения необходимого объема выборки при заданной ее ошибке. Третий тип задач, которые могут быть решены с использованием предельной ошибки выборки, – это определение вероятности, с которой можно гарантировать, что ошибка выборки не выйдет за заданные пределы.

Таблица 6.9

Формулы для нахождения необходимой численности выборки при разных способах отбора

Способ отбора	Оцениваемый параметр	Повторный отбор	Бесповторный отбор
Собственно случайный и механический	Средняя	$n = \frac{t^2\sigma^2}{\Delta^2}$	$n = \frac{Nt^2\sigma^2}{N\Delta^2 + t^2\sigma^2}$
	Доля	$n = \frac{t^2 pq}{\Delta^2}$	$n = \frac{Nt^2 pq}{N\Delta^2 + t^2 pq}$
Типический	Средняя	$n = \frac{t^2\overline{\sigma_i^2}}{\Delta^2}$	$n = \frac{Nt^2\overline{\sigma_i^2}}{N\Delta^2 + t^2\overline{\sigma_i^2}}$
	Доля	$n = \frac{t^2\overline{p_i q_i}}{\Delta^2}$	$n = \frac{Nt^2\overline{p_i q_i}}{N\Delta^2 + t^2\overline{p_i q_i}}$
Серийный	Средняя	$s = \frac{t^2\delta_{\bar{x}}^2}{\Delta^2}$	$s = \frac{St^2\delta_{\bar{x}}^2}{(S-1)\Delta^2 + t^2\delta_{\bar{x}}^2}$
	Доля	$s = \frac{t^2\delta_p^2}{\Delta^2}$	$s = \frac{St^2\delta_p^2}{(S-1)\Delta^2 + t^2\delta_p^2}$

Примечание. При серийном отборе на основе приведенных формул определяется число серий, которое необходимо обследовать, так как они являются единицей наблюдения при данном способе отбора.

Пример. При проведении 10-процентного выборочного обследования предприятий оптовой торговли одного из регионов было установлено, что в среднем на одно предприятие приходилось 16 работников при среднем квадратическом отклонении 12 человек. Обследовано 100 предприятий. С какой вероятностью можно утверждать, что относительная предельная ошибка выборки не превысит 5%?

Так как $\Delta_{\text{отн}} = \frac{\Delta}{\tilde{x}}100\%$, то $\Delta = \frac{\Delta_{\text{отн}}\,\tilde{x}}{100\%} = \frac{5 \cdot 16}{100} = 0{,}8$. Средняя ошибка при проведении обследования

$$\mu = \sqrt{\frac{\sigma^2}{n}\left(1 - \frac{n}{N}\right)} = \sigma\sqrt{\frac{1}{n}\left(1 - \frac{n}{N}\right)} = 12\sqrt{\frac{1}{100}(1 - 0{,}1)} \cong 1{,}14.$$

Коэффициент доверия $t = \frac{\Delta}{\mu} = \frac{0{,}8}{1{,}14} \cong 0{,}7$.

По таблице значений интеграла вероятностей Лапласа (см. Приложение 2) находим: при $t = 0{,}7$ вероятность $P \cong 0{,}516$. Следовательно, с вероятностью 0,516 можно гарантировать, что в результате проведенного обследования относительная предельная ошибка выборки не превысит 5%.

6.5. Малая выборка

Приведенные в параграфах 6.2–6.4 формулы средней ошибки выборки показывают, что ее величина зависит от объема выборки n, степени колеблемости изучаемого признака в генеральной совокупности и способа отбора. Для собственно случайной повторной выборки

$$\mu = \sqrt{\frac{\sigma_г^2}{n}} = \sqrt{\frac{\sigma^2}{n-1}}. \qquad (6.16)$$

Если объем выборки достаточно велик, единицей в знаменателе можно пренебречь. На практике иногда приходится отбирать из генеральной совокупности небольшое число единиц. В этом случае использование в формуле (6.16) вместо $(n-1)$ величины n может значительно повлиять на результат, т.е. занизить среднюю ошибку выборки. Как правило, выборка считается *малой*, если обследуется не более 30 единиц. Таким образом, ***средняя ошибка малой выборки*** при собственно случайном или механическом повторном отборе рассчитывается по формуле

$$\mu_{м.в} = \sqrt{\frac{\sigma^2}{n-1}}. \qquad (6.17)$$

В условиях малой выборки дисперсия выборочной совокупности не может рассматриваться в качестве оценки генеральной дисперсии.

Второе отличие заключается в том, что в выборках большого объема вероятность появления определенного нормированного отклонения выборочной средней от генеральной $t = \frac{\tilde{x} - \bar{x}}{\mu}$ подчиняется нормальному закону распределения независимо от того,

как распределены единицы в генеральной совокупности. Как следует из центральной предельной теоремы, предположение о нормальном распределении всех возможных значений выборочной средней и соответствующей величины t справедливо только при значительном объеме выборки.

В условиях малой выборки характер распределения единиц в генеральной совокупности оказывает влияние на вероятность появления той или иной ошибки выборки. В условиях нормально распределенной генеральной совокупности при $n < 30$ нормированное отклонение выборочной средней от генеральной $t = \dfrac{\tilde{x} - \bar{x}}{\mu_{\text{м.в}}}$ и соответствующая вероятность подчинены закону распределения Стьюдента, открытому в 1908 г. английским математиком У. Госсетом (псевдоним – Стьюдент).

Графически распределение Стьюдента имеет вид одновершинной кривой, которая симметрична относительно оси ординат и при увеличении объема выборки приближается к кривой нормального распределения (рис. 6.2). При $n > 100$ вероятность наступления того или иного значения t, найденная в соответствии с распределением Стьюдента, практически совпадает с соответствующей величиной интеграла вероятностей Лапласа. При $30 \leqslant n \leqslant 100$ расхождения между указанными значениями невелики, поэтому на практике данное распределение используется лишь для $n < 30$.

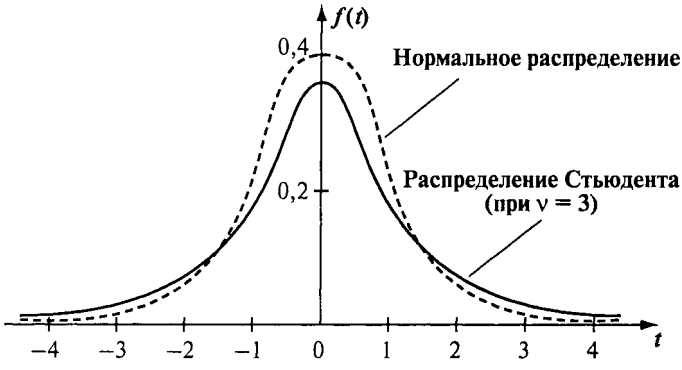


Рис. 6.2. Кривые нормального распределения и распределения Стьюдента (при ν = 3)

Согласно распределению Стьюдента, плотность распределения для нормированного отклонения выборочной средней от генеральной определяется по формуле

$$S(t) = \frac{\Gamma\left(\frac{\nu + 1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{\pi\nu}}\left(1 + \frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}}, \qquad (6.18)$$

где ν – число степеней свободы.

При определении выборочной дисперсии необходимо знать среднее значение признака, поэтому $\nu = n - 1$.

Гамма-функция имеет вид

$$\Gamma(u) = \int_0^\infty x^{u-1}e^{-x}dx. \qquad (6.19)$$

Последовательно подставляя в формулу (6.19) вместо u значения $\frac{\nu + 1}{2}$ и $\frac{\nu}{2}$, можно получить значения гамма-функции, которые необходимо использовать при расчете плотности распределения ошибок малой выборки.

Вероятность того, что нормированное отклонение выборочной средней от генеральной не превысит заданного значения t, будет равна площади, ограниченной кривой распределения Стьюдента и осью абсцисс в интервале от $-\infty$ до t:

$$S(t) = \frac{\Gamma\left(\frac{\nu + 1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{\pi\nu}}\int_{-\infty}^{t}\left(1 + \frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}} dt. \qquad (6.20)$$

Формула (6.20) свидетельствует о том, что в условиях малой выборки вероятность появления той или иной ошибки зависит не только от t, но и от объема выборки, так как $\nu = n - 1$. Чем меньше n, тем медленнее указанная кривая приближается к оси абсцисс (см. рис. 6.2). Следовательно, при малой выборке вероятность больши́х отклонений выборочной средней от генеральной более высока.

Найденное по формуле (6.20) значение функции $S(t)$ – это вероятность того, что фактический коэффициент доверия $t_{\text{факт}} \leqslant t$. Следовательно, $(1 - S(t))$ – вероятность того, что $t_{\text{факт}} > t$. Если

рассматривать абсолютную величину нормированного отклонения, т.е. $|t_{\text{факт}}|$, то вероятность ее выхода за заданные пределы t

$$P(|t_{\text{факт}}| > |t|) = 2[1 - S(t)].$$

Вероятность того, что это отклонение будет находиться в пределах от $-t$ до $+t$,

$$P(|t_{\text{факт}}| \leqslant |t|) = 1 - P(|t_{\text{факт}}| > |t|) = \\ = 1 - 2[1 - S(t)] = 2S(t) - 1. \qquad (6.21)$$

Кроме того, эта величина равна вероятности попадания среднего значения признака в генеральной совокупности в пределы $\tilde{x} \pm \Delta_{\text{м.в}}$ ($\tilde{x} - \Delta_{\text{м.в}} \leqslant \bar{x} \leqslant \tilde{x} + \Delta_{\text{м.в}}$), где $\Delta_{\text{м.в}}$ – предельная ошибка малой выборки.

Величины $S(t)$ для разных значений t и ν табулированы, поэтому в каждом конкретном случае нет необходимости выполнять расчет по формуле (6.20). Соответствующие таблицы могут быть представлены двумя способами:

- искомая вероятность $P(t)$ или $S(t)$ находится на пересечении строки, соответствующей значению коэффициента доверия t, и столбца, соответствующего числу степеней свободы ν или объему выборки n, так как $\nu = n - 1$ (см. Приложение 3);
- в клетках таблицы указывается значение коэффициента доверия t, соответствующее определенному числу степеней свободы ν и некоторым наиболее часто употребляемым значениям доверительной вероятности $P(t)$ (0,90; 0,95; 0,99) или уровню значимости, равному $\alpha = 1 - P(t)$ (0,10; 0,05; 0,01) (см. Приложение 9).

В статистических таблицах вероятности $S(t)$ и $P(t)$ не тождественны:

$S(t)$ – вероятность того, что фактическое значение нормированного отклонения выборочной средней от генеральной будет не больше, чем табличное значение, т.е. $S(t) = P(t_{\text{факт}} \leqslant t)$;

$P(t)$ – вероятность того, что $t_{\text{факт}}$ по абсолютной величине не превосходит значение $|t|$, т.е. $P(t) = P(|t_{\text{факт}}| \leqslant |t|) = 2S(t) - 1$.

Пример. В результате выборочной проверки налоговой инспекцией 10 промышленных предприятий города средняя доля документально неоформленных работ на них составила 17%. Определить вероятность того, что в генеральной совокупности доля документально неоформленных работ не превышает 25%.

Для нахождения средней ошибки малой выборки необходимо знать ее дисперсию:

$$\sigma^2 = w(1 - w) = 0{,}17(1 - 0{,}17) = 0{,}1411.$$

В таком случае

$$\mu_{\text{м.в}} = \sqrt{\frac{w(1-w)}{n-1}} = \sqrt{\frac{0{,}1411}{10-1}} \cong 0{,}125,$$

$$\Delta_{\text{м.в}} \leqslant p - w = 0{,}25 - 0{,}17 = 0{,}08.$$

Следовательно, $t = \dfrac{\Delta_{\text{м.в}}}{\mu_{\text{м.в}}} = \dfrac{0{,}08}{0{,}125} \cong 0{,}64$, отсюда $P(t_{\text{факт}} \leqslant 0{,}64) = S(t)$ при $\nu = n - 1 = 9$. По таблице Приложения 3 находим: при $t = 0{,}64$ и $n = 10$ вероятность $S(t) = 0{,}718$. Таким образом, с вероятностью 0,718 можно утверждать, что доля документально неоформленных работ на всех предприятиях города не превышает 25%.

Пример. При выборочном обследовании налоговой инспекцией 15 обменных пунктов города было установлено, что разница между курсом покупки и курсом продажи в среднем составляет 84 коп. за 1 долл. при среднем квадратическом отклонении 10 коп. С вероятностью 0,95 определить пределы, в которых находится разница между курсом покупки и курсом продажи валюты во всех обменных пунктах города.

Средняя ошибка малой выборки

$$\mu_{\text{м.в}} = \sqrt{\frac{\sigma^2}{n-1}} = \sqrt{\frac{100}{15-1}} = 2{,}67 \text{ коп.} \cong 3 \text{ коп.}$$

Так как $P(t) = 0{,}95$, то $S(t) = \dfrac{P(t)+1}{2} = \dfrac{0{,}95+1}{2} = 0{,}975$.

При $n = 15$ (т.е. $\nu = 15 - 1 = 14$) и $S(t) = 0{,}975$ по таблице Приложения 3 определяем, что значение t находится между 2,1 и 2,2. Более точно установить эту величину позволяет таблица Приложения 9: при $\nu = 14$ и $\alpha = 1 - P(t) = 0{,}05$ соответствующее значение $t = 2{,}145$. Следовательно, предельная ошибка выборки

$$\Delta_{\text{м.в}} = t\mu_{\text{м.в}} = 2{,}145 \cdot 2{,}67 = 5{,}73 \text{ коп.} \cong 6 \text{ коп.}$$

Пределы, в которых находится разница между курсом покупки и курсом продажи валюты, при найденном значении предельной ошибки выборки составляют

$$\tilde{x} - \Delta_{\text{м.в}} \leqslant \bar{x} \leqslant \tilde{x} + \Delta_{\text{м.в}},$$

т.е.

$$84 - 6 \leqslant \bar{x} \leqslant 84 + 6,$$

или

$$78 \text{ коп.} \leqslant \bar{x} \leqslant 90 \text{ коп.}$$

С вероятностью 0,95 можно утверждать, что в генеральной совокупности разница между курсом покупки и курсом продажи валюты составляет от 78 до 90 коп.

Выводы, сделанные на основе малой выборки, справедливы лишь при нормальном распределении значений изучаемого признака в генеральной совокупности. Поэтому использование малой выборки для оценки доли и средней в генеральной совокупности целесообразно в том случае, если исследователь не располагает необходимыми ресурсами для проведения выборки большего объема или если выборочное обследование связано с порчей единиц наблюдения (например, при проверке качества продуктов питания).

6.6. Распространение результатов выборочного наблюдения на генеральную совокупность

На заключительном этапе выборочного обследования решается вопрос о возможности распространения полученных результатов на генеральную совокупность. При этом учитываются два основных обстоятельства:

- насколько адекватно представлена генеральная совокупность в выборке, т.е. не изменилась ли в результате обследования структура запланированной ее основы, соблюдены ли основные пропорции между типическими группами в выборочной и генеральной совокупности. Вероятность возникновения таких нарушений достаточно велика в том случае, если единицей наблюдения является человек (например, он может отказаться отвечать на вопросы анкеты и т.п.).

  Для восстановления исходных пропорций генеральной совокупности проводится корректировка выборки либо путем отсечения части единиц, доля которых в выборке непропорционально велика по сравнению с долей в генеральной совокупности, либо путем многократного использования результатов наблюдения за единицами тех групп, которые недостаточно широко представлены в выборке;
- какова степень соответствия фактически полученной относительной ошибки выборки запланированному ее уровню. Фактическое значение относительной ошибки определяется путем сопоставления абсолютной величины предельной ошибки выборки, полученной в результате обследования, со средним уровнем признака, рассчитанным на основе выборки, т.е. $\Delta_{отн} = \frac{\Delta}{\bar{x}} 100\%$ (или для доли $\Delta_{отн} = \frac{\Delta}{w} 100\%$).

Если выборка адекватна генеральной совокупности и фактическая относительная ошибка выборки незначительно отличается от запланированного ее уровня, то на основе проведенного обследования можно оценить пределы, в которых находится среднее значение изучаемого признака (или доли) в генеральной совокупности, а также указать его возможное значение для совокупности в целом.

Оценивая пределы для среднего значения показателя в генеральной совокупности, необходимо указывать вероятность, с которой эти пределы гарантируются. Однако в официальных статистических публикациях пределы, как правило, не указываются, поскольку в них принята такая степень точности, что величины $\tilde{x} - \Delta$ и $\tilde{x} + \Delta$ с вероятностью, близкой к единице, практически совпадают. Так, при публикации результатов выборочных обследований домашних хозяйств по проблемам занятости средний возраст безработных приведен с точностью до десятых года (например, 34,4 года в 1996 г.), поскольку с вероятностью, близкой к единице, предельная ошибка выборки меньше 0,05 года.

Общее значение изучаемого показателя для совокупности в целом определяется двумя способами: методом прямого счета и методом коэффициентов.

Если в результате обследования получены верхняя и нижняя границы изучаемого признака в расчете на единицу совокупности, т.е. найдены величины $\tilde{x} - \Delta \leqslant \bar{x} \leqslant \tilde{x} + \Delta$, то с соответствующей вероятностью можно найти эти границы для совокупности в целом. Так как N – число единиц в генеральной совокупности, искомые пределы таковы:

$$N(\tilde{x} - \Delta) \leqslant N\bar{x} \leqslant N(\tilde{x} + \Delta).$$

Например, зная по результатам выборочного бюджетного обследования пределы для среднедушевого дохода на одного члена семьи, можно определить границы, в которых находится общая сумма доходов всего населения.

Метод коэффициентов используется для получения по данным выборки значений показателей, которые непосредственно не наблюдались, но тесно связаны с величинами, зафиксированными в ходе выборочного обследования. Этот метод используется также для уточнения данных сплошного наблюдения с помощью дополнительно проведенного выборочного обследования.

Пример. По данным переписи предприятий розничной торговли города установлено, что их общее число N_0 составило 350 единиц. Дополнительно проведенное выборочное обследование по-

казало, что из 54 торговых предприятий (n_1) бланк сплошного обследования заполнен по 50 единицам (n_0). В таком случае скорректированное общее число объектов генеральной совокупности

$$N_1 = N_0 \frac{n_1}{n_0},$$

где $\frac{n_1}{n_0}$ – коэффициент пересчета, основанный на данных выборочного обследования.

Итак,

$$\frac{n_1}{n_0} = \frac{54}{50} = 1{,}08.$$

Уточненное число предприятий розничной торговли

$$N_1 = 350 \cdot 1{,}08 = 378 \text{ единиц.}$$

6.7. Общие понятия и схема статистической проверки гипотез

Результаты выборочных наблюдений широко используются в статистике для проверки предположений, выдвигаемых в отношении характера или параметров распределения случайной величины в генеральной совокупности. Такие предположения, которые планируется проверить с помощью специальных статистических методов, называются ***статистическими гипотезами.***

Проверка статистической гипотезы заключается в том, чтобы оценить, можно ли считать случайным расхождение между выдвинутой гипотезой и результатами выборочного наблюдения. Такая оценка всегда носит вероятностный характер. Если расхождение между эмпирическими и теоретическими значениями не выходит за пределы случайной ошибки, то можно считать, что с заданной вероятностью выдвинутая гипотеза не опровергается. При этом справедливость самой гипотезы не доказывается, а лишь делается вывод о том, можно ли ее считать допустимой или необходимо отвергнуть.

Например, для санитарного контроля проводится мониторинг, в ходе которого устанавливается степень соответствия фактического содержания вредных веществ в атмосфере предельно допустимой концентрации (ПДК). Обозначим ПДК какого-либо вредного вещества, например двуокиси углерода, через x_0, а фактическую концентрацию, установленную в результате мониторинга, через x. Требуется проверить справедливость гипотезы о том, что содержа-

ние вредного вещества в атмосфере города можно признать допустимым. Если эта гипотеза не подтверждается, т.е. окажется, что $x > x_0$, то необходимы дополнительные меры по охране атмосферного воздуха.

Проверяемая гипотеза называется *основной* и обозначается через H_0. Суть проверки – убедиться в отсутствии систематической ошибки между исследуемым параметром генеральной совокупности и заданным его значением, т.е. проверяется гипотеза о нулевом расхождении между ними, поэтому основную гипотезу называют также *нулевой*.

При записи содержание гипотезы отделяется от символа H_0 двоеточием. В приведенном примере суть проверяемой гипотезы может быть представлена следующим образом:

$$H_0: \ x \leqslant x_0.$$

Гипотеза, *альтернативная* основной, обозначается через H_1. В нашем случае альтернативной является гипотеза о том, что содержание вредного вещества в атмосфере города превышает ПДК, т.е. $H_1: x > x_0$.

Выдвигаемые гипотезы могут быть простыми и сложными. *Простая* гипотеза однозначно характеризует оцениваемый параметр генеральной совокупности. Например, $H_0: x = x_0$, т.е. степень загрязнения воздуха точно соответствует ПДК. *Сложная* гипотеза определяет область возможных значений исследуемого параметра. Так, выдвинутая ранее гипотеза $H_0: x \leqslant x_0$ является сложной.

Поскольку при проверке гипотезы используются данные выборочного наблюдения, вывод о ее допустимости носит вероятностный характер, т.е. не исключена возможность ошибки. При этом могут возникать следующие ошибки:

- ***ошибка первого рода*** – если в результате проверки делается вывод о необходимости отклонить нулевую гипотезу, которая в действительности верна;
- ***ошибка второго рода*** – если нулевая гипотеза не отклоняется, хотя на самом деле она ошибочна.

Для того чтобы сделать вывод о соответствии результатов выборочного наблюдения выдвинутой гипотезе, необходимо принять определенный критерий, т.е. правила, в соответствии с которыми устанавливается, при каких результатах выборочного обследования основная гипотеза не может быть отклонена, а при каких от нее необходимо отказаться. Например, при проверке гипотезы о среднем значении признака в генеральной совокупности $H_0: \bar{x} = a$ в качестве критерия (θ) можно использовать среднее значение

признака в выборке $\tilde{x}$, отклонение выборочной средней от a (т.е. $\tilde{x} - a$), а также нормированное отклонение $t = \dfrac{\tilde{x} - a}{\mu}$.

Из множества значений статистического критерия необходимо выделить такое их подмножество, при попадании в которое выборочной характеристики основная гипотеза должна быть отклонена. Это подмножество называется ***критической областью.*** Ее границы устанавливаются таким образом, чтобы вероятность попадания в нее значений выборочной характеристики при условии справедливости выдвинутой гипотезы была величиной достаточно малой. Напомним, что указанная вероятность называется уровнем значимости критерия и обозначается через α. Если значение критерия попадает в критическую область при верной нулевой гипотезе, то эта гипотеза должна быть отвергнута, т.е. будет допущена ошибка первого рода, вероятность которой равна α. Уменьшая уровень значимости, мы снижаем вероятность появления ошибки первого рода. Однако если основная гипотеза неверна, то, уменьшая α, мы увеличиваем область допустимых значений и, соответственно, вероятность появления ошибки второго рода. Устанавливая уровень значимости, необходимо стремиться к минимизации возможных потерь, связанных с возникновением этих ошибок. Обычно уровень значимости принимается равным 0,05; 0,01; 0,005; 0,001. Если нулевая гипотеза верна, то вероятность ее принятия равна $(1 - \alpha)$.

Точка, разделяющая критическую область и область принятия нулевой гипотезы, называется ***критической.***

Обозначив вероятность ошибки второго рода через β, можно определить вероятность того, что при использовании для оценки гипотезы определенного критерия неверная гипотеза не будет принята. Эта вероятность определяет мощность критерия, и она равна $(1 - \beta)$.

Стремление увеличить мощность критерия при неизменном объеме выборки приводит к расширению критической области, т.е. повышает вероятность ошибки первого рода. Снизить вероятность их появления можно, лишь увеличив объем выборки, что практически не всегда возможно.

При заданном уровне значимости критическая область может быть определена как *односторонняя* или *двусторонняя* в зависимости от сформированной альтернативной гипотезы.

Предположим, что проверке подлежит гипотеза о среднем значении признака в генеральной совокупности H_0: $\bar{x} = a$, а в качестве критерия θ принята выборочная средняя.

Альтернативная гипотеза может быть представлена следующим образом: H_1: $\bar{x} < a$, H_1: $\bar{x} > a$ или H_1: $\bar{x} \neq a$.

1. H_1: $\bar{x} < a$. При достаточно большом объеме выборки распределение возможных значений выборочной средней приближается к нормальному распределению. При случайном расхождении между выборочными и генеральной средней они должны быть сгруппированы около величины $\bar{x} = a$. Если же среднее значение признака, полученное на основе выборки, значительно меньше, чем a, то выдвинутая гипотеза должна быть отклонена. В таком случае критическая область является *левосторонней* (рис. 6.3).

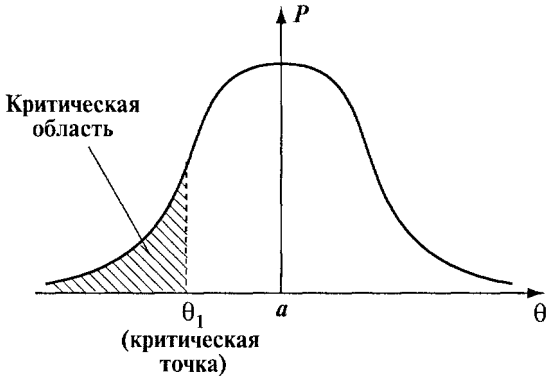


Рис. 6.3. Левосторонняя критическая область

Обозначив общую площадь, ограниченную кривой, отражающей распределение выборочной средней, через S, а площадь критической области через S_a, получаем: $\alpha = S_a/S = P(\theta < \theta_1)$. Принимая уровень значимости $\alpha = 0{,}05$, получаем $P(\theta < \theta_1) = 0{,}05$. Если в качестве критерия используется значение выборочной средней $\tilde{x}$, а критическая точка θ_1 представлена через случайную ошибку выборки, то можно записать:

$$P(\tilde{x} < a - t\mu) = \frac{1}{2} - \frac{1}{\sqrt{2\pi}} \int_{-t}^{0} e^{-\frac{t^2}{2}} dt = 0{,}05. \qquad (6.22)$$

Следовательно, $\frac{1}{\sqrt{2\pi}} \int_{-t}^{0} e^{-\frac{t^2}{2}} dt = 0{,}45$. В таком случае значение интеграла вероятностей Лапласа в пределах от $-t$ до $+t$

$$P(t) = \frac{1}{\sqrt{2\pi}} \int_{-t}^{+t} e^{-\frac{t^2}{2}} dt = \frac{2}{\sqrt{2\pi}} \int_{-t}^{0} e^{-\frac{t^2}{2}} dt = 0{,}9.$$

По таблице Приложения 2 находим соответствующее значение $t = 1{,}64$.

Таким образом, значение критерия в критической точке $\theta_1 = a - 1{,}64\mu$. Если по результатам выборочного обследования окажется, что $\tilde{x} > a - 1{,}64\mu$, то выдвинутая гипотеза H_0: $\bar{x} = a$ не отвергается. В противном случае ее необходимо отклонить.

2. H_1: $\bar{x} > a$. Критическая область при такой альтернативной гипотезе является *правосторонней* (рис. 6.4).

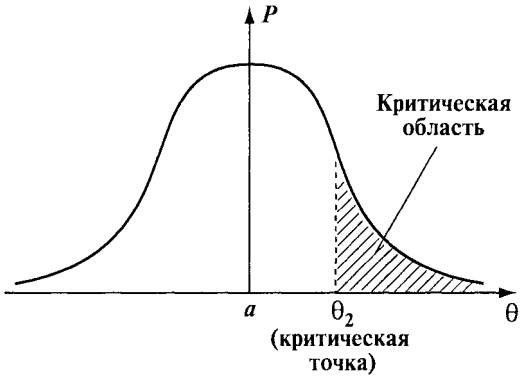


Рис. 6.4. Правосторонняя критическая область

При $\alpha = 0{,}05$ вероятность $P(\tilde{x} > a + t\mu) = 0{,}05$, т.е.

$$P(\tilde{x} > a + t\mu) = \frac{1}{2} - \frac{1}{\sqrt{2\pi}}\int_0^t e^{-\frac{t^2}{2}}dt = 0{,}05. \qquad (6.23)$$

Из этого следует, что $\frac{1}{\sqrt{2\pi}}\int_0^t e^{-\frac{t^2}{2}}dt = 0{,}45$, а соответствующее значение $t = 1{,}64$. В результате выбранный критерий в критической точке принимает значение $\theta_2 = a + 1{,}64\mu$.

3. H_1: $\bar{x} \neq a$. При такой формулировке альтернативной гипотезы строится двусторонняя критическая область, вероятность попадания в которую при $\alpha = 0{,}05$ равна сумме двух вероятностей:

$$P(\tilde{x} < \theta_1) + P(\tilde{x} > \theta_2) = 0{,}05.$$

Для того чтобы на основе одного уравнения найти две неизвестные критические точки, вероятности попадания в правую и левую ветви критической области принимаются равными (рис. 6.5), т.е.

$$P(\tilde{x} < \theta_1) = P(\tilde{x} > \theta_2) = \frac{0{,}05}{2} = 0{,}025.$$

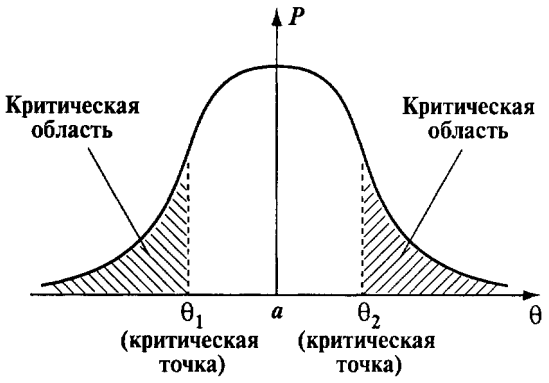


Рис. 6.5. Двусторонняя критическая область

Следовательно, критические точки θ_1 и θ_2 должны располагаться на равном расстоянии от величины a, т.е. $\theta_1 = a - t\mu$, $\theta_2 = a + t\mu$. В таком случае для нахождения t-статистики можно использовать значение интеграла вероятностей Лапласа в пределах от $-t$ до $+t$:

$$P(|\tilde{x} - a| > t\mu) = 1 - \frac{1}{\sqrt{2\pi}} \int_{-t}^{+t} e^{-\frac{t^2}{2}} dt = 0{,}05. \qquad (6.24)$$

Следовательно, $\frac{1}{\sqrt{2\pi}} \int_{-t}^{+t} e^{-\frac{t^2}{2}} dt = 0{,}95$, откуда табличное значение $t = 1{,}96$ (см. Приложение 2).

В результате критические точки имеют следующие значения:

$$\theta_1 = a - 1{,}96\mu, \qquad \theta_2 = a + 1{,}96\mu.$$

Полученные значения t при $\alpha = 0{,}05$ справедливы лишь в том случае, если объем выборочной совокупности достаточно велик. Если же выборка малая, то для нахождения t-статистики при заданном уровне значимости α необходимо использовать распределение Стьюдента (см. параграф 6.5).

Общий порядок проверки статистических гипотез таков:

- формулируется основная (проверяемая) и альтернативная гипотезы;
- выбирается статистический критерий для проверки справедливости гипотезы;
- определяются критическая область и область допустимых значений, значения критерия в критических точках;

- проводится выборочное обследование, по результатам которого рассчитывается фактическое значение выбранного критерия;
- на основе сравнения фактического и критического значений критерия делается вывод о правдоподобности или необходимости отклонения выдвинутой гипотезы.

В зависимости от вида проверяемых гипотез (о среднем значении, законе распределения, взаимосвязи признаков и т.д.) выбираются разные критерии (*t*-статистика (или коэффициент доверия), *t*-статистика Стьюдента, χ^2-критерий Пирсона; *F*-критерий Фишера и др.), которые подразделяются на параметрические и непараметрические. Для проведения оценки с использованием *параметрических* критериев необходимо знать закон распределения генеральной совокупности. *Непараметрические* критерии могут применяться при любом законе распределения, но при этом сохраняется главное условие – независимость испытаний при формировании выборочной совокупности, на основе которой проверяется выдвинутая гипотеза.

6.8. Проверка гипотез о средней и о доле

Гипотезы о средней

В статистической практике наиболее часто проверяются два вида гипотез о средних величинах:

- гипотеза о равенстве средней величины установленному нормативу;
- гипотеза о равенстве средних значений признака двух совокупностей.

Общий подход к проверке гипотезы о равенстве среднего значения признака в генеральной совокупности некоторой величине H_0: $\bar{x} = a$ описан в параграфе 6.7. В качестве критерия в этом случае целесообразно использовать нормированное отклонение выборочной средней от заданной величины:

$$t = \frac{\tilde{x} - a}{\mu}, \qquad (6.25)$$

где μ – средняя квадратическая ошибка выборочной средней, т.е. средняя ошибка выборки.

При большом объеме выборки ($n \geqslant 30$) средняя ошибка выборки μ рассчитывается по формуле $\mu = \sqrt{\frac{\sigma^2}{n}}$, а при $n < 30$ – по

формуле $\mu = \sqrt{\frac{\sigma^2}{n-1}}$. Если полученное по результатам обследования фактическое значение t-статистики меньше табличного, т.е. $t_{факт} < t$, то гипотеза не отклоняется. В противном случае нулевую гипотезу следует отклонить.

Пример. При оценке влияния изменений в налоговой политике на платежеспособность предприятий одного из регионов установлено, что до указанных изменений средний коэффициент покрытия по этим предприятиям соответствовал нормативу, равному 2. После внесения изменений в действующую налоговую систему было проведено выборочное обследование 49 предприятий региона, в результате которого установлено, что средний коэффициент покрытия на них составил 1,7 при среднем квадратическом отклонении 0,6.

Выдвигаемая нулевая гипотеза состоит в том, что изменения в проводимой налоговой политике существенно не повлияли на платежеспособность предприятий региона, т.е. коэффициент покрытия остался на прежнем уровне: H_0: $\bar{x} = 2$. В качестве альтернативной может быть рассмотрена гипотеза о том, что указанные изменения повлияли на степень платежеспособности предприятий: H_1: $\bar{x} \neq 2$.

Для проверки выдвинутой гипотезы примем уровень значимости $\alpha = 0{,}05$. Так как вероятность $P(|\bar{x} - 2| \geqslant t\mu) = 0{,}05$, а $n > 30$, то для значения интеграла вероятностей Лапласа $\frac{1}{\sqrt{2\pi}} \int_{-t}^{+t} e^{-\frac{t^2}{2}} dt =$ $= 0{,}95$ находим табличное значение t-статистики: $t = 1{,}96$ (см. Приложение 2).

Фактическое значение t-статистики

$$t_{факт} = \frac{|\bar{x} - a|}{\mu} = \frac{|\bar{x} - a|}{\sqrt{\frac{\sigma^2}{n}}} = \frac{|1{,}7 - 2|}{\sqrt{\frac{0{,}36}{49}}} = 3{,}5.$$

Так как $t_{факт} > t$, то выдвинутая гипотеза отклоняется, т.е. изменения в налоговой системе повлияли на платежеспособность предприятий региона.

Для того чтобы сделать более определенный вывод о характере этих изменений, альтернативную гипотезу сформулируем следующим образом: изменения в налоговой системе привели к снижению платежеспособности предприятий региона, т.е. H_1: $\bar{x} < 2$.

Зададим для этого случая уровень значимости $\alpha = 0{,}01$. Вероятность $P(\bar{x} < 2 - t\mu) = 0{,}01$, следовательно, значение интеграла вероятностей Лапласа в пределах от $-t$ до 0 равно $\frac{1}{\sqrt{2\pi}}\int_{-t}^{0} e^{-\frac{t^2}{2}} dt =$ $= \frac{1}{2} - 0{,}01 = 0{,}49$, а в пределах от $-t$ до $+t$ соответственно 0,98. По таблице Приложения 2 находим для данной вероятности значение t-статистики: $t = 2{,}33$. Так как $t_{факт} > t$, то нулевая гипотеза должна быть отклонена, т.е. с вероятностью 0,99 можно считать, что изменения в налоговой системе привели к снижению платежеспособности предприятий региона.

Если для проверки выдвинутой гипотезы используется малая выборка, то значение t-статистики определяется с помощью распределения Стьюдента. При этом степень обоснованности вывода зависит от того, насколько распределение генеральной совокупности соответствует нормальному закону.

Гипотеза о равенстве средних значений признака двух совокупностей выдвигается часто для того, чтобы проверить влияние какого-либо фактора на среднюю. Обозначим среднее значение признака в этих совокупностях через $\bar{x}_1$ и $\bar{x}_2$, а дисперсии в генеральных совокупностях – соответственно $\sigma^2_{г1}$ и $\sigma^2_{г2}$. В таком случае нулевая гипотеза может быть представлена следующим образом: H_0: $\bar{x}_1 = \bar{x}_2$. Для ее проверки проводится выборочное обследование, при котором объем выборки из первой совокупности составляет n_1, а из второй – n_2. Обозначим соответствующие значения средних в этих выборках через $\tilde{x}_1$ и $\tilde{x}_2$, а дисперсии – σ^2_1 и σ^2_2. В качестве критерия при проверке этой гипотезы принимается t-статистика, фактическое значение которой по результатам выборочного обследования рассчитывается по формуле

$$t_{факт} = \frac{|\tilde{x}_1 - \tilde{x}_2|}{\mu_{\tilde{x}_1 - \tilde{x}_2}}, \qquad (6.26)$$

где $\mu_{\tilde{x}_1 - \tilde{x}_2} = \sqrt{\frac{\sigma^2_{г1}}{n_1} + \frac{\sigma^2_{г2}}{n_2}}$ – стандартная ошибка разности выборочных средних.

Предположим, что дисперсии в двух совокупностях равны, т.е. $\sigma^2_{г1} = \sigma^2_{г2} = \sigma^2_г$. Следовательно,

$$\mu_{\tilde{x}_1 - \tilde{x}_2} = \sqrt{\frac{\sigma^2_г (n_1 + n_2)}{n_1 n_2}}. \qquad (6.27)$$

Если дисперсии в выборочных совокупностях известны, то они могут быть использованы для оценки общей дисперсии. Расчет проводится по формуле средней арифметической взвешенной, где в качестве весов выступает число степеней свободы в каждой выборке ($\nu = n - 1$):

$$\sigma_г^2 = \frac{\sigma_1^2(n_1 - 1) + \sigma_2^2(n_2 - 1)}{(n_1 - 1) + (n_2 - 1)} = \frac{\sigma_1^2(n_1 - 1) + \sigma_2^2(n_2 - 1)}{n_1 + n_2 - 2}.$$

Так как $\sigma_1^2 = \frac{\sum(x_{i1} - \tilde{x}_1)^2}{n_1 - 1}$, а $\sigma_2^2 = \frac{\sum(x_{i2} - \tilde{x}_2)^2}{n_2 - 1}$, то

$$\sigma_г^2 = \frac{\sum(x_{i1} - \tilde{x}_1)^2 + \sum(x_{i2} - \tilde{x}_2)^2}{n_1 + n_2 - 2} = \frac{n_1\sigma_1^2 + n_2\sigma_2^2}{n_1 + n_2 - 2}.$$

Подставим полученное выражение в формулу (6.26), учитывая также формулу (6.27):

$$t_{факт} = \frac{|\tilde{x}_1 - \tilde{x}_2|\sqrt{n_1 + n_2 - 2}\sqrt{n_1 n_2}}{\sqrt{n_1\sigma_1^2 + n_2\sigma_2^2}\sqrt{n_1 + n_2}}. \qquad (6.28)$$

Сравнивая фактическое значение t-статистики, рассчитанное по формуле (6.28), с табличным (см. Приложение 9) при заданном уровне значимости, можно сделать вывод о необходимости согласиться с выдвинутой гипотезой или отклонить ее.

Пример. Для оценки влияния формы собственности на платежеспособность предприятий отрасли проведено выборочное обследование частных и государственных предприятий, в результате которого получены данные, приведенные в табл. 6.10.

Таблица 6.10

Форма собственности	Число обследованных предприятий n_i	Средний коэффициент покрытия x_i	Дисперсия в выборочной совокупности σ_i^2
Частная	16	1,8	0,25
Государственная	10	1,2	1,18

В качестве нулевой выдвинем гипотезу о независимости степени платежеспособности предприятий от формы собственности,

т.е. о равенстве коэффициентов покрытия на предприятиях указанных форм собственности: H_0: $\bar{x}_1 = \bar{x}_2$. Альтернативной является гипотеза H_1: $\bar{x}_1 \neq \bar{x}_2$. При проверке выдвинутой гипотезы примем уровень значимости $\alpha = 0{,}05$. Рассчитаем по формуле (6.28) фактическое значение t-статистики:

$$t_{\text{факт}} = \frac{|1{,}8 - 1{,}2|\sqrt{16 + 10 - 2}\sqrt{16 \cdot 10}}{\sqrt{16 \cdot 0{,}25 + 10 \cdot 1{,}18}\sqrt{16 + 10}} = \frac{37{,}181}{20{,}268} = 1{,}83.$$

Табличное значение найдем на основе распределения Стьюдента при $\alpha = 0{,}05$ и числе степеней свободы $\nu = 16 + 10 - 2 = 24$.

Так как $P(t) = 2S(t) - 1 = 1 - \alpha = 0{,}95$, то $S(t) = \frac{1{,}95}{2} = 0{,}975$. Соответствующее табличное значение $t = 2{,}0639$ (см. Приложение 9). Фактическое значение t-статистики меньше табличного, следовательно, с вероятностью 0,95 можно считать, что платежеспособность предприятий не зависит от принятой на них формы собственности.

Гипотезы о доле

Аналогичные два вида гипотез могут быть проверены и для доли:

- гипотеза о равенстве доли единиц, обладающих определенным признаком, нормативу;
- сравнение долей единиц, обладающих определенным признаком, в двух совокупностях.

Порядок проверки гипотез первого вида аналогичен порядку, приведенному для средней, т.е. проверяется гипотеза H_0: $p = a$, где p – доля единиц, обладающих изучаемым признаком в генеральной совокупности, a – норматив. Альтернативными могут быть гипотезы трех видов:

1) H_1: $p \neq a$; 2) H_1: $p > a$; 3) H_1: $p < a$.

В качестве критерия также может быть принято значение t-статистики. Фактическое значение величины t рассчитывается по формуле

$$t_{\text{факт}} = \frac{w - a}{\mu}, \qquad (6.29)$$

где w – доля изучаемого признака в выборке;
μ – средняя ошибка выборки для доли.

Для выборки большого объема $\mu = \sqrt{\frac{w(1-w)}{n}}$, для малой выборки $\mu = \sqrt{\frac{w(1-w)}{n-1}}$.

Табличное значение t-статистики, как и для средней, находится на основе интеграла вероятностей Лапласа или распределения Стьюдента (для малой выборки).

При сравнении долей единиц, обладающих определенным признаком, в двух совокупностях применяется схема, аналогичная приведенной ранее для проверки соответствующей гипотезы о средней. В качестве критерия можно использовать t-статистику. Фактическое значение критерия в этом случае рассчитывается по формуле

$$t_{\text{факт}} = \frac{w_1 - w_2}{\mu_{w_1 - w_2}}, \tag{6.30}$$

где w_1 и w_2 – доля единиц, обладающих изучаемым признаком, в сравниваемых выборках;

$\mu_{w_1 - w_2}$ – стандартная ошибка разности выборочных долей.

Стандартная ошибка выборки может быть рассчитана по формуле

$$\mu_{w_1 - w_2} = \sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}, \tag{6.31}$$

где p – доля признака в генеральной совокупности;

n_1 и n_2 – объем каждой из двух выборок.

Эта формула справедлива, если величина p в двух сравниваемых генеральных совокупностях одинакова. Так как при проверке нулевой гипотезы величина p неизвестна, в формуле (6.31) можно использовать ее оценку, полученную по результатам выборочного обследования:

$$p = \frac{m_1 + m_2}{n_1 + n_2} = \frac{w_1 n_1 + w_2 n_2}{n_1 + n_2},$$

где m_1 и m_2 – частота изучаемого признака в каждой из двух выборок.

Сравнение фактического и табличного значений t-статистики позволяет отклонить или не отклонить выдвинутую гипотезу. Для сравнения двух долей можно использовать также χ^2-критерий Пирсона (см. параграф 5.8).

Глава 7
СТАТИСТИЧЕСКОЕ ИЗУЧЕНИЕ КОРРЕЛЯЦИОННЫХ ВЗАИМОСВЯЗЕЙ

7.1. Понятие корреляционной зависимости

Один из наиболее общих законов объективного мира – закон всеобщей связи и зависимости между явлениями. Естественно, что, исследуя явления в самых различных областях, статистика неизбежно сталкивается с зависимостями как между количественными, так и между качественными показателями, признаками. Ее задача – обнаружить (выявить) такие зависимости и дать им количественную характеристику.

Среди взаимосвязанных признаков (показателей) одни могут рассматриваться как определенные факторы, влияющие на изменение других, а вторые – как следствие, результат влияния первых. Соответственно, первые, т.е. признаки, влияющие на изменение других, называют ***факторными***, а вторые – ***результативными***.

Говоря о взаимосвязи между отдельными признаками, следует различать два вида связи: функциональную и стохастическую (статистическую), частным случаем которой является корреляционная связь.

Связь между двумя переменными x и y называется ***функциональной***, если определенному значению переменной x строго соответствует одно или несколько значений другой переменной y, и с изменением значения x значение y меняется строго определенно.

Такие связи обычно встречаются в точных науках: математике, физике и др. Например, известно, что площадь квадрата равна квадрату его стороны, т.е. $S = a^2$. При увеличении стороны квадрата в 2 раза, его площадь увеличится в 4 раза. Это соотношение характерно для любого квадрата, т.е. эта связь проявляется постоянно для каждого единичного случая (квадрата). Это *жестко детерминированная* связь.

Детерминированные связи можно встретить и в области экономических явлений. Например, при простой сдельной оплате труда связь между оплатой труда y и количеством изготовленных изделий x при фиксированной расценке за одну деталь, например 5 руб., легко выразить формулой $y = 5x$.

Существуют и иного рода связи, встречающиеся в области экономических и некоторых других явлений, где взаимно действуют многие факторы, комбинация которых приводит к вариации значений результативного признака (показателя) при одинаковом значении факторного признака.

Так, например, при изучении зависимости урожайности определенной культуры от количества выпавших осадков (или внесенных в почву удобрений) последние будут рассматриваться как факторный признак, а урожайность – как результативный. Между ними нет жестко детерминированной связи, т.е. при одном и том же количестве выпавших осадков (или внесенных удобрений) урожайность в разных хозяйствах, на разных участках земли будет неодинаковой, так как кроме осадков (или удобрений) на урожайность влияет много других факторов (качество семян, густота посева, уход за посевами, своевременность уборки и др.), комбинация которых вызывает вариацию урожайности.

Там, где взаимодействует множество факторов, в том числе и случайных, выявить зависимости, рассматривая единичный случай, невозможно.

Такие связи можно обнаружить только при массовом наблюдении как ***статистические закономерности*** (на основе изучения особенностей распределения, поведения средних и других показателей). Выявленная таким образом связь именуется ***стохастической*** или ***статистической***.

Корреляционная связь – понятие более узкое, чем статистическая связь, это, как уже говорилось, частный случай статистической (стохастической) связи.

Предметом изучения статистики являются в основном стохастические, корреляционные связи.

Слово «корреляция» (от английского *correlation*) означает соотношение, соответствие. Оно удачно отражает особенность зависимости, при которой определенному значению одного факторного признака может соответствовать несколько значений результативного показателя. На основе этих значений можно определить среднюю величину последнего, соответствующую каждому конкретному значению одного факторного признака или ряда признаков.

Связь, проявляющаяся при большом числе наблюдений в виде определенной зависимости между *средним значением результативного признака* и *признаками-факторами*, называется ***корреляционной***. Другими словами, корреляционную связь условно можно рассматривать как своего рода функциональную связь средней величины одного признака (результативного) со значением дру-

гого (или других). При этом, если рассматривается связь средней величины результативного показателя y с одним признаком-фактором x, корреляция называется *парной*, а если факторных признаков два и более (x_1, x_2, ..., x_m) – *множественной*.

При изучении множественной корреляции вводится еще понятие *частной корреляции*, под которой понимается зависимость между результативным показателем y и одним из факторных признаков x_i в условиях, когда влияние на них остальных факторов, учитываемых на фиксированном уровне, устранено.

По характеру изменений x и y в парной корреляции различают прямую и обратную связь.

При *прямой* зависимости значения обоих признаков изменяются в одном направлении, т.е. с увеличением значений x увеличиваются и значения y, с уменьшением значений факторного признака уменьшаются и значения результативного признака. Например, с ростом годового дохода в семье увеличивается (при прочих равных условиях) сумма сбережений за год или при уменьшении расхода электроэнергии на единицу продукции снижается себестоимость продукции.

При *обратной* зависимости значения факторного и результативного признаков изменяются в разных направлениях: например, при росте производительности труда себестоимость единицы продукции снижается или при снижении себестоимости продукции прибыль на предприятиях увеличивается и т.п.

Изучение корреляционных связей сводится в основном к решению следующих задач:

- выявление наличия (или отсутствия) корреляционной связи между изучаемыми признаками. Эта задача может быть решена на основе параллельного сопоставления (сравнения) значений x и y в каждой из n единиц совокупности, а также с помощью группировок и путем построения и анализа специальных корреляционных таблиц;
- измерение тесноты связи между двумя (и более) признаками с помощью специальных коэффициентов. Эта часть исследования именуется *корреляционным анализом*;
- определение уравнения регрессии – математической модели, в которой среднее значение результативного признака y рассматривается как функция одной или нескольких переменных – факторных признаков. Эта часть исследования именуется *регрессионным анализом*.

Последовательность рассмотрения перечисленных задач, естественно, может меняться в каждом конкретном исследовании.

Общий термин «*корреляционно-регрессионный анализ*» подразумевает всестороннее исследование корреляционных связей, в том числе нахождение уравнений регрессии, измерение тесноты и направления связи, а также определение возможных ошибок как параметров уравнений регрессии, так и показателей тесноты связи.

Для решения этих задач в статистике разработаны и широко используются различные методы и показатели (коэффициенты), одни из которых простейшие, а другие более сложные, основанные на вероятностных математических оценках.

Использование тех или иных приемов, методов определяется конкретной целью исследования. Так, в одних случаях достаточно просто констатировать факт наличия связи, обнаружения ее на массовых данных, в других – требуется количественно оценить эту связь, выявить роль отдельных факторов в изменении сложного результативного показателя, использовать модели связи для прогнозирования и т.п. Для решения сложных задач корреляционно-регрессионного анализа разработаны специальные компьютерные программы.

Теория корреляции начала разрабатываться во второй половине XIX в. и особенного расцвета достигла в XX в. Основоположниками теории корреляции являются английские биометрики Ф. Гальтон и К. Пирсон. В России их идеи получили развитие в трудах А.А. Чупрова.

7.2. Методы выявления корреляционной связи

Корреляционная зависимость между двумя признаками как частный случай стохастической связи выражается в вариации результативного признака y, вызванной изменением определенного факторного признака x в условиях взаимодействия его с множеством других факторов, не учитываемых при исследовании, но имеющихся в реальности.

Для выявления наличия и характера такой связи в статистике используется ряд методов: рассмотрение параллельных данных (значений x и y в каждой из n единиц); графический метод; метод аналитических группировок и корреляционных таблиц; расчет коэффициентов корреляции.

7.2.1. Параллельное рассмотрение значений x и y в каждой из n единиц

При небольшом числе наблюдений наличие корреляционной связи между двумя признаками x и y часто можно выявить визуально, путем простого параллельного сравнения их значений у отдельных единиц.

Для этого единицы наблюдения располагают по возрастанию значений факторного признака x и затем сравнивают с ним поведение значений результативного признака y. Примером таких параллельных данных могут служить приведенные в табл. 7.1 условные показатели по 10 предприятиям (однотипным) о стоимости основных производственных фондов x и валовом выпуске продукции y. (Предприятия расположены по возрастанию значений x.)

Таблица 7.1

Основные показатели деятельности предприятий
(данные условные)

Предприятия	Основные производственные фонды, млн руб. x_i	Валовой выпуск продукции, млн руб. y_i	Знаки отклонений от средней величины	
			$x_i - \bar{x}$	$y_i - \bar{y}$
1	12	28	–	–
2	16	40	–	–
3	25	38	–	–
4	38	65	–	–
5	43	80	–	–
6	55	101	+	+
7	60	95	+	–
8	80	125	+	+
9	91	183	+	+
10	100	245	+	+
Σ	520	1000		

В приведенном примере по мере увеличения значений x увеличиваются и значения y, хотя в отдельных случаях после возрастания наблюдается и уменьшение значений результативного признака (например, 38 после 40, 95 после 101). В целом же можно говорить, что чем больше стоимость основных фондов, тем больше валовой выпуск продукции, т.е. связь между x и y прямая.

Такое «субъективное» суждение о наличии корреляционной связи обычно сопровождается расчетом того или иного показателя, используемого для измерения тесноты связи: коэффициента Фехнера, ранговых коэффициентов корреляции, линейного коэффициента корреляции.

Коэффициент Фехнера (коэффициент корреляции знаков) – простейший показатель тесноты связи. Он основан на сравнении поведения отклонений индивидуальных значений каждого признака (x и y) от своей средней величины. При этом во внимание

принимаются не величины отклонений $(x_i - \bar{x})$ и $(y_i - \bar{y})$, а их знаки («+» или «–»). Определив знаки отклонения от средней величины в каждом ряду, рассматривают все пары знаков и подсчитывают число их совпадений и несовпадений. Если совпадение знаков обозначить символом C, а несовпадений – H, то коэффициент Фехнера можно записать как отношение разности чисел пар совпадений и несовпадений знаков к их сумме, т.е. к общему числу наблюдаемых единиц:

$$K_{\Phi} = \frac{\sum C - \sum H}{\sum C + \sum H}. \qquad (7.1)$$

Очевидно, что если знаки всех отклонений по каждому признаку совпадут, то $\sum H = 0$ и тогда $K_{\Phi} = 1$. Это характеризует наличие прямой связи. Если все знаки не совпадут, то $\sum C = 0$ и тогда $K_{\Phi} = -1$ (обратная связь). Если же $\sum C = \sum H$, то $K_{\Phi} = 0$. Итак, как и любой показатель тесноты связи, коэффициент Фехнера может принимать значения от 0 до ± 1. При этом чем ближе значение к 1, тем больше (сильнее) теснота зависимости между x и y. Однако равенство коэффициента Фехнера единице ни в коей мере нельзя воспринимать как свидетельство функциональной зависимости между x и y.

Чтобы определить коэффициент Фехнера в нашем примере (см. табл. 7.1), рассчитаем средние величины в каждом ряду:

$$\bar{x} = \frac{\sum x}{n} = \frac{520}{10} = 52; \qquad \bar{y} = \frac{\sum y}{n} = \frac{1000}{10} = 100.$$

В двух последних графах табл. 7.1 приведены знаки отклонений каждого значения x и y от своей средней величины. Число совпадений знаков составило 9, а число несовпадений равно 1. Отсюда

$$K_{\Phi} = \frac{\sum C - \sum H}{\sum C + \sum H} = \frac{9 - 1}{9 + 1} = 0{,}8.$$

Обычно такое значение показателя тесноты связи характеризует сильную зависимость.

Следует иметь в виду, что поскольку коэффициент Фехнера зависит только от знаков и не учитывает величину самих отклонений x и y от их средних величин, то он практически характеризует не столько тесноту связи, сколько ее наличие и направление. Так, в рассматриваемом примере по значению и знаку коэффициента Фехнера можно сказать, что между x и y существует прямая корреляционная связь.

Наличие корреляционной связи можно подтвердить (или опровергнуть) также с помощью других показателей (коэффициентов), используемых в статистике для измерения тесноты связи (см. параграфы 7.3 и 7.4).

Корреляционную зависимость для наглядности можно изобразить графически. Для этого, имея n взаимосвязанных пар значений x и y, пользуясь прямоугольной системой координат, каждую такую пару изображают в виде точки на плоскости с координатами x и y. Соединяя последовательно нанесенные точки, получают ломаную линию, именуемую *эмпирической линией регрессии.*

На рис. 7.1 изображена эмпирическая линия регрессии по данным табл. 7.1.

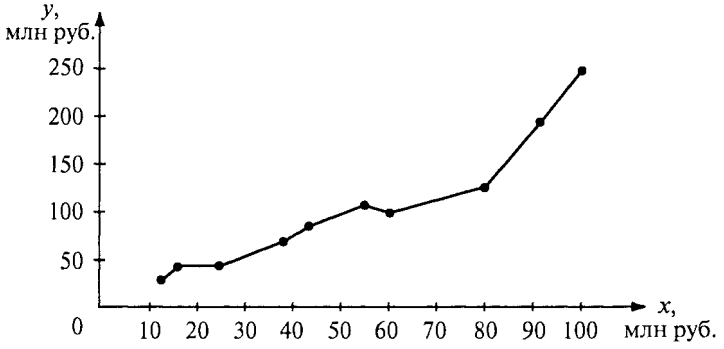


Рис. 7.1. Эмпирическая линия регрессии

7.2.2. Метод группировок

При большом числе наблюдений для выявления корреляционной связи между двумя количественными показателями x и y удобнее пользоваться методом группировок.

Чтобы выявить наличие корреляционной связи между двумя признаками, проводится группировка единиц совокупности по факторному признаку x и для каждой выделенной группы рассчитывается среднее значение результативного признака $\bar{y}_j$. Если результативный признак y зависит от факторного x, то в изменении среднего значения результативного признака $\bar{y}_j$ будет прослеживаться определенная закономерность.

Примером такой группировки могут служить данные об издержках обращения предприятий оптовой торговли с различным товарооборотом (табл. 7.2).

Таблица 7.2

Распределение уровня издержек обращения по группам предприятий оптовой торговли в апреле 1995 г.

Оптовый товарооборот, млн руб.	Количество предприятий	Издержки обращения, % к оптовому товарообороту
Менее 25	9362	46,0
26—50	3633	26,5
51—100	3618	24,4
101—200	3261	23,0
201—500	3034	17,6
Более 500	3100	16,9

Источник: Российский статистический ежегодник. 1996.

В последней графе табл. 7.2 приведены средние величины, рассчитанные на основе индивидуальных данных об издержках отдельных предприятий каждой группы.

Данные таблицы свидетельствуют о снижении среднего показателя издержек обращения от группы к группе, т.е. чем крупнее предприятия оптовой торговли (по объему товарооборота), тем меньше издержки обращения.

Таким образом, с помощью простой аналитической группировки можно выявить наличие зависимости между рассматриваемыми показателями: объемом товарооборота как показателем размера предприятий и средним уровнем издержек обращения.

Результаты группировки единиц совокупности могут быть оформлены и по-иному, в виде таблицы, в которой приведено комбинационное распределение единиц совокупности по двум признакам. Такие таблицы называют *таблицами взаимной сопряженности.*

Если в таблице оба признака, по которым дано распределение единиц совокупности, количественные, то такая таблица взаимной сопряженности называется *корреляционной.*

Корреляционная таблица строится по типу «шахматной», т.е. в подлежащем таблицы выделяются группы по факторному признаку x, в сказуемом — по результативному y или наоборот, а в клетках таблицы на пересечении x и y показано число случаев совпадения каждого значения x с соответствующим значением y. Общий вид такой таблицы показан на условном распределении 40 единиц по признакам x и y (табл. 7.3). (В качестве x может рассматриваться, например, стаж работы (число лет), а в качестве y — производительность труда (число изделий, вырабатываемых в час одним рабочим), $n = 40$ — число рабочих.)

Таблица 7.3

Корреляционная таблица

Значение признака x_j	Значение признака y_i				*Итого* (число единиц) $f_x = f_j$	Среднее значение по группам $\bar{y}_j$
	5	10	15	20		
1	1	3	—	—	4	8,75
3	2	3	7	—	12	12,08
5	—	3	9	4	16	15,31
7	—	—	5	3	8	16,87
Итого (число единиц) $f_y = f_i$	3	9	21	7	$\Sigma f = 40$	14,00

В первой строке значению факторного признака $x = 1$ один раз соответствует значение $y = 5$ и три раза $y = 10$. Аналогично во второй строке, где $x = 3$, два раза этому значению соответствует $y = 5$, три раза $y = 10$ и семь раз $y = 15$ и т.д.

В итоговой строке показано распределение всех 40 единиц по признаку y, поэтому и частоты обозначены как f_y (иногда их обозначают m_y). В итоговой графе (столбце) показано распределение тех же 40 единиц, но по признаку x – отсюда и обозначение частот f_x (или m_x). Каждая частота внутри таблицы – это f_{xy} (или m_{xy}). Если x считать факторным признаком, то для каждого j-го значения x по строке можно рассчитать среднее значение результативного признака, т.е. $\bar{y}_j$.

Так, по первой строке $\bar{y}_1 = (5 \cdot 1 + 10 \cdot 3)/4 = 8{,}75$; по второй строке $\bar{y}_2 = (5 \cdot 2 + 10 \cdot 3 + 15 \cdot 7)/12 = 12{,}08$; по третьей строке $\bar{y}_3 = (10 \cdot 3 + 15 \cdot 9 + 20 \cdot 4)/16 = 15{,}31$ и т.д. Это групповые средние результативного признака. Они приведены в последней графе табл. 7.3. Общую же среднюю для результативного показателя получим по распределению итоговой строки:

$$\bar{y} = \frac{\sum y_i f_y}{\sum f_y} = \frac{5 \cdot 3 + 10 \cdot 9 + 15 \cdot 21 + 20 \cdot 7}{40} = \frac{560}{40} = 14.$$

Как видно из таблицы, по мере увеличения значений x групповые средние значений y, т.е. $\bar{y}_j$, тоже увеличиваются от группы к группе, что позволяет сделать вывод о том, что между x и y существует корреляционная связь.

О наличии и направлении связи можно судить и по «внешнему виду» таблицы, т.е. по расположению в ней частот.

Так, если числа (частоты) расположены (разбросаны) в клетках таблицы беспорядочно, то это чаще всего свидетельствует либо об отсутствии связи между группировочными признаками, либо о их незначительной зависимости.

Если же частоты сконцентрированы ближе к одной из диагоналей и центру таблицы, образуя своего рода эллипс, то это почти всегда свидетельствует о наличии зависимости между x и y, близкой к линейной. Расположение по диагонали из верхнего левого угла в нижний правый свидетельствует о прямой линейной зависимости между показателями x и y, а из нижнего левого угла в верхний правый – об обратной.

Анализируя характер распределения частот в табл. 7.3, можно сделать вывод, что между показателями x и y существует прямая линейная зависимость.

Примером обратной зависимости может служить распределение, характеризующее зависимость между себестоимостью зерна и урожайностью зерновых по условным данным 80 хозяйств (табл. 7.4).

Таблица 7.4

Распределение 80 хозяйств по урожайности зерновых x и себестоимости 1 ц зерна y

Урожайность зерновых, ц/га x_j	Себестоимость 1 ц зерна, руб. y_i				*Итого* (число хозяйств) $f_x = f_j$	Средняя себестоимость 1 ц зерна по группам (строчкам), руб. $\bar{y}_j$
	До 130 ($y_1 = 125$)	130—140 ($y_2 = 135$)	140—150 ($y_3 = 145$)	Свыше 150 ($y_4 = 155$)		
А	1	2	3	4	5	6
До 15	—	—	—	2	2	155,0
15—17	—	1	2	3	6	148,3
17—19	—	—	7	1	8	146,2
19—21	—	8	8	—	16	140,0
21—23	2	20	12	—	34	137,9
23—25	1	8	1	—	10	135,0
Свыше 25	3	1	—	—	4	127,5
Итого (число хозяйств) $f_y = f_i$	6	38	30	6	$\Sigma f = 80$	$\bar{y} = 139{,}5$

В данной таблице каждому значению (интервалу значений) признака x соответствует ряд значений y и частоты расположены в форме эллипса, поэтому можно предположить, что распределение в таблице не случайно, что между x и y существует стохастическая связь. Однако наличие стохастической связи еще не означает наличие корреляционной связи. Последняя, напомним, проявляется только в изменении среднего значения результативного признака при изменении значений факторного признака.

В нашем примере средние значения себестоимости 1 ц зерна (см. графу 6 в табл. 7.4) снижаются от группы к группе, т.е. чем выше урожайность зерновых, тем ниже себестоимость.

Следовательно, между x и y существует обратная корреляционная зависимость.

Таким образом, наличие корреляционной связи одновременно означает наличие стохастической связи. Вместе с тем при наличии стохастической связи корреляционная связь может и отсутствовать, если групповые средние результативного признака в силу определенных причин окажутся одинаковыми.

При построении эмпирической линии регрессии по данным корреляционной таблицы в качестве x принимаются значения середины интервалов факторного признака, а в качестве y – групповые средние результативного показателя, т.е. $\bar{y}_j$. Воспользовавшись данными табл. 7.4, получим эмпирическую линию регрессии, приведенную на рис. 7.2. График наглядно иллюстрирует снижение себестоимости зерна по мере увеличения урожайности.

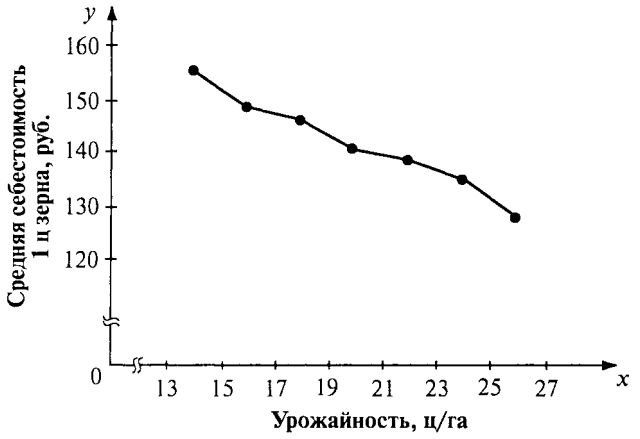


Рис. 7.2. Эмпирическая линия регрессии y по x

Когда взаимосвязанные показатели x и y представлены, как в нашем примере, в корреляционной таблице, предпочтительнее на графике по исходным данным строить «корреляционное поле», а затем на его фоне по средним значениям y – эмпирическую линию регрессии.

Корреляционное поле представляет, по существу, ту же корреляционную таблицу, в клетках которой вместо чисел (частот) проставлено соответствующее число точек (рис. 7.3).

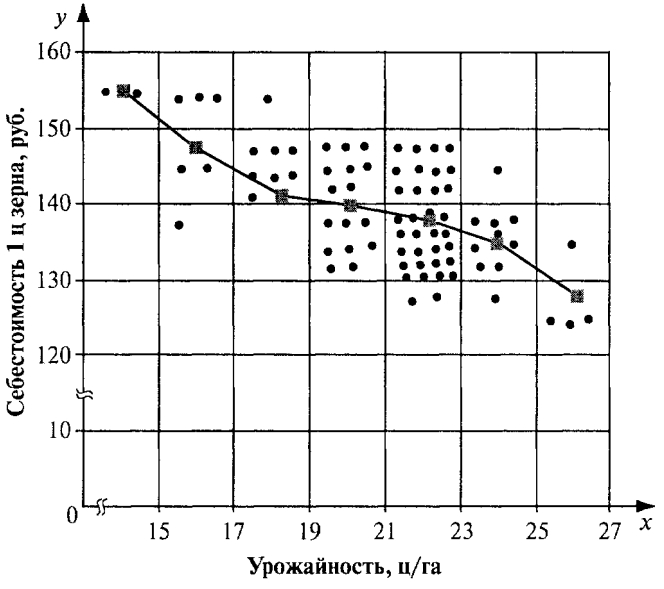


Рис. 7.3. Корреляционное поле и линия средних значений $\bar{y}_j$

Корреляционное поле отражает не только общую зависимость между x и y, но и концентрацию индивидуальных точек вокруг линии регрессии показателя $\bar{y}_j$.

На основе аналитических группировок и корреляционных таблиц можно не только выявить наличие зависимости между двумя коррелируемыми показателями, но и измерить тесноту этой связи, в частности, с помощью *эмпирического корреляционного отношения* (см. параграф 5.4)

$$\eta_{эмп} = \sqrt{\frac{\delta^2}{\sigma_y^2}}. \qquad (7.2)$$

Здесь δ^2 и σ_y^2 соответственно межгрупповая и общая дисперсии результативного признака, рассчитываемые как

$$\delta^2 = \frac{\sum_1^m (\bar{y}_j - \bar{y})^2 f_j}{\sum f_j} \quad \text{и} \quad \sigma_y^2 = \frac{\sum_1^k (y_i - \bar{y})^2 f_i}{\sum f_i},$$

где m – число групп по факторному признаку x;
k – число групп по результативному признаку y;
$\bar{y}_j$ – средние значения результативного признака по группам;
$\bar{y}$ – общее среднее значение результативного признака;
y_i – индивидуальные значения результативного признака;
$f_j = f_x$ – частота в j-й группе x;
$f_i = f_y$ – частота в i-й группе y.

Напомним, что квадрат эмпирического корреляционного отношения, т.е. $\eta^2_{\text{эмп}} = \frac{\delta^2}{\sigma_y^2}$, именуется эмпирическим коэффициентом детерминации.

В нашем примере (см. табл. 7.4)

$$\bar{y} = \frac{\sum y_i f_i}{\sum f_i} = \frac{125 \cdot 6 + 135 \cdot 38 + 145 \cdot 30 + 155 \cdot 6}{80} = 139{,}5.$$

Отсюда межгрупповая дисперсия

$$\delta^2 = \frac{\sum (\bar{y}_j - \bar{y})^2 f_j}{\sum f_j} = \frac{1}{80}\Big[(155 - 139{,}5)^2 \cdot 2 + (148{,}3 - 139{,}5)^2 \cdot 6 +$$
$$+ (146{,}2 - 139{,}5)^2 \cdot 8 + (140 - 139{,}5)^2 \cdot 16 + (137{,}9 - 139{,}5)^2 \cdot 34 +$$
$$+ (135 - 139{,}5)^2 \cdot 10 + (127{,}5 - 139{,}5)^2 \cdot 4\Big] = 27{,}78.$$

Общую дисперсию результативного признака рассчитаем по формуле

$$\sigma^2 = \sigma_y^2 = \frac{\sum (y_i - \bar{y})^2 f_i}{\sum f_i} =$$
$$= \frac{(125-139{,}5)^2 \cdot 6 + (135-139{,}5)^2 \cdot 38 + (145-139{,}5)^2 \cdot 30 + (155-139{,}5)^2 \cdot 6}{80} =$$
$$= 54{,}75.$$

Такой же результат получим по формуле

$$\sigma_y^2 = \overline{y^2} - (\bar{y})^2 = 19515 - 139{,}5^2 = 54{,}75,$$

предварительно рассчитав значения $\overline{y^2}$:

$$\overline{y^2} = \frac{\sum y_i^2 f_i}{\sum f_i} = \frac{125^2 \cdot 6 + 135^2 \cdot 38 + 145^2 \cdot 30 + 155^2 \cdot 6}{80} = 19515.$$

Отсюда эмпирический коэффициент детерминации по данным группировки, приведенной в табл. 7.4,

$$\eta^2_{\text{эмп}} = \frac{\delta^2}{\sigma_y^2} = \frac{27{,}78}{54{,}75} = 0{,}5074.$$

Извлекая квадратный корень из значения коэффициента детерминации, получаем значение эмпирического корреляционного отношения:

$$\eta_{\text{эмп}} = \sqrt{0{,}5074} = -0{,}71$$

(учитывая, что связь между x и y обратная, значение $\eta_{\text{эмп}}$ взято со знаком «–»).

Полученное значение $\eta = -0{,}71$ характеризует тесноту связи выше средней, поэтому можно сделать вывод о наличии существенной обратной связи между урожайностью и себестоимостью зерна.

7.2.3. Изучение связи между качественными признаками на основе таблиц сопряженности

Построение таблиц, в которых дается комбинационное распределение единиц совокупности по двум признакам, применимо не только к количественным, но и к неколичественным, т.е. качественным, или атрибутивным, признакам (пол, образование, семейное положение, профессия, форма собственности, вид заболеваний, вид преступлений и т.п.).

Качественные признаки, взаимосвязи между ними, их влияние на другие показатели (в том числе и количественные) особенно часто приходится изучать при проведении различных социологических исследований путем опроса или анкетирования.

В таких случаях о зависимости между теми или иными показателями (признаками) судят по комбинационному распределению единиц совокупности (респондентов) по двум изучаемым признакам. Это комбинационное распределение обычно оформляется

в виде таблиц сопряженности. Последние могут иметь разную размерность.

Простейшая форма таблицы взаимной сопряженности – таблица «четырех полей» (четырехклеточная). В ней по каждому признаку выделяется только две группы, чаще всего по альтернативному принципу («да» – «нет», «хорошо» – «плохо» и т.д.). Примером такой таблицы служит табл. 7.5, в которой приведены условные данные о распределении 500 опрошенных человек по двум показателям: наличие (отсутствие) у них прививки против гриппа и факт заболевания (незаболевания) гриппом во время его эпидемии.

Таблица 7.5

Таблица «четырех полей»

Группа лиц	Число лиц		
	заболевших гриппом	не заболевших гриппом	*Итого*
Сделавших прививку	30 (*a*)	270 (*b*)	300
Не сделавших прививку	120 (*c*)	80 (*d*)	200
Итого	150	350	500

Нетрудно заметить, что среди сделавших прививку подавляющее большинство (270 из 300, или 90%) не заболели гриппом, а среди не сделавших большая часть заболела (120 из 200, или 60%). Таким образом, можно предположить, что прививка положительно влияет на предупреждение заболевания; другими словами, можно предположить, что распределение в таблице (*a*, *b*, *c*, *d*) не случайно и существует стохастическая зависимость между группировочными признаками.

Однако выводы о зависимости, сделанные на глаз, часто могут быть ненадежными, ошибочными. Суждение о зависимости должно подкрепляться определенными статистическими критериями, например критерием Пирсона χ^2. Он позволяет судить о случайности (или неслучайности) распределения в таблицах взаимной сопряженности, а следовательно, и об отсутствии или наличии зависимости между признаками группировки в таблице. Чтобы воспользоваться критерием Пирсона χ^2, в таблице взаимной сопряженности наряду с эмпирическими частотами (или частостями) записывают теоретические (гипотетические) частоты, рассчитываемые исходя из так называемой нулевой гипотезы, т.е. предположения о том, что распределение внутри таблицы случайно и, следовательно, зависимость между признаками группировки отсутствует.

Следует иметь в виду, что при случайном распределении распределение частот в каждой строке (или графе) таблицы соответствует (пропорционально) распределению частот в итоговой строке (или графе). Поэтому теоретические частоты (частости) по строкам (или графам) рассчитывают пропорционально распределению единиц в итоговой строке (или графе).

Так, например, в табл. 7.5 в итоговой строке число заболевших гриппом составило 150 из 500, т.е. их доля – 0,3, а доля не заболевших – соответственно 0,7. Следовательно, теоретические частоты в первой строке для заболевших составят 0,3 от 300 (итог первой строки), т.е. $0{,}3 \cdot 300 = 90$, а для не заболевших 0,7 от 300, т.е. $0{,}7 \cdot 300 = 210$. Соответственно, по второй строке: для заболевших $0{,}3 \cdot 200 = 60$, а для не заболевших $0{,}7 \cdot 200 = 140$.

Перепишем табл. 7.5 еще раз в упрощенном виде с эмпирическими и теоретическими (в скобках) частотами:

Группа	I (да)	II (нет)	Σ
I (да)	30 (90)	270 (210)	300
II (нет)	120 (60)	80 (140)	200
Σ	150	350	500

На сопоставлении эмпирических и теоретических частот и основан критерий Пирсона χ^2, рассчитываемый по одной из формул:

$$\chi^2 = \sum_j \sum_i \frac{(f_{ij} - f'_{ij})^2}{f'_{ij}} \quad \text{или} \quad \chi^2 = \sum_j \sum_i \frac{f_{ij}^2}{f'_{ij}} - N, \qquad (7.3)$$

где f_{ij} и f'_{ij} – соответственно эмпирические и теоретические частоты по группам (иногда эти частоты обозначают как m_{ij} и m'_{ij}, но разная символика не меняет сути);

$N = \sum f_{ij}$ – общее число единиц совокупности.

Вторая формула (7.3) непосредственно выведена из первой.

Рассчитаем χ^2 для данных табл. 7.5 (см. таблицу выше).

По первой формуле (7.3)

$$\chi^2 = \sum_j \sum_i \frac{(f_{ij} - f'_{ij})^2}{f'_{ij}} =$$

$$= \frac{(30-90)^2}{90} + \frac{(270-210)^2}{210} + \frac{(120-60)^2}{60} + \frac{(80-140)^2}{140} = 142{,}85.$$

Такой же результат получим, рассчитав по второй формуле (7.3):

$$\chi^2 = \sum_j \sum_i \frac{f_{ij}^2}{f'_{ij}} - N = \left(\frac{30^2}{90} + \frac{270^2}{210} + \frac{120^2}{60} + \frac{80^2}{140} \right) - 500 = 142{,}85.$$

Рассчитанное (фактическое) значение χ^2 сопоставим с табличным (критическим, пороговым), определяемым по таблице Приложения 4 для заданного уровня значимости α (обычно α принимают равным 0,05 или 0,01) и числа степеней свободы $\nu = (k_1 - 1)(k_2 - 1)$, где k_1 и k_2 – число групп по одному и второму признакам группировки или, что то же самое, число строк и число граф в таблице.

В рассматриваемом примере $\nu = (2 - 1)(2 - 1) = 1$. Приняв уровень значимости $\alpha = 0{,}05$, по таблице Приложения 4 находим $\chi^2_{\text{табл}} = 3{,}84$.

Поскольку рассчитанное нами $\chi^2_{\text{факт}} > \chi^2_{\text{табл}}$, то выдвинутая нулевая гипотеза о случайном распределении отвергается, т.е. распределение не случайно, значит, существует стохастическая зависимость между такими показателями, как наличие (отсутствие) прививки и заболевание гриппом.

При независимости признаков часто́ты теоретического и эмпирического распределений совпадают, т.е. их разность $(f_{ij} - f'_{ij})$ и χ^2 равны нулю. Чем больше различия между теоретическими и эмпирическими частотами, тем больше значение χ^2 и вероятность того, что оно превысит критическое табличное значение, допустимое для случайных расхождений при принятии нулевой гипотезы.

Аналогично рассчитываются теоретические частоты и χ^2 в таблицах большей размерности, как, например, в табл. 7.6, где приведено распределение 200 опрошенных по двум признакам: сфере их деятельности и степени удовлетворенности оплатой своего труда. По каждому признаку выделено по три группы, т.е. это таблица размерности 3×3. Теоретические частоты в каждой строке рассчитаны пропорционально итоговой строке, т.е. в соотношении 0,35, 0,33 и 0,32.

Такую таблицу трудно проанализировать на глаз, хотя видно, что в бюджетных НИИ и на государственных предприятиях большинство работающих не удовлетворено оплатой своего труда, а в коммерческих структурах, наоборот, большинство довольно, т.е. распределение свидетельствует о наличии стохастической связи. Чтобы подтвердить или опровергнуть этот факт, воспользуемся критерием χ^2.

Таблица 7.6

Сфера деятельности	Численность работников, давших ответ на вопрос об удовлетворенности оплатой своего труда			
	Совсем не удовлетворен	Не совсем удовлетворен	Полностью удовлетворен	*Итого* (работников)
Бюджетные НИИ	22 (17,5)	20 (16,5)	8 (16)	50
Государственные предприятия	36 (28)	30 (26,4)	14 (25,6)	80
Коммерческие структуры	12 (24,5)	16 (23,1)	42 (22,4)	70
Итого (работников)	70	66	64	200
Доля работников	0,35	0,33	0,32	1,00

Для табл. 7.6 расчет χ^2 (с целью установить, существует ли связь между ответами 200 опрошенных человек на вопрос об удовлетворенности оплатой труда и сферой их деятельности) проводится аналогично:

$$\chi^2 = \frac{(22 - 17{,}5)^2}{17{,}5} + \frac{(20 - 16{,}5)^2}{16{,}5} + \frac{(8 - 16)^2}{16} + \frac{(36 - 28)^2}{28} +$$

$$+ \frac{(30 - 26{,}4)^2}{26{,}4} + \frac{(14 - 25{,}6)^2}{25{,}6} + \frac{(12 - 24{,}5)^2}{24{,}5} + \frac{(16 - 23{,}1)^2}{23{,}1} +$$

$$+ \frac{(42 - 22{,}4)^2}{22{,}4} = 39{,}6.$$

Число степеней свободы для табл. 7.6, где три строки и три графы, $\nu = (3 - 1)(3 - 1) = 4$. Приняв уровень значимости $\alpha = 0{,}05$, по таблице Приложения 4 для $\nu = 4$ определим $\chi^2_{табл} = 9{,}49$. Так как $\chi^2_{факт} > \chi^2_{табл}$, то, как и в предыдущем примере, это подтверждает наличие зависимости между рассмотренными показателями.

Порой для расчета χ^2 удобно пользоваться не абсолютными частотами или соответствующими им частостями, сумма которых по таблице в целом равна 1, а *частостями, вычисленными для каждой строки отдельно*, т.е. дающими в сумме единицу по каждой строке в отдельности. Частости, рассчитанные по каждой строке в отдельности, называют *условными*, а рассчитанные по итоговой строке – *безусловными*. Эти частости сопоставляют и выносят суждение о наличии или отсутствии связи между признаками группировки.

При случайном распределении (т.е. отсутствии связи) условные частости (по каждой строке) совпадают по значению с безусловными частостями (по итоговой строке). И чем больше расхождения условных частостей от безусловных, тем больше связь (зависимость) между признаками группировки.

Если обозначить частость j-й графы условного распределения по i-й строке через $w_{j/i}$, а частость этой графы в итоговой строке (безусловного распределения) через w_j, то для условного распределения по каждой строке χ_i^2 рассчитывается по формуле

$$\chi_i^2 = f_i \sum_j \frac{(w_{j/i} - w_j)^2}{w_j}, \qquad (7.4)$$

а для совокупности в целом χ^2 рассчитывается как сумма χ_i^2 по всем строкам, т.е.

$$\chi^2 = \sum_i \chi_i^2.$$

Рассмотрим этот способ расчета χ^2 на примере табл. 7.6, для чего воспроизведем ее еще раз (в виде табл. 7.7), записав в ней рассчитанные по каждой строке частости условных распределений и в итоговой строке – частости безусловного распределения.

Таблица 7.7

Сфера деятельности	Численность работников, давших ответ на вопрос об удовлетворенности оплатой своего труда (в долях к итогу по строке)			
	Совсем не удовлетворен	Не совсем удовлетворен	Полностью удовлетворен	*Итого*
Бюджетные НИИ	0,440	0,400	0,160	1,000
Государственные предприятия	0,450	0,375	0,175	1,000
Коммерческие структуры	0,170	0,230	0,600	1,000
Итого	0,350	0,330	0,320	1,000

Как видно из табл. 7.7, частости условных распределений (по строкам) не совпадают с частостями безусловного распределения (по итоговой строке), т.е. это распределение вряд ли можно считать случайным. Проверим это по значению χ^2.

Для первой строки

$$\chi_1^2 = f_1 \sum_j \frac{(w_{j/1} - w_j)^2}{w_j} =$$

$$= 50\left(\frac{(0{,}44 - 0{,}35)^2}{0{,}35} + \frac{(0{,}4 - 0{,}33)^2}{0{,}33} + \frac{(0{,}16 - 0{,}32)^2}{0{,}32}\right) = 5{,}9.$$

Для второй строки

$$\chi_2^2 = 80\left(\frac{(0{,}45 - 0{,}35)^2}{0{,}35} + \frac{(0{,}375 - 0{,}33)^2}{0{,}33} + \frac{(0{,}175 - 0{,}32)^2}{0{,}32}\right) = 8.$$

И для третьей строки

$$\chi_3^2 = 70\left(\frac{(0{,}17 - 0{,}35)^2}{0{,}35} + \frac{(0{,}23 - 0{,}33)^2}{0{,}33} + \frac{(0{,}6 - 0{,}32)^2}{0{,}32}\right) = 25{,}7.$$

В целом же для всей совокупности

$$\chi^2 = \sum_i \chi_i^2 = 5{,}9 + 8 + 25{,}7 = 39{,}6,$$

т.е. то же значение χ^2, что и ранее (по абсолютным частотам). Соответственно, выводы о характере распределения остаются теми же. Находим $\chi^2_{\text{табл}}$:

$$\chi^2_{\text{табл}} = 9{,}49 \text{ (при } \nu = 4 \text{ и } \alpha = 0{,}05).$$

Так как фактическое χ^2 (39,6) больше табличного ($\chi^2_{\text{факт}} > \chi^2_{\text{табл}}$), то гипотеза о случайном распределении в табл. 7.6 отвергается. Следовательно, с вероятностью 0,95 $(1 - \alpha)$ можно утверждать, что зависимость между рассматриваемыми признаками группировки существует.

В корреляционном анализе недостаточно лишь выявить теми или иными методами наличие связи между исследуемыми показателями. Теснота такой связи может быть различной, поэтому весьма важно ее измерить, т.е. определить меру связи в каждом конкретном случае.

В статистике для этой цели разработан ряд показателей (коэффициентов), используемых как для количественных, так и для качественных (атрибутивных) признаков.

7.3. Показатели тесноты связи между двумя качественными признаками

Для измерения тесноты связи между группировочными признаками в таблицах взаимной сопряженности могут быть использованы такие показатели, как коэффициент ассоциации, коэффициент контингенции, коэффициенты взаимной сопряженности Пирсона и Чупрова.

Первые два могут применяться лишь для «четырехклеточных» таблиц, а последние два – для таблиц любой размерности.

Применительно к таблице «четырех полей», частоты которых можно обозначить через a, b, c, d, ***коэффициент ассоциации*** выражается формулой

$$K_{ac} = \frac{ad - bc}{ad + bc}. \qquad (7.5)$$

Отметим его существенный недостаток: если в одной из четырех клеток отсутствует частота (т.е. равна 0), коэффициент ассоциации всегда будет равен по модулю 1, и тем самым преувеличена мера действительной связи.

Чтобы этого избежать, предлагается другой показатель – ***коэффициент контингенции:***

$$K_{конт} = \frac{ad - bc}{\sqrt{(a + b)(c + d)(a + c)(b + d)}}. \qquad (7.6)$$

Рассчитаем указанные коэффициенты для распределения, приведенного в табл. 7.5:

$$K_{ac} = \frac{ad - bc}{ad + bc} = \frac{30 \cdot 80 - 270 \cdot 120}{30 \cdot 80 + 270 \cdot 120} = -0{,}862;$$

$$K_{конт} = \frac{ad - bc}{\sqrt{(a + b)(c + d)(a + c)(b + d)}} = \frac{30 \cdot 80 - 270 \cdot 120}{\sqrt{300 \cdot 200 \cdot 150 \cdot 350}} = -0{,}534.$$

Коэффициент контингенции по значению всегда меньше коэффициента ассоциации.

Связь считается достаточно значительной и подтвержденной, если $|K_{ac}| > 0{,}5$ или $|K_{конт}| > 0{,}3$.

Поэтому в нашем примере оба коэффициента характеризуют достаточно большую обратную зависимость между исследуемыми признаками.

Если по каждому из двух группировочных взаимосвязанных признаков выделяется больше двух групп, то теснота связи между

качественными признаками измеряется с помощью ***коэффициентов взаимной сопряженности Пирсона*** или ***Чупрова***, рассчитываемых на основе показателя χ^2:

коэффициент Пирсона

$$K_{\text{П}} = C = \sqrt{\frac{\chi^2}{\chi^2 + n}}, \tag{7.7}$$

коэффициент Чупрова

$$K_{\text{Ч}} = T = \sqrt{\frac{\chi^2}{n\sqrt{(k_1 - 1)(k_2 - 1)}}}, \tag{7.8}$$

где n – число единиц наблюдения;

k_1 и k_2 – соответственно число строк и граф в таблице.

Оба коэффициента можно записать по-иному, если числитель и знаменатель в каждой дроби разделить на n и обозначить частное как φ^2, который и является показателем взаимной сопряженности.

Итак, если ввести обозначение $\frac{\chi^2}{n} = \varphi^2$, то коэффициент взаимной сопряженности Пирсона

$$C = \sqrt{\frac{\varphi^2}{\varphi^2 + 1}}, \tag{7.9}$$

а коэффициент взаимной сопряженности Чупрова

$$K_{\text{Ч}} = \sqrt{\frac{\varphi^2}{\sqrt{(k_1 - 1)(k_2 - 1)}}}. \tag{7.10}$$

Причем φ^2 можно рассчитать самостоятельно, без расчета χ^2:

$$\varphi^2 = \sum \frac{f_{ij}^2}{f_i f_j} - 1,$$

или

$$\varphi^2 = \sum \frac{m_{xy}^2}{m_x m_y} - 1 \text{ (при другом обозначении частот)},$$

т.е. φ^2 можно определить как сумму отношений квадратов частот каждой клетки таблицы к произведению итоговых частот по соответствующей строке и графе за минусом единицы.

Так, по данным табл. 7.5

$$\varphi^2 = \left(\frac{30^2}{300 \cdot 150} + \frac{270^2}{300 \cdot 350} + \frac{120^2}{200 \cdot 150} + \frac{80^2}{200 \cdot 350}\right) - 1 = 0{,}2857.$$

Такой же результат получим, разделив рассчитанное ранее значение $\chi^2 = 142{,}85$ (см. подпараграф 7.2.3) на $n = 500$, т.е.

$$\varphi^2 = \frac{\chi^2}{n} = \frac{142{,}5}{500} = 0{,}2857.$$

Отсюда коэффициент взаимной сопряженности Пирсона по формуле (7.9)

$$C = \sqrt{\frac{\varphi^2}{\varphi^2 + 1}} = \sqrt{\frac{0{,}2857}{1{,}2857}} = 0{,}46$$

или по формуле (7.7)

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}} = \sqrt{\frac{142{,}85}{142{,}85 + 500}} = 0{,}46.$$

Рассчитывать коэффициент Чупрова для таблицы «четырех полей» не рекомендуется, так как при числе степеней свободы $\nu = (2 - 1)(2 - 1) = 1$ он будет больше коэффициента Пирсона. Для таблиц же другой размерности (с числом строк и граф больше двух) коэффициент взаимной сопряженности Чупрова всегда меньше коэффициента Пирсона.

Рассчитаем коэффициенты взаимной сопряженности Пирсона и Чупрова и для табл. 7.6, где $n = 200$, $\chi^2 = 39{,}6$, $k_1 = 3$ и $k_2 = 3$.

Коэффициент взаимной сопряженности Пирсона

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}} = \sqrt{\frac{39{,}6}{39{,}6 + 200}} = 0{,}406$$

или, если $\varphi^2 = \frac{\chi^2}{n} = \frac{39{,}6}{200} = 0{,}198$, то

$$C = \sqrt{\frac{\varphi^2}{\varphi^2 + 1}} = \sqrt{\frac{0{,}198}{0{,}198 + 1}} = 0{,}406.$$

Коэффициент взаимной сопряженности Чупрова

$$K_{\text{ч}} = \sqrt{\frac{\chi^2}{n\sqrt{(k_1 - 1)(k_2 - 1)}}} = \sqrt{\frac{39{,}6}{200\sqrt{(3 - 1)(3 - 1)}}} = 0{,}31$$

или $K_{\text{ч}} = \sqrt{\frac{\varphi^2}{\sqrt{(k_1 - 1)(k_2 - 1)}}} = \sqrt{\frac{0{,}198}{\sqrt{(3 - 1)(3 - 1)}}} = 0{,}31.$

По полученным значениям коэффициентов взаимной сопряженности Пирсона и Чупрова можно сделать вывод о том, что связь (зависимость) между признаками, положенными в основу группировки в табл. 7.6, средняя.

Итак, на основе метода группировок теснота связи между двумя качественными показателями (признаками) может быть измерена с помощью таких коэффициентов, как коэффициент ассоциации $K_{\text{ас}}$, коэффициент контингенции $K_{\text{конт}}$, коэффициенты взаимной сопряженности Пирсона C и Чупрова $K_{\text{ч}}$. Все они рассчитываются по данным распределения n единиц совокупности в таблицах взаимной сопряженности.

7.4. Показатели тесноты связи между двумя количественными признаками

Связь между количественными признаками измеряется через их вариацию. *Измерить зависимость (связь) между двумя коррелируемыми величинами – значит определить, насколько вариация результативного признака обусловлена вариацией факторного признака.*

В качестве показателей тесноты связи между количественными признаками, кроме упоминавшегося ранее коэффициента Фехнера (см. подпараграф 7.2.1), наиболее часто используются линейный коэффициент корреляции, коэффициенты корреляции рангов, коэффициент конкордации, а также эмпирическое и теоретическое корреляционное отношение (см. подпараграф 7.2.2 и параграф 7.6).

7.4.1. Линейный коэффициент корреляции

Для измерения тесноты связи между двумя количественными признаками x и y наиболее широко используется линейный коэффициент корреляции r.

Как явствует из его названия, он применим лишь в случае линейной зависимости между признаками. Если форма связи между

x и y еще не определена, его рассчитывают с целью получить ответ на вопрос, можно ли считать зависимость линейной.

Как и коэффициент Фехнера, линейный коэффициент корреляции может быть построен на основе отклонений индивидуальных значений x и y от соответствующей средней величины. Но в отличие от K_{Φ} в линейном коэффициенте корреляции учитываются не только знаки, но и значения отклонений $(x - \bar{x})$ и $(y - \bar{y})$, выраженные для сопоставимости в единицах среднего квадратического отклонения каждого признака, т.е. как нормированные отклонения t:

$$t_x = \frac{x - \bar{x}}{\sigma_x} \quad \text{и} \quad t_y = \frac{y - \bar{y}}{\sigma_y}.$$

Линейный коэффициент корреляции представляет собой среднюю величину из произведений нормированных отклонений для x и y:

$$r = \frac{\sum\left(\frac{x - \bar{x}}{\sigma_x}\right)\left(\frac{y - \bar{y}}{\sigma_y}\right)}{n}. \tag{7.11}$$

Вынеся σ_x и σ_y за знак суммы (как постоянные величины), получим другой вид формулы линейного коэффициента корреляции:

$$r = \frac{\sum(x - \bar{x})(y - \bar{y})}{n\sigma_x\sigma_y}. \tag{7.12}$$

Числитель формулы (7.12), деленный на n, т.е.

$$\frac{\sum(x - \bar{x})(y - \bar{y})}{n} = \overline{(x - \bar{x})(y - \bar{y})},$$

представляет собой среднее произведение отклонений значений двух признаков от их средних, именуемое их ***ковариацией***. Поэтому можно сказать, что линейный коэффициент корреляции представляет собой частное от деления ковариации между x и y на произведение их средних квадратических отклонений.

Путем несложных математических преобразований можно получить и другие модификации формулы линейного коэффициента корреляции.

В частности, учитывая, что

$$\frac{\sum(x - \bar{x})(y - \bar{y})}{n} = \frac{\sum xy - \sum\bar{x}y - \sum x\bar{y} - \sum\bar{x}\bar{y}}{n} =$$

$$= \overline{xy} - \bar{x}\bar{y} - \bar{x}\bar{y} + \bar{x}\bar{y} = \overline{xy} - \bar{x}\bar{y},$$

формулу (7.12) можно привести к виду

$$r = \frac{\overline{xy} - \bar{x}\bar{y}}{\sigma_x \sigma_y}. \qquad (7.13)$$

Еще одно выражение для линейного коэффициента корреляции получим, преобразовав в формуле (7.12) знаменатель:

$$r = \frac{\sum(x - \bar{x})(y - \bar{y})}{\sqrt{\sum(x - \bar{x})^2 \sum(y - \bar{y})^2}}. \qquad (7.14)$$

Иногда линейный коэффициент корреляции удобно рассчитывать по итоговым значениям (суммам) исходных переменных:

$$r = \frac{n\sum xy - \sum x \sum y}{\sqrt{\left[n\sum x^2 - (\sum x)^2\right]\left[n\sum y^2 - (\sum y)^2\right]}} \qquad (7.15)$$

или

$$r = \frac{\sum xy - \sum x \dfrac{\sum y}{n}}{\sqrt{\left[\sum x^2 - \dfrac{(\sum x)^2}{n}\right]\left[\sum y^2 - \dfrac{(\sum y)^2}{n}\right]}}.$$

Линейный коэффициент корреляции можно рассчитать и по формуле

$$r = a_1 \frac{\sigma_x}{\sigma_y}, \qquad (7.16)$$

где a_1 – коэффициент регрессии в уравнении связи (см. параграф 7.5);

σ_x и σ_y – соответственно среднее квадратическое отклонение в ряду x и в ряду y.

Линейный коэффициент корреляции может принимать значения от -1 до $+1$, причем знак определяется в ходе решения.

Например, если $\overline{xy} > \bar{x}\bar{y}$, то r [по формуле (7.13)] будет положительным, что характеризует п р я м у ю зависимость между x и y. Если $\overline{xy} < \bar{x}\bar{y}$, то r будет со знаком «–», что означает о б р а т н у ю связь между x и y. Если $\overline{xy} = \bar{x}\bar{y}$, то r будет равен нулю, что означает о т с у т с т в и е линейной зависимости между x и y. Коэффициент корреляции, равный единице ($r = 1$), означа-

ет функциональную зависимость между x и y. Следовательно, всякое промежуточное значение r от 0 до 1 характеризует степень приближения корреляционной связи между x и y к функциональной.

Таким образом, коэффициент корреляции при линейной зависимости служит как мерой тесноты связи, так и показателем, характеризующим степень приближения корреляционной зависимости между x и y к линейной. Поэтому близость значения r к 0 в одних случаях может означать отсутствие связи между x и y, а в других свидетельствовать о том, что зависимость не линейная.

Рассмотрим расчет линейного коэффициента корреляции на примере.

Пример. Имеются данные по восьми фирмам о часовой оплате труда x и уровне текучести кадров y (табл. 7.8). Необходимо измерить тесноту связи между x и y.

Таблица 7.8

Расчетная таблица для определения линейного коэффициента корреляции

№ п/п	Часовая оплата труда, руб. x	Уровень текучести кадров, % y	x^2	xy	y^2
1	30	34	900	1020	1156
2	40	35	1600	1400	1225
3	50	33	2500	1650	1089
4	60	28	3600	1680	784
5	70	20	4900	1400	400
6	80	24	6400	1920	576
7	90	15	8100	1350	225
8	100	11	10000	1100	121
Σ	520	200	38000	11520	5576
Средняя величина	65 ($\bar{x}$)	25 ($\bar{y}$)	4750 ($\overline{x^2}$)	1440 ($\overline{xy}$)	697 ($\overline{y^2}$)

Предположив линейную зависимость между ними, воспользуемся формулой (7.13), для чего сначала рассчитаем σ_x и σ_y (расчет необходимых показателей приведен в табл. 7.8):

$$\sigma_x = \sqrt{\overline{x^2} - (\bar{x})^2} = \sqrt{4750 - (65)^2} = 22{,}9;$$

$$\sigma_y = \sqrt{\overline{y^2} - (\bar{y})^2} = \sqrt{697 - (25)^2} = 8{,}48.$$

Линейный коэффициент корреляции

$$r = \frac{\overline{xy} - \bar{x}\bar{y}}{\sigma_x \sigma_y} = \frac{1440 - 65 \cdot 25}{22{,}9 \cdot 8{,}48} = -0{,}95.$$

Аналогичный результат получим, воспользовавшись формулой (7.15):

$$r = \frac{n\sum xy - \sum x \sum y}{\sqrt{\left[n\sum x^2 - (\sum x)^2\right]\left[n\sum y^2 - (\sum y)^2\right]}} =$$

$$= \frac{8 \cdot 11520 - 520 \cdot 200}{\sqrt{\left[8 \cdot 38000 - (520)^2\right]\left[8 \cdot 5576 - (200)^2\right]}} = -0{,}95.$$

Чтобы воспользоваться формулой (7.14), по исходным данным рассчитаем отклонения от средних, их квадраты и произведения, как это показано в табл. 7.9.

Таблица 7.9

Расчетная таблица для определения линейного коэффициента корреляции

№ п/п	x	y	$x - \bar{x}$	$y - \bar{y}$	$(x - \bar{x})(y - \bar{y})$	$(x - \bar{x})^2$	$(y - \bar{y})^2$
1	30	34	−35	9	−315	1225	81
2	40	35	−25	10	−250	625	100
3	50	33	−15	8	−120	225	64
4	60	28	−5	3	−15	25	9
5	70	20	5	−5	−25	25	25
6	80	24	15	−1	−15	225	1
7	90	15	25	−10	−250	625	100
8	100	11	35	−14	−490	1225	196
Σ	520	200	0	0	−1480	4200	576

Напомним, что по данным таблицы $\bar{x} = \frac{520}{8} = 65$, $\bar{y} = \frac{200}{8} = 25$.

Отсюда

$$r = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sqrt{\sum (x - \bar{x})^2 \sum (y - \bar{y})^2}} = \frac{-1480}{\sqrt{4200 \cdot 576}} = -0{,}95.$$

Таким образом, по всем формулам получен один и тот же результат: $r = -0{,}95$, что позволяет сделать вывод о том, что между оплатой труда x и уровнем текучести кадров y существует сильная обратная связь, т.е. с увеличением оплаты труда текучесть кадров снижается.

Проверка коэффициента корреляции на значимость (существенность)

Интерпретируя значение коэффициента корреляции, следует иметь в виду, что он рассчитан для ограниченного числа наблюдений и подвержен случайным колебаниям, как и сами значения x и y, на основе которых он рассчитан, т.е., как любой выборочный показатель, он содержит случайную ошибку и не всегда однозначно отражает действительно реальную связь между изучаемыми показателями.

Для того чтобы оценить существенность (значимость) самого r и, соответственно, реальность измеряемой связи между x и y, необходимо рассчитать среднюю квадратическую ошибку коэффициента корреляции σ_r.

Оценка существенности (значимости) линейного коэффициента корреляции основана на сопоставлении значения r с его средней квадратической ошибкой:

$$\frac{|r|}{\sigma_r}.$$

Укажем **особенности расчета** этого критерия в зависимости от числа наблюдений (объема выборки) – n.

1. Если число наблюдений достаточно велико ($n > 50$) и есть основания полагать, что выборка осуществлена из нормальной совокупности, то средняя ошибка коэффициента корреляции рассчитывается по следующей приближенной формуле:

$$\sigma_r = \frac{1 - r^2}{\sqrt{n}}. \qquad (7.17)$$

Обычно при большом n, если коэффициент корреляции r превышает свою среднюю ошибку σ_r больше чем в 3 раза (т.е. $\frac{|r|}{\sigma_r} > 3$), он считается значимым (существенным), а связь – реальной.

Задавшись определенной вероятностью, можно определить доверительные пределы (границы) r. Так, при вероятности 0,95, для

которой коэффициент доверия $t = 1{,}96$ (см. Приложение 2), доверительные границы r составят

$$r \pm 1{,}96\frac{1 - r^2}{\sqrt{n}}.$$

При вероятности 0,997, для которой коэффициент доверия $t = 3$ (см. Приложение 2), доверительные границы r составят

$$r \pm 3\frac{1 - r^2}{\sqrt{n}} = r \pm 3\sigma_r.$$

Поскольку значение r не может превышать единицу, то в случае, если $r + 3\sigma_r > 1$, следует указывать только нижний предел, т.е. утверждать, что реальный r не менее чем $r - 3\sigma_r$.

2. При небольшом числе наблюдений ($n < 30$) средняя ошибка линейного коэффициента корреляции определяется как

$$\sigma_r = \frac{\sqrt{1 - r^2}}{\sqrt{n - 2}}, \qquad (7.18)$$

а значимость r проверяется на основе t-критерия Стьюдента. При этом выдвигается и проверяется нулевая гипотеза о равенстве коэффициента корреляции нулю, т.е. гипотеза об отсутствии связи между x и y в генеральной совокупности. Для этого определяется расчетное значение критерия:

$$t_{\text{расч}} = \frac{|r|}{\sigma_r} = \frac{r\sqrt{n - 2}}{\sqrt{1 - r^2}}$$

и сопоставляется с $t_{\text{табл}}$.

Если нулевая гипотеза верна, т.е. $r = 0$, то распределение t-критерия подчиняется закону Стьюдента (с заданными параметрами: уровнем значимости α, принимаемым обычно за 0,05, и числом степеней свободы $\nu = n - 2$). Поэтому в каждом конкретном случае по таблице распределения t-критерия Стьюдента (см. Приложение 9) находится критическое значение t, которое допустимо при справедливости нулевой гипотезы, и с ним сравнивается фактическое (расчетное) значение t.

Если $t_{\text{расч}} > t_{\text{табл}}$, то нулевая гипотеза отвергается и линейный коэффициент считается значимым, а связь между x и y – реальной.

Если $t_{\text{расч}} < t_{\text{табл}}$, то нулевая гипотеза не отвергается и коэффициент корреляции считается незначимым, т.е. считается, что связь между x и y отсутствует, и значение r, отличное от нуля, получено случайно.

Проверим на значимость линейный коэффициент корреляции, рассчитанный по данным табл. 7.8. Так как $n = 8$, $r = -0{,}95$, средняя ошибка коэффициента корреляции

$$\sigma_r = \frac{\sqrt{1 - r^2}}{\sqrt{n - 2}} = \frac{\sqrt{1 - 0{,}95^2}}{\sqrt{8 - 2}} = 0{,}13.$$

Отсюда

$$t_{\text{расч}} = \frac{|r|}{\sigma_r} = \frac{0{,}95}{0{,}13} = 7{,}3.$$

По таблице Приложения 9 находим $t_{\text{табл}}$ (при $\alpha = 0{,}05$ и числе степеней свободы $\nu = n - 2 = 6$):

$$t_{\text{табл}} = 2{,}4469.$$

Так как полученное $t_{\text{расч}} = 7{,}3$ больше $t_{\text{табл}} = 2{,}4469$, то нулевая гипотеза об отсутствии связи между x и y в генеральной совокупности отвергается, т.е. мы делаем вывод, что коэффициент корреляции значим и существенно отличается от нуля, подтверждая тем самым реальную связь между x и y.

7.4.2. Коэффициенты корреляции рангов

Наряду с линейным коэффициентом корреляции r для измерения тесноты связи между двумя коррелируемыми признаками часто используются менее точные, но более простые по расчету непараметрические показатели, к числу которых, кроме коэффициента Фехнера (см. подпараграф 7.2.1), относятся коэффициенты корреляции рангов (или ранговые коэффициенты корреляции) Спирмэна (ρ) и Кендэла (τ).

Оба показателя, названные именами английских ученых, предложивших эти коэффициенты, основаны на корреляции не самих значений коррелируемых признаков, а их рангов.

Ранг – это порядковый номер, присваиваемый каждому индивидуальному значению x и y (отдельно) в ранжированном ряду. Оба признака необходимо ранжировать (нумеровать) в одном и том же порядке: от меньших значений к большим и наоборот. Чаще нумерация (присвоение ранга) от 1 до n идет по возрастанию значений признака. Если встречается несколько одинаковых значений x (или y), то каждому из них присваивается ранг, равный частному от деления суммы рангов (мест в ряду), приходящихся на эти значения, на число равных значений.

Например, если после значения признака, получившего ранг 3, следуют по возрастанию два одинаковых значения (т.е. значения, занимающие 4-е и 5-е места, одинаковы), то им обоим присваивается ранг, равный 4,5, так как $(4 + 5)/2 = 4{,}5$. Если бы за рангом 3 следовали три равных значения признака, то им всем был бы присвоен одинаковый ранг 5, так как $(4 + 5 + 6)/3 = 5$.

Ранги признаков x и y обозначают символами N_x и N_y (иногда R_x и R_y). Суждение о связи между изменениями значений x и y основано на сравнении поведения рангов по двум признакам параллельно. Если у каждой пары x и y ранги совпадают, это характеризует максимально тесную п р я м у ю связь. Если же наблюдается полная противоположность рангов, т.е. в одном ряду ранги возрастают от 1 до n, а в другом – убывают от n до 1, это максимально возможная о б р а т н а я связь.

При общей идее перехода от самих значений признаков к их рангам подходы к измерению тесноты связи при корреляции рангов у Спирмэна и Кендэла несколько отличаются, что находит отражение в предложенных ими формулах.

Для расчета коэффициента Спирмэна значения признаков x и y нумеруют (отдельно) в порядке возрастания от 1 до n, т.е. им присваивают определенный ранг (N_x и N_y) – порядковый номер в ранжированном ряду. Затем для каждой пары рангов находят их разность (обозначается как $d = N_x - N_y$), и квадраты этой разности суммируют.

Коэффициент корреляции рангов Спирмэна

$$\rho = 1 - \frac{6\sum d^2}{n^3 - n}, \quad \text{или} \quad \rho = 1 - \frac{6\sum d^2}{n(n^2 - 1)}, \qquad (7.19)$$

где d – разность рангов x и y;

n – число наблюдаемых пар значений x и y.

Примечание. Формула (7.19) представляет собой не что иное, как модификацию одной из формул линейного коэффициента корреляции, где вместо x и y рассматриваются их ранги в виде ряда натуральных чисел от 1 до n, для которых, как известно, средняя величина равна $n(n + 1)/2$, а сумма квадратов отклонений чисел натурального ряда от их средней величины равна $(n^3 - n)/12$. После определенных преобразований формулы линейного коэффициента корреляции и замены значений x и y характеристиками натурального ряда получается формула (7.19).

Коэффициент корреляции рангов Спирмэна может принимать значения от 0 до ± 1.

Когда ранги двух признаков полностью совпадают, т.е. каждое значение $N_x = N_y$, то $\sum d^2 = 0$. Соответственно, $\rho = 1$, что характеризует, как уже указывалось, максимально тесную п р я м у ю связь.

Если ранги двух признаков имеют строго противоположное направление, т.е. первому рангу x соответствует n-й (последний) ранг y, второму – $(n-1)$-й ранг y и т.д., то в этом случае максимальная величина $\sum d^2$ равна $\frac{n(n^2-1)}{3}$ и, следовательно, $\frac{6\sum d^2}{n(n^2-1)}$ может иметь максимальное значение 2. Тогда по формуле Спирмэна $\rho = -1$, что характеризует полную (максимально тесную) обратную связь между изменениями значений x и y.

Если же связь между x и y отсутствует, то, очевидно, должно соблюдаться равенство $\sum d^2 = \frac{n(n^2-1)}{6}$, и тогда $\rho = 0$.

Следует иметь в виду, что, поскольку коэффициент Спирмэна учитывает разность только рангов, а не самих значений x и y, он менее точен по сравнению с линейным коэффициентом. Поэтому его крайние значения (1 или 0) нельзя безоговорочно расценивать как свидетельство функциональной связи или полного отсутствия зависимости между x и y.

Во всех других случаях, т.е. когда ρ не принимает крайних значений, он довольно близок к r. Если же учесть простоту его расчета, то становится понятным, почему многие исследователи отдают ему предпочтение, особенно на начальном этапе выявления наличия связи между изучаемыми показателями.

Рассмотрим расчет коэффициента корреляции рангов Спирмэна по данным о часовой оплате труда x и уровне текучести кадров y (см. табл. 7.8). Исходные данные и расчет необходимых показателей приведены в табл. 7.10.

Таблица 7.10

Расчетная таблица для определения коэффициента корреляции рангов Спирмэна

x	y	Ранги		Разность рангов $d = N_x - N_y$	d^2
		N_x	N_y		
30	34	1	7	–6	36
40	35	2	8	–6	36
50	33	3	6	–3	9
60	28	4	5	–1	1
70	20	5	3	2	4
80	24	6	4	2	4
90	15	7	2	5	25
100	11	8	1	7	49
$n = 8$					$\sum d^2 = 164$

Подставим в формулу (7.19) рассчитанные значения $\sum d^2 = 164$ и $n = 8$:

$$\rho = 1 - \frac{6\sum d^2}{n(n^2 - 1)} = 1 - \frac{6 \cdot 164}{8(64 - 1)} = -0{,}952.$$

Полученное значение коэффициента корреляции рангов Спирмэна ($\rho = -0{,}952$) свидетельствует о сильной обратной связи между x и y.

Формула (7.19) применима строго теоретически только тогда, когда отдельные значения x (и y), а следовательно, и их ранги не повторяются. Для случая повторяющихся (связанных) рангов есть другая, более сложная формула, скорректированная на число повторяющихся рангов. Однако опыт показывает, что результаты расчетов по скорректированной формуле для связанных рангов мало отличаются от результатов, полученных по формуле для неповторяющихся рангов. Поэтому на практике формула (7.19) успешно применяется как для неповторяющихся, так и для повторяющихся рангов.

Коэффициент корреляции рангов Кендэла τ строится несколько по-другому, хотя его расчет также начинается с ранжирования значений признаков x и y.

Ранги x (N_x) располагают строго *в порядке возрастания* и параллельно записывают соответствующее каждому N_x значение N_y.

Поскольку N_x записаны строго по возрастанию, то ставится задача определить меру соответствия последовательности N_y «правильному» следованию N_x. При этом для каждого N_y последовательно определяют число следующих за ним рангов, превышающих его значение, и число рангов, меньших по значению. Первые («правильное» следование) учитываются как баллы со знаком «+», и их сумма обозначается буквой P. Вторые («неправильное» следование) учитываются как баллы со знаком «–», и их сумма обозначается буквой Q.

Очевидно, что максимальное значение P достигается в том случае, если ранги y (N_y) совпадают с рангами x (N_x) и в каждом ряду представляют ряд натуральных чисел от 1 до n. Тогда после первой пары значений $N_x = 1$ и $N_y = 1$ число превышения данных значений рангов составит $(n - 1)$, после второй пары, где $N_x = 2$ и $N_y = 2$, соответственно $(n - 2)$ и т.д. Таким образом, если ранги x и y совпадают и число пар рангов равно n, то

$$P_{\max} = (n - 1) + (n - 2) + \ldots + 3 + 2 + 1 = \frac{n(n - 1)}{2}.$$

Если последовательность рангов y имеет обратную тенденцию по отношению к последовательности рангов x, то Q будет иметь такое же максимальное значение по модулю:

$$|Q_{\max}| = \frac{n(n-1)}{2}.$$

Если же ранги y не совпадают с рангами x, то суммируются все положительные и отрицательные баллы ($S = P + Q$); отношение данной суммы S к максимальному значению одного из слагаемых и представляет собой ***коэффициент корреляции рангов Кендэла*** τ, т.е.

$$\tau = \frac{S}{\frac{n(n-1)}{2}}, \quad \text{или} \quad \tau = \frac{2S}{n(n-1)}. \qquad (7.20)$$

Рассмотрим расчет коэффициента корреляции рангов Кендэла на примере табл. 7.11, где x и y изменяются в одном направлении.

Пример. Предположим, по 10 хозяйствам имеются данные об урожайности картофеля y (ц/га) и о количестве внесенных на 1 га удобрений x (кг). Необходимо измерить тесноту связи между изменениями x и y с помощью коэффициента корреляции рангов Кендэла. Исходные данные и необходимые расчеты приведены в табл. 7.11.

Таблица 7.11

Расчетная таблица для определения коэффициента корреляции рангов Кендэла

x	y	Ранги		Подсчет баллов	
		N_x	N_y	«+»	«–»
138	218	1	1	9	0
175	240	2	3	7	1
190	232	3	2	7	0
196	280	4	6	4	2
200	260	5	4	5	0
235	310	6	9	1	3
250	290	7	7	2	1
260	278	8	5	2	0
275	300	9	8	1	0
290	320	10	10	—	—
$n = 10$				$P = 38$	$Q = -7$

Поясним, как происходит подсчет баллов.

Поскольку ранги x, т.е. N_x, даны строго в порядке возрастания, подсчет баллов ведем, наблюдая за изменением N_y. Так, после первой пары девять значений N_y больше 1 и ни одного меньше 1. Поэтому в первой строке стоит 9 в графе со знаком «+» и 0 в графе со знаком «–».

После второй пары, где $N_y = 3$, наблюдается семь случаев, когда ранги y превышают значение 3, и один ранг ($N_y = 2$) по значению меньше 3. Соответственно во второй строке записана цифра 7 в графе со знаком «+» и 1 в графе со знаком «–» и т.д.

В итоге $P = 38$, $Q = -7$, а $S = P + Q = 38 - 7 = 31$.

Отсюда коэффициент корреляции рангов Кендэла

$$\tau = \frac{2S}{n(n-1)} = \frac{2 \cdot 31}{10 \cdot 9} = 0{,}69.$$

Полученное значение рангового коэффициента корреляции Кендэла характеризует довольно большую (выше средней) тесноту связи между изменениями x и y.

Аналогично рассчитывается τ и для случая противоположной направленности рангов x и y, как, например, в табл. 7.10. Чтобы рассчитать по данным табл. 7.10 коэффициент Кендэла, перепишем значения рангов x и y еще раз в табл. 7.12 и определим P и Q.

Таблица 7.12

Расчетная таблица

N_x	N_y	Подсчет баллов	
		«+»	«–»
1	7	1	6
2	8	0	6
3	6	0	5
4	5	0	4
5	3	1	2
6	4	0	2
7	2	0	1
8	1	—	—
$n = 8$		$P = 2$	$Q = -26$

Согласно данным табл. 7.12 при «правильном» следовании рангов x в ряду y только в двух случаях наблюдается превышение предыдущего ранга: это значение 8 после первой пары, где $N_y = 7$, и 4 после пятой пары, где $N_y = 3$. Во всех остальных случаях последующие ранги меньше рассматриваемого в каждой паре N_x

и N_y. Так, за первой парой следует шесть рангов, значение которых меньше $N_y = 7$, за второй парой также следует шесть рангов, значение которых меньше $N_y = 8$, за третьей парой – пять рангов, которые меньше $N_y = 6$, и т.д.

Таким образом, в сумме $P = 2$, $Q = -26$, а $S = P + Q = 2 - 26 = -24$. Отсюда

$$\tau = \frac{2S}{n(n-1)} = \frac{2 \cdot (-24)}{8 \cdot 7} = -0{,}857.$$

Полученное отрицательное значение коэффициента Кендэла характеризует сильную обратную связь между x и y.

Формула коэффициента корреляции рангов Кендэла (7.20) применяется для случаев, когда отдельные значения признака (как x, так и y) не повторяются и, следовательно, их ранги не объединены.

Если же встречается несколько одинаковых значений x (или y), т.е. ранги повторяются, становятся с в я з а н н ы м и, коэффициент корреляции рангов Кендэла определяется по формуле

$$\tau = \frac{S}{\sqrt{\left[\frac{n(n-1)}{2} - U_x\right]\left[\frac{n(n-1)}{2} - U_y\right]}}, \qquad (7.21)$$

где S – фактическая общая сумма баллов при оценке +1 каждой пары рангов с одинаковым порядком изменения и –1 каждой пары рангов с обратным порядком изменения;

$U_x = U_y = \frac{\sum t(t-1)}{2}$ – число баллов, корректирующих (уменьшающих) максимальную сумму баллов за счет повторений (объединений) t рангов в каждом ряду.

Отметим, что случаи следования одинаковых повторяющихся рангов (в любом ряду) оцениваются баллом 0, т.е. они не учитываются при расчете ни со знаком «+», ни со знаком «–».

Рассмотрим расчет коэффициента корреляции Кендэла для связанных рангов по следующим условным данным (табл. 7.13), где x – стоимость основных фондов (млн руб.), а y – выпуск продукции (млн руб.) у 10 предприятий одной отрасли.

Сначала определим ранги значений для признака x. Минимальному значению $x = 13$ присваивается ранг 1. Следующим за ним двум одинаковым значениям $x = 15$, занимающим 2-е и 3-е места,

Таблица 7.13

Расчетная таблица

x	y	N_x	N_y	Подсчет баллов	
				«+»	«–»
13	31	1	2	8	1
15	30	2,5	1	7	0
15	32	2,5	3	7	0
16	33	4	4,5	5	0
18	33	6	4,5	3	0
18	34	6	6	3	0
18	35	6	7,5	2	0
19	35	8	7,5	2	0
20	38	9	9,5	0	0
22	38	10	9,5	—	—
$n = 10$				$P = 37$	$Q = -1$

присваиваем каждому ранг 2,5 (так как $(2 + 3)/2 = 2{,}5$); ранг 4 присваивается значению $x = 16$. Каждому из трех одинаковых значений $x = 18$, занимающих 5, 6 и 7-е места (ранги) в ряду, присваивается ранг 6 – средняя величина из суммы их рангов, т.е. $(5 + 6 + 7)/3 = 6$. Поскольку дальше нет одинаковых значений x, то следующим трем значениям x (19, 20, 22) соответственно присваиваются ранги 8, 9 и 10.

Аналогично определены и ранги y.

Подсчет баллов со знаками «+» и «–» проводится описанным ранее методом с одной лишь оговоркой. Например, подсчитывая число «правильных» и «неправильных» следований после второй пары рангов ($N_x = 2{,}5$ и $N_y = 1$), третью пару не учитываем ни со знаком «+», ни со знаком «–», так как значение $N_x = 2{,}5$ повторяет значение N_x рассматриваемой второй пары. Так же и в других случаях. Например, рассматривая пятую пару ($N_x = 6$ и $N_y = 4{,}5$), по той же причине не учитываем шестую и седьмую пары, у которых $N_x = 6$. Рассматривая седьмую пару ($N_x = 6$ и $N_y = 7{,}5$), не учитываем восьмую пару, у которой $N_y = 7{,}5$ повторяет значение $N_y = 7{,}5$ седьмой пары, и т.д.

Подсчитав все баллы, получим $P = 37$, $Q = -1$, а $S = P + Q = 37 - 1 = 36$. Максимальная сумма баллов равна

$$\frac{n(n-1)}{2} = \frac{10 \cdot 9}{2} = 45.$$

Далее рассчитаем поправки U_x и U_y:

$$U_x = \frac{\sum t(t-1)}{2} = \frac{2(2-1)+3(3-1)}{2} = 4$$

(t – число повторяющихся (связанных) рангов в ряду x, а именно: два ранга со значением 2,5 и три ранга со значением 6);

$$U_y = \frac{\sum t(t-1)}{2} = \frac{2(2-1)+2(2-1)+2(2-1)}{2} = 3$$

(t – соответственно число связанных рангов в ряду y, а именно: два ранга со значением 4,5, два – со значением 7,5 и два – со значением 9,5).

Отсюда коэффициент корреляции рангов Кендэла для случая связанных рангов

$$\tau = \frac{S}{\sqrt{\left[\frac{n(n-1)}{2} - U_x\right]\left[\frac{n(n-1)}{2} - U_y\right]}} = \frac{36}{\sqrt{(45-4)(45-3)}} = 0{,}867.$$

Полученный результат позволяет сделать вывод о значительном соответствии последовательности рангов двух переменных, а следовательно, о большой зависимости между изменениями рассматриваемых показателей x и y.

Перечислим преимущества ранговых коэффициентов корреляции Спирмэна и Кендэла: они легко вычисляются, с их помощью можно изучать и измерять связь не только между количественными, но и между качественными (атрибутивными) признаками, ранжированными определенным образом. Кроме того, при использовании ранговых коэффициентов корреляции не требуется знать форму связи изучаемых явлений.

7.4.3. Коэффициент конкордации

Если число ранжируемых признаков (факторов) больше двух, то для измерения тесноты связи между ними можно использовать предложенный М. Кендэлом и Б. Смитом ***коэффициент конкордации*** (множественный коэффициент ранговой корреляции)

$$W = \frac{12S}{m^2(n^3 - n)}, \qquad (7.22)$$

где S – сумма квадратов отклонений суммы m рангов от их средней величины;

m – число ранжируемых признаков;

n – число ранжируемых единиц (число наблюдений).

Формула (7.22) применяется для случая, когда ранги по каждому признаку не повторяются. Если же есть связанные ранги, то коэффициент конкордации рассчитывается с учетом числа таких повторяющихся (связанных) рангов по каждому фактору:

$$W = \frac{12S}{m^2(n^3 - n) - m\sum_{1}^{m}(t^3 - t)}, \qquad (7.23)$$

где t – число одинаковых рангов по каждому признаку.

Рассмотрим расчет коэффициента конкордации для случая, когда по каждому признаку ранги не повторяются.

Пример. Предположим, по четырем опрошенным семьям получены данные о их годовом доходе, числе детей и сбережениях за год (графы А, 1–3 табл. 7.14). В этой же таблице приведены расчеты всех показателей, необходимых для определения коэффициента конкордации между x_1, x_2 и x_3. В нашем примере $m = 3$, а $n = 4$.

Таблица 7.14

Расчетная таблица для определения коэффициента конкордации

Порядковый номер семьи	Годовой доход семьи, тыс. руб. x_1	Число детей в семье x_2	Сбережения за год, тыс. руб. x_3	Ранги каждого фактора R_{ij}			Сумма рангов $\sum_{1}^{m} R_{ij}$	Квадрат суммы рангов $\left(\sum_{1}^{m} R_{ij}\right)^2$	Квадрат отклонений суммы рангов от их средней величины $\left(\sum_{1}^{m} R_{ij} - T\right)^2$
				R_{1j}	R_{2j}	R_{3j}			
А	1	2	3	4	5	6	7	8	9
1	130	2	12,5	1	2	1	4	16	12,25
2	135	1	13,1	2	1	2	5	25	6,25
3	138	3	14,2	3	3	4	10	100	6,25
4	140	4	13,6	4	4	3	11	121	12,25
Σ							30	262	$S = 37$

Обозначив через R_{ij} ранг i-го фактора у j-й единицы, ранжируем каждый из трех факторов (графы 4–6 табл. 7.14), а затем найдем сумму рангов по каждой строке и итог по графе 7. Расчет S можно выполнить двояко: по итогам граф 7 и 8 или 7 и 9 табл. 7.14.

Первый способ. Определив сумму рангов по строкам (графа 7), возведем каждую из них в квадрат и просуммируем (графа 8). В нашем примере эта сумма равна 262, т.е.

$$\sum_{1}^{n}\left(\sum_{1}^{m} R_{ij}\right)^2 = 262.$$

Затем итог графы 7 возведем в квадрат $\left(\sum_{1}^{n}\sum_{1}^{m} R_{ij}\right)^2$ и разделим на n. Это частное вычтем из итога графы 8:

$$S = \sum_{1}^{n}\left(\sum_{1}^{m} R_{ij}\right)^2 - \frac{\left(\sum_{1}^{n}\sum_{1}^{m} R_{ij}\right)^2}{n} = 262 - \frac{30^2}{4} = 37.$$

Второй способ. Разделив итог графы 7 на n, найдем среднюю величину суммы рангов. Обозначив ее через T, получим

$$T = \frac{\sum_{1}^{n}\sum_{1}^{m} R_{ij}}{n} = \frac{30}{4} = 7{,}5.$$

Затем определим S как сумму квадратов отклонений суммы рангов каждой строки от их средней величины:

$$S = \sum_{1}^{n}\left(\sum_{1}^{m} R_{ij} - T\right)^2 = (4 - 7{,}5)^2 + (5 - 7{,}5)^2 + (10 - 7{,}5)^2 + (11 - 7{,}5)^2 = 37.$$

Этот расчет приведен в графе 9. Ее итог и есть искомая сумма S.

Рассчитав S любым из двух способов и подставив его значение в формулу (7.22), найдем коэффициент конкордации:

$$W = \frac{12S}{m^2(n^3 - n)} = \frac{12 \cdot 37}{3^2(4^3 - 4)} = 0{,}82.$$

Коэффициент конкордации W может принимать значения от 0 до 1. Полученное значение $W = 0{,}82$ позволяет сделать вывод о сильной зависимости между тремя рассмотренными показателями. Однако, чтобы это утверждение не было ошибочным, W проверяется на существенность (значимость).

Существенность коэффициента конкордации оценивается критерием χ^2, рассчитываемым по формуле

$$\chi^2 = \frac{12S}{mn(n - 1)} \qquad (7.24)$$

(при отсутствии связанных рангов)

или

$$\chi^2 = \frac{12S}{mn(n - 1) - \frac{\sum_{1}^{m}(t^3 - t)}{n - 1}} \qquad (7.25)$$

(при наличии связанных рангов).

Фактическое значение χ^2 сравнивается с табличным, соответствующим принятому уровню значимости α (0,05 или 0,01) и числу степеней свободы $\nu = n - 1$.

Если $\chi^2_{\text{факт}} > \chi^2_{\text{табл}}$, то W – существен (значим).

В нашем примере, где нет связанных рангов,

$$\chi^2_{\text{факт}} = \frac{12 \cdot 37}{3 \cdot 4(4-1)} = 12{,}3;$$

$$\chi^2_{\text{табл}} = 7{,}81 \ (\alpha = 0{,}05 \text{ и } \nu = 4 - 1 = 3).$$

Так как $\chi^2_{\text{факт}} > \chi^2_{\text{табл}}$, то на уровне значимости $\alpha = 0{,}05$ можно признать W существенным.

Коэффициент конкордации особенно часто используется в экспертных оценках, например, для того, чтобы определить степень согласованности мнений экспертов о важности того или иного оцениваемого показателя или составить рейтинг отдельных единиц по какому-либо признаку.

В формуле (7.22) в этих случаях m означает число экспертов, а n – число ранжируемых единиц (или признаков).

Рассмотрим расчет коэффициента конкордации при наличии связанных рангов.

Пример. Предположим, два эксперта ($m = 2$) ранжировали четыре признака ($n = 4$), влияющие на определенный результат, по их важности (табл. 7.15).

Таблица 7.15

Экспертная оценка приоритетности признаков

Факторный признак x_i	Ранг, установленный экспертом		Сумма рангов по каждому признаку	Квадрат суммы рангов
	первым	вторым		
1	2	3	4	5
x_1	1	1,5	2,5	6,25
x_2	2,5	1,5	4,0	16,00
x_3	2,5	4	6,5	42,25
x_4	4	3	7	49,00
Σ	10	10	20	113,50

Учитывая наличие связанных рангов, для расчета коэффициента конкордации используем формулу (7.23).

В нашем примере $m = 2$, $n = 4$, следовательно,

$$S = 113{,}5 - \frac{20^2}{4} = 13{,}5.$$

Так как и у первого эксперта два связанных ранга (2,5 и 2,5) и у второго два (1,5 и 1,5), то $\sum_{1}^{2}(t^3 - t) = (2^3 - 2) + (2^3 - 2) = 12$.

Подставляя все рассчитанные значения в формулу (7.23), получаем

$$W = \frac{12S}{m^2(n^3 - n) - m\sum_{1}^{m}(t^3 - t)} = \frac{12 \cdot 13{,}5}{2^2(4^3 - 4) - 2 \cdot 12} = 0{,}75.$$

Хотя значение W довольно большое, проверим его на значимость с помощью χ^2 (для связанных рангов):

$$\chi^2_{\text{факт}} = \frac{12S}{mn(n - 1) - \frac{\sum(t^3 - t)}{n - 1}} = \frac{12 \cdot 13{,}5}{2 \cdot 4 \cdot 3 - \frac{12}{3}} = 8{,}1.$$

Для $\nu = n - 1 = 3$ и $\alpha = 0{,}05$ значение $\chi^2_{\text{табл}} = 7{,}81$ (см. таблицу Приложения 4).

Так как $\chi^2_{\text{факт}} > \chi^2_{\text{табл}}$ (8,1 > 7,81), то коэффициент конкордации W следует признать значимым.

7.5. Нахождение уравнений регрессии между двумя признаками

Измерить корреляционную связь между признаками x и y и найти форму этой связи, ее аналитическое выражение (математическую модель) – две важные, неразрывные и дополняющие друг друга задачи корреляционно-регрессионного анализа. Их можно рассматривать в разной последовательности. В настоящем учебнике сначала рассмотрены методы выявления корреляционной связи и измерения ее тесноты, а теперь перейдем к нахождению уравнений связи (регрессии).

Найти уравнение регрессии – значит по эмпирическим (фактическим) данным математически описать изменения взаимно коррелируемых величин.

Уравнение регрессии должно определить, каким будет среднее значение результативного признака y при том или ином значении факторного признака x, если остальные факторы, влияющие на y и не связанные с x, не учитывать, т.е. абстрагироваться от них. Другими словами, уравнение регрессии можно рассматривать как вероятностную гипотетическую функциональную связь средней величины результативного признака y со значениями факторного признака x.

Уравнение регрессии можно также назвать *теоретической линией регрессии.* Рассчитанные по уравнению регрессии значения результативного признака называются *теоретическими,* обычно обозначаются $\bar{y}_x$ (читается: «игрек, выравненный по x») и рассматриваются как функция от x, т.е. $\bar{y}_x = f(x)$. (Иногда для простоты записи вместо $\bar{y}_x$ пишут y' или $\hat{y}$.)

Найти в каждом конкретном случае тип функции, с помощью которой можно наиболее адекватно отразить ту или иную зависимость между признаками x и y, – одна из основных задач регрессионного анализа.

Выбор теоретической линии регрессии часто обусловлен формой эмпирической линии регрессии; теоретическая линия как бы сглаживает изломы эмпирической линии регрессии. Кроме того, необходимо учитывать природу изучаемых показателей и специфику их взаимосвязей.

Для аналитической связи между x и y могут использоваться следующие простые виды уравнений:

а) $\bar{y}_x = a_0 + a_1 x$ (прямая);

б) $\bar{y}_x = a_0 + a_1 x + a_2 x^2$ (парабола 2-го порядка);

в) $\bar{y}_x = a_0 + a_1 \frac{1}{x}$ (гипербола);

г) $\bar{y}_x = a_0 a_1^x$ (показательная функция*);

д) $\bar{y}_x = a_0 + a_1 \lg x$ (логарифмическая функция);

е) $\bar{y}_x = \frac{d}{1 + e^{a_0 + a_1 x}}$ (логистическая функция) и др.

Обычно зависимость, выражаемую уравнением прямой, называют *линейной* (или *прямолинейной*), а все остальные – *криволинейными.*

Выбрав тип функции, по эмпирическим данным определяют параметры уравнения. При этом отыскиваемые параметры должны быть такими, при которых рассчитанные по уравнению теоретические значения результативного признака $\bar{y}_x$ были бы максимально близки к эмпирическим данным (или, что то же самое, минимально от них отличались).

Существует несколько методов нахождения параметров уравнения регрессии. Наиболее часто используется *метод наименьших*

* Показательную функцию можно привести к линейному виду, перейдя к логарифмам исходных данных: $\lg \bar{y}_x = \lg a_0 + x \lg a_1$.

квадратов (МНК). Его суть заключается в следующем требовании: искомые теоретические значения результативного признака $\bar{y}_x$ должны быть такими, при которых бы обеспечивалась минимальная сумма квадратов их отклонений от эмпирических значений, т.е.

$$S = \sum(y - \bar{y}_x)^2 \rightarrow \min \qquad (7.26)$$

(минимизируются квадраты отклонений, поскольку $\sum(y - \bar{y}_x) = 0$).

Поставив данное условие, легко определить, при каких значениях a_0, a_1 и т.д. для каждой аналитической кривой эта сумма квадратов отклонений будет минимальной.

7.5.1. Парная линейная регрессия

Линейная зависимость – наиболее часто используемая форма связи между двумя коррелируемыми признаками, и выражается она, как указывалось ранее, при парной корреляции уравнением прямой:

$$\bar{y}_x = a_0 + a_1 x. \qquad (7.27)$$

Гипотеза именно о линейной зависимости между x и y выдвигается в том случае, если значения результативного и факторного признаков возрастают (или убывают) одинаково, примерно в арифметической прогрессии.

Параметры a_0 и a_1 отыскиваются по МНК следующим образом. Согласно требованию МНК при линейной зависимости в формуле (7.26) вместо $\bar{y}_x$ записываем его конкретное выражение: $a_0 + a_1 x$. Тогда

$$S = \sum(y - a_0 - a_1 x)^2 \rightarrow \min.$$

Дальнейшее решение сводится к задаче на экстремум, т.е. к определению того, при каком значении a_0 и a_1 функция двух переменных S может достигнуть минимума.

Как известно, для этого надо найти частные производные S по a_0 и a_1, приравнять их к нулю и после элементарных преобразований решить систему двух уравнений с двумя неизвестными.

В соответствии с изложенным найдем частные производные

$$\begin{cases} \dfrac{\partial S}{\partial a_0} = 2\sum(y - a_0 - a_1 x)(-1) = 0, \\ \dfrac{\partial S}{\partial a_1} = 2\sum(y - a_0 - a_1 x)(-x) = 0. \end{cases}$$

Сократив каждое уравнение на -2, раскрыв скобки и перенеся члены с x в одну сторону, а с y – в другую, получим

$$\begin{cases} na_0 + a_1 \sum x = \sum y, \\ a_0 \sum x + a_1 \sum x^2 = \sum xy. \end{cases} \tag{7.28}$$

Эта система называется *системой нормальных уравнений* МНК для линейного уравнения регрессии.

Для решения системы (7.28) по эмпирическим данным определяем число единиц наблюдения n, сумму значений факторного признака $\sum x$, сумму их квадратов $\sum x^2$, а также сумму значений результативного признака $\sum y$ и сумму произведений $\sum xy$.

Подставив все эти суммы в систему нормальных уравнений, найдем параметры искомой прямой (линейного уравнения регрессии).

При этом указанные суммы можно определить двумя способами:

- по данным о значениях x и y каждой единицы совокупности (по списку);
- по сгруппированным данным, представленным в виде корреляционной или иной таблицы.

Расчет параметров уравнения регрессии по индивидуальным данным

Рассмотрим расчет параметров уравнения регрессии между стоимостью основных фондов x и валовым выпуском продукции y по данным табл. 7.1, которые были использованы при расчете коэффициента Фехнера (см. подпараграф 7.2.1).

Исходные данные и расчет необходимых сумм показаны в табл. 7.16.

Предположим, что зависимость между показателями x и y линейная, т.е. $\bar{y}_x = a_0 + a_1 x$. Параметры a_0 и a_1 этого уравнения найдем, решив систему нормальных уравнений (7.28). Подставив в нее необходимые суммы, рассчитанные в табл. 7.16, получим

$$\begin{cases} 10a_0 + 520a_1 = 1000, \\ 520a_0 + 35624a_1 = 70244. \end{cases}$$

Решив последнюю систему уравнений, найдем, что $a_0 = -10{,}24$, $a_1 = 2{,}12$. Отсюда искомое уравнение регрессии y по x будет

$$\bar{y}_x = -10{,}24 + 2{,}12x.$$

Таблица 7.16

Расчетная таблица для определения параметров уравнения регрессии по индивидуальным данным

Основные фонды, млн руб. x	Валовой выпуск продукции, млн руб. y	x^2	xy	$\bar{y}_x = -10{,}24 + 2{,}12x$
12	28	144	336	15
16	40	256	640	24
25	38	625	950	43
38	65	1444	2470	70
43	80	1849	3440	81
55	101	3025	5555	106
60	95	3600	5700	117
80	125	6400	10000	159
91	183	8281	16653	183
100	245	10000	24500	202
$\sum x = 520$	$\sum y = 1000$	$\sum x^2 = 35624$	$\sum xy = 70244$	$\sum \bar{y}_x = 1000$

Подставляя в данное уравнение последовательно значения x (12, 16, 25 и т.д.), находим теоретические (выравненные) значения результативного признака, т.е. $\bar{y}_x$, которые показывают, каким теоретически должен быть средний объем валового выпуска продукции при данной стоимости основных фондов x_i (при прочих равных условиях для всех предприятий). Теоретические значения результативного признака $\bar{y}_x$ приведены в последней графе табл. 7.16 (с округлением до целых).

Для нахождения a_0 и a_1 при линейной зависимости могут быть предложены готовые формулы.

Так, на основе определителей 2-го порядка из системы нормальных уравнений (7.28) получим

$$a_1 = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - \sum x \sum x}. \qquad (7.29)$$

Формулу (7.29) для a_1 можно представить и по-иному. Если в системе нормальных уравнений каждое уравнение разделить на n, получим

$$\begin{cases} a_0 + a_1\bar{x} = \bar{y}, \\ a_0\bar{x} + a_1\overline{x^2} = \overline{xy}. \end{cases}$$

Отсюда

$$a_1 = \frac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - (\bar{x})^2} = \frac{\overline{xy} - \bar{x}\bar{y}}{\sigma_x^2}, \qquad (7.30)$$

$$a_0 = \bar{y} - a_1\bar{x}. \qquad (7.31)$$

В рассматриваемом примере найдем параметр a_1 по формуле (7.29):

$$a_1 = \frac{10 \cdot 70244 - 520 \cdot 1000}{10 \cdot 35624 - 520 \cdot 520} = 2{,}12.$$

Рассчитав $\bar{x} = \frac{\sum x}{n} = \frac{520}{10} = 52$ и $\bar{y} = \frac{1000}{10} = 100$, легко найти a_0:

$$a_0 = \bar{y} - a_1\bar{x} = 100 - 2{,}12 \cdot 52 = -10{,}24,$$

т.е. результат, как и следовало ожидать, тот же.

Параметр a_1, т.е. коэффициент при x, в уравнении линейной регрессии называется *коэффициентом регрессии.*

Коэффициент регрессии показывает, на сколько (в абсолютном выражении) изменяется значение результативного признака y при изменении факторного признака x на единицу.

Наряду с коэффициентом регрессии в экономическом анализе часто используется показатель *эластичности* изменения результативного признака относительно факторного.

Коэффициент эластичности $Э$ показывает, на сколько процентов изменяется в среднем результативный признак y при изменении факторного признака x на 1%. Обычно $Э$ рассчитывают как отношение прироста (в %) результативного признака к приросту (в %) факторного признака.

Более точно коэффициент эластичности определяют на основе уравнений регрессии:

$$Э = \frac{\partial \bar{y}_x}{\partial x}\frac{x}{\bar{y}_x}, \qquad (7.32)$$

где $\frac{\partial \bar{y}_x}{\partial x}$ – первая производная уравнения регрессии y по x.

Коэффициент эластичности для большинства форм связи – величина переменная, т.е. изменяется с изменением значений фактора x. Так, для линейной зависимости $\bar{y}_x = a_0 + a_1 x$

$$Э = a_1 \frac{x}{a_0 + a_1 x}. \qquad (7.33)$$

Применительно к рассмотренному уравнению регрессии, выражающему зависимость объема валового выпуска от стоимости основных фондов ($\bar{y}_x = -10{,}24 + 2{,}12x$), коэффициент эластичности

$$Э = \frac{2{,}12x}{-10{,}24 + 2{,}12x}.$$

Подставляя в данное выражение разные значения x, получаем и разные значения $Э$. Так, например, при $x = 50$ коэффициент эластичности $Э = 1{,}11$, а при $x = 80$ соответственно $Э = 1{,}09$ и т.д. Это значит, что при увеличении основных фондов x с 50 до 50,5 млн руб., т.е. на 1%, валовой выпуск y возрастет в среднем на 1,11% прежнего уровня; при увеличении x с 80 до 80,8 млн руб., т.е. на 1%, y возрастет на 1,09% и т.д.

Расчет параметров уравнения регрессии по сгруппированным данным

Когда наблюдение ведется над большим числом пар значений x и y, то, как указывалось ранее, данные удобнее располагать в виде аналитической или корреляционной таблицы, где указаны распределения по x и по y и, соответственно, их частоты f_x и f_y. При этом $\sum f_x = \sum f_y = n$ – общее число наблюдений.

При составлении и решении системы нормальных уравнений в этих случаях все суммы значений x и y, их произведений должны учитываться вместе с их весом, а именно:

$$\begin{cases} na_0 + a_1 \sum x f_x = \sum y f_y, \\ a_0 \sum x f_x + a_1 \sum x^2 f_x = \sum xy f_{xy}. \end{cases}$$

Рассмотрим расчет сумм, необходимых для решения данной системы при работе с корреляционной таблицей, на примере табл. 7.3, где приведено условное распределение 40 единиц по признакам x и y. В табл. 7.17 воспроизведены исходные условные данные и расчет необходимых сумм в дополнительных графах и строках.

Итак, $n = 40$ $(\sum f_x = \sum f_y = n)$, $\sum x f_x = 176$, $\sum x^2 f_x = 904$, $\sum y f_y = 560$, $\sum y^2 f_y = 8500$, $\sum xy f_{xy} = 2640$.

Таблица 7.17

Расчетная таблица для нахождения параметров уравнения регрессии по сгруппированным данным

x	y				Итого f_x	xf_x	x^2f_x	xyf_{xy}
	5	10	15	20				
1	1	3	—	—	4	4	4	35
3	2	3	7	—	12	36	108	435
5	—	3	9	4	16	80	400	1225
7	—	—	5	3	8	56	392	945
Итого f_y	3	9	21	7	40	176	904	2640
yf_y	15	90	315	140	$\sum yf_y = 560$			
y^2f_y	75	900	4725	2800	$\sum y^2f_y = 8500$			

По заголовкам дополнительных граф (и строк) очевиден способ расчета указанных произведений и их сумм в таблице. Возможно, следует пояснить расчет $\sum xyf_{xy}$. Эту величину можно вычислить по-разному. В табл. 7.17 сначала по каждой строке были найдены произведения, а затем их общая сумма. Так,

по первой строке $xyf_{xy} = 1 \cdot 5 \cdot 1 + 1 \cdot 10 \cdot 3 = 35$,

по второй строке $xyf_{xy} = 3 \cdot 5 \cdot 2 + 3 \cdot 10 \cdot 3 + 3 \cdot 15 \cdot 7 = 435$

и т.д.

Подставим полученные суммы в систему:

$$\begin{cases} 40a_0 + 176a_1 = 560, \\ 176a_0 + 904a_1 = 2640. \end{cases}$$

Решив систему, находим параметры: $a_0 = 8{,}06$; $a_1 = 1{,}35$. Отсюда искомое уравнение

$$\bar{y}_x = 8{,}06 + 1{,}35x.$$

По данным корреляционной таблицы легко рассчитать и линейный коэффициент корреляции, в частности по формуле

$$r = \frac{\overline{xy} - \bar{x}\,\bar{y}}{\sigma_x \sigma_y},$$

где σ_x и σ_y – соответственно среднее квадратическое отклонение в ряду x и в ряду y.

Из табл. 7.17 находим:

$$\overline{xy} = \frac{2640}{40} = 66, \quad \bar{x} = \frac{176}{40} = 4{,}4, \quad \bar{y} = \frac{560}{40} = 14,$$

$$\overline{x^2} = \frac{904}{40} = 22{,}6, \quad \overline{y^2} = \frac{8500}{40} = 212{,}5.$$

Далее определяем

$$\sigma_x = \sqrt{\overline{x^2} - (\bar{x})^2} = \sqrt{22{,}6 - 4{,}4^2} = 1{,}8,$$

$$\sigma_y = \sqrt{\overline{y^2} - (\bar{y})^2} = \sqrt{212{,}5 - 14^2} = 4{,}06.$$

Отсюда

$$r = \frac{66 - 4{,}4 \cdot 14}{1{,}8 \cdot 4{,}06} = 0{,}6,$$

т.е. между x и y связь средняя (умеренная).

Сопряженные уравнения

Часто зависимость между коррелируемыми показателями x и y такова, что каждый из них можно рассматривать в качестве и факторного, и результативного признака. Такими показателями, например, могут быть производительность и оплата труда.

Если первый показатель обозначить x, а второй – y, то уравнение регрессии можно записать и как y по x, т.е. $\bar{y}_x$, и как x по y, т.е. $\bar{x}_y$. В случае линейной зависимости это будет соответственно $\bar{y}_x = a_0 + a_1 x$ и $\bar{x}_y = a'_0 + a'_1 y$. Такие уравнения называются *сопряженными.*

При линейной зависимости между x и y коэффициенты корреляции для каждого из сопряженных уравнений можно записать соответственно как

$$r_{y/x} = a_1 \frac{\sigma_x}{\sigma_y}, \quad r_{x/y} = a'_1 \frac{\sigma_y}{\sigma_x}.$$

Значения этих коэффициентов равны, т.е. $r_{y/x} = r_{x/y}$; параметры же сопряженных уравнений, естественно, не одинаковы. Однако, зная уравнение регрессии y по x ($\bar{y}_x = a_0 + a_1 x$) и такие показатели, как $\bar{x}$, $\bar{y}$ и r, при необходимости можно легко записать сопряженное уравнение, т.е. x по y ($\bar{x}_y = a'_0 + a'_1 y$).

Как определить параметры a'_0 и a'_1? Из записанных выше формул $r_{y/x}$ и $r_{x/y}$ нетрудно заметить, что квадрат линейного коэффициента корреляции равен произведению коэффициентов регрессии сопряженных уравнений, т.е. $r^2 = a_1 a'_1$. Отсюда, зная коэффициент регрессии одного уравнения и значение коэффициента корреляции между x и y (r), легко определить коэффициент регрессии сопряженного уравнения.

Так, по уравнению $\bar{y}_x = 8{,}06 + 1{,}35x$, полученному по данным табл. 7.17, и зная, что при этом $r = 0{,}6$, можно определить $a'_1 = r^2/a_1 = 0{,}36/1{,}35 = 0{,}266$. Затем, зная, что $\bar{x} = 4{,}4$ и $\bar{y} = 14$, определяем $a'_0 = \bar{x} - a'_1 \bar{y} = 4{,}4 - 0{,}266 \cdot 14 = 0{,}676$. Отсюда искомое сопряженное уравнение регрессии x по y будет $\bar{x}_y = 0{,}676 + 0{,}266y$. Однако надо иметь в виду, что указанный метод расчета параметров сопряженного уравнения регрессии применим лишь при предположении, что оба уравнения линейные.

7.5.2. Параболическая корреляция

Эмпирическая линия регрессии, отражающая на графике зависимость между x и y, не всегда дает основание для выдвижения гипотезы о линейной зависимости. Характер ломаной линии может быть самым различным.

Например, если рассматривать зависимость урожайности какой-либо сельскохозяйственной культуры (y) от количества выпавших осадков (x), то вполне можно предположить, что с увеличением x будет возрастать и y, но может наступить момент, когда с увеличением x (избыток осадков) урожайность начнет падать. На графике такие данные дадут ломаную, близкую к параболе 2-го порядка.

Вообще при выборе вида аппроксимирующей функции, отражающей зависимость между коррелируемыми показателями x и y, там, где это возможно, необходимо руководствоваться и логическими рассуждениями (как показано выше). Если же это невозможно, то основным обоснованием выбора той или иной аналитической формулы для уравнения регрессии является форма (вид) эмпирической линии регрессии.

Если при равномерном возрастании x значения y возрастают или убывают ускоренно либо возрастают, а затем убывают, то чаще всего в этом случае зависимость между коррелируемыми величинами может быть выражена в виде параболы 2-го порядка

$$\bar{y}_x = a_0 + a_1 x + a_2 x^2.$$

Параметры данного уравнения находят по *методу наименьших квадратов*. Напомним, что его суть сводится к нахождению параметров такой математической модели (уравнения регрессии), при которой обеспечивается минимальная сумма квадратов отклонений эмпирических (фактических) значений результативного показателя (y) от теоретических ($\bar{y}_x$), рассчитанных по уравнению регрессии.

Так, если $\bar{y}_x = a_0 + a_1 x + a_2 x^2$, должно соблюдаться следующее требование:

$$S = \sum (y - a_0 - a_1 x - a_2 x^2)^2 \rightarrow \min.$$

Найдя частные производные данной функции по a_0, a_1 и a_2 и приравняв их к нулю, после несложных алгебраических преобразований получим следующую систему нормальных уравнений:

$$\begin{cases} n a_0 + a_1 \sum x + a_2 \sum x^2 = \sum y, \\ a_0 \sum x + a_1 \sum x^2 + a_2 \sum x^3 = \sum xy, \\ a_0 \sum x^2 + a_1 \sum x^3 + a_2 \sum x^4 = \sum x^2 y. \end{cases} \qquad (7.34)$$

Решив эту систему, находим параметры искомого уравнения параболы 2-го порядка.

Рассмотрим конкретный пример нахождения уравнения регрессии в форме параболы 2-го порядка, отражающего зависимость между урожайностью озимой пшеницы (y) и количеством внесенных органических удобрений (x) по данным наблюдения на пяти участках (данные условные). Исходные данные и все необходимые для решения системы суммы приведены в табл. 7.18.

Таблица 7.18

Расчетная таблица для определения параметров параболы 2-го порядка

Внесено органических удобрений, т/га x	Урожайность, ц/га y	x^2	x^3	x^4	xy	x^2y	$\bar{y}_x$
1	2	3	4	5	6	7	8
1	16	1	1	1	16	16	16,2
2	19	4	8	16	38	76	18,5
3	20	9	27	81	60	180	20,4
4	22	16	64	256	88	352	21,9
5	23	25	125	625	115	375	23,0
$n = 5$	100	55	225	979	317	1199	100

По данным табл. 7.18 записываем систему уравнений:

$$\begin{cases} 5a_0 + 15a_1 + 55a_2 = 100, \\ 15a_0 + 55a_1 + 225a_2 = 317, \\ 55a_0 + 225a_1 + 979a_2 = 1199. \end{cases}$$

Решив эту систему, получим:

$$a_0 = 13{,}41, \quad a_1 = 2{,}98, \quad a_2 = -0{,}214.$$

Отсюда искомое уравнение

$$\bar{y}_x = 13{,}41 + 2{,}98x - 0{,}214x^2.$$

Подставляя в него последовательно значения x получаем теоретические значения результативного показателя, т.е. $\bar{y}_x$ (приведены в графе 8 табл. 7.18).

Расчет параметров существенно упрощается, если вместо значений x пользоваться их отклонениями от средней величины, т.е. $(x - \bar{x})$. Тогда система уравнений примет вид

$$\begin{cases} na_0 + a_1\sum(x - \bar{x}) + a_2\sum(x - \bar{x})^2 = \sum y, \\ a_0\sum(x - \bar{x}) + a_1\sum(x - \bar{x})^2 + a_2\sum(x - \bar{x})^3 = \sum(x - \bar{x})y, \\ a_0\sum(x - \bar{x})^2 + a_1\sum(x - \bar{x})^3 + a_2\sum(x - \bar{x})^4 = \sum(x - \bar{x})^2 y. \end{cases}$$

Поскольку $\sum(x - \bar{x}) = 0$, а также $\sum(x - \bar{x})^3 = 0$, то эту систему можно упростить:

$$\begin{cases} na_0 + a_2\sum(x - \bar{x})^2 = \sum y, \\ a_1\sum(x - \bar{x})^2 = \sum(x - \bar{x})y, \\ a_0\sum(x - \bar{x})^2 + a_2\sum(x - \bar{x})^4 = \sum(x - \bar{x})^2 y. \end{cases} \qquad (7.35)$$

Откуда a_1 рассчитывается непосредственно из второго уравнения, а a_0 и a_2 – путем решения системы двух уравнений (первого и третьего) с двумя неизвестными.

Однако надо иметь в виду, что найденные таким образом параметры не являются окончательными (искомыми), так как они определены для уравнения регрессии y по $(x - \bar{x})$. Только подставив в последнее выражение значение $\bar{x}$, после несложных алгебраических преобразований получим искомое уравнение y по x.

Проиллюстрируем этот способ на том же примере, переписав исходные данные и дополнив их необходимыми расчетами в табл. 7.19.

Таблица 7.19

Расчетная таблица для определения параметров параболы 2-го порядка при замене x на $(x - \bar{x})$

x	y	$x - \bar{x}$	$(x - \bar{x})^2$	$(x - \bar{x})^4$	$y(x - \bar{x})$	$y(x - \bar{x})^2$	$\bar{y}_x$
1	16	–2	4	16	–32	64	16,2
2	19	–1	1	1	–19	19	18,5
3	20	0	0	0	0	0	20,4
4	22	1	1	1	22	22	21,9
5	23	2	4	16	46	92	23,0
$\sum x = 15$	$\sum y = 100$	0	10	34	17	197	100

Подставим полученные суммы в систему (7.35):

$$\begin{cases} 5a_0 + 10a_2 = 100, \\ 10a_1 = 17, \\ 10a_0 + 34a_2 = 197. \end{cases}$$

Отсюда $a_1 = \mathbf{1,7}$, а параметры a_0 и a_2 находим из

$$\begin{cases} 5a_0 + 10a_2 = 100, \\ 10a_0 + 34a_2 = 197, \end{cases} \quad \text{или} \quad \begin{cases} a_0 + 2a_2 = 20, \\ a_0 + 3{,}4a_2 = 19{,}7, \end{cases}$$

$$1{,}4a_2 = -0{,}3.$$

Итак, $a_2 = \mathbf{-0,214}$, $a_0 = 20 - 2(-0{,}214) = \mathbf{20,428}$.

Уравнение y по $(x - \bar{x})$ будет иметь вид

$$\bar{y}_{(x-\bar{x})} = 20{,}428 + 1{,}7(x - \bar{x}) - 0{,}214(x - \bar{x})^2.$$

Так как $\bar{x} = 3$, то после несложных алгебраических преобразований последнего уравнения получим искомое уравнение регрессии y по x:

$$\bar{y}_x = 20{,}428 + 1{,}7(x - 3) - 0{,}214(x - 3)^2 = 20{,}428 + 1{,}7x - 5{,}1 - 0{,}214x^2 + 1{,}284x - 1{,}926 = \mathbf{13{,}4 + 2{,}984x - 0{,}214x^2},$$

т.е. результат тот же (небольшие расхождения в сотых и тысячных знаках у первых двух параметров – результат округлений при расчете a_2).

Теоретические значения результативного признака $\bar{y}_x$ можно найти не применяя последнее уравнение. Проще использовать предпоследнее уравнение: подставив в него значения $(x - 3)$, т.е. отклонения от средней величины, можно сразу получить $\bar{y}_x$. Естественно, результат не изменится.

Зависимость между x и y может выражаться также уравнением параболы более высокого порядка.

7.5.3. Гиперболическая корреляция

Обратная зависимость между двумя признаками может выражаться либо уравнением прямой (т.е. линейной регрессии) с отрицательным коэффициентом регрессии, либо уравнением гиперболы

$$\bar{y}_x = a_0 + a_1 \frac{1}{x}$$

(уравнение гиперболы предпочтительнее использовать в тех случаях, когда значение результативного признака, равное нулю, ли-

шено смысла, что теоретически возможно при обратной линейной зависимости).

Согласно МНК система для нахождения параметров гиперболы a_0 и a_1 будет иметь вид

$$\begin{cases} na_0 + a_1\sum\frac{1}{x} = \sum y, \\ a_0\sum\frac{1}{x} + a_1\sum\left(\frac{1}{x}\right)^2 = \sum\frac{y}{x}. \end{cases} \tag{7.36}$$

Проиллюстрируем нахождение уравнения регрессии в форме гиперболы на конкретном примере.

Пример. В табл. 7.20 (графы 1, 2) по пяти предприятиям, выпускающим одноименную продукцию, приведены условные данные о размерах выпуска x (тыс. единиц) и себестоимости единицы продукции y (руб.). Требуется найти уравнение регрессии y по x.

Расчет всех необходимых для решения задачи показателей показан в графах 3–5 табл. 7.20.

Таблица 7.20

Расчетная таблица для определения параметров уравнения гиперболы по индивидуальным данным

x	y	$\frac{1}{x}$	$\left(\frac{1}{x}\right)^2$	$\frac{y}{x}$	$\bar{y}_x$
1	2	3	4	5	6
5	72	0,200	0,040000	14,40	72,2
8	68	0,125	0,015625	8,50	67,6
10	64	0,100	0,010000	6,40	66,0
12	66	0,083	0,006889	5,50	65,0
14	65	0,071	0,005041	4,64	64,3
Σ	335	0,579	0,077555	39,44	335,1

Предположим, что зависимость между x и y можно выразить функцией

$$\bar{y}_x = a_0 + a_1\frac{1}{x}.$$

Определим параметры данной гиперболы, используя систему нормальных уравнений (7.36). Подставив в данную систему рассчитанные в табл. 7.20 суммы $\sum y$, $\sum\frac{1}{x}$, $\sum\left(\frac{1}{x}\right)^2$, $\sum\frac{y}{x}$, получим

$$\begin{cases} 5a_0 + 0{,}579a_1 = 335, \\ 0{,}579a_0 + 0{,}077555a_1 = 39{,}44. \end{cases}$$

Решив систему, имеем $a_0 = 59{,}9$, $a_1 = 61{,}6$. Следовательно, $\bar{y}_x = 59{,}9 + 61{,}6\frac{1}{x}$.

Подставляя в данное уравнение значения $x = 5, 8, 10, 12, 14$, получаем выравненные (теоретические) показатели себестоимости единицы продукции ($\bar{y}_x$). Они приведены в последней графе табл. 7.20.

Если параметры рассчитываются по сгруппированным данным, т.е. при наличии частот, система уравнений принимает вид

$$\begin{cases} na_0 + a_1 \sum \frac{f_x}{x} = \sum y f_y, \\ a_0 \sum \frac{f_x}{x} + a_1 \sum \frac{f_x}{x^2} = \sum \frac{y f_y}{x}. \end{cases} \qquad (7.37)$$

Если исходные данные представлены в форме корреляционной таблицы, решается система уравнений (7.37).

Рассмотрим расчет параметров гиперболы по данным табл. 7.4, где приведено распределение 80 хозяйств по урожайности зерновых x и себестоимости зерна y.

Предположив зависимость между x и y в форме гиперболы, рассчитаем все необходимые для решения системы уравнений (7.37) суммы в дополнительных графах табл. 7.21.

Таблица 7.21

Расчетная таблица для нахождения параметров уравнения гиперболы

Середина интервала x	Середина интервала y				*Итого* f_x	$\frac{f_x}{x}$	$\frac{f_x}{x^2}$	$\Sigma y = yf_{xy}$	$\frac{1}{x}\Sigma y$	$\bar{y}_x$
	125	135	145	155						
14	—	—	—	2	2	0,143	0,010	310	22,143	161,1
16	—	1	2	3	6	0,375	0,023	890	55,625	152,7
18	—	—	7	1	8	0,444	0,025	1170	65,000	146,2
20	—	8	8	—	16	0,800	0,040	2240	112,000	140,9
22	2	20	12	—	34	1,545	0,070	4690	213,182	136,7
24	1	8	1	—	10	0,417	0,017	1350	56,25	133,1
26	3	1	—	—	4	0,154	0,006	510	19,615	130,0
Итого f_y	6	38	30	6	$\Sigma f = 80$	3,878	0,191	**11160**	543,815	
yf_y	750	5130	4350	930	$\Sigma yf_y = \mathbf{11160}$					

Из табл. 7.21 имеем: $n = 80$, $\sum \frac{f_x}{x} = 3{,}878$, $\sum \frac{f_x}{x^2} = 0{,}191$, $\sum y f_y =$

$= 11160$, $\sum \frac{y f_{xy}}{x} = 543{,}815$.

Подставим значения в систему уравнений (7.37):

$$\begin{cases} 80a_0 + 3{,}878a_1 = 11160, \\ 3{,}878a_0 + 0{,}191a_1 = 543{,}815. \end{cases}$$

Решив систему, получим: $a_0 = 93{,}92$ и $a_1 = 940{,}28$. Таким образом, искомое уравнение гиперболы

$$\bar{y}_x = 93{,}92 + 940{,}28\frac{1}{x}.$$

Подставляя в него последовательно значения x (14, 16, 18 и т.д.), находим теоретические значения себестоимости $\bar{y}_x$ (показаны в последней графе табл. 7.21).

Сумма всех $\bar{y}_x$, найденная с учетом соответствующих весов по группам, т.е. по f_x, должна совпасть с суммой всех фактических y:

$$\sum y f_y = \sum \bar{y}_x f_x.$$

Итак,

$$\sum y f_y = 11160,$$

$$\sum \bar{y}_x f_x = 161{,}1 \cdot 2 + 152{,}7 \cdot 6 + 146{,}2 \cdot 8 + 140{,}9 \cdot 16 + 136{,}7 \cdot 34 + 133{,}1 \cdot 10 + 130{,}0 \cdot 4 = 11161{,}2.$$

Расхождение на 1,2 – результат округлений.

Выбор той или иной формы уравнения регрессии не всегда прост и однозначен, так как зависимость между одними и теми же признаками x и y с большей или меньшей точностью можно выразить несколькими формулами (уравнениями регрессии). Предпочтение следует отдавать тому уравнению, параметры которого рассчитываются более просто и при этом имеют определенную экономическую (логическую) интерпретацию.

7.6. Теоретическое корреляционное отношение как универсальный показатель тесноты связи

Как уже отмечалось, нахождение уравнения регрессии и измерение тесноты связи между двумя (или более) показателями – две неразрывно связанные и дополняющие друг друга стороны исследования корреляционных зависимостей в статистике.

Измерить тесноту связи между коррелируемыми величинами — значит определить, насколько вариация результативного признака обусловлена вариацией факторного (факторных) признака.

Ранее были рассмотрены показатели, с помощью которых можно выявить наличие корреляционной связи между двумя признаками x и y и измерить тесноту этой связи: коэффициент Фехнера, ранговые коэффициенты корреляции Спирмэна и Кендэла, линейный коэффициент корреляции и др.

Наряду с ними существует универсальный показатель — *корреляционное отношение* (или *коэффициент корреляции по Пирсону*), применимое ко всем случаям корреляционной зависимости независимо от формы этой связи.

Следует различать эмпирическое корреляционное отношение и теоретическое.

Как уже отмечалось ранее (см. с. 210–212), *эмпирическое корреляционное отношение* рассчитывается по аналитической группировке (или корреляционной таблице) на основе правила сложения дисперсий как корень квадратный из отношения межгрупповой дисперсии результативного признака δ^2 к общей дисперсии результативного признака σ_y^2, т.е.

$$\eta_{\text{эмп}} = \sqrt{\frac{\delta^2}{\sigma_y^2}}, \quad \text{или} \quad \eta_{\text{эмп}} = \sqrt{\frac{\sum(\bar{y}_j - \bar{y})^2}{\sum(y_i - \bar{y})^2}}.$$

Теоретическое корреляционное отношение $\eta_{\text{теор}}$ определяется на основе выравненных (теоретических) значений результативного признака $\bar{y}_x$, рассчитанных по уравнению регрессии (для любой формы связи).

Теоретическое корреляционное отношение представляет собой относительную величину, получаемую в результате сравнения среднего квадратического отклонения в ряду теоретических значений результативного признака со средним квадратическим отклонением в ряду эмпирических значений (или корень квадратный из отношения дисперсий теоретического и эмпирического ряда значений результативного признака).

Так как суммы теоретических и эмпирических значений результативного признака совпадают, т.е. $\sum\bar{y}_x = \sum y$, то и среднее значение признака у этих рядов одинаково — $\bar{y}$.

Если обозначить дисперсию эмпирического ряда игреков через σ_y^2 (или D_y), а теоретического ряда — через δ^2 (или $D_{\bar{y}_x}$), то каждая из них выразится формулами

$$D_y = \sigma_y^2 = \frac{\sum(y_i - \bar{y})^2}{n}, \quad \text{а} \quad D_{\bar{y}_x} = \delta^2 = \frac{\sum(\bar{y}_x - \bar{y})^2}{n}.$$

Сравнивая вторую дисперсию с первой, получим *теоретический коэффициент детерминации*

$$\eta^2_{\text{теор}} = \frac{D_{\bar{y}_x}}{D_y} = \frac{\delta^2}{\sigma_y^2}, \quad \text{или} \quad \eta^2_{\text{теор}} = \frac{\sum(\bar{y}_x - \bar{y})^2}{\sum(y_i - \bar{y})^2}.$$

Если учесть, что σ_y^2 (или D_y) – дисперсия эмпирического ряда игреков – характеризует вариацию результативного признака за счет в с е х факторов, включая и фактор x, т.е. измеряет о б щ у ю вариацию величины y, а дисперсия теоретического ряда, т.е. δ^2 (или $D_{\bar{y}_x}$) характеризует вариацию результативного признака за счет вариации т о л ь к о фактора x (при прочих равных условиях), то отношение второй дисперсии к первой, т.е. коэффициент детерминации $\eta^2_{\text{теор}} = \frac{\delta^2}{\sigma_y^2}$, показывает, какую долю в общей дисперсии результативного признака занимает дисперсия, выражающая влияние вариации фактора x на вариацию y.

Извлекая корень квадратный из коэффициента детерминации, получаем теоретическое корреляционное отношение

$$\eta_{\text{теор}} = \sqrt{\frac{\delta^2}{\sigma_y^2}}, \quad \text{или} \quad \eta_{\text{теор}} = \sqrt{\frac{\sum(\bar{y}_x - \bar{y})^2}{\sum(y_i - \bar{y})^2}}.$$

В основе исчисления и эмпирического и теоретического корреляционного отношения лежит правило сложения дисперсий, согласно которому в первом случае (при расчете η по группировке) общая дисперсия равна сумме межгрупповой дисперсии и средней из групповых, т.е. $\sigma^2 = \delta^2 + \overline{\sigma^2}$. Во втором случае (при расчете η по уравнению регрессии) в качестве межгрупповой дисперсии выступает дисперсия теоретических значений результативного признака, т.е. $\delta^2 = D_{\bar{y}_x}$, которую можно назвать *факторной* $\delta^2_{\text{фактор}}$, поскольку она отражает влияние фактора x на вариацию y, а вместо средней из групповых дисперсий принимается *остаточная дисперсия* $\sigma^2_{\text{ост}}$, отражающая влияние на вариацию результативного признака всех остальных факторов (кроме x), не учтенных в модели (в уравнении регрессии), т.е. остаточная дисперсия отражает необъясненные расхождения между эмпирическими и

теоретическими значениями результативного признака и рассчитывается по формуле

$$\sigma^2_{\text{ост}} = \frac{\sum(y_i - \bar{y}_x)^2}{n}.$$

Таким образом, общая дисперсия эмпирического ряда y равна сумме факторной и остаточной дисперсий:

$$\sigma^2_y = \delta^2_{\text{фактор}} + \sigma^2_{\text{ост}},$$

а теоретическое корреляционное отношение

$$\eta_{\text{теор}} = \sqrt{\frac{\delta^2_{\text{фактор}}}{\sigma^2_y}}. \qquad (7.38)$$

Факторную дисперсию можно выразить как $\delta^2_{\text{фактор}} = \sigma^2_y - \sigma^2_{\text{ост}}$. Подставив это выражение в формулу (7.38), получим еще одну формулу для вычисления корреляционного отношения:

$$\eta_{\text{теор}} = \sqrt{\frac{\sigma^2_y - \sigma^2_{\text{ост}}}{\sigma^2_y}} = \sqrt{1 - \frac{\sigma^2_{\text{ост}}}{\sigma^2_y}}. \qquad (7.39)$$

В последнем виде корреляционное отношение при криволинейной форме связи обычно называют *индексом корреляции.*

Корреляционное отношение (индекс корреляции) может находиться в пределах от 0 до 1, что хорошо видно из формул (7.38) и (7.39).

Если результативный признак всецело зависит от фактора x (т.е. связь функциональная), то выравненные (теоретические) значения результативного признака $\bar{y}_x$ совпадают с эмпирическими y. Тогда $\delta^2_{\text{фактор}} = \sigma^2_y$, или $\sigma^2_{\text{ост}} = 0$, и корреляционное отношение $\eta = 1$, что означает полную зависимость вариации y от вариации x.

Если же фактор x не оказывает никакого влияния на вариацию y, то общая дисперсия σ^2_y совпадает с дисперсией остаточной $\sigma^2_{\text{ост}}$, т.е. $\sigma^2_y = \sigma^2_{\text{ост}}$, и в этом случае $\eta = 0$. Это означает, что признак y не коррелирован с фактором x.

Таким образом, чем ближе значение η к 1, тем теснее связь между вариацией y и x. И наоборот, чем ближе η к 0, тем зависимость слабее. Обычно при $\eta < 0{,}3$ говорят о малой зависимости между коррелируемыми величинами, при $0{,}3 < \eta < 0{,}6$ – о средней, при $0{,}6 < \eta < 0{,}8$ – о зависимости выше средней и при $\eta > 0{,}8$ – о большой, сильной зависимости.

Корреляционное отношение применимо как для парной, так и для множественной корреляции независимо от формы связи. В этом смысле его можно назвать универсальным показателем тесноты связи.

Покажем расчет теоретического корреляционного отношения как меры тесноты связи по данным табл. 7.18. Исходные данные и расчет дополнительных показателей, необходимых для исчисления η, приведены в табл. 7.22.

Таблица 7.22

Расчетная таблица для нахождения корреляционного отношения

Внесено органических удобрений, т/га x	Урожайность, ц/га: фактическая y	Урожайность, ц/га: рассчитанная по уравнению регрессии $\bar{y}_x$	$y-\bar{y}$	$(y-\bar{y})^2$	$\bar{y}_x-\bar{y}$	$(\bar{y}_x-\bar{y})^2$	$y-\bar{y}_x$	$(y-\bar{y}_x)^2$
1	2	3	4	5	6	7	8	9
1	16	16,2	−4	16	−3,8	14,44	−0,2	0,04
2	19	18,5	−1	1	−1,5	2,25	0,5	0,25
3	20	20,4	0	0	0,4	0,16	−0,4	0,16
4	22	21,9	2	4	1,9	3,61	0,1	0,01
5	23	23,0	3	9	3,0	9,00	0	0
$\sum x = 15$	$\sum y = 100$	$\sum \bar{y}_x = 100$	0	30	0	29,46	0	0,46

В данном примере общая средняя урожайность

$$\bar{y} = \frac{\sum y}{n} = \frac{100}{5} = 20 \text{ ц/га}.$$

Общая дисперсия (дисперсия ряда эмпирических значений результативного признака)

$$\sigma_y^2 = \frac{\sum (y - \bar{y})^2}{n} = \frac{30}{5} = 6,$$

факторная дисперсия (дисперсия ряда теоретических значений результативного признака)

$$\delta^2_{\text{фактор}} = \frac{\sum (\bar{y}_x - \bar{y})^2}{n} = \frac{29,46}{5} = 5,892.$$

Отсюда теоретическое корреляционное отношение

$$\eta_{\text{теор}} = \sqrt{\frac{\delta^2_{\text{фактор}}}{\sigma^2_y}} = \sqrt{\frac{5{,}892}{6}} = 0{,}99.$$

Данное значение $\eta_{\text{теор}} = 0{,}99$ характеризует очень тесную зависимость изменения урожайности от изменения количества внесенных удобрений.

Такой же результат получим, используя формулу индекса корреляции (7.39). Данные для остаточной дисперсии $\sigma^2_{\text{ост}}$ рассчитаны в графе 9 табл. 7.22.

Вообще в таблице должно соблюдаться следующее равенство:

$$\sum(y - \bar{y})^2 = \sum(\bar{y}_x - \bar{y})^2 + \sum(y - \bar{y}_x)^2.$$

В нашем примере незначительные расхождения ($30 \neq 29{,}46 + 0{,}46$) объясняются округлением значений параметров уравнения регрессии и самих $\bar{y}_x$.

Итак, остаточная дисперсия в нашем примере равна

$$\sigma^2_{\text{ост}} = \frac{\sum(y - \bar{y}_x)^2}{n} = \frac{0{,}46}{6} = 0{,}077.$$

Отсюда

$$\eta_{\text{теор}} = \sqrt{1 - \frac{\sigma^2_{\text{ост}}}{\sigma^2_y}} = \sqrt{1 - \frac{0{,}077}{6}} = 0{,}99.$$

Как уже отмечалось, теоретическое корреляционное отношение позволяет измерять тесноту зависимости при любой форме связи.

Нетрудно доказать, что при линейной зависимости теоретическое корреляционное отношение тождественно линейному коэффициенту корреляции, т.е. $\eta_{\text{теор}} = r$. Для этого преобразуем формулу

$$\eta_{\text{теор}} = \sqrt{\frac{\delta^2_{\text{фактор}}}{\sigma^2_y}}, \quad \text{где} \quad \delta^2_{\text{фактор}} = \sqrt{\frac{\sum(\bar{y}_x - \bar{y})^2}{n}}.$$

Учитывая, что при линейной зависимости $\bar{y}_x = a_0 + a_1 x$ и $\bar{y} = a_0 + a_1 \bar{x}$,

$$\delta^2_{\text{фактор}} = \frac{\sum(a_0 + a_1 x - a_0 - a_1 \bar{x})^2}{n} = \frac{a_1^2 \sum(x - \bar{x})^2}{n} = a_1^2 \sigma_x^2.$$

Отсюда

$$\eta_{\text{теор}} = \sqrt{\frac{a_1^2\sigma_x^2}{\sigma_y^2}} = a_1\frac{\sigma_x}{\sigma_y} = r.$$

Линейный коэффициент корреляции в виде $r = a_1\dfrac{\sigma_x}{\sigma_y}$ выступает в роли стандартизированного коэффициента регрессии, т.е. показывает, на сколько «сигм» изменится в среднем y при увеличении x на одну «сигму» (среднее квадратическое отклонение в ряду x).

Из формулы $r = a_1\dfrac{\sigma_x}{\sigma_y}$ путем преобразований и замены a_1 можно получить и другие модифицированные формулы линейного коэффициента корреляции, уже рассмотренные в параграфе 7.4.

Так, например, согласно формуле (7.30)

$$a_1 = \frac{\overline{xy} - \bar{x}\,\bar{y}}{\sigma_x^2}.$$

Тогда

$$r = a_1\frac{\sigma_x}{\sigma_y} = \frac{\overline{xy} - \bar{x}\,\bar{y}}{\sigma_x\sigma_y},$$

т.е. мы имеем формулу линейного коэффициента корреляции (7.13), которую раньше получили другим способом.

7.7. Оценка существенности коэффициента регрессии и уравнения связи

Рассчитанные для ограниченного числа наблюдений параметры уравнения регрессии не являются единственно возможными, строго однозначными, поскольку представляют собой лишь оценку реальных параметров связи в генеральной совокупности.

По этой причине в каждом конкретном случае, найдя по эмпирическим данным параметры (оценки) уравнения регрессии, определяют их среднюю ошибку μ_{a_i} и с заданной вероятностью пределы, в которых эти параметры могут находиться. Затем параметры проверяют на существенность (значимость).

Рассмотрим случай линейной зависимости, т.е. $\bar{y}_x = a_0 + a_1x$.

Расчет ошибок параметров a_0 и a_1 основан на использовании остаточной дисперсии, характеризующей расхождение (отклоне-

ние) между эмпирическими и теоретическими значениями результативного признака.

Средняя ошибка параметра a_0

$$\mu_{a_0} = \frac{\sigma_{ост}}{\sqrt{n-2}}, \tag{7.40}$$

а средняя ошибка параметра a_1, т.е. коэффициента регрессии,

$$\mu_{a_1} = \frac{\sigma_{ост}}{\sigma_x\sqrt{n-2}}, \tag{7.41}$$

где $\sigma_{ост} = \sqrt{\dfrac{\sum(y-\bar{y}_x)^2}{n}}$.

Среднюю ошибку параметров a_0 и a_1 можно записать и по-другому, на основе следующих преобразований.

Выразим остаточную дисперсию как разность между общей дисперсией результативного признака и межгрупповой (факторной):

$$\sigma^2_{ост} = \sigma^2_y - \delta^2_{фактор}.$$

Разделив обе части равенства на σ^2_y, получим

$$\frac{\sigma^2_{ост}}{\sigma^2_y} = 1 - r^2.$$

Отсюда

$$\sigma^2_{ост} = \sigma^2_y(1-r^2), \text{ или } \quad \sigma_{ост} = \sigma_y\sqrt{1-r^2}.$$

Подставив последнее выражение в формулы (7.40) и (7.41), получим

$$\mu_{a_0} = \frac{\sigma_y\sqrt{1-r^2}}{\sqrt{n-2}}, \tag{7.42}$$

$$\mu_{a_1} = \frac{\sigma_y\sqrt{1-r^2}}{\sigma_x\sqrt{n-2}}. \tag{7.43}$$

Рассчитав среднюю ошибку параметра и задавшись определенной вероятностью, а следовательно, и коэффициентом доверия t, можно определить доверительные интервалы для каждого параметра как $a_i \pm t\mu_{a_i}$.

Значимость параметра проверяется путем сопоставления его значения со средней ошибкой. Обозначим это соотношение как t:

$$t_{a_i} = \frac{a_i}{\mu_{a_i}}.$$

Тогда

$$t_{a_0} = \frac{a_0}{\mu_{a_0}} = \frac{a_0\sqrt{n-2}}{\sigma_y\sqrt{1-r^2}}, \tag{7.44}$$

$$t_{a_1} = \frac{a_1}{\mu_{a_1}} = \frac{a_1\sigma_x\sqrt{n-2}}{\sigma_y\sqrt{1-r^2}}. \tag{7.45}$$

По значению t (в зависимости от объема наблюдений) и судят о значимости параметра.

Особенно важно определять значимость параметра при x, т.е. для коэффициента регрессии (a_1), поскольку при этом определяется существенность самого фактора x, влияние его на вариацию результативного показателя (y).

При большом числе наблюдений ($n > 30$) параметр a_i считается значимым, если $t_{a_i} > 3$.

Если выборка малая, т.е. $n < 30$, фактическое (расчетное) t сопоставляется с табличным (критическим) t-критерием Стьюдента, определяемым для числа степеней свободы $\nu = n - 2$ и заданного уровня значимости α (0,05 или 0,01) по Приложению 9.

Если $t_{\text{факт}} > t_{\text{табл}}$, то параметр считается значимым.

В примере, приведенном на с. 247–249 ($\bar{y}_x = 8{,}06 + 1{,}35x$), где $n = 40$, $a_1 = 1{,}35$, $r = 0{,}6$, $\sigma_x = 1{,}8$ и $\sigma_y = 4{,}06$, средняя ошибка параметра a_0 (свободного члена) равна [см. формулу (7.42)]

$$\mu_{a_0} = \frac{\sigma_y\sqrt{1-r^2}}{\sqrt{n-2}} = \frac{4{,}06\sqrt{1-0{,}6^2}}{\sqrt{40-2}} = 0{,}522,$$

а

$$t_{a_0} = \frac{a_0}{\mu_{a_0}} = \frac{8{,}06}{0{,}522} = 15{,}4.$$

В свою очередь средняя ошибка коэффициента регрессии (a_1) будет

$$\mu_{a_1} = \frac{\sigma_y\sqrt{1-r^2}}{\sigma_x\sqrt{n-2}} = \frac{4{,}06\sqrt{1-0{,}6^2}}{1{,}8\sqrt{40-2}} = 0{,}29,$$

а

$$t_{a_1} = \frac{a_1}{\mu_{a_1}} = \frac{1{,}35}{0{,}29} = 4{,}65.$$

Так как $t_{\text{факт}} > 3$ и для a_0, и для a_1, делаем вывод о значимости параметров.

В примере, приведенном на с. 244 ($\bar{y}_x = -10{,}24 + 2{,}12x$), где $n = 10$, для проверки значимости параметра $a_1 = 2{,}12$ фактическое (расчетное) t надо сравнить с табличным.

По данным табл. 7.16 рассчитаем сначала линейный коэффициент корреляции $r = a_1 \dfrac{\sigma_x}{\sigma_y}$ и необходимые для него значения σ_x и σ_y по формулам $\sigma_x = \sqrt{\overline{x^2} - (\bar{x})^2}$ и $\sigma_y = \sqrt{\overline{y^2} - (\bar{y})^2}$. Воспользуемся суммами, полученными в табл. 7.16: $\sum x = 520$; $\sum y = 1000$; $\sum x^2 = 35624$. Недостающую сумму квадратов игреков определим дополнительно:

$$\sum y^2 = 28^2 + 40^2 + 38^2 + 65^2 + 80^2 + 101^2 + 95^2 + 125^2 + 183^2 + 245^2 = 142818.$$

Отсюда $\bar{x} = 52$, $\bar{y} = 100$, $\overline{x^2} = 3562{,}4$, $\overline{y^2} = 14281{,}8$.

Следовательно,

$$\sigma_x = \sqrt{\overline{x^2} - (\bar{x})^2} = \sqrt{3562{,}4 - 52^2} = 29{,}3,$$

$$\sigma_y = \sqrt{\overline{y^2} - (\bar{y})^2} = \sqrt{14281{,}8 - 100^2} = 65{,}4,$$

откуда

$$r = a_1 \frac{\sigma_x}{\sigma_y} = 2{,}12 \frac{29{,}3}{65{,}4} = 0{,}95.$$

Определим $t_{\text{факт}}$ для a_1:

$$t_{a_1} = \frac{a_1 \sigma_x \sqrt{n-2}}{\sigma_y \sqrt{1 - r^2}} = \frac{2{,}12 \cdot 29{,}3 \sqrt{10 - 2}}{65{,}4 \sqrt{1 - 0{,}95^2}} = 8{,}6.$$

Далее по таблице Приложения 9 для числа степеней свободы $\nu = n - 2 = 10 - 2 = 8$ и уровня значимости $\alpha = 0{,}05$ найдем $t_{\text{табл}} = 2{,}306$. Так как $t_{\text{факт}} = 8{,}6$ больше табличного, можно сделать вывод о значимости коэффициента регрессии a_1.

Наряду с проверкой значимости отдельных параметров осуществляется проверка значимости уравнения регрессии в целом или, что то же самое, проверка адекватности модели.

Эта задача решается путем расчета F-критерия Фишера и сопоставления его с табличным (критическим). F-критерий представляет собой отношение дисперсии теоретических значений результативного признака (факторной дисперсии) к остаточной дисперсии, каждая из которых рассчитана на одну степень свободы:

$$F = \frac{\delta^2_{\text{фактор}}/(m-1)}{\sigma^2_{\text{ост}}/(n-m)}, \tag{7.46}$$

где m – число параметров в уравнении регрессии;
$(m-1)$ – число степеней свободы для факторной дисперсии (теоретических значений y);
n – число наблюдений;
$(n-m)$ – число степеней свободы для остаточной дисперсии.

Часто вместо числа параметров уравнения регрессии m принимается число коэффициентов регрессии k, которое на единицу меньше m, т.е. $k = m - 1$.

В этом случае формула F-критерия записывается в виде

$$F = \frac{\delta^2_{\text{фактор}}/k}{\sigma^2_{\text{ост}}/(n-k-1)} = \frac{\delta^2_{\text{фактор}}}{\sigma^2_{\text{ост}}} \frac{n-k-1}{k}.$$

Не рассчитывая $\delta^2_{\text{фактор}}$ и $\sigma^2_{\text{ост}}$, порой удобнее пользоваться суммами квадратов соответствующих отклонений:

$$F = \frac{\sum(\bar{y}_x - \bar{y})^2}{\sum(y_i - \bar{y}_x)^2} \frac{n-k-1}{k}.$$

Если $\delta^2_{\text{фактор}}$ и $\sigma^2_{\text{ост}}$ разделить на общую дисперсию в эмпирическом ряду σ^2_y, получим соответственно

$$\frac{\delta^2_{\text{фактор}}}{\sigma^2_y} = \eta^2 \quad (\text{или } r^2 \text{ для линейной связи}),$$

$$\frac{\sigma^2_{\text{ост}}}{\sigma^2_y} = \frac{\sigma^2_y - \delta^2_{\text{фактор}}}{\sigma^2_y} = 1 - \eta^2 \quad (\text{или } 1 - r^2).$$

Тогда

$$F = \frac{r^2}{1-r^2} \frac{n-m}{m-1} \quad \text{или} \quad F = \frac{r^2}{1-r^2} \frac{n-k-1}{k}. \tag{7.47}$$

Расчетное F сопоставляется с табличным (критическим), определяемым для числа степеней свободы $\nu_1 = m - 1$ и $\nu_2 = n - m$ и заданного уровня значимости α. Если $F_{\text{расч}} > F_{\text{табл}}$, уравнение значимо.

Проверим на значимость рассмотренное ранее уравнение регрессии $\bar{y}_x = -10{,}24 + 2{,}12x$, для которого $r = 0{,}95$, $n = 10$, $m = 2$. Для данного уравнения

$$F_{\text{расч}} = \frac{0{,}95^2}{1 - 0{,}95^2} \frac{10 - 2}{2 - 1} = 74.$$

Находим табличное F (см. Приложение 8). Для $\alpha = 0{,}05$, $\nu_1 = 1$ и $\nu_2 = 8$ получаем $F_{\text{табл}} = 5{,}32$. Так как $F_{\text{расч}} > F_{\text{табл}}$, то уравнение значимо.

7.8. Множественная корреляция

При решении практических задач исследователи сталкиваются с тем, что корреляционные связи не ограничиваются связями между двумя признаками: результативным y и факторным x. В действительности результативный признак зависит от нескольких факторных. Например, инфляция тесно связана с динамикой потребительских цен, розничным товарооборотом, численностью безработных, объемами экспорта и импорта, курсом доллара, количеством денег в обращении, объемом промышленного производства и другими факторами.

В условиях действия множества факторов показатели парной корреляции оказываются условными и неточными. Количественно оценить влияние различных факторов на результат, определить форму и тесноту связи между результативным признаком y и факторными признаками x_1, x_2, ..., x_k можно методами множественной (многофакторной) корреляции.

Многофакторный корреляционно-регрессионный анализ сводится к решению следующих задач:

- обосновать взаимосвязи факторов, влияющих на исследуемый показатель;
- определить степень влияния каждого фактора на результативный признак путем построения модели-уравнения множественной регрессии, которая позволяет установить, в каком направлении и на какую величину изменится результативный показатель при изменении каждого фактора, входящего в модель;
- количественно оценить тесноту связи между результативным признаком и факторами.

Математически задача сводится к *нахождению аналитического выражения, наилучшим образом описывающего связь факторных признаков с результативным*, т.е. к отысканию функции

$$\bar{y}_{1, 2, \ldots, k} = f(x_1, x_2, \ldots, x_k).$$

Выбрать форму связи довольно сложно. Эта задача на практике основывается на априорном теоретическом анализе изучаемого явления и подборе известных типов математических моделей.

Среди многофакторных регрессионных моделей выделяют *линейные* (относительно независимых переменных) и *нелинейные*. Наиболее простыми для построения, анализа и экономической интерпретации являются многофакторные линейные модели, которые содержат независимые переменные только в первой степени:

$$\bar{y}_x = a_0 + a_1x_1 + a_2x_2 + \ldots + a_kx_k,$$

где a_0 – свободный член;

$a_1, a_2, \ldots, a_k$ – коэффициенты регрессии;

$x_1, x_2, \ldots, x_k$ – факторные признаки.

Если связь между результативным признаком и анализируемыми факторами нелинейна, то выбранная для ее описания нелинейная многофакторная модель (степенная, показательная и т.д.) может быть сведена к линейной путем линеаризации.

Параметры уравнения множественной регрессии, как и парной, рассчитываются методом наименьших квадратов, при этом решается система нормальных уравнений с $(k + 1)$ неизвестным:

$$\begin{cases} a_0 n + a_1 \sum\limits_{i=1}^{n} x_{i1} + a_2 \sum\limits_{i=1}^{n} x_{i2} + \ldots + a_k \sum\limits_{i=1}^{n} x_{ik} = \sum\limits_{i=1}^{n} y_i, \\ a_0 \sum\limits_{i=1}^{n} x_{i1} + a_1 \sum\limits_{i=1}^{n} x_{i1}^2 + a_2 \sum\limits_{i=1}^{n} x_{i1}x_{i2} + \ldots + a_k \sum\limits_{i=1}^{n} x_{i1}x_{ik} = \sum\limits_{i=1}^{n} y_i x_{i1}, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ a_0 \sum\limits_{i=1}^{n} x_{ik} + a_1 \sum\limits_{i=1}^{n} x_{i1}x_{ik} + a_2 \sum\limits_{i=1}^{n} x_{i2}x_{ik} + \ldots + a_k \sum\limits_{i=1}^{n} x_{ik}^2 = \sum\limits_{i=1}^{n} y_i x_{ik}, \end{cases}$$

где x_{ij} – значение j-го факторного признака в i-м наблюдении;

y_i – значение результативного признака в i-м наблюдении $(i = \overline{1, n})$.

Как правило, прежде чем найти параметры уравнения множественной регрессии, определяют и анализируют парные коэффициенты корреляции. При этом систему нормальных уравнений можно видоизменить таким образом, чтобы при вычислении параметров регрессии использовать уже найденные парные коэф-

фициенты корреляции. Для этого в уравнении регрессии заменим переменные y, x_1, x_2, ..., x_k переменными t_j, полученными следующим образом:

$$t_{iy} = \frac{y_i - \bar{y}}{\sigma_y}, \quad t_{ij} = \frac{x_{ij} - \bar{x}_j}{\sigma_{x_j}} \quad (i = \overline{1, n}, \quad j = \overline{1, k}).$$

Эта процедура называется *стандартизацией переменных*. В результате осуществляется переход от натурального масштаба переменных x_{ij} к центрированным и нормированным отклонениям t_{ij}. В стандартизированном масштабе среднее значение признака равно 0, а среднее квадратическое отклонение равно 1, т.е. $\bar{t}_j = 0$, $\sigma_{t_j} = 1$.

При переходе к стандартизированному масштабу переменных уравнение множественной регрессии принимает вид

$$t_y = \beta_1 t_1 + \beta_2 t_2 + \ldots + \beta_k t_k,$$

где β_j ($j = \overline{1, k}$) – коэффициенты регрессии.

В этом уравнении β-коэффициенты представляют собой стандартизированные коэффициенты множественной корреляции. Смысл их легко понять, если в уравнении регрессии вместо каждого t_j, кроме какого-либо одного, подставить его среднее значение ($\bar{t}_j = 0$). Тогда соответствующий β-коэффициент будет характеризовать изменение исследуемого показателя в зависимости от изменения одного фактора при постоянном уровне остальных. Иными словами, β-коэффициент показывает, на какую часть сигмы (σ_y) изменилось бы значение результата, если бы соответствующий j-й фактор изменился на сигму (σ_{x_j}), а прочие факторы не изменились.

Кроме того, β-коэффициенты позволяют оценить степень воздействия факторных признаков на результат. В силу того что все β-коэффициенты выражены в одинаковых единицах измерения, при $\beta_2 > \beta_3$ фактор x_2 сильнее влияет на результативный признак, чем фактор x_3.

Свободный член в стандартизированном уравнении отсутствует, так как

$$a_0 = \bar{y} - a_1\bar{x}_1 - a_2\bar{x}_2 - \ldots - a_k\bar{x}_k.$$

Параметры уравнения множественной регрессии в натуральном масштабе и уравнения регрессии в стандартизированном виде взаимосвязаны:

$$a_j = \frac{\sigma_y}{\sigma_{x_j}}\beta_j \quad (j = \overline{1, k}).$$

Нетрудно заметить, что это обычная формула коэффициента регрессии, выраженного через линейный коэффициент корреляции.

Стандартизированные коэффициенты множественной регрессии β_j также вычисляют методом наименьших квадратов, который приводит к системе нормальных уравнений

$$\begin{cases} r_{y1} = \beta_1 + r_{12}\beta_2 + \ldots + r_{1k}\beta_k, \\ r_{y2} = r_{21}\beta_1 + \beta_2 + \ldots + r_{2k}\beta_k, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ r_{yk} = r_{k1}\beta_1 + r_{k2}\beta_2 + \ldots + \beta_k, \end{cases}$$

где $r_{yj} = \dfrac{1}{n}\sum_{i=1}^{n} t_{iy}t_{ij}$ – парный коэффициент корреляции результативного признака y с j-м факторным;

$r_{jl} = \dfrac{1}{n}\sum_{i=1}^{n} t_{ij}t_{il}$ – парный коэффициент корреляции j-го факторного признака с l-м факторным.

После того как получено уравнение множественной регрессии (в стандартизированном или натуральном масштабе), необходимо *измерить тесноту связи между результативным признаком и факторными признаками.* Для измерения степени совокупного влияния отобранных факторов на результативный признак рассчитывают совокупный коэффициент детерминации R^2 и совокупный коэффициент множественной корреляции R – общие показатели тесноты связи многих признаков независимо от формы связи.

Совокупный коэффициент детерминации R^2 характеризует долю вариации результативного признака, обусловленную изменением всех факторов, входящих в уравнение множественной регрессии.

Приведем несколько формул для расчета совокупного коэффициента детерминации.

1. Аналогично индексу парной корреляции используется соотношение факторной и общей дисперсий (или остаточной и общей дисперсий)

$$R^2_{y, x_1, x_2, \ldots, x_k} = \frac{\delta^2_{\text{фактор}}}{\sigma^2_y}, \quad \text{или} \quad R^2_{y, x_1, x_2, \ldots, x_k} = 1 - \frac{\sigma^2_{\text{ост}}}{\sigma^2_y},$$

где $\delta^2_{\text{фактор}} = \dfrac{1}{n}\sum_{i=1}^{n}((\bar{y}_x)_i - \bar{y})^2$ – факторная дисперсия, характеризующая вариацию результативного признака, обусловленную вариацией включенных в анализ факторов;

σ_y^2 – общая дисперсия результативного признака;

$\sigma_{ост}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - (\bar{y}_x)_i)^2 = \sigma_y^2 - \delta_{фактор}^2$ – остаточная дисперсия, характеризующая отклонения фактических уровней результативного признака y_i от рассчитанных по уравнению множественной регрессии $(\bar{y}_x)_i$.

2. При линейной форме связи расчет совокупного коэффициента детерминации можно выполнить, используя парные коэффициенты корреляции:

$$R^2_{y, x_1, x_2, \ldots, x_k} = \frac{a_1 r_{y1}\sigma_{x_1} + a_2 r_{y2}\sigma_{x_2} + \ldots + a_k r_{yk}\sigma_{x_k}}{\sigma_y},$$

где a_1, a_2, ..., a_k – параметры уравнения множественной регрессии в натуральном масштабе.

3. Легко вычислить совокупный коэффициент детерминации, используя уравнение регрессии в стандартизированном виде:

$$R^2_{y, x_1, x_2, \ldots, x_k} = \beta_1 r_{y1} + \beta_2 r_{y2} + \ldots + \beta_k r_{yk}.$$

Совокупный коэффициент множественной корреляции R представляет собой квадратный корень из совокупного коэффициента детерминации R^2. Пределы изменения совокупного коэффициента множественной корреляции: $0 \leqslant R \leqslant 1$. Чем ближе R к 1, тем точнее уравнение множественной линейной регрессии отражает реальную связь. Иначе говоря, среди отобранных факторов присутствуют те, которые решающим образом влияют на результативный. Малое значение R можно объяснить либо тем, что в уравнение множественной регрессии не включены существенно влияющие на результат факторы, либо тем, что установленная линейная форма зависимости не отражает реальной взаимосвязи признаков. Добиться адекватности модели множественной регрессии эмпирическим данным возможно, соответственно, либо включением в уравнение регрессии дополнительных, ранее не учитываемых факторов, либо построением нелинейной модели множественной регрессии.

Совокупный коэффициент множественной корреляции зависит не только от корреляции результативного признака с факторными, но и от корреляции факторных признаков между собой. Наличие между двумя факторами весьма тесной линейной связи (парный коэффициент корреляции r_{jl} превышает по абсолютной

величине 0,8) называется ***коллинеарностью***, а между несколькими факторами – ***мультиколлинеарностью***.

Поясним эту зависимость для случая двух факторных признаков. Если при двухфакторном анализе в формуле расчета совокупного коэффициента корреляции выразить β-коэффициенты через парные коэффициенты корреляции, получим следующее выражение:

$$R_{y, x_1, x_2} = \sqrt{\frac{r_{y1}^2 + r_{y2}^2 - 2r_{y1}r_{y2}r_{12}}{1 - r_{12}^2}}. \qquad (7.48)$$

Из формулы (7.48) следует, что чем теснее связь результата с каждым фактором в отдельности, тем выше совокупный множественный коэффициент корреляции R_{y, x_1, x_2} при условии, что составляющая этой формулы, содержащая r_{12} – парный коэффициент корреляции, мала, т.е. корреляция между факторами отсутствует.

Вообще говоря, отсутствие корреляционной связи между факторными признаками и наличие тесной связи между результативным и факторными признаками – условие включения этих факторных признаков в регрессионную модель.

Допустим, мы получили зависимость результативного признака только от фактора x_1: уравнение регрессии $\bar{y}_x = a_0 + a_1x_1$. Включая дополнительный факторный признак x_2, мы полнее объясняем результат, т.е. уравнение регрессии $\bar{y}_x = a_0 + a_1x_1 + a_2x_2$ точнее описывает функцию. Однако если новая переменная x_2 коррелирует с x_1, то чем теснее эта корреляция, тем более вероятно, что включение дополнительного фактора в уравнение регрессии существенно не увеличит совокупный коэффициент множественной корреляции. Иначе говоря, чем выше межфакторная корреляция, тем ближе совокупный коэффициент множественной корреляции R_{y, x_1, x_2} по значению к максимальному из парных коэффициентов корреляции. В этом случае включение в уравнение регрессии дополнительного фактора нецелесообразно. Если в модель включены мультиколлинеарные факторы, то уравнение регрессии неадекватно отражает реальные экономические взаимосвязи. В этом случае экономическая интерпретация коэффициентов регрессии и корреляции затруднена.

Проблема отбора факторных признаков и проблема мультиколлинеарности могут быть решены на основе многомерных статистических методов анализа (например, с помощью пошаговой регрессии).

Наряду с измерением совместного влияния отобранных факторов на результативный признак важно определить воздействие каждого фактора при элиминировании его взаимосвязи с остальными (что возможно, когда последние зафиксированы на постоянном уровне). Для решения данной задачи применяют ***частные коэффициенты детерминации***

$$R^2_{y, x_k (x_1, x_2, \ldots, x_{k-1})} = \frac{R^2_{y, x_1, x_2, \ldots, x_k} - R^2_{y, x_1, x_2, \ldots, x_{k-1}}}{1 - R^2_{y, x_1, x_2, \ldots, x_{k-1}}}, \qquad (7.49)$$

где $R^2_{y, x_k (x_1, x_2, \ldots, x_{k-1})}$ – частный коэффициент детерминации, характеризующий воздействие k-го фактора при элиминировании его взаимосвязи с остальными факторами;

$R^2_{y, x_1, x_2, \ldots, x_k}$ – коэффициент множественной детерминации, отражающий влияние всех включенных в анализ факторов;

$R^2_{y, x_1, x_2, \ldots, x_{k-1}}$ – коэффициент множественной детерминации, отражающий влияние всех факторов, кроме одного, воздействие которого отражает частный коэффициент детерминации.

Частный коэффициент детерминации характеризует долю вариации результативного признака, обусловленную воздействием данного фактора, при элиминировании его взаимосвязи с остальными факторами, включенными в анализ. В зависимости от количества факторов, влияние которых исключается, частные коэффициенты детерминации могут быть первого порядка (при исключении влияния одного фактора), второго порядка (при исключении влияния двух факторов) и т.д.

Частный коэффициент корреляции есть квадратный корень из частного коэффициента детерминации.

Для того чтобы оценить сравнительную силу влияния факторов, по каждому фактору рассчитывают ***частные коэффициенты эластичности***

$$Э_j = \frac{\Delta x_j}{\bar{x}_j} : \frac{\Delta y}{\bar{y}} = a_j \frac{\bar{x}_j}{\bar{y}},$$

где $\bar{x}_j$ – среднее значение j-го факторного признака;

$\bar{y}$ – среднее значение результативного признака;

a_j – коэффициент регрессии при j-м факторном признаке.

Расчет коэффициента эластичности дополняет экономический анализ. Данный коэффициент показывает, на сколько процентов следует ожидать изменения результативного показателя при изменении фактора на 1% и неизменном значении других факторов.

Сумма частных коэффициентов эластичности $\sum_{j=1}^{k} Э_j$ позволяет оценить эластичность в целом при совокупном изменении факторов.

Рассмотрим методику корреляционно-регрессионного анализа на примере статистической обработки данных по предприятиям электросвязи (табл. 7.23).

Таблица 7.23

Основные производственные показатели предприятий электросвязи

Номер предприятия	Чистая прибыль, тыс. руб. y	Численность обслуживаемого населения, млн чел. x_1	Рентабельность, % x_2
1	197	4,9	20
2	254	5,1	22
3	112	6,5	10
4	145	3,7	21
5	176	4,0	25
6	76	2,5	19

В качестве результативного признака возьмем чистую прибыль y. Основные факторы, влияющие на ее формирование: численность населения, обслуживаемого предприятием электросвязи x_1, и рентабельность x_2. Линейная форма зависимости между признаками постулируется, и, следовательно, задача сводится к отысканию параметров уравнения

$$\bar{y}_x = a_0 + a_1 x_1 + a_2 x_2.$$

При линейной форме связи множественный корреляционно-регрессионный анализ проводится на основе информации о средних значениях признаков $\bar{y}$, $\bar{x}_1$, $\bar{x}_2$, их средних квадратических отклонениях σ_y, σ_{x_1}, σ_{x_2} и парных коэффициентах корреляции r_{y1}, r_{y2}, r_{12}.

Построим уравнение двухфакторной регрессии в стандартизированном масштабе и рассчитаем показатели тесноты связи (табл. 7.24).

Таблица 7.24

Расчетная таблица для определения параметров уравнения регрессии

y	x_1	x_2	x_1^2	x_2^2	x_1x_2	yx_1	yx_2	y^2
197	4,9	20	24,0	400	98	965	3940	38809
254	5,1	22	26,0	484	112	1295	5588	64516
112	6,5	10	42,3	100	65	728	1120	12544
145	3,7	21	13,7	441	78	537	3045	21025
176	4,0	25	16,0	625	100	704	4400	30976
76	2,5	19	6,3	361	48	190	1444	5776
$\Sigma y =$ $= 960$	$\Sigma x_1 =$ $= 26,7$	$\Sigma x_2 =$ $= 117$	$\Sigma x_1^2 =$ $= 128,3$	$\Sigma x_2^2 =$ $= 2411$	$\Sigma x_1x_2 =$ $= 501$	$\Sigma yx_1 =$ $= 4419$	$\Sigma yx_2 =$ $= 19537$	$\Sigma y^2 =$ $= 173646$

Используя итоги расчетной таблицы (см. табл. 7.24) и известные формулы для расчета средних, дисперсий и парных коэффициентов корреляции:

$$\bar{x} = \frac{\sum x}{n}, \quad \sigma^2 = \overline{x^2} - (\bar{x})^2, \quad r_{yx} = \frac{\overline{xy} - \bar{x}\bar{y}}{\sigma_x \sigma_y},$$

вычислим показатели, необходимые для отыскания β-коэффициентов:

$\bar{y} = 160$ тыс. руб., $\sigma_y = 57,8$ тыс. руб.;

$\bar{x}_1 = 4,45$ млн чел., $\sigma_{x_1} = 1,2513$ млн чел.;

$\bar{x}_2 = 19,5\%$, $\sigma_{x_2} = 4,6458\%$;

$r_{y1} = 0,3392$, $r_{y2} = 0,5071$, $r_{12} = -0,5806$.

Система нормальных уравнений в стандартизированном виде может быть записана так:

$$\begin{cases} 0,3392 = \beta_1 - 0,5806\beta_2, \\ 0,5071 = -0,5806\beta_1 + \beta_2. \end{cases}$$

Решая эту систему, находим: $\beta_1 = 0,9558$, $\beta_2 = 1,062$.

Таким образом, можно записать уравнение регрессии в стандартизированном виде:

$$t_y = 0,9558t_1 + 1,062t_2.$$

Коэффициенты при t_j показывают, что большее воздействие на чистую прибыль предприятия электросвязи оказывает рентабельность ($\beta_2 > \beta_1$). С ее ростом на σ при постоянной численности обслуживаемого населения чистая прибыль увеличивается на 1,062 своего среднего квадратического отклонения.

Переход от стандартизированного уравнения регрессии к уравнению регрессии в натуральном масштабе осуществляется по формулам

$$a_j = \frac{\sigma_y}{\sigma_{x_j}}\beta_j, \quad a_0 = \bar{y} - a_1\bar{x}_1 - a_2\bar{x}_2.$$

Найдем параметры искомого уравнения:

$$a_1 = \frac{\sigma_y}{\sigma_{x_1}}\beta_1 = \frac{57,8}{1,2513}0,9558 = 44,15,$$

$$a_2 = \frac{\sigma_y}{\sigma_{x_2}}\beta_2 = \frac{57,8}{4,6458}1,062 = 13,21,$$

$$a_0 = \bar{y} - a_1\bar{x}_1 - a_2\bar{x}_2 = 160 - 44,15 \cdot 4,45 - 13,21 \cdot 19,5 = -294.$$

Уравнение зависимости чистой прибыли предприятий электросвязи от численности обслуживаемого населения и рентабельности имеет вид

$$\bar{y}_x = -294 + 44,15x_1 + 13,21x_2.$$

Оно показывает, что с ростом численности обслуживаемого населения на 1 млн чел. при исключении влияния второго фактора (рентабельности) чистая прибыль возрастает на 44,15 тыс. руб., а при неизменной численности населения с ростом рентабельности на 1% чистая прибыль повысится на 13,21 тыс. руб.

Коэффициент множественной детерминации для нашего примера окажется равным

$$R^2_{y,\,x_1,\,x_2} = \beta_1 r_{y1} + \beta_2 r_{y2} = 0,9558 \cdot 0,3392 + 1,062 \cdot 0,5071 = 0,8627.$$

Отсюда коэффициент множественной корреляции

$$R_{y,\,x_1,\,x_2} = \sqrt{R^2_{y,\,x_1,\,x_2}} = \sqrt{0,8627} = 0,929.$$

Полученные значения коэффициентов множественной корреляции и детерминации, близкие к 1, свидетельствуют о том, что при построении двухфакторной модели учтены важные факторы увеличения чистой прибыли. При дополнительном включении факторов в анализ (для данного числа предприятий) может увеличиться совокупный коэффициент детерминации и, соответственно, уменьшиться остаточная дисперсия, доля которой в нашем примере мала:

$$\sigma^2_{\text{ост}} = 1 - R^2_{y,\,x_1,\,x_2} = 1 - 0,8627 = 0,1373.$$

Следовательно, на долю неучтенных факторов приходится не более 13,73% дисперсии результативного признака.

Рассчитаем эластичность по каждому фактору и по их совокупности:

$$Э_1 = a_1 \frac{\bar{x}_1}{\bar{y}} = 44{,}15 \frac{4{,}45}{160} = 1{,}23,$$

$$Э_2 = a_2 \frac{\bar{x}_2}{\bar{y}} = 13{,}21 \frac{19{,}5}{160} = 1{,}61,$$

$$\sum_{j=1}^{2} Э_j = 2{,}84.$$

Эластичность по каждому фактору и в целом по совокупности больше 1, значит, чистая прибыль увеличивается в большей степени, чем факторы. С увеличением каждого фактора на 1% следует ожидать увеличения чистой прибыли на 2,84%.

Важный этап в построении моделей множественной регрессии и проведении корреляционно-регрессионного анализа – *статистическая оценка точности и надежности параметров корреляции.*

Для оценки значимости каждого коэффициента регрессии необходимо рассчитать значение *t*-критерия Стьюдента (отношение коэффициента регрессии к его средней ошибке):

$$t^j_{\text{расч}} = \frac{|a_j|}{\sigma_{a_j}}.$$

Средняя (стандартная) ошибка коэффициента регрессии может быть найдена приближенно:

$$\sigma^2_{a_j} = \frac{\sigma^2_y}{k},$$

где σ^2_y – дисперсия результативного признака;

k – число факторных признаков.

Коэффициент регрессии считается статистически значимым, если $t^j_{\text{расч}}$ превышает $t_{\text{табл}}$ – табличное (теоретическое) значение *t*-критерия Стьюдента для заданного уровня значимости α и $(n - k - 1)$ степеней свободы:

$$t^j_{\text{расч}} > t^{\alpha, n-k-1}_{\text{табл}}.$$

Напомним, что подобным образом рассчитывается t-критерий Стьюдента и для оценки значимости коэффициента парной корреляции:

$$t_{\text{расч}} = \frac{|r_{yj}|}{\sigma_{r_{yj}}},$$

где $\sigma_{r_{yj}} = \sqrt{\frac{1 - r_{yj}^2}{n - 2}}$.

Коэффициент парной корреляции считается значимым, если

$$t_{\text{расч}} > t_{\text{табл}}^{\alpha, n-k-1}.$$

Вывод об адекватности всей модели и правильности выбора формы связи можно проверить с помощью F-критерия:

$$F_{\text{расч}} = \frac{R^2}{1 - R^2} \frac{n - k - 1}{k},$$

где R^2 – совокупный коэффициент множественной детерминации.

Величина $F_{\text{табл}}$ находится по таблицам (см. Приложение 8) при заданном уровне значимости α и числе степеней свободы $\nu_1 = k$, $\nu_2 = n - k - 1$.

F-критерий представляет собой соотношение оценок факторной и случайной дисперсий, рассчитанных на одну степень свободы. Число степеней свободы для факторной дисперсии $\nu_1 = k$, для случайной $\nu_2 = n - k - 1$. Если $F_{\text{расч}} > F_{\text{табл}}$, связь признается существенной.

Если факторных признаков много и число значений каждого из них велико, для корреляционно-регрессионного анализа применяют стандартные статистические программы.

Рассмотрим процедуру вычислений параметров множественной линейной регрессии на примере статистической программы *Stadia*. В качестве исходных данных вновь используем информацию о предприятиях электросвязи (см. табл. 7.23). Однако теперь возьмем 14 предприятий и количество факторных признаков увеличим до 7.

Для программы *Stadia* исходную информацию требуется представить в виде матрицы с размерами $(k + 1) \times n$, в которой первые k столбцов соответствуют независимым переменным x_j ($j = \overline{1, k}$), а $(k + 1)$-й столбец – результативному признаку (табл. 7.25). При этом каждый столбец содержит n значений факторов. В нашем

Таблица 7.25

Исходные данные о предприятиях электросвязи

Номер предприятия	Капитализация, млн руб. x_1	Реализация, тыс. руб. x_2	Количество линий, тыс. x_3	Численность населения, млн чел. x_4	Рентабельность, % x_5	Цифровизация, % x_6	Износ, % x_7	Чистая прибыль, тыс. руб. y
1	10369	2861	4009	8,6	26	11	49	736
2	3823	1006	1827	4,9	20	22	35	197
3	2662	1154	632	5,1	22	17	41	254
4	2328	1088	1192	6,5	10	13	39	112
5	2295	696	618	3,7	21	22	43	145
6	1615	715	569	4,0	25	24	41	176
7	1519	662	563	4,3	8	17	42	50
8	869	408	423	2,5	19	28	39	76
9	845	455	387	2,6	10	15	47	44
10	773	409	443	2,4	20	39	35	81
11	751	343	347	1,6	12	15	45	40
12	730	402	287	1,6	18	28	43	73
13	727	355	330	2,7	11	9	46	41
14	604	380	210	1,1	8	10	37	30

случае $k = 7$, $n = 14$, результативным признаком y является чистая прибыль, x_1, x_2, ..., x_7 – факторные признаки.

Подсчет ведется для регрессионной модели

$$\bar{y}_x = a_0 + \sum_{j=1}^{7} a_j x_j.$$

Результаты расчета получают в следующем виде:

МНОЖЕСТВЕННАЯ ЛИНЕЙНАЯ РЕГРЕССИЯ

Коэфф.	a_0	a_1	a_2	a_3	a_4	a_5	a_6	a_7
Значение	-162	0,00937	0,259	-0,000697	-22,4	6,51	-0,572	1,76
Ст. ошиб.	59,7	0,0193	0,0488	0,0294	5,35	1,35	0,882	1,21
Значим.	0,0341	0,646	0,0023	0,98	0,0062	0,0034	0,546	0,196

Множеств. R	R^2	Ст. ошиб.	F	Значим.
0,99859	0,99718	14,289	303	0

Гипотеза 1: <Регрессионная модель адекватна экспериментальным данным>

Итак, модель зависимости чистой прибыли для 14 предприятий электросвязи от перечисленных семи факторных признаков согласно результатам вычислений выглядит следующим образом:

$$\bar{y}_x = -162 + 0{,}00937x_1 + 0{,}259x_2 - 0{,}000697x_3 - 22{,}4x_4 + \\ + 6{,}51x_5 - 0{,}572x_6 + 1{,}76x_7.$$

Выбранные признаки практически полностью описывают вариацию результативного фактора, так как совокупный коэффициент множественной детерминации $R^2 = 0{,}99718$.

Однако полученная модель слишком громоздка. На практике обычно ограничиваются меньшим количеством наиболее существенных факторов. Отбор этих факторов можно осуществить с помощью процедуры *пошаговой регрессии*. Сущность этого метода состоит в последовательном включении факторов в уравнение регрессии и последующей проверке их значимости.

На первом этапе рассчитываются парные коэффициенты корреляции, представленные в виде корреляционной матрицы:

КОРРЕЛЯЦИОННАЯ МАТРИЦА

	x_1	x_2	x_3	x_4	x_5	x_6	x_7
x_2	0,982						
x_3	0,981	0,955					
x_4	0,851	0,903	0,844				
x_5	0,533	0,506	0,472	0,412			
x_6	−0,273	−0,305	−0,251	−0,281	0,455		
x_7	0,34	0,333	0,272	0,201	−0,012	0,512	
y	0,974	0,97	0,931	0,811	0,652	−0,194	0,357

Из этой матрицы следует, что факторы x_1, x_2, x_3, x_4 мультиколлинеарны, так как их парные коэффициенты корреляции больше 0,8. Для получения адекватной модели необходимо устранить мультиколлинеарность. С этой целью переменные x_1, x_2, x_3 выводятся из рассмотрения. К остальным переменным применяется процедура пошаговой регрессии.

На первом шаге этой процедуры в модель автоматически вводится переменная x_4, так как она имеет наибольший по абсолютной величине коэффициент корреляции с результативным признаком: $|r_{y4}| = 0{,}811$.

*** **МЕТОД ВКЛЮЧЕНИЯ.** Шаг № 1, введена переменная: x_4

Коэфф.	a_0	a_1
Значение	−115	71,1
Ст. ошиб.	62	14,8
Значим.	0,0855	0,0006

Множеств. R	R^2	Ст. ошиб.	F	Значим.
0,81138	0,65834	111,19	23,1	0

Гипотеза 1: <Регрессионная модель адекватна экспериментальным данным>

Измен. R^2	F	Значим.
0,658	23,1	0,0006

Переменные в уравнении

Переменн.	Коэфф. B	Ст. ош. B	Бета	F	Значим.
x_4	71,1	14,8	0,811	23,1	0,0006

Переменные не в уравнении

Переменн.	Коэфф. B	Ст. ош. B	Бета	F	Значим.
x_5	11	4,46	0,383	6,08	0,0299
x_6	0,784	3,97	0,0362	0,039	0,841
x_7	8,61	7,18	0,203	1,44	0,255

В результате получаем однофакторную регрессионную модель:

$$\bar{y}_x = -115 + 71{,}1x_4,$$

при этом $R^2 = 0{,}65834$, т.е. вариацией фактора x_4 объясняется 65,8% вариации результативного признака.

Для уточнения модели регрессии на в т о р о м шаге в рассмотрение автоматически включается следующая переменная x_5 с наибольшим частным коэффициентом корреляции.

*** **МЕТОД ВКЛЮЧЕНИЯ.** Шаг № 2, введена переменная: x_5

Коэфф.	a_0	a_1	a_3
Значение	-245	57,2	11
Ст. ошиб.	74	13,6	4,46
Значим.	0,0069	0,0017	0,0299

Множеств. R	R^2	Ст. ошиб.	F	Значим.
0,88319	0,78002	93,185	19,5	0,0004

Гипотеза 1: <Регрессионная модель адекватна экспериментальным данным>

Измен. R^2	F	Значим.
0,122	6,08	0,0299

Переменные в уравнении

Переменн.	Коэфф. B	Ст. ош. B	Бета	F	Значим.
x_4	57,2	13,6	0,653	17,7	0,0017
x_5	11	4,46	0,383	6,08	0,0299

Переменные не в уравнении

Переменн.	Коэфф. B	Ст. ош. B	Бета	F	Значим.
x_6	-7,61	3,71	-0,351	4,21	0,065
x_7	10,3	5,57	0,244	3,45	0,0904

Двухфакторная регрессионная модель имеет вид

$$\bar{y}_x = -245 + 57{,}2x_4 + 11x_5.$$

В результате включения дополнительного фактора совокупный множественный коэффициент детерминации R^2 возрос до 0,78002.

Увеличение R^2 свидетельствует о целесообразности включения x_5 в модель.

На этом процедура пошаговой регрессии для нашего случая прекращается, поскольку ни одна из оставшихся переменных не обеспечивает заданное минимальное значение F-критерия. Кроме того, включение оставшихся переменных в уравнение существенно не увеличивает совокупный коэффициент множественной детерминации R^2.

Таким образом, мы получили, что результативный признак для 14 предприятий электросвязи – чистая прибыль – в основном зависит от двух факторов: численности населения, обслуживаемого предприятием x_4, и рентабельности x_5.

Глава 8
АНАЛИЗ РЯДОВ ДИНАМИКИ

8.1. Понятие о рядах динамики. Их виды

Изучение изменения различных явлений во времени – одна из важнейших задач статистики. Решается эта задача путем составления и анализа так называемых рядов динамики (иногда их также называют временны́ми или хронологическими рядами).

Ряд динамики представляет собой числовые значения определенного статистического показателя в последовательные моменты или периоды времени (т.е. расположенные в хронологическом порядке).

Числовые значения того или иного статистического показателя, составляющие ряд динамики, называют ***уровнями*** ряда и обычно обозначают через y. Первый член ряда y_0 (или y_1) называют *начальным уровнем*, а последний y_n – *конечным*. Моменты или периоды времени, к которым относятся уровни, обозначают через t.

Ряды динамики, как правило, представляют в виде таблицы или графически. При графическом изображении ряда динамики на оси абсцисс строится шкала времени t, а на оси ординат – шкала уровней ряда y (арифметическая или иногда логарифмическая).

Одна из первых задач изучения рядов динамики – выявить основную тенденцию (закономерность) в изменении уровней ряда, именуемую ***трендом***. Закономерность в изменении уровней ряда в одних случаях проявляется довольно наглядно, в других – может затушевываться колебаниями, вызванными случайными и неслучайными причинами.

В табл. 8.1 приведен временной ряд числа фермерских хозяйств в России за 1991–2002 гг. Изображенные на рис. 8.1 эти данные наглядно иллюстрируют бурный рост числа фермерских хозяйств с 1991 по 1996 г. и замедленный рост (даже снижение) в последующие годы.

Данные табл. 8.2 и рис. 8.2 наглядно иллюстрируют снижение добычи угля за период 1991–2002 гг.

Таблица 8.1

Фермерские хозяйства в России (на 1 января)

Год	Число хозяйств, тыс.	Год	Число хозяйств, тыс.
1991	4,4	1997	278,6
1992	49,0	1998	274,3
1993	182,8	1999	270,2
1994	270,0	2000	261,1
1995	279,2	2001	261,7
1996	280,1	2002	265,6

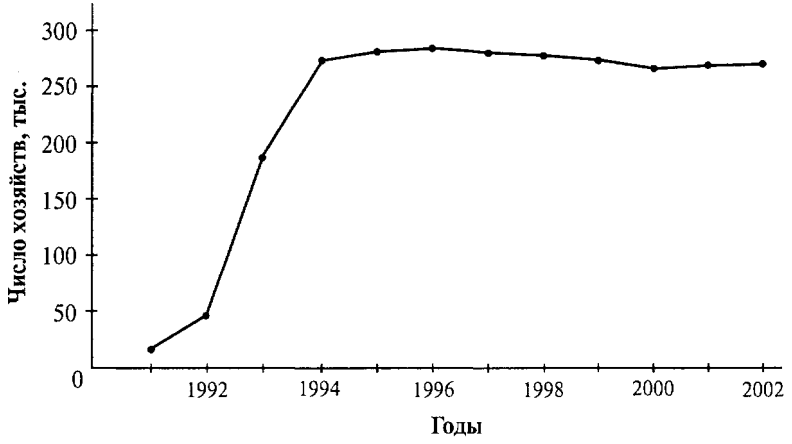


Рис. 8.1. Динамика числа фермерских хозяйств в России за 1991–2002 гг.

Таблица 8.2

Добыча угля в России

Год	Добыто млн т	Год	Добыто млн т
1991	353	1997	244
1992	337	1998	232
1993	306	1999	250
1994	272	2000	258
1995	263	2001	270
1996	257	2002	253

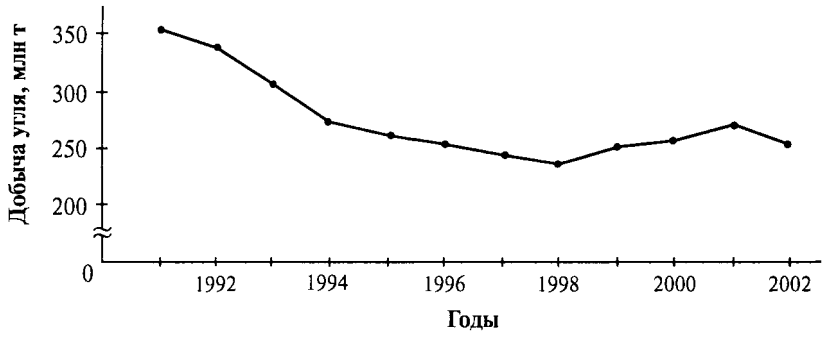


Рис. 8.2. Добыча угля в России в 1991–2002 гг.

Несколько иной характер изменений уровня ряда наблюдается в табл. 8.3 и на рис. 8.3.

Данные табл. 8.3 и рис. 8.3 свидетельствуют о значительных колебаниях уровней по месяцам при общем снижении рождаемости.

Таблица 8.3

Численность родившихся в России по месяцам за 1997 г.

Месяц	Число родившихся, тыс. чел.	Месяц	Число родившихся, тыс. чел.
Январь	108,3	Июль	118,2
Февраль	103,4	Август	103,2
Март	105,4	Сентябрь	104,4
Апрель	109,9	Октябрь	103,6
Май	106,2	Ноябрь	90,5
Июнь	105,9	Декабрь	104,4

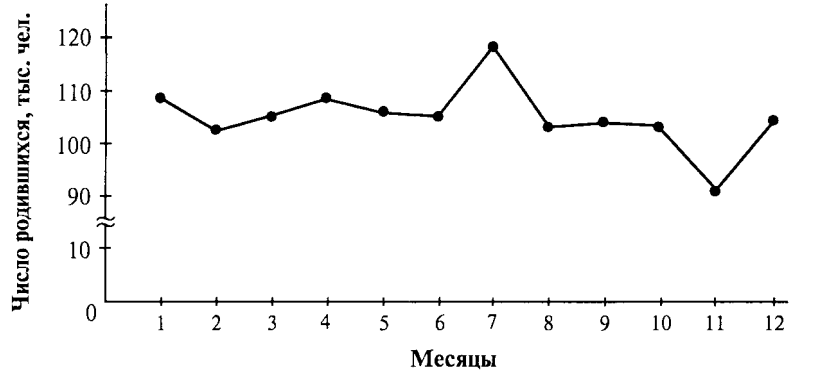


Рис. 8.3. Динамика рождаемости по месяцам за 1997 г.

Аналогичные колебания уровней ряда, но по годам, наблюдаются в динамике средней урожайности зерновых в России (табл. 8.4 и рис. 8.4).

Таблица 8.4

Средняя урожайность зерновых в России (во всех категориях хозяйств)

Год	Урожайность, ц/га	Год	Урожайность, ц/га
1991	14,4	1997	16,5
1992	17,2	1998	12,9
1993	16,3	1999	14,4
1994	14,4	2000	15,6
1995	11,6	2001	19,4
1996	12,9		

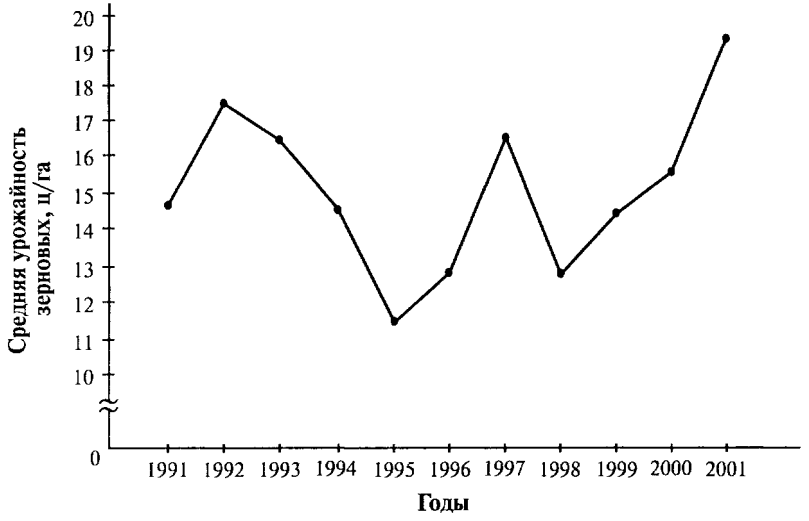


Рис. 8.4. Динамика урожайности зерновых в России за 1991–2001 гг.

В табл. 8.5 и на рис. 8.5 приведены данные об уровне рентабельности промышленной продукции в 1993–1991 гг.

Таблица 8.5

Уровень рентабельности продукции в промышленности России

Год	1993	1994	1995	1996	1997	1998	1999	2000	2001
Рентабельность продукции, %	32,0	19,5	20,1	9,2	9,0	12,7	25,5	24,7	18,5

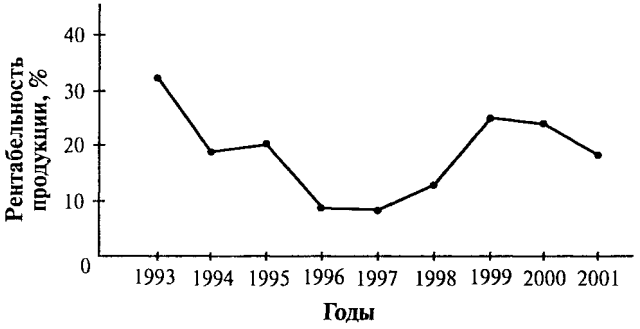


Рис. 8.5. Динамика уровня рентабельности продукции в промышленности России за 1993–2001 гг.

В приведенных примерах рядов динамики отражен различный характер изменения во времени отдельных показателей.

Уровни любого ряда – это результат взаимодействия самых разных факторов, одни из которых действуют длительно, другие – кратковременно, одни – главные (определяют тенденцию изменения), другие – случайные (затушевывают ее) и т.д.

Поэтому, чтобы сделать правильные выводы о закономерностях развития того или иного показателя, надо суметь отделить главную тенденцию изменения (тренд) от колебаний, вызванных случайными кратковременными причинами. Для этого ряды динамики подвергают обработке.

Виды рядов динамики

В одних рядах уровни выражены абсолютными показателями (см. табл. 8.2, 8.3), в других – средними (см. табл. 8.4) или относительными (см. табл. 8.5). В зависимости от вида показателей уровней ряда и ряды динамики также подразделяют на ***ряды абсолютных, относительных*** и ***средних величин (показателей)***.

На основе рядов абсолютных величин образуются ряды динамики относительных и средних величин, поэтому ряды абсолют-

ных величин рассматривают как исходные, а ряды относительных и средних величин – как производные.

Ряды относительных величин могут характеризовать: темпы роста (или снижения) определенного показателя; изменение удельного веса того или иного показателя в совокупности (например, удельного веса (доли) городского населения или доли приватизированных предприятий в той или иной отрасли); изменение показателей интенсивности отдельных явлений (например, производство продукции на душу населения, уровень рождаемости и смертности на 1000 человек населения) и др.

Примерами рядов динамики средних величин служат данные о среднегодовой численности занятых в экономике (или безработных), о средней заработной плате в отдельных отраслях, о среднем размере пенсий, о средней урожайности отдельных сельскохозяйственных культур и др.

Кроме того, уровни рядов динамики могут относиться к определенным моментам времени (датам) или же периодам (интервалам). Так, в табл. 8.1 уровни характеризуют значение изучаемого показателя (число фермерских хозяйств) по состоянию на 1 января каждого года; а в табл. 8.2 и 8.3 уровни (добыча угля и число родившихся) являются показателями, полученными как итог за указанный период (год или месяц).

В соответствии с этим в статистике различают моментные и интервальные ряды динамики.

Моментным называется ряд, уровни которого характеризуют значение показателя (явления) по состоянию на определенные моменты времени (дату).

Интервальным называется ряд, уровни которого характеризуют значение показателя, достигнутое за определенный период (интервал) времени.

Отметим отличительную особенность интервальных рядов абсолютных величин: их уровни можно дробить и складывать (суммировать). Так, зная добычу угля по годам, можно разделить каждый уровень на 12 и получить новые данные – о среднемесячной добыче угля за указанный период. Или же, суммируя данные о численности родившихся по месяцам, можно получить численность родившихся за год. Подобные действия с уровнями моментного ряда лишены смысла.

Суммируя уровни интервальных рядов абсолютных величин, можно строить ряды с нарастающим итогом. Примером такого ряда служит ряд, приведенный в табл. 8.6.

Таблица 8.6

Строительство в России автомобильных дорог с твердым покрытием общего типа

Год	Построено автомобильных дорог, тыс. км	Построено автомобильных дорог (нарастающим итогом, начиная с 1991 г.), тыс. км
1991	10,3	10,3
1992	6,4	16,7
1993	7,9	24,6
1994	6,5	31,1
1995	7,5	38,6
1996	5,7	44,3
1997	4,7	49,0
1998	5,0	54,0
1999	5,4	59,4
2000	6,6	66,0
2001	4,4	70,4

8.2. Сопоставимость уровней и смыкание рядов динамики

Одно из требований, которые предъявляются к анализируемым рядам динамики, – сопоставимость уровней ряда.

Несопоставимость уровней может возникнуть по разным причинам. Перечислим основные из них:

- изменение границ территории, к которой отнесены те или иные показатели;
- изменение методологии учета или расчета показателей. Например, если в одни годы средняя урожайность какой-либо сельскохозяйственной культуры рассчитывалась с засеянной площади, а в другие – с убранной, то уровни несопоставимы. Или если в одни годы производительность труда в промышленности определялась в расчете на одного рабочего, а в другие – на одного работника промышленно-производственного персонала, то сравнивать такие данные или соединять их в один ряд нельзя – они несопоставимы;
- изменение даты учета. Например, учет скота в течение ряда лет проводился по состоянию на 1 октября, а затем – на 1 января. Если соединить в один ряд данные о численности скота за ряд лет с разной датой учета, получим несопоставимые уровни;
- изменение единиц измерения или счета. Так, например, нельзя сравнивать данные о производстве ткани, если за одни годы

они приведены в погонных метрах, а за другие – в квадратных. Или, например, если меняется масштаб цен (как это произошло в России), то нельзя стоимостные показатели за одни годы приводить в старых, а за другие – в новых ценах;

- различная продолжительность периодов, к которым относятся уровни. Например, нельзя строить ряд, где одни уровни являются месячными показателями, а другие – квартальными или годовыми.

Могут быть и другие причины несопоставимости.

Однако в зависимости от цели исследования выводы о сопоставимости данных могут быть различными. Так, изменение границ территории не всегда служит препятствием для сравнения данных в старых и новых границах. Например, если с изменением границ какой-то области ставится задача определить изменение численности населения (или объема производства промышленной продукции) в данной области именно в связи с изменением ее территории, то не только можно, но и до́лжно сопоставлять данные (о численности населения или объеме производства) в разных границах. Если же ставится задача охарактеризовать темпы естественного прироста населения (или развития промышленности), то сравниваемые показатели должны относиться к одним и тем же территориальным границам.

Следовательно, прежде чем анализировать уровни ряда динамики, надо, исходя из цели исследования, убедиться в их сопоставимости. Если данные несопоставимы, необходимо добиться их сопоставимости, прибегнув к дополнительным расчетам.

Смыкание рядов динамики

Под *смыканием рядов динамики* понимают объединение в один ряд (более длинный) двух или нескольких рядов, уровни которых исчислены по разным методологиям или в разных границах. При этом для осуществления такого смыкания необходимо, чтобы данные для одного из периодов (переходного) были исчислены по двум методологиям. Покажем это на примере. Предположим, по одной из областей России имеются данные о численности безработных, определенные за 1993–1995 гг. на 1 октября, а за 1995–1997 гг. – на конец марта (табл. 8.7).

Чтобы проанализировать динамику численности безработных за 1993–1997 гг., необходимо сомкнуть (объединить) приведенные в табл. 8.7 два ряда в один, а чтобы уровни нового ряда были сопоставимы, следует пересчитать данные за 1993 и 1994 гг. по состоянию на конец марта.

Таблица 8.7

Численность безработных в одной из областей России

Год	1993	1994	1995	1996	1997
Численность безработных, тыс. чел.:					
на 1 октября	20	22,5	25	—	—
на конец марта	—	—	27	29	32,5
Сомкнутый ряд абсолютных величин на конец марта, тыс. чел.	21,6	24,3	27	29	32,5
Сомкнутый ряд относительных величин, % к 1995 г.	80,0	90,0	100	107,4	120,4

Для этого на основе данных за 1995 г., определенных на две даты, рассчитываем отношение между ними: 27/25 = 1,08. Умножая на этот коэффициент данные за 1993–1994 гг., делаем их сопоставимыми с последующими уровнями. Сомкнутый ряд динамики (в абсолютных величинах) показан в средней части табл. 8.7.

Можно применить и другой способ смыкания рядов, дающий результат в относительных величинах. Так, например, уровни 1995 г. (для него имеются данные учета безработных на две даты) принимаются за 100%, а остальные пересчитываются в процентах к ним: соответственно за 1993 и 1994 гг. – к 25 тыс. чел., а за 1996 и 1997 гг. – к 27 тыс. чел. В результате получаем сомкнутый (сопоставимый) ряд динамики численности безработных в процентах к 1995 г., т.е. в относительных величинах (приведен в нижней части табл. 8.7).

Приведение рядов к одному основанию

Переход к относительным величинам целесообразно осуществлять и при параллельном анализе динамики нескольких показателей (или одного и того же показателя по разным объектам), если по абсолютным данным трудно выявить особенности развития. В таких случаях уровни всех рассматриваемых рядов приводятся в процентах (или коэффициентах) к уровню одного и того же периода или момента времени (либо иной базе сравнения). Этот прием перехода от абсолютных показателей к относительным именуют в статистике *приведением рядов к одному основанию.*

Рассмотрим его на примере данных, приведенных в табл. 8.8.

Во всех рядах заметно снижение уровней с 1992 по 1998 г., а затем снова повышение. Однако сделать вывод об интенсивности снижения и повышения по отдельным видам продукции визуально затруднительно.

Таблица 8.8

Динамика объема производства некоторых видов продукции в России за 1991–2001 гг.

Год	Добыча угля, млн т	Добыча нефти (без газового конденсата), млн т	Добыча природного газа, млрд $м^3$	Производство электроэнергии всеми электростанциями, млрд кВт/ч
1991	353	452	608	1068
1992	337	390	609	1008
1993	306	345	588	957
1994	272	310	581	876
1995	263	298	570	860
1996	257	293	575	847
1997	245	297	544	834
1998	232	294	564	827
1999	250	295	564	846
2000	258	313	555	878
2001	270	337	551	891

Для наглядности приведем все четыре ряда к одному основанию, для чего примем уровни 1991 г. в каждом ряду за 100% (табл. 8.9).

Таблица 8.9

Динамика объема производства некоторых видов продукции в России
(в % к 1991 г.)

Год	Уголь	Нефть	Природный газ	Электроэнергия
1991	100,0	100,0	100,0	100,0
1992	95,5	86,3	100,2	94,4
1993	86,7	76,3	96,7	89,6
1994	77,0	68,5	95,6	82,0
1995	74,5	65,9	93,8	80,5
1996	72,8	64,8	94,6	79,3
1997	69,4	65,7	89,5	78,1
1998	65,7	65,0	92,8	77,4
1999	70,8	65,3	92,8	79,2
2000	73,1	69,2	91,3	82,1
2001	76,5	74,6	90,6	83,4

Нетрудно заметить, что данные табл. 8.9, где все ряды приведены к одному основанию, легче интерпретировать, анализиро-

вать. Итак, самое большое снижение объема производства произошло к 1998 г. в добыче нефти и угля; к 2001 г. она (добыча) несколько повысилась и составила соответственно 74,6 и 76,5% по отношению к уровню 1991 г. Меньше всего за указанный период изменялась добыча природного газа.

Обычно ряды динамики приводят к одному основанию и тогда, когда сравнивают за несколько лет один и тот же показатель в разных странах, оцениваемый в соответствующей валюте.

Таблица 8.10 содержит данные о валовом внутреннем продукте (ВВП) в ряде стран, приведенные к одному основанию (уровень 1990 г. принят за 100), что облегчает параллельное сравнение данных.

Таблица 8.10

Динамика ВВП в ряде стран (уровень 1990 г. = 100)

Год	Россия	Великобритания	США	Норвегия	Франция	Япония
1990	100	100	100	100	100	100
1991	95	98	99	102	101	104
1992	81	98	102	107	102	105
1993	74	100	106	109	101	105
1994	65	103	110	115	103	106
1995	62	107	111	120	105	107
1996	60	109	115	126	107	112
1997	61	113	119	132	109	113
1998	58	115	124	134	113	109
1999	61	118	129	135	116	110
2000	66	122	136	138	120	113

8.3. Основные показатели изменения уровней ряда

Анализ рядов динамики начинается с определения того, как именно изменяются уровни ряда (увеличиваются, уменьшаются или остаются неизменными) в абсолютном и относительном выражении. Чтобы проследить за направлением и размером изменений уровней во времени, для рядов динамики рассчитывают такие показатели, как:

- абсолютные приросты (изменения) уровней;
- темпы роста;
- темпы прироста (снижения) уровней.

Абсолютный прирост (абсолютное изменение) уровней рассчитывается как разность между двумя уровнями ряда. Он показыва-

ет, на сколько (в единицах измерения показателей ряда) уровень одного периода больше или меньше уровня какого-либо предшествующего периода, и, следовательно, может иметь знак «+» (при увеличении уровней) или «–» (при уменьшении уровней).

В зависимости от базы сравнения абсолютные приросты могут рассчитываться как цепные и как базисные.

Вычитая из каждого уровня предыдущий (${}_{\text{ц}}\Delta y = y_i - y_{i-1}$), получаем абсолютные изменения уровней ряда за отдельные периоды как *цепные*. Вычитая из каждого уровня начальный (${}_{\text{б}}\Delta y = y_i - y_0$), получаем накопленные итоги прироста (изменения) показателя с начала изучаемого периода, т.е. абсолютные изменения рассчитываются как *базисные*.

Если значения цепных абсолютных приростов (изменений) постоянны, то уровни ряда изменяются равномерно. Если же абсолютные приросты от периода к периоду возрастают (или убывают), то уровни изменяются ускоренно (или замедленно). В этом случае можно рассчитать *показатель ускорения* как разность между двумя смежными цепными абсолютными приростами: $\Delta_\Delta = \Delta_i - \Delta_{i-1}$.

Наряду с абсолютными изменениями уровней ряда важно измерить также их относительное изменение.

Темп роста (изменения) $T_{\text{р}}$ – относительный показатель, рассчитываемый как процентное отношение двух уровней ряда.

Темпы роста как относительные величины могут выражаться в виде коэффициентов, т.е. простого кратного отношения (если база сравнения принимается за единицу), и в процентах (если база сравнения принимается за 100 единиц). Чаще всего, говоря о темпах, имеют в виду отношение уровней в процентах.

Выраженные в коэффициентах темпы роста показывают, во сколько раз уровень данного периода больше уровня базы сравнения или какую часть его составляет. При процентном выражении темп роста показывает, сколько процентов составляет уровень данного периода от уровня базы сравнения.

В зависимости от базы сравнения *коэффициенты роста* ($k_{\text{р}}$) могут рассчитываться как *цепные*, когда каждый уровень сопоставляется с уровнем предыдущего периода (${}_{\text{ц}}k_{\text{р}} = y_i / y_{i-1}$), и как *базисные*, когда все уровни сопоставляются с уровнем одного какого-то периода, принятого за базу сравнения (часто это начальный уровень ряда: ${}_{\text{б}}k_{\text{р}} = y_i / y_0$).

Между цепными и базисными коэффициентами роста существует связь, позволяющая при необходимости переходить от цепных к базисным и наоборот, в частности:

- произведение цепных коэффициентов роста равно базисному;
- результат деления двух базисных коэффициентов равен цепному (промежуточному).

Темп прироста (снижения) $T_{пр}$ – относительный показатель, показывающий, на сколько процентов данный уровень больше (или меньше) другого, принимаемого за базу сравнения. Показатель $T_{пр}$ можно рассчитать двояко:

- путем вычитания 100% из темпа роста (снижения), т.е. $T_{пр} = T_р - 100\%$;
- как процентное отношение абсолютного прироста к тому уровню, по сравнению с которым рассчитан абсолютный прирост.

Так, темп прироста за год будет равен $T_{пр} = \frac{ц\Delta y}{y_{i-1}} 100\%$.

Иногда для анализа рассчитывается такой показатель, как ***абсолютное значение 1% прироста*** α – отношение абсолютного прироста уровня к темпу прироста (за соответствующий период):

$$\alpha = \frac{\Delta y}{T_{пр}} = \frac{y_i - y_{i-1}}{\frac{y_i - y_{i-1}}{y_{i-1}} 100\%} = 0{,}01 y_{i-1}.$$

Абсолютное значение 1% прироста равно одной сотой предыдущего уровня.

Для базисных абсолютных приростов и темпов прироста расчет α не имеет смысла, так как при сравнении всех накопленных приростов с одним и тем же первоначальным уровнем y_0 для всех периодов будет получаться одно и то же значение 1% прироста.

Иногда приходится сопоставлять темпы роста или темпы прироста за одни и те же отрезки времени по двум показателям или по одному показателю, но относящемуся к разным территориям (странам, регионам и т.п.) или объектам.

Отношение темпов роста (или прироста) по двум динамическим рядам (в одинаковые отрезки времени) называют *коэффициентом опережения.*

В табл. 8.11 рассчитаны все упомянутые выше показатели изменения уровней ряда на примере производства яиц в России за 1997–2001 гг. Все они характеризуют неуклонное повышение из года в год производства яиц в России за указанный период.

Таблица 8.11

Основные показатели изменения уровней ряда
(на примере производства яиц в России)

Год	1997	1998	1999	2000	2001
Производство яиц, млрд шт. (уровни ряда y)	32,2	32,7	33,1	34,1	35,2
Абсолютные приросты Δ, млрд шт.:					
цепные (по годам)	—	0,5	0,4	1,0	1,1
базисные (с 1997 г.)	—	0,5	0,9	1,9	3,0
Темпы роста базисные (по отношению к 1997 г.):					
коэффициенты	1	1,016	1,028	1,059	1,093
проценты	100,0	101,6	102,8	105,9	109,3
Темпы роста цепные (по отношению к предыдущему году):					
коэффициенты	—	1,016	1,012	1,030	1,032
проценты	—	101,6	101,2	103,0	103,2
Темпы прироста, %:					
ежегодные (цепные)	—	1,6	1,2	3,0	3,2
к 1997 г.	—	1,6	2,8	5,9	9,3
Абсолютное значение 1% прироста, млрд шт.	—	0,312	0,333	0,333	0,344

8.4. Исчисление средних показателей в рядах динамики

Каждый ряд динамики можно рассматривать как некую совокупность n меняющихся во времени показателей, которые можно обобщать в виде средних величин. Такие обобщенные (средние) показатели особенно необходимы при сравнении изменений того или иного показателя в разные периоды, в разных странах и т.д.

Обобщенной характеристикой динамического ряда может служить прежде всего ***средний уровень ряда*** $\bar{y}$. Поскольку средняя величина в данном случае рассчитывается из меняющихся во времени показателей, то она называется *средней хронологической.*

Для разных видов рядов динамики средний уровень рассчитывается неодинаково.

Так, в интервальном ряду абсолютных величин с равными периодами (интервалами) средний уровень рассчитывается как средняя арифметическая простая из уровней ряда:

$$\bar{y} = \frac{\sum y_i}{n},$$

где y_i — отдельные уровни ряда;
n — число уровней.

В примере, приведенном в табл. 8.11, средний годовой уровень производства яиц в России за 5 лет (1997–2001 гг.) составил

$$\bar{y} = \frac{\sum y_i}{n} = \frac{32{,}2 + 32{,}7 + 33{,}1 + 34{,}1 + 35{,}2}{5} = 33{,}46 \text{ млрд шт.}$$

Аналогично определяется средний уровень в рядах средних величин. Так, по данным табл. 8.4, где представлена динамика средней урожайности зерновых по годам, среднюю урожайность, например, за 1991–1995 гг. рассчитаем как

$$\bar{y} = \frac{14{,}4 + 17{,}2 + 16{,}3 + 14{,}4 + 11{,}6}{5} = 14{,}8 \text{ ц/га},$$

а за 1996–2001 гг. соответственно как

$$\bar{y} = \frac{12{,}9 + 16{,}5 + 12{,}9 + 14{,}4 + 15{,}6 + 19{,}4}{6} = 15{,}3 \text{ ц/га}.$$

Несколько по-иному рассчитывается средний уровень для моментных рядов. Например, если имеется моментный ряд, содержащий n уровней $(y_1, y_2, \ldots, y_n)$ с равными промежутками между датами (моментами), то такой ряд легко преобразовать в ряд средних величин. При этом показатель (уровень) на начало каждого периода одновременно является показателем на конец предыдущего периода. Тогда средняя величина показателя для каждого периода (промежутка между датами) может быть рассчитана как полусумма значений y на начало и конец периода, т.е. как $\bar{y}_i = \frac{y_i + y_{i+1}}{2}$. Количество таких средних будет $(n - 1)$. Как указывалось ранее, для рядов средних величин средний уровень рассчитывается по средней арифметической. Следовательно, можно записать

$$\bar{y} = \frac{\frac{y_1 + y_2}{2} + \frac{y_2 + y_3}{2} + \ldots + \frac{y_{n-2} + y_{n-1}}{2} + \frac{y_{n-1} + y_n}{2}}{n-1}.$$

После преобразования числителя получаем

$$\bar{y} = \frac{\frac{y_1}{2} + y_2 + y_3 + \ldots + y_{n-1} + \frac{y_n}{2}}{n-1} = \frac{\frac{y_1 + y_n}{2} + \sum_{i=2}^{n-1} y_i}{n-1}. \qquad (8.1)$$

Эта средняя известна в статистике как *средняя хронологическая для моментных рядов.*

Рассмотрим ее расчет на конкретных примерах.

Пример. Имеются следующие данные об остатках вкладов населения в банках России в первом полугодии 1997 г. (на начало месяца):

Месяц	1	2	3	4	5	6	7
Сумма вкладов y, трлн руб.	127,6	129,7	132,7	133,8	135,4	137,1	139,8

Средний остаток вкладов населения за первое полугодие 1997 г. [по формуле (8.1)] составил

$$\bar{y} = \frac{\frac{127,6 + 139,8}{2} + 129,7 + 132,7 + 133,8 + 135,4 + 137,1}{7 - 1} =$$

$$= 133,73 \text{ трлн руб.}$$

В случае неравных промежутков между датами среднюю хронологическую для моментного ряда можно рассчитать как среднюю арифметическую из средних значений уровней на каждую пару моментов, взвешенных по величине расстояний (отрезков времени) между датами, т.е.

$$\bar{y} = \frac{\left(\frac{y_1 + y_2}{2}\right)t_1 + \left(\frac{y_2 + y_3}{2}\right)t_2 + \ldots + \left(\frac{y_{n-1} + y_n}{2}\right)t_{n-1}}{t_1 + t_2 + \ldots + t_{n-1}} =$$

$$= \frac{\sum\left(y_i + y_{i+1}\right)t_i}{2\sum t_i}. \qquad (8.2)$$

Пример. Пусть имеются следующие данные о наличии товарных остатков на складе за 2003 г.:

Дата учета	01.01.03	01.03.03	01.06.03	01.11.03	01.01.04
Остатки товаров y, тыс. руб.	126	130	138	150	160

Средний месячный остаток товаров за 2003 г. [по формуле (8.2)] составит

$$\bar{y} = \frac{(126 + 130)2 + (130 + 138)3 + (138 + 150)5 + (150 + 160)2}{2(2 + 3 + 5 + 2)} =$$

$$= \frac{3376}{24} = 140,67 \text{ тыс. руб.}$$

В данном случае предполагается, что в промежутках между датами уровни принимали разные значения, и мы из двух известных (y_i и y_{i+1}) определяем средние, из которых затем уже рассчитываем общую среднюю для всего анализируемого периода.

Если же предполагается, что каждое значение y_i остается неизменным до следующего $(i + 1)$-го момента, т.е. известна точная дата изменения уровней, то расчет можно осуществлять по формуле

$$\bar{y} = \frac{\sum y_i t_i}{\sum t_i},$$

где t_i – время, в течение которого уровень y_i оставался неизменным.

Кроме среднего уровня в рядах динамики рассчитываются и другие средние показатели.

Средний абсолютный прирост (изменение) уровней рассчитывается как средняя арифметическая простая из отдельных цепных приростов, т.е.

$$\bar{\Delta}_y = \frac{\sum \left(_{\text{ц}}\Delta_y\right)_i}{n},$$

или на основе накопленного абсолютного прироста за n периодов:

$$\bar{\Delta}_y = \frac{y_n - y_0}{n}.$$

Так, средний годовой абсолютный прирост производства яиц в России за 1998–2001 гг. (см. табл. 8.11) составил

$$\bar{\Delta}_y = \frac{\sum \left(_{\text{ц}}\Delta_y\right)_i}{n} = \frac{0{,}5 + 0{,}4 + 1{,}0 + 1{,}1}{4} = \frac{3}{4} = 0{,}75 \text{ млрд шт.}$$

или

$$\bar{\Delta}_y = \frac{y_n - y_0}{n} = \frac{35{,}2 - 32{,}2}{4} = 0{,}75 \text{ млрд шт.}$$

Примечание. Уровень 1997 г. обозначен через y_0 как базисный для расчета приростов начиная с 1998 г. Период, для которого усредняется показатель годового прироста, составляет 4 года, с 1998 по 2001 г. включительно.

Особое значение в анализе рядов динамики придается расчету средних темпов (коэффициентов) роста.

Наиболее часто ***средний темп роста*** рассчитывается как средняя геометрическая из цепных темпов роста, т.е. рассчитанных в каждый период по отношению к предыдущему.

Основанием для использования средней геометрической служат следующие рассуждения. Пусть имеется определенный ряд динамики с уровнями

$$y_0, y_1, y_2, y_3, \ldots, y_n.$$

Цепные коэффициенты роста k_i для каждого периода составят:

$$k_1 = \frac{y_1}{y_0}, \quad k_2 = \frac{y_2}{y_1}, \quad \ldots, \quad k_n = \frac{y_n}{y_{n-1}}.$$

На основании этого каждый уровень можно выразить через предыдущий или y_0 (базисный):

$$y_1 = y_0 k_1; \quad y_2 = y_1 k_2 = y_0 k_1 k_2; \quad y_3 = y_2 k_3 = y_0 k_1 k_2 k_3; \ \ldots;$$
$$y_n = y_{n-1} k_n = y_0 k_1 k_2 k_3 \cdots k_n,$$

т.е. конечный уровень y_n равняется базисному y_0, умноженному на произведение цепных коэффициентов роста.

Рассчитывая средний коэффициент роста, мы предполагаем, что замена индивидуальных коэффициентов роста k_i средними $\bar{k}$ обеспечивает достижение одинакового значения конечного уровня y_n.

Так как

$$y_n = y_0 k_1 k_2 \cdots k_n \quad \text{и} \quad y_n = y_0 \underbrace{\bar{k}\,\bar{k} \cdots \bar{k}}_{n} = y_0 (\bar{k})^n,$$

то

$$y_0 k_1 k_2 \cdots k_n = y_0 (\bar{k})^n.$$

Отсюда

$$\bar{k} = \sqrt[n]{k_1 k_2 \cdots k_n}, \tag{8.3}$$

т.е. средний коэффициент роста равен корню n-й степени из произведения n цепных коэффициентов роста. Заметим, что это и есть средняя геометрическая из n цепных коэффициентов роста. Если выражать темп роста в процентах, то

$$\bar{T}_p = \sqrt[n]{k_1 k_2 \cdots k_n}\, 100\%.$$

Используя выражение $y_n = y_0 (\bar{k})^n$, получаем другую формулу для расчета среднего коэффициента роста, тождественную формуле (8.3):

$$\bar{k} = \sqrt[n]{\frac{y_n}{y_0}}. \tag{8.4}$$

Таким образом, если средний темп (коэффициент) роста ориентирован на достижение определенного конечного уровня, используются следующие формулы:

$$\bar{k} = \sqrt[n]{\prod k_{i/(i-1)}}, \qquad \bar{k} = \sqrt[n]{\frac{y_n}{y_0}},$$

где $k_{i/(i-1)}$ – цепные коэффициенты роста;
n – число коэффициентов (или число периодов (лет, месяцев), за которые определяется средний коэффициент);
$\prod$ – знак произведения;
y_0 и y_n – соответственно начальный (базисный) и конечный абсолютные уровни.

Поскольку отношение $\frac{y_n}{y_0}$ – базисный коэффициент роста, то формула (8.4) применима не только для абсолютных уровней, но и для коэффициентов роста, приведенных к одной и той же базе.

По данным табл. 8.11 средний годовой коэффициент роста производства яиц в России за 1998–2001 гг. составил:

по формуле (8.3)

$$\bar{k} = \sqrt[n]{\prod k_{i/(i-1)}} = \sqrt[4]{1{,}016 \cdot 1{,}012 \cdot 1{,}03 \cdot 1{,}032} =$$

$$= \sqrt[4]{1{,}093} = 1{,}022 \text{ (т.е. } \bar{T} = 102{,}2\%);$$

по формуле (8.4)

$$\bar{k} = \sqrt[n]{\frac{y_n}{y_0}} = \sqrt[4]{\frac{35{,}2}{32{,}2}} = \sqrt[4]{1{,}093} = 1{,}022 \text{ (т.е. } \bar{T} = 102{,}2\%).$$

Таким образом, среднегодовой темп роста производства яиц за 1998–2001 гг. составил 102,2%, а средний темп прироста – 2,2%.

Однако надо иметь в виду, что средний темп роста, рассчитанный по формулам (8.3) и (8.4), зависит от значений крайних уровней ряда. Одинаковый темп роста можно получить для рядов с совершенно различным характером изменения, но с одинаковыми значениями крайних уровней. Поэтому, прежде чем рассчитывать средний темп роста определенного показателя за какой-либо период, нужно тщательно проанализировать, целесообразно ли вычислять темпы роста в отдельные отрезки времени. В случае необходимости «длинные» и неодинаковые по характеру изменения периоды следует разбить на более однородные части (с похожей динамикой уровней), для которых расчет средних темпов роста будет иметь смысл.

Говоря о среднем геометрическом коэффициенте роста, следует отметить еще одну его особенность. Так, если на основе значений y_0 и среднего геометрического коэффициента (темпа) роста рассчитать за все периоды «теоретические» уровни ($y'_1 = y_0\bar{k}$, $y'_2 = y_0(\bar{k})^2$, ..., $y'_n = y_0(\bar{k})^n$), то сумма последних не будет совпадать с суммой всех фактических уровней ($\sum_1^n y_i$), хотя значения y_0 и y_n в обоих рядах совпадут.

Вместе с тем при расчете среднего коэффициента (темпа) роста порой более важно ориентироваться на достижение общей суммы уровней, а не только конечного уровня. Например, когда речь идет о динамике таких показателей, как вложение инвестиций, ввод в действие жилой площади, строительство автомобильных дорог и т.п., важно определить средний темп роста, при котором достигается суммарное значение показателя за анализируемый период, а не только конечный уровень. При таком подходе каждый уровень можно выразить через y_0 и $\bar{k}$ следующим образом:

$$y_1 = y_0\bar{k}, \quad y_2 = y_0(\bar{k})^2, \quad \ldots, \quad y_n = y_0(\bar{k})^n.$$

Тогда, если ориентироваться на то, что суммы фактических и расчетных уровней должны совпадать, можно записать:

$$y_1 + y_2 + \ldots + y_n = y_0\bar{k} + y_0(\bar{k})^2 + \ldots + y_0(\bar{k})^n,$$

или

$$\sum_1^n y_i = y_0\left[\bar{k} + (\bar{k})^2 + (\bar{k})^3 + \ldots + (\bar{k})^n\right].$$

Отсюда

$$\frac{\sum_1^n y_i}{y_0} = \bar{k} + (\bar{k})^2 + (\bar{k})^3 + \ldots + (\bar{k})^n. \qquad (8.5)$$

Формула (8.5) условно названа *средней параболической*, а рассчитанное по ней $\bar{k}$ – *средним параболическим коэффициентом (темпом) роста*. Эта формула предложена статистиком из Саратова профессором Л.С. Казинцом в книге «Темпы роста и абсолютные приросты» (1975). Он же составил таблицу, в которой для отдельных периодов ($n = 2 \div 10$) определено значение среднего параболического темпа роста $\bar{k}_{\text{параб}}$, соответствующее тому или

иному отношению суммы уровней за период к базисному уровню $\frac{\sum y_i}{y_0}$. Эта таблица* приведена в Приложении 10.

По ней определяется средний параболический темп роста показателя, обеспечивающий получение суммы фактических уровней за период.

Рассмотрим конкретные примеры расчета среднего параболического темпа роста.

Пример 1. Имеются следующие данные по России о вводе в действие жилой площади за 1985–1990 гг.:

Год	1985	1986	1987	1988	1989	1990
Введено млн м^2	62,6	66,2	72,8	72,3	70,4	61,7

Надо определить средний годовой коэффициент (темп) роста ввода в действие жилой площади за 1986–1990 гг.

Сначала рассчитаем средний геометрический темп роста по формуле (8.4), ориентируясь на достигнутый конечный уровень 1990 г.:

$$\bar{k} = \sqrt[n]{\frac{y_n}{y_0}} = \sqrt[5]{\frac{61,7}{62,6}} = 0,997 \text{ (или } 99,7\%).$$

т.е. согласно данному расчету ввод в действие жилой площади в России в 1986–1990 гг. снижался ежегодно на 0,3%. Вместе с тем каждый год (кроме последнего) уровни повышались. Очевидно, что в данном случае расчет среднего годового темпа роста надо выполнять, ориентируясь на общую сумму ввода в действие жилья за весь период (5 лет), т.е. по средней параболической [см. формулу (8.5)]:

$$\bar{k} + (\bar{k})^2 + (\bar{k})^3 + \ldots + (\bar{k})^n = \frac{\sum_1^n y_i}{y_0}.$$

В нашем примере $y_0 = 62,6$, а

$$\sum_1^5 y_i = 66,2 + 72,8 + 72,3 + 70,4 + 61,7 = 343,4.$$

---

* Данная таблица, составленная Казинцом для $\bar{k} > 1$, дополнена расчетами Г.Л. Громыко для $\bar{k} < 1$.

Находим отношение $\frac{\sum_{1}^{5} y_i}{y_0} = \frac{343,4}{62,6} = 5,485.$

Обращаемся далее к таблице Приложения 10 и в графе, где $n = 5$, ищем значение, близкое к полученному нами отношению. В данном случае это 5,468. Этому отношению соответствует $\bar{k} = 1,03$ (или $\bar{T} = 103\%$), что означает увеличение ввода в действие жилой площади в указанный период ежегодно в среднем на 3%. Это и есть средний параболический темп прироста. В нашем примере он отражает реальную картину относительного изменения уровней.

Аналогично решается задача и при снижении уровней.

Пример 2. Ввод в действие жилой площади в России за 1990–1995 гг. характеризовался следующими данными:

Год	1990	1991	1992	1993	1994	1995
Введено млн м^2	61,7	49,4	41,5	41,8	39,2	41,0
	y_0	$\sum_{1}^{5} y_i = 212,9$				

Для расчета среднего параболического темпа роста находим отношение

$$\frac{\sum_{1}^{5} y_i}{y_0} = \frac{212,9}{61,7} = 3,45.$$

По таблице Приложения 10 в графе, где $n = 5$, находим ближайшее к 3,45 значение табличного отношения: это 3,463. Ему соответствует $\bar{k} = 0,88$, что означает ежегодное снижение ввода жилья на 12%.

Таким образом, рассчитывая средние темпы роста (снижения), следует четко определять, что достигается при этом среднем темпе: конечный уровень показателя y_n или же сумма уровней за весь период $\sum_{1}^{n} y_i$. В первом случае для расчета $\bar{k}$ используют среднюю геометрическую, а во втором – среднюю параболическую.

Средние темпы прироста рассчитываются на основе средних темпов роста путем вычитания из последних 100%:

$$\bar{T}_{\text{пр}} = \bar{T}_{\text{р}} - 100\%.$$

Так, в примере 1 средний темп роста ввода жилья в России за 1986–1990 гг. составил 103%. Следовательно, средний годовой темп прироста составил 103 – 100 = 3%.

В примере 2, где за 1991–1995 гг. $\overline{T}_{р} = 88\%$, средний годовой темп снижения ввода жилья составил $\overline{T}_{пр} = 88 - 100 = -12\%$.

8.5. Методы выявления основной тенденции (тренда) в рядах динамики

Как уже отмечалось, уровни ряда динамики формируются под влиянием взаимодействия многих факторов, одни из которых, будучи основными, главными, определяют закономерность, тенденцию развития, другие – случайные – вызывают колебания уровней.

Можно сказать, что динамика ряда включает три компоненты:

- долговременное движение (так называемый тренд);
- кратковременное систематическое движение (например, сезонные колебания);
- несистематическое случайное движение, вызывающее колебания уровней относительно тренда.

Изучая ряды динамики, исследователи пытаются разделить эти компоненты и выявить основную закономерность развития явления в отдельные периоды, т.е. выявить общую тенденцию в изменении уровней рядов, освобожденную от действия случайных факторов. С этой целью (устранить колебания, вызванные случайными причинами) ряды динамики подвергают обработке.

Существует несколько методов обработки рядов динамики, помогающих выявить основную тенденцию изменения уровней ряда, а именно: метод укрупнения интервалов, метод скользящей средней и аналитическое выравнивание. Во всех методах вместо фактических уровней при обработке ряда рассчитываются иные (расчетные) уровни, в которых тем или иным способом взаимопогашается действие случайных факторов и тем самым уменьшается колеблемость уровней. Последние в результате становятся как бы «выравненными», «сглаженными» по отношению к исходным фактическим данным. Такие методы обработки рядов называются *сглаживанием* или *выравниванием* рядов динамики.

Метод укрупнения интервалов

Простейший метод сглаживания уровней ряда – *укрупнение интервалов*, для которых определяется итоговое значение или

средняя величина исследуемого показателя. Этот метод особенно эффективен, если первоначальные уровни ряда относятся к коротким промежуткам времени. Например, если имеются данные о ежесуточной погрузке грузов по какой-либо железной дороге за месяц, то, естественно, в таком ряду возможны значительные колебания уровней, так как чем меньше период, за который приводятся данные, тем больше влияние случайных факторов.

Чтобы устранить это влияние, рекомендуется укрупнить интервалы времени, например до 5 или 10 дней, и для этих укрупненных интервалов рассчитать общий или среднесуточный объем погрузок (соответственно по пятидневкам или декадам). В ряду с укрупненными интервалами времени закономерность изменения уровней будет более наглядной.

Пусть, например, имеются следующие данные о выпуске продукции на предприятии по месяцам за год (в сопоставимых ценах):

Месяц	1	2	3	4	5	6	7	8	9	10	11	12
Выпуск продукции, млн руб.	5,1	5,4	5,2	5,3	5,6	5,8	5,6	5,9	6,1	6,0	5,9	6,2

Укрупним интервалы до трех месяцев и рассчитаем суммарный и среднемесячный выпуск продукции по кварталам. Новые данные будут выглядеть следующим образом:

Квартал	Выпуск продукции, млн руб.	
	общий	среднемесячный
I	15,7	5,23
II	16,7	5,57
III	17,6	5,87
IV	18,1	6,03

Очевидно, что новые данные более четко выражают закономерность изменения выпуска продукции за год – увеличение из квартала в квартал.

Метод скользящей средней

По сути *метод скользящей средней* несколько схож с предыдущим, но в данном случае фактические уровни заменяются средними уровнями, рассчитанными для последовательно подвижных (скользящих) укрупненных интервалов, охватывающих m уровней ряда.

Например, если принять $m = 3$, то сначала рассчитывается средняя величина из первых трех уровней, затем находится сред-

няя величина из второго, третьего и четвертого уровнеи, потом из третьего, четвертого и пятого и т.д., т.е. каждый раз в сумме трех уровней появляется один новый уровень, а два остаются прежними. Это и обусловливает взаимопогашение случайных колебаний в средних уровнях. Рассчитанные из m членов скользящие средние относятся к середине (центру) каждого рассматриваемого интервала.

Рассмотрим данный метод сглаживания на конкретном примере, характеризующем динамику выпуска продукции за 12 месяцев на одном из предприятий (табл. 8.12). Сглаживание будем осуществлять по трем членам (уровням).

Таблица 8.12

Расчет скользящей средней по трем членам

Месяц	Выпуск продукции, тыс. ед.	Скользящая сумма трех уровней	Скользящая средняя из трех уровней
Январь	82	—	—
Февраль	79	246	82,0
Март	85	244	81,3
Апрель	80	243	81,0
Май	78	234	78,0
Июнь	76	226	75,3
Июль	72	226	75,3
Август	78	218	72,7
Сентябрь	68	216	72,0
Октябрь	70	212	70,7
Ноябрь	74	210	70,0
Декабрь	66	—	—

Сглаженный ряд (см. последнюю графу табл. 8.12) более наглядно показывает тенденцию к снижению уровней из месяца в месяц, которая в исходном ряду несколько затушевывалась скачкообразными колебаниями уровней. Эффект сглаживания, устраняющего колебания уровней за счет случайных причин, хорошо виден также при графическом изображении фактических и сглаженных уровней (рис. 8.6).

Сглаживание методом скользящей средней можно проводить по любому числу членов m, но удобнее, если m – нечетное число, так как в этом случае скользящая средняя сразу относится к конкретной временной точке – середине (центру) интервала. Если же m – четное, то скользящая средняя относится к промежутку между временны́ми точками: например, при сглаживании по четырем членам средняя из первых четырех уровней будет находить-

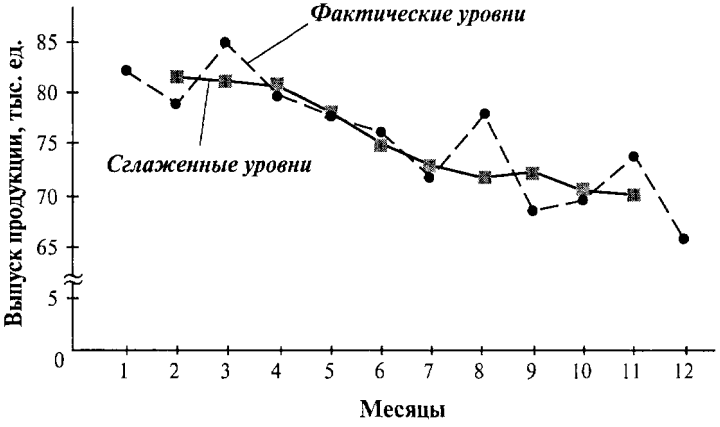


Рис. 8.6. Динамика выпуска продукции за год

ся между второй и третьей датой, следующая средняя – между третьей и четвертой и т.д. Тогда, чтобы сглаженные уровни относились непосредственно к конкретным временны́м точкам (датам), из каждой пары смежных промежуточных значений скользящих средних находят среднюю арифметическую, которую и относят к определенной дате (периоду). Такой прием двойного расчета сглаженных уровней называется *центрированием.*

Недостатком метода скользящей средней является то, что сглаженный ряд «укорачивается» по сравнению с фактическим с двух концов: при нечетном m на $(m - 1)/2$ с каждого конца, а при четном – на $m/2$ с каждого конца. Применяя этот метод, надо помнить, что он сглаживает (устраняет) лишь случайные колебания. Если же, например, ряд содержит сезонную волну, она сохранится и после сглаживания методом скользящей средней.

Кроме того, этот метод сглаживания, как и укрупнение интервалов, является механическим, эмпирическим и не позволяет выразить общую тенденцию изменения уровней в виде математической модели.

Аналитическое выравнивание

Более совершенный метод обработки рядов динамики в целях устранения случайных колебаний и выявления тренда – *выравнивание уровней ряда по аналитическим формулам* (или *аналитическое выравнивание*). Суть аналитического выравнивания заключается в замене эмпирических (фактических) уровней y_i теоретическими $\hat{y}_i$, которые рассчитаны по определенному уравнению,

принятому за математическую модель тренда, где теоретические уровни рассматриваются как функция времени: $\hat{y}_t = f(t)$.

При этом каждый фактический уровень y_i рассматривается как сумма двух составляющих: $y_i = f(t) + \varepsilon_t$, где $f(t) = \hat{y}_t$ – систематическая составляющая, отражающая тренд и выраженная определенным уравнением, а ε_t – случайная величина, вызывающая колебания уровней вокруг тренда.

Задача аналитического выравнивания сводится к следующему:

- определение на основе фактических данных вида (формы) гипотетической функции $\hat{y}_t = f(t)$, способной наиболее адекватно отразить тенденцию развития исследуемого показателя;
- нахождение по эмпирическим данным параметров указанной функции (уравнения);
- расчет по найденному уравнению теоретических (выравненных) уровней.

В аналитическом выравнивании наиболее часто используются следующие простейшие функции:

- *линейная (прямая)*: $\hat{y}_t = a_0 + a_1 t$;
- *показательная*: $\hat{y}_t = a_0 a_1^t$;
- *гиперболическая*: $\hat{y}_t = a_0 + \frac{a_1}{t}$;
- *парабола 2-го (или более высокого) порядка*: $\hat{y}_t = a_0 + a_1 t + a_2 t^2$;
- *ряд Фурье*: $\hat{y}_t = a_0 + \sum_{k=1}^{m} (a_k \cos kt + b_k \sin kt)$.

Здесь $\hat{y}_t$ – теоретические (выравненные) уровни (читается: «игрек, выравненный по t»), t – условное обозначение времени (1, 2, 3, ...), a_0, a_1, a_2 – параметры аналитической функции, k – число гармоник (при выравнивании по ряду Фурье).

Выбор той или иной функции для выравнивания ряда динамики осуществляется, как правило, на основании графического изображения эмпирических данных, дополняемого содержательным анализом особенностей развития исследуемого показателя (явления) и специфики разных функций, их возможности отразить те или иные нюансы развития. Определенную вспомогательную роль при выборе аналитической функции играют также механические приемы сглаживания (укрупнение интервалов и метод скользящей средней). Частично устраняя случайные колебания, они по-

могают более точно определить тренд и выбрать адекватную модель (уравнение) для аналитического выравнивания.

Кроме того, в результате многолетнего опыта использования аналитического выравнивания рядов динамики наработаны некоторые правила или, вернее, условия использования перечисленных простых уравнений, которыми полезно руководствоваться при выборе функции.

1. Так, выравнивание по прямой линии (линейной функции) $\hat{y}_t = a_0 + a_1 t$ эффективно для рядов, уровни которых изменяются примерно в арифметической прогрессии, т.е. когда первые разности уровней (абсолютные приросты) $\Delta = y_i - y_{i-1}$ более или менее постоянны.

2. Если вторые разности уровней (ускорения) более или менее постоянны, то такое развитие хорошо описывается параболой 2-го порядка $\hat{y}_t = a_0 + a_1 t + a_2 t^2$. Если постоянны n-е разности уровней, можно использовать параболу n-го порядка $\hat{y}_t = a_0 + a_1 t + a_2 t^2 + \ldots + a_n t^n$, позволяющую «улавливать» перегибы, изломы в кривой, смену направлений изменения уровней. Парабола 2-го порядка отражает развитие с ускоренным или замедленным изменением уровней ряда.

3. Если при последовательном расположении t (меняющемся в арифметической прогрессии) значения уровней меняются в геометрической прогрессии, т.е. цепные коэффициенты роста примерно постоянны, то такое развитие можно отразить показательной функцией $\hat{y}_t = a_0 a_1^t$.

4. Если обнаружено замедленное снижение уровней ряда, которые по логике не могут снизиться до нуля, для описания характера тренда выбирают гиперболу вида $\hat{y}_t = a_0 + \frac{a_1}{t}$ и т.д.

Если по тем или иным причинам уровни эмпирического ряда трудно математически описать одной функцией, следует разбить исследуемый период на отдельные части и затем выровнять каждую часть по соответствующей кривой.

Нередко один и тот же эмпирический ряд можно выровнять по разным аналитическим формулам (например, по линейной функции и гиперболе или по линейной и показательной функции) и получить при этом довольно близкие результаты. Чтобы решить вопрос о том, использование какой кривой дает лучший результат, обычно сопоставляют суммы квадратов отклонений эмпирических уровней от теоретических, рассчитанных по разным функциям, т.е. $\sum(y - \hat{y}_t)^2$. Та функция, при которой эта сумма

квадратов меньше, считается более адекватной, приемлемой. Однако сравнивать непосредственно суммы квадратов отклонений можно в том случае, если сравниваемые уравнения имеют одинаковое число параметров. Если же число параметров m разное, то каждую сумму квадратов делят на разность $(n - m)$, выступающую в роли числа степеней свободы, и сравнивают уже квадраты отклонений уровней, рассчитанные на одну степень свободы (т.е. остаточные дисперсии на одну степень свободы).

Параметры искомых уравнений при аналитическом выравнивании могут быть определены по-разному. Чаще всего их определяют, решая систему нормальных уравнений, полученных методом наименьших квадратов. Но возможны и другие приемы.

Рассмотрим выравнивание рядов динамики по некоторым аналитическим функциям и методы определения их параметров.

Выравнивание по линейной функции $\hat{y}_t = a_0 + a_1 t$. Воспользуемся данными о производстве мяса в России за 1991–1995 гг. (табл. 8.13) и попытаемся определить закономерность изменения уровней в данном периоде в виде уравнения тренда, т.е. осуществим аналитическое выравнивание ряда.

Таблица 8.13

Производство мяса в России за 1991–1995 гг.

Год	1991	1992	1993	1994	1995	*Итого*
Условное обозначение года t	1	2	3	4	5	
Производство мяса y, млн т	9,4	8,3	7,5	6,8	5,9	Σy = = 37,9

Поскольку в данном ряду уровни меняются примерно в арифметической прогрессии, есть все основания принять уравнение тренда в виде линейной функции. Наша задача – определить параметры a_0 и a_1 искомого уравнения по эмпирическим данным.

Существует несколько методов определения параметров гипотетической линейной функции тренда. Рассмотрим их в порядке возрастания сложности.

1. Самым простым является метод составления и решения системы двух уравнений *по значениям двух конкретных уровней ряда*. Используем его в нашем примере (см. табл. 8.13). Так, ведя отсчет времени от первого года (1991), обозначим условно все временны́е точки через t и придадим им значения 1, 2, 3, 4, 5. Взяв лю-

бые два значения уровней, можно по этим двум точкам построить уравнение прямой (как бы провести прямую через две точки) $\hat{y}_t = a_0 + a_1 t$.

Например, если взять первый и четвертый уровни, т.е. 1991 и 1994 гг., для которых t соответственно равно 1 и 4, а уровни – 9,4 и 6,8, можно записать следующую систему:

$$\begin{cases} a_0 + a_1 \cdot 1 = 9{,}4, \\ a_0 + a_1 \cdot 4 = 6{,}8. \end{cases}$$

Решая эту систему, находим, что $a_1 = -0{,}87$, а $a_0 = 10{,}27$. Отсюда искомое уравнение тренда будет $\hat{y}_t = 10{,}27 - 0{,}87t$. Это приближенная модель тренда. Подставляя в уравнение значения $t = 1, 2, 3, 4, 5$, получаем выравненные (теоретические) значения уровней:

$$\hat{y}_1 = 9{,}4; \quad \hat{y}_2 = 8{,}53; \quad \hat{y}_3 = 7{,}66; \quad \hat{y}_4 = 6{,}79; \quad \hat{y}_5 = 5{,}92.$$

Этот метод отыскания параметров прост, но он не дает однозначного ответа. Если взять значения других двух уровней, то параметры уравнения будут несколько иные. Так, если взять второй и пятый уровни, получим систему уравнений

$$\begin{cases} a_0 + 2a_1 = 8{,}3, \\ a_0 + 5a_1 = 5{,}9. \end{cases}$$

Решив эту систему, получим уравнение тренда $\hat{y}_t = 9{,}9 - 0{,}8t$, в соответствии с которым теоретические уровни будут иметь следующие значения, несколько отличающиеся от полученных выше при $t = 1$ и $t = 4$:

$$\hat{y}_1 = 9{,}1; \quad \hat{y}_2 = 8{,}3; \quad \hat{y}_3 = 7{,}5; \quad \hat{y}_4 = 6{,}7; \quad \hat{y}_5 = 5{,}9.$$

Кроме того, сумма выравненных уровней и в первом и во втором случае расчета по двум точкам не совпадает с суммой эмпирических значений. Чтобы значения параметров были меньше подвержены случайностям, можно усреднить параметры двух уравнений, найденные по разным парам точек. Тогда в нашем примере $a_0 = (10{,}27 + 9{,}9)/2 = 10{,}085$ и $a_1 = [-0{,}87 + (-0{,}8)]/2 = -0{,}835$, а уравнение тренда с усредненными параметрами $\hat{y}_t = 10{,}085 - 0{,}835t$, по которому теоретические уровни будут иметь следующие значения:

$$\hat{y}_1 = 9{,}25; \quad \hat{y}_2 = 8{,}415; \quad \hat{y}_3 = 7{,}58; \quad \hat{y}_4 = 6{,}745; \quad \hat{y}_5 = 5{,}91.$$

2. Другой метод нахождения параметров линейного тренда заключается в следующем: эмпирический ряд разбивают на две ча-

сти (желательно равные) и для каждой из них определяют *суммарные значения уровней и времени.* При этом выдвигается требование, чтобы суммы эмпирических и теоретических (выравненных) уровней были равны или, что одно и то же, чтобы сумма о т к л о н е н и й фактических уровней от теоретических, рассматриваемых как средние, была равна нулю, т.е.

$$\sum(y-\hat{y}_t) = \sum(y-a_0-a_1t) = 0,$$

где y – эмпирические (фактические) уровни;

$\hat{y}_t = a_0 + a_1t$ – теоретические уровни.

Раскрыв скобку и перенеся в правую часть равенства $\sum y$, получим $na_0 + a_1\sum t = \sum y$.

Рассчитав такие суммарные показатели для двух частей исследуемого ряда, получим систему двух уравнений, решение которой и даст параметры искомого уравнения тренда.

Приме́ним этот метод к нашему примеру (см. табл. 8.13).

Разобьем ряд на две части: 1) 1991–1993 гг.; 2) 1994–1995 гг.

Для первой части $\sum_1 t = 1 + 2 + 3 = 6$, для второй $\sum_2 t = 4 + 5 = 9$. Соответственно, для первой части $\sum_1 y = 9{,}4 + 8{,}3 + 7{,}5 = 25{,}2$, для второй $\sum_2 y = 6{,}8 + 5{,}9 = 12{,}7$.

Таким образом, можно записать

$$\begin{cases} 3a_0 + 6a_1 = 25{,}2, \\ 2a_0 + 9a_1 = 12{,}7. \end{cases}$$

Решив эту систему, находим: $a_0 = 10{,}04$ и $a_1 = -0{,}82$.

Отсюда уравнение тренда

$$\hat{y}_t = 10{,}04 - 0{,}82t.$$

Подставив в уравнение значения $t = 1, 2, 3, 4, 5$, получим теоретические (выравненные) уровни:

$\hat{y}_1 = 9{,}22$; $\hat{y}_2 = 8{,}4$; $\hat{y}_3 = 7{,}58$; $\hat{y}_4 = 6{,}76$; $\hat{y}_5 = 5{,}94$,

сумма которых ($\sum\hat{y}_t = 37{,}9$) совпадает с суммой эмпирических уровней, что не наблюдалось, когда уравнение строилось по двум уровням (точкам). Во всех найденных уравнениях a_1 (коэффициент при t) характеризует средний годовой абсолютный прирост (в данном случае убыль) производства мяса. И хотя значения a_1 несколько отличаются в отдельных уровнях, полученных при выборе разных точек (периодов), все они близки к –0,8 млн т.

3. Наиболее распространенный метод нахождения параметров аналитического уравнения при выравнивании рядов динамики –

метод наименьших квадратов (МНК) (см. подпараграф 7.5.1). При этом методе учитываются все эмпирические уровни и должна обеспечиваться минимальная сумма квадратов отклонений эмпирических значений уровней y от теоретических $\hat{y}_t$, т.е.

$$\sum (y - \hat{y}_t)^2 \to \min.$$

В частности, при выравнивании по прямой вида $\hat{y}_t = a_0 + a_1 t$ параметры a_0 и a_1 определяются путем решения системы нормальных уравнений, полученной методом наименьших квадратов (с заменой x на t),

$$\begin{cases} na_0 + a_1 \sum t = \sum y, \\ a_0 \sum t + a_1 \sum t^2 = \sum yt, \end{cases}$$

где n – количество уровней ряда;
t – порядковый номер в условном обозначении периода или момента времени;
y – уровни эмпирического ряда.

Определим этим методом (МНК) параметры линейного тренда в рассматриваемом примере (см. табл. 8.13), для чего исходные данные и все расчеты необходимых сумм представим в табл. 8.14.

Таблица 8.14

Расчет теоретических уровней линейного тренда

Год	Производство мяса, млн т y	Условное обозначение времени t	t^2	yt	Выравненные (теоретические) уровни $\hat{y}_t = 10{,}13 - 0{,}85t$	$(y - \hat{y}_t)^2$
1991	9,4	1	1	9,4	9,28	0,0144
1992	8,3	2	4	16,6	8,43	0,0169
1993	7,5	3	9	22,5	7,58	0,0064
1994	6,8	4	16	27,2	6,73	0,0049
1995	5,9	5	25	29,5	5,88	0,0004
$n = 5$	$\sum y = 37{,}9$	$\sum t = 15$	$\sum t^2 = 55$	$\sum yt = 105{,}2$	$\sum \hat{y}_t = 37{,}9$	$\sum (y - \hat{y}_t)^2 = 0{,}0430$

Приняв в качестве гипотетической функции теоретических уровней прямую $\hat{y}_t = a_0 + a_1 t$, определим параметры последней, для чего решим систему нормальных уравнений, в которую под-

ставлены найденные в итоговой (последней) строке табл. 8.14 суммы:

$$\begin{cases} 5a_0 + 15a_1 = 37{,}9, \\ 15a_0 + 55a_1 = 105{,}2. \end{cases}$$

Решение этой системы возможно любыми известными читателю способами. Существуют готовые формулы для a_1 и a_0:

$$a_1 = \frac{n\sum yt - \sum t \sum y}{n\sum t^2 - (\sum t)^2}, \quad a_0 = \frac{\sum y \sum t^2 - \sum t \sum yt}{n\sum t^2 - (\sum t)^2}.$$

Для a_0 есть еще более простой расчет:

$$a_0 = \bar{y} - a_1 \bar{t}, \quad \text{или} \quad a_0 = \frac{\sum y}{n} - a_1 \frac{\sum t}{n}. \tag{8.6}$$

В нашем примере

$$a_1 = \frac{5 \cdot 105{,}2 - 15 \cdot 37{,}9}{5 \cdot 55 - 15^2} = -0{,}85, \quad a_0 = \frac{37{,}9}{5} - (-0{,}85)\frac{15}{5} = 10{,}13.$$

Отсюда искомое уравнение тренда

$$\hat{y}_t = 10{,}13 - 0{,}85t. \tag{8.7}$$

Подставляя в полученное уравнение значения $t = 1, 2, 3, 4, 5$, определяем теоретические уровни (см. предпоследнюю графу табл. 8.14).

Сравнивая значения эмпирических и теоретических (выравненных) уровней, видим, что они очень близки, т.е. можно сказать, что найденное уравнение весьма удачно характеризует основную тенденцию изменения уровней именно как линейную функцию. Об этом свидетельствует и сумма квадратов отклонений y от $\hat{y}_t$ (см. последнюю графу табл. 8.14), на основе которой рассчитывается среднее квадратическое отклонение от тренда $\sigma = \sqrt{\frac{\sum (y - \hat{y}_t)^2}{n - m}}$ (где m – число параметров в уравнении тренда), используемое для оценки адекватности подобранной линии тренда.

Система нормальных уравнений и, соответственно, расчет параметров a_0 и a_1 упрощаются, если отсчет времени ведется от середины ряда. Например, при нечетном числе уровней серединная точка (год, месяц) принимается за нуль. Тогда предшествующие периоды обозначаются соответственно -1, -2, -3 и т.д., а следующие за средним (центральным) – соответственно 1, 2, 3 и т.д.

При четном числе уровней два серединных момента (периода) времени обозначают -1 и $+1$, а все последующие и предыдущие, соответственно, через два интервала: ± 3, ± 5, ± 7 и т.д.

При таком порядке отсчета времени (от середины ряда) $\sum t = 0$, поэтому система нормальных уравнений упрощается до следующих двух уравнений, каждое из которых решается самостоятельно:

$$\begin{cases} na_0 = \sum y & \Rightarrow \quad a_0 = \dfrac{\sum y}{n}, \\ a_1 \sum t^2 = \sum yt & \Rightarrow \quad a_1 = \dfrac{\sum yt}{\sum t^2}. \end{cases} \tag{8.8}$$

Используем этот прием выравнивания для нашего примера (табл. 8.15).

Таблица 8.15

Расчет теоретических уровней (при счете времени от середины ряда)

Год	Производство мяса, млн т y	t	t^2	yt	Выравненные уровни $\hat{y}_t = 7{,}58 - 0{,}85t$
1991	9,4	-2	4	$-18{,}8$	9,28
1992	8,3	-1	1	$-8{,}3$	8,43
1993	7,5	0	0	0	7,58
1994	6,8	1	1	6,8	6,73
1995	5,9	2	4	11,8	5,88
Σ	37,9	0	10	$-8{,}5$	37,90

Подставив найденные суммы в указанные уравнения, получим (с учетом того, что $n = 5$):

$$5a_0 = 37{,}9 \quad \Rightarrow \quad a_0 = \frac{37{,}9}{5} = 7{,}58,$$
$$10a_1 = -8{,}5 \quad \Rightarrow \quad a_1 = -0{,}85.$$

Отсюда искомое уравнение тренда

$$\hat{y}_t = 7{,}58 - 0{,}85t. \tag{8.7а}$$

В последней графе табл. 8.15 приведены теоретические уровни, рассчитанные по уравнению (8.7а) (путем подстановки в него значений $t = -2, -1, 0, 1, 2$). Они полностью совпадают с теоретическими уровнями, рассчитанными в табл. 8.14 по уравнению (8.7).

Коэффициент регрессии в уравнениях (8.7) и (8.7а), естественно, имеет одинаковое значение ($a_1 = -0{,}85$) и характеризует сред-

нее годовое изменение (уменьшение) производства мяса в России за период 1991–1995 гг. В то же время параметр a_0 (свободный член) различен, поскольку отсчет ведется от разного периода. Поэтому каждый раз, записывая уравнение тренда, необходимо указывать, от какой временно́й точки ведется счет.

Рассмотрим пример упрощенного решения системы уравнений при четном числе уровней. В табл. 8.16 приведены исходные данные о производстве яиц в России за 1996–2001 гг. и расчет показателей, необходимых для определения параметров уравнения тренда в форме линейной функции $\hat{y}_t = a_0 + a_1 t$.

Таблица 8.16

Выравнивание ряда динамики по линейной функции (при счете времени от середины ряда и четном числе уровней)

Год	Производство яиц, млрд шт. y	t	t^2	yt	Выравненные уровни $\hat{y}_t = 33{,}2 + 0{,}32t$
1996	31,9	−5	25	−159,5	31,6
1997	32,2	−3	9	−96,6	32,2
1998	32,7	−1	1	−32,7	32,9
1999	33,1	1	1	33,1	33,5
2000	34,1	3	9	102,3	34,2
2001	35,2	5	25	176,0	34,8
$n = 6$	$\sum y = 199{,}2$	$\sum t = 0$	$\sum t^2 = 70$	$\sum yt = 22{,}6$	$\sum \hat{y}_t = 199{,}2$

Поскольку $\sum t = 0$, то для нахождения параметров a_0 и a_1 используем формулы (8.8):

$$a_0 = \frac{\sum y}{n} = \frac{199{,}2}{6} = 33{,}2 \quad \text{и} \quad a_1 = \frac{\sum yt}{\sum t^2} = \frac{22{,}6}{70} = 0{,}32.$$

Отсюда искомое уравнение тренда

$$\hat{y}_t = 33{,}2 + 0{,}32t.$$

Подставляя в данное уравнение значения $t = -5, -3, -1, 1, 3, 5$, получаем теоретические значения уровней (см. последнюю графу табл. 8.16).

Выравнивание по показательной функции $\hat{y}_t = a_0 a_1^t$. Как уже отмечалось, выравнивание по показательной функции проводится, в основном, когда уровни ряда меняются в геометрической прогрессии, т.е. цепные коэффициенты роста более или менее постоянны.

Нетрудно заметить, что логарифм показательной функции представляет собой линейную функцию $\lg \hat{y}_t = \lg a_0 + t \lg a_1$. Поэтому, если заменить уровни ряда их логарифмами, параметры a_0 и a_1 можно определить (через их логарифмы), решая следующую систему нормальных уравнений, полученную методом наименьших квадратов:

$$\begin{cases} n \lg a_0 + \lg a_1 \sum t = \sum \lg y, \\ \lg a_0 \sum t + \lg a_1 \sum t^2 = \sum t \lg y \end{cases} \tag{8.9}$$

или при счете от середины ряда (когда $\sum t = 0$)

$$\begin{cases} n \lg a_0 = \sum \lg y, \\ \lg a_1 \sum t^2 = \sum t \lg y. \end{cases} \tag{8.10}$$

Откуда

$$\lg a_0 = \frac{\sum \lg y}{n} \quad \text{и} \quad \lg a_1 = \frac{\sum t \lg y}{\sum t^2}.$$

Рассмотрим выравнивание по показательной функции на условном примере, характеризующем численность населения одного из городов России за 1996–2002 гг. (табл. 8.17).

Таблица 8.17

Выравнивание уровней ряда по показательной функции

Год	Численность населения на 1 января, тыс. чел. y	$\lg y$	t	t^2	$t \lg y$	$\lg \hat{y}_t$	Выравненные уровни $\hat{y}_t$
1996	108,3	2,0346	−3	9	−6,1038	2,0362	108,6
1997	111,8	2,0484	−2	4	−4,0968	2,0482	111,8
1998	115,1	2,0611	−1	1	−2,0611	2,0602	114,9
1999	118,5	2,0737	0	0	0	2,0722	118,1
2000	121,7	2,0853	1	1	2,0853	2,0842	121,4
2001	124,7	2,0955	2	4	4,1910	2,0962	124,8
2002	128,0	2,1072	3	9	6,3216	2,1082	128,3
$n = 7$	$\sum y =$ $= 828{,}1$	$\sum \lg y =$ $= 14{,}5058$	$\sum t = 0$	$\sum t^2 =$ $= 28$	$\sum t \lg y =$ $= 0{,}3362$	$\sum \lg \hat{y}_t =$ $= 14{,}5054$	$\sum \hat{y}_t =$ $= 827{,}9$

Логарифмируя уровни ряда y и ведя счет от середины ряда, рассчитываем в табл. 8.17 все необходимые суммы, на основе которых определяются сначала логарифмы параметров a_0 и a_1, а затем и сами параметры уравнения тренда:

$$\lg a_0 = \frac{\sum \lg y}{n} = \frac{14{,}5058}{7} = 2{,}0722, \text{ отсюда } a_0 = 118;$$

$$\lg a_1 = \frac{\sum t \lg y}{\sum t^2} = \frac{0{,}3362}{28} = 0{,}012, \text{ отсюда } a_1 = 1{,}028.$$

Следовательно,

$$\lg \hat{y}_t = 2{,}0722 + 0{,}012t,$$

а искомое уравнение

$$\hat{y}_t = 118 \cdot 1{,}028^t.$$

Для расчета выравненных уровней удобнее пользоваться формулой логарифмов, т.е. $\lg \hat{y}_t = 2{,}0722 + 0{,}012t$. Подставляя в эту формулу значения $t = -3, -2, -1, 0, 1, 2, 3$, находим $\lg \hat{y}_t$, а затем по таблицам логарифмов $\hat{y}_t$.

Так, для 1996 г.

$$\lg \hat{y}_t = 2{,}0722 + 0{,}012(-3) = 2{,}0362, \text{ отсюда } \hat{y}_t = 108{,}6;$$

для 1997 г.

$$\lg \hat{y}_t = 2{,}0722 + 0{,}012(-2) = 2{,}0482, \text{ отсюда } \hat{y}_t = 111{,}8 \text{ и т.д.}$$

Логарифмы выравненных уровней и сами уровни приведены в двух последних графах табл. 8.17. Судя по тому, что эмпирические уровни y весьма близки к теоретическим $\hat{y}_t$, можно сделать вывод о том, что показательная функция подходит для отражения данного тренда.

При выравнивании по показательной функции значение параметра a_1 практически характеризует средний темп роста исследуемого показателя в рассматриваемый период. Так, в нашем примере $a_1 = 1{,}028$ означает, что численность населения города за 1996–2002 гг. увеличивалась ежегодно в среднем в 1,028 раза (или, если перевести в проценты и вычесть 100%, можно сказать, что средний годовой темп прироста населения за указанный период составлял 2,8%).

Выравнивание по параболе 2-го порядка $\hat{y}_t = a_0 + a_1 t + a_2 t^2$. Парабола 2-го порядка как уравнение тренда может быть использована для выравнивания таких рядов, уровни которых сначала возрастают, а затем снижаются (или наоборот), или в рядах, где

вторые разности уровней примерно постоянны. Обычно при этом ориентируются по графическому изображению эмпирических данных.

Параметры искомого уравнения тренда a_0, a_1 и a_2 определяют, решая систему нормальных уравнений, полученных методом наименьших квадратов:

$$\begin{cases} na_0 + a_1 \sum t + a_2 \sum t^2 = \sum y, \\ a_0 \sum t + a_1 \sum t^2 + a_2 \sum t^3 = \sum yt, \\ a_0 \sum t^2 + a_1 \sum t^3 + a_2 \sum t^4 = \sum yt^2 \end{cases} \qquad (8.11)$$

или при счете от середины ряда (когда $\sum t = 0$)

$$\begin{cases} na_0 + a_2 \sum t^2 = \sum y, \\ a_1 \sum t^2 = \sum yt, \\ a_0 \sum t^2 + a_2 \sum t^4 = \sum yt^2. \end{cases} \qquad (8.12)$$

Выравнивание по параболе 2-го порядка проиллюстрируем на примере данных о числе незанятых граждан России, состоявших на учете в службах занятости в первой половине 1996 г. Исходные данные и расчет показателей, необходимых для решения системы нормальных уравнений (8.12), приведены в табл. 8.18.

Таблица 8.18

Выравнивание ряда динамики по параболе 2-го порядка (при счете времени от середины ряда)

Месяц	Число незанятых в 1996 г. на конец месяца, млн чел. y	t	t^2	t^4	yt	yt^2	Выравненные уровни $\hat{y}_t = 3{,}01 + 0{,}018t - 0{,}009t^2$
Январь	2,70	−5	25	625	−13,50	67,50	2,70
Февраль	2,87	−3	9	81	−8,61	25,83	2,88
Март	2,97	−1	1	1	−2,97	2,97	2,98
Апрель	3,06	1	1	1	3,06	3,06	3,02
Май	2,97	3	9	81	8,91	26,73	2,98
Июнь	2,87	5	25	625	14,35	71,75	2,88
$n = 6$	$\sum y =$ = 17,44	$\sum t =$ = 0	$\sum t^2 =$ = 70	$\sum t^4 =$ = 1414	$\sum yt =$ = 1,24	$\sum yt^2 =$ = 197,84	$\sum \hat{y}_t =$ = 17,44

Подставляем в систему (8.12) полученные в табл. 8.18 итоговые показатели:

$$\begin{cases} 6a_0 + 70a_2 = 17{,}44, \\ 70a_1 = 1{,}24, \\ 70a_0 + 1414a_2 = 197{,}84. \end{cases}$$

Отсюда

$$a_1 = \frac{1{,}24}{70} = 0{,}018.$$

Решая совместно первое и третье уравнения системы, находим $a_2 = -0{,}009$ и $a_0 = 3{,}01$.

Следовательно, искомое уравнение тренда

$$\hat{y}_t = 3{,}01 + 0{,}018t - 0{,}009t^2.$$

Подставляя в него значения $t = -5, -3, -1, 1, 3, 5$, определяем теоретические (выравненные) уровни (см. последнюю графу табл. 8.18). Сравнивая их с эмпирическими уровнями, отмечаем, что они почти полностью совпадают, т.е. парабола 2-го порядка — вполне адекватная функция для отражения основной тенденции (тренда) изменения уровней за исследуемый период, что подтверждает и рис. 8.7.

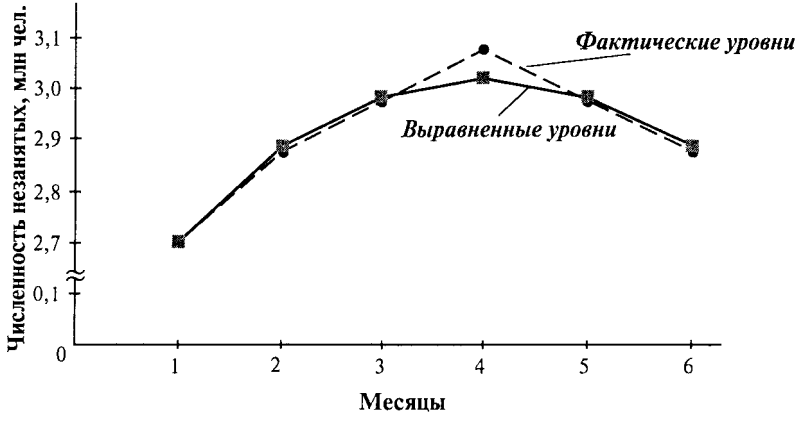


Рис. 8.7. Фактические и выравненные уровни числа незанятых, состоявших на учете в службах занятости в первой половине 1996 г.

Парабола 2-го порядка может иметь и иной вид. Так, если значения уровней сначала убывают, а затем возрастают, то вместо выпуклости кривая будет иметь вогнутость. Рассмотрим выравнивание по такой кривой данных о добыче угля в России за 1995–2001 гг. Исходные данные и расчет показателей, необходимых для решения системы нормальных уравнений (8.12), приведены в табл. 8.19.

Таблица 8.19

Выравнивание ряда динамики по параболе 2-го порядка (при счете времени от середины ряда)

Год	Добыча угля, млн т y	t	t^2	t^4	yt	yt^2	Выравненные уровни $\hat{y}_t = 241{,}6 + t + 3t^2$
1995	263	−3	9	81	−789	2367	265,6
1996	257	−2	4	16	−514	1028	251,6
1997	245	−1	1	1	−245	245	243,6
1998	232	0	0	0	0	0	241,6
1999	250	1	1	1	250	250	245,6
2000	258	2	4	16	516	1032	255,6
2001	270	3	9	81	810	2430	271,6
$n = 7$	$\sum y =$ $= 1775$	$\sum t =$ $= 0$	$\sum t^2 =$ $= 28$	$\sum t^4 =$ $= 196$	$\sum yt =$ $= 28$	$\sum yt^2 =$ $= 7352$	$\sum \hat{y}_t =$ $= 1775{,}2$

Подставляем в систему (8.12) полученные в табл. 8.19 итоговые показатели:

$$\begin{cases} 7a_0 + 28a_2 = 1775, \\ 28a_1 = 28, \\ 28a_0 + 196a_2 = 7352. \end{cases}$$

Решив систему, имеем $a_1 = 1$, $a_2 = 3$, $a_0 = 241{,}57 \approx 241{,}6$.

Следовательно, искомое уравнение тренда добычи угля в России за 1995–2001 гг. (при отсчете времени от середины ряда) составило

$$\hat{y}_t = 241{,}6 + t + 3t^2.$$

Подставляя в него значения $t = -3, -2, -1, 0, 1, 2, 3$, находим теоретические (выравненные) уровни (см. последнюю графу табл. 8.19). На рис. 8.8 приведены фактические и выравненные уровни добычи угля в России за 1995–2001 гг.

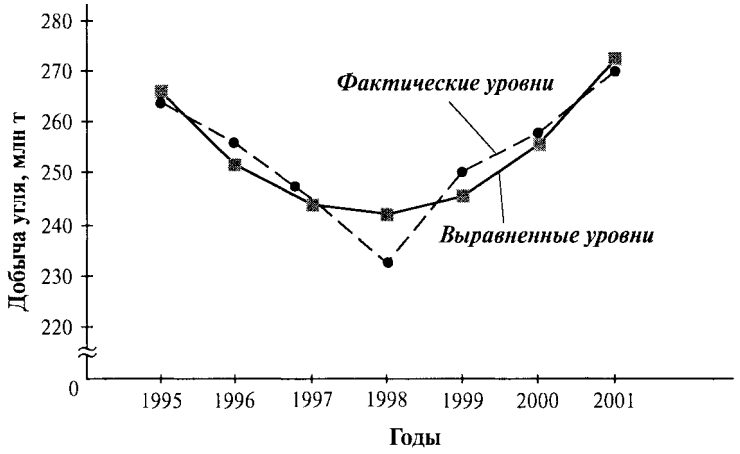


Рис. 8.8. Фактические и выравненные уровни добычи угля в России за 1995–2001 гг.

Выравнивание по гиперболе $\hat{y}_t = a_0 + a_1 \frac{1}{t}$. Гипербола как уравнение тренда может быть использована для выравнивания таких рядов, уровни которых сначала резко снижаются, а затем это снижение замедляется.

Параметры искомого уравнения в виде гиперболы, т.е. a_0 и a_1, определяем, решая следующую систему нормальных уравнений, полученных методом наименьших квадратов:

$$\begin{cases} na_0 + a_1 \sum \frac{1}{t} = \sum y, \\ a_0 \sum \frac{1}{t} + a_1 \sum \left(\frac{1}{2}\right)^2 = \sum \frac{y}{t}. \end{cases} \qquad (8.13)$$

В качестве примера выравнивания по гиперболе могут быть использованы приводимые в табл. 8.20 данные о производстве шерсти (в физическом весе) в России за период 1995–2001 гг.

Учитывая характер изменения уровней ряда, выдвигаем гипотезу о гиперболическом тренде $\hat{y}_t = a_0 + a_1 \frac{1}{t}$. Необходимые для решения системы уравнений (8.13) суммы $\sum \frac{1}{t}$, $\sum \left(\frac{1}{t}\right)^2$, $\sum y$, $\sum \frac{y}{t}$ рассчитаны в табл. 8.20. Подставляя их в систему (8.13), получаем

$$\begin{cases} 7a_0 + 2{,}59a_1 = 399, \\ 2{,}59a_0 + 1{,}5099a_1 = 184{,}2. \end{cases}$$

Таблица 8.20

Выравнивание данных о производстве шерсти в России по гиперболе

Год	Производство шерсти, млн т y	t	$\frac{1}{t}$	$\left(\frac{1}{t}\right)^2$	$\frac{y}{t}$	Выравненные уровни $\hat{y}_t = 32{,}5 + \frac{66{,}3}{t}$
1995	93	1	1,00	1,0000	93,0	98,8
1996	77	2	0,50	0,2500	38,5	65,5
1997	61	3	0,33	0,1089	20,3	54,5
1998	48	4	0,25	0,0625	12,0	49,0
1999	40	5	0,20	0,0400	8,0	45,8
2000	40	6	0,17	0,0289	6,7	43,5
2001	40	7	0,14	0,0196	5,7	42,0
$n = 7$	$\sum y = 399$		$\sum \frac{1}{t} = 2{,}59$	$\sum \left(\frac{1}{t}\right)^2 = 1{,}5099$	$\sum \frac{y}{t} = 184{,}2$	$\sum \hat{y}_t = 399{,}2$

Решив систему, имеем $a_0 = 32{,}5$ и $a_1 = 66{,}3$. Отсюда уравнение тренда

$$\hat{y}_t = 32{,}5 + \frac{66{,}3}{t}.$$

Подставляя в данное уравнение значения $t = 1, 2, 3, 4, 5, 6, 7$, определяем теоретические (выравненные) уровни ряда. Они показаны в последней графе табл. 8.20 и на рис. 8.9. (Расхождение между $\sum y$ и $\sum \hat{y}_t$ на 0,2 вызвано округлениями при расчетах параметров a_0 и a_1 и самих $\hat{y}_t$.)

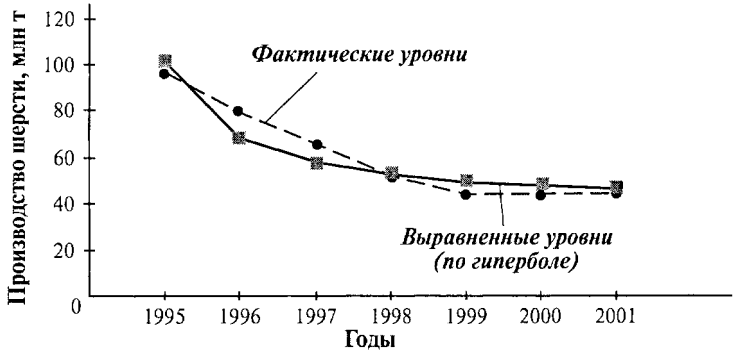


Рис. 8.9. Динамика производства шерсти в России

Выравнивание с помощью ряда Фурье. Особое место в аналитическом выравнивании динамических рядов занимает выравнивание с помощью ряда Фурье, в котором уровни можно выразить как функцию времени следующим уравнением:

$$\hat{y}_t = a_0 + \sum_{k=1}^{m} (a_k \cos kt + b_k \sin kt). \qquad (8.14)$$

Выравнивание по формуле (8.14) рекомендуется проводить в тех случаях, когда в эмпирическом ряду наблюдается периодичность изменения уровней. В этом случае периодические колебания уровней динамического ряда можно представить в виде синусоидальных колебаний. Поскольку последние представляют собой гармонические колебания, то синусоиды, полученные при выравнивании по ряду Фурье, называют *гармониками* различных порядков. Показатель k в уравнении (8.14) определяет число гармоник. Обычно при выравнивании по ряду Фурье рассчитывают несколько гармоник (чаще не более 4) и затем уже определяют, с каким числом гармоник ряд Фурье наилучшим образом отражает изменения уровней ряда.

При выравнивании по ряду Фурье периодические колебания уровней динамического ряда представлены в виде суммы нескольких синусоид (гармоник), наложенных друг на друга.

Так, при $k = 1$ ряд Фурье будет иметь вид

$$\hat{y}_t = a_0 + a_1 \cos t + b_1 \sin t,$$

а при $k = 2$, соответственно,

$$\hat{y}_t = a_0 + a_1 \cos t + b_1 \sin t + a_2 \cos 2t + b_2 \sin 2t$$

и т.д.

Параметры уравнения теоретических уровней, определяемого рядом Фурье, находят, как и в других случаях, методом наименьших квадратов. Приведем без вывода формулы, используемые для исчисления параметров ряда Фурье:

$$a_0 = \frac{\sum y}{n}; \quad a_k = \frac{2\sum y \cos kt}{n}; \quad b_k = \frac{2\sum y \sin kt}{n}. \qquad (8.15)$$

Последовательные значения t обычно определяются от 0 с увеличением (приростом), равным $\frac{2\pi}{n}$, где n – число уровней эмпирического ряда.

Например, при $n = 10$ временны́е точки t можно записать следующим образом:

$$0;\ \frac{2\pi}{10}\cdot 1;\ \frac{2\pi}{10}\cdot 2;\ \frac{2\pi}{10}\cdot 3;\ \frac{2\pi}{10}\cdot 4;\ \frac{2\pi}{10}\cdot 5;\ \frac{2\pi}{10}\cdot 6;\ \frac{2\pi}{10}\cdot 7;\ \frac{2\pi}{10}\cdot 8;\ \frac{2\pi}{10}\cdot 9,$$

или (после сокращения)

$$0;\ \frac{\pi}{5};\ \frac{2\pi}{5};\ \frac{3\pi}{5};\ \frac{4\pi}{5};\ \pi;\ \frac{6\pi}{5};\ \frac{7\pi}{5};\ \frac{8\pi}{5};\ \frac{9\pi}{5}.$$

При $n = 12$ значения t, соответственно, будут

$$0;\ \frac{\pi}{6};\ \frac{\pi}{3};\ \frac{\pi}{2};\ \frac{2\pi}{3};\ \frac{5\pi}{6};\ \pi;\ \frac{7\pi}{6};\ \frac{4\pi}{3};\ \frac{3\pi}{2};\ \frac{5\pi}{3};\ \frac{11\pi}{6}.$$

Значения $\sin kt$ и $\cos kt$ удобно расположить в таблице. Например, в табл. 8.21 приведены значения $\sin kt$ и $\cos kt$ ($k = 1$ и $k = 2$) для $n = 12$.

Выравнивание по ряду Фурье часто дает хорошие результаты в рядах, содержащих сезонную волну.

Проиллюстрируем выравнивание по ряду Фурье на условном примере данных о продаже зимней одежды в одном из районов в 2002 г. В табл. 8.22 приведены исходные данные (графы 1–3) и расчет показателей, необходимых для получения уравнений первой и второй гармоник ($k = 1$ и $k = 2$). Итак,

$$a_0 = \frac{\sum y}{n} = \frac{552}{12} = 46,$$

$$a_1 = \frac{2\sum y\cos t}{n} = \frac{\sum y\cos t}{6} = \frac{-66{,}24}{6} = -11{,}04,$$

$$b_1 = \frac{2\sum y\sin t}{n} = \frac{\sum y\sin t}{6} = \frac{34{,}43}{6} = 5{,}74.$$

Отсюда

$$_1\hat{y}_t = 46 - 11{,}04\cos t + 5{,}74\sin t.$$

Подставляя в данное уравнение значения $\cos t$ и $\sin t$ (из табл. 8.21), получаем теоретические значения объема продажи зимней одежды по месяцам, показанные в графе 6 табл. 8.22. Как видно, теоретические значения $_1\hat{y}_t$, рассчитанные по уравнению первой гармоники, заметно отличаются от эмпирических y. Поэтому попытаемся определить уравнение второй гармоники, т.е.

$$_2\hat{y}_t = a_0 + a_1\cos t + b_1\sin t + a_2\cos 2t + b_2\sin 2t.$$

Таблица 8.21

Значения sin *kt* и cos *kt* (для *n* = 12)

t	cos t	cos $2t$	sin t	sin $2t$
0	1	1	0	0
$\pi/6$	0,866	0,5	0,5	0,866
$\pi/3$	0,5	–0,5	0,866	0,866
$\pi/2$	0	–1	1	0
$2\pi/3$	–0,5	–0,5	0,866	–0,866
$5\pi/6$	–0,866	0,5	0,5	–0,866
π	–1	1	0	0
$7\pi/6$	–0,866	0,5	–0,5	0,866
$4\pi/3$	–0,5	–0,5	–0,866	0,866
$3\pi/2$	0	–1	–1	0
$5\pi/3$	0,5	–0,5	–0,866	–0,866
$11\pi/6$	0,866	0,5	–0,5	–0,866

Таблица 8.22

Выравнивание по ряду Фурье

Месяц	t	Продано зимней одежды, тыс. руб. y	$y \cos t$	$y \sin t$	${}_1\hat{y}_t$	$y \cos 2t$	$y \sin 2t$	${}_2\hat{y}_t$
1	2	3	4	5	6	7	8	9
1	0	37	37,0	0	35	37	0	37,9
2	$\pi/6$	40	34,64	20,0	39,3	20	34,64	39,6
3	$\pi/3$	44	22,0	38,1	45,5	–22	38,1	43,0
4	$\pi/2$	52	0	52,0	51,7	–52	0	43,8
5	$2\pi/3$	46	–23,0	39,84	56,5	–23	–39,84	56,2
6	$5\pi/6$	70	–60,62	35,0	58,4	35	–60,6	61,0
7	π	60	–60,0	0	57,0	60	0	59,9
8	$7\pi/6$	48	–41,57	–24,0	52,7	24	41,57	53,0
9	$4\pi/3$	46	–23,0	–39,84	46,5	–23	39,84	44,0
10	$3\pi/2$	38	0	–38,0	39,3	–38	0	36,4
11	$5\pi/3$	36	18,0	–31,17	35,5	–18	–31,18	35,2
12	$11\pi/6$	35	30,31	–17,5	33,6	17,5	–30,31	36,2
$n = 12$	Σ	552	–66,24	34,43	551,0	17,5	–7,78	551,2

Расчеты, необходимые для нахождения параметров a_2 и b_2, также приведены в табл. 8.22.

Итак,

$$a_2 = \frac{\sum y \cos 2t}{6} = \frac{17,5}{6} = 2,9; \quad b_2 = \frac{\sum y \sin 2t}{6} = \frac{-7,78}{6} = -1,3.$$

Отсюда уравнение второй гармоники

$$_2\hat{y}_t = 46 - 11{,}04\cos t + 5{,}74\sin t + 2{,}9\cos 2t - 1{,}3\sin 2t.$$

Подставляя в данное уравнение значения cos t, sin t, cos $2t$, sin $2t$ (см. табл. 8.21), получаем теоретические значения $_2\hat{y}_t$ (см. последнюю графу табл. 8.22).

Нетрудно заметить, что теоретические значения $_2\hat{y}_t$, рассчитанные по уравнению второй гармоники, более близки к эмпирическим уровням, чем $_1\hat{y}_t$. Об этом свидетельствует и сумма квадратов отклонений теоретических значений от эмпирических: $\sum(y - {}_1\hat{y}_t)^2 =$ $= 286{,}88$, $\sum(y - {}_2\hat{y}_t)^2 = 232{,}22$.

Аналогично рассчитывают параметры уравнения с применением третьей и четвертой гармоник и проверяют близость теоретических значений к эмпирическим.

В заключение отметим, что выравнивание играет важную роль в анализе рядов динамики. Правильный подбор типа кривой для определения тренда представляет не только теоретический, но и практический интерес, в частности при прогнозировании.

Однако обработка рядов динамики (любым способом) дает эффект только при достаточно большом числе уровней ряда.

Следует отметить, что найденные уравнения тренда часто используют для прогнозирования *методом экстраполяции*, т.е. распространения в будущее закономерности развития, выявленной в прошлом, в исследованном периоде. Однако экстраполировать ряд по уравнению тренда можно только тогда, когда есть уверенность в том, что выявленная и описанная уравнением тренда закономерность развития устойчива и сохранится в будущем, т.е. что условия, в которых происходили изучаемые явления в определенном периоде в прошлом, стабильны и предположительно не изменятся в ближайшем будущем, на которое экстраполируется ряд.

8.6. Измерение колеблемости в рядах динамики

Как уже отмечалось, уровни ряда динамики формируются под влиянием различных взаимодействующих факторов, одни из которых определяют тенденцию развития, а другие – колеблемость (вариацию).

Изучение колеблемости в рядах динамики как предмета исследования часто является самостоятельной задачей в статистике.

Колебания уровней ряда могут носить разный характер. Исследователи временны́х рядов всегда пытались классифицировать факторы, вызывающие те или иные колебания, и, соответственно, выделить типы колебаний. Большинство авторов чаще всего выделяют (наряду с трендом) *циклические* (долгопериодические), *сезонные* (обнаруживаемые в рядах, где данные приведены за кварталы или месяцы) и *случайные* колебания.

Для измерения колеблемости уровней в рядах динамики могут использоваться показатели, аналогичные показателям вариации признака, рассмотренным в главе 5:

- размах, или амплитуда, отклонений отдельных уровней от их средней (по модулю) или от тренда;
- среднее линейное отклонение d (по модулю) отдельных уровней от общей средней или от тренда;
- среднее квадратическое отклонение σ отдельных уровней от общей средней или от тренда;
- относительный показатель колеблемости уровней, аналогичный коэффициенту вариации,

$$V = \frac{\sigma}{\bar{y}} 100\%.$$

При этом важно учитывать, относительно какого показателя (уровня) исследуется колеблемость. Например, можно исследовать колеблемость вокруг среднего уровня ряда $\bar{y}$, который на графике выразится прямой, параллельной оси абсцисс. А можно исследовать колебания уровней вокруг линии тренда (или скользящей средней). Различный характер таких колебаний наглядно виден на графике (рис. 8.10).

Рассмотрим традиционный случай расчета среднего квадратического отклонения отдельных уровней y_i от общего среднего уровня ряда $\bar{y}$:

$$\sigma = \sqrt{\frac{\sum (y_i - \bar{y})^2}{n}}. \qquad (8.16)$$

В данном случае величина $\sum (y_i - \bar{y})^2$ характеризует сумму квадратов отклонений фактических уровней от общей средней за счет в с е х факторов, формирующих уровни, как основных, определяющих тренд, так и случайных.

Задача исследования колебаний уровней в рядах динамики сводится к разложению общей колеблемости на составляющие и

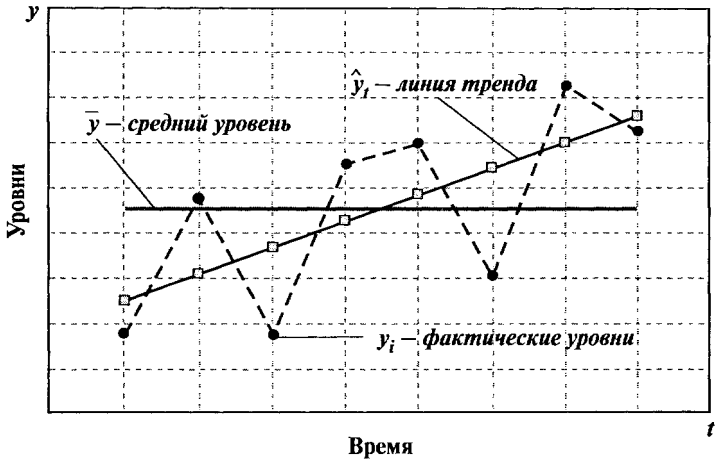


Рис. 8.10. Колебания фактических уровней y_i относительно среднего уровня $\bar{y}$ и линии тренда $\hat{y}_t$

выделению именно тех колебаний, которые интересуют исследователя.

Для решения этой задачи требуется разложить общую сумму квадратов отклонений от средней $\sum(y_i - \bar{y})^2$ на составляющие.

Имея фактические (эмпирические) уровни ряда y_i и уровни, выравненные по определенному тренду, $\hat{y}_t$, можно рассчитать следующие суммы квадратов отклонений:

1) $\sum(y_i - \bar{y})^2$ – общую сумму квадратов отклонений фактических уровней от их общей средней;
2) $\sum(\hat{y}_t - \bar{y})^2$ – сумму квадратов отклонений за счет тренда (за счет фактора времени);
3) $\sum(y_i - \hat{y}_t)^2$ – сумму квадратов отклонений за счет случайных факторов.

Согласно правилу сложения вариации и правилу сложения дисперсий первая сумма равна сумме двух последних:

$$\sum(y_i - \bar{y})^2 = \sum(\hat{y}_t - \bar{y})^2 + \sum(y_i - \hat{y}_t)^2.$$

Отсюда, пользуясь величиной $\sum(\hat{y}_t - \bar{y})^2$, можно определить среднее квадратическое отклонение уровней ряда за счет тренда (фактора времени).

В свою очередь, используя $\sum(y_i - \hat{y}_t)^2$, можно определить среднее квадратическое отклонение уровней за счет случайных факторов. Чем меньше эта сумма, тем ближе фактические уровни к линии тренда. Это означает, что линия тренда подобрана удачно, т.е. адекватна эмпирическим данным. Поэтому среднее квадратическое отклонение, рассчитанное на основе данной суммы квадратов отклонений от тренда, одновременно рассматривается как *средняя квадратическая ошибка уравнения тренда.* При этом поскольку разные уравнения тренда имеют различное число параметров m, средняя квадратическая ошибка уравнения тренда σ (или $\sigma_{ост}$) рассчитывается путем деления $\sum(y_i - \hat{y}_t)^2$ не на n, а на $(n - m)$, т.е. на число степеней свободы:

$$\sigma = \sqrt{\frac{\sum(y_i - \hat{y}_t)^2}{n - m}}. \qquad (8.17)$$

Если уровни ряда являются месячными или квартальными показателями и несут на себе влияние сезонности, то в общей сумме квадратов отклонений уровней ряда от их средней $\sum(y_i - \bar{y})^2$ можно выделить также составляющую, характеризующую сезонные колебания.

8.7. Выявление и измерение сезонных колебаний

В рядах динамики, уровни которых являются месячными или квартальными показателями, наряду со случайными колебаниями часто наблюдаются *сезонные колебания*, под которыми понимается периодически повторяющиеся из года в год повышение и снижение уровней в отдельные месяцы или кварталы.

Сезонным колебаниям подвержены внутригодовые уровни многих показателей. Так, расход электроэнергии в летние месяцы значительно меньше, чем в зимние. Потребление мяса больше в зимние месяцы, производство некоторых видов продуктов (сахара, растительного масла и др.), связанных с переработкой сельскохозяйственной продукции, увеличивается в месяцы, непосредственно следующие за окончанием уборки урожая, рыночные цены на овощи в отдельные месяцы далеко не одинаковы и т.д.

При графическом изображении таких рядов сезонные колебания проявляются в повышении и снижении уровней в определенные месяцы (или кварталы).

В качестве иллюстрации рядов с сезонными колебаниями могут служить данные о производстве растительного масла в России за 1992–1993 гг. по месяцам (табл. 8.23) и их графическое изображение (рис. 8.11).

Таблица 8.23

Производство растительного масла в России в 1992—1993 гг. по месяцам, тыс. т

Год	Месяц											
	1	2	3	4	5	6	7	8	9	10	11	12
1992	109,5	102,7	86,6	82,3	76,6	70,0	57,6	24,5	36,3	70,7	95,2	104,5
1993	97,6	95,5	114,2	101,3	105,6	94,6	75,2	38,6	38,9	78,7	96,5	111,0

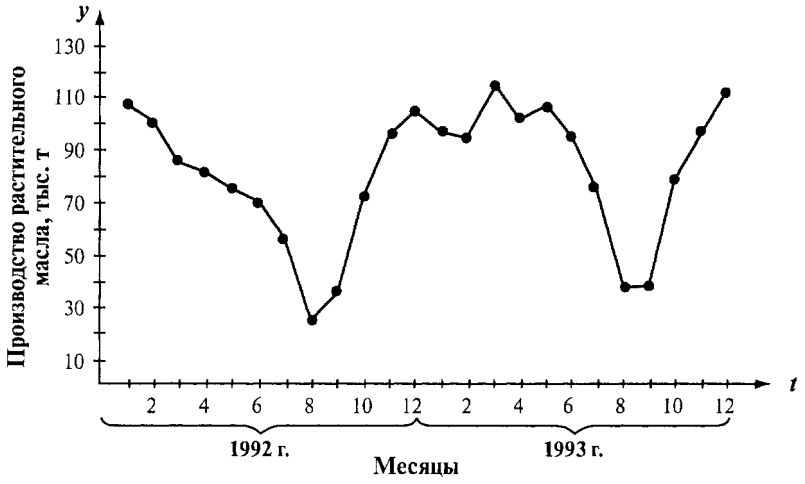


Рис. 8.11. Динамика производства растительного масла в России за 1992—1993 гг. по месяцам

Вместо месячных показателей могут быть квартальные. Если колебания не случайны, они сохраняются и в квартальных уровнях, как это показано в табл. 8.24, где месячные данные нашего примера преобразованы в квартальные, и на рис. 8.12.

Таблица 8.24

Производство растительного масла в России в 1992—1993 гг. по кварталам

Год	1992				1993			
Квартал	I	II	III	IV	I	II	III	IV
Произведено тыс. т	298,8	228,9	118,4	270,4	307,3	301,5	152,7	286,2

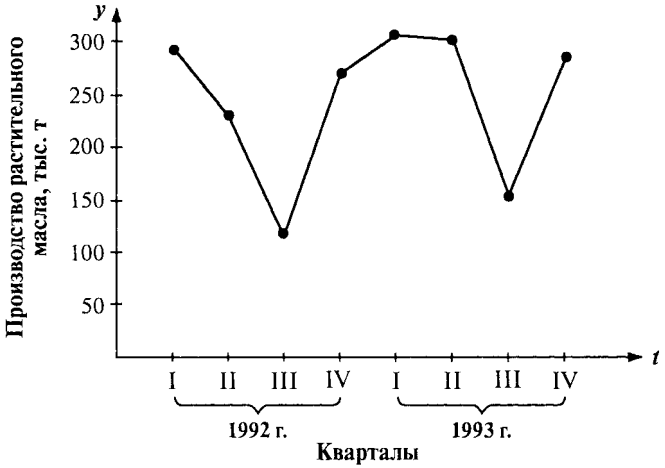


Рис. 8.12. Динамика производства растительного масла в России за 1992—1993 гг. по кварталам

Наблюдение за сезонными колебаниями позволяет, с одной стороны, устранить их там, где они нежелательны (например, можно более равномерно использовать в течение года строительных рабочих), с другой стороны, решить ряд практических задач (например, определить потребности в рабочей силе, оборудовании и сырье в тех отраслях, где влияние сезонности велико).

При изучении рядов динамики, содержащих «сезонную волну», ее выделяют из общей колеблемости уровней и измеряют. Существует ряд методов для решения этой задачи. Все они основаны на сравнении фактических уровней каждого месяца (или квартала) со средним уровнем, предполагающим равномерное распределение годового показателя по месяцам (или кварталам), либо со сглаженными скользящими средними или выравненными по уравнению тренда. При этом для измерения «сезонной волны» рассчитывают либо абсолютные разности (отклонения) фактических уровней от среднего уровня (или от выравненных), либо отношения месячных уровней к среднему месячному уровню за год, так называемые ***индексы сезонности***:

$$I_{\text{сез}} = \frac{y_i}{\bar{y}} 100\%. \qquad (8.18)$$

В табл. 8.25 показан расчет индексов сезонности и абсолютных отклонений уровней от среднего на примере данных о производстве растительного масла в России в 1992 г.

Таблица 8.25

Сезонные колебания производства растительного масла в России в 1992 г.

Месяц	Производство масла, тыс. т y_i	Индекс сезонности, % к среднему месячному уровню $\frac{y_i}{\bar{y}}100\%$	Абсолютное отклонение от среднего месячного уровня $y_i - \bar{y}$	Абсолютное отклонение, % к среднему месячному уровню $(y_i - \bar{y})/\bar{y}$	$(I_{ссз} - 100\%)^2$	$(y_i - \bar{y})^2$
1	2	3	4	5	6	7
Январь	109,5	143,4	33,125	43,4	1883,56	1097,266
Февраль	102,7	134,5	26,325	34,5	1190,25	693,006
Март	86,6	113,4	10,225	13,4	179,56	104,551
Апрель	82,3	107,8	5,925	7,8	60,84	35,106
Май	76,6	100,3	0,225	0,3	0,09	0,051
Июнь	70,0	91,6	−6,375	−8,4	70,56	40,641
Июль	57,6	75,4	−18,375	−24,6	605,16	352,501
Август	24,5	32,1	−51,875	−67,9	4610,41	2691,018
Сентябрь	36,3	47,5	−40,075	−52,5	2756,25	1606,006
Октябрь	70,7	92,6	−5,675	−7,4	54,76	32,206
Ноябрь	95,2	124,6	18,825	24,6	605,16	354,381
Декабрь	104,5	136,8	28,125	36,8	1354,24	791,016
Итого	916,5	1200	0	0	13370,84	7797,747

Средний месячный уровень за год

$$\bar{y} = \frac{\sum y_i}{n} = \frac{916,5}{12} = 76,375 \text{ тыс. т.}$$

В графе 3 табл. 8.25 индексы сезонности рассчитаны как процентное отношение фактического уровня каждого месяца y_i к среднему месячному $\bar{y}$ за год, т.е. по формуле (8.18). В графе 4 приведены абсолютные отклонения уровней каждого месяца от среднего месячного за год, а в графе 5 – эти же отклонения в процентах к среднему месячному уровню. Нетрудно видеть, что данные графы 5 представляют собой разность между индексом сезонности и 100%. Другими словами, независимо от того, как учитываются различия в месячных уровнях, измерение сезонности в конечном счете сводится к расчету индексов сезонности.

Графическое изображение индексов сезонности (рис. 8.13) наглядно показывает форму, характер «сезонной волны» относительно среднего месячного уровня за год, принимаемого за 100%.

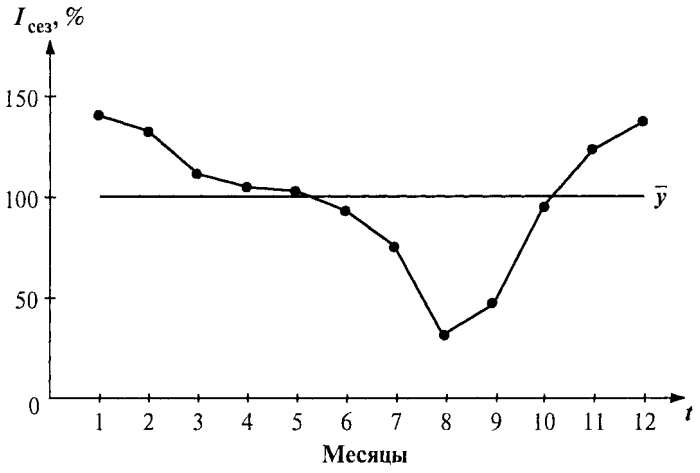


Рис. 8.13. Индексы сезонности производства растительного масла в России в 1992 г.

Данные табл. 8.25 и рис. 8.13 показывают, что минимальный объем производства растительного масла в 1992 г. приходился на август, а максимальный – на январь.

Для характеристики силы (меры) колеблемости уровней динамического ряда из-за сезонной неравномерности часто предлагается использовать ***среднее квадратическое отклонение индексов сезонности*** (в процентах) от 100%, т.е.

$$\sigma_{сез} = \sqrt{\frac{\sum (I_{сез} - 100\%)^2}{n}}. \qquad (8.19)$$

В нашем примере сумма $\sum (I_{сез} - 100\%)^2 = 13370{,}84$ рассчитана в графе 6 (см. табл. 8.25). Тогда

$$\sigma_{сез} = \sqrt{\frac{13370{,}84}{12}} = 33{,}38\%.$$

Этот же результат можно получить и по-другому, как коэффициент вариации (колеблемости):

$$V = \frac{\sigma}{\bar{y}} 100\%,$$

где $\sigma = \sqrt{\frac{\sum(y_i - \bar{y})^2}{n}}$ – среднее квадратическое отклонение уровней ряда.

В табл. 8.25 сумма квадратов отклонений от среднего уровня $\sum(y_i - \bar{y})^2$ рассчитана в графе 7, среднее значение уровня $\bar{y} = 76{,}375$. Отсюда

$$\sigma = \sqrt{\frac{\sum(y_i - \bar{y})^2}{n}} = \sqrt{\frac{7797{,}747}{12}} = 25{,}5 \text{ тыс. т},$$

$$V = \frac{\sigma}{\bar{y}} 100\% = \frac{25{,}5}{76{,}375} 100 = 33{,}38\%,$$

т.е. результаты двух показателей ($\sigma_{сез}$ и V) идентичны.

Важно отметить, что в формуле (8.16) сумма квадратов отклонений месячных уровней от общего среднего уровня, т.е. $\sum(y_i - \bar{y})^2$, измеряет колеблемость за счет всех факторов, а не только за счет сезонной неравномерности, поэтому, пользуясь ею для измерения колеблемости ряда из-за сезонной неравномерности, не следует переоценивать ее значение.

Конечно, там, где колебания в основном определены влиянием сезонности, расчет по формуле (8.16) их учитывает. Однако при этом не исключается и влияние случайных колебаний.

В рассмотренном методе расчета индексов сезонности $I_{сез} = \frac{y_i}{\bar{y}} 100\%$ использовались данные одного года. Этот метод довольно прост, но в силу элемента случайности месячные данные одного года недостаточно надежны для определения меры сезонных колебаний. Поэтому рекомендуется пользоваться месячными (или квартальными) данными за ряд лет (в основном за 3 года, хотя не исключена возможность использования данных за 2 года, а также за период более 3 лет).

Расчет индексов сезонности за ряд лет

При наличии месячных данных за ряд лет расчет индексов сезонности можно осуществить по-разному. Рассмотрим несколько способов.

1. По данным ряда лет рассчитывается среднее значение уровня для каждого месяца $\bar{y}_i$, а также средний месячный уровень за весь период $\bar{y}$. Затем определяются индексы сезонности как процентное отношение средних уровней для каждого ме-

сяца к общему среднему месячному уровню всего ряда (за все годы), т.е. по формуле

$$I_{\text{сез}i} = \frac{\bar{y}_i}{\bar{y}} 100\%. \qquad (8.20)$$

Например, по данным табл. 8.23 за 2 года получим следующие средние уровни по месяцам:

в январе $\bar{y}_1 = (109{,}5 + 97{,}6)/2 = 103{,}55$ тыс. т;

в феврале $\bar{y}_2 = (102{,}7 + 95{,}5)/2 = 99{,}1$ тыс. т;

в марте $\bar{y}_3 = (86{,}6 + 114{,}2)/2 = 100{,}4$ тыс. т и т.д.

Средний месячный уровень за 2 года

$$\bar{y} = \frac{\sum_{1}^{24} y_i}{24} = 82{,}0175 \text{ тыс. т} \cong 82 \text{ тыс. т.}$$

Отсюда индексы сезонности $I_{\text{сез}i} = \frac{\bar{y}_i}{\bar{y}} 100\%$:

в январе $I_{\text{сез}1} = \frac{103{,}55}{82} 100 = 126{,}3\%$;

в феврале $I_{\text{сез}2} = \frac{99{,}1}{82} 100 = 120{,}5\%$;

в марте $I_{\text{сез}3} = \frac{100{,}4}{82} 100 = 122{,}5\%$ и т.д.

Данный метод используется в основном в тех случаях, когда уровни одноименных месяцев в разные годы отличаются незначительно.

2. Если же наблюдается тенденция к увеличению или снижению уровней из года в год, то эффективнее рассчитывать индексы сезонности по следующей схеме.

Для каждого года отдельно рассчитываются индексы сезонности по формуле (8.18), т.е. как $I_{\text{сез}i} = \frac{y_i}{\bar{y}} 100\%$, а затем из индексов одноименных месяцев находится средняя арифметическая.

Покажем этот метод на примере данных табл. 8.23. Рассчитаем индексы сезонности для 1993 г. так же, как для 1992 г.

В 1993 г. средний месячный уровень $\bar{y}$ составил 87,3 тыс. т. Отсюда месячные индексы сезонности 1993 г.:

в январе $\frac{97,6}{87,3}100 = 111,8\%$;

в феврале $\frac{95,5}{87,3}100 = 109,4\%$;

в марте $\frac{114,2}{87,3}100 = 130,8\%$ и т.д.

Зная месячные индексы сезонности за 1992 г. (см. табл. 8.25) и за 1993 г., определяем из них для каждого месяца среднюю арифметическую, которую и принимаем в качестве обобщенной меры сезонных колебаний:

в январе $(143,4 + 111,8)/2 = 127,6\%$;
в феврале $(134,5 + 109,4)/2 = 121,95\%$;
в марте $(113,4 + 130,8)/2 = 122,1\%$ и т.д.

3. Следующий прием измерения сезонных колебаний при наличии тренда в данных за ряд лет основан на сопоставлении фактических месячных (или квартальных) уровней либо со сглаженными методом скользящей средней, либо с выравненными по определенной аналитической формуле.

В первом случае месячные данные за ряд лет сглаживаются 12-месячной скользящей средней (при квартальных данных — 4-квартальной скользящей средней). Затем фактические уровни каждого месяца (или квартала) выражают в процентах к скользящей средней.

На основе таких отношений (индексов сезонности) за ряд лет находится средняя арифметическая для каждого месяца (или квартала). Полученные усредненные индексы сезонности и являются искомыми, характеризующими «сезонную волну».

Аналогично рассчитываются индексы сезонности и во втором случае, на основе сопоставления фактических уровней с выравненными по аналитической формуле. Здесь та же последовательность расчетов с той лишь разницей, что вместо сглаженных скользящих средних сначала находится уравнение тренда и по нему рассчитываются выравненные (теоретические) уровни $\hat{y}_t$. Затем определяется отношение фактических уровней к выравненным, т.е. рассчитываются индексы сезонности для каждого месяца (или квартала):

$$I_{\text{сез}i} = \frac{y_i}{\hat{y}_t}. \qquad (8.21)$$

Поскольку за n лет отдельные месяцы повторяются, значения месячных индексов сезонности для отдельных лет усредняются.

Рассмотрим этот метод расчета индексов сезонности по отношению к тренду на условном примере динамики объема строительных работ в одном из регионов по кварталам за 3 года. Исходные данные и последующие расчеты показаны в табл. 8.26.

Предполагая, что фактические уровни y_i (см. графу 2 табл. 8.26) имеют линейный тренд $\hat{y}_t = a_0 + a_1 t$, и ведя счет времени от начала ряда ($t = 1, 2, 3, \ldots$), подсчитываем все необходимые суммы в таблице (графы 2–5). По этим суммам и определяем параметры a_0 и a_1, решая систему нормальных уравнений

$$\begin{cases} na_0 + a_1 \sum t = \sum y, \\ a_0 \sum t + a_1 \sum t^2 = \sum yt, \end{cases} \quad \text{т.е.} \quad \begin{cases} 12a_0 + 78a_1 = 198, \\ 78a_0 + 650a_1 = 1388{,}8, \end{cases}$$

или сразу по формулам

$$a_1 = \frac{\sum yt - \frac{\sum y}{n}\sum t}{\sum t^2 - \frac{(\sum t)^2}{n}} = \frac{1388{,}8 - \frac{198}{12}78}{650 - \frac{78^2}{12}} = 0{,}71,$$

$$a_0 = \frac{\sum y}{n} - a_1 \frac{\sum t}{n} = \frac{198}{12} - 0{,}71\frac{78}{12} = 11{,}9.$$

Отсюда уравнение тренда

$$\hat{y}_t = 11{,}9 + 0{,}71t.$$

Подставляя в него значения $t = 1, 2, \ldots, 12$, находим выравненные уровни $\hat{y}_t$ (с точностью до одной десятой) (см. графу 6 табл. 8.26).

Отношения фактических уровней y_i (графа 2) к выравненным (теоретическим) $\hat{y}_t$ (графа 6) и являются индексами сезонности (графа 7) по отношению к тренду.

Поскольку квартальные индексы в разные годы различны, они усредняются. Например, для I квартала $\bar{I}_{\text{I}} = (89{,}7 + 89{,}6 + 85{,}8)/3 = 88{,}4$, для II квартала $\bar{I}_{\text{II}} = (91{,}7 + 96{,}3 + 97{,}4)/3 = 95{,}1$ и т.д.

Усредненные значения записываются в качестве искомых индексов сезонности для всех трех лет (графа 8).

Умножая выравненные уровни на средние индексы сезонности, получаем теоретические (выравненные) уровни с учетом «сезонной волны» (графа 9).

Таблица 8.26

Расчет величин для определения индексов сезонности по отношению к тренду

Год	Квартал	Выполнено работ, млн руб. y_i	t	t^2	yt	Выравненные уровни $\hat{y}_t = 11{,}9 + 0{,}71t$	Индекс сезонности, % $I_{сез\,i} = y_i / \hat{y}_t$	Средний индекс сезонности $\bar{I}$	Выравненные уровни с учетом сезонности $\tilde{y}_t = \hat{y}_t \bar{I}$
А	1	2	3	4	5	6	7	8	9
2001	I	11,3	1	1	11,3	12,6	89,7	88,4	11,1
	II	12,2	2	4	24,4	13,3	91,7	95,1	12,6
	III	17,5	3	9	52,5	14,0	125,0	121,3	17,0
	IV	14,4	4	16	57,6	14,7	98,0	95,1	13,9
2002	I	13,8	5	25	69,0	15,4	89,6	88,4	13,6
	II	15,6	6	36	93,6	16,2	96,3	95,1	15,4
	III	20,2	7	49	141,4	16,9	119,5	121,3	20,5
	IV	17,4	8	64	139,2	17,6	99,0	95,1	16,7
2003	I	15,7	9	81	141,3	18,3	85,8	88,4	16,1
	II	18,4	10	100	184,0	18,9	97,4	95,1	17,9
	III	23,5	11	121	258,5	19,7	119,3	121,3	23,9
	IV	18,0	12	144	216,0	20,4	88,2	95,1	19,4
Σ	12	198,0	78	650	1388,8	198,0	1199,5 (1200)	1199,7 (1200)	198,1

Прогнозирование с учетом индекса сезонности

Индексы сезонности используются и при прогнозировании. Так, зная уравнение тренда и средние индексы сезонности, можно продлить наш ряд, т.е. спрогнозировать квартальные уровни, например, в 2004 г. при условии, что выявленная для 2001–2003 гг. закономерность развития устойчива и сохранится в прогнозируемом периоде. Как уже указывалось, этот метод продления в будущее закономерности (тенденции), выявленной в прошлом, называется *экстраполяцией.*

В общем виде

$$\hat{y}_{прогноз} = f(t)\,\bar{I}_{сез}.$$

В нашем примере

$$\hat{y}_{прогноз} = (a_0 + a_1 t)\,\bar{I}_{сез} = (11{,}9 + 0{,}71t)\,\bar{I}_{сез}.$$

Подставляя соответствующие значения t и индексов сезонности, получаем следующий прогноз на 2004 г.:

I квартал $\hat{y}_{I} = (11{,}9 + 0{,}71 \cdot 13) \cdot 0{,}884 = 18{,}7$ млн руб.;

II квартал $\hat{y}_{II} = (11{,}9 + 0{,}71 \cdot 14) \cdot 0{,}951 = 20{,}8$ млн руб.;

III квартал $\hat{y}_{III} = (11{,}9 + 0{,}71 \cdot 15) \cdot 1{,}213 = 27{,}4$ млн руб.;

IV квартал $\hat{y}_{IV} = (11{,}9 + 0{,}71 \cdot 16) \cdot 0{,}951 = 22{,}1$ млн руб.

Рассмотренная схема учета «сезонной волны» (умножение тренда на индекс сезонности) является *мультипликативной.*

Возможна и другая схема учета сезонной волны – *аддитивная*, когда к тренду прибавляется средняя величина абсолютных отклонений.

Чтобы экстраполировать ряд, приведенный в табл. 8.26, по аддитивной схеме, определим отклонение фактических уровней от выравненных и выполним все необходимые расчеты в табл. 8.27.

По аддитивной схеме $\hat{y}_{\text{прогноз}} = f(t) + \overline{(y_i - \hat{y}_t)}$ = тренд + средние отклонения по кварталам.

Таблица 8.27

Экстраполяция ряда по аддитивной схеме

Год	Квартал	Фактические уровни y_i	Выравненные уровни (тренд) $\hat{y}_t$	Абсолютное отклонение от тренда $y_i - \hat{y}_t$	Среднее отклонение по кварталам $\overline{y_i - \hat{y}_t}$	Выравненные уровни с учетом сезонности $\tilde{y}_t = \hat{y}_t + \overline{(y_i - \hat{y}_t)}$
А	1	2	3	4	5	6
2001	I	11,3	12,6	−1,3	−1,83	10,77
	II	12,2	13,3	−1,1	−0,73	12,57
	III	17,5	14,0	3,5	3,53	17,53
	IV	14,4	14,7	−0,3	−0,97	13,73
2002	I	13,8	15,4	−1,6	−1,83	13,57
	II	15,6	16,2	−0,6	−0,73	15,47
	III	20,2	16,9	3,3	3,53	20,43
	IV	17,4	17,6	−0,2	−0,97	16,63
2003	I	15,7	18,3	−2,6	−1,83	16,47
	II	18,4	18,9	−0,5	−0,73	18,17
	III	23,5	19,7	3,8	3,53	23,23
	IV	18,0	20,4	−2,4	−0,97	19,43
Σ		198,0	198,0	0	0	198,0

В нашем примере прогноз на 2004 г., выполненный по аддитивной схеме, даст следующие показатели по кварталам:

I квартал $\hat{y}_{I} = (11{,}9 + 0{,}71 \cdot 13) - 1{,}83 = 19{,}3$ млн руб.;

II квартал $\hat{y}_{II} = (11{,}9 + 0{,}71 \cdot 14) - 0{,}73 = 21{,}05$ млн руб.;

III квартал $\hat{y}_{III} = (11{,}9 + 0{,}71 \cdot 15) + 3{,}53 = 26{,}08$ млн руб.;

IV квартал $\hat{y}_{IV} = (11{,}9 + 0{,}71 \cdot 16) - 0{,}97 = 22{,}29$ млн руб.

Результаты прогнозирования, полученные по мультипликативной и аддитивной схемам, несколько отличаются, но эти различия не столь значительны. Вообще точечный прогноз весьма ненадежное дело. Обычно для прогнозируемых показателей с заданной вероятностью определяются интервалы «от и до», которые учитывают среднюю квадратическую ошибку уравнения тренда.

Разложение общей суммы квадратов отклонений фактических уровней от их средней

Как уже отмечалось, при анализе рядов динамики с наличием тренда и сезонных колебаний важно выделить в общей колеблемости фактических данных долю отдельных составляющих (тренда, сезонности и случайных колебаний). Эту задачу можно решить путем разложения общей суммы квадратов отклонений фактических уровней от среднего уровня ряда за весь период, т.е. $\sum(y_i - \bar{y})^2$, на отдельные составляющие. Так, если принять следующие обозначения для разных уровней:

y_i – фактические уровни ряда,

$\bar{y}$ – средний уровень ряда,

$\hat{y}_t$ – тренд (теоретические уровни, рассчитанные по аналитической функции),

$\tilde{y}_t = \hat{y}_t \bar{I}_{сез}$ – тренд с учетом сезонности,

то интерпретация следующих сумм будет такова:

1) $\sum(y_i - \bar{y})^2$ – общая сумма квадратов отклонений фактических уровней от их средней;
2) $\sum(\hat{y}_t - \bar{y})^2$ – сумма квадратов отклонений за счет тренда;
3) $\sum(\hat{y}_t - \tilde{y}_t)^2$ – сумма квадратов отклонений за счет сезонности;
4) $\sum(y_i - \tilde{y}_t)^2$ – сумма квадратов отклонений за счет случайных колебаний.

Общая сумма квадратов должна быть равна сумме трех последних сумм, т.е.

$$\sum(y_i - \bar{y})^2 = \sum(\hat{y}_t - \bar{y})^2 + \sum(\hat{y}_t - \tilde{y}_t)^2 + \sum(y_i - \tilde{y}_t)^2.$$

Проиллюстрируем это на нашем примере, для чего выпишем в отдельную таблицу исходные данные y_i и все рассчитанные нами уровни $\hat{y}_t$, $\tilde{y}_t$, а также квадраты соответствующих отклонений (табл. 8.28).

Таблица 8.28

Расчет величин для разложения общей суммы квадратов отклонений фактических уровней от их средней ($\bar{y}$ = 16,5)

Год	Квартал	Фактические уровни y_i	Выравненные уровни (тренд) $\hat{y}_t$	Выравненные уровни с учетом сезонности $\tilde{y}_t$	$(y_i-\bar{y})^2$	$(\hat{y}_t-\bar{y})^2$	$(\hat{y}_t-\tilde{y}_t)^2$	$(y_i-\tilde{y}_t)^2$
А	1	2	3	4	5	6	7	8
2001	I	11,3	12,6	11,1	27,04	15,21	2,25	0,04
	II	12,2	13,3	12,6	18,49	10,24	0,49	0,16
	III	17,5	14,0	17,0	1,00	6,25	9,00	0,25
	IV	14,4	14,7	13,9	4,41	3,24	0,64	0,25
2002	I	13,8	15,4	13,6	7,29	1,21	3,24	0,04
	II	15,6	16,2	15,4	0,81	0,09	0,64	0,04
	III	20,2	16,9	20,5	13,69	0,16	12,96	0,09
	IV	17,4	17,6	16,7	0,81	1,21	0,81	0,49
2003	I	15,7	18,3	16,1	0,64	3,24	4,84	0,16
	II	18,4	18,9	17,9	3,61	5,76	1,00	0,25
	III	23,5	19,7	23,9	49,00	10,24	17,64	0,16
	IV	18,0	20,4	19,4	2,25	15,21	1,00	1,96
Σ		198,0	198,0	198,1	129,04	72,06	54,51	3,89

Итак, в результате расчетов получаем

$$\sum(y_i - \bar{y})^2 = 129{,}04, \qquad \sum(\hat{y}_t - \bar{y})^2 = 72{,}06,$$
$$\sum(\hat{y}_t - \tilde{y}_t)^2 = 54{,}01, \qquad \sum(y_i - \tilde{y}_t)^2 = 3{,}89.$$

Сумма слагаемых (72,06; 54,51; 3,89) равна 130,46. Незначительное отличие этой суммы от 129,04 – это результат округлений на всех этапах расчета выравненных уровней.

На основе полученных данных можно сделать вывод, что случайные колебания в исходном ряду были весьма незначительны. Основные факторы колеблемости уровней исследуемого ряда – тренд и сезонность.

8.8. Автокорреляция в рядах динамики

Во многих рядах динамики можно наблюдать зависимость t-го уровня y_t от предшествующих y_{t-1}. Например, численность населения за определенный год зависит (при прочих равных условиях) от численности в предшествующие годы; то же можно сказать и о поголовье скота, численность которого в каждый год зависит от численности поголовья в предшествующие годы; урожайность сельскохозяйственных культур в отдельные годы также может быть связана с урожайностью в предшествующие периоды и т.д.

Зависимость между последовательными (соседними) уровнями ряда динамики называется в статистике ***автокорреляцией***. Исследование рядов на автокорреляцию – одна из частных, но важных задач при статистическом изучении рядов динамики. В частности, если установлено наличие автокорреляции, то эту зависимость можно выразить *уравнением авторегрессии.* В отдельных случаях приходится устранять влияние автокорреляции на взаимосвязь между исследуемыми показателями. Так возникает необходимость измерения автокорреляции.

Измерить автокорреляцию между уровнями ряда можно с помощью ***коэффициента автокорреляции*** r_a, исчисляемого по формуле парного линейного коэффициента корреляции

$$r = \frac{\overline{xy} - \bar{x}\bar{y}}{\sigma_x \sigma_y}.$$

Коэффициент автокорреляции r_a можно рассчитывать либо между соседними уровнями, либо между уровнями, сдвинутыми на любое число единиц времени m. Этот сдвиг, именуемый *временны́м лагом*, определяет порядок коэффициента автокорреляции: 1-го порядка при $m = 1$, т.е. между соседними уровнями; 2-го порядка при $m = 2$, т.е. при сдвиге уровней на два периода, и т.д.

Рассмотрим коэффициент автокорреляции 1-го порядка.

Если исходные фактические уровни ряда, относящиеся к определенному моменту времени (или периоду) t, обозначить через y_t, то сдвинутые уровни (в зависимости от направления сдвига) соответственно обозначают y_{t+1} или y_{t-1}. Тогда формулу коэффициента автокорреляции можно записать в двух вариантах:

$$r_a = \frac{\overline{y_t y_{t+1}} - \bar{y}_t \bar{y}_{t+1}}{\sigma_{y_t} \sigma_{y_{t+1}}}; \qquad (8.22)$$

$$r_a = \frac{\overline{y_t y_{t-1}} - \bar{y}_t \bar{y}_{t-1}}{\sigma_{y_t} \sigma_{y_{t-1}}}. \qquad (8.23)$$

Мы отдаем предпочтение формуле (8.23), поэтому все дальнейшие рассуждения и расчеты будут связаны с ней.

Нетрудно представить, что при достаточно большом числе уровней ряда n значения средних уровней и средних квадратических отклонений у исходного и сдвинутого рядов практически совпадают, т.е. $\bar{y}_t \cong \bar{y}_{t-1}$ и $\sigma_{y_t} \cong \sigma_{y_{t-1}}$.

Используя эти равенства и отдавая предпочтение средней $\bar{y}_t$ и дисперсии $\sigma^2_{y_t}$, рассчитанным для всех n членов исходного ряда, получим приближенную формулу коэффициента автокорреляции

$$r_a = \frac{\overline{y_t y_{t-1}} - (\bar{y}_t)^2}{\sigma^2_{y_t}} \qquad (8.24)$$

или тождественную ей

$$r_a = \frac{\sum y_t y_{t-1} - n(\bar{y}_t)^2}{\sum y_t^2 - n(\bar{y}_t)^2}. \qquad (8.25)$$

Чтобы иметь возможность пользоваться формулами (8.24) и (8.25) для коротких рядов, у которых первый и последний уровни отличаются незначительно, сдвинутый (укороченный) ряд условно дополняют, принимая $y_1 = y_n$ (чтобы сдвинутый ряд не укорачивался и чтобы средний уровень и дисперсия одного ряда были соответственно равны среднему уровню и дисперсии второго ряда).

Рассмотрим расчет коэффициента автокорреляции на примере.

Пример. Предположим, известны данные о поголовье коров в одном из регионов. Исходные данные и расчет необходимых величин для подстановки в формулы (8.24) и (8.25) приведены в табл. 8.29 (дополненные данные в сдвинутом ряду взяты в скобки).

Таблица 8.29

Расчет величин для определения коэффициента автокорреляции 1-го порядка

Год	Поголовье коров на начало года, тыс. голов (фактические уровни) y_t	Уровни, сдвинутые на один год y_{t-1}	$y_t y_{t-1}$	y_t^2
1993	4,2	(5,3)	22,26	17,64
1994	4,0	4,2	16,80	16,00
1995	4,3	4,0	17,20	18,49
1996	4,2	4,3	18,06	17,64
1997	4,3	4,2	18,06	18,49
1998	4,4	4,3	18,92	19,36
1999	4,5	4,4	19,80	20,25
2000	4,8	4,5	21,60	23,04
2001	5,0	4,8	24,00	25,00
2002	5,3	5,0	26,50	28,09
Σ	45,0	45,0	203,20	204,0

По итоговым данным табл. 8.29 находим:

$$\bar{y}_t = \frac{45}{10} = 4{,}5; \quad (\bar{y}_t)^2 = 4{,}5^2 = 20{,}25; \quad \overline{y_t^2} = \frac{204}{10} = 20{,}4;$$

$$\sigma^2_{y_t} = \overline{y_t^2} - (\bar{y}_t)^2 = 20{,}4 - 20{,}25 = 0{,}15; \quad \overline{y_t y_{t-1}} = \frac{203{,}2}{10} = 20{,}32.$$

Подставляя полученные значения в формулу (8.24), имеем

$$r_a = \frac{\overline{y_t y_{t-1}} - (\bar{y}_t)^2}{\sigma^2_{y_t}} = \frac{20{,}32 - 20{,}25}{0{,}15} = 0{,}47.$$

Тот же результат получим и по формуле (8.25):

$$r_a = \frac{\sum y_t y_{t-1} - n(\bar{y}_t)^2}{\sum y_t^2 - n(\bar{y}_t)^2} = \frac{203{,}2 - 10 \cdot 4{,}5^2}{204 - 10 \cdot 4{,}5^2} = 0{,}47.$$

Найденное значение коэффициента автокорреляции само по себе еще не говорит о наличии или отсутствии автокорреляции. Его необходимо сравнить с критическим.

Существуют специальные таблицы, в которых для разного числа членов ряда n и разных уровней значимости α определена критическая область проверяемой нулевой гипотезы (об отсутствии автокорреляции между уровнями ряда). Одна из таких таблиц, составленная Р. Андерсоном, приведена в Приложении 7.

Фактическое значение коэффициента автокорреляции r_a сравнивается с табличным (критическим) при 5- или 1-процентном уровне значимости. Если фактическое (расчетное) значение r_a меньше табличного, то гипотеза об отсутствии автокорреляции в ряду может быть принята. Если же фактическое значение r_a больше табличного, то нулевая гипотеза отвергается и делается вывод о наличии автокорреляции.

Сравним рассчитанное в нашем примере значение коэффициента автокорреляции $r_a = 0{,}47$ с табличным при 5-процентном уровне значимости. По таблице Приложения 7 находим, что для $n = 10$ при $\alpha = 0{,}05$ критическое значение коэффициента автокорреляции равно 0,360. Так как рассчитанное нами значение r_a (0,47) больше табличного (0,360), то с вероятностью $P = 0{,}95$ ($P = 1 - \alpha$) можно сделать вывод о наличии автокорреляции в исследуемом ряду (хотя с вероятностью 0,99 такой вывод сделать нельзя, поскольку для $\alpha = 0{,}01$ критическое значение коэффициента корреляции равно 0,525).

(При отрицательном коэффициенте автокорреляции фактическое и табличное значения сравниваются по модулю.)

Измерение автокорреляции между остаточными величинами

Часто приходится решать вопрос о наличии или отсутствии автокорреляции не между уровнями ряда, а между их отклонениями от тренда или от среднего уровня, т.е. между так называемыми *остаточными величинами.* Сумма таких остаточных величин и средняя из них равны нулю.

Из формулы (8.25) видно, что для рядов, у которых средний уровень равен нулю ($\bar{y} = 0$), коэффициент автокорреляции примет следующий вид:

$$r_a = \frac{\sum y_t y_{t-1}}{\sum y_t^2}.$$

Поскольку обычно через y обозначают уровни ряда, то во избежание путаницы в обозначениях для остаточных величин предпочтительнее использовать символ ε_t. Тогда формула *коэффициента автокорреляции для остаточных величин* примет вид

$$r_a = \frac{\sum_{t=2}^{n} \varepsilon_t \varepsilon_{t-1}}{\sum_{t=1}^{n} \varepsilon_t^2}. \qquad (8.26)$$

Кроме показателя r_a [см. формулу (8.26)], для обнаружения автокорреляции между соседними остаточными величинами часто используется критерий, разработанный Дурбиным и Ватсоном (в иной транскрипции Дарбиным и Уотсоном) и носящий их имена.

Критерий Дурбина – Ватсона, обозначаемый как d (иногда DW), рассчитывается по формуле

$$d = \frac{\sum_{t=2}^{n} (\varepsilon_t - \varepsilon_{t-1})^2}{\sum_{t=1}^{n} \varepsilon_t^2}. \qquad (8.27)$$

Этот показатель можно связать с формулой (8.26) коэффициента автокорреляции для остаточных величин. Так, если предположить, что $\sum_{t=2}^{n} \varepsilon_t^2 \cong \sum_{t=2}^{n} \varepsilon_{t-1}^2$, то, возведя в квадрат числитель критерия d, можно записать

$$d = \frac{2\sum_{t=2}^{n} \varepsilon_t^2 - 2\sum_{t=2}^{n} \varepsilon_t \varepsilon_{t-1}}{\sum_{t=1}^{n} \varepsilon_t^2} \cong 2\left(1 - \frac{\sum_{t=2}^{n} \varepsilon_t \varepsilon_{t-1}}{\sum_{t=1}^{n} \varepsilon_t^2}\right) = 2(1 - r_a) = 2 - 2r_a.$$

Очевидно, что если автокорреляция отсутствует, т.е. $r_a = 0$, то $d = 2$. Если же имеет место полная автокорреляция, т.е. r_a равен 1 или -1, то значение d будет соответственно 0 или 4.

Более точное суждение об отсутствии автокорреляции в остаточных величинах ε_t дает таблица, в которой для разного числа наблюдений n и числа независимых переменных в уравнении регрессии v определены верхние d_2 и нижние d_1 критические границы критерия d. Такая таблица приведена в Приложении 5.

Для проверки нулевой гипотезы об отсутствии автокорреляции в остаточных величинах рассчитанное фактическое значение d сравнивается с табличными d_1 и d_2:

1) если $d > d_2$ (до $4 - d_2$), гипотеза об отсутствии автокорреляции принимается;

2) если $d < d_1$, гипотеза об отсутствии автокорреляции отвергается;

3) если $d_1 \leqslant d \leqslant d_2$ или $(4 - d_2) \leqslant d \leqslant (4 - d_1)$, ничего определенного сказать нельзя и требуется дальнейшее исследование (например, уточнение уравнения тренда, увеличение числа наблюдений и пр.);

4) если $d > (4 - d_1)$, имеет место отрицательная автокорреляция.

Для иллюстрации расчета критерия Дурбина – Ватсона d, а также r_a воспользуемся данными табл. 8.15 (о производстве мяса в России).

На основе фактических и выравненных уровней рассчитаем остаточные величины $\varepsilon_t = y_t - \hat{y}_t$ и проверим их на автокорреляцию. Все расчеты показаны в табл. 8.30.

Таблица 8.30

Расчет величин для исчисления коэффициента автокорреляции r_a и критерия Дурбина – Ватсона d

Год	Производство мяса, млн т y_t	Выравненные уровни $\hat{y}_t$	Остаточные величины $\varepsilon_t = y_t - \hat{y}_t$	ε_{t-1}	$\varepsilon_t \varepsilon_{t-1}$	ε_t^2	$(\varepsilon_t - \varepsilon_{t-1})^2$
1991	9,4	9,28	0,12	—	—	0,0144	—
1992	8,3	8,43	–0,13	0,12	–0,0156	0,0169	0,0625
1993	7,5	7,58	–0,08	–0,13	0,0104	0,0064	0,0025
1994	6,8	6,73	0,07	–0,08	–0,0056	0,0049	0,0225
1995	5,9	5,88	0,02	0,07	0,0014	0,0004	0,0025
Σ	37,9	37,90	0		–0,0094	0,0430	0,0900

При аналитическом выравнивании рядов динамики остаточные величины проверяются на автокорреляцию. Цель проверки – определить адекватность подобранной функции (линии тренда), используемой для отражения тенденции развития в исследуемый период. Если в остаточных величинах обнаруживается автокорреляция, это признак неадекватности выбранного уравнения тренда.

Итак, коэффициент автокорреляции для ε_t по данным табл. 8.30

$$r_a = \frac{\sum_{t=2}^{n} \varepsilon_t \varepsilon_{t-1}}{\sum_{t=1}^{n} \varepsilon_t^2} = \frac{-0{,}0094}{0{,}043} = -0{,}218.$$

Далее обращаемся к таблице Приложения 7 и находим, что для $n = 5$ и $\alpha = 0{,}05$ критическое значение отрицательного коэффициента автокорреляции равно –0,753. Так как рассчитанное фактическое значение r_a (по модулю) меньше критического (0,218 < 0,753), делаем вывод об отсутствии автокорреляции в остаточных величинах.

Рассчитаем для этой же цели критерий Дурбина – Ватсона:

$$d = \frac{\sum_{t=2}^{n} (\varepsilon_t - \varepsilon_{t-1})^2}{\sum_{t=1}^{n} \varepsilon_t^2} = \frac{0{,}09}{0{,}043} = 2{,}09.$$

Полученное значение d близко к 2, что свидетельствует об отсутствии автокорреляции в остаточных величинах.

К такому же выводу придем, обратившись к таблице Приложения 5. Она начинается с $n = 15$, но все, что относится к $n = 15$, может быть использовано и для $n < 15$. Поскольку в нашем примере для выравнивания использовалась линейная функция с о д н о й переменной t, то в таблице Приложения 5 находим значение d_2 в графе, где $v = 1$. Для $n = 15$ верхняя граница $d_2 = 1{,}36$. Рассчитанное же нами $d = 2{,}09$. Так как $d > d_2$ (и не превосходит величину $4 - d_2 = 2{,}64$), гипотеза об отсутствии автокорреляции в остаточных величинах принимается, чем подтверждается и адекватность уравнения тренда.

Нахождение уравнения авторегрессии

В рядах динамики, в которых обнаружена автокорреляция между уровнями ряда, каждый уровень y_t можно рассматривать как функцию предыдущих значений уровней. Уравнение, выражающее эту зависимость, называется *уравнением авторегрессии.*

Наиболее простой формой зависимости между соседними уровнями ряда может служить линейная функция, выраженная уравнением

$$\hat{y}'_t = a_0 + a_1 y_{t-1}. \qquad (8.28)$$

Уравнение регрессии, которое связывает исходные уровни ряда с теми же уровнями, сдвинутыми на определенный лаг, определяется по общим правилам регрессионного анализа.

Параметры уравнения авторегрессии (8.28) с лагом в один год находим, решая систему нормальных уравнений

$$\begin{cases} na_0 + a_1 \sum y_{t-1} = \sum y_t, \\ a_0 \sum y_{t-1} + a_1 \sum y_{t-1}^2 = \sum y_t y_{t-1}. \end{cases}$$

При этом следует иметь в виду, что поскольку сдвинутый ряд y_{t-1} содержит на один уровень меньше, чем исходный ряд, то все расчеты сумм необходимо проводить для одного и того же числа членов ряда, а именно для $(n-1)$.

Продолжим рассмотрение примера, приведенного в табл. 8.29, и найдем для него уравнение авторегрессии. Скорректировав с учетом сдвига итоговые данные, рассчитанные в табл. 8.29, получим следующие значения величин, необходимых для решения системы нормальных уравнений:

$$n = 9;\ \sum y_t = 40{,}8;\ \sum y_{t-1} = 39{,}7;\ \sum y_{t-1}^2 = 175{,}91;\ \sum y_t y_{t-1} = 180{,}94.$$

Подставив их в систему уравнений, получим

$$\begin{cases} 9a_0 + 39{,}7a_1 = 40{,}8, \\ 39{,}7a_0 + 175{,}91a_1 = 180{,}94, \end{cases}$$

откуда находим $a_0 = -0{,}87$ и $a_1 = 1{,}225$.

Таким образом, авторегрессионная модель будет иметь вид

$$\hat{y}'_t = -0{,}87 + 1{,}225 y_{t-1}.$$

Подставляя в найденное уравнение значения уровней y_{t-1}, находим $\hat{y}'_t$, т.е. теоретическое поголовье коров для каждого года на основе данных за предыдущий год (табл. 8.31).

Из табл. 8.31 видно, что начиная с 1997 г. теоретические уровни, рассчитанные по авторегрессионной модели 1-го порядка, практически совпадают с фактическими уровнями, т.е. найденное линейное уравнение достаточно хорошо отражает характер зависимости между последовательными уровнями ряда.

Таблица 8.31

Год	Поголовье коров, тыс. голов		Год	Поголовье коров, тыс. голов	
	фактическое y_t	теоретическое $\hat{y}'_t$		фактическое y_t	теоретическое $\hat{y}'_t$
1993	4,2	—	1998	4,4	4,4
1994	4,0	4,3	1999	4,5	4,5
1995	4,3	4,0	2000	4,8	4,6
1996	4,2	4,4	2001	5,0	5,0
1997	4,3	4,3	2002	5,3	5,3

Более сложной формой линейной авторегрессионной зависимости будет такая, при которой значение уровня в каждый момент t, т.е. $\hat{y}'_t$, характеризуется зависимостью одновременно от нескольких предшествующих уровней, т.е.

$$\hat{y}'_t = f(y_{t-1}, y_{t-2}, \ldots, y_{t-m}),$$

или

$$\hat{y}'_t = a_0 + a_1 y_{t-1} + a_2 y_{t-2} + \ldots + a_m y_{t-m},$$

где m — число уровней ряда, включенных в уравнение в качестве переменных и определяющих порядок авторегрессии.

Авторегрессионные модели различного порядка можно оценить с помощью остаточных дисперсий, рассчитываемых между фактическими и теоретическими уровнями, исчисленными по уравнениям авторегрессии разного порядка. Предпочтение следует отдать уравнению авторегрессии с таким числом m, при котором остаточная дисперсия минимальна.

8.9. Корреляция рядов динамики

Во многих экономических исследованиях приходится изучать динамику нескольких показателей одновременно, т.е. рассматривать параллельно несколько динамических рядов. Естественно, что в этих случаях можно встретить ряды, у которых колебания уровней взаимообусловлены. Например, динамика рыночных цен на какую-либо продукцию земледелия в известной степени связана с динамикой урожайности данной культуры. В свою очередь, динамика урожайности или валовых сборов зависит от динамики количества осадков. Динамика перевозок грузов определенным образом зависит от динамики производства промышленной и сельскохозяйственной продукции и т.п.

При изучении таких рядов динамики, естественно, возникает необходимость измерить зависимость между ними, вернее, определить, насколько изменения уровней одного ряда зависят от изменения уровней другого ряда. Эта задача решается обычно путем *коррелирования* рядов динамики, т.е. путем исчисления ***коэффициента корреляции*** между уровнями двух рядов:

$$r = \frac{\overline{xy} - \bar{x}\bar{y}}{\sigma_x \sigma_y} \quad \text{или} \quad r = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sqrt{\left[\sum x^2 - \frac{(\sum x)^2}{n}\right]\left[\sum y^2 - \frac{(\sum y)^2}{n}\right]}}.$$

Однако при этом возникает следующая проблема. Если показатели ряда x и ряда y рассматривать как функцию времени, т.е. $x = f(t)$ и $y = f(t)$, то при однонаправленности их трендов можно получить большое значение коэффициента корреляции между x и y даже тогда, когда они независимы, именно в силу однонаправленности их изменения.

Поэтому, прежде чем коррелировать ряды динамики, необходимо установить, возможна ли связь между исследуемыми показателями x и y. Ответ на этот вопрос дает логический (качественный) анализ.

Кроме того, одно из условий корреляции – независимость отдельных значений переменных множества x, так же как и множества y. Для рядов динамики это равнозначно отсутствию автокорреляции между уровнями ряда, т.е. коррелировать уровни рядов динамики можно лишь в том случае, если в каждом из них отсутствует автокорреляция. Следовательно, прежде чем коррелировать уровни рядов динамики, необходимо проверить каждый ряд на автокорреляцию [по формуле (8.25)].

Коррелирование уровней рядов динамики

Как установить, возможна ли корреляция уровней двух рядов динамики? Рассмотрим конкретный пример.

Пример. Пусть имеются следующие данные по одному из регионов за 5 лет (табл. 8.32).

По логике расход кормов на одну голову скота, несомненно, влияет на продукцию выращивания скота (его прирост, привес). Поставим задачу измерить тесноту этой связи. Однако прежде чем коррелировать уровни, проверим каждый ряд на автокорреляцию.

Таблица 8.32

Расход кормов и продукция выращивания скота в расчете на одну условную голову крупного рогатого скота, кг

Год	1998	1999	2000	2001	2002
Расход кормов (кормовых единиц) x	30,4	30,3	28,8	29,3	30,2
Продукция выращивания скота y	118	111	102	103	95

Сначала рассмотрим ряд показателей x – расход кормов. Все расчеты по нему приведены в табл. 8.33.

Таблица 8.33

Расчет величин для проверки наличия автокорреляции в ряду x

Год	x_t	x_{t-1}	$x_t x_{t-1}$	x_t^2
1998	30,4	(30,2)	918,08	924,16
1999	30,3	30,4	921,12	918,09
2000	28,8	30,3	872,64	829,44
2001	29,3	28,8	843,84	858,49
2002	30,2	29,3	884,86	912,04
Σ	149,0	149,0	4440,54	4442,22

Среднее значение уровней

$$\bar{x}_t = \frac{\sum x_t}{n} = \frac{149}{5} = 29{,}8.$$

Подставляя рассчитанные показатели в формулу (8.25), получаем

$$r_a = \frac{\sum x_t x_{t-1} - n\left(\bar{x}_t\right)^2}{\sum x_t^2 - n\left(\bar{x}_t\right)^2} = \frac{4440{,}54 - 5 \cdot 29{,}8^2}{4442{,}22 - 5 \cdot 29{,}8^2} = 0{,}168.$$

По таблице Приложения 7 определяем предельное (критическое) значение коэффициента автокорреляции для $n = 5$ и $\alpha = 0{,}05$. Оно равно 0,253. Так как рассчитанный нами $r_a = 0{,}168$ меньше табличного, делаем вывод об отсутствии автокорреляции между уровнями в ряду x.

Аналогичные выводы получим и для уровней ряда y. (Расчеты в целях экономии места опускаем.)

Следовательно, в примере, приведенном в табл. 8.32, можно коррелировать у р о в н и рядов x и y.

Расчет величин, необходимых для исчисления линейного коэффициента корреляции между x и y, показан в табл. 8.34.

Таблица 8.34

Расчет величин для определения коэффициента корреляции между *x* и *y*

Год	x	y	x^2	xy	y^2
1998	30,4	118	924,16	3587,2	13924
1999	30,3	111	918,09	3363,3	12321
2000	28,8	102	829,44	2937,6	10404
2001	29,3	103	858,49	3017,9	10609
2002	30,2	95	912,04	2869,0	9025
Σ	149	529	4442,22	15775,0	56283

Воспользовавшись полученными в табл. 8.34 суммами, рассчитаем линейный коэффициент корреляции между x и y по следующей формуле:

$$r_{xy} = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sqrt{\left[\sum x^2 - \frac{(\sum x)^2}{n}\right]\left[\sum y^2 - \frac{(\sum y)^2}{n}\right]}} =$$

$$= \frac{15775 - \frac{149 \cdot 529}{5}}{\sqrt{\left[4442{,}22 - \frac{149^2}{5}\right]\left[56283 - \frac{529^2}{5}\right]}} = 0{,}43.$$

Полученное значение коэффициента корреляции (0,43) характеризует среднюю тесноту связи между изменением продукции выращивания скота y и расходом кормов на одну голову x.

Исключение автокорреляции в рядах динамики

Если между уровнями ряда (при корреляции рядов динамики) существует автокорреляция, она должна быть устранена.

Есть несколько способов исключения автокорреляции в рядах динамики. Рассмотрим два из них: коррелирование отклонений от выравненных уровней и коррелирование последовательных разностей.

1. Коррелирование отклонений от выравненных уровней. Один из способов исключения автокорреляции заключается в том, что коррелируются не сами уровни, а отклонения фактических уровней от выравненных, отражающих тренд, т.е. коррелируются

остаточные величины. Для этого каждый ряд динамики выравнивают по определенной характерной для него аналитической формуле (т.е. находят $\hat{x}_t$ и $\hat{y}_t$), затем из эмпирических уровней вычитают выравненные (т.е. находят $d_x = x - \hat{x}_t$ и $d_y = y - \hat{y}_t$) и определяют тесноту связи между отклонениями d_x и d_y. Формулу *коэффициента корреляции между остаточными величинами* можно записать в следующем виде:

$$r = \frac{\sum d_x d_y}{\sqrt{\sum d_x^2 \sum d_y^2}}. \qquad (8.29)$$

Рассмотрим измерение тесноты связи между остаточными величинами на конкретном примере.

Пример. Предположим, по одному из районов имеются данные о поголовье коров x и производстве молока y за 10 лет (табл. 8.35). Измерить корреляцию между изменением уровней в двух рядах x и y.

Рассчитаем линейный коэффициент корреляции между уровнями x и y по формуле

$$r_{xy} = \frac{\overline{xy} - \overline{x}\,\overline{y}}{\sigma_x \sigma_y}.$$

Он равен 0,95. Однако можно предположить, что такое большое значение r обусловлено тем, что оба ряда имеют однонаправленный тренд.

Проверяем уровни каждого ряда на автокорреляцию и убеждаемся, что она присутствует и в одном, и в другом ряду. Поэтому коэффициент корреляции между x и y, равный 0,95, явно преувеличен и не может рассматриваться как показатель тесноты связи между колебаниями уровней двух рядов. Следовательно, надо исключить автокорреляцию в каждом ряду.

Чтобы исключить влияние автокорреляции в каждом ряду, выровняем и один и второй ряд по уравнению прямой. Не приводя расчеты параметров, запишем уравнение тренда для каждого ряда при условии отсчета времени от середины ряда (для 1998 г. $t = -1$, для 1999 г. $t = 1$):

$$\hat{x}_t = 4{,}51 + 0{,}07t\,, \qquad \hat{y}_t = 12{,}2 + 0{,}22t.$$

Подставляя в каждое уравнение значения $t = -9, -7, -5, \ldots$, получаем выравненные значения $\hat{x}_t$ и $\hat{y}_t$. Эти значения приведены в табл. 8.35. Там же показаны отклонения фактических уровней от выравненных (с точностью до десятых) и расчет величин, необходимых для исчисления коэффициента корреляции между d_x и d_y.

Таблица 8.35

Расчет остаточных величин

Год	Поголовье коров, тыс. голов x	Производство молока, тыс. т y	Выравненные значения		Остаточные величины ε_t		d_x^2	d_y^2	$d_x d_y$
			$\hat{x}_t$	$\hat{y}_t$	$d_x = x - \hat{x}_t$	$d_y = y - \hat{y}_t$			
1994	4,0	10,0	3,9	10,2	0,1	−0,2	0,01	0,04	−0,02
1995	4,0	10,2	4,0	10,7	0	−0,5	0	0,25	0
1996	4,2	11,5	4,2	11,1	0	0,4	0	0,16	0
1997	4,3	11,8	4,3	11,5	0	0,3	0	0,09	0
1998	4,4	12,0	4,4	12,0	0	0	0	0	0
1999	4,5	12,6	4,6	12,4	−0,1	0,2	0,01	0,04	−0,02
2000	4,6	12,8	4,7	12,9	−0,1	−0,1	0,01	0,01	0,01
2001	4,8	13,1	4,9	13,3	−0,1	−0,2	0,01	0,04	0,02
2002	5,0	13,6	5,0	13,7	0	−0,1	0	0,01	0
2003	5,3	14,4	5,1	14,2	0,2	0,2	0,04	0,04	0,04
Σ	45,1	122,0	45,1	122,0	0	0	0,08	0,68	0,03

Примечание. Чтобы различать остаточные величины ε_t для разных рядов, приняты обозначения d_x и d_y.

Находим коэффициент корреляции между остаточными величинами d_x и d_y:

$$r_{d_x d_y} = \frac{\sum d_x d_y}{\sqrt{\sum d_x^2 \sum d_y^2}} = \frac{0{,}03}{\sqrt{0{,}08 \cdot 0{,}68}} = 0{,}128.$$

Судя по значению рассчитанного коэффициента корреляции, можно сказать, что зависимость между коррелируемыми величинами незначительна (в исследуемом периоде).

При исчислении коэффициента корреляции между остаточными величинами предполагается, что отклонения фактических уровней от выравненных (т.е. d_x и d_y) являются случайными величинами, не зависящими от времени, т.е. что между ними отсутствует автокорреляция. Однако, если недостаточно точно подобрано уравнение тренда или по другим причинам, остаточные величины могут содержать автокорреляцию. Поэтому, прежде чем коррелировать отклонения d_x и d_y, необходимо убедиться, что между этими остаточными величинами автокорреляция отсутствует. В целях проверки используют коэффициент автокорреляции

для рядов с нулевым значением среднего уровня r_a или критерии Дурбина – Ватсона d:

$$r_a = \frac{\sum_{t=2}^{n} \varepsilon_t \varepsilon_{t-1}}{\sum_{t=1}^{n} \varepsilon_t^2}, \qquad d = \frac{\sum_{t=2}^{n} (\varepsilon_t - \varepsilon_{t-1})^2}{\sum_{t=1}^{n} \varepsilon_t^2}.$$

В рассмотренном примере проверка d_x и d_y на автокорреляцию показала, что в каждом из них автокорреляция отсутствует. Поэтому вполне правомерно их коррелировать.

2. Коррелирование последовательных разностей. Исключить влияние тренда при коррелировании рядов динамики можно и другим способом, в частности путем корреляции последовательных разностей уровней каждого ряда. Алгебраически легко показать, что при переходе от уровней к их разностям исключается влияние общей тенденции на колеблемость.

Если исходить из того, что каждый фактический уровень – это результат влияния главной тенденции (тренда) и случайных остаточных факторов, т.е. $y = \hat{y}_t + \varepsilon_t$, где $\hat{y}_t$ – выравненное значение, определяющее тренд, а ε_t – отклонение фактического уровня от выравненного значения, то при изменении ряда по прямой $\hat{y}_t = a_0 + a_1 t$, обозначая последовательно моменты времени через $t = 1, 2, 3, \ldots$, можно записать:

для $t = 1$ $\quad y_1 = a_0 + a_1 + \varepsilon_1$;
$t = 2$ $\quad y_2 = a_0 + 2a_1 + \varepsilon_2$;
$t = 3$ $\quad y_3 = a_0 + 3a_1 + \varepsilon_3$;
$t = 4$ $\quad y_4 = a_0 + 4a_1 + \varepsilon_4$ и т.д.

Найдем первые разности уровней:

$$\Delta_1' = y_2 - y_1 = a_1 + (\varepsilon_2 - \varepsilon_1);$$
$$\Delta_2' = y_3 - y_2 = a_1 + (\varepsilon_3 - \varepsilon_2);$$
$$\Delta_3' = y_4 - y_3 = a_1 + (\varepsilon_4 - \varepsilon_3) \quad \text{и т.д.}$$

Так как во всех этих разностях присутствует одна и та же постоянная величина a_1, то очевидно, что колебания рассчитанных разностей Δ зависят только от ε_t, т.е. влияние общей тенденции (тренда) механически исключается.

Если уровни ряда изменяются по параболе 2-го порядка, т.е. если $\hat{y}_t = a_0 + a_1 t + a_2 t^2$, то получим:

для $t = 1$ $\quad y_1 = a_0 + a_1 + a_2 + \varepsilon_1$;
$t = 2$ $\quad y_2 = a_0 + 2a_1 + 4a_2 + \varepsilon_2$;
$t = 3$ $\quad y_3 = a_0 + 3a_1 + 9a_2 + \varepsilon_3$;
$t = 4$ $\quad y_4 = a_0 + 4a_1 + 16a_2 + \varepsilon_4$ и т.д.

Найдем первые разности уровней:

$$\Delta_1' = y_2 - y_1 = a_1 + 3a_2 + (\varepsilon_2 - \varepsilon_1);$$
$$\Delta_2' = y_3 - y_2 = a_1 + 5a_2 + (\varepsilon_3 - \varepsilon_2);$$
$$\Delta_3' = y_4 - y_3 = a_1 + 7a_2 + (\varepsilon_4 - \varepsilon_3) \quad \text{и т.д.}$$

Как видно, первые разности содержат кроме постоянного a_1 еще и переменные слагаемые: $3a_2$, $5a_2$, $7a_2$ и $(\varepsilon_2 - \varepsilon_1)$, $(\varepsilon_3 - \varepsilon_2)$, $(\varepsilon_4 - \varepsilon_3)$ и т.д.

Для того чтобы устранить влияние общей тенденции, на основе первых разностей рассчитаем вторые разности:

$$\Delta_1'' = \Delta_2' - \Delta_1' = 2a_2 + (\varepsilon_3 - 2\varepsilon_2 + \varepsilon_1);$$
$$\Delta_2'' = \Delta_3' - \Delta_2' = 2a_2 + (\varepsilon_4 - 2\varepsilon_3 + \varepsilon_2) \quad \text{и т.д.}$$

Как видно из расчетов, различие вторых разностей определяется только ε_t, так как $2a_2$ – величина постоянная во всех вторых разностях.

Таким образом, если возникает необходимость определить корреляцию между двумя рядами с исключением влияния общей тенденции в каждом ряду, можно коррелировать последовательные разности уровней:

при изменении уровней по прямой – первые разности;

при изменении по параболе 2-го порядка – вторые разности;

при изменении по параболе n-го порядка – n-е разности.

Формула *коэффициента корреляции разностей* по аналогии с формулой (8.29) имеет вид

$$r_{\Delta_x\Delta_y} = \frac{\sum \Delta_x \Delta_y}{\sqrt{\sum \Delta_x^2 \sum \Delta_y^2}}. \qquad (8.30)$$

Пример. Требуется рассчитать коэффициент корреляции между последовательными разностями по данным о количестве внесенных минеральных удобрений на 1 га под зерновыми культурами и об урожайности зерновых культур в хозяйствах России за 1992–1996 гг. Исходные данные и расчет необходимых показателей приведены в табл. 8.36.

Подставляя в формулу (8.30) итоговые показатели, полученные в табл. 8.36, имеем

$$r_{\Delta_x\Delta_y} = \frac{\sum \Delta_x \Delta_y}{\sqrt{\sum \Delta_x^2 \sum \Delta_y^2}} = \frac{70{,}9}{\sqrt{585 \cdot 13{,}95}} = 0{,}785.$$

Таблица 8.36

Расчет величин для коррелирования последовательных разностей

Год	Внесено удобрений на 1 га, кг x	Урожайность зерновых, ц/га y	$\Delta_x = x_i - x_{i-1}$	$\Delta_y = y_i - y_{i-1}$	Δ_x^2	Δ_y^2	$\Delta_x \Delta_y$
1992	52	17,2	—	—	—	—	—
1993	46	16,3	−6	−0,9	36	0,81	5,4
1994	24	14,4	−22	−1,9	484	3,61	41,8
1995	16	11,6	−8	−2,8	64	7,84	22,4
1996	17	12,9	1	1,3	1	1,69	1,3
Σ					585	13,95	70,9

Рассчитанное значение $r_{\Delta_x \Delta_y}$ свидетельствует о сильном влиянии изменения количества вносимых удобрений на изменение урожайности зерновых. (Кстати, и в ряду x, и в ряду y автокорреляция отсутствует, т.е. можно было коррелировать сами уровни, а не их разности. Линейный коэффициент корреляции между уровнями x и y равен 0,96, т.е. $r_{xy} = 0{,}96$.)

Корреляция рядов с лагом

Изучая корреляцию между рядами динамики, следует иметь в виду, что во многих случаях изменения уровней одного ряда могут вызвать изменение уровней другого ряда только через определенный интервал времени. Например, увеличение (или снижение) производства многих товаров в данном периоде вызовет увеличение (или снижение) объема товарооборота через определенный промежуток времени, увеличение числа браков в данном году может привести к увеличению числа родившихся через год и т.д.

Поэтому, чтобы правильно оценить влияние изменения уровней одного ряда на другой, необходимо сдвигать один ряд относительно другого на определенный промежуток времени (лаг) и коррелировать ряды с лагом. Предварительный логический анализ должен помочь исследователю определить этот лаг.

Скользящие коэффициенты корреляции

Коэффициент корреляции, отражающий тесноту связи (зависимости) между изменениями уровней двух рядов, служит своего рода средним, обобщающим показателем для конкретного периода.

Однако для длительного периода эта зависимость не является постоянной, она может меняться во времени. Поэтому, чтобы судить о том, в какие периоды зависимость между изменениями уровней двух рядов слабее или сильнее, рекомендуется рассчитывать с е р и ю скользящих коэффициентов корреляции для определенного интервала (по аналогии с расчетом скользящей средней при выравнивании динамических рядов). На основе расчета скользящих коэффициентов корреляции можно выявить те периоды, когда зависимость усиливается или уменьшается. Зная такие периоды, легче объяснить изменение этой зависимости в конкретных условиях (экономических или др.) отмеченного периода.

Определение уравнения регрессии для связанных рядов динамики

Ряды динамики, уровни которых могут рассматриваться у одних как результативные, а у других – как факторные, называют *связанными.*

Для таких связанных рядов не только измеряют корреляцию между ними (рассмотренными методами), но и при необходимости находят уравнение регрессии, аналогично тому, как это решается в пространственных совокупностях.

Однако в этих случаях, чтобы устранить или уменьшить автокорреляцию, в уравнение регрессии дополнительно вводится фактор времени t, причем в линейной форме. Например, если зависимость между y и x предположительно линейная, уравнение регрессии примет вид

$$\hat{y}_{x,t} = a_0 + a_1 x + a_2 t.$$

Если предполагается, что зависимость может быть выражена функцией параболы 2-го порядка, то

$$\hat{y}_{x,t} = a_0 + a_1 x + a_2 x^2 + a_3 t$$

и т.д.

Метод включения фактора времени в уравнение регрессии для связанных рядов динамики был предложен Фришем и Боу и известен под их именами.

Уравнение регрессии для связанных рядов может выражать модель изменения уровней одного ряда (результативного) от н е с к о л ь к и х других, например: изменение объема перевозок грузов от изменения производства промышленной и сельскохозяйственной продукции, от изменения тарифов и других факторов.

Корреляцию между такими рядами можно рассматривать как множественную и применять к ним все приемы исследования

множественной корреляции, с учетом фактора времени (см. параграф 7.8).

В заключение еще раз отметим, что, прежде чем коррелировать ряды динамики, следует подвергнуть их тщательному анализу и установить логическую связь между рассматриваемыми показателями. В противном случае (при формальном подходе) можно рассчитать довольно высокий коэффициент корреляции даже при отсутствии зависимости (в силу простого поступательного параллельного изменения во времени двух показателей).

8.10. Анализ рядов динамики и прогнозирование

Изучая и анализируя ряды динамики, исследователи всегда стремились на основе выявленных особенностей изменения явлений в прошлом предугадать поведение рядов в будущем, т.е. пытались строить различные прогнозы путем экстраполяции (продления) рядов.

Экстраполяцию ряда динамики можно осуществить различными способами. Однако независимо от применяемого способа экстраполяции обязательно предполагается, что закономерность (тенденция) изменения, выявленная для определенного периода в прошлом, сохранится на ограниченном отрезке времени в будущем. Поэтому любому прогнозированию в виде экстраполяции ряда должно предшествовать тщательное изучение длительных рядов динамики, которое позволило бы определять тенденцию изменения. Поскольку тенденция развития также может изменяться, то данные, полученные путем экстраполяции ряда, надо рассматривать как вероятностные, как своего рода оценки.

Перечислим некоторые простейшие **приемы экстраполяции рядов динамики**, помогающие прогнозировать те или иные показатели.

1. Если при анализе ряда динамики обнаруживается, что абсолютные приросты уровней примерно постоянны, можно рассчитать средний абсолютный прирост (как среднюю арифметическую) и последовательно прибавить его к последнему уровню ряда столько раз, на сколько периодов экстраполируется ряд.

2. Если за исследуемый ряд лет (или другие периоды) годовые коэффициенты роста остаются более-менее постоянными, можно рассчитать средний коэффициент роста и умножить последний уровень ряда на средний коэффициент роста в степени, соответствующей периоду экстраполяции.

3. Учитывая, что между изменениями нескольких показателей существует зависимость, можно экстраполировать один ряд ди-

намики на основе сведений об изменении второго ряда, связанного с ним.

Так, определив зависимость между изменением объема капитальных вложений и объемом выпускаемой продукции в той или иной отрасли, можно экстраполировать данные о производстве продукции на основе данных о намеченных капиталовложениях; зная, какой будет численность детей через t лет (по таблицам смертности), можно определить возможное потребление детских товаров и т.д.

4. Можно экстраполировать ряды на основе выравнивания их по определенной аналитической формуле. Зная уравнение для теоретических уровней и подставляя в него значения t за пределами исследованного ряда, можно рассчитать для данных t вероятностные уровни $\hat{y}_t$.

Так как, выравнивая ряды динамики по аналитическим формулам, мы главным образом определяем тренд, то при прогнозировании иногда целесообразно, выровняв ряд по той или иной формуле и определив тренд, найти отклонение фактических уровней от выравненных. Затем определить закономерность (тренд) изменения во времени этих отклонений, т.е. найти для их изменения свою формулу, свой тренд. После этого экстраполировать оба ряда, накладывая их друг на друга.

Пользуясь этим методом, следует помнить, что экстраполяция динамического ряда на основе уравнения, полученного при выравнивании, только тогда может дать оценки, близкие к реальным значениям, когда в эмпирическом ряду невелики случайные колебания, измеряемые средним квадратическим отклонением разности $(y - \hat{y}_t)$, и между случайными отклонениями отсутствует автокорреляция.

5. Иногда при прогнозировании можно экстраполировать авторегрессионную функцию уровней ряда. При этом методе изучаемый ряд динамики анализируют с точки зрения автокорреляции.

Очевидно, что чем больше автокорреляция между уровнями ряда, тем больше оснований для расчета будущих показателей на основе имеющихся. При этом автокорреляция должна быть исчислена для разных лагов между уровнями. Установив наличие автокорреляции между уровнями ряда (с определенным лагом), можно найти уравнение, выражающее эту автокорреляционную зависимость, и, пользуясь им, экстраполировать ряд.

Данный перечень методов прогнозирования не является исчерпывающим, приведены лишь простейшие методы экстраполяции.

Однако хорошо известно, что те или иные «предсказания» статистики иногда не только не подтверждаются, но прямо противоположны действительному ходу изменения изучаемых показателей. Это доказывает, что прогнозирование, основанное только на обработке данных наблюдения, слишком рискованно, если оно не учитывает множества взаимосвязанных фактов и моментов, которые способны изменить тенденцию развития в будущем.

Прогнозы могут строиться на длительный период (долгосрочные прогнозы) и на небольшие отрезки времени (краткосрочные прогнозы). Естественно, что и методы прогнозирования при этом могут и должны различаться. Так, при долгосрочном прогнозе урожайности (на 5–10 лет) следует исходить из динамики средней многолетней урожайности и экстраполировать найденную для нее аппроксимирующую функцию. Для краткосрочных же прогнозов более важно исследовать влияние факторов, определяющих изучаемый показатель. Например, при прогнозировании урожайности в текущем году важно изучить состояние на определенный момент многих факторов, влияющих на урожайность (количество влаги в почве весной, количество внесенных удобрений, качество семян и т.п.), и, зная зависимость урожайности от них в виде уравнения связи, установленного по данным наблюдения в прошлом, строить прогноз. В этом случае прогноз базируется как бы на факторах-симптомах, т.е. по состоянию отдельных факторов на данный период определяется состояние изучаемого показателя в будущем.

Экономическое прогнозирование невозможно без хорошего знания изучаемого явления и владения различными методами обработки динамических рядов, которые в каждом отдельном случае помогли бы обнаружить общую закономерность изменения, периодичность в повышении или снижении уровней (если она имеет место), случайные колебания, автокорреляцию и корреляцию между отдельными рядами.

Глава 9
ЭКОНОМИЧЕСКИЕ ИНДЕКСЫ

9.1. Общее понятие об индексах. Их виды

Среди методов статистического анализа особое и весьма важное место занимает *индексный метод.*

Слово «индекс» (*index*) в переводе с латинского означает показатель, указатель. В статистике под ***индексом*** понимается относительная величина, характеризующая соотношение значений определенного показателя во времени, пространстве, а также сравнение фактических данных с планом или иным нормативом.

В зависимости от базы сравнения индексы можно подразделить на ***динамические*** (отражающие изменение явления во времени) и ***территориальные*** (используемые для пространственных, межрегиональных сопоставлений различных показателей).

Чаще всего термин «индекс» ассоциируется с понятием относительного изменения какого-либо показателя во времени.

Показатель, изменение которого характеризуется индексом, называют *индексируемой величиной.* Последняя содержится в названии самого индекса, например: индекс цен, индекс заработной платы, индекс физического объема продукции и т.д.

Индексный метод имеет свою терминологию и символику. Обычно используются следующие обозначения индексируемых величин:

- q — количество (объем) какого-либо товара, продукции в натуральном выражении;
- p — цена единицы товара;
- pq — стоимость продукции, или товарооборот;
- c (или z) — себестоимость единицы продукции;
- t — затраты времени на производство единицы продукции, трудоемкость;
- w — выработка продукции в единицу времени или на одного работника (производительность труда*);

* Иногда рекомендуется символом w обозначать производительность труда в натуральном выражении $\left(w = \frac{q}{T}\right)$, а в стоимостном выражении (в сопоставимых ценах) использовать другой символ, например $V\left(V = \frac{pq}{T}\right)$, или наоборот.

$T = tq$ – общие затраты времени на производство продукции или численность работников;

y – урожайность отдельных сельскохозяйственных культур;

$П$ (или S) – посевная площадь под отдельными культурами и т.д.

Поскольку индексы рассчитываются путем сравнения значений определенного показателя за два периода, то, чтобы различать, к какому периоду относятся индексируемые величины, возле каждого символа справа ставятся подстрочные знаки: 0 – для базисного периода (база сравнения) и 1 – для отчетного (текущего) периода.

Если же рассчитываются индексы для ряда периодов, то обычно каждая индексируемая величина, отнесенная к определенному периоду, снабжается его подстрочным символом. Например, данные о количестве произведенного продукта за пять лет можно обозначить как q_1, q_2, q_3, q_4, q_5.

По степени охвата элементов совокупности индексы делятся на индивидуальные и общие (сводные).

Индивидуальные индексы, обозначаемые символом i, характеризуют относительное изменение отдельного единичного элемента сложной совокупности (например, изменение цены на молоко или хлеб, изменение урожайности ячменя или пшеницы (яровой или озимой), изменение объема добычи нефти или газа и т.д.). Исходя из принятых обозначений индексируемых величин, нетрудно записать формулы индивидуальных индексов для различных показателей:

$i_q = \frac{q_1}{q_0}$ – индекс объема одного определенного продукта (товара);

$i_p = \frac{p_1}{p_0}$ – индекс цены определенного продукта;

$i_c = \frac{c_1}{c_0}$ – индекс себестоимости единицы отдельного продукта;

$i_w = \frac{w_1}{w_0}$ – индекс производительности труда (по отдельным видам продукции);

$i_T = \frac{T_1}{T_0}$ – индекс численности работников (или общих затрат времени на производство продукции);

$i_y = \frac{y_1}{y_0}$ – индекс урожайности отдельной культуры и т.д.

Все индивидуальные индексы показывают, каково соотношение между отчетным (со знаком «1») и базисным (со знаком «0») показателями или во сколько раз увеличилась или уменьшилась индексируемая величина.

Все индивидуальные индексы по сути являются относительными величинами динамики или коэффициентами (темпами) роста (снижения), рассмотренными в главе 8.

В расчете индивидуальных индексов проблем практически нет. Однако в области экономических явлений наряду с индивидуальными индексами, характеризующими изменения единичных элементов, возникает необходимость расчета сводных относительных величин, обобщающих изменения определенного показателя в сложной совокупности, отдельные элементы которой несопоставимы (в физических единицах) и поэтому непосредственно не могут суммироваться.

Например, нельзя суммировать в физических единицах тонны нефти с тоннами стали или метрами ткани, киловатт-часами электроэнергии и т.п. Так же, как нельзя непосредственно суммировать цены на разные товары (на мясо, картофель, молоко, хлеб, обувь, одежду и т.п.).

Для обобщения относительного изменения определенного показателя в сложной совокупности рассчитываются ***общие (сводные) индексы***, обозначаемые символом I и характеризующие относительное изменение индексируемой величины (показателя) в целом по сложной совокупности, отдельные элементы которой несоизмеримы в физических единицах.

Например, по данным Госкомстата России, цены на продовольственные товары в декабре 2002 г. составляли 102,2% по отношению к предыдущему месяцу (ноябрю) и 111,0% по отношению к декабрю 2001 г.; продукция промышленности в 2003 г. составила 107,7% по отношению к предыдущему году, а продукция сельского хозяйства – соответственно 101,5%.

Все эти данные – общие индексы, поскольку в приведенных примерах речь идет об изменениях определенного показателя в сложных совокупностях (цен на продовольственные товары, объема продукции промышленности, сельского хозяйства).

Общие индексы могут различаться по ширине охвата совокупности. Так, например, наряду с общим индексом объема продукции всей промышленности исчисляются индексы объема выпуска по отдельным отраслям. Последние, будучи по своей природе общими индексами, выступают в отношении индекса по всей промышленности в роли групповых (частных) индексов. Иногда груп-

повые индексы называют «субиндексами» по отношению к общеитоговому («тотальному») индексу.

Именно построение общих индексов, относящихся к сложным (агрегированным) совокупностям, связано с определенными проблемами, решение которых и составляет суть индексного метода (индексной теории).

Общие индексы широко используются в статистической практике на различных уровнях – от предприятия до национальной экономики в целом, везде, где требуется обобщить изменения определенного показателя по сложной совокупности.

С помощью общих индексов характеризуется изменение цен на потребительские товары, изменение уровня жизни, развитие производства отдельных отраслей и экономики в целом и многое другое.

Общие индексы позволяют, с одной стороны, обобщать изменения индексируемой величины у отдельных единиц (элементов) или частей совокупности и, с другой стороны, определять (измерять) влияние изменения отдельных факторов на изменение результативного показателя явления в целом (например, влияние изменения урожайности на изменение валового сбора той или иной сельскохозяйственной культуры, влияние изменения цен на изменение товарооборота).

Эти функции индексов явились основанием развития двух концепций индексной теории: синтетической (обобщающей) и аналитической.

Сторонники синтетической концепции трактуют общий индекс как показатель среднего изменения (или изменения в среднем) индексируемой величины в целом по совокупности, а сторонники аналитической теории – как показатель, характеризующий изменение значения результативного показателя за счет изменения индексируемой величины.

С нашей точки зрения, такое деление искусственно и не способствует развитию индексной теории. Представляется, что именно сочетание этих двух функций общих индексов, т.е. возможность решения обеих задач, как обобщения индивидуальных изменений у отдельных элементов, так и выявления роли отдельных факторов в изменении результативного показателя, и является их достоинством и спецификой, позволяющими выделять индексный метод в особый метод статистического исследования, анализа.

Как уже отмечалось, с помощью индексов можно охарактеризовать относительное изменение самых различных показателей.

Эти показатели (индексируемые величины) могут иметь разный характер. Одни являются объемными (количественными); другие условно можно назвать качественными: они представляют собой показатели, определяемые на какую-то единицу (цена единицы товара, себестоимость единицы продукции, урожайность с 1 га и т.д.). В соответствии с этим и индексы можно подразделить на индексы *количественных* показателей (индекс физического объема производства, индекс продаж акций и т.п.) и *качественных* (индекс цен, индекс себестоимости, индекс урожайности, индекс заработной платы и др.).

Каждый из этих индексов имеет свои особенности, но любой общий индекс может быть исчислен двояко: как *агрегатный* и как *средний из индивидуальных*. Рассмотрим оба способа построения (исчисления) общих индексов.

9.2. Агрегатные индексы

Агрегатный способ построения (исчисления) общих индексов сводится к выражению с помощью определенных соизмерителей итогового (суммарного) значения несопоставимых в физических единицах показателей в сложной совокупности («агрегате») и последующему сопоставлению такой суммы в отчетном и базисном периодах.

Рассмотрим построение агрегатного индекса на примере индекса физического объема как наиболее типичного для количественных показателей и на примере индекса цен как наиболее типичного для качественных показателей.

1. Агрегатный индекс физического объема. Допустим, известны данные о производстве различной несоизмеримой в физических единицах продукции на одном предприятии за два периода и необходимо с помощью общего индекса охарактеризовать относительное изменение объема всей продукции в отчетном периоде по сравнению с объемом в предшествующем (базисном) периоде.

Неоднородную продукцию, не допускающую непосредственного суммирования, можно с помощью определенных соизмерителей выразить в одинаковых единицах измерения и, определив в них общий объем изучаемой продукции в отчетном и базисном периодах, найти отношение этих общих объемов.

Чаще всего в качестве такого соизмерителя выступает ц е н а за единицу продукции. Умножая цены на количество произведен-

ной продукции, получаем стоимостное (ценностное) выражение продукции каждого вида, которое допускает суммирование.

Кроме цены, соизмерителем в отдельных случаях может служить себестоимость единицы продукции или затраты труда на единицу продукции.

Общий индекс, полученный путем сопоставления итоговых показателей, количественно выражающих сложное явление в отчетном и базисном периодах с помощью соизмерителей, называют *агрегатным*. Соответственно, и способ исчисления общего индекса таким путем (через соизмерители) называется агрегатным.

Обозначая объем продукции (товаров) через q, а цены – через p, можно представить стоимость продукции в базисном периоде как $\sum q_0 p_0$, а в отчетном – как $\sum q_1 p_1$. Сопоставляя эти два показателя, получим индекс стоимости

$$I_{pq} = \frac{\sum q_1 p_1}{\sum q_0 p_0}, \qquad (9.1)$$

который показывает относительное изменение стоимости продукции как за счет изменения цен, так и за счет изменения объема отдельных товаров.

Если же продукцию двух сравниваемых периодов оценить в одних и тех же неизменных ценах, то очевидно, что стоимость продукции двух периодов будет отличаться лишь за счет изменения объема продукции. Поэтому общий индекс, исчисленный как отношение стоимости продукции двух периодов в одних и тех же ценах, называют *агрегатным индексом физического объема* (обозначается I_q или $I_{\text{ф.об}}$).

В агрегатном индексе физического объема в качестве соизмерителя различных товаров принимаются цены базисного периода p_0 или цены, неизменные в течение ряда лет p. (Такие цены называют также сопоставимыми.) Соответственно, и формулу агрегатного индекса физического объема можно записать двояко*:

$$I_q = \frac{\sum q_1 p_0}{\sum q_0 p_0} \quad \text{или} \quad I_q = \frac{\sum q_1 p}{\sum q_0 p}, \qquad (9.2)$$

где q_0 и q_1 – объем продукции различных видов соответственно в базисном и отчетном периоде.

---

* Здесь и далее формулы индексов записываются упрощенно, без дополнительных значков, обозначающих, что q и p относятся к конкретному j-му товару, т.е. всегда предполагается суммирование по всем товарам.

Заметим, что суммы в числителе и знаменателе формулы (9.2) имеют вполне реальный смысл:

$\sum q_0 p_0$ – стоимость продукции базисного периода в базисных ценах;

$\sum q_1 p_0$ – стоимость продукции отчетного периода в базисных ценах.

Таким образом, чтобы исчислить общий индекс физического объема агрегатным способом, продукцию базисного и отчетного периодов оценивают в одних и тех же сопоставимых (базисных) ценах и делят второй показатель на первый.

Внешней отличительной особенностью любого агрегатного индекса является то, что и в числителе, и в знаменателе данного индекса имеется сумма произведений двух показателей, один из которых меняется, т.е. выступает в роли индексируемой величины, а второй остается неизменным, т.е. выступает в роли соизмерителя (или весов).

Разность между числителем и знаменателем агрегатного индекса характеризует изменение в абсолютном выражении сложного (результативного) показателя за счет изменения индексируемой величины.

Проиллюстрируем расчет агрегатного индекса физического объема на условном примере.

Пример. Предположим, фирма выпускает три вида неоднородной продукции. Данные об их производстве и ценах на них за два периода приведены в табл. 9.1 (графы А, 1–4).

Таблица 9.1

Товар	Выработано тыс. единиц		Цена за единицу товара, руб.		Стоимость продукции в базисных ценах, тыс. руб.	
	базисный период q_0	отчетный период q_1	базисный период p_0	отчетный период p_1	базисный период $q_0 p_0$	отчетный период $q_1 p_0$
А	1	2	3	4	5	6
X	80	60	13	16	1040	780
Y	50	30	18	20	900	540
Z	40	35	6	8	240	210
Σ	—	—	—	—	2180	1530

Чтобы рассчитать агрегатный индекс физического объема, определяем общую стоимость продукции базисного и отчетного пе-

риодов в одних и тех же базисных ценах (см. графы 5 и 6) и сопоставляем вторую с первой:

$$I_q = \frac{\sum q_1 p_0}{\sum q_0 p_0} = \frac{1530}{2180} = 0{,}702 \text{ (или } 70{,}2\%).$$

Это означает, что общий объем (выпуск) продукции в отчетном периоде по сравнению с базисным составил 70,2% (или уменьшился на 29,8% (70,2 – 100)).

Вычитая из числителя знаменатель ($\sum q_1 p_0 - \sum q_0 p_0 = 1530 - 2180 = -650$), определяем, что в абсолютном выражении за счет уменьшения выпуска стоимость продукции в отчетном периоде уменьшилась на 650 тыс. руб.

Как уже отмечалось, при построении агрегатного индекса физического объема могут использоваться и другие соизмерители. Так, например, если принять в качестве соизмерителей себестоимость единицы продукции в базисном периоде c_0, то агрегатный индекс физического объема можно записать как $I_q = \frac{\sum q_1 c_0}{\sum q_0 c_0}$. Тогда разность между числителем и знаменателем $\sum q_1 c_0 - \sum q_0 c_0$ покажет, как изменились общие затраты (издержки) на производство в связи с изменением выпуска продукции.

Если в качестве соизмерителей принять затраты времени на единицу продукции в базисном периоде t_0, то формула агрегатного индекса физического объема будет иметь вид $I_q = \frac{\sum q_1 t_0}{\sum q_0 t_0}$, а разность $\sum q_1 t_0 - \sum q_0 t_0$ будет характеризовать изменение общих затрат времени на производство продукции за счет изменения объема выпуска.

2. Агрегатный индекс цен. По аналогии с индексом физического объема для определенного набора товаров (продуктов) может быть построен и агрегатный индекс цен (индекс качественного показателя). При этом рассуждения остаются теми же: если нельзя суммировать цены на различные товары, то можно суммировать и сопоставлять стоимости этих товаров.

Однако, сопоставляя два значения стоимости pq, мы должны показать изменение последней лишь за счет изменения цен p, т.е. необходимо устранить влияние изменения количества производимой (или реализуемой) в разные периоды продукции q на стоимостный показатель продукции. Для этого один и тот же коли-

чественный набор продуктов надо оценить в ценах отчетного и базисного периодов и затем сопоставить первую величину со второй. Таким образом, в агрегатном индексе цен индексируемой величиной является, естественно, цена p, а соизмерителем (вернее, весами) – количество произведенных (реализованных) товаров q, принятое на уровне базисного или отчетного периода.

Агрегатная формула общего индекса цен была впервые предложена в 1864 г. немецким ученым Э. Ласпейресом. Он предлагал строить агрегатный индекс цен, приняв в качестве весов продукцию базисного периода q_0:

$$I_p^{\text{Л}} = \frac{\sum q_0 p_1}{\sum q_0 p_0}. \qquad (9.3)$$

В таком виде, т.е. построенный по продукции базисного периода, этот индекс известен как ***индекс цен Ласпейреса***.

В 1874 г. другой немецкий ученый, Г. Пааше, предложил строить агрегатный индекс цен по продукции текущего периода q_1:

$$I_p^{\text{П}} = \frac{\sum q_1 p_1}{\sum q_1 p_0}. \qquad (9.4)$$

Такой индекс, т.е. построенный по продукции текущего периода, известен как ***индекс цен Пааше***.

Примечание. Впоследствии названия этих разных методов расчета индексов цен (с весами базисного или текущего периодов) были механически распространены и на другие индексы (физического объема и пр.), т.е. если в сводном индексе использованы веса базисного периода, то говорят, что индекс построен по методу Ласпейреса, а при весах текущего периода – по методу Пааше.

На практике используются формулы индексов цен и Ласпейреса и Пааше, хотя они и дают разные результаты. (По значению индекс Ласпейреса, как правило, больше индекса Пааше.)

Каждый из этих индексов имеет свои особенности, которым отдается предпочтение в конкретных условиях использования.

Так, индекс цен Ласпейреса удобен для оперативной (недельной, месячной, квартальной) информации об изменении цен на определенный фиксированный набор товаров, когда пересчет каждый раз на текущий набор (количество) товаров сопряжен с большими затратами труда и времени. По формуле Ласпейреса рассчитывают индекс потребительских цен (ИПЦ).

В то же время формуле Пааше отдается предпочтение, когда индекс цен рассматривается в системе с индексом стоимости и индексом физического объема. В этом случае, чтобы обеспечи-

вать взаимосвязь между индексом стоимости $\frac{\sum q_1 p_1}{\sum q_0 p_0}$, индексом физического объема $\frac{\sum q_1 p_0}{\sum q_0 p_0}$ и индексом цен $\frac{\sum q_1 p_1}{\sum q_1 p_0}$, последний обязательно должен строиться по продукции текущего периода, т.е. как индекс Пааше. В противном случае равенство

$$\frac{\sum q_1 p_1}{\sum q_0 p_0} = \frac{\sum q_1 p_0}{\sum q_0 p_0} \frac{\sum q_1 p_1}{\sum q_1 p_0}$$

не будет достигнуто.

Кроме того, при расчете индекса цен по формуле Пааше, вычитая из числителя знаменатель, легко определить в абсолютном выражении сумму потерь (или прибыли) за счет изменения цен на продукцию отчетного (текущего) периода.

Рассмотрим расчет агрегатных индексов цен на примере.

Пример. Пусть имеются данные о реализации продукции одним из хозяйств за два периода (табл. 9.2, графы А, В, 1–4).

Таблица 9.2

Продукт	Единица измерения	Базисный период		Отчетный период		Стоимость товаров базисного периода, руб.		Стоимость товаров отчетного периода, руб.	
		Продано единиц q_0	Цена за единицу, руб. p_0	Продано единиц q_1	Цена за единицу, руб. p_1	в базисных ценах $q_0 p_0$	в отчетных ценах $q_0 p_1$	в базисных ценах $q_1 p_0$	в отчетных ценах $q_1 p_1$
А	В	1	2	3	4	5	6	7	8
Свинина	кг	1000	80	900	96	80000	96000	72000	86400
Картофель	кг	3000	8	4000	10	24000	30000	32000	40000
Молоко	л	5000	10	6000	11	50000	55000	60000	66000
Σ	—	—	—	—	—	154000	181000	164000	192400

Легко определить изменение цен в отчетном периоде по сравнению с базисным по каждому отдельному продукту, рассчитав индивидуальные индексы цен:

а) по свинине $i_p = \frac{96}{80} = 1{,}2$ (или 120%);

б) по картофелю $i_p = \frac{10}{8} = 1{,}25$ (или 125%);

в) по молоку $i_p = \frac{11}{10} = 1{,}1$ (или 110%).

Чтобы определить, как в среднем изменились цены на все продукты (или какова средняя величина изменения цен на все продукты), рассчитаем сводный (общий) индекс цен в форме агрегатного индекса:

1) по формуле Ласпейреса $I_p^{Л}$;

2) по формуле Пааше $I_p^{П}$.

Необходимые для их расчета суммы (оценка продукции в базисном и отчетном периодах в ценах двух периодов) приведены в графах 5–8 табл. 9.2.

Итак,

$$I_p^{Л} = \frac{\sum q_0 p_1}{\sum q_0 p_0} = \frac{181000}{154000} = 1{,}175 \text{ (или } 117{,}5\%);$$

$$I_p^{П} = \frac{\sum q_1 p_1}{\sum q_1 p_0} = \frac{192400}{164000} = 1{,}173 \text{ (или } 117{,}3\%),$$

т.е. по формуле Ласпейреса цены по всем продуктам выросли в среднем на 17,5% (117,5 – 100), а по формуле Пааше – на 17,3% (117,3 – 100).

Расхождение не очень большое (на 0,2 процентных пункта), но все же есть. Какому же индексу отдать предпочтение? Думается, что на таком уровне исследования (по отдельному хозяйству и совокупности хозяйств) предпочтение следует отдать индексу Пааше, поскольку он показывает реальное изменение стоимости продукции, реализованной в отчетном периоде, за счет изменения цен. В этом индексе числитель $\sum q_1 p_1$ – реальная величина, фактическая выручка, полученная от реализации продукции в отчетном периоде, а знаменатель $\sum q_1 p_0$ – условная величина, показывающая, какой была бы выручка, если бы продукция отчетного периода продавалась по базисным ценам. Разность между ними, т.е. $\sum q_1 p_1 - \sum q_1 p_0 = 192400 - 164000 = 28400$ руб., показывает в данном случае, какую прибыль дополнительно получило хозяйство при реализации продукции в отчетном периоде за счет роста цен.

В формуле же индекса цен Ласпейреса в знаменателе содержится реальная выручка (стоимость) от реализации в базисном периоде $\sum q_0 p_0$, а в числителе – условная величина $\sum q_0 p_1$, характеризующая, какой была бы выручка от реализации продукции базисного периода по ценам отчетного периода. Разность $\sum q_0 p_1 - \sum q_0 p_0$ практически не представляет интереса, так как эта величина слишком отвлеченная: она показывает, насколько изменилась бы выручка (стоимость) в прошлом (базисном) периоде, если бы базисная продукция была реализована по текущим (отчетным) ценам.

Кроме того, при расчете индекса цен по формуле Пааше, как уже отмечалось, легко увязываются изменения трех взаимосвязанных показателей: стоимости (выручки), объема реализации и цен. Так, по данным табл. 9.2 индекс стоимости продукции

$$I_{pq} = \frac{\sum q_1 p_1}{\sum q_0 p_0} = \frac{192400}{154000} = 1{,}249 \text{ (или } 124{,}9\%),$$

т.е. стоимость продукции (выручка от продажи) в отчетном периоде увеличилась на 24,9% (124,9 – 100), что составило в абсолютном выражении 38400 руб.:

$$\sum q_1 p_1 - \sum q_0 p_0 = 192400 - 154000 = 38400 \text{ руб.}$$

Индекс физического объема реализации

$$I_q = \frac{\sum q_1 p_0}{\sum q_0 p_0} = \frac{164000}{154000} = 1{,}065 \text{ (или } 106{,}5\%).$$

В абсолютном выражении увеличение стоимости за счет изменения объема реализации составило 10000 руб.:

$$\sum q_1 p_0 - \sum q_0 p_0 = 164000 - 154000 = 10000 \text{ руб.}$$

Таким образом, имеет место увязка индексов (относительного изменения показателей):

$$I_{pq} = I_p I_q$$

(в нашем примере $1{,}249 = 1{,}173 \cdot 1{,}065$),

а также абсолютных изменений:

$$\sum q_1 p_1 - \sum q_0 p_0 = (\sum q_1 p_1 - \sum q_1 p_0) + (\sum q_1 p_0 - \sum q_0 p_0)$$

(в нашем примере $38400 = 28400 + 10000$),

т.е. общее изменение стоимости продукции равно сумме приростов за счет изменения цен и за счет изменения объема.

В начале XX в. американский экономист И. Фишер предложил вместо формул индексов цен Ласпейреса и Пааше использовать среднюю геометрическую из них, т.е. корень квадратный из произведения индексов цен Ласпейреса и Пааше:

$$I_p^{\Phi} = \sqrt{\frac{\sum q_0 p_1}{\sum q_0 p_0} \frac{\sum q_1 p_1}{\sum q_1 p_0}}. \qquad (9.5)$$

В нашем примере индекс цен Фишера будет равен

$$I_p^{\Phi} = \sqrt{1{,}175 \cdot 1{,}173} = 1{,}174 \text{ (или } 117{,}4\%).$$

Этот индекс Фишер назвал «*идеальным*», поскольку в нем не отдается предпочтение ни продукции базисного периода, ни продукции текущего периода.

Кроме того, этот индекс «обратим» во времени, т.е. если рассчитывать индекс базисного периода к отчетному, он будет равен обратной величине первоначального индекса (т.е. отчетного периода к базисному). Другими словами, перемножение таких «обратных» индексов дает единицу:

$$\sqrt{\frac{\sum q_0 p_1}{\sum q_0 p_0} \frac{\sum q_1 p_1}{\sum q_1 p_0}} \sqrt{\frac{\sum q_0 p_0}{\sum q_0 p_1} \frac{\sum q_1 p_0}{\sum q_1 p_1}} = 1.$$

Однако индекс Фишера из-за его формальности и трудности экономической интерпретации используется редко, в основном при территориальных сопоставлениях.

Мы рассмотрели расчет агрегатных индексов физического объема и цен как наиболее типичных представителей агрегатных индексов соответственно для количественных и качественных индексируемых показателей.

По аналогии можно записать агрегатные индексы для некоторых других показателей.

1. Так, по данным о выпуске q и себестоимости c отдельных видов продукции за два периода можно рассчитать аналогично индексу цен Пааше агрегатный индекс себестоимости

$$I_c = \frac{\sum q_1 c_1}{\sum q_1 c_0}. \qquad (9.6)$$

В этом индексе себестоимость отдельных товаров c – индексируемая величина, а продукция отчетного периода q_1 – вес.

Данный индекс показывает, как меняются в относительном выражении общие затраты на производство $\sum qc$ за счет изменения себестоимости отдельных товаров.

2. Для группы однородных культур (например, зерновых) общий индекс урожайности в агрегатной форме выразится формулой

$$I_y = \frac{\sum y_1 П_1}{\sum y_0 П_1}, \qquad (9.7)$$

где y_0 и y_1 – урожайность отдельных культур соответственно в базисном и текущем периоде;

$П_1$ – посевная площадь под отдельными культурами в текущем периоде.

В целом числитель индекса I_y характеризует фактический валовой сбор данной группы культур в текущем периоде, а знаменатель представляет собой условную величину – валовой сбор группы культур с площади текущего периода при базисном уровне урожайности. Таким образом, индекс урожайности в агрегатном виде характеризует изменение валового сбора на фиксированной площади за счет изменения урожайности сельскохозяйственных культур.

3. Представляется целесообразным остановиться еще на одном индексе – производительности труда.

Если обозначить объем произведенной продукции через Q (в натуральном измерении для однородной продукции это q, а для разнородной продукции в стоимостном выражении это $\sum pq$ в неизменных ценах), а затраты времени на ее производство через T (человекочасы, человекодни, человекомесяцы или средняя численность работников в месяц), то производительность труда можно измерить количеством продукции w (в натуральном или стоимостном выражении в неизменных ценах), вырабатываемой в единицу времени, либо затратами рабочего времени t на единицу продукции.

Первый показатель называют *прямым показателем производительности труда*, а второй – *обратным* или *трудоемкостью*, т.е. прямой показатель производительности труда $w = Q/T$, а обратный (трудоемкость) $t = T/Q$.

Индивидуальные индексы для указанных показателей рассчитываются по следующим формулам:

$$i_w = \frac{w_1}{w_0} \quad \text{и} \quad i_t = \frac{t_1}{t_0}.$$

Однако i_t – это индекс затрат времени на единицу продукции, или индекс трудоемкости. Если же на основе данных t характери-

зуется изменение производительности труда, то берется величина, обратная индексу трудоемкости, или просто базисная трудоемкость сопоставляется с текущей:

$$i_w = \frac{w_1}{w_0} = \frac{t_0}{t_1}.$$

Сводный же индекс производительности труда в агрегатной форме рассчитывается по формуле

$$I_w = \frac{\sum w_1 T_1}{\sum w_0 T_1}, \qquad (9.8)$$

где w_0 и w_1 – выработка продукции в единицу рабочего времени (или одним рабочим) соответственно в базисном и отчетном периодах: в натуральном выражении при однородной продукции или в стоимостном выражении (в сопоставимых ценах) при разнородной продукции;

$\sum w_1 T_1 = \sum Q_1$ – фактический объем продукции, произведенной в отчетном периоде (в натуральном или стоимостном выражении в сопоставимых ценах);

$\sum w_0 T_1$ – условная величина, показывающая, каким был бы выпуск продукции в отчетном периоде при численности работников отчетного периода, но базисной производительности труда.

Агрегатный индекс производительности труда можно выразить через показатель трудоемкости t:

$$I_w = \frac{\sum q_1 t_0}{\sum q_1 t_1}, \qquad (9.9)$$

где q_1 – выпуск продукции отдельных видов в натуральном выражении в отчетном периоде;

$\sum q_1 t_0$ – условные затраты времени на выпуск продукции отчетного периода при базисной трудоемкости;

$\sum q_1 t_1$ – затраты времени на весь объем продукции в отчетном периоде.

Агрегатные индексы легко интерпретировать, поэтому они считаются основной формой общих (сводных) индексов, но не единственной.

9.3. Средние индексы из индивидуальных (групповых)

Общие индексы могут быть исчислены не только как агрегатные, но и как средние из индивидуальных или групповых. Например, если имеются данные об изменении цен на конкретные товары, то, естественно, из таких индивидуальных индексов могут быть рассчитаны общие (сводные) индексы как средние величины, причем взвешенные.

Поскольку существует несколько форм (видов) средних величин, то при расчете средних индексов прежде всего возникает вопрос о форме средней и о весах.

В статистической практике средние индексы рассчитываются преимущественно в форме среднего арифметического и среднего гармонического индексов:

$$\bar{I}_{\text{арифм}} = \frac{\sum if}{\sum f} \quad \text{и} \quad \bar{I}_{\text{гарм}} = \frac{\sum M}{\sum \frac{M}{i}},$$

где i – индивидуальные индексы изучаемого показателя (индексируемой величины);

f и M – веса соответственно в среднем арифметическом и среднем гармоническом индексе.

Веса для среднего арифметического и среднего гармонического индексов определяются исходя из тождества их агрегатному, который, как указывалось, является основной формой общего индекса. При этом для каждого конкретного индекса веса особые.

Рассмотрим индекс физического объема и индекс цен.

1. Индекс физического объема. Если речь идет об индексе физического объема, то при исчислении среднего арифметического индекса должно выполняться следующее тождество:

$$\frac{\sum i_q f}{\sum f} = \frac{\sum q_1 p_0}{\sum q_0 p_0}.$$

Это тождество будет иметь место, если $f = q_0 p_0$. Тогда

$$\bar{I}_{\text{арифм}} = \frac{\sum i_q q_0 p_0}{\sum q_0 p_0} = \frac{\sum \frac{q_1}{q_0} q_0 p_0}{\sum q_0 p_0} = \frac{\sum q_1 p_0}{\sum q_0 p_0}.$$

Таким образом, общий индекс физического объема в форме *среднего арифметического индекса* будет иметь вид

$$\bar{I}_q = \frac{\sum i_q q_0 p_0}{\sum q_0 p_0}. \qquad (9.10)$$

Аналогично, определяя веса среднего гармонического индекса объема, следует помнить о необходимости соблюдения условия

$$\frac{\sum M}{\sum \frac{M}{i_q}} = \frac{\sum q_1 p_0}{\sum q_0 p_0}.$$

Это равенство будет соблюдено, если $M = q_1 p_0$. Тогда

$$I_q = \frac{\sum q_1 p_0}{\sum \frac{q_1 p_0}{i_q}} = \frac{\sum q_1 p_0}{\sum \frac{q_1 p_0}{q_1} q_0} = \frac{\sum q_1 p_0}{\sum q_0 p_0},$$

т.е. *средний гармонический индекс объема* можно записать как

$$\bar{I}_q = \frac{\sum q_1 p_0}{\sum \frac{q_1 p_0}{i_q}}. \qquad (9.11)$$

Преобразовать агрегатный индекс физического объема в средний арифметический и средний гармонический можно и путем следующих простых подстановок.

Исходя из индивидуального индекса объема $i_q = q_1/q_0$ выражаем продукцию отчетного периода: $q_1 = i_q q_0$. Подставляя данное выражение в числитель агрегатной формулы, получаем общий индекс физического объема в форме среднего арифметического индекса:

$$I_q = \frac{\sum q_1 p_0}{\sum q_0 p_0} = \frac{\sum i_q q_0 p_0}{\sum q_0 p_0}.$$

Аналогично, выражая продукцию базисного периода как $q_0 = q_1/i_q$, осуществляем замену в знаменателе агрегатного индекса физического объема. В результате получаем общий индекс физического объема в форме среднего гармонического индекса:

$$I_q = \frac{\sum q_1 p_0}{\sum q_0 p_0} = \frac{\sum q_1 p_0}{\sum \frac{q_1 p_0}{i_q}}.$$

Такое преобразование наглядно показывает тождество между агрегатным индексом и средним арифметическим и средним гармоническим индексами физического объема.

Как видно из приведенных формул, весами индивидуальных индексов объема в среднем арифметическом индексе служит стоимость продукции базисного периода в базисных (или сопоставимых) ценах q_0p_0, а в среднем гармоническом индексе – стоимость продукции отчетного периода в базисных (или сопоставимых) ценах q_1p_0.

При решении конкретных задач выбор той или иной формы среднего индекса определяется прежде всего тем, какие исходные данные имеются в распоряжении исследователя. Так, если известны индивидуальные индексы объема и стоимость продукции базисного периода в базисных ценах, т.е. q_0p_0, общий индекс физического объема можно рассчитать как средний арифметический из индивидуальных.

Рассмотрим расчет такого индекса на условном примере.

Пример. Имеются данные об изменении выпуска продукции трех видов на каком-либо предприятии (табл. 9.3). Требуется найти общий индекс физического объема.

Индивидуальные индексы физического объема отдельных видов продукции (изделий) приведены в графе 1. Рассчитаем из них общий индекс физического объема как средний арифметический, приняв в качестве весов стоимость продукции базисного периода, т.е.

$$\bar{I}_q = \frac{\sum i_q q_0 p_0}{\sum q_0 p_0} = \frac{1{,}03 \cdot 500 + 1{,}06 \cdot 800 + 1{,}04 \cdot 700}{500 + 800 + 700} =$$

$$= 1{,}0455 \text{ (или } 104{,}55\%).$$

Это означает, что в целом по всем изделиям выпуск увеличился на 4,55% (104,55 – 100).

Таблица 9.3

Выпуск продукции на предприятии

Изделие	Индивидуальный индекс объема продукции $i_q = \frac{q_1}{q_0}$	Стоимость продукции базисного периода в базисных ценах, тыс. руб. q_0p_0	Доля изделий в общей стоимости продукции базисного периода $d_0 = \frac{q_0p_0}{\sum q_0p_0}$
А	1	2	3
X	1,03	500	0,25
Y	1,06	800	0,40
Z	1,04	700	0,35
Σ	—	2000	1,00

Вместо абсолютных данных о стоимости отдельных изделий в базисном периоде можно принимать их долю (удельный вес) в общей стоимости, т.е.

$$d_0 = \frac{q_0 p_0}{\sum q_0 p_0}.$$

Тогда формула среднего арифметического индекса из индивидуальных будет иметь вид $\bar{I}_q = \sum i_q d_0$, поскольку $\sum d_0 = 1$.

В нашем примере расчет по этой формуле дает тот же результат, что и по абсолютным данным:

$$\bar{I}_q = \sum i_q d_0 = 1{,}03 \cdot 0{,}25 + 1{,}06 \cdot 0{,}4 + 1{,}04 \cdot 0{,}35 = 1{,}0455 \text{ (или } 104{,}55\%).$$

(Если d_0 выражено в процентах, формула среднего арифметического индекса такова: $\bar{I}_q = \frac{\sum i_q d_0}{100\%}$.)

Пример. Пусть наряду с индивидуальными индексами объема отдельных товаров известна стоимость продукции отчетного периода в базисных ценах (табл. 9.4).

Таблица 9.4

Выпуск продукции на предприятии

Изделие	Индивидуальный индекс физического объема $i_q = \frac{q_1}{q_0}$	Стоимость продукции отчетного периода в базисных ценах, тыс. руб. $q_1 p_0$
X	0,98	392
Y	0,96	912
Z	0,95	684

В этом случае общий индекс физического объема следует рассчитывать как средний гармонический, где весами служит стоимость отдельных видов продукции отчетного периода в базисных ценах:

$$\bar{I}_q = \frac{\sum q_1 p_0}{\sum \frac{q_1 p_0}{i_q}} = \frac{392 + 912 + 684}{\frac{392}{0{,}98} + \frac{912}{0{,}96} + \frac{684}{0{,}95}} = 0{,}96 \text{ (или } 96\%),$$

т.е. в целом объем продукции уменьшился на 4% (96 – 100).

Однако надо отметить, что применительно к физическому объему средний гармонический индекс практически не применяется. Для динамики физического объема используется агрегатная форма индекса либо средний арифметический индекс, тождественный агрегатному.

Веса для средних индексов по другим показателям также определяются исходя из тождества их агрегатному индексу.

2. Индекс цен. Применительно к индексам цен возможны два варианта взвешивания и для среднего арифметического, и для среднего гармонического индексов, в зависимости от того, по отношению к какому агрегатному индексу рассматривается их тождество: к индексу Ласпейреса или Пааше.

В случае если за исходную принимается формула Ласпейреса $I_p^{Л} = \frac{\sum p_1 q_0}{\sum p_0 q_0}$, заменяя в числителе p_1 на $i_p p_0$ (из $i_p = p_1/p_0$), получим *средний арифметический индекс цен*:

$$\bar{I}_p^{Л} = \frac{\sum i_p p_0 q_0}{\sum p_0 q_0}, \qquad (9.12)$$

где весами служит стоимость отдельных групп продукции базисного периода $p_0 q_0$ (или их доля в общей стоимости продукции базисного периода $d_0 = \frac{p_0 q_0}{\sum p_0 q_0}$. Тогда средний арифметический индекс цен будет $\bar{I}_p^{Л} = \sum i_p d_0$ или $\bar{I}_p^{Л} = \frac{\sum i_p d_0}{100\%}$, если d_0 выражено в процентах. Данная форма среднего индекса используется в статистической практике в России при расчете ИПЦ – индекса потребительских цен).

Если же исходить из агрегатной формулы индекса цен Пааше $I_p^{П} = \frac{\sum p_1 q_1}{\sum p_0 q_1}$, то тождественный ему средний арифметический индекс

$$\bar{I}_p^{П} = \frac{\sum i_p p_0 q_1}{\sum p_0 q_1}, \qquad (9.13)$$

т.е. в этом случае весами для индивидуальных индексов должна служить стоимость продукции отчетного периода в базисных ценах, что не совсем удобно для расчетов на практике.

Чтобы записать формулы среднего гармонического индекса цен, из $i_p = p_1/p_0$ выражаем цену базисного периода: $p_0 = p_1/i_p$ и подставляем это выражение в знаменатель агрегатного индекса цен.

Тогда *средний гармонический индекс цен* по Ласпейресу будет иметь вид

$$\bar{I}_p^{\text{Л}} = \frac{\sum p_1 q_0}{\sum \frac{p_1 q_0}{i_p}}, \tag{9.14}$$

а по Пааше, соответственно,

$$\bar{I}_p^{\text{П}} = \frac{\sum p_1 q_1}{\sum \frac{p_1 q_1}{i_p}}. \tag{9.15}$$

В формуле (9.14) весами служит стоимость отдельных видов продукции базисного периода в отчетных (текущих) ценах $p_1 q_0$, а в формуле (9.15) – стоимость продукции текущего периода в текущих ценах $p_1 q_1$. Формула (9.15) более предпочтительна для практического использования, чем (9.14).

Средние арифметические и средние гармонические индексы являются своего рода модификациями агрегатных индексов, т.е. применительно к индексам цен должны соблюдаться следующие равенства:

а) по Ласпейресу: $\dfrac{\sum p_1 q_0}{\sum p_0 q_0} = \dfrac{\sum i_p p_0 q_0}{\sum p_0 q_0} = \dfrac{\sum p_1 q_0}{\sum \frac{p_1 q_0}{i_p}}$;

б) по Пааше: $\dfrac{\sum p_1 q_1}{\sum p_0 q_1} = \dfrac{\sum i_p p_0 q_1}{\sum p_0 q_1} = \dfrac{\sum p_1 q_1}{\sum \frac{p_1 q_1}{i_p}}$.

Отсюда следует, что значения среднего арифметического и среднего гармонического индексов цен будут совпадать лишь тогда, когда их веса определены из тождества одной и той же агрегатной формуле (по Ласпейресу или Пааше).

Однако средние индексы по Ласпейресу не тождественны одноименным средним индексам по Пааше, как не тождественны и сами агрегатные индексы цен Ласпейреса и Пааше. Поэтому при

расчете сводного (общего) индекса как среднего из индивидуальных необходимо точно указывать, модификацией какого агрегатного индекса является используемый средний индекс, так как это определяет его веса.

Рассмотрим расчет средних индексов цен на условном примере.

Пример. Предположим, имеются данные о реализации продукции какой-либо фирмы за два периода (табл. 9.5, графы А, 1–3).

Таблица 9.5

Данные о реализации продукции фирмы

Товар	Выручка от реализации, тыс. руб.		Изменение цен в отчетном периоде по сравнению с базисным $i_p = \frac{p_1}{p_0}$	Стоимость продукции базисного периода в отчетных ценах $i_p p_0 q_0 = p_1 q_0$	Стоимость продукции отчетного периода в базисных ценах $\frac{p_1 q_1}{i_p} = p_0 q_1$
	в базисном периоде $p_0 q_0$	в отчетном периоде $p_1 q_1$			
А	1	2	3	4	5
X	260	309	1,03	267,8	300
Y	520	636	1,06	551,2	600
Z	420	473	1,10	462,0	430
Σ	1200	1418	—	1281,0	1330

Требуется определить общий (сводный) индекс цен.

Возможно двоякое решение этой задачи.

1. Если для индивидуальных индексов цен (графа 3) в качестве весов использовать фактическую стоимость отдельных товаров в базисном периоде $p_0 q_0$, то сводный индекс цен можно рассчитать как средний арифметический, тождественный агрегатному индексу Ласпейреса, т.е. по формуле (9.10):

$$\bar{I}_p^{\text{Л}} = \frac{\sum i_p p_0 q_0}{\sum p_0 q_0} = \frac{1{,}03 \cdot 260 + 1{,}06 \cdot 520 + 1{,}1 \cdot 420}{260 + 520 + 420} =$$

$$= 1{,}0675 \text{ (или } 106{,}75\%).$$

Это означает, что цены по всем товарам в наборе базисного периода выросли в среднем на 6,75% (106,75 – 100).

2. Если по исходным данным использовать в качестве весов стоимость продукции отчетного периода $p_1 q_1$, то общий индекс

будет рассчитан как средний гармонический индекс цен (по Пааше):

$$\bar{I}_p^{\text{П}} = \frac{\sum p_1 q_1}{\sum \frac{p_1 q_1}{i_p}} = \frac{309 + 636 + 473}{\frac{309}{1{,}03} + \frac{636}{1{,}06} + \frac{473}{1{,}1}} = 1{,}0662 \text{ (или } 106{,}62\%),$$

т.е. по этому индексу цены в среднем выросли на 6,62% (106,62 – – 100).

Как и следовало ожидать, в первом случае (по Ласпейресу) значение индекса несколько выше, чем во втором (по Пааше). Расхождение между ними вызвано тем, что в первом индексе (по Ласпейресу) изменение цен рассматривается по составу продукции базисного периода q_0, а во втором (по Пааше) – по продукции отчетного периода q_1.

Какому индексу отдать предпочтение, зависит от цели исследования. Если нас интересует не только относительное изменение цен, но и абсолютная величина изменения агрегатного показателя (стоимости pq), вызванного изменением цен, то следует отдать предпочтение среднему гармоническому индексу, тождественному агрегатному индексу Пааше, т.е. $\bar{I}_p^{\text{П}}$.

Так, в нашем примере (где знаменатель $\sum \frac{p_1 q_1}{i_p} = \sum p_0 q_1$) разность между числителем и знаменателем $\sum p_1 q_1 - \sum p_0 q_1 = 1418 - - 1330 = 88$ тыс. руб. – реальная дополнительная выручка, полученная фирмой в отчетном периоде вследствие изменения цен на продукцию отчетного периода q_1.

Как правило, при изучении динамики цен на уровне отдельных фирм, отраслей предпочтение отдается агрегатному индексу Пааше или тождественному ему среднему гармоническому индексу цен с весами $p_1 q_1$.

Средний арифметический индекс цен с весами $p_0 q_0$, т.е. тождественный агрегатному индексу цен Ласпейреса, в статистической практике используется в основном при расчете сводного индекса потребительских цен (ИПЦ), публикуемого ежемесячно по фиксированному набору товаров и услуг. В этом индексе (ИПЦ) в качестве весов для индексов цен по отдельным группам товаров (продовольственным, непродовольственным, платным услугам) принимается структура потребительских расходов в базисном периоде по указанным группам товаров и услуг, т.е. их доля $d_{p_0 q_0}$.

Расчет ИПЦ показан на конкретном примере.

Пример. Имеются следующие сведения об изменении цен и тарифов по группам товаров и услуг в 2001 г. и о структуре потребительских расходов в 2000 и 2001 гг. по данным обследования домашних хозяйств (табл. 9.6).

Рассчитать сводный индекс потребительских цен.

Таблица 9.6

Товары и услуги населению	Изменение цен и тарифов в 2001 г., % к декабрю 2000 г.	Структура потребительских расходов населения, %	
		2000 г. ($d_{p_0q_0}$)	2001 г. ($d_{p_1q_1}$)
Продукты питания	+17,8	49,4	48,4
Алкогольные напитки	+12,6	2,5	2,4
Непродовольственные товары	+12,7	34,3	34,4
Платные услуги населению	+36,9	13,8	14,8
Всего	?	100,0	100,0

Сводный индекс потребительских цен можно рассчитать по следующим формулам:

$$\text{ИПЦ} = \frac{\sum i_p p_0 q_0}{\sum p_0 q_0}; \quad \text{ИПЦ} = \frac{\sum i_p d_{p_0 q_0}}{\sum d_{p_0 q_0}}.$$

Так как в нашем примере отсутствуют абсолютные данные о стоимости товаров по группам в базисном периоде ($q_0 p_0$), но имеется их доля в общих потребительских расходах населения ($d_{p_0 q_0}$), то мы используем вторую формулу:

$$\text{ИПЦ} = \frac{\sum i_p d_{p_0 q_0}}{\sum d_{p_0 q_0}} = \frac{1{,}178 \cdot 49{,}4 + 1{,}126 \cdot 2{,}5 + 1{,}127 \cdot 34{,}3 + 1{,}369 \cdot 13{,}8}{100} =$$

$$= 1{,}185565 \approx 1{,}186 \text{ (или } 118{,}6\%).$$

Сводный индекс потребительских цен, равный 118,6%, означает, что по всем группам товаров и услуг цены и тарифы выросли на 18,6% (118,6 – 100 = 18,6).

В данном параграфе, как и в параграфе 9.2, рассмотрены только принципы построения общих (сводных) индексов как агрегатных или средних из индивидуальных.

В статистической практике рассчитываются индексы цен различного характера, в различных отраслях и на различных уровнях. Это индексы цен производителей промышленной продукции, индексы цен на приобретаемые виды материально-технических ресурсов, индексы цен на продукцию, реализованную сельскохозяйственными предприятиями, индексы цен на приобретенную сельскохозяйственными предприятиями промышленную продукцию и услуги, индексы цен по капитальным вложениям и т.д.

Для расчета сводных индексов на разных уровнях органами статистики разработаны рекомендации, которые касаются выбора формы индекса, весов, набора товаров и пр. (см. сборник Госкомстата России «Цены в России. 1998»).

9.4. Индексы переменного и фиксированного составов. Индекс структурных сдвигов

При изучении качественных показателей часто приходится рассматривать изменение во времени (или пространстве) с р е д н е й величины индексируемого показателя для определенной однородной совокупности. Например, в статистических сборниках публикуются данные о динамике средних цен на определенные продукты, средней урожайности зерновых культур, средней номинальной заработной плате в отдельных отраслях экономики и т.д.

Будучи сводной характеристикой качественного показателя, средняя величина складывается как под влиянием значений показателя у индивидуальных элементов (единиц), из которых состоит объект, так и под влиянием соотношения их весов («структуры» объекта).

Если любой качественный индексируемый показатель обозначить через x, а его веса – через f, то динамику среднего показателя можно отразить как за счет изменения обоих факторов (x и f), так и за счет каждого фактора отдельно. В результате получим три различных индекса:

- индекс переменного состава;
- индекс фиксированного состава;
- индекс структурных сдвигов.

Индекс переменного состава отражает динамику среднего показателя (для однородной совокупности) за счет изменения *индексируемой величины* x у отдельных элементов (частей целого) и за счет изменения *весов* f, по которым взвешиваются отдельные значения x. Любой индекс переменного состава – это отношение двух

средних величин для однородной совокупности (за два периода или по двум территориям):

$$I_{\text{п.с}} = \bar{x}_1 : \bar{x}_0 = \frac{\sum x_1 f_1}{\sum f_1} : \frac{\sum x_0 f_0}{\sum f_0}. \qquad (9.16)$$

Свое название этот индекс получил потому, что он характеризует динамику средних величин не только за счет изменения индексируемой величины у отдельных элементов (частей целого), но и за счет изменения удельного веса этих частей в общей совокупности, т.е. изменения с о с т а в а совокупности.

Например, средняя себестоимость определенного вида продукции, выпускаемой на разных предприятиях, зависит как от уровня себестоимости на отдельных предприятиях, так и от количества продукции, выпускаемой этими предприятиями. Поэтому индекс себестоимости переменного состава отражает изменение средней себестоимости определенного продукта как за счет изменения себестоимости на каждом предприятии, так и за счет изменения удельного веса отдельных предприятий в общем выпуске продукции.

Индекс фиксированного состава отражает динамику среднего показателя лишь за счет изменения *индексируемой величины x*, при фиксировании весов на уровне, как правило, отчетного периода f_1:

$$I_{\text{ф.с}} = \frac{\sum x_1 f_1}{\sum f_1} : \frac{\sum x_0 f_1}{\sum f_1}. \qquad (9.17)$$

Другими словами, индекс фиксированного состава исключает влияние изменения структуры (состава) совокупности на динамику средних величин, т.е. он характеризует динамику с р е д н и х величин, рассчитанных для двух периодов по одной и той же фиксированной структуре весов.

По аналогии можно показать динамику среднего показателя лишь за счет изменения *весов f* при фиксировании индексируемой величины на уровне базисного периода x_0. Такой индекс условно назван ***индексом структурных сдвигов*** ($I_{\text{стр}}$):

$$I_{\text{стр}} = \frac{\sum x_0 f_1}{\sum f_1} : \frac{\sum x_0 f_0}{\sum f_0}. \qquad (9.18)$$

Если от абсолютных весов перейти к относительным весам ($d = \frac{f_i}{\sum f_i}$ и $\sum d = 1$), формулы индексов средних величин примут следующий вид:

$$I_{\text{п.с}} = \frac{\sum x_1 d_1}{\sum x_0 d_0}; \quad I_{\text{ф.с}} = \frac{\sum x_1 d_1}{\sum x_0 d_1}; \quad I_{\text{стр}} = \frac{\sum x_0 d_1}{\sum x_0 d_0}.$$

Все три формулы отражают динамику среднего показателя определенной индексируемой величины x, но в каждой из них видно, влияние какого фактора учитывается при динамике среднего показателя.

Нетрудно заметить, что индекс переменного состава есть произведение индекса фиксированного состава на индекс структурных сдвигов. Таким образом, индекс структурных сдвигов можно рассчитать путем деления индекса переменного состава на индекс фиксированного состава:

$$I_{\text{стр}} = I_{\text{п.с}} : I_{\text{ф.с}}.$$

Как отмечалось в параграфе 9.1, для обозначения различных показателей в индексном методе используется определенная символика. Пользуясь ею (вместо x и f), можно записать формулы индексов переменного и фиксированного составов, а также индекса структурных сдвигов для конкретных индексируемых показателей.

Рассмотрим эти формулы для некоторых конкретных показателей.

1. Индекс себестоимости. Предположим, что определенный вид продукции производится на нескольких предприятиях. Если обозначить себестоимость единицы продукции через c, а выпуск продукции отдельных предприятий (как веса) через q, можно следующим образом записать формулу *индекса себестоимости переменного состава*:

$$I^{c}_{\text{п.с}} = \frac{\sum c_1 q_1}{\sum q_1} : \frac{\sum c_0 q_0}{\sum q_0} = \bar{c}_1 : \bar{c}_0. \qquad (9.19)$$

Он характеризует изменение средней себестоимости единицы данной продукции по совокупности предприятий за счет изменения c и q на каждом предприятии.

Индекс себестоимости фиксированного состава, характеризующий динамику средних показателей при одной и той же фиксированной структуре совокупности q_1, выразится формулой

$$I^{c}_{\text{ф.с}} = \frac{\sum c_1 q_1}{\sum q_1} : \frac{\sum c_0 q_1}{\sum q_1}. \qquad (9.20а)$$

После сокращения на $\sum q_1$ этот индекс принимает вид формулы агрегатного индекса себестоимости:

$$I^{c}_{\text{ф.с}} = \frac{\sum c_1 q_1}{\sum c_0 q_1}. \qquad (9.20б)$$

В этом индексе устранено влияние структурного фактора (удельного веса отдельных предприятий в общем выпуске продукции) на динамику средней себестоимости; он практически определяет среднее изменение себестоимости данного вида продукции по совокупности предприятий.

Индекс фиксированного состава, в отличие от индекса переменного состава, не может выходить за пределы значений групповых (индивидуальных) индексов, так как является средним из них.

Индекс структурных сдвигов применительно к показателю себестоимости выражается формулой

$$I^{c}_{\text{стр}} = \frac{\sum c_0 q_1}{\sum q_1} : \frac{\sum c_0 q_0}{\sum q_0}. \qquad (9.21)$$

Он характеризует изменение средней себестоимости за два периода, рассчитанной для разной структуры совокупности q и при постоянной себестоимости на уровне базисного периода c_0.

Как уже отмечалось, этот индекс можно получить и путем деления индекса переменного состава на индекс фиксированного состава:

$$I^{c}_{\text{стр}} = I^{c}_{\text{п.с}} : I^{c}_{\text{ф.с}} = \left(\frac{\sum c_1 q_1}{\sum q_1} : \frac{\sum c_0 q_0}{\sum q_0} \right) : \frac{\sum c_1 q_1}{\sum c_0 q_1} = \frac{\sum c_0 q_1}{\sum q_1} : \frac{\sum c_0 q_0}{\sum q_0}.$$

Рассмотрим расчет индексов себестоимости переменного и фиксированного составов, а также индекса структурных сдвигов на конкретном примере.

Пример. Допустим, имеются данные о выпуске и себестоимости одноименного продукта по трем предприятиям (табл. 9.7).

Таблица 9.7

Номер предприятия	Базисный период			Отчетный период		
	Выпуск продукции		Себестоимость единицы продукции, руб.	Выпуск продукции		Себестоимость единицы продукции, руб.
	тыс. единиц	в долях к итогу		тыс. единиц	в долях к итогу	
	q_0	d_0	c_0	q_1	d_1	c_1
1	10	0,50	15	10	0,40	14,2
2	6	0,30	13	7	0,28	12,5
3	4	0,20	10	8	0,32	9,5
Σ	20	1,00	($\bar{c}_0$ = 13,4)	25	1,00	($\bar{c}_1$ = 12,22)

Требуется определить изменение себестоимости единицы продукции на каждом предприятии, а также в целом по всем предприятиям с помощью индексов: а) переменного состава, б) фиксированного состава, в) структурных сдвигов.

Расчет индивидуальных индексов себестоимости продукции по каждому предприятию дает следующие результаты:

$$\text{по 1-му предприятию } i_c = \frac{14{,}2}{15} = 0{,}947 \text{ (или 94,7\%)};$$

$$\text{по 2-му предприятию } i_c = \frac{12{,}5}{13} = 0{,}961 \text{ (или 96,1\%)};$$

$$\text{по 3-му предприятию } i_c = \frac{9{,}5}{10} = 0{,}95 \text{ (или 95\%)}.$$

Чтобы определить индекс себестоимости переменного состава ($I^c_{\text{п.с}} = \bar{c}_1 : \bar{c}_0$), рассчитаем среднюю себестоимость единицы данного вида продукции по трем предприятиям в отчетном и базисном периодах как среднюю арифметическую взвешенную. Итак, средняя себестоимость в базисном периоде

$$\bar{c}_0 = \frac{\sum c_0 q_0}{\sum q_0} = \frac{15 \cdot 10 + 13 \cdot 6 + 10 \cdot 4}{10 + 6 + 4} = \frac{268}{20} = 13{,}4 \text{ руб.},$$

а средняя себестоимость в отчетном периоде

$$\bar{c}_1 = \frac{\sum c_1 q_1}{\sum q_1} = \frac{14{,}2 \cdot 10 + 12{,}5 \cdot 7 + 9{,}5 \cdot 8}{10 + 7 + 8} = \frac{305{,}5}{25} = 12{,}22 \text{ руб.}$$

Сопоставляя их, получаем индекс себестоимости переменного состава:

$$I^c_{\text{п.с}} = \bar{c}_1 : \bar{c}_0 = 12{,}22 : 13{,}4 = 0{,}912 \text{ (или 91,2\%)},$$

т.е. средняя себестоимость единицы изделия снизилась на 8,8% (91,2 – 100).

Если бы выпуск продукции по отдельным предприятиям оставался без изменения или изменился всюду пропорционально, т.е. если бы удельный вес каждого предприятия в выпуске продукции оставался неизменным, то тогда, очевидно, снижение средней себестоимости на 8,8% можно было бы объяснить только снижением себестоимости на каждом предприятии. Фактически же в нашем примере менялась не только себестоимость на каждом предприятии, но и удельный вес каждого предприятия в общем выпуске продукции.

Следовательно, снижение средней себестоимости на 8,8% достигнуто за счет изменения двух факторов (c и q).

В нашем примере общий индекс переменного состава меньше, чем каждый из индивидуальных индексов, т.е. снижение средней себестоимости (8,8%) оказалось больше, чем снижение себестоимости на отдельных предприятиях (5,3; 3,9 и 5%). Очевидно, что это можно объяснить изменением структуры выпуска, в частности увеличением доли 3-го предприятия, имеющего самую низкую себестоимость.

Чтобы исключить влияние изменения структуры совокупности на динамику средних величин, рассчитаем индекс себестоимости фиксированного состава по одной из двух формул (9.20) (приняв в качестве фиксированной структуру выпуска отчетного периода q_1):

$$\text{а) } I^c_{\text{ф.с}} = \frac{\sum c_1 q_1}{\sum q_1} : \frac{\sum c_0 q_1}{\sum q_1} = 12{,}22 : \frac{15 \cdot 10 + 13 \cdot 7 + 10 \cdot 8}{10 + 7 + 8} =$$

$$= 12{,}22 : 1{,}84 = 0{,}952 \text{ (или } 95{,}2\%);$$

$$\text{б) } I^c_{\text{ф.с}} = \frac{\sum c_1 q_1}{\sum c_0 q_1} = \frac{14{,}2 \cdot 10 + 12{,}5 \cdot 7 + 9{,}5 \cdot 8}{15 \cdot 10 + 13 \cdot 7 + 10 \cdot 8} = \frac{305{,}5}{321} =$$

$$= 0{,}952 \text{ (или } 95{,}2\%).$$

Индекс фиксированного состава, характеризуя среднее изменение себестоимости на всех трех предприятиях, не может выходить за пределы значений индивидуальных индексов себестоимости по отдельным предприятиям, что подтверждается и нашим примером.

Влияние структурного фактора на динамику средней себестоимости отразим с помощью индекса структурных сдвигов, рассчитав его двояко:

а) как отношение индексов себестоимости переменного и фиксированного составов:

$$I^c_{\text{стр}} = I^c_{\text{п.с}} : I^c_{\text{ф.с}} = 0{,}912 : 0{,}952 = 0{,}958 \text{ (или } 95{,}8\%);$$

б) по формуле (9.21):

$$I^c_{\text{стр}} = \frac{\sum c_0 q_1}{\sum q_1} : \frac{\sum c_0 q_0}{\sum q_0} = 12{,}84 : 13{,}4 = 0{,}958 \text{ (или } 95{,}8\%).$$

Данный результат означает, что на 4,2% (95,8 – 100) средняя себестоимость снизилась за счет структурного фактора, в частности за счет увеличения доли продукции на 3-м предприятии с более низкой себестоимостью и за счет уменьшения доли выпуска на 1-м предприятии с более высокой себестоимостью.

Вместо абсолютных данных о выпуске продукции q можно воспользоваться в качестве весов их долями по предприятиям d. Тогда

$$I^{c}_{\text{п.с}} = \frac{\sum c_1 d_1}{\sum c_0 d_0} = \frac{14{,}2 \cdot 0{,}4 + 12{,}5 \cdot 0{,}28 + 9{,}5 \cdot 0{,}32}{15 \cdot 0{,}5 + 13 \cdot 0{,}3 + 10 \cdot 0{,}2} =$$

$$= \frac{12{,}22}{13{,}4} = 0{,}912 \text{ (или } 91{,}2\%);$$

$$I^{c}_{\text{ф.с}} = \frac{\sum c_1 d_1}{\sum c_0 d_1} = \frac{12{,}22}{15 \cdot 0{,}4 + 13 \cdot 0{,}28 + 10 \cdot 0{,}32} =$$

$$= \frac{12{,}22}{12{,}84} = 0{,}952 \text{ (или } 95{,}2\%);$$

$$I^{c}_{\text{стр}} = \frac{\sum c_0 d_1}{\sum c_0 d_0} = \frac{12{,}84}{13{,}4} = 0{,}958 \text{ (или } 95{,}8\%),$$

т.е. ответы получены такие же, что и при расчете по абсолютным весам.

Аналогично можно записать формулы индексов переменного и фиксированного составов и для других показателей.

2. Индекс цен. Динамику средних цен одного вида продукции характеризуют следующие индексы:

1) *индекс цен переменного состава*

$$I^{p}_{\text{п.с}} = \frac{\sum p_1 q_1}{\sum q_1} : \frac{\sum p_0 q_0}{\sum q_0} = \bar{p}_1 : \bar{p}_0. \qquad (9.22)$$

Этот индекс показывает, как изменилась средняя цена определенного вида товара, реализованная по разным ценам на разных рынках, за счет двух факторов: p – изменения цен на отдельных рынках и q – изменения количества (доли) товаров, реализованных на разных рынках;

2) *индекс цен фиксированного состава*

$$I^{p}_{\text{ф.с}} = \frac{\sum p_1 q_1}{\sum q_1} : \frac{\sum p_0 q_1}{\sum q_1}, \qquad (9.23)$$

или (после сокращения на $\sum q_1$)

$$I^{p}_{\text{ф.с}} = \frac{\sum p_1 q_1}{\sum p_0 q_1}.$$

Этот индекс, устраняя влияние структурного фактора на динамику средних цен, определяет среднее изменение цен на данный товар на всех рынках, т.е. по всей совокупности реализованной продукции;

3) *индекс структурных сдвигов*

$$I^{p}_{\text{стр}} = I^{p}_{\text{п.с}} : I^{p}_{\text{ф.с}} = \frac{\sum p_0 q_1}{\sum q_1} : \frac{\sum p_0 q_0}{\sum q_0}. \qquad (9.24)$$

Данный индекс характеризует изменение средней цены товара за счет структурного фактора, т.е. изменения долей продукции, реализованной по разным ценам.

В формулах (9.22)–(9.24) p_0 и p_1 – цены реализации товара в разных пунктах в базисном и отчетном периодах; q_0 и q_1 – количество реализованной продукции в разных пунктах в базисном и отчетном периодах.

(Расчет индексов цен переменного, фиксированного составов и индекса структурных сдвигов приведен далее, на с. 407–410.)

3. Индекс урожайности. По такой же схеме можно записать индексы, характеризующие динамику средней урожайности группы однородных культур (например, зерновых):

1) *индекс урожайности переменного состава*

$$I^{y}_{\text{п.с}} = \bar{y}_1 : \bar{y}_0 = \frac{\sum y_1 \Pi_1}{\sum \Pi_1} : \frac{\sum y_0 \Pi_0}{\sum \Pi_0}, \qquad (9.25)$$

где $\bar{y}_0$ и $\bar{y}_1$ – средняя урожайность соответственно в базисном и отчетном периоде;

y_0 и y_1 – урожайность отдельных культур соответственно в базисном и отчетном периоде;

Π_0 и Π_1 – площадь под отдельными культурами соответственно в базисном и отчетном периоде.

Индекс урожайности переменного состава отражает изменение средней урожайности группы однородных культур (например, зерновых) как за счет изменения урожайности отдельных культур, так и за счет изменения структуры посевных площадей. Аналогично речь может идти о динамике средней урожайности какой-либо одной культуры по группе хозяйств, районов и т.д.;

2) *индекс урожайности фиксированного состава*

$$I^{y}_{\text{ф.с}} = \frac{\sum y_1 \Pi_1}{\sum \Pi_1} : \frac{\sum y_0 \Pi_1}{\sum \Pi_1}, \qquad (9.26)$$

где $\frac{\sum y_1 \Pi_1}{\sum \Pi_1}$ – реальная средняя урожайность данной группы культур в отчетном периоде;

$\frac{\sum y_0 \Pi_1}{\sum \Pi_1}$ – условная величина, характеризующая, какой была бы средняя урожайность в базисном периоде при отчетной структуре посевных площадей.

Сократив на $\sum \Pi_1$ [см. формулу (9.26)], получим индекс урожайности фиксированного состава в форме агрегатного индекса:

$$I^{y}_{\text{ф.с}} = \frac{\sum y_1 \Pi_1}{\sum y_0 \Pi_1};$$

3) разделив индекс урожайности переменного состава на индекс фиксированного состава, получим *индекс структурных сдвигов*, характеризующий динамику средней урожайности за счет изменения структуры посевных площадей:

$$I^{y}_{\text{стр}} = \frac{\sum y_0 \Pi_1}{\sum \Pi_1} : \frac{\sum y_0 \Pi_0}{\sum \Pi_0}. \qquad (9.27)$$

4. Индекс производительности труда. При изучении динамики средней производительности труда (по совокупности предприятий, отраслей и пр.) индексы переменного и фиксированного составов (а также индекс структурных сдвигов) могут иметь разный вид в зависимости от того, в каких единицах измерения выражена продукция (Q), на основе которой рассчитывается уровень производительности труда (выработка на одного работника или в определенную единицу времени: человекочас, человекодень), а также от того, каким показателем характеризуется производительность труда: прямым (по выработке – w) или обратным (по трудоемкости – t). Можно выделить следующие основные виды индексов производительности труда.

Для прямых показателей (по выработке):

а) *натуральные индексы* (для однородной продукции, допускающей суммирование в натуральном выражении ($\sum q$), и где $w = \frac{q}{T}$):

$$I^{w}_{\text{п.с}} = \overline{w}_1 : \overline{w}_0 = \frac{\sum q_1}{\sum T_1} : \frac{\sum q_0}{\sum T_0} = \frac{\sum w_1 T_1}{\sum T_1} : \frac{\sum w_0 T_0}{\sum T_0};$$

$$I^{w}_{\text{ф.с}} = \frac{\sum w_1 T_1}{\sum T_1} : \frac{\sum w_0 T_1}{\sum T_1} = \frac{\sum w_1 T_1}{\sum w_0 T_1};$$

$$I^{w}_{\text{стр}} = \frac{\sum w_0 T_1}{\sum T_1} : \frac{\sum w_0 T_0}{\sum T_0}, \quad \text{или} \quad I_{\text{стр}} = I_{\text{п.с}} : I_{\text{ф.с}};$$

б) *стоимостные индексы* (для разнородной продукции, оцениваемой в стоимостном выражении в неизменных ценах как $\sum Q = \sum pq$, и где w определяется в стоимостном выражении как $w = \frac{\sum pq}{T}$):

$$I^{w}_{\text{п.с}} = \overline{w}_1 : \overline{w}_0 = \frac{\sum q_1 p_0}{\sum T_1} : \frac{\sum q_0 p_0}{\sum T_0} = \frac{\sum w_1 T_1}{\sum T_1} : \frac{\sum w_0 T_0}{\sum T_0};$$

$$I^{w}_{\text{ф.с}} = \frac{\sum w_1 T_1}{\sum T_1} : \frac{\sum w_0 T_1}{\sum T_1} = \frac{\sum w_1 T_1}{\sum w_0 T_1};$$

$$I^{w}_{\text{стр}} = \frac{\sum w_0 T_1}{\sum T_1} : \frac{\sum w_0 T_0}{\sum T_0}, \quad \text{или} \quad I_{\text{стр}} = I_{\text{п.с}} : I_{\text{ф.с}};$$

(Расчет стоимостных индексов производительности труда приведен на с. 413–415.)

Для обратных показателей (по трудоемкости):

$$I^{w}_{\text{п.с}} = \bar{t}_0 : \bar{t}_1 = \frac{\sum T_0}{\sum q_0} : \frac{\sum T_1}{\sum q_1} = \frac{\sum t_0 q_0}{\sum q_0} : \frac{\sum t_1 q_1}{\sum q_1};$$

$$I^{w}_{\text{ф.с}} = \frac{\sum q_1 t_0}{\sum q_1} : \frac{\sum q_1 t_1}{\sum q_1} = \frac{\sum q_1 t_0}{\sum q_1 t_1}$$

или как средний арифметический индекс (по методу академика С.Г. Струмилина)

$$I^{w}_{\text{ф.с}} = \frac{\sum i T_1}{\sum T_1} = \frac{\sum \left(\frac{T_0}{q_0} : \frac{T_1}{q_1} \right) T_1}{\sum T_1};$$

$$I^{w}_{\text{стр}} = \frac{\sum t_0 q_0}{\sum q_0} : \frac{\sum t_0 q_1}{\sum q_1}, \quad \text{или} \quad I_{\text{стр}} = I_{\text{п.с}} : I_{\text{ф.с}}.$$

5. Индекс заработной платы. При изучении динамики средней заработной платы (по совокупности предприятий, отраслей и т.п.), обозначив среднюю заработную плату одного работника в отдельных подразделениях через x, а среднесписочную численность занятых через T, можно записать следующие формулы индексов:

$$I^{\text{з.п}}_{\text{п.с}} = \overline{x}_1 : \overline{x}_0 = \frac{\sum x_1 T_1}{\sum T_1} : \frac{\sum x_0 T_0}{\sum T_0};$$

$$I^{\text{з.п}}_{\text{ф.с}} = \frac{\sum x_1 T_1}{\sum T_1} : \frac{\sum x_0 T_1}{\sum T_1} = \frac{\sum x_1 T_1}{\sum x_0 T_1};$$

$$I^{\text{з.п}}_{\text{стр}} = \frac{\sum x_0 T_1}{\sum T_1} : \frac{\sum x_0 T_0}{\sum T_0}, \quad \text{или} \quad I_{\text{стр}} = I_{\text{п.с}} : I_{\text{ф.с}}.$$

Перечень конкретных индексов переменного и фиксированного составов, а также индексов структурных сдвигов можно было бы продолжить, но, думается, и перечисленных достаточно, чтобы понять их природу и назначение при изучении динамики средних показателей.

В заключение отметим, что индекс переменного состава для любого качественного показателя равен произведению индекса фиксированного состава на индекс структурных сдвигов (структуры):

$$I_{\text{п.с}} = I_{\text{ф.с}} I_{\text{стр}}.$$

Об этом следует помнить при расчете указанных индексов.

9.5. Цепные и базисные индексы

Если известны данные за несколько периодов (больше двух), по ним может быть построен ряд (система) индексов: либо с постоянной для всех базой сравнения, либо с переменной.

Ряд индексов, каждый из которых рассчитан по отношению к предыдущему периоду, называют ***цепными индексами***, а ряд индексов с постоянной базой сравнения – ***базисными***.

В табл. 9.8 приведены в качестве примера цепные и базисные индивидуальные индексы цен и физического объема.

Таблица 9.8

Цепные и базисные индивидуальные индексы

Название индивидуального индекса	Цепные индексы	Базисные индексы
Индекс цен	$\frac{p_1}{p_0}; \frac{p_2}{p_1}; \ldots; \frac{p_t}{p_{t-1}}$	$\frac{p_1}{p_0}; \frac{p_2}{p_0}; \ldots; \frac{p_t}{p_0}$
Индекс физического объема	$\frac{q_1}{q_0}; \frac{q_2}{q_1}; \ldots; \frac{q_t}{q_{t-1}}$	$\frac{q_1}{q_0}; \frac{q_2}{q_0}; \ldots; \frac{q_t}{q_0}$

Между цепными и базисными индексами существует определенная взаимосвязь, что позволяет переходить от одних индексов к другим.

Так, перемножая последовательно цепные индексы, можно получить базисные индексы, например:

$$\frac{p_1}{p_0}\frac{p_2}{p_1} = \frac{p_2}{p_0}; \quad \frac{p_1}{p_0}\frac{p_2}{p_1}\frac{p_3}{p_2} = \frac{p_3}{p_0} \quad \text{и т.д.}$$

или

$$\frac{q_1}{q_0}\frac{q_2}{q_1} = \frac{q_2}{q_0}; \quad \frac{q_1}{q_0}\frac{q_2}{q_1}\frac{q_3}{q_2} = \frac{q_3}{q_0} \text{ и т.д.}$$

В свою очередь, отношение двух последовательных базисных индексов дает цепной индекс, например:

$$\frac{p_2}{p_0} : \frac{p_1}{p_0} = \frac{p_2}{p_1} \quad \text{или} \quad \frac{q_3}{q_0} : \frac{q_2}{q_0} = \frac{q_3}{q_2} \quad \text{и т.д.}$$

Цепные и базисные индексы могут быть построены и для общих индексов. При этом последние могут иметь постоянные и переменные веса.

Если, например, известны данные по предприятию о выпуске q нескольких видов продукции (A, $Б$, $В$ и т.д.) и о ценах p на нее за четыре периода, то при вычислении цепных и базисных общих индексов физического объема и цен можно по-разному решать вопрос о весах (соизмерителях).

Так, при расчете *цепных индексов физического объема* по агрегатной формуле продукцию всех периодов можно оценить в одних и тех же ценах (предположим, в ценах первого периода p_1). Тогда такие цепные индексы будут выглядеть следующим образом:

$$I_{q_{2/1}} = \frac{\sum q_2 p_1}{\sum q_1 p_1}; \quad I_{q_{3/2}} = \frac{\sum q_3 p_1}{\sum q_2 p_1}; \quad I_{q_{4/3}} = \frac{\sum q_4 p_1}{\sum q_3 p_1}.$$

Так как все эти индексы имеют одни и те же соизмерители p_1, они являются индексами с п о с т о я н н ы м и весами.

Вычисляя цепные индексы физического объема, можно было поступить и по-другому: для каждого периода строить индекс объема, принимая в качестве весов цены предыдущего периода. Тогда

$$I_{q_{2/1}} = \frac{\sum q_2 p_1}{\sum q_1 p_1}; \quad I_{q_{3/2}} = \frac{\sum q_3 p_2}{\sum q_2 p_2}; \quad I_{q_{4/3}} = \frac{\sum q_4 p_3}{\sum q_3 p_3}.$$

Эти индексы, построенные по разным соизмерителям, являются индексами физического объема с п е р е м е н н ы м и весами.

Аналогично можно записать в двух вариантах и агрегатные индексы цен.

Цепные индексы цен:

а) с п о с т о я н н ы м и весами (примем в качестве них q_1):

$$I_{p_{2/1}} = \frac{\sum p_2 q_1}{\sum p_1 q_1}; \quad I_{p_{3/2}} = \frac{\sum p_3 q_1}{\sum p_2 q_1}; \quad I_{p_{4/3}} = \frac{\sum p_4 q_1}{\sum p_3 q_1};$$

б) с переменными весами (веса текущего периода):

$$I_{p_{2/1}} = \frac{\sum p_2 q_2}{\sum p_1 q_2}; \quad I_{p_{3/2}} = \frac{\sum p_3 q_3}{\sum p_2 q_3}; \quad I_{p_{4/3}} = \frac{\sum p_4 q_4}{\sum p_3 q_4}.$$

Для общих (агрегатных) индексов переход от цепных индексов к базисным строго математически возможен лишь для индексов с постоянными весами.

Например, на основе записанных выше цепных индексов физического объема с постоянными весами путем перемножения их легко получить соответствующий *базисный индекс физического объема*:

$$\frac{\sum q_2 p_1}{\sum q_1 p_1} \frac{\sum q_3 p_1}{\sum q_2 p_1} = \frac{\sum q_3 p_1}{\sum q_1 p_1} \quad \text{или} \quad \frac{\sum q_2 p_1}{\sum q_1 p_1} \frac{\sum q_3 p_1}{\sum q_2 p_1} \frac{\sum q_4 p_1}{\sum q_3 p_1} = \frac{\sum q_4 p_1}{\sum q_1 p_1}.$$

То же самое для индексов цен с постоянными весами:

$$\frac{\sum p_2 q_1}{\sum p_1 q_1} \frac{\sum p_3 q_1}{\sum p_2 q_1} \frac{\sum p_4 q_1}{\sum p_3 q_1} = \frac{\sum p_4 q_1}{\sum p_1 q_1},$$

т.е. произведение цепных индексов цен с постоянными весами дает *базисный индекс цен* (четвертого периода к первому).

Перемножение же цепных индексов цен с переменными весами не дает базисного индекса, т.е. математически такого равенства не будет:

$$\frac{\sum p_2 q_2}{\sum p_1 q_2} \frac{\sum p_3 q_3}{\sum p_2 q_3} \frac{\sum p_4 q_4}{\sum p_3 q_4} \neq \frac{\sum p_4 q_4}{\sum p_1 q_4}.$$

Поэтому переход от цепных индексов к базисным правомочен только для индексов с постоянными весами. Если же это применяется к индексам с переменными весами, то оговаривается условность такого перехода и предполагается, что структура (состав) «агрегата», для которого вычисляется индекс, мало подвержена изменениям.

9.6. Взаимосвязанные индексы и определение роли отдельных факторов в динамике сложных (результативных) показателей

Как уже отмечалось, общие индексы позволяют, во-первых, характеризовать динамику индексируемого показателя в сложной совокупности и, во-вторых, измерять влияние отдельных факто-

ров на динамику сложных показателей. По существу, возможность решения второй задачи заложена в самóм построении общих индексов в агрегатной форме.

Рассматривая ряд статистических показателей, можно заметить, что многие из них взаимосвязаны, и эта взаимосвязь, в частности, носит мультипликативный характер, т.е. проявляется в том, что один показатель представляет собой произведение ряда других. Например, товарооборот можно представить как произведение количества реализованной продукции на цену (qp), валовой сбор той или иной культуры – как произведение урожайности на площадь ($y\Pi$), объем выпуска продукции – как произведение численности работников на их производительность труда ($q = Tw$) и т.д.

Все показатели – сомножители в указанных произведениях – могут рассматриваться как факторы, которые определяют значение сложного (результативного) показателя. Изменение результативного показателя может происходить за счет изменения всех факторов, его определяющих. Например, товарооборот может измениться как за счет изменения количества (объема) реализованных товаров, так и за счет изменения цен. На изменение валового сбора может оказывать влияние изменение и посевных площадей, и урожайности. Объем выпущенной продукции на любом предприятии может меняться как за счет изменения численности работников, так и за счет изменения их производительности труда и т.д.

Поэтому при анализе изменения сложных показателей важно определить, какова роль отдельных факторов в этом изменении. Решая данную задачу, применяют метод абстракции, неизбежный при изучении социально-экономических явлений.

Чтобы выявить относительное влияние отдельного фактора на динамику сложного показателя, необходимо в результативном показателе, представленном в виде произведения нескольких факторов, исследуемый фактор рассматривать как п е р е м е н н ы й, а остальные считать п о с т о я н н ы м и. Так, если определенный показатель K можно представить как произведение двух факторов a и b, то отношение $\frac{a_1 b_0}{a_0 b_0}$ должно показывать изменение показателя K за счет фактора a, а отношение $\frac{a_0 b_1}{a_0 b_0}$ – изменение K за счет фактора b. В свою очередь, отношение $\frac{a_1 b_1}{a_0 b_0}$ покажет изменение результативного показателя за счет обоих факторов.

Однако при таком обособлении каждого фактора и абстрагировании от влияния прочих факторов важно решить вопрос: на уровне какого периода (базисного или отчетного) следует рассматривать факторы, принимаемые за постоянные. Теоретически здесь возможно несколько решений:

1) независимо от того, в какой последовательности изучается влияние индексируемых факторов, *постоянные факторы рассматриваются на уровне базисного периода*, т.е. как показано выше;

2) *постоянные факторы рассматриваются на уровне отчетного периода*, т.е. $\frac{a_1 b_1}{a_0 b_1}$ (влияние фактора a) и $\frac{a_1 b_1}{a_1 b_0}$ (влияние фактора b);

3) *каждый из уже исследованных факторов при определении влияния других (последующих) факторов рассматривается на уровне отчетного периода*, т.е. если влияние фактора a определять отношением $\frac{a_1 b_0}{a_0 b_0}$, то влияние фактора b в этом случае должно определяться отношением $\frac{a_1 b_1}{a_1 b_0}$.

Очевидно, что чем больше факторов – сомножителей в результативном показателе, тем больше вариантов подобного рода отношений. Следовательно, прежде чем рассматривать обособленное влияние каждого фактора, необходимо обосновать, почему прочие факторы приняты на том или ином уровне.

В статистической практике **схема построения факторных индексов** такова: если результативный показатель можно представить как произведение объемного (количественного) и качественного факторов, то в случае, когда определяется влияние изменения объемного показателя на результативный, качественный показатель фиксируется на уровне базисного периода, а в случае, когда определяется влияние изменения качественного показателя, объемный (рассматриваемый как постоянный) фиксируется на уровне отчетного периода.

Возвращаясь к **агрегатному способу построения общих индексов**, еще раз подчеркнем, что любой агрегатный индекс построен по принципу обособленного рассмотрения влияния изменения отдельных факторов на изменение сложного показателя.

Так, индекс физического объема в агрегатном виде $I_q = \frac{\sum p_0 q_1}{\sum p_0 q_0}$ показывает, как изменяется стоимость определенного круга продукции (сложный показатель) за счет изменения количества про-

дукции (при фиксировании цен на уровне базисного периода), т.е. индекс физического объема рассматривается как факторный по отношению к индексу стоимости.

Таким образом, если сложные показатели представляют собой произведение двух (или более) факторов, то индексы, рассчитанные для таких взаимосвязанных показателей, должны находиться в той же зависимости, что и сами показатели.

Например, если товарооборот можно представить как произведение количества проданных товаров на их цену, то и индекс товарооборота должен равняться произведению индекса количества товара (физического объема товарооборота) на индекс цен. Аналогично индекс объема продукции будет равен произведению индекса числа работников на индекс производительности труда, а индекс валового сбора отдельных культур – произведению индекса посевной площади на индекс урожайности и т.д.

Эта взаимосвязь наглядно проявляется между индивидуальными индексами. Так, для товарооборота pq, цены p и количества определенного продукта q можно записать следующее соотношение их индексов:

$$\frac{p_1q_1}{p_0q_0} = \frac{p_1}{p_0}\,\frac{q_1}{q_0},$$

для объема продукции q, числа работников T и производительности труда $w = q/T$ – соотношение

$$\frac{q_1}{q_0} = \frac{w_1T_1}{w_0T_0} = \frac{w_1}{w_0}\,\frac{T_1}{T_0},$$

для валового сбора V определенной культуры, ее урожайности y и посевной площади $П$ – соотношение

$$\frac{V_1}{V_0} = \frac{y_1П_1}{y_0П_0} = \frac{y_1}{y_0}\,\frac{П_1}{П_0}.$$

Если речь идет не об индивидуальных индексах, а об общих, то факторные индексы должны строиться с таким расчетом, чтобы сохранялась необходимая взаимосвязь между факторными и результативными индексами. Однако при этом возможно несколько решений.

Так, для тех же индексов товарооборота, цен и физического объема эта взаимосвязь может быть обеспечена двояко:

$$1)\ \frac{\sum p_1q_1}{\sum p_0q_0} = \frac{\sum p_1q_1}{\sum p_0q_1}\,\frac{\sum p_0q_1}{\sum p_0q_0},\ \text{т.е. } I_{pq} = I_p^{\text{П}}I_q^{\text{Л}};$$

$$2)\ \frac{\sum p_1q_1}{\sum p_0q_0} = \frac{\sum p_1q_0}{\sum p_0q_0}\,\frac{\sum p_1q_1}{\sum p_1q_0},\ \text{т.е. } I_{pq} = I_p^{\text{Л}}I_q^{\text{П}}.$$

Хотя в обоих случаях обеспечена взаимосвязь, соответствующие индексы цен и объема первого и второго вариантов не равнозначны и, рассматриваемые как факторные индексы, не одинаково отражают влияние указанных факторов на изменение товарооборота. Следовательно, требуется обосновать, почему в каждом из факторных индексов p и q фиксируются на том или ином уровне.

Как уже указывалось, в статистической практике при индексировании качественных показателей отдается предпочтение фиксированию объемных показателей на уровне отчетного периода и, наоборот, при индексировании объемных показателей качественные, выступающие в роли соизмерителей, принимаются на уровне базисного периода. Исходя из этого первый вариант записи следует признать более предпочтительным.

Аналогично взаимосвязь между индексами валового сбора, урожайности и посевной площади можно записать в виде следующего равенства:

$$\frac{\sum y_1 П_1}{\sum y_0 П_0} = \frac{\sum y_1 П_1}{\sum y_0 П_1} \frac{\sum y_0 П_1}{\sum y_0 П_0}. \qquad (9.28)$$

В данном случае индекс урожайности $\frac{\sum y_1 П_1}{\sum y_0 П_1}$, отражающий изменения валового сбора за счет изменения урожайности, построен при фиксировании посевных площадей на уровне отчетного периода, а индекс посевных площадей $\frac{\sum y_0 П_1}{\sum y_0 П_0}$, показывающий изменение валового сбора за счет изменения посевных площадей, построен при фиксировании урожайности на уровне базисного периода. Таким образом, в одном из двух факторных индексов веса фиксируются на уровне отчетного периода, а в другом – на уровне базисного периода. Только при этом они в произведении дадут индекс результативного показателя.

Следует иметь в виду, что в отдельных случаях возможны различные формы записи связи между отдельными показателями или между их индексами. Форма записи существенно влияет на интерпретацию выводов о роли отдельных факторов в относительном и абсолютном изменении результативного показателя.

Например, валовой сбор $\sum yП$ можно выразить как $\bar{y}\sum П$ (из $\bar{y} = \frac{\sum yП}{\sum П}$). Естественно, $\sum yП = \bar{y}\sum П$.

При последней записи взаимосвязанные индексы валового сбора, средней урожайности и посевных площадей будут выглядеть следующим образом:

$$I_V = \frac{\bar{y}_1 \sum П_1}{\bar{y}_0 \sum П_0} = \frac{\bar{y}_1 \sum П_1}{\bar{y}_0 \sum П_1} \frac{\bar{y}_0 \sum П_1}{\bar{y}_0 \sum П_0} = \frac{\bar{y}_1}{\bar{y}_0} \frac{\sum П_1}{\sum П_0}, \qquad (9.29)$$

т.е. индекс валового сбора равен произведению индекса средней урожайности на индекс общей посевной площади.

При равенстве индексов валового сбора в формулах (9.28) и (9.29) интерпретация факторных индексов в них неодинакова. Так, влияние структурного фактора (изменения структуры посевных площадей) в формуле (9.28) отражено в агрегатном индексе посевных площадей $\frac{\sum y_0 П_1}{\sum y_0 П_0}$ совместно с изменением площади, а в формуле (9.29) – в индексе средней урожайности $\frac{\bar{y}_1 \sum П_1}{\bar{y}_0 \sum П_1} = \frac{\bar{y}_1}{\bar{y}_0}$ (индексе переменного состава). В свою очередь, индекс $\frac{\bar{y}_0 \sum П_1}{\bar{y}_0 \sum П_0} = \frac{\sum П_1}{\sum П_0}$ в формуле (9.29) характеризует относительное изменение валового сбора только за счет изменения посевных площадей (без учета структурных изменений).

Поэтому во всех случаях, когда рассчитывается среднее значение индексируемого качественного показателя, важно при построении факторных индексов учитывать, какова цель исследования: определить влияние количественного фактора в чистом виде или в сочетании с изменением его структуры.

Таким образом, для того чтобы раскрыть различные стороны динамики сложных показателей, используется не один, а целый ряд индексов, различных по построению и содержанию, но вместе с тем взаимосвязанных и взаимодополняющих. Поэтому можно говорить о системе индексов, используемых при анализе динамики различных показателей.

Особо рассмотрим **построение индексов в случае трех факторов** (множителей). В области экономических явлений можно встретить результативные показатели, зависящие от трех факторов и более. Так, стоимость материальных затрат, связанных с производством определенной продукции, зависит от количества выпущенной продукции, удельного расхода того или иного материала и цен на него. Изменение любого из названных факторов повле-

чет за собой изменение результативного показателя. Следовательно, изучая динамику такого сложного результативного показателя, важно правильно построить систему взаимосвязанных индексов.

Эта задача сложнее, чем в случае взаимодействия двух факторов, так как, определяя влияние изменения одного из трех факторов на изменение результативного показателя, приходится фиксировать в качестве неизменных два фактора (при этом каждый из них можно фиксировать на разных уровнях).

Построение системы индексов для различных показателей будет различно. Все зависит от объекта исследования. Если обозначить количество выработанной продукции через q, удельные расходы материалов на единицу продукции через m, а цены на материалы через p, то результативный показатель z – общая стоимость материальных затрат – выразится как $\sum qmp$, а его изменение будет

$$I_z = \frac{\sum q_1 m_1 p_1}{\sum q_0 m_0 p_0}.$$

Факторные индексы при этом должны быть построены так, чтобы они были взаимосвязанными, т.е. составляли систему. Эта взаимосвязь обеспечивается, в частности, при таком построении факторных индексов:

$$\frac{\sum q_1 m_1 p_1}{\sum q_0 m_0 p_0} = \frac{\sum q_1 m_0 p_0}{\sum q_0 m_0 p_0} \frac{\sum q_1 m_1 p_0}{\sum q_1 m_0 p_0} \frac{\sum q_1 m_1 p_1}{\sum q_1 m_1 p_0}.$$

Три множителя в правой части равенства – это соответственно:

1) индекс объема q, построенный при фиксировании удельного расхода материалов и цен на уровне базисного периода;

2) индекс удельных расходов m, построенный при фиксировании объема продукции на уровне отчетного периода и фиксировании цен на уровне базисного периода;

3) индекс цен p, построенный при фиксировании объема продукции и удельного расхода материалов на уровне отчетного периода.

(Если в указанных формулах вычесть из числителя каждой дроби ее знаменатель, можно получить сумму абсолютного прироста результативного показателя за счет соответствующего фактора.)

При такой системе взаимосвязанных индексов в первом факторном индексе (объема) фиксированные показатели приняты на уровне базисного периода, поскольку оба они (и удельный расход, и цены) качественные показатели. Во втором факторном

индексе (удельных расходов) влияние изменения норм на общие затраты указано уже с учетом изменившегося объема, т.е. на отчетный выпуск продукции (цены по-прежнему неизменны). И наконец, третий факторный индекс (цен) построен с фиксированием на уровне отчетного периода как объема продукции, так и удельных расходов.

Очевидно, что чем больше взаимосвязанных факторов определяют результативный показатель, тем сложнее процесс разложения по факторам и тем более обоснованным должно быть построение каждого факторного индекса.

9.7. Разложение абсолютных приростов по факторам

В конкретных исследованиях значение индексов как относительных величин намного возрастает, если они дополнены абсолютными величинами.

Построение взаимосвязанных индексов позволяет одновременно решать обе задачи: определять изменение сложного (результативного) показателя как в относительном, так и в абсолютном выражении за счет влияния отдельных факторов.

Расчет изменения сложного показателя за счет изменения отдельных факторов в абсолютном выражении называют *разложением абсолютного прироста (убыли) по факторам.*

Рассмотрим простейшую двухфакторную модель результативного показателя и покажем разные подходы (методы) к разложению абсолютных изменений сложных (результативных) показателей по факторам на конкретных примерах.

Пример. Пусть имеются данные по фермерскому хозяйству о продаже картофеля на трех рынках области (табл. 9.9, графы А, 1–4).

Таблица 9.9

Данные о продаже картофеля на трех рынках

Рынок	Цена 1 кг, руб.		Продано кг		Выручка от продажи, руб.		p_0q_1
	март p_0	апрель p_1	март q_0	апрель q_1	март p_0q_0	апрель p_1q_1	
А	1	2	3	4	5	6	7
1	8	10	800	1000	6400	10000	8000
2	9	12	600	800	5400	9600	7200
3	10	14	600	700	6000	9800	7000
Σ	—	—	2000	2500	17800	29400	22200

Требуется определить относительное и абсолютное изменение выручки от продажи картофеля в апреле по сравнению с выручкой в марте и разложить абсолютное изменение по факторам.

Общая выручка в каждом месяце может быть представлена как сумма произведений цены на количество проданного по данной цене картофеля, т.е. как $\sum pq$. Эти данные за соответствующие месяцы приведены в графах 5, 6 табл. 9.9: $\sum p_0 q_0 = 17800$ руб. и $\sum p_1 q_1 = 29400$ руб.

Отсюда относительное изменение выручки от продажи

$$I_{pq} = \frac{\sum p_1 q_1}{\sum p_0 q_0} = \frac{29400}{17800} = 1{,}651 \text{ (или } 165{,}1\%).$$

В абсолютном выражении изменение выручки составило

$$\Delta pq = \sum p_1 q_1 - \sum p_0 q_0 = 29400 - 17800 = 11600 \text{ руб.}$$

Поскольку общая выручка pq зависит от физического объема реализации и от цен, то индексы двух последних показателей, являясь факторными по отношению к стоимости, покажут, в какой степени каждый фактор повлиял на относительное изменение результативного показателя pq.

Рассчитав в графе 7 необходимые условные данные о стоимости продукции, реализованной в апреле, в ценах марта, т.е. $\sum p_0 q_1$, определим индекс физического объема I_q и индекс цен I_p:

$$I_q = \frac{\sum p_0 q_1}{\sum p_0 q_0} = \frac{22200}{17800} = 1{,}247 \text{ (или } 124{,}7\%);$$

$$I_p = \frac{\sum p_1 q_1}{\sum p_0 q_1} = \frac{29400}{22200} = 1{,}324 \text{ (или } 132{,}4\%).$$

Это означает, что общая выручка от продажи картофеля в апреле по сравнению с мартовским показателем возросла на 24,7% за счет изменения количества реализованной продукции и на 32,4% за счет изменения цен. Причем рост выручки на 24,7% вызван изменением не только общего количества проданного картофеля, но и доли реализации на отдельных рынках, т.е. изменением структуры продаж на трех рынках.

В целом же относительное изменение стоимости равно произведению факторных индексов, т.е. агрегатные индексы результативного и факторных показателей связаны мультипликативно:

$$I_{pq} = I_p I_q.$$

В нашем примере

$$\frac{\sum p_1q_1}{\sum p_0q_0} = \frac{\sum p_1q_1}{\sum p_0q_1} \frac{\sum p_0q_1}{\sum p_0q_0} = 1{,}324 \cdot 1{,}247 = 1{,}651.$$

Так как индексы взаимосвязаны, абсолютный прирост выручки можно разложить по факторам согласно следующей схеме (**первый метод**):

$$\sum p_1q_1 - \sum p_0q_0 = (\sum p_1q_1 - \sum p_0q_1) + (\sum p_0q_1 - \sum p_0q_0)$$

или

$$\Delta pq = {}_p\Delta pq + {}_q\Delta pq,$$

где $\Delta pq = \sum p_1q_1 - \sum p_0q_0$ – общее изменение (прирост) выручки (стоимости) от продажи картофеля;

${}_p\Delta pq = \sum p_1q_1 - \sum p_0q_1$ – увеличение выручки за счет изменения цен на отдельных рынках;

${}_q\Delta pq = \sum p_0q_1 - \sum p_0q_0$ – увеличение выручки за счет изменения объема (количества) и доли (структуры) продаж на трех рынках.

Так, в нашем примере $\Delta pq = 29400 - 17800 = 11600$ руб., ${}_p\Delta pq = 29400 - 22200 = 7200$ руб., ${}_q\Delta pq = 22200 - 17800 = 4400$ руб. В целом $11600 = 7200 + 4400$.

Это наиболее распространенная схема разложения абсолютного прироста результативного показателя (при двухфакторной мультипликативной модели последнего). Однако при таком разложении структурный фактор, т.е. изменение доли продаж на разных рынках, остается невыделенным, его влияние отражено совместно с изменением объема количественного показателя.

Если ставится задача вычленить в абсолютном приросте результативного показателя слагаемое, обусловленное изменением структуры совокупности (структурными сдвигами), используют другой метод разложения. Он применим к однородным совокупностям, для которых можно рассчитать средние значения индексируемого качественного показателя, их индексы переменного и фиксированного составов, а также индекс структуры (структурных сдвигов) [по аналогии с формулой (9.29)].

Рассмотрим **второй метод** разложения абсолютного прироста по факторам на том же примере (см. табл. 9.9).

Так как данные табл. 9.9 относятся к однородной продукции (картофелю), то для каждого месяца можно рассчитать среднюю цену на картофель $\bar{p}$, а именно:

в марте $$\bar{p}_0 = \frac{\sum p_0 q_0}{\sum q_0} = \frac{17800}{2000} = 8{,}9 \text{ руб.};$$

в апреле $$\bar{p}_1 = \frac{\sum p_1 q_1}{\sum q_1} = \frac{29400}{2500} = 11{,}76 \text{ руб.}$$

Сопоставляя средние цены на картофель за два месяца, получаем *индекс цен переменного состава*:

$$I^p_{\text{п.с}} = \frac{\sum p_1 q_1}{\sum q_1} : \frac{\sum p_0 q_0}{\sum q_0} = \bar{p}_1 : \bar{p}_0 = 11{,}76 : 8{,}9 = 1{,}321 \text{ (или } 132{,}1\%),$$

который показывает, что средняя цена на картофель в апреле по сравнению со средней ценой в марте возросла на 32,1%.

Учитывая природу этого индекса, мы понимаем, что средняя цена изменилась как за счет изменения цен на отдельных рынках, так и за счет изменения доли продаж на них, т.е. за счет изменения структуры продаж на рынках.

Чтобы устранить влияние структурного фактора на динамику средних цен, рассчитывают *индекс цен фиксированного состава*:

$$I^p_{\text{ф.с}} = \frac{\sum p_1 q_1}{\sum q_1} : \frac{\sum p_0 q_1}{\sum q_1} = \frac{29400}{2500} : \frac{22200}{2500} = 11{,}76 : 8{,}88 = 1{,}324$$

или, если сократить на $\sum q_1$,

$$I^p_{\text{ф.с}} = \frac{\sum p_1 q_1}{\sum p_0 q_1} = \frac{29400}{22200} = 1{,}324 \text{ (или } 132{,}4\%).$$

Нетрудно заметить, что по значению этот индекс фиксированного состава совпадает с индексом цен, рассчитанным по агрегатной форме при рассмотрении первого метода.

Разделив индекс цен переменного состава на индекс фиксированного состава, получим так называемый *индекс структуры* (структурных сдвигов), характеризующий изменение средней цены на картофель за счет изменения структуры продаж, т.е. доли продаж на отдельных рынках:

$$I^p_{\text{стр}} = I^p_{\text{п.с}} : I^p_{\text{ф.с}} = 1{,}321 : 1{,}324 = 0{,}998 \text{ (или } 99{,}8\%).$$

Этот же результат можно получить непосредственно по формуле (9.24):

$$I^{p}_{\text{стр}} = \frac{\sum p_0 q_1}{\sum q_1} : \frac{\sum p_0 q_0}{\sum q_0} = 8{,}88 : 8{,}9 = 0{,}998,$$

т.е. на 0,2% средняя цена на картофель возросла за счет изменения структуры продаж.

Средние величины, рассчитанные именно в этих трех индексах (цен), используются при втором методе разложения абсолютных приростов по факторам. Суть этого метода такова.

Поскольку средняя цена картофеля в базисном и отчетном периодах рассчитывалась соответственно по формулам

$$\bar{p}_0 = \frac{\sum p_0 q_0}{\sum q_0} \quad \text{и} \quad \bar{p}_1 = \frac{\sum p_1 q_1}{\sum q_1},$$

общая стоимость картофеля на трех рынках в базисном и отчетном периодах будет соответственно

$$\sum p_0 q_0 = \bar{p}_0 \sum q_0 \quad \text{и} \quad \sum p_1 q_1 = \bar{p}_1 \sum q_1.$$

Следовательно, изменение выручки можно представить как результат изменения средней цены $\bar{p}$ и изменения общего объема проданного картофеля $\sum q$, т.е.

$$\frac{\sum p_1 q_1}{\sum p_0 q_0} = \frac{\bar{p}_1 \sum q_1}{\bar{p}_0 \sum q_1} \frac{\bar{p}_0 \sum q_1}{\bar{p}_0 \sum q_0}.$$

Рассматривая изменение каждого фактора при втором постоянном и вычитая из числителя знаменатель, определяем абсолютный прирост выручки за счет соответствующего фактора.

Итак, абсолютный прирост выручки как результативного показателя

$$\Delta pq = \sum p_1 q_1 - \sum p_0 q_0 = 29400 - 17800 = 11600 \text{ руб.},$$

в том числе по факторам:

1) за счет изменения общего объема проданного картофеля $\sum q$:

$$_{\sum q}\Delta pq = (\sum q_1 - \sum q_0)\bar{p}_0 = (2500 - 2000) \cdot 8{,}9 = 4450 \text{ руб.}$$

Этот же результат можно получить и по-иному, умножая выручку в базисном периоде $\sum p_0 q_0$ на коэффициент прироста объема продаж картофеля в отчетном периоде. Так, если $\frac{\sum q_1}{\sum q_0} = \frac{2500}{2000} = 1{,}25$, то

$$_{\sum q}\Delta pq = \sum p_0 q_0 \left(\frac{\sum q_1}{\sum q_0} - 1 \right) = 17800 (1{,}25 - 1) = 4450 \text{ руб.};$$

2) за счет изменения средней цены на картофель $\bar{p}$:

$$_{\bar{p}}\Delta pq = (\bar{p}_1 - \bar{p}_0)\sum q_1 = (11{,}76 - 8{,}9)\cdot 2500 = 7150 \text{ руб.}$$

Однако, как уже отмечалось, изменение средней цены зависит от изменения цен и доли реализации картофеля на отдельных рынках. Поэтому абсолютный прирост выручки в размере 7150 руб. можно, в свою очередь, подразделить на две части, для чего используются значения средних цен (они рассчитаны ранее в индексах фиксированного состава и структуры).

В нашем примере результат, полученный в п. 2 (7150 руб.), можно разложить на два слагаемых:

1) прирост выручки за счет изменения цен на отдельных рынках

$$_{p}\Delta pq = \left(\frac{\sum p_1 q_1}{\sum q_1} - \frac{\sum p_0 q_1}{\sum q_1}\right)\sum q_1 = (11{,}76 - 8{,}88)\cdot 2500 = 7200 \text{ руб.}$$

Этот же результат можно получить по-иному (после сокращения на $\sum q_1$):

$$_{p}\Delta pq = \sum p_1 q_1 - \sum p_0 q_1 = 29400 - 22200 = 7200 \text{ руб.};$$

2) прирост выручки за счет изменения структуры (доли) продаж на отдельных рынках

$$_{\text{стр.}q}\Delta pq = \left(\frac{\sum p_0 q_1}{\sum q_1} - \frac{\sum p_0 q_0}{\sum q_0}\right)\sum q_1 = (8{,}88 - 8{,}9)\cdot 2500 = -50 \text{ руб.}$$

В сумме 7200 и (–50) дают 7150.

Таким образом, в итоге общий абсолютный прирост выручки Δpq, равный 11600 руб., мы разложили по факторам на три слагаемых:

1) 4450 руб. – за счет изменения объема реализации картофеля;
2) 7200 руб. – за счет изменения цен на картофель (на всех рынках);
3) –50 руб. – за счет изменения доли реализованного картофеля на отдельных рынках, т.е. за счет структурного фактора.

Нетрудно заметить, что сумма первого и третьего слагаемых, т.е. 4450 + (–50), равна результату, полученному при разложении с помощью первого метода как $\sum p_0 q_1 - \sum p_0 q_0 = 4400$ руб., т.е. как прирост выручки за счет изменения и объема и структуры проданного на трех рынках картофеля.

Рассмотрим еще один пример и одновременно ознакомимся с расчетом индексов производительности труда.

Пример. Предположим, имеются данные по трем предприятиям о выпуске продукции и численности работников за два периода (табл. 9.10, графы А, 1–4).

Определить:

1) изменение производительности труда по каждому предприятию в отдельности и в целом по трем предприятиям в виде индексов переменного и фиксированного составов, а также индекса структурных сдвигов;
2) изменение общего выпуска продукции по трем предприятиям в абсолютном выражении, разложив последний по факторам: а) за счет изменения общей численности работников; б) за счет изменения производительности труда на отдельных предприятиях; в) за счет изменения структурного фактора – доли работников на отдельных предприятиях.

Таблица 9.10

Предприятие	Выпуск продукции в сопоставимых ценах*, тыс. руб.		Средняя численность работников, чел.		Средняя выработка на 1 работника в сопоставимых ценах, тыс. руб.		Индекс производительности труда $i = \frac{w_1}{w_0}$	$w_0 T_1$
	Базисный период $Q_0 = w_0 T_0$	Отчетный период $Q_1 = w_1 T_1$	Базисный период T_0	Отчетный период T_1	Базисный период w_0	Отчетный период w_1		
А	1	2	3	4	5	6	7	8
1	1500	912	500	320	3,0	2,85	0,95	960
2	3100	2256	620	480	5,0	4,70	0,94	2400
3	7040	6080	880	800	8,0	7,60	0,95	6400
$\sum$	11640	9248	2000	1600	($\bar{w}_0$ = = 5,82)	($\bar{w}_1$ = = 5,78)	($I^w_{\text{п.с}}$ = = 0,993)	9760

* Выпуск продукции в сопоставимых ценах, обозначенный нами через Q, практически представляет $Q_0 = \sum q_0 p_0$ и $Q_1 = \sum q_1 p_0$.

1. В графах 5–7 табл. 9.10 рассчитаны производительность труда w и ее индексы по отдельным предприятиям. В итоговой строке графы 7 приведен индекс, характеризующий динамику

среднеи производительности труда по трем предприятиям, т.е. *индекс переменного состава*, рассчитанный по формуле

$$I^{w}_{\text{п.с}} = \overline{w}_1 : \overline{w}_0 = \frac{\sum Q_1}{\sum T_1} : \frac{\sum Q_0}{\sum T_0} = \frac{\sum w_1 T_1}{\sum T_1} : \frac{\sum w_0 T_0}{\sum T_0} =$$

$$= \frac{9248}{1600} : \frac{11640}{2000} = 5{,}78 : 5{,}82 = 0{,}993 \text{ (или } 99{,}3\%).$$

В абсолютном выражении снижение средней производительности труда составило $\overline{w}_1 - \overline{w}_0 = 5{,}78 - 5{,}82 = -0{,}04$ тыс. руб.

Значение индекса переменного состава в данном примере выходит за пределы индивидуальных значений по отдельным предприятиям, т.е. на отдельных предприятиях производительность труда снизилась на 5 и 6%, а средняя производительность (согласно индексу переменного состава) – всего на 0,7%. Очевидно, что это результат влияния структурного фактора, в частности увеличения доли работников на третьем предприятии, где более высокая производительность труда.

Чтобы устранить влияние структурного фактора на динамику средней производительности труда, рассчитаем *индекс фиксированного состава* (для чего предварительно определим в графе 8 необходимый для расчета индекса условный показатель $\sum w_0 T_1$ – выпуск продукции при численности работников, зафиксированной в отчетном периоде, и базисном уровне производительности труда):

$$I^{w}_{\text{ф.с}} = \frac{\sum w_1 T_1}{\sum T_1} : \frac{\sum w_0 T_1}{\sum T_1} = \frac{9248}{1600} : \frac{9760}{1600} =$$

$$= 5{,}78 : 6{,}1 = 0{,}9475 \text{ (или } 94{,}75\%).$$

Применяя формулу агрегатного индекса производительности труда, т.е. после сокращения на $\sum T_1$, получаем тот же результат:

$$I^{w}_{\text{ф.с}} = \frac{\sum w_1 T_1}{\sum w_0 T_1} = \frac{9248}{9760} = 0{,}9475 \text{ (или } 94{,}75\%).$$

Как и следовало ожидать, индекс фиксированного состава не вышел за пределы индивидуальных индексов, т.е. он показывает, что в среднем по всем предприятиям производительность труда снизилась на 5,25%.

В абсолютном выражении за счет изменения w на отдельных предприятиях средняя производительность труда снизилась на $5{,}78 - 6{,}1 = -0{,}32$ тыс. руб.

Чтобы отразить влияние структурного фактора на динамику средней производительности труда, рассчитаем *индекс структурных сдвигов*:

$$I^{w}_{\text{стр}} = \frac{\sum w_0 T_1}{\sum T_1} : \frac{\sum w_0 T_0}{\sum T_0} = 6{,}1 : 5{,}82 = 1{,}048 \text{ (или } 104{,}8\%)$$

или

$$I^{w}_{\text{стр}} = I^{w}_{\text{п.с}} : I^{w}_{\text{ф.с}} = 99{,}3 : 94{,}75 = 1{,}048.$$

Данный индекс означает, что за счет изменения доли работников на отдельных предприятиях средняя производительность труда выросла на 4,8% (или в абсолютном выражении на 6,1 – 5,82 = = 0,28 тыс. руб.).

Далее на основе полученных данных *разложим абсолютное изменение выпуска продукции ΔQ по факторам.*

2. Абсолютное изменение выпуска продукции

$$\Delta Q = \sum Q_1 - \sum Q_0 = 9248 - 11640 = -2392 \text{ тыс. руб.},$$

в том числе:

а) за счет изменения общей численности работников

$${}_{\sum T}\Delta Q = (\sum T_1 - \sum T_0)\,\overline{w}_0 = (1600 - 2000) \cdot 5{,}82 = -2328 \text{ тыс. руб.};$$

б) за счет изменения производительности труда на отдельных предприятиях

$${}_{w}\Delta Q = \left(\frac{\sum w_1 T_1}{\sum T_1} - \frac{\sum w_0 T_1}{\sum T_1}\right)\sum T_1 = (5{,}78 - 6{,}1)\cdot 1600 = -512 \text{ тыс. руб.}$$

или сразу

$${}_{w}\Delta Q = \sum w_1 T_1 - \sum w_0 T_1 = 9248 - 9760 = -512 \text{ тыс. руб.};$$

в) за счет изменения доли работников на предприятиях с разной производительностью труда, т.е. за счет структурного фактора,

$${}_{\text{стр.}T}\Delta Q = \left(\frac{\sum w_0 T_1}{\sum T_1} - \frac{\sum w_0 T_0}{\sum T_0}\right)\sum T_1 = (6{,}1 - 5{,}82)\cdot 1600 = 448 \text{ тыс. руб.}$$

В целом

ΔQ	=	${}_{\sum T}\Delta Q$	+	${}_{w}\Delta Q$	+	${}_{\text{стр.}T}\Delta Q$
–2392	=	–2328	–	512	+	448
(общее изменение выпуска)		(за счет изменения численности работников)		(за счет изменения производительности труда на отдельных предприятиях)		(за счет изменения доли работников на отдельных предприятиях)

Таким образом, мы разложили абсолютное изменение объема выпуска продукции по трем факторам с выделением влияния изменения структуры.

Однако можно было разложить абсолютное изменение выпуска на основе взаимосвязи следующих трех агрегатных индексов (т.е. по первому методу):

$$\frac{\sum w_1 T_1}{\sum w_0 T_0} = \frac{\sum w_1 T_1}{\sum w_0 T_1} \frac{\sum w_0 T_1}{\sum w_0 T_0},$$

где $\frac{\sum w_1 T_1}{\sum w_0 T_0}$ – индекс общего объема выпуска как результативного показателя;

$\frac{\sum w_1 T_1}{\sum w_0 T_1}$ – индекс производительности труда как факторный индекс по отношению к индексу общего объема, отражающий изменение объема выпуска за счет изменения производительности труда на отдельных предприятиях;

$\frac{\sum w_0 T_1}{\sum w_0 T_0}$ – индекс численности работников как факторный по отношению к индексу общего объема, отражающий изменение объема выпуска за счет изменения численности работников с учетом их распределения по предприятиям.

Тогда

$$\sum w_1 T_1 - \sum w_0 T_0 = (\sum w_1 T_1 - \sum w_0 T_1) + (\sum w_0 T_1 - \sum w_0 T_0).$$

В нашем примере получаем

$$9248 - 11640 = (9248 - 9760) + (9760 - 11640)$$

или

$$-2392 = -512 - 1880.$$

В общем виде разложение можно записать так:

$$\Delta Q = {}_w\Delta Q + {}_T\Delta Q.$$

При данном разложении уменьшение выпуска за счет изменения (снижения) производительности труда на отдельных предприятиях (${}_w\Delta Q = -512$ тыс. руб.) совпадает с показателем предыдущего разложения (см. п. 2б на с. 415).

Что касается второго слагаемого ${}_T\Delta Q = \sum w_0 T_1 - \sum w_0 T_0 =$ $= -1880$ тыс. руб., то оно учитывает влияние изменения и численности, и доли работников (структурного фактора) на из-

менение выпуска продукции, и поэтому равно по значению сумме первой и третьей составляющих предыдущего разложения (см. пп. 2а и 2в на с. 415):

$$\Sigma_T \Delta Q = -2328 \text{ тыс. руб.} \quad \text{и} \quad {}_{\text{стр.}T}\Delta Q = 448 \text{ тыс. руб.}$$

Таким образом, можно по-разному разложить абсолютные приросты по факторам, поэтому необходимо четко представлять, какие результаты будут получены, если применить тот или иной метод.

Мы рассмотрели простейший случай разложения по факторам абсолютного прироста, когда результативный показатель представляет собой двухфакторную мультипликативную модель.

Чем больше взаимосвязанных факторов определяют результативный показатель, тем сложнее процесс разложения абсолютных приростов по факторам и тем более обоснованным должно быть построение каждого факторного индекса (в частности, фиксирование весов на уровне базисного или отчетного периода).

В статистическом анализе прием разложения абсолютных приростов результативного показателя по факторам играет важную роль и вместе с тем содержит элемент условности.

9.8. Проблемы и методы исчисления территориальных индексов

Проблемы исчисления территориальных индексов во многом схожи с проблемами исчисления традиционных (динамических) индексов цен и физического объема (см. параграфы 9.2–9.7). В обоих случаях измеряются соотношения уровней цен или физического объема товаров и услуг в различных периодах (*динамические индексы*) или в различных странах, регионах (*территориальные индексы*). Однако расчет территориальных индексов нередко более сложен по следующим причинам:

1) различия в структуре цен и количества товаров между странами, как правило, гораздо значительнее, чем между периодами в рамках одной страны (особенно, если периоды не отстоят слишком далеко друг от друга). Это обусловлено особенностями экономики различных стран (например, в разных странах неодинакова степень субсидирования товаров и услуг, степень распространения нерыночных видов деятельности и т.д.; некоторые социальные услуги могут предоставляться в одних странах на платной основе, в других – на бесплатной);

2) территориальные (международные) сопоставления нередко осуществляются одновременно для группы стран (например, для

стран Европейского союза или СНГ), поэтому необходимо согласовывать индексы, исчисленные для всей группы.

Для того чтобы исчислять территориальные индексы и сопоставлять данные, которые относятся к различным странам и регионам, статистики вынуждены конструировать и использовать особые формулы. Разрабатывая эти формулы, ученые руководствуются положениями двух теорий индексов: аксиоматической и экономической.

В *аксиоматической теории индексов* сформулирован ряд требований к индексам с точки зрения формальной логики (например, требования факторной пробы, обратимости во времени, тождественности и др.). Так, требование тождественности означает, что если цены в отчетном периоде не изменились по сравнению с ценами в базисном периоде, то общий индекс цен должен быть равен единице независимо от изменения физического объема. Другое требование этой теории – пропорциональность индексов – означает, что если цены отчетного периода выросли в k раз по сравнению с ценами в базисном периоде, то средний индекс также должен увеличиться в k раз независимо от изменения физического объема.

В *экономической теории индексов* содержится концептуальная основа для поиска «истинного» индекса. Так, истинный индекс цен можно получить, сопоставив расходы потребителей в текущем и базисном периодах при условии, что они обеспечивают «равную пользу» (равную полезность) потребителям при разных ценах, т.е. фактические расходы потребителей сравниваются с условными, гипотетическими, которые при разных ценах в двух периодах обеспечивают «равную пользу». Это сравнение и должно обеспечить отыскание «истинного» индекса цен. Заметим, что экономическая теория индексов достаточно абстрактна, поскольку статистики не оперируют категориями пользы или полезности, а имеют дело с конкретными товарами и услугами. Тем не менее теория выражает некий общий теоретический подход к разработке индексов.

В специальной литературе не прекращается дискуссия о том, насколько обоснованы аксиоматическая и экономическая теории индексов, можно ли применять положения этих теорий в статистической практике, каковы относительные достоинства и недостатки различных индексных формул. Аксиоматическую теорию критикуют за то, что в ней предполагается отсутствие связи между изменением цен и изменением физического объе-

ма. Экономическую теорию критикуют за абстрактный характер, за то, что невозможно использовать ее выводы в практической деятельности.

Основные требования к территориальным индексам таковы:

1) *характерность весов.* Согласно этому требованию для показателей двух стран A и B в качестве весов должны использоваться цены (физический объем товаров) стран A и B (или средние из них), но не цены (физический объем) какой-либо третьей страны C;

2) *независимость от выбора базисной страны* (требование обратимости индексов во времени, адаптированное к территориальным сопоставлениям). Это требование в математической форме можно записать следующим образом:

$$I^{A/B} I^{B/A} = 1,$$

где $I^{A/B}$ – индекс цен (физического объема) страны A по отношению к стране B;

$I^{B/A}$ – индекс цен (физического объема) страны B по отношению к стране A;

3) *транзитивность* (требование циркулярности индексов, адаптированное к территориальным сопоставлениям). В математической форме это требование можно записать следующим образом:

$$I^{A/B} = I^{A/C} : I^{B/C},$$

где $I^{A/C}$ – индекс цен (физического объема) страны A по отношению к стране C;

$I^{B/C}$ – индекс цен (физического объема) страны B по отношению к стране C.

Суть требования транзитивности состоит в том, что индекс, полученный для некоторой пары стран A и B путем прямого сопоставления их цен (физического объема), должен быть равен этому же индексу, полученному косвенным путем, т.е. делением индекса $I^{A/C}$ на индекс $I^{B/C}$;

4) *аддитивность.* Согласно требованию аддитивности индексы цен (физического объема), исчисленные для всей совокупности товаров и услуг (например, для ВВП в целом), должны быть четко согласованы с индексами, исчисленными для всех групп этой совокупности. При этом предполагается, что показатели ВВП, исчисленные в ценах базисного периода, должны быть равны сумме компонентов ВВП, также исчисленных в ценах базисного периода;

5) *требование факторной пробы.* Согласно требованию факторной пробы произведение индекса цен и индекса физического объема должно быть равно индексу стоимости. В математической форме это требование можно записать следующим образом:

$$I_p^{A/B} I_q^{A/B} = I_{pq}^{A/B},$$

где $I_p^{A/B}$ – индекс цен показателя (например, ВВП) страны A по отношению к стране B;

$I_q^{A/B}$ – индекс физического объема показателя (например, ВВП) страны A по отношению к стране B;

$I_{pq}^{A/B}$ – индекс стоимости показателя (например, ВВП) страны A по отношению к стране B.

В теории и практике международных сопоставлений различают прямые парные и многосторонние сопоставления.

Прямые парные сопоставления проводятся для какой-либо изолированной пары стран (например, для России и США, Англии и Франции и т.д.). На них не влияют показатели третьих стран. Таким образом, для прямых парных сопоставлений важным оказывается требование характерности весов, а также требования факторной пробы и независимости от выбора базисной страны.

Многосторонние сопоставления проводятся одновременно для группы стран. Для многосторонних сопоставлений особое значение имеет требование транзитивности индексов, т.е. согласованность результатов расчетов индексов цен и физического объема для всей группы стран.

Как прямые парные, так и многосторонние сопоставления имеют свою специфику, поэтому для их проведения используют различные формулы индексов.

Прямые парные сопоставления

Для проведения прямых парных сопоставлений ВВП и паритетов покупательной способности (ППС) валют используют индексы Ласпейреса, Пааше, средней геометрической невзвешенной, а также формулу Фишера.

Расчет индексов по схеме прямых парных сопоставлений ВВП и ППС проводится в несколько этапов:

1) ВВП сопоставляемых стран A и B подразделяется на однородные товарные группы (как правило, 300–350 групп);

2) для каждой товарной группы подбирается несколько идентичных товаров-представителей с ценами, что дает возможность исчислить индивидуальные индексы цен для всех отобранных товаров-представителей (i_1, i_2, i_3, ..., i_n). При этом следует иметь в виду, что товары-представители должны быть идентичны не только по техническим параметрам (например, содержание металла в руде или мощность мотора автомобиля), но также по ряду факторов, влияющих на цену, таких, как условия реализации и наличие сопутствующих услуг, предоставление гарантии и гарантийное обслуживание, размер упаковки и условия платежа и т.д.;

3) для каждой товарной группы по индивидуальным индексам цен на товары-представители исчисляется средний индекс цен. Для этой цели применяется формула *средней геометрической невзвешенной*

$$\bar{i} = \sqrt[n]{i_1 i_2 i_3 \cdots i_n},$$

что связано с необходимостью обеспечить независимость индексов от выбора базисной страны (формула средней арифметической не обеспечивает этого требования). Кроме того, на практике, как правило, отсутствуют данные о весах товаров-представителей, поэтому используется средняя геометрическая невзвешенная;

4) далее исчисляются средние индексы цен (физического объема) для ВВП в целом. Для расчета средних индексов цен (физического объема) можно использовать различные формулы. Вначале применяют традиционные *формулы Ласпейреса* и *Пааше*:

$$I_{Л.р}^{A/B} = \frac{\sum i_s w_B}{\sum w_B}, \qquad I_{П.р}^{A/B} = \frac{\sum w_A}{\sum \frac{w_A}{i_s}},$$

где $I_{Л.р}^{A/B}$ – индекс цен по формуле Ласпейреса стран A и B;

$I_{П.р}^{A/B}$ – индекс цен по формуле Пааше стран A и B;

i_s – средний индекс цен для товарной группы s.

В качестве весов w для исчисления индекса цен по формуле Ласпейреса используются данные о товарной структуре ВВП (т.е. о доле отдельных товарных групп в ВВП) базисной страны (в нашем случае страны B). В качестве весов w для исчисления индекса цен по формуле Пааше используются данные о товарной структуре ВВП страны A.

Так как нет оснований предпочесть один индекс другому и на практике удобно иметь одно значение, средний индекс цен исчисляют по *формуле Фишера*

$$I_{\text{Ф}.p}^{A/B} = \sqrt{I_p^{\text{Л}} I_p^{\text{П}}}.$$

Искомый индекс физического объема ВВП стран A и B получают делением индекса стоимости ВВП этих стран на индекс цен Фишера:

$$I_q^{A/B} = I_{pq}^{A/B} : I_{\text{Ф}.p}^{A/B}.$$

Тот же самый результат можно получить другим способом: последовательно сопоставить ВВП двух стран A и B соответственно в ценах стран A и B (при этом получим два индекса физического объема по формулам Ласпейреса и Пааше) и исчислить средний индекс физического объема по формуле Фишера, т.е.

$$I_{\text{Л}.q}^{A/B} = \frac{\sum q_A p_B}{\sum q_B p_B}, \qquad I_{\text{П}.q}^{A/B} = \frac{\sum q_A p_A}{\sum q_B p_A}, \qquad I_{\text{Ф}.q}^{A/B} = \sqrt{I_q^{\text{Л}} I_q^{\text{П}}},$$

где $I_{\text{Л}.q}^{A/B}$ – индекс физического объема по формуле Ласпейреса стран A и B;

$I_{\text{П}.q}^{A/B}$ – индекс физического объема по формуле Пааше стран A и B;

$I_{\text{Ф}.q}^{A/B}$ – индекс физического объема по формуле Фишера стран A и B.

В специальной литературе применение формулы Фишера нередко обосновывается необходимостью устранить так называемый *эффект Гершенкрона*, который состоит в том, что индексы Ласпейреса, как правило, больше индексов Пааше. По мнению ряда специалистов, эффект Гершенкрона свидетельствует о том, что оба индекса (и Ласпейреса, и Пааше) искажают «истинное» значение индекса, которое, возможно, находится между ними.

Необходимо отметить, что эффект Гершенкрона наблюдается при о т р и ц а т е л ь н о й зависимости между объемами произведенной (реализованной) продукции и ценами. Другими словами, эффект Гершенкрона проявляется тогда, когда с ростом производства того или иного товара снижаются цены или темп роста цен на этот товар.

Таким образом, применение формулы Фишера при прямых парных сопоставлениях обеспечивает однозначный результат для каждой пары стран и позволяет получить индексы, которые удов-

летворяют требованиям характерности весов, обратимости во времени, а также факторной пробы. Однако индексы, полученные по формуле Фишера, не удовлетворяют требованию аддитивности, а цепные индексы Фишера – требованию транзитивности, чрезвычайно важному для многосторонних территориальных (международных) сопоставлений.

Многосторонние сопоставления

Для проведения многосторонних сопоставлений ВВП и ППС разработаны формулы индексов, которые удовлетворяют требованию транзитивности. Среди них индексы, исчисленные по формулам ЭКШ, Гири – Камиса, Уолша и Геларди.

1. Чаще других используется *формула ЭКШ* (в названии использованы начальные буквы фамилий трех статистиков, предложивших этот индекс: венгров Элтетэ и Кэвеша и поляка Шульца). Разработчики индекса ЭКШ стремились получить такой средний индекс для каждой пары стран, участвующих в многостороннем сопоставлении, который бы удовлетворял требованию транзитивности и минимально отклонялся от первоначального индекса Фишера для данной пары, поскольку последний отвечает требованию характерности весов. Другими словами, идея индекса ЭКШ – достичь известного компромисса между различными требованиями к индексам. Формула индекса ЭКШ такова:

$$I_{\text{ЭКШ}}^{A/B} = \sqrt[n]{\left(F^{A/B}\right)^2 \left(F^{A/j} F^{j/B}\right)},$$

где $I_{\text{ЭКШ}}^{A/B}$ – индекс ЭКШ для стран A и B;
$F^{A/B}$ – индекс Фишера для стран A и B;
$F^{A/j}$ – индекс Фишера для стран A и j;
$F^{j/B}$ – индекс Фишера для стран j и B;
n – число стран, участвующих в сопоставлении.

Таким образом, индекс ЭКШ для стран A и B представляет собой среднюю геометрическую из индексов Фишера (исчисленных прямым и косвенным путем, т.е. через третью страну) для этих стран; при этом прямой индекс Фишера $F^{A/B}$ берется с весом 2.

Поясним сказанное на примере. Предположим, что в сопоставлении принимают участие четыре страны: A, B, C и D. Тогда индекс ЭКШ для стран A и B

$$I_{\text{ЭКШ}}^{A/B} = \sqrt[4]{\left(F^{A/B}\right)^2 \left(F^{A/C} F^{C/B}\right)\left(F^{A/D} F^{D/B}\right)},$$

а индекс ЭКШ для стран A и C

$$I_{\text{ЭКШ}}^{A/C} = \sqrt[4]{\left(F^{A/C}\right)^2\left(F^{A/B}F^{B/C}\right)\left(F^{A/D}F^{D/C}\right)}.$$

Аналогично можно рассчитать индексы ЭКШ для любой пары стран.

Индексы ЭКШ транзитивны, независимы от выбора базисной страны и в наименьшей мере отклоняются от прямого индекса Фишера. Однако индекс ЭКШ не удовлетворяет требованию аддитивности.

2. Другой важный индекс, применяемый при многосторонних международных сопоставлениях, определяется по *формуле Гири – Камиса*. Эта формула позволяет исчислить средние международные цены на различные группы товаров, выраженные в единицах условной международной валюты, а также ППС валют всех стран, участвующих в многосторонних сопоставлениях, по отношению друг к другу и к условной международной валюте. Таким образом, применение метода Гири – Камиса позволяет исчислить ВВП всех стран, участвующих в сопоставлении, в единых средних международных ценах. Это, в свою очередь, обеспечивает основу для исчисления индекса физического объема ВВП различных стран.

Индексы, полученные по формуле Гири – Камиса, удовлетворяют требованиям транзитивности, независимости от выбора базисной страны, факторной пробы и аддитивности, однако не удовлетворяют требованию характерности весов. Кроме того, исчисление индексов Гири – Камиса весьма сложно, так как предполагает сбор большого объема данных, применение компьютеров.

Например, если в сопоставлении ВВП принимает участие 10 стран и их ВВП разбит на 300 товарных групп, то применение формулы Гири – Камиса потребует решения системы 310 уравнений с 310 неизвестными (300 неизвестных – это средние международные цены, а 10 неизвестных – это ППС валют стран по отношению к некоторой условной международной валюте).

3. Еще один метод территориальных (международных) сопоставлений, для которого разработана особая форма индекса, носит название *метода Уолша*. Формула индекса Уолша имеет следующий вид:

$$I_{\text{У}}^{A/B} = \prod (i_p^{A/B})^{w_s},$$

где $i_p^{A/B}$ – средний индекс цен для товарной группы s в стране A по сравнению со страной B;

w_s – средняя доля товарной группы s для всей совокупности стран, принимающих участие в сопоставлении.

Таким образом, по формуле Уолша рассчитывается средний геометрический индекс, взвешенный по средним весам для группы стран, участвующих в сопоставлении; в качестве этих средних весов выступают средние (для всей совокупности стран) доли товарных групп в соответствующих показателях (например, в ВВП).

Расчет индекса Уолша рассмотрим на условном числовом примере.

Пример. В табл. 9.11 представлены данные о структуре ВВП стран A, B и C в разбивке на три товарные группы: 1, 2 и 3. В этой же таблице подсчитаны средние доли товарных групп 1, 2, 3, а также в отдельной колонке указаны средние индексы цен по всем трем товарным группам.

Таблица 9.11

Товарная структура ВВП стран *A*, *B* и *C*

Товарная группа	Доля в ВВП стран			В среднем для трех стран	Средние индексы цен	
	A	B	C		A/B	B/C
1	0,2	0,3	0,4	0,3	1,2	1,4
2	0,4	0,6	0,5	0,5	1,4	1,7
3	0,4	0,1	0,1	0,2	1,8	1,1
Σ	1,0	1,0	1,0	1,0	$I_p^{A/B} = 1,5$	$I_p^{B/C} = 1,6$

Исходя из этих условий средний индекс цен ВВП страны A по отношению к стране B

$$I_p^{A/B} = 1,2^{0,3} \cdot 1,4^{0,5} \cdot 1,8^{0,2} = 1,5,$$

а средний индекс цен страны B по отношению к стране C

$$I_p^{B/C} = 1,4^{0,3} \cdot 1,7^{0,5} \cdot 1,1^{0,2} = 1,6.$$

Индексы Уолша транзитивны и независимы от выбора базисной страны, однако они не удовлетворяют требованию аддитивности, а также в меньшей мере, чем индексы ЭКШ, удовлетворяют требованию характерности весов.

4. В практике международных сопоставлений ВВП, проводимых в рамках Европейского союза, в течение нескольких лет применялся *метод Геларди* (названный так в честь итальянского статистика, предложившего этот метод). В его основе лежит ис-

числение индексов физического объема ВВП различных стран с помощью оценки ВВП в средних международных ценах, получаемых по формуле средней геометрической невзвешенной. Таким образом, метод Геларди схож с методом Гири – Камиса, однако в отличие от него средние международные цены исчисляются здесь по формуле средней геометрической невзвешенной (а не по формуле средней арифметической взвешенной, как в методе Гири – Камиса).

Индексы, исчисленные методом Геларди, отвечают требованиям транзитивности, аддитивности, независимости от выбора базисной страны, однако полученные средние международные цены не имеют ясного экономического содержания.

Главная особенность формулы Геларди – отсутствие взвешивания, т.е. всем странам независимо от их размера или уровня экономического развития придается равный вес. По мысли Геларди, это устраняет искажающее влияние эффекта Гершенкрона на результаты расчета, поскольку применение весов означало бы, что средние международные цены тяготеют к ценам больши́х (или богатых) стран. Если больши́е (или богатые) страны приняты в качестве базисных, то при сравнении с ними других стран будут получены относительно более высокие результаты, так как индексы Ласпейреса дают более высокие значения, чем индексы Пааше. Однако не все специалисты разделяют эту точку зрения.

5. При территориальном сопоставлении макроэкономических показателей широко применяется также *метод цепных индексов.* Например, если в рамках некоторой группы стран ВВП всех стран сопоставляется с ВВП какой-либо одной страны, принятой за базу сравнения, то ВВП всех стран, кроме базисной, сравниваются с помощью цепных индексов.

Так, если сопоставление проводится для стран A, B, C, D и в качестве базисной принята страна B, тогда можно исчислить ряд индексов, характеризующих соотношение ВВП всех стран по сравнению с ВВП страны B:

$$I^{A/B},\ I^{C/B},\ I^{D/B}.$$

Далее для получения индексов ВВП стран (кроме базисной), например, A и D применяется цепной метод:

$$I^{A/D} = I^{A/B} : I^{D/B}.$$

В качестве иллюстрации приводится табл. 9.12 с результатами сопоставления ВВП стран СНГ и Монголии по данным за 2000 г.

Таблица 9.12

Валовой внутренний продукт в 2000 г.
(исчислено по паритету покупательной способности валют)

Страна	ВВП всего, млрд руб.	Удельный вес страны в общем объеме ВВП стран СНГ, % (СНГ11 = 100%)	ВВП на душу населения, руб.	Индекс физического объема на душу населения (в среднем по СНГ11 = 100%)
Азербайджан	139,7	1,6	17624	45,8
Армения	53,0	0,6	13952	36,3
Беларусь	337,0	3,8	33682	87,5
Грузия	75,4	0,9	16311	42,4
Казахстан	479,3	5,4	32235	83,8
Кыргызстан	51,6	0,6	10496	27,3
Молдова	39,3	0,4	10809	28,1
Россия	7275,4	81,8	49984	129,1
Таджикистан	36,4	0,4	5877	15,3
Туркменистан	144,0	1,6	29880	77,7
Всего по странам СНГ (11 стран*)	8891,3	100,0	38476	100,0
Монголия	28,9	0,3	12127	31,5

* С учетом данных по Узбекистану.

Изложенное выше далеко не исчерпывает всех проблем построения территориальных индексов, а представляет систематизацию наиболее важных теоретических положений.

ПРИЛОЖЕНИЯ
(математико-статистические таблицы и основные формулы)

ПРИЛОЖЕНИЕ 1

Значения функции $\varphi(t) = \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}$

t	0	1	2	3	4	5	6	7	8	9
0,0	3989	3989	3989	3988	3986	3984	3982	3980	3977	3973
0,1	3970	3965	3961	3956	3951	3945	3939	3932	3925	3918
0,2	3910	3902	3894	3885	3876	3867	3857	3847	3836	3825
0,3	3814	3802	3790	3778	3765	3752	3739	3725	3712	3697
0,4	3683	3668	3653	3637	3621	3605	3589	3572	3555	3538
0,5	3521	3503	3485	3467	3448	3429	3410	3391	3372	3352
0,6	3332	3312	3292	3271	3251	3230	3209	3187	3166	3144
0,7	3123	3101	3079	3056	3034	3011	2989	2966	2943	2920
0,8	2897	2874	2850	2827	2803	2780	2756	2732	2709	2685
0,9	2661	2637	2613	2589	2565	2541	2516	2492	2468	2444
1,0	2420	2396	2371	2347	2323	2299	2275	2251	2227	2203
1,1	2179	2155	2131	2107	2083	2059	2036	2012	1989	1965
1,2	1942	1919	1895	1872	1849	1826	1804	1781	1758	1736
1,3	1714	1691	1669	1647	1626	1604	1582	1561	1539	1518
1,4	1497	1476	1456	1435	1415	1394	1374	1354	1334	1315
1,5	1295	1276	1257	1238	1219	1200	1182	1163	1145	1127
1,6	1109	1092	1074	1057	1040	1023	1006	0989	0973	0957
1,7	0940	0925	0909	0893	0878	0863	0848	0833	0818	0804
1,8	0790	0775	0761	0748	0734	0721	0707	0694	0681	0669
1,9	0656	0644	0632	0620	0608	0596	0584	0573	0562	0551
2,0	0540	0529	0519	0508	0498	0488	0478	0468	0459	0449
2,1	0440	0431	0422	0413	0404	0396	0387	0379	0371	0363
2,2	0355	0347	0339	0332	0325	0317	0310	0303	0297	0290
2,3	0283	0277	0270	0264	0258	0252	0246	0241	0235	0229
2,4	0224	0219	0213	0203	0203	0198	0194	0189	0184	0180
2,5	0175	0171	0167	0163	0158	0154	0151	0147	0143	0139
2,6	0136	0132	0129	0126	0122	0119	0116	0113	0110	0107
2,7	0104	0101	0099	0096	0093	0091	0088	0086	0084	0081
2,8	0079	0077	0075	0073	0071	0069	0067	0065	0063	0061
2,9	0060	0058	0056	0055	0053	0051	0050	0048	0047	0046
3,0	0044	0043	0042	0040	0039	0038	0037	0036	0035	0034
4,0	0001	0001	0001	0000	0000	0000	0000	0000	0000	0000

Примечание. В таблице даны мантиссы значений функции (0,...).

Значения 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	0717
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	3759
0,5	3829	3899	3969	4039	4108	4177	4245	4313	4381	4448
0,6	4515	4581	4647	4713	4778	4843	4907	4971	5035	5098
0,7	5161	5223	5285	5346	5407	5467	5527	5587	5646	5705
0,8	5763	5821	5878	5935	5991	6047	6102	6157	6211	6265
0,9	6319	6372	6424	6476	6528	6579	6629	6679	6729	6778
1,0	6827	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	8789	8812	8836	8859	8882
1,6	8904	8926	8948	8969	8990	9011	9031	9051	9070	9089
1,7	9109	9127	9146	9164	9182	9199	9216	9233	9249	9265
1,8	9281	9297	9312	9327	9342	9357	9371	9385	9399	9412
1,9	9425	9439	9451	9464	9476	9488	9500	9512	9523	9534
2,0	9545	9556	9566	9576	9586	9596	9606	9615	9625	9634
2,1	9643	9651	9660	9668	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	9865	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	99889	99892	99896	99900
3,3	99903	—	—	—	—	—	—	—	—	—
3,4	99933	—	—	—	—	—	—	—	—	—
3,5	99953	—	—	—	—	—	—	—	—	—
4,0	99994	—	—	—	—	—	—	—	—	—
5,0	99999	—	—	—	—	—	—	—	—	—

$S(t)$ в распределении Стьюдента

t \ ν	1	2	3	4	5	6–7	8–10	11–15	16–25	26–30	∞
0,0	500	500	500	500	500	500	500	500	500	500	500,000
0,1	532	535	537	537	538	538	539	539	539	539	539,827
0,2	563	570	573	574	576	575	578	578	578	578	579,259
0,3	593	606	608	610	612	613	615	616	616	616	617,911
0,4	621	636	642	645	647	649	651	652	653	654	655,421
0,5	648	667	674	678	681	683	685	687	689	689	691,462
0,6	672	695	705	710	713	715	718	721	722	724	725,746
0,7	694	723	733	739	742	746	749	752	754	756	758,036
0,8	715	746	759	766	770	774	778	781	783	785	788,144
0,9	733	768	783	790	795	800	804	808	811	813	815,939
1,0	750	789	804	813	818	823	828	832	835	838	841,344
1,1	765	807	824	834	839	844	850	854	858	860	864,333
1,2	779	824	842	852	858	864	870	874	878	881	884,930
1,3	791	838	858	868	875	881	887	892	896	899	903,199
1,4	803	852	872	883	890	896	902	907	912	915	919,243
1,5	813	864	885	896	903	909	916	921	925	928	933,192
1,6	822	875	896	908	915	921	928	933	937	940	945,200
1,7	831	884	906	918	925	932	938	943	948	951	955,434
1,8	839	893	915	927	934	941	947	952	956	959	964,069
1,9	846	901	923	935	942	948	955	960	964	967	971,283
2,0	852	908	930	942	949	955	962	967	970	973	977,249
2,1	858	915	937	948	955	961	967	972	976	978	982,135
2,2	864	921	942	954	960	966	972	977	980	982	986,096
2,3	870	926	948	958	965	971	977	981	984	986	989,275
2,4	874	931	952	963	969	975	980	984	987	989	991,802
2,5	879	935	956	976	973	978	983	987	989	991	993,790
2,6	883	939	960	970	976	981	986	989	991	993	995,338
2,7	887	943	963	973	979	983	988	991	993	995	996,533
2,8	891	946	966	976	981	985	990	993	995	996	997,444
2,9	894	949	969	978	983	987	991	994	996	997	998,134
3,0	898	952	971	980	985	989	993	995	997	997	998,650
3,1	901	955	973	982	987	990	994	996	997	998	999,032
3,2	904	957	975	984	988	991	995	997	998	998	999,312
3,3	906	960	977	985	989	992	995	997	998	999	999,516
3,4	909	962	979	986	990	993	996	998	998	999	999,663

t \ ν	1	2	3	4	5	6–7	8–10	11–15	16–25	26–30	∞
3,5	911	964	980	988	991	993	997	998	999	—	999,767
3,6	914	965	982	989	992	994	997	998	—	—	999,840
3,7	916	967	983	990	993	995	998	999	—	—	999,892
3,8	918	969	984	990	994	996	998	999	—	—	999,927
3,9	920	970	985	991	994	996	998	999	—	—	999,951
4,0	922	971	986	992	995	997	998	—	—	—	999,968
4,1	924	973	987	993	995	997	999	—	—	—	999,979
4,2	926	974	988	993	996	998	999	—	—	—	999,986
4,3	927	975	988	994	996	998	999	—	—	—	999,991
4,4	929	976	989	994	996	998	—	—	—	—	999,994
4,5	930	977	989	995	997	998	—	—	—	—	999,996
4,6	932	978	990	995	997	998	—	—	—	—	999,997
4,7	933	979	991	995	997	999	—	—	—	—	999,998
4,8	935	980	991	996	998	999	—	—	—	—	999,999
4,9	936	980	992	996	998	999	—	—	—	—	999,999
5,0	937	981	992	996	998	999	—	—	—	—	999,999
5,1	938	982	993	996	998	—	—	—	—	—	999,999
5,2	940	982	993	997	998	—	—	—	—	—	999,999
5,3	941	983	993	997	998	—	—	—	—	—	999,999
5,4	942	984	994	997	998	—	—	—	—	—	—
5,5	943	984	994	997	999	—	—	—	—	—	—
5,6	943	984	994	997	999	—	—	—	—	—	—
5,7	945	985	995	998	999	—	—	—	—	—	—
5,8	946	986	995	998	999	—	—	—	—	—	—
5,9	947	986	995	998	999	—	—	—	—	—	—
6,0	947	987	995	998	—	—	—	—	—	—	—

Значения χ^2-критерия Пирсона при уровне значимости 0,10, 0,05, 0,01 и числе степеней свободы ν

df (ν)	0,10	0,05	0,01	*df* (ν)	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	34,17	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	15,99	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	31,41	37,57				

Значения критерия Дурбина – Ватсона при 5%-ном уровне существенности (для положительной автокорреляции)*

Число наблюдений n	$v = 1$		$v = 2$		$v = 3$	
	d_1	d_2	d_1	d_2	d_1	d_2
15	1,08	1,36	0,95	1,54	0,82	1,75
16	1,10	1,37	0,98	1,54	0,86	1,73
17	1,13	1,38	1,02	1,54	0,90	1,71
18	1,16	1,39	1,05	1,53	0,93	1,69
19	1,18	1,40	1,08	1,53	0,97	1,68
20	1,20	1,41	1,10	1,54	1,00	1,68
30	1,35	1,49	1,28	1,57	1,21	1,65
50	1,50	1,59	1,46	1,63	1,42	1,67

* v – число переменных в уравнении регрессии.

Значения функции $P(\lambda)$

λ	P	λ	P
0,30	1,0000	1,10	0,1777
0,35	9997	1,20	1122
0,40	9972	1,30	0681
0,45	9874	1,40	0397
0,50	9639	1,50	0222
0,55	9228	1,60	0120
0,60	8643	1,70	0062
0,65	7920	1,80	0032
0,70	7112	1,90	0015
0,75	6272	2,00	0007
0,80	5441	2,10	0003
0,85	4653	2,20	0001
0,90	3927	2,30	0001
0,95	3275	2,40	0000
1,00	2700	2,50	0000

Критические значения коэффициентов автокорреляции (r_a) при уровнях значимости $\alpha = 0{,}05$ и $\alpha = 0{,}01$

Объем выборки n	Положительные значения		Отрицательные значения	
	$\alpha = 0{,}05$	$\alpha = 0{,}01$	$\alpha = 0{,}05$	$\alpha = 0{,}01$
5	0,253	0,297	−0,753	−0,798
6	0,345	0,447	−0,708	−0,863
7	0,370	0,510	−0,674	−0,799
8	0,371	0,531	−0,625	−0,764
9	0,366	0,533	−0,593	−0,737
10	0,360	0,525	−0,564	−0,705
11	0,353	0,515	−0,539	−0,679
12	0,348	0,505	−0,516	−0,655
13	0,341	0,495	−0,497	−0,634
14	0,335	0,485	−0,479	−0,615
15	0,328	0,475	−0,462	−0,597
20	0,299	0,432	−0,399	−0,524

Значения *F*-критерия Фишера при уровне значимости 0,05

df_2	df_1 (ν_1)																
(ν_2)	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	9,12	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	3,23
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	2,70	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	2,30
13	4,67	3,81	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

df_2	df_1 (ν_1)																
(ν_2)	1	2	3	4	5	6	7	8	9	10	11	12	14	16	20	30	∞
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	1,76
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	2,28	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

Примечание: df_1 (ν_1) — число степеней свободы для большей дисперсии; df_2 (ν_2) — число степеней свободы для меньшей дисперсии.

Значения *t*-критерия Стьюдента при уровне значимости 0,10, 0,05, 0,01

df (ν)	α			*df* (ν)	α		
	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

Таблица для расчета средних коэффициентов роста (снижения) по средней параболической:

$$\bar{k} + (\bar{k})^2 + (\bar{k})^3 + \ldots + (\bar{k})^n = \frac{\sum_1^n y_i}{y_0}$$

$\bar{k}$ \ n	2	3	4	5	6	7	8	9	10
0,80	1,440	1,952	2,362	2,690	2,952	3,162	3,330	3,464	3,571
0,805	1,453	1,975	2,395	2,733	3,005	3,224	3,400	3,542	3,656
0,81	1,466	1,997	2,428	2,777	3,059	3,288	3,473	3,623	3,745
0,815	1,479	2,021	2,462	2,821	3,114	3,353	3,548	3,706	3,836
0,82	1,492	2,043	2,495	2,866	3,170	3,419	3,623	3,790	3,927
0,825	1,506	2,067	2,530	2,913	3,228	3,488	3,703	3,880	4,025
0,83	1,519	2,091	2,566	2,960	3,287	3,558	3,783	3,970	4,125
0,835	1,532	2,114	2,601	3,006	3,345	3,628	3,864	4,051	4,216
0,84	1,546	2,139	2,637	3,055	3,406	3,701	3,949	4,157	4,332
0,845	1,559	2,162	2,672	3,103	3,467	3,775	4,035	4,254	4,440
0,85	1,572	2,186	2,708	3,152	3,529	3,849	4,121	4,352	4,549
0,855	1,586	2,211	2,745	3,202	3,593	3,927	4,213	4,457	4,665
0,86	1,600	2,236	2,783	3,253	3,658	4,006	4,350	4,563	4,784
0,865	1,613	2,260	2,820	3,304	3,723	4,086	4,399	4,670	4,905
0,87	1,627	2,285	2,858	3,356	3,790	4,167	4,495	4,781	5,030
0,875	1,641	2,311	2,897	3,410	3,859	4,252	4,595	4,896	5,159
0,88	1,654	2,335	2,935	3,463	3,927	4,391	4,750	5,066	5,344
0,885	1,668	2,361	2,975	3,518	3,998	4,423	4,799	5,132	5,427
0,89	1,682	2,384	3,011	3,572	4,069	4,511	4,905	5,255	5,567
0,895	1,696	2,413	3,055	3,629	4,143	4,603	5,014	5,383	5,713
0,90	1,710	2,439	3,095	3,685	4,216	4,694	5,124	5,511	5,859
0,905	1,724	2,465	3,136	3,743	4,292	4,790	5,240	5,647	6,015
0,91	1,738	2,491	3,177	3,801	4,369	4,886	5,356	5,784	6,173
0,915	1,752	2,518	3,219	3,861	4,447	4,984	5,476	5,925	6,337
0,92	1,776	2,544	3,260	3,919	4,525	5,083	5,596	6,068	6,502
0,925	1,781	2,572	3,304	3,981	4,608	5,187	5,723	6,219	6,677
0,93	1,795	2,599	3,347	4,043	4,690	5,292	5,852	6,373	6,857
0,935	1,809	2,627	3,391	4,105	4,774	5,398	5,982	6,529	7,039
0,94	1,824	2,655	3,436	4,170	4,860	5,508	6,117	6,690	7,228
0,945	1,838	2,682	3,479	4,233	4,945	5,618	6,254	6,855	7,423

$\bar{k}$ \ n	2	3	4	5	6	7	8	9	10
0,95	1,852	2,709	3,523	4,297	5,032	5,730	6,394	7,024	7,624
0,955	1,867	2,738	3,570	4,364	5,123	5,847	6,539	7,200	7,831
0,96	1,881	2,766	3,615	4,430	5,212	5,963	6,684	7,376	8,041
0,965	1,896	2,795	3,662	4,499	5,306	6,086	6,838	7,563	8,264
0,97	1,911	2,823	3,708	4,566	5,399	6,207	6,990	7,751	8,488
0,975	1,926	2,852	3,756	4,637	5,496	6,334	7,150	7,947	8,723
0,98	1,940	2,881	3,803	4,707	5,593	6,461	7,311	8,145	8,962
0,985	1,955	2,911	3,852	4,779	5,693	6,592	7,478	8,351	9,211
0,99	1,970	2,940	3,900	4,851	5,792	6,724	7,647	8,560	9,464
0,995	1,985	2,970	3,950	4,925	5,896	6,861	7,822	8,778	9,729
1,000	2,000	3,000	4,000	5,000	6,000	7,000	8,000	9,000	10,000
1,005	2,015	3,030	4,050	5,075	6,106	7,142	8,182	9,228	10,279
1,01	2,030	3,060	4,101	5,152	6,214	7,286	8,368	9,462	10,567
1,015	2,045	3,091	4,152	5,230	6,323	7,433	8,560	9,703	10,863
1,02	2,060	3,122	4,204	5,308	6,434	7,583	8,754	9,949	11,168
1,025	2,076	3,152	4,256	5,388	6,547	7,736	8,954	10,200	11,480
1,03	2,091	3,184	4,309	5,468	6,662	7,892	9,159	10,464	11,808
1,035	2,106	3,215	4,362	5,550	6,779	8,052	9,368	10,731	12,141
1,04	2,122	3,246	4,416	5,633	6,898	8,214	9,583	11,006	12,487
1,045	2,137	3,278	4,471	5,717	7,019	8,380	9,802	11,288	12,841
1,05	2,152	3,310	4,526	5,802	7,142	8,549	10,026	11,578	13,207
1,055	2,168	3,342	4,581	5,888	7,267	8,721	10,256	11,875	13,583
1,06	2,184	3,375	4,637	5,975	7,394	8,898	10,492	12,181	13,972
1,065	2,199	3,407	4,694	6,064	7,523	9,076	10,731	12,493	14,371
1,07	2,215	3,440	4,751	6,153	7,654	9,260	10,978	12,817	14,784
1,075	2,231	3,473	4,808	6,244	7,788	9,447	11,230	13,148	15,209
1,08	2,246	3,506	4,867	6,336	7,923	9,636	11,487	13,486	15,649
1,085	2,262	3,540	4,925	6,418	8,061	9,831	11,752	13,835	16,096
1,09	2,278	3,573	4,985	6,502	8,200	10,028	12,020	14,192	16,560
1,095	2,294	3,607	5,044	6,619	8,342	10,230	12,297	14,560	17,038
1,10	2,310	3,641	5,105	6,715	8,487	10,436	12,579	14,937	17,531
1,105	2,326	3,675	5,166	6,814	8,634	10,646	12,868	15,324	18,038
1,11	2,342	3,710	5,228	6,913	8,783	10,860	13,164	15,722	18,561
1,115	2,358	3,744	5,290	7,013	8,935	11,077	13,466	16,130	19,100
1,12	2,374	3,779	5,353	7,115	9,089	11,300	13,776	16,549	19,654
1,125	2,391	3,814	5,416	7,218	9,245	11,526	14,092	16,978	20,226
1,13	2,407	3,850	5,480	7,322	9,405	11,757	14,416	17,420	20,814
1,135	2,423	3,885	5,545	7,428	9,566	11,993	14,747	17,873	21,420

$\bar{k}$ \ n	2	3	4	5	6	7	8	9	10
1,14	2,439	3,921	5,610	7,535	9,730	12,233	15,085	18,337	22,044
1,145	2,456	3,957	5,676	7,644	9,897	12,477	15,432	18,814	22,688
1,15	2,472	3,993	5,742	7,754	10,067	12,727	15,786	19,304	23,349
1,155	2,489	4,030	5,809	7,865	10,239	12,981	16,148	19,806	24,031
1,16	2,506	4,066	5,877	7,977	10,414	13,240	16,518	20,321	24,733
1,165	2,522	4,103	5,945	8,091	10,591	13,504	16,897	20,850	25,455
1,17	2,539	4,140	6,014	8,207	10,772	13,772	17,285	21,393	26,200
1,175	2,556	4,178	6,084	8,324	10,955	14,047	17,681	21,950	26,966
1,18	2,572	4,215	6,154	8,442	11,141	14,327	18,068	22,521	27,755
1,185	2,589	4,253	6,225	8,562	11,330	14,612	18,500	23,107	28,567
1,19	2,606	4,291	6,296	8,683	11,523	14,902	18,923	23,709	29,403
1,195	2,623	4,329	6,369	8,806	11,718	15,198	19,375	24,326	30,264
1,20	2,640	4,368	6,442	8,930	11,916	15,499	19,799	24,959	31,151

Значения функции e^{-x}

x	e^{-x}	x	e^{-x}	x	e^{-x}	x	e^{-x}	x	e^{-x}
0,00	1,0000	0,40	0,6703	0,80	0,4493	1,20	0,3012	1,60	0,2019
01	0,9900	41	6637	81	4449	21	2982	61	1999
02	9802	42	6570	82	4404	22	2952	62	1979
03	9704	43	6505	83	4360	23	2923	63	1959
04	9608	44	6440	84	4317	24	2894	64	1940
05	9512	45	6376	85	4274	25	2865	65	1920
06	9418	46	6313	86	4232	26	2837	66	1901
07	9324	47	6250	87	4190	27	2808	67	1882
08	9231	48	6188	88	4148	28	2780	68	1864
09	9139	49	6126	89	4107	29	2753	69	1845
10	9048	50	6065	90	4066	30	2725	70	1827
11	8958	51	6005	91	4025	31	2698	71	1809
12	8869	52	5945	92	3985	32	2671	72	1791
13	8781	53	5886	93	3916	33	2645	73	1773
14	8694	54	5827	94	3906	34	2618	74	1755
15	8607	55	5769	95	3867	35	2592	75	1738
16	8521	56	5712	96	3829	36	2567	76	1720
17	8437	57	5655	97	3791	37	2541	77	1703
18	8353	58	5599	98	3753	38	2516	78	1686
19	8270	59	5543	99	3716	39	2491	79	1670
20	8187	60	5488	1,00	3679	40	2466	80	1653
21	8106	61	5434	01	3642	41	2441	81	1637
22	8025	62	5379	02	3606	42	2417	82	1620
23	7945	63	5326	03	3570	43	2393	83	1604
24	7866	64	5273	04	3535	44	2369	84	1588
25	7788	65	5220	05	3499	45	2346	85	1572
26	7711	66	5169	06	3465	46	2322	86	1557
27	7634	67	5117	07	3430	47	2299	87	1541
28	7558	68	5066	08	3396	48	2276	88	1526
29	7483	69	5016	09	3362	49	2254	89	1511
30	7408	70	4966	10	3329	50	2231	90	1496
31	7334	71	4916	11	3296	51	2209	91	1481
32	7261	72	4868	12	3263	52	2187	92	1466
33	7189	73	4819	13	3230	53	2165	93	1451
34	7118	74	4771	14	3198	54	2144	94	1437
35	7047	75	4724	15	3166	55	2122	95	1424
36	6977	76	4677	16	3135	56	2101	96	1409
37	6907	77	4630	17	3104	57	2080	97	1395
38	6839	78	4584	18	3073	58	2060	98	1381
39	6771	79	4538	19	3042	59	2039	99	1367
								2,00	1353

Основные формулы

Средние величины

Средняя арифметическая:

- простая (для несгруппированных данных) $\bar{x} = \dfrac{\sum x_i}{n}$;
- взвешенная (для сгруппированных данных) $\bar{x} = \dfrac{\sum x_i f_i}{\sum f_i}$.

Средняя гармоническая:

- простая $\bar{x} = \dfrac{n}{\sum \dfrac{1}{x_i}}$;
- взвешенная $\bar{x} = \dfrac{\sum V_i}{\sum \dfrac{V_i}{x_i}}$ или $\bar{x} = \dfrac{\sum M_i}{\sum \dfrac{M_i}{x_i}}$.

Средняя геометрическая:

- простая $\bar{x} = \sqrt[n]{x_1 x_2 \cdots x_n} = \sqrt[n]{\prod x_i}$;
- взвешенная $\bar{x} = \sqrt[\sum f_i]{x_1^{f_1} x_2^{f_2} \cdots x_n^{f_n}} = \sqrt[\sum f_i]{\prod (x_i)^{f_i}}$.

Средняя квадратическая:

- простая $\bar{x} = \sqrt{\dfrac{\sum x_i^2}{n}}$;
- взвешенная $\bar{x} = \sqrt{\dfrac{\sum x_i^2 f_i}{\sum f_i}}$.

Структурные средние (в интервальном вариационном ряду)

Мода (в рядах с равными интервалами)

$$Mo = x_0 + h \frac{f_{Mo} - f_{Mo-1}}{(f_{Mo} - f_{Mo-1}) + (f_{Mo} - f_{Mo+1})},$$

где x_0 – нижняя граница модального интервала; h – величина модального интервала; f_{Mo}, f_{Mo-1}, f_{Mo+1} – частоты или частости соответственно модального, предмодального и послемодального интервалов.

Медиана

$$Me = x_0 + h\frac{\frac{1}{2}\sum f - F_{Me-1}}{f_{Me}},$$

где $\frac{1}{2}\sum f$ – порядковый номер медианы; F_{Me-1} – накопленная частота или частость до медианного интервала.

Квартили:

- первый $Q_1 = x_{Q_1} + h\frac{\frac{1}{4}\sum f - F_{Q_1-1}}{f_{Q_1}}$;
- второй $Q_2 = Me$;
- третий $Q_3 = x_{Q_3} + h\frac{\frac{3}{4}\sum f - F_{Q_3-1}}{f_{Q_3}}$.

Децили:

- первый $D_1 = x_{D_1} + h\frac{\frac{1}{10}\sum f - F_{D_1-1}}{f_{D_1}}$;
- второй $D_2 = x_{D_2} + h\frac{\frac{2}{10}\sum f - F_{D_2-1}}{f_{D_2}}$;

⋮

- девятый $D_9 = x_{D_9} + h\frac{\frac{9}{10}\sum f - F_{D_9-1}}{f_{D_9}}$.

Децильный коэффициент дифференциации признака в вариационном ряду ($K_д$ или *ДКД*)

$$ДКД = \frac{D_9}{D_1}.$$

Показатели вариации

Размах вариации

$$R = x_{\max} - x_{\min}.$$

Среднее линейное (абсолютное) отклонение:

- для несгруппированных данных $\bar{d} = \dfrac{\sum |x_i - \bar{x}|}{n}$;
- для сгруппированных данных $\bar{d} = \dfrac{\sum |x_i - \bar{x}| f_i}{\sum f_i}$.

Дисперсия вариационного признака:

- для несгруппированных данных $\sigma^2 = \dfrac{\sum (x_i - \bar{x})^2}{n}$;
- для сгруппированных данных $\sigma^2 = \dfrac{\sum (x_i - \bar{x})^2 f_i}{\sum f_i}$.

Среднее квадратическое отклонение:

- для несгруппированных данных $\sigma = \sqrt{\dfrac{\sum (x_i - \bar{x})^2}{n}}$;
- для сгруппированных данных $\sigma = \sqrt{\dfrac{\sum (x_i - \bar{x})^2 f_i}{\sum f_i}}$.

Коэффициент вариации (относительный показатель вариации)

$$V = \frac{\sigma}{\bar{x}} 100\% \quad \text{или} \quad V = \frac{\bar{d}}{Me} 100\%.$$

Правило сложения дисперсий (для совокупности, разбитой на группы):

$$\sigma^2 = \overline{\sigma^2} + \delta^2.$$

Средняя из групповых дисперсий

$$\overline{\sigma^2} = \frac{\sum \sigma_i^2 n_i}{\sum n_i},$$

где n_i – число единиц в i-й группе.

Межгрупповая дисперсия

$$\delta^2 = \frac{\sum (\bar{x}_i - \bar{x}_{\text{общ}})^2 n_i}{\sum n_i}.$$

Эмпирический коэффициент детерминации

$$\eta^2_{\text{эмп}} = \frac{\delta^2}{\sigma^2}.$$

Эмпирическое корреляционное отношение

$$\eta_{\text{эмп}} = \sqrt{\frac{\delta^2}{\sigma^2}}.$$

Дисперсия альтернативного признака

$$\sigma_p^2 = pq.$$

Общая дисперсия доли (для совокупности, разбитой на группы)

$\sigma_{\bar{p}}^2 = \bar{p}(1-\bar{p})$ и $\sigma_{\bar{p}}^2 = \overline{\sigma_{p_i}^2} + \delta_{p_i}^2$ (по правилу сложения дисперсий)

Средняя из групповых дисперсий доли

$$\overline{\sigma_{p_i}^2} = \overline{p_i(1-p_i)} = \frac{\sum p_i(1-p_i)n_i}{\sum n_i}.$$

Межгрупповая дисперсия доли

$$\delta_{p_i}^2 = \frac{\sum (p_i - \bar{p})^2 n_i}{\sum n_i}.$$

Моменты распределения

Условные моменты

$$M_k^A = \frac{\sum (x_i - A)^k f_i}{\sum f_i}.$$

Начальные моменты ($A = 0$)

$$M_k = \frac{\sum x_i^k f_i}{\sum f_i}$$

1) первого порядка $M_1 = \dfrac{\sum x_i f_i}{\sum f_i} = \bar{x}$;

2) второго порядка $M_2 = \dfrac{\sum x_i^2 f_i}{\sum f_i} = \overline{x^2}$;

3) третьего порядка $M_3 = \dfrac{\sum x_i^3 f_i}{\sum f_i} = \overline{x^3}$;

4) четвертого порядка $M_4 = \dfrac{\sum x_i^4 f_i}{\sum f_i} = \overline{x^4}$.

Центральные моменты ($A = \bar{x}$)

$$\mu_k = \frac{\sum (x_i - \bar{x})^k f_i}{\sum f_i}$$

1) $\mu_1 = \dfrac{\sum (x_i - \bar{x}) f_i}{\sum f_i} = 0;$

2) $\mu_2 = \dfrac{\sum (x_i - \bar{x})^2 f_i}{\sum f_i} = \sigma^2;$

3) $\mu_3 = \dfrac{\sum (x_i - \bar{x})^3 f_i}{\sum f_i};$

4) $\mu_4 = \dfrac{\sum (x_i - \bar{x})^4 f_i}{\sum f_i};$

$\mu_2 = M_2 - M_1^2;$

$\mu_3 = M_3 - 3M_1M_2 + 2M_1^3;$

$\mu_4 = M_4 - 4M_1M_3 + 6M_1^2M_2 - 3M_1^4.$

Нормированные моменты

$$r_k = \frac{\mu_k}{\sigma_k}$$

1) $r_1 = \dfrac{\mu_1}{\sigma} = 0;$

2) $r_2 = \dfrac{\mu_2}{\sigma^2};$

3) $r_3 = \dfrac{\mu_3}{\sigma^3};$

4) $r_4 = \dfrac{\mu_4}{\sigma^4}.$

Коэффициент асимметрии (скошенности):

а) Пирсона $As = \dfrac{\bar{x} - Mo}{\sigma};$

б) нормированный момент третьего порядка $r_3 = \dfrac{\mu_3}{\sigma^3} = As.$

Эксцесс распределения (крутизна распределения)

$$Ex = \frac{\mu_4}{\sigma^4} - 3.$$

Кривая нормального распределения

$$y_t = \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} = \varphi(t) \qquad \left(t = \frac{x - \bar{x}}{\sigma}\right).$$

Теоретические частоты при выравнивании ряда:

а) по кривой нормального распределения $f_{\text{теор}} = m' = \dfrac{Nh}{\sigma}\varphi(t)$

б) по кривой Пуассона $f_{\text{теор}} = m' = NP(x) = N\dfrac{a^x e^{-a}}{x!}$ $(a = \bar{x})$.

Критерии согласия:

а) Пирсона χ^2 (хи-квадрат)

$$\chi^2 = \sum \frac{(m_{\text{эмп}} - m_{\text{теор}})^2}{m_{\text{теор}}} \quad \text{или} \quad \chi^2 = \sum \frac{(f_{\text{эмп}} - f_{\text{теор}})^2}{f_{\text{теор}}},$$

где $m = f$ – частоты;

б) Романовского $K_{\text{Р}} = \dfrac{\left|\chi^2 - \nu\right|}{\sqrt{2\nu}}$;

в) Колмогорова $\lambda = \dfrac{D}{\sqrt{N}}$, где $D = \max\left|F_{\text{эмп}} - F_{\text{теор}}\right|$.

Показатели концентрации

Коэффициент Джини

$$G = \sum p_i q_{i+1} - \sum p_{i+1} q_i.$$

Коэффициент Герфиндаля

$$H = \sum \left(\frac{x_i m_i}{\sum x_i m_i}\right)^2 \quad \text{или} \quad H = \sum \left(\frac{x_i w_i}{\sum x_i w_i}\right)^2.$$

Коэффициент Лоренца

$$L = \frac{1}{2}\sum \left| w_i - \frac{x_i w_i}{\sum x_i w_i} \right|.$$

Выборочный метод

Средняя ошибка выборки:

Вид выборки	Повторный отбор	Бесповторный отбор
Собственно случайная и механическая:		
для средней ($\tilde{x}$)	$\mu = \dfrac{\sigma}{\sqrt{n}}$	$\mu = \sqrt{\dfrac{\sigma^2}{n}\left(1 - \dfrac{n}{N}\right)}$
для доли (w)	$\mu = \sqrt{\dfrac{w(1-w)}{n}}$	$\mu = \sqrt{\dfrac{w(1-w)}{n}\left(1 - \dfrac{n}{N}\right)}$
Типическая (при отборе, пропорциональном объему групп):		
для средней ($\tilde{x}$)	$\mu = \sqrt{\dfrac{\overline{\sigma^2}}{n}}$	$\mu = \sqrt{\dfrac{\overline{\sigma^2}}{n}\left(1 - \dfrac{n}{N}\right)}$
для доли (w)	$\mu = \sqrt{\dfrac{\overline{w(1-w)}}{n}}$	$\mu = \sqrt{\dfrac{\overline{w(1-w)}}{n}\left(1 - \dfrac{n}{N}\right)}$
Серийная (гнездовая)*:		
для средней ($\tilde{x}$)	$\mu = \sqrt{\dfrac{\delta_{\bar{x}}^2}{s}}$	$\mu = \sqrt{\dfrac{\delta_{\bar{x}}^2}{s}\left(\dfrac{S-s}{S-1}\right)}$
для доли (w)	$\mu = \sqrt{\dfrac{\delta_w^2}{s}}$	$\mu = \sqrt{\dfrac{\delta_w^2}{s}\left(\dfrac{S-s}{S-1}\right)}$

* При серийной выборке повторный отбор практически не применяется.

Предельная ошибка выборки $\Delta = t\mu$,

где t – коэффициент доверия.

Средняя ошибка малой выборки $\mu_{\text{м.в}} = \sqrt{\dfrac{\sigma^2}{n-1}}$.

Предельная ошибка малой выборки $\Delta_{\text{м.в}} = t\mu_{\text{м.в}}$.

(При этом $P(|\tilde{x} - \bar{x}| \leqslant t\mu) = 2S(t) - 1$.)

Необходимый объем выборки:

Вид выборки	Повторный отбор	Бесповторный отбор
Собственно случайная и механическая:		
для средней ($\bar{x}$)	$n = \dfrac{t^2\sigma^2}{\Delta^2}$	$n = \dfrac{Nt^2\sigma^2}{N\Delta^2 + t^2\sigma^2}$
для доли (p)	$n = \dfrac{t^2 pq}{\Delta^2}$	$n = \dfrac{Nt^2 pq}{N\Delta^2 + t^2 pq}$
Типическая:		
для средней ($\bar{x}$)	$n = \dfrac{t^2\overline{\sigma^2}}{\Delta^2}$	$n = \dfrac{Nt^2\overline{\sigma^2}}{N\Delta^2 + t^2\overline{\sigma^2}}$
для доли (p)	$n = \dfrac{t^2\overline{pq}}{\Delta^2}$	$n = \dfrac{Nt^2\overline{pq}}{N\Delta^2 + t^2\overline{pq}}$
Серийная:		
для средней ($\bar{x}$)	$s = \dfrac{t^2\delta_{\bar{x}}^2}{\Delta^2}$	$s = \dfrac{St^2\delta_{\bar{x}}^2}{(S-1)\Delta^2 + t^2\delta_{\bar{x}}^2}$
для доли (p)	$s = \dfrac{t^2\delta_p^2}{\Delta^2}$	$s = \dfrac{St^2\delta_p^2}{(S-1)\Delta^2 + t^2\delta_p^2}$

Примечание. При неизвестном значении pq в формулах для n берется максимум этого показателя: $\max(pq) = 0{,}25$ (т.е. в случае когда $p = q$). Если для определения n задана относительная ошибка выборки ($\Delta_{отн}$), то во всех приведенных в таблице формулах берется относительный показатель вариации – коэффициент вариации V и тогда, например, для собственно случайной выборки

$$n = \frac{Nt^2V^2}{N\Delta_{отн}^2 + t^2V^2}.$$

Средняя ошибка разности двух выборочных средних

$$\mu_{разн} = \sqrt{\mu_1^2 + \mu_2^2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}.$$

Изучение корреляционных взаимосвязей

Однофакторные модели связи (уравнения регрессии «*y* по *x*»):

- линейная (по прямой) $\bar{y}_x = a_0 + a_1 x$.

  Система нормальных уравнений для нахождения параметров a_0 и a_1

  $$\begin{cases} na_0 + a_1 \sum x = \sum y, \\ a_0 \sum x + a_1 \sum x^2 = \sum xy \end{cases}$$

  или $a_1 = \dfrac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2}$, $\quad a_0 = \bar{y} - a_1 \bar{x}$;

- по параболе 2-го порядка $\bar{y}_x = a_0 + a_1 x + a_2 x^2$.

  Система нормальных уравнений для нахождения параметров

  $$\begin{cases} na_0 + a_1 \sum x + a_2 \sum x^2 = \sum y, \\ a_0 \sum x + a_1 \sum x^2 + a_2 \sum x^3 = \sum xy, \\ a_0 \sum x^2 + a_1 \sum x^3 + a_2 \sum x^4 = \sum x^2 y; \end{cases}$$

- гиперболическая $\bar{y}_x = a_0 + a_1 \dfrac{1}{x}$.

  Система нормальных уравнений для нахождения параметров

  $$\begin{cases} na_0 + a_1 \sum \dfrac{1}{x} = \sum y, \\ a_0 \sum \dfrac{1}{x} + a_1 \sum \left(\dfrac{1}{x}\right)^2 = \sum \dfrac{y}{x}. \end{cases}$$

Многофакторная прямолинейная модель связи

$$\bar{y}_{1,\,2,\,\ldots,\,k} = f(x_1, x_2, \ldots, x_k) = a_0 + a_1 x_1 + a_2 x_2 + \ldots + a_k x_k.$$

***t*-критерий Стьюдента** для проверки значимости параметров уравнения регрессии

$$t_{\text{расч}} = \frac{a_i}{\mu_{a_i}}.$$

$t_{\text{расч}}$ сопоставляется с $t_{\text{табл}}$ (α; $\nu = n - k - 1$) или (α; $\nu = n - m$), где k – число факторных признаков, m – число параметров в уравнении регрессии.

***F*-критерий Фишера** для проверки значимости уравнений регрессии

$$F_{\text{расч}} = \frac{\delta^2_{\text{фактор}} / (m-1)}{\sigma^2_{\text{ост}} / (n-m)},$$

где m – число параметров в уравнении регрессии,

$$F_{\text{расч}} = \frac{\delta^2_{\text{фактор}} / k}{\sigma^2_{\text{ост}} / (n-k-1)},$$

где k – число факторных признаков.

$F_{\text{расч}}$ сопоставляется с $F_{\text{табл}}$ $(\alpha;\ \nu_1 = m-1;\ \nu_2 = n-m)$ или $(\alpha;\ \nu_1 = k;\ \nu_2 = n-k-1)$.

Показатели тесноты связи между двумя количественными признаками

Линейный коэффициент корреляции

$$r = a_1 \frac{\sigma_x}{\sigma_y}; \quad r = \frac{\overline{xy} - \bar{x}\,\bar{y}}{\sigma_x \sigma_y}; \quad r = \frac{\sum (x-\bar{x})(y-\bar{y})}{n\sigma_x \sigma_y};$$

$$r = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sqrt{\sum (x-\bar{x})^2 \sum (y-\bar{y})^2}}; \quad r = \frac{n\sum xy - \sum x \sum y}{\sqrt{\left[n\sum x^2 - (\sum x)^2\right]\left[n\sum y^2 - (\sum y)^2\right]}}.$$

Средняя квадратическая ошибка линейного коэффициента корреляции

а) при $n > 50$ $\sigma_r = \dfrac{1-r^2}{\sqrt{n}}$ (r считается значимым, если $\dfrac{|r|}{\sigma_r} > 3$);

б) при $n < 30$ значимость r проверяется по t-критерию Стьюдента: $t_{\text{расч}} = \dfrac{r\sqrt{n-2}}{\sqrt{1-r^2}}$ сопоставляется с $t_{\text{табл}}$ $(\alpha;\ \nu = n-2)$.

Если $t_{\text{расч}} > t_{\text{табл}}$, то r – значим.

Корреляционное отношение (теоретическое)

$$\eta_{\text{теор}} = \sqrt{\frac{\delta^2_{\text{фактор}}}{\sigma^2}}; \quad \eta_{\text{теор}} = \sqrt{1 - \frac{\sigma^2_{\text{ост}}}{\sigma^2}}.$$

Коэффициент Фехнера

$$K_{\Phi} = \frac{\sum C - \sum H}{\sum C + \sum H}.$$

Коэффициент корреляции рангов

а) Спирмэна

$$\rho = 1 - \frac{6\sum d^2}{n(n^2-1)};$$

б) Кендэла:

- при отсутствии связанных (одинаковых) рангов $\tau = \dfrac{2S}{n(n-1)}$;
- при наличии связанных рангов

$$\tau = \frac{S}{\sqrt{\left[\dfrac{n(n-1)}{2} - U_x\right]\left[\dfrac{n(n-1)}{2} - U_y\right]}}.$$

Коэффициент конкордации (для измерения тесноты связи между двумя (и более) признаками):

- при отсутствии связанных рангов $W = \dfrac{12S}{m^2(n^3 - n)}$;
- при наличии связанных рангов $W = \dfrac{12S}{m^2(n^3 - n) - m\sum(t^3 - t)}$.

Показатели тесноты связи между двумя качественными признаками

Коэффициент ассоциации

$$K_{ac} = \frac{ad - bc}{ad + bc}.$$

Коэффициент контингенции

$$K_{конт} = \frac{ad - bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}.$$

для «четырехклеточных» таблиц

a	b
c	d

Коэффициенты взаимной сопряженности:

а) Пирсона

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}} \quad \text{или} \quad C = \sqrt{\frac{\varphi^2}{\varphi^2 + 1}};$$

б) Чупрова

$$K_{ч} = \sqrt{\frac{\chi^2}{n\sqrt{(k_1 - 1)(k_2 - 1)}}} \quad \text{или} \quad K_{ч} = \sqrt{\frac{\varphi^2}{\sqrt{(k_1 - 1)(k_2 - 1)}}}.$$

где $\varphi^2 = \dfrac{\chi^2}{n}$

Ряд динамики: $y_0, y_1, y_2, \ldots, y_n$.

Коэффициенты роста:

- базисные $k_{pi} = \dfrac{y_i}{y_0}$;
- цепные $k_{pi} = \dfrac{y_i}{y_{i-1}}$.

Темп роста $T_p = k_p \cdot 100\%$.

Темп прироста $T_{пр} = T_p - 100\%$.

Средний уровень ряда динамики (средняя хронологическая):

- в интервальном ряду абсолютных величин и в ряду средних величин

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n};$$

- в моментном ряду с равноотстоящими уровнями

$$\bar{y} = \frac{\frac{1}{2}y_1 + y_2 + \ldots + y_{n-1} + \frac{1}{2}y_n}{n-1} = \frac{\frac{y_1 + y_n}{2} + \sum_{i=2}^{n-1} y_i}{n-1}.$$

Средний коэффициент (темп) роста (за период с 1-го по n-й):

- средняя геометрическая из цепных коэффициентов роста

$\bar{k} = \sqrt[n]{k_1 k_2 \cdots k_n}$ ($\bar{k}$ обеспечивает достижение конечного уровня ряда: $y_n = y_0(\bar{k})^n$)

или тождественная ей

$\bar{k} = \sqrt[n]{\dfrac{y_n}{y_0}}$ (под корнем отношение абсолютных уровней, или базисный коэффициент роста);

- средняя параболическая $\bar{k} + (\bar{k})^2 + (\bar{k})^3 + ... + (\bar{k})^n = \dfrac{\sum\limits_{i=1}^{n} y_i}{y_0}$

  ($\bar{k}$ обеспечивает достижение суммы уровней за анализируемый период n: $\sum\limits_{i=1}^{n} y_i = y_0\left[\bar{k} + (\bar{k})^2 + ... + (\bar{k})^n\right]$.

Коэффициент автокорреляции между уровнями ряда

$$r_a = \frac{\overline{y_t y_{t-1}} - \bar{y}_t \bar{y}_{t-1}}{\sigma_{y_t} \sigma_{y_{t-1}}}$$

или приближенно

$$r_a = \frac{\overline{y_t y_{t-1}} - (\bar{y}_t)^2}{\sigma^2_{y_t}} = \frac{\sum y_t y_{t-1} - n(\bar{y}_t)^2}{\sum y_t^2 - n(\bar{y}_t)^2}.$$

Коэффициент автокорреляции между остаточными величинами ($\varepsilon_t = y_t - \hat{y}_t$):

- коэффициент Андерсона $r_a = \dfrac{\sum\limits_{t=2}^{n} \varepsilon_t \varepsilon_{t-1}}{\sum\limits_{t=1}^{n} \varepsilon_t^2}$;
- критерий Дурбина – Ватсона $d = \dfrac{\sum\limits_{t=2}^{n} \left(\varepsilon_t - \varepsilon_{t-1}\right)^2}{\sum\limits_{t=1}^{n} \varepsilon_t^2}$.

Измерение корреляции между уровнями двух рядов динамики:

- при отсутствии автокорреляции в каждом ряду $r_{xy} = \dfrac{\overline{xy} - \overline{xy}}{\sigma_x \sigma_y}$;
- при наличии автокорреляции в каждом ряду:

  $r_{d_x d_y} = \dfrac{\sum d_x d_y}{\sqrt{\sum d_x^2 \sum d_y^2}}$ – коэффициент корреляции отклонений уровней ряда от тренда ($d_x = x_i - \hat{x}_t$ и $d_y = y_i - \hat{y}_t$);

  $r_{\Delta_x \Delta_y} = \dfrac{\sum \Delta_x \Delta_y}{\sqrt{\sum \Delta_x^2 \sum \Delta_y^2}}$.

Экономические индексы

Индивидуальные индексы:

- физического объема продукции $i_q = \frac{q_1}{q_0}$;
- цен $i_p = \frac{p_1}{p_0}$;
- себестоимости $i_c = \frac{c_1}{c_0}$ (или $i_z = \frac{z_1}{z_0}$);
- урожайности $i_y = \frac{y_1}{y_0}$;
- производительности труда:
  а) через выработку продукции (w) в единицу времени $i_w = \frac{w_1}{w_0}$;
  б) через затраты времени (t) на единицу продукции $i_w = \frac{t_0}{t_1}$;
- стоимости продукции $i_{pq} = \frac{p_1 q_1}{p_0 q_0}$.

Общие (сводные) индексы в агрегатной форме:

- физического объема $I_q = \frac{\sum q_1 p_0}{\sum q_0 p_0}$;
- цен:
  а) Пааше $I_p^{\text{П}} = \frac{\sum p_1 q_1}{\sum p_0 q_1}$;
  б) Ласпейреса $I_p^{\text{Л}} = \frac{\sum p_1 q_0}{\sum p_0 q_0}$;
- стоимости продукции или товарооборота $I_{pq} = \frac{\sum p_1 q_1}{\sum p_0 q_0}$;
- себестоимости продукции $I_c = \frac{\sum c_1 q_1}{\sum c_0 q_1}$;
- издержек производства $I_{cq} = \frac{\sum c_1 q_1}{\sum c_0 q_0}$;

- производительности труда:

  а) через прямой показатель $w = \frac{Q}{T}$

  $I_w = \frac{\sum w_1 T_1}{\sum w_0 T_1}$ (для однородной продукции, измеряемой в натуральном выражении, и разнородной продукции, измеряемой в стоимостном выражении в сопоставимых ценах);

  б) через трудоемкость $t = \frac{T}{q}$

  $I_{\text{пр.тр}} = \frac{\sum t_0 q_1}{\sum t_1 q_1}$;

- затрат времени на единицу продукции $I_t = \frac{\sum t_1 q_1}{\sum t_0 q_1}$;
- общих затрат времени на производство продукции $I_{tq} = \frac{\sum t_1 q_1}{\sum t_0 q_0}$.

Средний арифметический индекс физического объема продукции

$$\bar{I}_q = \frac{\sum i_q q_0 p_0}{\sum q_0 p_0}.$$

Средний гармонический индекс цен

$$\bar{I}_p = \frac{\sum p_1 q_1}{\sum \frac{p_1 q_1}{i_p}}$$ (тождественный агрегатному индексу цен Пааше).

Индекс потребительских цен

$$\text{ИПЦ} = \frac{\sum i_p p_0 q_0}{\sum p_0 q_0}$$ (среднеарифметическая форма, тождественная агрегатному индексу цен Ласпейреса)

или

$$\text{ИПЦ} = \frac{\sum i_p d_{p_0 q_0}}{\sum d_{p_0 q_0}},$$

где $d_{p_0 q_0}$ – доля отдельных групп товаров в общем объеме товарооборота базисного периода.

«Идеальный» индекс цен Фишера

$$I_p^{\Phi} = \sqrt{\frac{\sum p_1 q_0}{\sum p_0 q_0} \frac{\sum p_1 q_1}{\sum p_0 q_1}} = \sqrt{I_p^{\text{Л}} I_p^{\text{П}}}.$$

Взаимосвязи некоторых индексов

$$I_{pq} = I_p^{\text{П}} I_q = \frac{\sum p_1 q_1}{\sum p_0 q_1} \frac{\sum q_1 p_0}{\sum q_0 p_0};$$

$$I_{cq} = I_c I_q = \frac{\sum c_1 q_1}{\sum c_0 q_1} \frac{\Sigma q_1 c_0}{\sum q_0 c_0};$$

$$I_{qt} = I_q I_t = \frac{\sum q_1 t_0}{\sum q_0 t_0} \frac{\sum q_1 t_1}{\sum q_1 t_0}.$$

Индексы переменного и фиксированного составов

Общая формула индекса переменного состава (отражает изменение средней величины индексируемого показателя за счет двух факторов: x и f)

$$I_{\text{п.с}} = \bar{x}_1 : \bar{x}_0 = \frac{\sum x_1 f_1}{\sum f_1} : \frac{\sum x_0 f_0}{\sum f_0},$$

где x – индексируемая величина; f – веса при расчете средних.

Общая формула индекса фиксированного (постоянного) состава (отражает изменение средней величины индексируемого показателя только за счет изменения x)

$$I_{\text{ф.с}} = \frac{\sum x_1 f_1}{\sum f_1} : \frac{\sum x_0 f_1}{\sum f_1}, \quad \text{или} \quad I_{\text{ф.с}} = \frac{\sum x_1 f_1}{\sum x_0 f_1}.$$

Общая формула индекса структурных сдвигов (отражает изменение средней величины индексируемого показателя только за счет изменения f – структуры совокупности)

$$I_{\text{стр}} = \frac{\sum x_0 f_1}{\sum f_1} : \frac{\sum x_0 f_0}{\sum f_0} \quad \text{или} \quad I_{\text{стр}} = I_{\text{п.с}} : I_{\text{ф.с}}.$$

СПИСОК РЕКОМЕНДУЕМОЙ ЛИТЕРАТУРЫ

1. *Адамов В.Е.* Факторный индексный анализ. – М.: Статистика, 1977.

2. *Айвазян С.А., Енюков И.С., Мешалкин Л.Д.* Прикладная статистика. – М.: Финансы и статистика, 1983.

3. *Айвазян С.А., Мхитарян В.С.* Теория вероятностей и прикладная статистика. – Т. I. – М.: ЮНИТИ, 2001.

4. *Аллен Р.* Экономические индексы. – М.: Статистика, 1980.

5. *Андерсон Т.* Статистический анализ временных рядов: Пер. с англ. – М.: Мир, 1983.

6. *Боярский А.Я.* Теоретические исследования по статистике: Сб. науч. тр. – М.: Статистика, 1974.

7. *Венецкий И.Г., Венецкая В.И.* Основные математико-статистические понятия и формулы в экономическом анализе. – М.: Статистика, 1974.

8. *Джини К.* Средние величины: Пер. с итал. – М.: Статистика, 1970.

9. *Дружинин Н.К.* Выборочный метод и его применение в социально-экономических исследованиях. – М.: Статистика, 1968.

10. *Дубров А.М., Мхитарян В.С., Трошин Л.И.* Многомерные статистические методы. – М.: Финансы и статистика, 1998.

11. *Елисеева И.И., Рукавишников В.О.* Группировка, корреляция, распознавание образов. – М.: Статистика, 1977.

12. *Елисеева И.И., Юзбашев М.М.* Общая теория статистики. – М.: Финансы и статистика, 2004.

13. *Ефимова М.Р., Петрова Е.В., Румянцев В.Н.* Общая теория статистики: Учебник. – 2-е изд., испр. и доп. – М.: ИНФРА-М, 2004.

14. *Йейтс Ф.* Выборочный метод в переписях и обследованиях: Пер. с англ. – М.: Статистика, 1965.

15. *Лукомский Я.И.* Теория корреляции и ее применение к анализу производства. – М.: Госстатиздат, 1961.

16. *Макарова Н.В., Трофимец В.Я.* Статистика в Excel. – М.: Финансы и статистика, 2002.

17. Методологические положения по статистике. – Вып. 3/Госкомстат России. – М., 2000.

18. Теория статистики/Под ред. проф. Р.А. Шмойловой. – М.: Финансы и статистика, 2003.

19. *Фишер И.* Построение индексов: Пер. с англ. – М.: Изд-во ЦСУ СССР, 1928.

20. *Четыркин Е.М.* Статистические методы прогнозирования. – М.: Финансы и статистика, 1983.

21. Эконометрика: Учебник/Под ред. чл.-кор. РАН И.И. Елисеевой. – М.: Финансы и статистика, 2002.

ПРЕДМЕТНЫЙ УКАЗАТЕЛЬ

ОГЛАВЛЕНИЕ

Громыко Галина Леонтьевна
Воробьев Александр Николаевич
Иванов Юрий Николаевич
Казаринова Светлана Евгеньевна
Карасева Лариса Алексеевна
Мамий Ирина Петровна
Матюхина Ирина Николаевна

Учебное издание

ТЕОРИЯ СТАТИСТИКИ

Учебник

Редактор *И.В. Мартынова*
Корректор *М.В. Литвинова*
Оригинал-макет подготовлен
в Издательском Доме «ИНФРА-М»

Художественное оформление серии выполнено
Издательством Московского университета и издательством «Проспект»
по заказу Московского университета

ЛР № 070824 от 21.01.93 г.

Сдано в набор 10.06.2004. Подписано в печать 23.09.2004.
Формат 60×90/16. Бумага офсетная. Гарнитура «Newton».
Печать офсетная. Усл. печ. л. 30,0. Уч.-изд. л. 29,16.
Тираж 5000 экз. Заказ 4839.
Цена свободная.

Издательский Дом «ИНФРА-М»,
127214, Москва, Дмитровское ш., 107.
Тел.: (095) 485-71-77. Факс: (095) 485-53-18.
E-mail: books @ infra-m.ru
http://www.infra-m.ru

ОАО "Тверской полиграфический комбинат"
170024, г. Тверь, пр-т Ленина, 5. Телефон: (0822) 44-42-15
Интернет/Home page - www.tverpk.ru Электронная почта (E-mail) -sales@ tverpk.ru

ISBN 5-16-002158-2