КОЛИЧЕСТВЕННЫЕ МЕТОДЫ В ИСТОРИЧЕСКИХ ИССЛЕДОВАНИЯХ

ББК 63.2
К60

Авторы:
Л. И. Бородкин, И. М. Гарскова, Т. Ф. Изместьева, И. Д. Ковальченко, Л. В. Милов, И. М. Промахина, Н. Б. Селунская

Редакционная коллегия:
И. Д. Ковальченко (ответственный редактор), И. М. Гарскова, Н. Б. Селунская

Рецензенты:
кафедра всеобщей истории Тартуского государственного университета (зав. кафедрой проф. *Х. М. Лиги*); ст. научный сотрудник Института истории СССР АН СССР *Б. М. Клосс*

К60 **Количественные** методы в исторических исследованиях: Учеб. пособие для студ. вузов, обуч-ся по спец. «История» / Гарскова И. М. Изместьева Т. Ф. Милов Л. В. и др. Под ред. И. Д. Ковальченко.— М.: Высш. шк., 1984. — 384 с., ил.

В пер.: 1 р. 10 к.

Пособие состоит из трех разделов. В первом характеризуются общие проблемы применения количественных методов в исторических исследованиях; во втором излагаются основы математико-статистического анализа; в третьем подводятся итоги применения количественных методов в исторических исследованиях в советской и зарубежной историографии и определяются перспективы их использования. В пособии впервые рассматриваются вопросы применения количественных методов для анализа исторического материала.

К $\frac{0501000000-201}{001(01)-84}$ 31—84 ББК 63.2 902.9

Ирина Михайловна Гарскова, Тамара Федоровна Изместьева, Леонид Васильевич Малов и др.

КОЛИЧЕСТВЕННЫЕ МЕТОДЫ В ИСТОРИЧЕСКИХ ИССЛЕДОВАНИЯХ

Заведующая редакцией Т. Г. Липкина. Редактор Н. В. Павлова. Младшие редакторы О. Т. Ускова и Г. М. Фирсова. Художник В. И. Мешалкин. Художественный редактор Т. А. Коленкова. Технический редактор З. А. Муслимова. Корректор Р. К. Косинова.

ИБ № 4196

Изд. № ИСТ — 316. Сдано в набор 28.11.83. Подп. в печать 23.02.84. А—06021. Формат 84×108¹/₃₂. Бум. тип. № 2. Гарнитура литературная. Печать высокая. Объем 20,16 усл. печ. л. 20,16 усл. кр.-отт. 22,13 уч.-изд. л. Тираж 7000 экз. Зак. № 2060. Цена 1 руб. 10 коп.

Издательство «Высшая школа», 101430, Москва, ГСП-4, Неглинная ул., д. 29/14

Московская типография № 8 Союзполиграфпрома при Государственном комитете СССР по делам издательств, полиграфии и книжной торговли. 101898, Москва, Центр, Хохловский пер., 7.

© Издательство «Высшая школа», 1984

ВВЕДЕНИЕ

Отличительной чертой развития науки в эпоху научно-технической революции является ее все углубляющаяся математизация и машинизация. Сейчас уже нет таких областей науки, в которые в той или иной мере не вторглись математические методы, электронные вычислительные машины и другая техника. Безусловно, этому во многом содействуют успехи в развитии прикладной математики и электронной, прежде всего вычислительной, техники. Даже первые поколения ЭВМ позволяли выполнять такой объем вычислений, который невозможен в условиях применения обычной, «малой» счетной техники. Новейшие же ЭВМ обладают прямо-таки безграничными памятью и быстродействием. Это наряду с разработкой целого ряда новых методов вычислительной математики открыло возможность постановки и решения новых исследовательских задач.

Однако было бы ошибочно приписывать интенсивный процесс математизации науки лишь успехам в развитии математики и ЭВМ, как это иногда делают. Основой этого процесса являются внутренние тенденции, присущие развитию самой науки. Именно эти тенденции и вообще потребности человеческой практики, составной частью которой является наука, стимулируют развитие и самой математики, и электронной техники.

Прежде всего следует указать на то, что в развитии науки всегда имела место тенденция к выявлению количественных характеристик изучаемых явлений и процессов объективного мира природы и общества, к количественной оценке черт, свойств и сущности этих явлений, а следовательно, и к применению тех или иных математических методов обработки и анализа конкретных данных. Объективной основой этой тенденции является то, что повсюду в природе и обществе имеет место принцип единства количества и качества. Этот принцип не имеет исключений. В объективном мире не существует неких «чистых» качеств или количеств, не связанных с качеством. Они всегда находятся в единстве. Следовательно, сущность того или иного явления, которая и выражает его качественную определенность, будет раскрыта в полной мере только тогда, когда будет выявлена количественная мера данного качества. Именно поэтому К. Маркс считал, как отмечал П. Лафарг, что «наука только тогда достигает совершенства, когда ей удается пользоваться математикой»[1]. Это важное положение относится ко всем наукам.

Понятно, что стремление к математизации усиливалось по мере развития науки, углубления исследований и усложнения решаемых задач. К середине XX в. развитие науки достигло больших успехов, а во второй половине XX в. в ней произошли и происходят качественные внутренние сдвиги, которые и являются главной причиной все более интенсивного проникновения в науку математических и машинных методов исследования.

Один из этих сдвигов состоит в том, что теперь во всех областях науки исследователям, с одной стороны, приходится оперировать огромным объемом уже накопленных знаний, а с другой — привлекать все больше новых фактических данных, чтобы продвинуться вперед в изучении тех или иных явлений и процессов. Науковеды полагают, что в XX в. человечество накопило вдвое больше знаний, чем за всю свою предшествующую историю, а в ближайшие примерно 15 лет объем современных знаний будет удвоен. Все это выдвигает сложные задачи не только в области систематизации, хранения и использования накопленных знаний, но и в сфере проведения новых научных исследований. Решение этих задач невозможно без широкого применения количественных и машинных методов выявления, обработки и анализа конкретно-научных данных.

Еще более важной чертой в развитии современной науки является все более явственное тяготение к интегральному, системному, т. е. целостному, рассмотрению изучаемых объектов, явлений и процессов. Развитие науки на всех ее этапах характеризуется наличием двух тенденций: с одной стороны, к детальному и углубленному, т. е. дифференцированному изучению отдельных аспектов, сторон и черт явлений и процессов, а с другой — к их обобщенному, целостному освещению. Но в те или иные периоды та или иная из этих тенденций имеет преобладающее значение. Системный, целостный подход — отличительная черта современной науки.

Суть этого подхода в том, что исследуемый объект, явление или процесс рассматривается как некая целостная сис-

[1] Воспоминания о К. Марксе и Ф. Энгельсе. М., 1956, с. 66.

тема, обладающая соответствующей структурой, которая характеризуется тем или иным строением, т. е. совокупностью элементов, а также определенным типом взаимосвязи этих элементов и присущих им черт и свойств. Главная задача исследования состоит в раскрытии этого строения и взаимосвязей и в выявлении их качественного своеобразия. Для решения подобных задач применяются различные методы, получившие название системно-структурного анализа. Конечным итогом такого анализа является построение модели исследуемого явления. Системный анализ требует многостороннего подхода к объекту, учета различных его характеристик и их сопоставления, что в наиболее полной мере может быть достигнуто при обращении к количественным показателям и их математической обработке.

Системный подход и структурный анализ неизбежно приводят к комплексному, или, как его еще иначе называют, междисциплинарному исследованию, т. е. к применению принципов и методов различных наук. Сведение их воедино и получение общего итога комплексного исследования также оказывается наиболее эффективным при использовании количественных данных и математических методов их обработки.

Таким образом, все более широкое распространение в современной науке количественных методов прежде всего диктуется внутренними потребностями ее развития. Эти потребности в полной мере проявляются и в исторической науке.

В исторической науке в последние десятилетия также произошел информационный взрыв. Быстро растет объем конкретного материала. В нашей стране ежегодно издается по истории, археологии и этнографии до 20 тыс. работ.

В связи с необходимостью привлечения все большего объема фактических данных перед историками стоит задача как вовлечения в научный оборот новых источников, так и повышения информативной отдачи всех видов источников.

Историки все более широко обращаются к массовым источникам, характеризующим массовые явления и процессы исторического развития и очень часто содержащим разнообразные сведения о них. Круг массовых исторических источников весьма разнообразен, а объем содержащихся в них конкретных сведений огромен. Особенно много таких источников по новой и новейшей истории. К их числу относятся различного рода статистические материалы и обследования, делопроизводственная документация центральных и местных органов управления и различных учреждений, документы личного учета, справочные материалы, данные конкретно-социологических обследований и многие другие [1]. Они позволяют развернуто и углубленно изучать историко-демографические процессы, социально-экономическое развитие, социальную структуру, политику, культуру, общественное самосознание. При этом такое изучение может вестись на микро- и макроуровнях, т. е. начиная с первичных ячеек и кончая всей системой данных отношений в масштабах обширных регионов или целой страны, а также охватывая различные периоды времени. Так, в последние годы советские историки все более широко привлекают массовые источники для изучения важных явлений и процессов исторического развития нашей страны. К их числу относятся писцовые книги XVII в. — эти своеобразные средневековые земельные кадастры, охватившие многие тысячи поместий и вотчин и позволяющие изучать земельные отношения и феодальную ренту, т. е. важнейшие атрибуты феодально-крепостнической системы. Для конца XVIII в. подобную же роль играют «Экономические примечания» к Генеральному межеванию, в которых по нескольким десяткам губерний европейской России описаны все поселения с указанием численности населения, земельных угодий, форм эксплуатации крестьян, неземледельческих занятий и другие сведения. Важным источником для изучения разложения феодальных и генезиса капиталистических отношений во второй половине XVIII — первой половине XIX в. являются делопроизводственные и статистические материалы вотчинных фондов и особенно подворные описи крестьянских хозяйств.

Весьма обширен комплекс массовых исторических источников по истории России эпохи капитализма. В более чем 100 тыс. уставных грамот и выкупных актов, оформлявших условия выхода крестьян из крепостной зависимости, содержатся разнообразные данные о помещичьем и крестьянском хозяйстве. В эпоху капитализма сложилась определенная система государственного статистического учета и был накоплен огромный статистический материал. Наиболее ценными здесь являются обширные сведения земельных (1877, 1887, 1905 гг.), сельскохозяйственных (1916 и

---

[1] Обзор массовых источников по социально-экономической истории нашей страны нового и новейшего времени дан в работах: Массовые источники по социально-экономической истории России периода капитализма. М., 1979; Массовые источники по социально-экономической истории советского общества. М., 1979.

1917 гг.), промышленных (1900, 1908, 1912 гг.) и других переписей и обследований.

Уникальный комплекс статистических данных, составляющих только в сводном виде несколько тысяч томов публикаций, был собран органами местного самоуправления — земствами. Они незаменимы при изучении крестьянского и помещичьего хозяйства, буржуазной аграрной эволюции, положения крестьян, социальной структуры деревни и других вопросов. Этот комплекс давно и широко используется историками.

В последнее время было обращено особое внимание на делопроизводственную документацию акционерных компаний, отдельных предприятий и банков. В частности, очень ценные сведения содержат описания помещичьих имений, заложенных в земельных банках, и крестьянских хозяйств, обращавшихся за ссудами в эти банки. Эти сведения имеются по многим десяткам тысяч помещичьих и сотням тысяч крестьянских хозяйств. Из делопроизводственных материалов промышленных предприятий обширные сведения о рабочем классе, его составе, источниках пополнения, положении и вообще социальном облике дают паспортные и расчетные книги рабочих и другие документы личного учета.

Разнообразные данные о состоянии крестьянского хозяйства и положении крестьян, о классовой борьбе в деревне, о роли общины в жизни деревни, об идеологии и социальной психологии крестьян содержат такие массовые документы делопроизводства низших и местных органов управления крестьянством, как мирские приговоры, договоры и сделки, жалобы и прошения крестьян.

Еще более обширен комплекс массовых источников по истории советского общества. Для изучения основных тенденций, закономерностей и особенностей общественного развития в советский период незаменимыми являются материалы как текущей статистики, так и всякого рода единовременных переписей и обследований, которые стали регулярно проводиться сразу же после Октябрьской социалистической революции. Для изучения численности, социального состава, движения населения, его культурного уровня и других вопросов первостепенное значение имеют переписи населения. Основу для освещения хода социально-экономического развития создают наряду с текущей статистикой промышленные, сельскохозяйственные и профессиональные переписи, материалы о планировании экономического и социального развития и о ходе выполнения планов. Анализ хода социально-экономического и вообще общественного развития в его исходном, а следовательно, и наиболее массовом звене, т. е. на уровне отдельных предприятий, совхозов, колхозов, учреждений науки и культуры, требует привлечения текущих делопроизводственных материалов. Среди них особое значение имеют годовые отчеты, синтезирующие итоги этой деятельности, отраженные в других документах. Первичные делопроизводственные документы предприятий и учреждений насчитывают десятки и сотни тысяч единиц.

Для изучения социальной структуры советского общества, характеристики общего облика, профессионального и культурного развития рабочего класса и интеллигенции, трудовых коллективов в различных областях производства, управления, науки и культуры ценнейшие сведения имеют листки учета кадров и другие документы личного учета. Для анализа состава выборных партийных, советских и общественных органов и выявления общественно-политической активности различных классов и социальных слоев важные данные содержат анкеты делегатов съездов Советов, партийных конференций и съездов и других выборных форумов.

Указанный перечень далеко не исчерпывает даже основных видов массовых исторических источников, которые имеются в распоряжении историков, занимающихся изучением отечественной истории. Со многими видами массовых источников имеют дело историки и при изучении исторического развития зарубежных стран.

Массовые исторические источники по своей внутренней природе не только допускают, но и требуют применения количественных и машинных методов обработки и анализа содержащихся в них данных. Собственно только посредством этих методов такие источники и можно достаточно широко вовлечь в научный оборот и эффективно использовать для изучения многих явлений и процессов исторического развития. Следует также иметь в виду, что с интенсивным развитием автоматизированных систем управления (АСУ) в различных областях производства и других сферах общественной жизни с самого начала все данные об их функционировании кодируются и шифруются, фиксируются в виде определенных цифровых или иных систем и предназначаются для машинной обработки. Понятно, что такие данные обычными, «ручными» методами нельзя не только обработать, но даже и прочитать, т. е. узнать их содержание. В скором времени эти формализованные материалы будут занимать важное место среди источников по новейшей истории.

Значительную роль количественные и машинные методы играют при решении задач, связанных с повышением информативной отдачи источников. Эта проблема стояла перед историками всегда. Но ее острота возрастает в связи с углублением исторических исследований, постановкой новых задач и ростом объема необходимых конкретно-исторических данных. При изучении сколько-нибудь сложных исторических явлений и процессов историк, даже при наличии разнообразных источников, как правило, испытывает недостаток тех или иных необходимых данных. Такое положение объясняется следующим. Современники, творцы источников собирают сведения об общественной жизни, преследуя определенные цели. Этим прежде всего и определяется отбор фиксируемых данных. Историки же, изучая прошлое на основе этих источников, ставят иные цели. Поэтому они часто и не находят в источниках прямых сведений об интересующих сторонах и чертах изучаемых явлений. Это расхождение усугубляется по мере развития науки, которое ведет к постановке все новых и новых исследовательских задач. Но указанное противоречие между тем, что непосредственно, явно отражено источником, и тем, что необходимо историку, отнюдь не является неразрешимым. Наряду с непосредственно выраженной источники содержат еще и информацию скрытую. Она характеризует многообразные взаимосвязи, присущие явлениям исторической действительности. Эти взаимосвязи не выражены непосредственно. Поэтому и информация о них называется скрытой. Для ее выявления необходимы специальная обработка и анализ данных, непосредственно представленных в источниках. Особенно много скрытой информации содержат массовые исторические источники. При выявлении скрытой информации путем анализа взаимосвязей большую помощь историкам могут оказать количественные и машинные методы обработки исходных данных источников.

Но и в исторической науке главным фактором, диктующим необходимость обращения к количественным и машинным методам, является потребность в более глубоком и точном раскрытии сущности исторических явлений и процессов. Историков в этой связи все чаще не удовлетворяют примерные оценки тех или иных черт этих явлений и часто гипотетические предположения об их сущности, основанные на описательном анализе. Если, например, на начальных стадиях изучения тех или иных явлений при характеристике свойственных им изменений можно ограничиться такими оценками, как «интенсивные», «быстрые», «замедленные» и т. д., то при углубленном анализе этих изменений подобные оценки оказываются недостаточными и возникает потребность в более строгом и точном их выражении, т. е. в переходе к выявлению их количественной меры. Иначе говоря, в результате развития исторической науки все более отчетливо проявляется необходимость в раскрытии количественной меры соответствующих качеств. При этом такая необходимость обнаруживается не только по отношению к явлениям и процессам, характеризуемым массовыми историческими источниками, которые либо непосредственно содержат количественные данные, либо включают сведения, легко поддающиеся количественному выражению, но и по отношению к сугубо индивидуальным явлениям и описательным источникам, таким, например, как литературные, публицистические и другие произведения.

Системный подход к историческим явлениям и процессам прежде всего требует раскрытия внутренней структуры соответствующих систем, выявления механизма закономерностей и особенностей их функционирования. Для успешного решения этих задач необходимо применение методов структурного анализа, начиная с простейшей количественной обработки данных и кончая построением сложных моделей этих систем, основанных на применении различных математических методов.

Таким образом, потребность применения количественных методов в исторических исследованиях (или их квантификация, как иначе называют этот процесс) диктуется прежде всего присущими самой исторической науке тенденциями развития.

С рубежа 50—60-х годов XX в. началось все более расширяющееся применение историками разных стран количественных и машинных методов обработки и анализа конкретно-исторических данных. В настоящее время в ряде областей исторических исследований, прежде всего в таких, как историко-демографические, социально-экономические и историко-социологические, применение количественных методов и ЭВМ получило широкое распространение. В этой связи естественно возникла потребность овладения историками методами количественного анализа и машинной обработки данных. Появилась и необходимость в создании специальных пособий по применению количественных методов в исторических исследованиях. Некоторые такие работы опубликованы как у нас в стране, так и за рубежом. Характер этих работ различен. Одни из них[1] представляют собой не учебные пособия, а популярные работы, в которых показывается эффективность применения количественных методов в исторической науке и характеризуется тип задач, при решении которых целесообразно применение этих методов. Такие работы весьма полезны, но они рассчитаны на читателя, который уже имеет определенные подготовку и опыт применения количественных методов. Они не ставят своей задачей помочь в непосредственном овладении методикой и техникой применения тех или иных математико-статистических методов обработки и анализа количественных данных.

Другой тип работ по своему характеру ближе к учебным пособиям. В них основное внимание уделяется рассмотрению методов математико-статистической обработки и анализа исторических данных с раскрытием условий, методики и техники их использования. Несколько таких работ издано в США и Англии[2]. Однако в них не уделено должного внимания общим, теоретико-методологическим проблемам применения количественных методов в исторических исследованиях. Между тем, именно теоретико-методологические посылки и принципы, на основе которых строится количественный анализ в исторических, как, впрочем, и во всех других исследованиях, определяют правомерность и эффективность его применения при решении тех или иных задач. Основные расхождения между историками марксистами и немарксистами в применении количественных методов, как и вообще в исторических исследованиях, идут именно по линии теоретико-методологической. Очевидно, что пособие по применению количественных методов в исторических исследованиях не должно ограничиваться методическими и техническими сторонами этого применения, но и включать характеристику основных методологических проблем.

Далее, при определении характера пособия учитывался опыт преподавания студентам исторического факультета Московского университета курса количественных методов.

---

[1] См., например: *Миронов Б. Н., Степанов З. В.* Историк и математика. (Математические методы в исторических исследованиях). Л., 1975.

[2] Характеристику этих работ см.: *Гарскова И. М.* Количественные методы и ЭВМ для историка. (Обзор англо-американских изданий).— В сб.: «Математические методы в социально-экономических и археологических исследованиях». М., 1981.

Этот курс включает две самостоятельные части. Вначале читается курс «Количественные методы в исторических исследованиях». В нем характеризуются методы исторического исследования и место в них количественных методов, освещаются методологические проблемы применения количественных методов и показываются основные итоги и перспективы применения количественных методов и ЭВМ в отечественных и зарубежных исследованиях по истории. Затем преподается (в форме лекций и практических занятий) курс «Основы математико-статистического анализа», который ведут математики и который имеет целью прежде всего ознакомление с основными разделами математической статистики. В итоге изучения указанных курсов студент должен овладеть тем минимумом знаний, который позволяет квалифицированно изучать исторические исследования, в которых применены количественные методы, а также необходим для того, чтобы начать самостоятельную работу по применению тех или иных из этих методов.

Исходя из сказанного была избрана следующая структура настоящего учебного пособия. В первой его части рассматриваются основные методологические проблемы применения количественных методов в исторической науке — место количественных методов в исторических исследованиях, вопросы измерения исторических явлений и выявления репрезентативности количественных данных, методологические аспекты моделирования исторических явлений и процессов и некоторые другие. Именно в понимании прежде всего этих проблем, как указывалось, и расходятся марксистский и немарксистский подходы к применению количественных методов в исторических исследованиях. В рамках учебного пособия нет возможности, да и необходимости, последовательно характеризовать различия в этом подходе. Главным здесь является изложение тех принципов применения количественных методов, которые вытекают из материалистически-диалектической, марксистско-ленинской теории и методологии исторического познания.

Во второй части работы характеризуются основные методы математико-статистического анализа. Круг их в основном совпадает с тем, что включают в себя существующие учебные пособия по математической статистике. При освещении этих методов авторы акцентировали внимание на двух моментах. Во-первых, на раскрытии логически-содержательной сути рассматриваемых методов, что необходимо для их правильного применения как в части выбора тех задач, для решения которых эти методы можно использовать, так и при интерпретации результатов, полученных при обработке данных соответствующими методами.

Во-вторых, это показ методики применения этих методов и техники вычислений различных показателей и коэффициентов на конкретных исторических примерах. Понятно, что техника вычислений могла быть продемонстрирована лишь применительно к тем методам математико-статистического анализа, которые не требуют сложных и громоздких расчетов. Когда подобные расчеты невозможны в ручном исполнении или занимают много времени, обращаются к помощи ЭВМ, на которых соответствующие расчеты выполняются по стандартным программам. В пособии нет математических доказательств. Их можно при желании найти в соответствующей литературе. Для понимания логической сути характеризуемых методов и проведения соответствующих расчетов достаточно знания математики в объеме средней школы.

В третьей части работы освещаются основные направления применения количественных методов в отечественных исследованиях по истории. Цель этого раздела состоит в том, чтобы на имеющемся опыте применения количественных методов при изучении различного рода исторических явлений и процессов, с одной стороны, показать их эффективность, а с другой — на конкретных проблемах раскрыть технику применения этих методов, начиная с постановки задачи и кончая интерпретацией результатов количественного анализа и формулированием выводов. Тем самым наглядно будет показана ведущая роль методологических принципов при применении количественных методов, сущность системного, целостного подхода к изучаемым явлениям и основные этапы, методика и техника системно-структурного анализа. Понятно, что все это может быть проиллюстрировано на небольшом числе примеров.

Таковы основные цели и структура настоящего пособия. Оно подготовлено сотрудниками кафедры источниковедения и историографии истории СССР исторического факультета Московского государственного университета, историками и математиками, которые разработали указанные выше курсы по количественным методам в исторических исследованиях и обеспечивают их преподавание. Поскольку подобное учебное пособие издается впервые, его авторы и редакторы будут признательны за все отклики и пожелания, которые несомненно будут учтены при дальнейшей работе над пособием.

1

МЕТОДОЛОГИЧЕСКИЕ ПРОБЛЕМЫ ПРИМЕНЕНИЯ КОЛИЧЕСТВЕННЫХ МЕТОДОВ В ИСТОРИЧЕСКОЙ НАУКЕ

ГЛАВА 1
МЕСТО КОЛИЧЕСТВЕННЫХ МЕТОДОВ В ИСТОРИЧЕСКИХ ИССЛЕДОВАНИЯХ

§ 1. Методы научного исследования. Их виды и содержание

Всякое научное познание, в том числе и историческое, имеет свой объект и предмет. Объектом исторического познания является вся многообразная совокупность проявлений общественной жизни с древнейших времен до современности. Однако на каждом этапе развития науки подвергаются изучению лишь определенные стороны или явления общественной жизни. Они и составляют предмет познания. При конкретно-историческом изучении явлений и процессов прошлого историк ставит определенную исследовательскую задачу, т. е. стремится раскрыть те или иные черты и свойства, присущие этим явлениям и процессам, выявить закономерности и особенности их функционирования, показать их связь с другими явлениями и процессами и их место и роль в общественном развитии и т. д.

Для проведения исследования и решения поставленной задачи историк, как и ученый в любой другой области науки, должен, во-первых, опираться на определенные представления, т. е. исходить из совокупности посылок и идей, которые позволяют раскрыть сущность исследуемых явлений и процессов. Совокупность этих посылок и идей составляет **теорию научного познания.** Теорией научного познания, которая дает возможность наиболее объективно и глубоко объяснить ход исторического развития, является, как известно, **исторический материализм.** На нем и базируются исследования историков-марксистов. Буржуазные теории исторического познания представляют собой разновидности идеализма или вульгарного материализма.

Во-вторых, для проведения любого исследования ученый должен опираться на определенные принципы, пути и способы познания явлений общественной жизни. Эти принципы, пути и способы, а также разработанные на их основе правила и процедуры исследования и его техника и образуют в совокупности соответствующий **метод исследования.** Иначе говоря, метод представляет собой в широком смысле тот арсенал принципов и средств, с помощью которого ученый получает необходимую информацию об изучаемых явлениях, обрабатывает и анализирует ее. Объяснение полученных в результате этого данных на основе соответствующих теоретических положений и дает в конечном счете новые научные знания об этих явлениях. Совокупность методов научного, в том числе и исторического, познания весьма разнообразна, и они имеют ряд уровней. Это **методы общефилософские, общенаучные, специально-научные** и **конкретно-**, или **проблемно-научные.**

Философские методы научного познания раскрывают общие принципы научного познания и тенденции его развития, способствуют его обогащению новыми общими и конкретными методами. Общефилософским марксистским методом познания является **материалистическая диалектика.** Ее основные принципы, требующие рассмотрения явлений во всей их сложной взаимосвязи, в движении, в единстве и борьбе присущих им противоположностей, являются исходными для всех других марксистских методов научного познания.

Общенаучными являются методы, которые применяются во многих науках. К числу таких методов относятся, например, анализ и синтез, индукция и дедукция, описательный и количественный анализ и др. Они отличаются от философских методов, которые служат для них основой, тем, что охватывают процесс познания не в целом, а лишь его отдельные аспекты и тенденции. Так, анализ и синтез раскрывают принципы исследования конкретных данных; индукция и дедукция показывают основные пути познания сущности явлений; описательные и количественные методы связаны со способами выражения информации об исследуемых объектах.

К **специально-научным методам** относятся такие, которые используются отдельными науками. Например, историко-сравнительный метод, хотя и применяется в ряде наук, но присущ прежде всего исторической науке. Историческими по своей сути являются синхронный и диахронный методы изучения явлений общественной жизни. При первом из них — синхронном — совокупность явлений и объектов рассматривается как определенная вариация в одном и том же временном срезе. При диахронном подходе явления рассматриваются в их историческом изменении и движении.

Каждая наука обладает определенной совокупностью именно ей присущих методов научного исследования. Эти методы базируются на определенных философских и общенаучных методах.

Конкретно-научные или **проблемно-научные** методы рассчитаны на изучение тех или иных отдельных сторон или черт действительности, которая является объектом соответствующей науки. Очевидно, что в исторической науке для раскрытия сути явлений экономических, социальных, политических, идеологических и т. д. необходимы различные методы исследования. Каждая наука наряду со специальными располагает обширным арсеналом конкретно-проблемных методов, рассчитанных на изучение определенных явлений и процессов.

Все указанные уровни и типы методов научного познания тесно взаимосвязаны, ибо в объективной действительности имеет место взаимосвязь всеобщего, особенного и единичного. Всякий метод познания включает в себя следующие три компонента. **Во-первых,** это основные принципы и пути, способы и приемы познания, которые являются теорией метода, или, как говорят иначе, его **методологией. Во-вторых,** для практической реализации этих принципов и способов необходима разработка соответствующих правил и процедур. Их совокупность составляет **методику** данного метода. **В-третьих,** для проведения любого исследования требуются определенные орудия и инструменты в широком их понимании. Они образуют **технику** научного исследования. Поэтому метод исследования представляет собой органическое сочетание методологии, методики и техники. Ведущим из этих компонентов является методология. Она оказывает определяющее влияние на методику и технику. Однако взаимосвязь этих компонентов диалектическая, т. е. идет как сверху вниз — от методологии к методике и технике, так и обратно. Поэтому на развитие и совершенствование методов познания оказывают воздействие все его компоненты. Например, применение новой техники исследования требует совершенствования методики и углубления конкретной методологии, т. е. общих принципов соответствующего метода.

§ 2. Количественный анализ и его место в исторических исследованиях

При научном изучении любых объектов, явлений и процессов, в том числе и исторических, основной целью является раскрытие их внутренней сущности посредством **качественного**, т. е. **содержательного** анализа. Качественный анализ представляет собой совокупность аналитических и синтетических процедур, имеющих своей целью выявление коренных свойств, основных закономерностей и особенностей возникновения и функционирования исследуемых объектов и явлений, что позволяет раскрыть их истинную внутреннюю суть и роль в объективной реальности. Качественный анализ всегда основывается на определенных теоретических посылках. Он играет ведущую роль на всех этапах научного исследования.

На основе качественного анализа происходит выбор объекта и предмета изучения и ставится исследовательская задача. Посредством качественного анализа определяется круг тех признаков и свойств рассматриваемых явлений и процессов, которые должны быть учтены и изучены, чтобы была решена исследовательская задача. На нем базируется выбор необходимых источников и методов исследования, а также выявляется соответствующая система фактов. Обработка и анализ этих фактов, обобщение полученной в результате информации и формулирование выводов и наблюдений также основываются на качественном, содержательном подходе.

Характер самого качественного анализа определяется той теорией и методологией научного познания, из которых он исходит. Наиболее широкий и глубокий, последовательно объективный сущностно-содержательный анализ общественных явлений может быть проведен только на основе материалистической теории и диалектического метода, присущих марксизму.

Качественный анализ, играющий ведущую роль в научных исследованиях, может быть осуществлен при наличии соответствующей информации, т. е. сведений об используемых объектах и явлениях действительности.

Поскольку объектом исторического познания является прошлое, которое невозможно наблюдать непосредственно, информация о нем извлекается из исторических источников. К ним относится вся совокупность вещественных, письменных, изобразительных и фонических памятников и материалов, в которых так или иначе отражено прошлое.

Существует два способа выражения информации об объективной реальности — **описательный** и **количественный.** Сами по себе описательный и количественный методы могут лишь охарактеризовать те или иные видимые черты и свойства исследуемых объектов, но не их внутреннюю, чаще всего скрытую суть. Эта суть на основе информации, представленной в описательном или количественном виде, раскрывается в результате сущностно-содержательного, качественного анализа.

Таким образом, описательный и количественный анализ характеризуют явление, а качественный анализ — сущность. В этом состоит их функциональное различие в процессе научного познания. Поскольку в действительности явление и сущность органически взаимосвязаны, постольку и в научном исследовании неразрывно взаимосвязаны описательный и количественный анализ, с одной стороны, и качественный — с другой. Фактически в научном исследовании имеет место либо **сущностно-описательный,** либо **сущностно-количественный** анализ. Никакого чистого качественного или самостоятельного описательного, или количественного анализа не существует. Может быть лишь их единство в указанном сочетании. В этой связи необходимо обратить внимание на необоснованность широко распространенного толкования качественного и количественного анализа как альтернативных методов исследования. В действительности же количественный анализ можно рассматривать как альтернативу не качественному, а описательному анализу. Между ними, действительно, имеют место существенные различия, к характеристике которых мы и переходим.

* *
*

Описательные методы являются вполне правомерной и наиболее широко распространенной формой исторических исследований. Это имеет свои объективные основания. Во-первых, подавляющая часть сохранившихся письменных источников имеет описательный характер. Поэтому естественно, что сведения этих источников используются историками в той форме, в какой они зафиксированы их творцами. Сведения вещественных и изобразительных источников, чтобы быть использованными в исследовании, также должны быть переведены в языковую, т. е. описательную, форму.

Во-вторых, многие явления общественной жизни на данной стадии развития науки либо не поддаются измерению вообще, либо это измерение может быть весьма приближенным и даже условным. Подробнее об этом будет сказано при рассмотрении проблем измерения исторических явлений и процессов.

Описательный анализ может использоваться при изучении любых проявлений общественной жизни. При наличии в источниках необходимых для решения исследовательской задачи сведений описательный анализ позволяет получить представительную и доказательную систему фактов. Наиболее эффективным и незаменимым описательный анализ является при освещении всякого рода индивидуальных явлений общественной жизни, ибо здесь имеет существенное значение учет деталей и нюансов, а также сложных, интегральных по своей сути явлений. Например, весьма затруднительно, точнее говоря, невозможно, выразить количественно такое сложное явление, как образ жизни.

Описательные методы позволяют выделить многие существенные черты исторических явлений и процессов, раскрыть их качественное своеобразие и причинную обусловленность, показать типичные и специфические формы их проявления и т. д.

Плюсом описательных методов является возможность последовательного и детального отражения явлений общественной жизни, яркой характеристики исторической эпохи, присущих ей настроений, убеждений и устремлений, а также массовых и индивидуальных исторических персонажей.

К достоинствам описательного анализа следует отнести и его техническую простоту. Доступность многих описательных по своей форме исторических источников (прежде всего таких, как мемуарная литература, периодическая печать, научные, публицистические и литературные произведения) и несложность простого воспроизведения их содержания порождают иллюзию, будто бы историческим исследованием может заниматься любой образованный человек. Причиной подобного заблуждения является непонимание или забвение того, что в исторической науке, как и во всех других науках, кроме описания явлений необходим еще анализ, раскрывающий их внутреннюю суть.

Однако определенная универсальность, широкий диапазон применения, незаменимость и высокая эффективность описательного анализа при изучении многих явлений общественной жизни не исключает и его ограниченности. Описательный анализ наряду с плюсами имеет и свои минусы.

Во-первых, описательные методы малопригодны для обобщенной характеристики исследуемых объектов, что при изучении массовых явлений и процессов может приводить либо к иллюстративности, т. е. к сосредоточению внимания лишь на отдельных из этих объектов, либо к фактографизму и детализации, т. е. к трудно реализуемому стремлению максимально детально описать как можно большее число их. В первом случае это может породить бездоказательность наблюдений и выводов, а во втором — смешение главного и второстепенно, существенного и несущественного, типичного и случайного.

Во-вторых, позволяя установить те или иные признаки и свойства рассматриваемых объектов, описательный анализ не показывает их абсолютной и относительной меры. Например, историк, изучая процесс разложения крестьянства в нашей стране в эпоху капитализма на основе тех или иных описательных данных, может, скажем, узнать о наличии в среде крестьянства качественно отличных по своему хозяйственному состоянию и социальному облику групп (сельские полупролетарии и пролетарии, середняки, мелкие капиталисты). Но на основе таких указаний нельзя определить численность и удельный вес этих групп.

В-третьих, описательный анализ, допуская возможность выявления взаимосвязей тех или иных явлений или черт и свойств определенных объектов, не измеряет силы воздействия одних факторов на другие. Применительно к тому же крестьянству описательные сведения могут указывать на большую доходность более крупных крестьянских хозяйств сравнительно с более мелкими. Но таких сведений недостаточно для выявления степени зависимости доходности крестьянского хозяйства от размеров его трудовых, земельных или иных ресурсов.

В силу указанных причин описательные методы в тех случаях, когда необходимо оценить масштабы, удельный вес и уровень развития, а также степень сходства и различия, тесноту и силу взаимосвязи и взаимодействия тех или иных объектов, явлений и процессов или их черт и свойств, обрекают исследователей на неточные, весьма приблизительные и даже неопределенные оценки типа «больше — меньше», «сильно-слабо», «существенно-несущественно» и т. д. В общем описательные методы не позволяют выявить количественную меру соответствующих качеств. Таковы сильные и слабые стороны описательного анализа.

Объективной основой для все более широкого распространения количественных методов в исторических исследованиях является, как уже отмечалось, потребность в установлении количественной меры явлений и процессов общественной жизни. Только в том случае, когда удается установить количественную меру соответствующего качества, исследование достигает наибольшей глубины.

Сущность количественного анализа состоит не в том, что исследователь оперирует определенными количественными, цифровыми показателями, как нередко полагают. Те или иные количественные показатели могут использоваться и при описательном анализе. Количественный анализ — это выявление и формирование системы численных характеристик изучаемых объектов, явлений и процессов, которые, будучи подвергнуты определенной математической обработке, создают основу для раскрытия количественной меры соответствующего качества.

Количественный анализ позволяет установить абсолютную и относительную меру рассматриваемых черт и свойств объектов и явлений и выявить степень или силу их проявления.

Теоретически количественные методы могут быть применены при изучении любых явлений и процессов объективной реальности, в том числе и исторической, так как всякому качеству присуще определенное количество. Однако практически такое применение возможно лишь постольку, поскольку удается получить количественные характеристики исследуемых явлений. Сами возможности измерения этих явлений определяются их природой и уровнем развития науки и техники и расширяются по мере прогресса последних.

Количественный анализ, как и описательный, имеет сферы, в которых без него просто невозможно обойтись и где он может дать наибольший эффект. К таким сферам в историческом развитии относятся прежде всего всякого рода массовые явления и процессы. Они представляют собой различные экономические, демографические, социальные, политические, культурные и другие системы с присущими им структурами, включающими большие совокупности объектов, обладающих многообразными свойствами. Эти системы находятся в постоянном изменении и движении. То или иное состояние этих систем или итог происходящих в них изменений складываются из многих состояний и изменений составляющих эти системы объектов, и выступают как соответствующие равнодействующие или определенные тенденции. Для выявления всего этого необходимо применение количественных методов.

Однако массовые явления и процессы представляют собой ту сферу проявлений общественной жизни, изучение которых непременно требует применения количественных методов. Но эти методы могут быть эффективными и при исследовании многих индивидуальных явлений. Такой эффект достигается в тех случаях, когда эти явления изучаются системно, т. е. рассматриваются как определенная совокупность элементов с соответствующими взаимосвязями. Подобными системными свойствами обладают очень многие индивидуальные явления общественной жизни.

Следовательно, практические возможности применения количественных методов в исторических исследованиях очень обширны, а открываемые этим применением перспективы углубления анализа исторических явлений и процессов — весьма существенны.

Наряду с этим и количественные методы анализа исторических явлений и процессов имеют свои минусы. Главный из них состоит в том, что широко применяемое при количественном анализе усреднение данных и отвлечение от частных черт и свойств исследуемых объектов сопряжено с потерей информации и может привести к чрезмерному абстрагированию.

Количественные методы значительно более трудоемки и сложны в применении и понимании, чем методы описательные. Это требует дополнительных усилий от исследователей. Впрочем, они многократно окупаются теми результатами, которые получаются при их применении.

В целом количественные методы являются несомненно более мощными, чем описательные методы анализа исторических явлений и процессов, ибо создают основу для более глубокого раскрытия их внутренней, качественной сущности.

Таким образом, и описательные, и количественные методы исследования исторических явлений и процессов представляют собой вполне правомерные и равнонеобходимые формы исторических исследований. Каждый из этих методов имеет свою сферу, в которой его применение является непременным или наиболее эффективным. Сильная сторона описательных методов — их конкретность и образность, а количественных — глубина и точность. Эти методы не противостоят один другому, а дополняют друг друга. Поэтому неправомерна абсолютизация того или иного из них и их противопоставление. В практике научных исследований наибольший успех достигается тогда, когда оба эти метода гармонично сочетаются. Правда, фактическое положение здесь таково, что описательными методами анализа владеют все историки, а количественными — лишь часть из них. Поэтому задача состоит в том, чтобы все историки овладели основными методами исторического исследования — и описательным, и количественным анализом. В этом залог углубления исторических исследований во всех их аспектах.

Следует подчеркнуть, что и описательные, и количественные методы являются лишь основой для качественного, содержательного анализа и только в сочетании с ним приводят к раскрытию внутренней сущности явлений и процессов исторического развития, т. е. к решению главной задачи исторического, как и всякого другого, исследования.

ГЛАВА 2
ИЗМЕРЕНИЕ ИСТОРИЧЕСКИХ ЯВЛЕНИЙ

§ 1. Общие принципы и методы измерения

Применение количественных методов при анализе явлений объективного мира требует наличия числовых характеристик этих явлений, т. е. их измерения.

Измерение в широком смысле — это способ установления отношений между объектами. Суть этого способа состоит в том, что информация об изучаемых объектах выражается в системе количественных, числовых характеристик.

Объективной основой для измерения является то, что всякое качество имеет определенную количественную меру. Поэтому в принципе все они могут быть измерены. Однако практические возможности такого измерения имеют свои границы, обусловливаемые как чрезмерным многообразием и сложностью явлений и присущих им черт и свойств, так и естественной для каждого этапа развития науки и практики ограниченностью методологических, методических и технических средств измерения. Особенно сложным является измерение общественных явлений.

Сущность, т. е. качественная определенность общественных явлений, проявляется в их свойствах или чертах, отражаемых соответствующими признаками. По форме своего выражения признаки подразделяются на **количественные** и **качественные**, т. е. атрибутивные. Количественные признаки показывают меру определенных свойств, а атрибутивные — лишь наличие этих свойств и их сравнительную интенсивность. Так, например, возраст человека —

это признак количественный, который указывает меру свойства, а пол — признак атрибутивный, который лишь описательно фиксирует сам характер свойства. В свою очередь атрибутивные, описательные признаки могут быть простыми и сложными. **Простые атрибутивные признаки** выражают однозначные, или одноаспектные, свойства объектов, такие, например, как пол, национальность, семейное положение и т. п. **Сложные атрибутивные признаки** характеризуют многоаспектные, интегральные свойства объектов и явлений. Сложные атрибутивные признаки, как правило, представляют собой интегральные оценки. Поэтому их еще называют **признаками-оценками.** Например, успеваемость школьников или студентов, качество той или иной продукции, интенсивность и эффективность классовой борьбы рабочих или крестьян и другие подобные признаки являются сложными атрибутивными признаками-оценками.

Понятно, что возможности и способы измерения количественных и атрибутивных признаков, с одной стороны, и простых и сложных атрибутивных признаков — с другой, являются разными. Как и всякое познание, измерение начинается с сущностно-содержательного, качественного (не путать с качественными, атрибутивными признаками) анализа изучаемых явлений и процессов и определения цели измерения. Каждый социальный объект неисчерпаем по своим особенностям и свойствам. Поэтому необходимо предварительное определение того, какие из этих свойств и черт должны измеряться. Выбор того или иного аспекта измерения определяется исследовательской задачей. Лишь установив цель измерения, можно решать вопрос о методе измерения и его единицах.

В силу того, что признаки исследуемых объектов различаются по своему характеру и потому; что цели измерения могут быть разными, существуют различные принципы и методы измерения. Совокупность тех или иных принципов измерения составляет **шкалу измерений. Шкала измерений — это правила придания рассматриваемому признаку числового значения.** Различные шкалы измерений дают разную по полноте и точности информацию об объекте и приводят к численным показателям, возможность математической обработки и анализа которых неодинакова.

Шкалы измерений. Вся применяемая в современной науке совокупность принципов и способов измерения сводится к четырем шкалам измерения.

Шкала наименований (номинальная). Измерение здесь состоит в классификации объектов по тем или иным признакам. Это наиболее простая форма приписывания чисел объектам. Числа здесь используются для определения принадлежности объекта к определенному классу. Например, население того или иного пункта можно разделить на мужчин и женщин и обозначить первых «1», а вторых «2». Или, среди населения можно выделить группы с образованием «низшим», «неполным средним», «средним», «неполным высшим» и «высшим». Представителей этих групп можно обозначить числовыми значениями — 1, 2, 3, 4, 5. Понятно, что в шкале наименований числа играют роль символов. Поэтому одни числовые значения могут быть заменены любыми другими, либо вообще вместо чисел можно взять иные символы (например, А, Б, В, Г, Д). Суть измерения по шкале наименований состоит в том, что объекты характеризуются тождеством внутри класса и различием между классами.

Шкала порядка. Принцип измерения здесь состоит в ранжировании объектов по тому или иному признаку. Измерение по шкале порядка показывает расположение объектов по интенсивности проявления в них определенного свойства, т. е. ранжирует их по принципу «больше — меньше», но без указания, насколько больше или насколько меньше. Поэтому интервалы между отдельными объектами остаются неизвестными. Так, например, можно расположить забастовки рабочих или волнения крестьян по интенсивности их борьбы. Число при измерении по шкале порядка обозначает место объекта в упорядоченном ряду. Это место можно считать рангом объекта. Числа эти могут быть заменены другими, но так, чтобы не было нарушений последовательности расположения объектов. Так, вместо 1, 2, 3 можно обозначить 11, 12, 13 или 5, 10, 15 и т. д.

Существуют две другие шкалы измерения, которые называются **метрическими** или **количественными.** Процедура измерения здесь состоит в определении числовых значений признаков посредством какой-либо меры или единицы измерения.

Одной из количественных, метрических шкал является **шкала интервалов.** Обладая всеми свойствами порядковой шкалы, шкала интервалов отличается от нее тем, что точно определяет величину интервала в принятых единицах измерения. Поэтому интервальная шкала, хотя также позволяет измерять лишь интенсивность свойств, показывает, насколько различной она была. Для этого выбирается, хотя и произвольно, единица измерения и условное начало отсчета. Примерами интервальных шкал могут служить температурные шкалы по Цельсию, Реомюру, Фаренгейту, а также различные шкалы календарного времени (например, старый юлианский и новый григорианский солнечные календари, лунный календарь и др.). Во всех подобных способах измерения точка отсчета является условной, а интервалы могут быть разными. Переход от одной интервальной шкалы к другой может осуществляться в результате линейных преобразований.

Пропорциональная шкала, или **шкала отношений**, представляет собой метрическую шкалу, по которой ведется счет и измерение **количественных признаков.** Для этой шкалы характерны метрические единицы измерения и начальная нулевая точка отсчета. Поэтому пропорциональная шкала позволяет установить, во сколько раз те или иные свойства одного объекта больше, чем у другого, т. е. выявить их точное количественное соотношение (отсюда и название — шкала отношений). В качестве единиц измерения могут выступать либо меры пространства, времени, площади, объема, веса и т. д., либо единица соответствующего объекта (такое-то количество или число фабрик, заводов, населения, рабочих и т. д.), а измерение осуществляется посредством счета, т. е. является измерением в собственном смысле.

Различные шкалы измерения связаны определенной зависимостью. Каждая последующая шкала включает в себя все предыдущие. Так, порядковая шкала является в то же время и номинальной, интервальная — порядковой и номинальной, пропорциональная — интервальной, порядковой и номинальной.

Измерение по шкале наименований и шкале порядка указывает лишь на различие свойств объектов или на разную интенсивность проявления этих свойств. Численные показатели номинальной и порядковой шкал не допускают арифметической обработки. В этом заключается ограниченность этих методов измерения. На этом основании классификацию и ранжирование иногда не считают измерением. Но поскольку по шкалам наименований и порядка все же можно установить определенные отношения между объектами, то классификация и ранжирование в таком смысле являются измерением.

Показатели измерения по интервальной и пропорциональной шкалам допускают арифметическую обработку, а следовательно, и возможность применения при их анализе других математических методов.

Таким образом, измерение может представлять собой сложную процедуру, в которой применяются различные шкалы и используются и преобразуются различные показатели измерения.

Единицы измерения. Важнейшим моментом в процедуре измерения является выбор его единиц. Единицы измерения, как и само измерение, являются категориями историчными. Они вырабатывались и апробировались в процессе длительного опыта. По мере роста потребности в измерении и его точности расширялась и совершенствовалась система единиц измерения и вырабатывались предъявляемые к ним требования.

Коль скоро всякое измерение имеет своей целью выявление соотношения объектов, их свойств и черт, выражаемых соответствующими признаками, то непременным условием корректности измерения являются **неизменность** и размерная **однородность**, т. е. **одномасштабность** единиц измерения в процессе измерения. Только при этом условии результаты измерения будут сопоставимы в пространстве и времени. От размерной однородности зависит и точность измерения. Например, различия в обеспеченности крестьян хлебом в тех или иных местностях в тот или иной период времени или в одной и той же местности в разное время можно выявить, определяя количество собранного хлеба в расчете на один двор. Но очевидно, что крестьянский двор не является равномасштабной единицей измерения, ибо средняя населенность двора в разных местностях могла быть неодинаковой и изменялась во времени. Поэтому более точные результаты даст учет количества хлеба на душу крестьянского населения, так как эта единица измерения обладает значительно большей размерной однородностью, равномасштабностью во времени и пространстве. Степень необходимой размерной однородности единиц измерения определяется целями измерения и зависит от характера измеряемых признаков.

Наиболее развернутая и стабильная система эталонных единиц измерения сложилась в естественных науках (метр, килограмм, секунда и т. д.). Некоторые из них применяются и при измерении общественных явлений. Но в целом многообразная совокупность проявлений общественной жизни в очень ограниченной степени может быть измерена посредством единиц измерения, применяющихся в естественных науках. Так, Ф. Энгельс отмечал, что «...совершенно невозможно выразить экономические отношения в физических мерах»[1]. В. И. Ленин также подчеркивал, что «нет ничего легче, как наклеить «энергетический» или «биолого-социологический» ярлык на явления вроде кризисов, революций, борьбы классов и т. п., но нет и ничего бесплоднее, схоластичнее, мертвее, чем это занятие»[2].

Для измерения общественных явлений применяются самые различные единицы. Выбор их производится исходя из природы изучаемых явлений и процессов, которая отражается в соответствующих признаках, а также зависит от поставленных задач, т. е. цели измерения, и его целесообразности. В принципе единица измерения должна ориентироваться на тот минимальный элемент или компонент, который лежит в основе структуры измеряемой совокупности. Например, обувь естественно измерять парами, автомобили штуками и т. д. Здесь единица измерения соответствует минимальному компоненту соответствующей совокупности объектов. Но исходя из указанного принципа минимальности единицы измерения, следует учитывать и практическую целесообразность. Так объем добычи различных сырьевых ресурсов (нефть, уголь, руда металлов и т. д.) теоретически можно измерять самыми мелкими единицами веса, ибо исходные элементы их структуры имеют молекулярный уровень. Но практически применение таких мелких единиц нецелесообразно и используют более крупные единицы (как правило, тонны).

Единицы измерения количественных признаков. При измерении количественных признаков широко применяются различные **натуральные** и **метрические** единицы измерения — штук вагонов или самолетов, голов скота, человек населения и т. п., либо тонн, метров и т. д. того-то. Но поскольку, будучи однотипными, измеряемые объекты могут сильно различаться по своим свойствам, постольку натуральные единицы измерения могут быть очень приближенными, а то и вовсе объединять в сущности разнородные объекты. Например, нефть, добываемая на разных промыслах, может отличаться по своему качеству, автомобили или самолеты разных марок неодинаковы по своей мощности и другим данным и т. д. Поэтому измерение добычи нефти только тоннами, а самолетов или автомобилей штуками является весьма приближенным. Чтобы добиться большей точности, применяют либо двойные единицы измерения (например, число товарных вагонов в штуках и в

---

[1] *Маркс К., Энгельс Ф.* Соч. 2-е изд., т. 35, с. 110.
[2] *Ленин В. И.* Полн. собр. соч., т. 18, с. 348.

тоннах грузоподъемности, количество скота в головах и в «живом весе» и т. д.), либо **условно-натуральные,** или **приведенные,** единицы измерения. Такой единицей обычно выступает одна из разновидностей того или иного типа продукции. К этой разновидности через определенные коэффициенты, учитывающие различия свойств, приводятся все ее другие виды. Применяются приведенные меры **массы** (молоко с учетом жирности, сахарная свекла с учетом содержания сахара, масса топлива в переводе на 7000-калорийное и т. д.), **мощности** (трактора в переводе на 15-сильные, электромоторы в переводе на 100-киловаттные и т. д.) и др. Широко используется измерение в приведенных **штуках** (консервы в условных банках, скот в переводе на крупный и т. д.). Вообще приведенные единицы применяются главным образом при измерении однотипных и взаимозаменяемых видов продукции.

Натуральные и метрические единицы измерения наряду с измерением натурально-вещественных ресурсов, фигурирующих в общественном развитии, широко применяются и при характеристике народонаселения, трудовых ресурсов и других демографических явлений. Здесь используются такие единицы, как человек (житель вообще), взрослый работник, человеко-день или человеко-час работы, болезни, отдыха и т. д. Для таких единиц измерения характерна еще большая разнородность содержания, чем для единиц измерения натурально-вещественных объектов, ибо люди, как социальные и индивидуальные существа, отличаются бесконечным многообразием присущих им черт, свойств, потребностей и т. д. Если, например, один человек в течение дня потратил два часа своего отдыха на занятие спортом, а другой — два часа на чтение художественной литературы, то при равной затрате времени на отдых содержание отдыха различно. И эти различия нельзя привести к одному знаменателю посредством каких-либо коэффициентов. Для получения сравнимых по содержанию численных показателей в подобных случаях измерение должно не просто фиксировать явление (в данном случае отдых), но и его структуру (виды отдыха).

Содержательная разнородность одних и тех же единиц измерения порождается чрезвычайным многообразием черт и свойств объектов и явлений общественной жизни, их многомерным характером. В этой связи возникла проблема разработки таких единиц измерения, которые бы позволяли интегрировать, обобщать натуральные и метрические единицы измерения, характеризующие различные объекты или их свойства и поэтому не поддающиеся прямому сопоставлению. Очевидно, например, что нельзя судить о сравнительном размахе производства тракторной и хлопчатобумажной промышленности по числу штук выпущенных тракторов и метров ткани. Добиться сравнимости здесь можно путем введения **стоимостных** единиц измерения. Стоимостные единицы (рубли) широко применяются при измерении многих явлений общественной жизни (прежде всего производственно-технических и социально-экономических) и позволяют сопоставлять и обобщать весьма разнообразные из этих явлений.

Применение стоимостных (денежных) единиц измерения также связано с определенными сложностями. Главная из них состоит в том, что установление цены (а именно оно лежит в основе стоимостных оценок) всегда является в определенной мере приближенным и даже субъективным, а сама цена — весьма подвижной. Поэтому при использовании стоимостных единиц измерения необходим учет точности данных о ценах и сопоставимости цен в стоимостном отношении.

В целом же стоимостные единицы измерения играют весьма важную роль как в самой общественной жизни, так и при ее изучении.

Широко применяемой интегральной единицей измерения разнородных по своему содержанию количественных признаков являются **индексы.** Их построение представляет собой достаточно сложную процедуру, основанную на применении математических методов обработки количественных данных.

Таким образом, единицы измерения количественных признаков общественных явлений и процессов являются весьма многообразными и усложняются с возрастанием неоднородности и масштабов измеряемых объектов (совершенно, например, очевидно, что рубль является значительно более сложной единицей измерения, чем метр или тонна).

Измерение качественных признаков. Несмотря на то что совокупность количественных признаков, характеризующих объекты, явления и процессы общественной жизни, является весьма обширной, еще более широк круг присущих им качественных, атрибутивных признаков. Эти признаки измеряются, как указывалось, прежде всего посредством номинальной и порядковой шкал. Измерение здесь является качественным и результаты его не могут подвергаться количественному анализу. В этой связи возникла важная проблема выработки количественных методов измерения качественных признаков (т. е. их измерения по интервальной и пропорциональной шкалам) с применением определенных единиц измерения. Простейшим методом измерения атрибутивных признаков является определение численности объектов с определенным свойством в соответствующей совокупности однотипных объектов посредством счета. Так, можно сосчитать количество на том или ином предприятии рабочих мужчин и женщин, рабочих, имеющих то или иное образование, владеющих той или иной специальностью, и т. д. Таким образом, качественные признаки могут быть выражены в показателях количества. Но информативность таких показателей ограниченная, ибо они не показывают размеров соответствующих свойств. Поэтому они хороши только при измерении простых атрибутивных признаков, когда соответствующее свойство однозначно и очевидно (как, например, пол рабочих).

Одна из наиболее актуальных проблем измерения общественных явлений состоит в выработке количественных единиц измерения сложных атрибутивных признаков, признаков-оценок. Эти признаки, как правило, являются комплексными. Они включают целый ряд свойств и черт и необходима единица измерения, которая будет их синтезировать.

Разработкой общих принципов и методов измерения сложных атрибутивных признаков занимается недавно возникшая прикладная математическая дисциплина **квалиметрия** (от латинского «квали» — качество, свойство, и греческого «метрео» — измерять).

Практика измерения сложных атрибутивных признаков привела к выработке такой единицы их количественного измерения, как баллы. Балл — это количественное выражение совокупных качественных, атрибутивных свойств объекта. Практически баллы всегда определяются на основе **экспертных оценок.**

Экспертная оценка комплексных атрибутивных признаков — труднейшая процедура в измерении общественных явлений. Во-первых, весьма сложны и различны сами атрибутивные признаки-оценки. В одних случаях они представляют собой совокупное выражение более простых слагаемых (компонент). Скажем, качество прочитанной лекции определяется знанием предмета и общей эрудицией лектора, его умением логично, понятно, убедительно и интересно подать материал и т. д. В других случаях качественная природа определяется внутренней субстанцией, соответствующим системным свойством, т. е. признаком, составляющим внутреннюю сущность явления. Так, стоимость, определяющаяся общественно необходимыми затратами труда, является внутренним скрытым, системным свойством товаров. Понятно, что и отыскание элементарных составных частей сложного свойства, в результате измерения которых можно получить интегральную оценку, и выявление внутренней субстанции системного свойства, которое может быть основой для такой оценки, чаще всего оказываются весьма затруднительными, а то и вообще невозможными. В последнем случае интегральный показатель может быть построен путем перехода от измерения непосредственных свойств к измерению их взаимосвязей. А это еще больше осложняет задачу.

Во-вторых, придание сложным атрибутивным признакам самих баллов всегда сохраняет элемент субъективности и остается в той или иной мере приблизительным.

В итоге балльные экспертные оценки являются наиболее неопределенными и неоднородными единицами измерения. И хотя они и поддаются математической обработке, они не обладают той точностью и определенностью, что натуральные или метрические единицы измерения. Это хорошо видно на следующем примере. Типичным методом экспертной балльной оценки сложного атрибутивного признака является оценка успеваемости школьников и студентов. В основе ее лежит содержательная градация учащихся по степени овладения ими теми или иными знаниями. Описательно эта степень оценивается как «отличная», «хорошая», «удовлетворительная», «плохая» и «очень плохая», т. е. сложный атрибутивный признак формализуется и представляется в системе качественных оценок. Затем эти оценки выражаются в баллах — 5, 4, 3, 2 и 1. Очевидно, что это выражение является условным, ибо указанные баллы можно заменить другими. Например, разницу между высшим и низшим уровнем данного качества можно взять не пяти-, а десятикратной и т. д. Поэтому баллы не обладают свойством показателей количества, выражаемых натуральными или метрическими единицами измерения. Так, например, если состав одной семьи равен 4 человекам, а второй — 2, то очевидно, что численность первой семьи в два раза больше второй. Если же знания одного студента оценены 4, а второго 2, то нельзя утверждать, что знания первого вдвое больше, чем знания второго.

Но несмотря на свою приблизительность и даже условность, измерение сложных атрибутивных признаков посредством экспертных балльных оценок все же позволяет значительно углубить изучение многих явлений общественной жизни сравнительно с их описательным анализом, ибо открывается путь пусть и к весьма неточной, но количественной оценке соответствующих качеств. Кроме того, не следует забывать, что указанные несовершенства измерения сложных качественных признаков определяются достигнутым уровнем науки и практики, которые непрерывно развиваются. Интенсивно разрабатываются и методы измерения посредством экспертных оценок[1].

Таковы основные общие проблемы измерения, которые имеют непосредственное отношение и к измерению исторических явлений.

§ 2. Измерение исторических явлений

Поскольку возможности и методы измерения количественных и качественных признаков различны, рассмотрим методы измерения исторических явлений с учетом этих различий.

Измерение количественных признаков. Количественные признаки, зафиксированные в исторических источниках, обычно выражены в тех или иных единицах измерения (натуральных или метрических). Они могут быть использованы исследователями либо переведены в другие единицы. Здесь важно учитывать, что в древности и средневековье и даже в новое время одни и те же единицы измерения были разномасштабными. Так, например, десятина, являвшаяся основной единицей измерения земельных площадей в нашей стране в досоветский период, вплоть до середины XIX в. имела разный масштаб. Существовала десятина казенная, равная 2400 кв. саженей (80×30) и десятина хозяйственная, равная 3200 кв. саженей (80×40).

Таким образом, требуется проверка размерности единиц измерения, указанных в источниках, и приведение их к одному масштабу.

При измерении количественных признаков исторических явлений могут быть получены два вида рядов количественных данных. Один из них — **вариационные ряды.** Это ряды данных, показывающие количественную меру того или иного признака у разных объектов, составляющих ту или

---

[1] См.: *Бешелев С. Д., Гурвич Ф. Г.* Экспертные оценки. М., 1973; Статистические методы анализа экспертных оценок. Сб. статей. М., 1977.

иную их совокупность. Например, возраст или зарплата каждого из рабочих того или иного цеха или предприятия либо численность населения в городах той или иной страны в определенный момент представляют собой вариационные ряды. Второй вид рядов количественных признаков — это **временные или динамические ряды.** Их образуют количественные показатели, характеризующие величину того или иного признака у одних и тех же объектов в разные моменты времени. Так, зарплата каждого из рабочих по месяцам или годам либо количество населения в городах за целый ряд лет образуют временные, динамические ряды показателей. Любые количественные признаки, как указывалось, могут измеряться по любой из четырех шкал. Но наиболее точным и естественным является их измерение по пропорциональной шкале.

Единицы учета. Важным моментом при измерении количественных признаков является выбор статистических единиц учета. Например, различные количественные показатели о крестьянском хозяйстве в России в эпоху капитализма, собранные земскими статистиками, указаны в источниках или могут быть представлены исследователем в расчете на крестьянский двор и на душу населения. Выбор статистической единицы измерения зависит от исследовательской задачи. Если историка интересует вопрос о том, каковы были размеры крестьянских хозяйств (по тем или иным признакам) в различных местностях или в одной и той же местности в разное время, необходимо рассматривать показатели на двор. Для сравнения же уровня развития крестьянского хозяйства необходимо брать показатели в расчете на душу населения, ибо они более соразмерны, чем показатели на двор.

При отсутствии ясности о том, какая единица учета необходима, следует использовать одновременно разные единицы и в процессе исследования выяснить, какие из них в наибольшей мере отвечают поставленной задаче.

Усреднение показателей. Измерение количественных признаков и получение системы необходимых данных обычно связано с тем или иным усреднением количественных показателей. Вопрос о том, каким должен быть масштаб усреднения используемых данных также является весьма важным при измерении исторических явлений. Так, те же данные земской статистики о крестьянском хозяйстве могут усредняться в масштабе отдельных селений, волостей (т. е. административных единиц, включающих ряд селений), уездов (состоящих обычно из 10—20 волостей) и губерний (в состав которых входило по 8—15 уездов). Очевидно, что масштаб усреднения, т. е. его пространственные и временные границы, во многом влияет на информативность показателей. Главное здесь состоит в том, что по мере укрупнения масштаба усреднения происходит все большее сглаживание различий и абстрагирование от несущественных из них. Это, с одной стороны, приводит к определенной потере первоначальной информации, но, с другой,— позволяет выявить наиболее типичное и существенное. И, наоборот, слишком мелкий масштаб усреднения будет затушевывать типичное и отражать только особенное. Поэтому масштаб усреднения должен выбираться таким, чтобы, раскрывая типичное и существенное, он вместе с тем позволял выявить и особенное. Этот выбор опять-таки зависит от исследовательской задачи, в частности от пространственного и временного размаха изучаемых явлений и процессов. Если, скажем, то же крестьянское хозяйство рассматривается в пределах всей европейской России, то усреднение не только в масштабе селений или волостей, но даже и уездов будет слишком мелким и может не позволить выявить основные черты. Здесь необходимо усреднение в масштабе губерний. Оно раскроет наиболее типичное и вместе с тем отразит особенное. При затруднительности определения необходимого масштаба усреднения следует использовать разные его масштабы.

В целом же при измерении количественных признаков исторических объектов необходим обоснованный выбор натуральных или метрических единиц измерения, статистических единиц учета и масштабов усреднения количественных показателей.

Итоги измерения количественных признаков обычно сводятся в таблицы, или, как их еще называют, матрицы показателей (см. табл. 1).

В строках такой таблицы содержатся значения различных признаков по одному и тому же объекту (в данном случае крестьянскому двору), а в столбцах — значение одного и того же признака у различных объектов. Подобная матрица может включать любое число объектов и характеризующих их признаков. Она удобна для последующей математической обработки количественных данных.

Восполнение пробелов. Существенным моментом в измерении совокупности объектов, характеризуемых тем или иным набором признаков, является полнота данных. Поэтому возникает проблема их восполнения. Существуют различные методы восполнения пробелов в количественных

Таблица 1. Показатели состояния крестьянского хозяйства (в таком-то селении, такого-то уезда и губернии в таком-то году)

Дворы, № п/п	Население, человек		На двор						
	мужчин	женщин	работников мужчин, чел.	всей земли, дес.	всего посевов, дес.	скота, голов		
						лошадей	коров	свиней
1	3	3	2	10	6	2	1	3
2	4	3	2	12	8	3	2	4
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.

данных. В основе их лежат соответствующие расчеты. Простейшим методом восполнения пробелов в показателях является замещение их средним значением данного признака во всей совокупности рассматриваемых объектов. Так, например, если по тому или иному крестьянскому двору отсутствуют сведения о количестве всей земли, то его можно принять равным среднему количеству земли на двор во всех других дворах.

В более сложных случаях широко применяются математические методы расчета недостающих данных.

Но всякое восполнение пробелов расчетными данными снижает точность количественных показателей, ибо любой расчет является приблизительным. Поэтому восполнение пробелов в показателях путем расчетов может быть, во-первых, ограниченным и, во-вторых, и это главное, должно непременно основываться на надежных исходных данных.

При необходимости получить матрицу показателей, не имеющую пробелов, и невозможности или нецелесообразности заполнения пробелов расчетными данными полноты матрицы добиваются следующим образом. При наличии объектов, по которым отсутствует ряд показателей, исключают из матрицы эти объекты. При наличии же признаков, по которым есть пробелы по ряду объектов, убирают эти признаки. Практически чаще всего делается и то и другое. Кроме того, следует иметь в виду, что существуют методы математической обработки количественных показателей с учетом пробелов в матрице данных.

Измерение качественных признаков посредством счета. Большинство признаков, характеризующих исторические объекты, являются качественными, атрибутивными. Простейшим методом измерения качественных признаков может быть счет, т. е. определение числа объектов с данным свойством в общей совокупности этих объектов. Например, можно сосчитать количество среди рабочих того или иного цеха, предприятия, отрасли промышленности, города и т. д. мужчин и женщин, семейных и одиноких, имеющих профессию или нет, число рабочих определенной профессии, с тем или иным образованием и т. д. Суммарные численности признаков могут быть переведены в частотное выражение, которое показывает долю объектов с данным признаком во всей их совокупности. Для этого общее число объектов принимается равным 1,0, а число объектов с данным признаком относится к общему их количеству. Если общее число рабочих равно 100, а число мужчин — 60, то частота (доля) мужчин будет составлять 0,6.

В отличие от измерения количественных признаков численные значения качественных признаков относятся не к каждому из объектов (объект лишь обладает или не обладает тем или иным свойством), а лишь к их совокупности, а сам количественный показатель не является метрическим, т. е. он не показывает величину свойства как такового, а лишь общее количество и долю объектов с данным свойством. Но несмотря на это, счет может широко применяться при измерении как простых, так и сложных качественных признаков.

При изучении простых качественных признаков, т. е. таких свойств, которые могут рассматриваться как имеющие равную интенсивность у всех обладающих ими объектов, счет фактически является единственно возможным методом измерения. При изучении подобных признаков важно установить доли объектов с тем или иным свойством в общей их совокупности. Так, если заводом выпускаются автомашины двух цветов — серые и черные, то при всем том, что каждая из черных или серых машин может иметь свои оттенки цвета, существенным будет учет именно доли машин с черным и серым цветом и вполне можно полагать, что все машины одного цвета являются одинаково черными либо серыми, т. е. обладают одинаковой интенсивностью свойства.

Счет является эффективным средством измерения и сложных качественных признаков, в том числе и признаков-оценок. Например, рабочих того или иного предприятия можно разделить на имеющих профессию и не имеющих ее. Это сложные качественные признаки, интенсивность которых у разных рабочих может быть неодинаковой, ибо как степень «профессионализма», так и степень «непрофессионализма» могут быть разными. Но при решении многих исследовательских задач различиями в интенсивности этих свойств можно пренебречь. В тех же случаях, когда необходима более детальная информация, следует в этих сложных качественных признаках попытаться выделить их более простые составные части. Так, среди рабочих, не имеющих профессии, можно выделить тех, кто имеет определенный опыт работы на промышленном предприятии, и тех, кто его не имеет, а среди рабочих с профессией выделить тех, кто имеет профессиональное образование, и тех, кто овладел специальностью на рабочем месте, и т. д. Разумеется возможность «разложения» сложного качественного признака на более простые, что увеличивает информацию, определяется и состоянием источниковой базы, т. е. наличием соответствующих сведений.

В общем, посредством счета могут измеряться любые качественные признаки в каких угодно совокупностях объектов, зафиксированные как в отдельных видах источников, так и в различных их совокупностях.

При изучении массовых явлений и процессов исторического развития без такого измерения невозможно обойтись, ибо только оно может дать не отдельные примеры, а систему количественных показателей о тех или иных свойствах этих явлений и процессов. Например, исследуя крестьянское или рабочее движение в той или иной стране в тот или иной период времени, посредством счета можно установить общее количество выступлений, затем выявить число и долю выступлений, вызванных теми или иными причинами, имеющими те или иные формы борьбы, преследовавших те или иные цели, характеризовавшиеся тем или иным составом участников и т. д., т. е. получить систематическую информацию о важных сторонах крестьянского и рабочего движения. Измерение посредством счета может широко применяться также для учета различных однотипных повторяющихся простых и сложных событий и явлений исторического развития (неурожаи, эпидемии, эпизоотии, пожары, войны, те или иные мероприятия в области внутренней и внешней политики и т. д.), что позволяет выявить их пространственную и временную интенсивность и тем самым воздействие на ход исторического развития.

Измерение качественных признаков путем счета может дать эффективные результаты и при исследовании сложных индивидуальных явлений общественной жизни, зафиксированных в нарративных (описательных) исторических источниках. Для того чтобы при изучении того или иного индивидуального события или явления можно было применить счет, необходимо рассмотреть его как определенную систему с присущей ей структурой, т. е. составными элементами и их взаимосвязями, которые могут быть предметом счета. Трудности измерения состоят в выявлении составных элементов рассматриваемого события или явления. Для преодоления их необходим предварительный тщательный анализ и разработка способов формализации содержания соответствующего явления.

При использовании нарративных исторических источников измерение можно применить для анализа не только их содержания, но и авторского стиля, что может послужить основой для атрибуции анонимных произведений. При таком подходе как определенная система рассматривается грамматический строй произведения, а в качестве его исходных элементов, подлежащих счету, будут выступать части речи (существительные, прилагательные, глаголы, наречия, союзы) в тех или иных формах и сочетаниях. Выделив систему этих форм и сочетаний, можно подвергнуть их счету и получить численные характеристики грамматического строя произведения. Последующая их обработка математическими методами позволит установить обобщенные количественные показатели этого стиля [1].

Таким образом, измерение качественных признаков посредством счета тех или иных исторических объектов, событий и явлений самих по себе, либо по наличию у них тех или иных свойств, либо на основе учета исходных элементов их внутренних структур может широко применяться в исторических исследованиях. Результаты счета могут служить информационной основой для раскрытия сущности рассматриваемых объектов, явлений и процессов как сами по себе, так и в сочетании с описательным изложением информации.

Контент-анализ. Возможность и необходимость широкого применения измерения и количественного анализа при рассмотрении качественных признаков общественных явлений привели к разработке специального метода формализации и измерения таких признаков. Этот метод получил название **контент-анализа.** Он применим при изучении простых и сложных качественных признаков и особенно эффекти-

---

[1] См.: *Бородкин Л. И., Милов Л. В., Морозова Л. Е.* К вопросу о формальном анализе авторских особенностей стиля в произведениях Древней Руси. — В сб.: Математические методы в историко-экономических и историко-культурных исследованиях.

вен при обработке больших по объему и разнообразных по содержанию видов или комплексов нарративных исторических источников (материалы прессы, различные описания и записки, всякого рода анкетные и подобные им сведения). Суть контент-анализа состоит в том, что исходя из исследовательской задачи и учета возможностей используемых описательных источников выделяется определенная система качественных признаков, характеризующих свойства объектов и явлений, т. е. вырабатывается определенный формуляр для обработки источников, имеющих вид анкеты. Затем проводится счет объектов или их элементов, обладающих этими признаками и находящихся в определенных сочетаниях. При необходимости (например, при обработке на ЭВМ) эти признаки могут кодироваться, т. е. получать условное числовое обозначение. В итоге описательная информация об исследуемых явлениях будет формализована и представлена в системе числовых значений, последние могут сводиться в различного рода таблицы, удобные для последующей обработки. Например, при измерении посредством счета качественных признаков, характеризующих совокупность тех или иных объектов, таблица может иметь следующую первоначальную форму (табл. 2).

Указанные в таблице числа являются кодами соответствующих признаков. По данным такой таблицы легко установить количество объектов с данным признаком и их долю в общей совокупности, а также те сочетания, в которых встречаются признаки, т. е. выявлять их распределение. Например, можно определить общее число и долю мужчин с тем или иным образованием, либо количество женатых мужчин с тем или иным образованием и т. д.

Таблица 2. Состав рабочих Н-ского завода в таком-то году

№ п/п	Пол		Образование			Семейное положение	
	мужской	женский	неполное среднее	среднее	незаконченное высшее	женат (замужем)	холост
код	11	12	21	22	23	61	62
1	11			22			62
2		12			23	61	
3	11		21			61	
Итого . .							

Таковы основные пути измерения качественных, атрибутивных признаков, зафиксированных в исторических источниках, в тех случаях, когда не возникает необходимости учета интенсивности проявления тех или иных свойств объектов, отражаемых этими признаками.

Проблемы экспертных оценок сложных качественных признаков. Одна из наиболее сложных проблем измерения общественных явлений прошлого и настоящего связана с оценкой интенсивности свойств, присущих объектам и явлениям и отражаемых сложными качественными признаками. Это причина того, что историки при изучении таких признаков избегают количественной оценки интенсивности, ограничиваясь ее описательной характеристикой. Так, например, исследуя остроту классовой борьбы рабочих, как правило, раскрывают ее через такие оценки, как «активная», «пассивная», «острая», «слабая» и т. п. Это, конечно, вносит определенные коррективы и градацию в характеристику интенсивности этой борьбы, но оставляет эту характеристику крайне неопределенной. Поэтому возникает потребность измерения интенсивности свойств, выраженных сложными качественными признаками. Такое измерение необходимо не только для углубления научных исследований явлений общественной жизни, но и для практической деятельности. Все это и породило такой метод измерения интенсивности тех или иных свойств объектов и явлений общественной жизни, как **экспертные оценки.** Суть этого метода, как было сказано выше, состоит в том, что первоначально на основе содержательного анализа выявляется определенная качественная градация интенсивности соответствующего свойства. Затем каждой ступени этой градации придается числовое значение в виде баллов, которые должны отражать степень различия интенсивности свойств на той или иной ее ступени. Поскольку и градация интенсивности свойств и ее количественная балльная оценка производится экспертами, т. е. специалистами в данной области, такой способ измерения сложных качественных признаков и получил название метода экспертных оценок.

Экспертные оценки все более широко применяются в различных областях практической человеческой деятельности. Возьмем ли мы балльные оценки успеваемости учащихся и студентов, тарифные разряды различных специальностей в промышленности или различные категории научных учреждений по оплате труда их сотрудников либо разные сорта товаров и изделий по их качеству, а следовательно, и ценам — во всех подобных случаях имеет место экспертная количественная оценка различий интенсивности тех или иных свойств, выраженных в сложных качественных признаках.

При научном исследовании явлений общественной жизни экспертные оценки наиболее широко используются в области конкретно-социологических исследований [1]. В исторической науке измерение сложных качественных признаков путем экспертных оценок пока не получило заметного распространения. А между тем при всей его ограниченности оно позволяет значительно углубить анализ исторических явлений и процессов. Очевидно, например, что при изучении того же рабочего движения его характер и значение будут раскрыты более полно и глубоко, если наряду с учетом числа забастовок, их длительности и количества участников будет измерена общая интенсивность борьбы рабочих. Одним из путей ее экспертной оценки может быть следующий. По своей общей интенсивности выступления рабочих можно разделить на ряд категорий. Например, можно выделить выступления: 1 — очень слабые, 2 — слабые, 3 — слабо-умеренные, 4 — умеренные, 5 — умеренно-сильные, 6 — сильные, 7 — очень сильные. Такова будет качественная, а именно порядковая шкала оценок. Для специалистов в области рабочего движения чаще всего не будет затруднительным ранжировать выступления рабочих по этой шкале.

Сложнее решить вопрос о количественной, балльной оценке указанных степеней интенсивности борьбы рабочих. Основная трудность состоит здесь в определении количественного интервала между низшим и высшим уровнями интенсивности рассматриваемого свойства. Он всегда будет иметь приблизительный характер. Но и при этом введение балльных оценок имеет большой смысл, ибо открывается путь, хотя и к крайне приближенной, но количественной оценке различий в проявлении сложных свойств. Очевидно, например, что различия в интенсивности борьбы рабочих в двух выступлениях, из которых одно было «очень слабым», а другое «очень сильным», будут более очевидными, если эта интенсивность в первом случае будет оценена в 1, а во втором в 7 баллов.

Это простейший подход к экспертной оценке интенсивности сложных качественных свойств исторических объектов и явлений. Более трудным, но точным будет подход, основанный на экспертной оценке составных компонентов

---

[1] См.: Социальные исследования: построение и сравнение показателей. М., 1978.

комплексных качественных признаков. Для этого сложные свойства, характеризуемые этими признаками, вначале должны быть расчленены на более простые составные части. Затем каждая из этих частей получает балльную оценку, а сумма баллов будет интегральным количественным показателем интенсивности проявления сложного качественного свойства.

Например, в таком сложном свойстве борьбы рабочих, как ее общая интенсивность, можно выделить в качестве основных составных компонентов: уровень сознательности рабочих; цели борьбы и характер требований; формы борьбы и степень ее организованности и последовательности. И хотя каждый из этих компонентов также является сложным качественным признаком, но оценить степень интенсивности этих свойств борьбы рабочих легче, чем интенсивность борьбы в целом. Так, в данном случае очевидно, что такой компонент, как уровень сознательности рабочих, является определяющим по отношению к двум другим. Чем более высоким будет этот уровень, тем более широкими будут цели борьбы и требования рабочих, а также более боевыми и организованными формы борьбы. Следовательно, оценка в баллах первого компонента должна быть более высокой, чем двух других. Вероятно, было бы правомерно принять его максимальный вес в баллах, равным весу второго и третьего компонентов вместе взятых. Можно, к примеру оценивать уровень сознательности рабочих в пределах до 10 баллов, а характер требований и формы борьбы — до 5 баллов.

Разумеется, экспертная оценка в баллах интенсивности отдельных свойств и здесь будет условной, но итоговый результат будет более точным и обоснованным. Допустим, что мы имеем дело с забастовкой 1000 человек рабочих на капиталистическом предприятии в условиях, когда самосознание рабочих было низким и они не понимали, что для радикального улучшения их положения необходима борьба за изменение всего общественного, в том числе и политического, строя. В силу этого их требования имели чисто экономический и узкий характер, а их борьба была неорганизованной, локальной, выжидательной и т. п. В то же время на другом предприятии забастовали 500 рабочих, которые осознавали свои классовые интересы, понимали связь их положения с общим характером буржуазного строя, а поэтому их требования носили широкий экономический и политический характер и выходили за рамки узких местных интересов, а их борьба отличалась высокой организованностью, последовательностью, бескомпромиссностью, была рассчитана на привлечение общественного мнения и т. д. Видимо, не вызовет особых возражений оценка в баллах по компонентам и в целом первого выступления: 3 (из 10) +2 (из 5) + +2 (из 5) =7, а второго: 10+5+5=20. И хотя в первом выступлении участвовало вдвое больше рабочих, общая интенсивность борьбы и ее общественная значимость во втором выступлении были намного большими.

Может возникнуть вопрос об эффективности и целесообразности продемонстрированной процедуры измерения интенсивности борьбы рабочих, поскольку вывод о том, что во втором выступлении она была намного выше, чем в первом, очевиден и без этого измерения. Подобное сомнение неправомерно даже в том случае, если измерение сложных качественных признаков имеет своей целью перевод в экспертные балльные оценки различий в интенсивности сложных свойств, которые выражены описательно и которые характеризуют индивидуальные объекты и явления, ибо различия в интенсивности проявления сложных свойств будут улавливаться и выражаться более точно. Еще более эффективным будет переход от описательной к количественной форме выражения информации о массовых объектах и явлениях, характеризуемых сложными качественными признаками. Возможность применения в обработке данных экспертных оценок математических методов открывает путь к значительному расширению информации об этих объектах и явлениях и углублению ее анализа. Такой эффект будет особенно существенным в тех случаях, когда удается сложное свойство расчленить на его составные компоненты и тем самым перейти к анализу взаимосвязи между ними. Так, если интенсивность борьбы рабочих оценивается в целом, то получается один количественный показатель, характеризующий ее. Если же он, как было показано, расчленен на три составные части, то наряду с этим показателем будет еще три новых. Но самое главное состоит в том, что теперь можно анализировать взаимосвязи между этими четырьмя показателями, что даст дополнительно еще шесть показателей. Понятно, что взаимосвязь интенсивности свойств можно анализировать и при описательном ее выражении. Но возможности такого анализа крайне ограничены и в плане охвата исследуемых объектов, и учета степени интенсивности соответствующего сложного свойства, и точности анализа. Переход же к количественной экспертной оценке интенсивности проявления этих свойств снимает все эти ограничения. При этом экспертные оценки интенсивности выражения сложных качественных свойств могут быть включены в систему других количественных показателей, характеризующих эти объекты и явления. Так, те же экспертные оценки интенсивности борьбы рабочих могут рассматриваться наряду с данными о количестве участников забастовок и их длительности, о составе рабочих, их положении и т. д.

Итак, измерение сложных качественных признаков исторических объектов и явлений, особенно массовых, посредством экспертных оценок открывает возможность для значительного углубления их анализа.

Таким образом, существует не только принципиальная возможность измерения самых различных по своему характеру и масштабам событий, явлений и процессов общественного развития в прошлом и настоящем, но разработан целый ряд эффективных принципов и методов и накоплен большой практический опыт такого измерения.

ГЛАВА 3
ОСНОВНЫЕ ПРОБЛЕМЫ МЕТОДОЛОГИИ КОЛИЧЕСТВЕННОГО АНАЛИЗА ИСТОРИЧЕСКИХ ЯВЛЕНИИ

Количественный анализ исторических явлений и процессов может дать эффективные результаты лишь в том случае, если он во всех своих звеньях основывается на последовательно научной теории и методологии исторического познания. Такой теорией и методологией является исторический материализм и диалектический метод. Именно они позволяют наиболее широко, глубоко и обоснованно подойти к объекту познания и постановке исследовательской задачи, выбору путей и методов ее решения, отбору фактических данных, их обработке и анализу, истолкованию и обобщению. Исторический материализм и диалектический метод представляют собой в этом плане общую теоретико-методологическую основу всего марксистского обществознания, т. е. всех общественных и гуманитарных наук, в том числе и науки исторической.

Количественный анализ как один из путей изучения советскими историками явлений и процессов общественного развития также опирается на общую марксистскую теорию и методологию исторического познания. Вместе с тем существует и ряд конкретно-методологических принципов, присущих именно количественному анализу.

§ 1. Постановка исследовательской задачи

Правильная постановка исследовательской задачи при любых, в том числе и количественных, методах ее решения прежде всего требует всестороннего подхода к изучаемым объектам, явлениям и процессам общественной жизни, рассмотрения их во всем многообразии, внутренней сложности, взаимосвязи и обусловленности, в развитии. Всякие отступления от этих принципов ограничивают познавательные возможности, либо вовсе могут приводить к ошибочным результатам.

Так, представители субъективистской методологии исторического познания, лежащей в основе так называемой событийной, описательной буржуазной историографии, ограничивают объект исторического познания изучением лишь индивидуальных и неповторимых исторических событий, которые, по их мнению, будто бы и исчерпывают основное содержание общественной жизни. Понятно, что при таком подходе к объекту исторического познания конкретные задачи исторической науки могут состоять лишь в воспроизведении, описании этих событий и субъективном истолковании их сути.

Сторонники другого направления немарксистской методологии исторического познания — структурализма, стремясь преодолеть ограниченность субъективистского подхода к объекту исторического познания и постановке исследовательских задач, основным в общественной жизни считают структуры, т. е. внутреннее строение и взаимосвязь элементов различных общественных систем. В анализе этих структур и в построении их моделей прежде всего с помощью количественных методов и усматривается главная задача исторической науки. Такой подход в некоторой мере преодолевает узость субъективистской методологии, ибо открывает путь к познанию массовых исторических явлений, допускает признание определенных закономерностей в историческом развитии, т. е. позволяет ставить более широкие исследовательские задачи и получать более обоснованные и глубокие их решения. Однако и структуралистская методология исторического познания ограничивает его объект и постановку исследовательских задач, ибо, во-первых, также акцентирует внимание, хотя и на очень важной, но лишь одной стороне общественной жизни, а именно на изучении структур, и, во-вторых, рассматривает общественные структуры как фактически внесоциальные, изначальные и самодовлеющие, внутренне неподвижные и неупорядоченные общественные системы.

Только марксизм допускает и требует всестороннего подхода к объекту исторического познания, рассмотрения всей сложной совокупности явлений общественной жизни «...лишь (α) исторически; (β) лишь в связи с другими; (γ) лишь в связи с конкретным опытом истории»[1]. Преимущества такого подхода очевидны. В самом деле, в объективной исторической действительности события, структура и развитие тесно взаимосвязаны. Всякое историческое событие, с одной стороны, несет в себе черты специфического и неповторимого, а с другой — общего и закономерного, тесно связанного с той или иной структурой и развитием. Точно так же всякая общественная система и присущая ей структура находятся в непрерывном движении и развитии, а совокупность самих структур представляет собой, несмотря на их многообразие, упорядоченную иерархию, в которой существует соподчиненность, идущая от базиса к надстройке. Наконец, развитие является не простой последовательностью событий и механической сменой одних структур другими, а внутренне закономерным, поступательно-прогрессивным процессом. Поэтому правомерен только целостный подход к объекту исторического познания. Это означает, что при постановке конкретной исследовательской задачи необходимо учитывать место изучаемого процесса или явления в общей их совокупности, с одной стороны, а рассматривая целое, не забывать о его составных элементах — с другой.

Важное значение при постановке исследовательской задачи имеет объективный учет того, что было достигнуто в результате предшествующего изучения анализируемых или аналогичных явлений и процессов. При этом не следует проявлять ни нигилизма, ни консерватизма по отношению к результатам, достигнутым предшественниками: другими словами, нельзя исходить из позиции либо полного отрицания, либо абсолютизации этих результатов. Любая односторонность будет сужать и обеднять исследовательскую задачу.

Далее, постановка исследовательской задачи не должна исходить из стремления получить субъективно желаемый итог. Она должна быть прежде всего направлена на расширение и углубление подхода к изучаемым явлениям и процессам на базе их теоретического, сущностно-содержатель-

---

[1] *Ленин В. И.* Полн. собр. соч., с. 49, с. 329.

ного анализа и всестороннего обобщения ранее полученных результатов, а также учета тех практических, в том числе и научных, потребностей, которые диктуют необходимость изучения тех или иных исторических явлений и процессов.

Игнорирование указанных методологических принципов может привести к той или иной ограниченности, просчетам и ошибкам в постановке исследовательской задачи, а следовательно, и в полученных результатах.

Приведем один пример. Большой интерес у американских историков и экономистов вызывает изучение факторов, обусловивших ликвидацию в прошлом веке системы рабства. В этой связи уделяется внимание сравнительному анализу двух экономических систем — системы плантационного хозяйства, основанного на рабском труде, и системы сельскохозяйственного производства на базе применения свободного наемного труда. В последнее время в ряде работ, основанных на применении количественных методов и ЭВМ, оспаривается традиционный вывод о том, что к середине XIX в. рабский труд в силу его низкой производительности стал общественно-нерентабельным. Сравнив рабский труд на юге со свободным трудом на севере, авторы этих работ пришли к выводу, что прибыльность рабского труда была выше вольнонаемного. Из этого был сделан вывод, что причины ликвидации рабства лежали не в области экономики, а в сфере морали и справедливости. Даже допуская, что проделанные расчеты правильны, указанный вывод остается необоснованным, а вопрос о сравнительной общественной эффективности рабского и вольнонаемного труда нерешенным. Причина тому — узкая постановка задачи и субъективная устремленность к опровержению ранее полученных результатов. Во-первых, вопрос об эффективности рабского и свободного труда рассматривался лишь в плане их прибыльности и выгодности для плантаторов и фермеров, использовавших этот труд. Что же касается общественной рентабельности и выгодности рабского труда, которая может не совпадать (и действительно не совпадала) с интересами плантаторов, то она оказалась обойденной. Поэтому в соответствующих расчетах авторы и ограничились определением прибыльности рабского труда и не выявили его производительности сравнительно с трудом свободным. Во-вторых, вопрос о рабском и свободном труде рассматривался без учета условий их функционирования и перспектив производственно-экономического прогресса, необходимости совершенствования технической базы производства и т. д. А между тем хорошо известно, что рабский труд оправдывал себя экономически только «...на обширных пространствах естественно плодородной почвы, требующей лишь примитивного труда», что «интенсивные культуры, зависящие не столько от плодородия почвы, сколько от вложенного капитала, образованности и инициативности работника, противоречат самому существу рабства» [1].

Так определенная субъективная заданность в постановке исследовательской задачи, а именно стремление опровергнуть ранее сделанные выводы, и методологическая ограниченность этой постановки, выразившаяся в отождествлении общественной рентабельности рабского труда с его выгодностью для плантаторов, привели к ошибочному итогу.

Таким образом, объективная, всесторонне обоснованная постановка исследовательской задачи — непременное условие успешного применения любых методов ее решения. Особенно важно учитывать это при применении количественных методов, когда для обработки и анализа конкретных данных нередко применяется сложный математический аппарат, что усложняет весь процесс изучения и затрудняет выявление корректности всех его стадий.

Сам круг исследовательских задач исторической науки, которые могут решаться на основе применения количественных методов, весьма обширен. Наиболее необходимы и эффективны эти методы при изучении всякого рода массовых явлений и процессов демографического, экономического, социального, политического и культурного развития. Многие стороны и тенденции, закономерности и особенности функционирования и развития массовых явлений и процессов вообще могут быть выявлены и объяснены лишь на основе применения количественных методов. Поэтому именно исследовательские задачи, которые не могут быть решены традиционными методами, и должны быть в центре внимания при использовании количественных методов в изучении таких явлений. Количественные методы находят все более широкое применение и при исследовании индивидуальных событий и явлений общественного развития. Здесь обращение к этим методам требует выявления тех сторон и черт этих явлений, которые обладают системными свойствами. Таким образом, применение количественных методов требует особенно пристального внимания к постановке исследовательской задачи.

---

[1] *Маркс К., Энгельс Ф.* Соч. 2-е изд., т. 15, с. 344.

§ 2. Репрезентативность количественных данных

Вслед за постановкой исследовательской задачи происходит выявление тех исторических источников, сведения которых могут быть использованы при ее решении. Эти сведения могут быть количественными и описательными. Количественные признаки при необходимости приводятся в единую систему, соответствующую поставленной задаче. Здесь может потребоваться перевод показателей в иные единицы измерения, пересчет на иные единицы учета и т. д. Качественные признаки подвергаются измерению также в соответствии с решаемой задачей. Но прежде чем приступать к обработке и анализу количественных данных, следует установить, во-первых, насколько эти данные достоверны и точны и, во-вторых, в какой мере они репрезентативны, представительны для решения поставленной задачи, т. е. в какой мере они качественно и количественно позволяют правильно раскрыть суть исследуемых явлений и процессов.

Достоверность данных выражается **в точности** измерения соответствующих признаков. Эта точность может варьировать в пределах от весьма приблизительных количественных оценок до полного соответствия показателя действительным размерам характеризуемых явлений. Разница между величиной, полученной в результате измерения, и истинным значением признака называется **ошибкой** или **погрешностью** измерения. Ошибки измерения могут быть качественными и количественными.

Качественные и количественные ошибки измерения вызываются самыми различными причинами. Можно выделить два типа таких причин. Во-первых, это ошибки, возникающие вследствие несостоятельности или ограниченности тех теоретико-методологических посылок, исходя из которых проводится измерение. Во-вторых, это ошибки, которые являются результатом неточности самих измерений (собственно ошибки измерения).

Совершенно очевидна тесная связь измерения с той теорией и методологией, на основе которых рассматривалась соответствующая область общественных явлений. Если теория еще не раскрыла даже самой общей сути явления либо дает ограниченное или ошибочное понимание этой сути, то нельзя рассчитывать на правильное отражение этой сути в соответствующих показателях. Так, например, буржуазная политическая экономия была не в состоянии раскрыть коренную суть капиталистического способа производства ввиду игнорирования такой его черты, как присвоение капиталистом прибавочной стоимости. Эта же теория допускает подмену нормы прибавочной стоимости нормой прибыли, отождествления выгодности и прибыльности производства для капиталиста с его общественной эффективностью и т. д. Ясно, что все это приводило и приводит к ошибкам и искажениям при измерении социально-экономических явлений и прежде всего при выделении тех признаков, которые отражают суть этих явлений.

К неадекватному отражению исторической действительности приводит не только несостоятельность социально-экономических и общественно-политических теорий, но и несовершенство статистических и математических посылок измерения. Понятия, выработанные в общей теории, отражающей ту или иную сторону действительности, для проведения измерения необходимо перевести на язык понятий статистических и математических. Такой перевод может породить ошибки даже при правильной общей теории, особенно в тех случаях, когда измеряются сложные явления и процессы. Такие явления, как производительные силы, производственные отношения, эффективность производства, жизненный уровень и многие другие, даже при ясном теоретическом понимании их сути и на современном этапе науки очень трудно выразить не только в одном, но даже и в ряде показателей. Еще более несовершенными были статистико-математические принципы, т. е. теории измерения общественных явлений в прошлом.

Общие теоретико-методологические посылки сказывались не только на адекватности самого измерения, но и на последующей сводке первичных данных. Так, ценнейшие первичные материалы, собранные русскими земскими статистиками в конце XIX и начале XX в., были в значительной мере испорчены в процессе сводки. Вслед за народниками земские статистики, считая развитие капитализма в России упадком и регрессом и игнорируя разложение крестьянства, абсолютизировали значение средних показателей и поэтому чаще всего прибегали к пообщинным сводкам данных и отказывались от выделения различных по своей хозяйственной состоятельности групп крестьянских дворов.

Таким образом, для оценки достоверности исходных количественных данных как в виде первичных показателей измерения, так и в форме их сводок необходимо выявление тех общеисторических и статистико-математических теоретико-методологических принципов, на основе которых проводилось измерение и получены результаты, зафиксированные в исторических источниках.

Неточности самих измерений могут порождаться ошибками в **регистрации** количественных значений признаков и в их **исчислениях.**

Ошибки регистрации могут быть систематическими и случайными. **Систематические ошибки** являются следствием проявления определенных причин, которые чаще всего могут быть установлены. Систематические ошибки измерения бывают **преднамеренными** и **непреднамеренными.** Первые из них, как правило, являются однонаправленными (систематическое занижение прибыли и завышение расходов промышленниками, занижение размеров феодальной ренты помещиками и т. д.). Непреднамеренные систематические ошибки часто связаны с округлениями (например, возраста), трудностью восстановления по памяти точных данных (например, расходов при бюджетных обследованиях) и другими причинами.

Случайные ошибки регистрации вызываются самыми различными причинами (небрежность или невнимательность регистраторов, неисправность измерительных приборов, несовершенство методов измерения и т. д.). Случайные ошибки имеют разнонаправленный характер (в одних случаях показатель завышается, а в других — занижается). При большом числе наблюдений они имеют нормальное распределение и взаимопогашаются.

Ошибки исчисления возникают при обработке количественных данных в результате многократных вычислительных операций с неточными исходными показателями, замены точных расчетов приближенными, многократных округлений и т. д.

В целом общее отклонение результата измерения от истинных размеров признака складывается из совокупности отклонений, вызванных разными причинами, и нередко может быть существенным. Поэтому для выявления достоверности используемых количественных данных важна проверка их точности.

Идеальным итогом такой проверки является установление числового значения ошибки как в целом, так и по составным ее компонентам. Однако практически чаще всего добиться этого оказывается невозможным и приходится ограничиваться лишь выявлением возможности ошибок и их природы и иногда весьма приблизительной оценкой их величины.

Вопрос о степени той точности, которой должны отвечать используемые количественные показатели, должен решаться исходя из исследовательской задачи. Такая точность называется **необходимой.** Она должна допускать возможность проведения требуемых этой задачей сопоставлений количественных показателей и выявления различий в количественной мере рассматриваемых признаков.

Допустим, например, что сравниваются доходы от земледелия двух крестьянских дворов и предполагается возможность ошибки в имеющихся сведениях об этих доходах в пределах до ±10%. Тогда при доходах одного двора в 100 руб., а другого в 110 руб. фактическая точность данных (у первого двора 90—110 руб., а у второго — 99—121 руб.) будет ниже той, которая необходима для выявления различий в доходах. Лишь при ошибке, равной ±4% и менее, можно утверждать, что доходы второго двора несколько превышали доходы первого, т. е. данные будут иметь необходимую для выявления различий точность.

Для определения качественной и количественной достоверности показателей, зафиксированных в исторических источниках, необходимо установить: кто (какое учреждение или лицо) и с какой целью собирал данные, по какой программе проводился сбор, как был организован сбор, откуда (или от кого) поступили сведения, кто непосредственно собирал их, как обрабатывались и обобщались первичные данные, какова была система проверки данных в процессе их сбора и обработки и т. д. Наряду с этим необходим учет существовавших в то время представлений о сущности явлений, о которых собирались факты, о принципах и методах измерения, а также выявление практических возможностей измерения.

Говоря о достоверности количественных данных, содержащихся в исторических источниках, следует отметить рост степени их достоверности по мере приближения к современности. Используя показатели, обладающие разной степенью достоверности, и имея в виду применение математических методов для их обработки и анализа, историк должен особенно осторожно относиться к данным с сомнительной достоверностью. Вместе с тем никоим образом не следует проявлять нигилизма по отношению к неточным данным и делить сведения источников на пригодные и непригодные для обработки и анализа, как это, к сожалению, нередко бывает. Задача состоит в том, чтобы, ясно представляя себе недостатки данных, найти пути их преодоления, пути повышения информативной отдачи этих данных.

При этом следует иметь в виду одну важную особенность, присущую многим количественным показателям, особенно тем, которые характеризуют массовые явления и процессы исторического развития. Дело в том, что погрешности в измерениях прежде всего возникают при определении числовых значений абсолютных размеров признаков. Но неточные в этом отношении количественные показатели могут быть основой для получения весьма точных относительных сравнительных показателей. Это объясняется тем, что при одинаковых возможностях и методах измерения степень неточности измерений будет примерно равной, следовательно, относительные показатели будут иметь высокую точность.

Приведем один пример. Важнейшим источником для изучения земледелия в России конца XVIII — первой половины XIX в. являются ежегодные ведомости губернаторских отчетов о посевах и сборах хлебов. Ведомости эти содержат сведения о количестве высеваемого зерна и размерах его сбора по губерниям. Данные по губерниям выводились из нескольких тысяч независимых показателей, поступавших с мест. Многие из них были очень неточны. И хотя ошибки измерения были случайными и взаимопогашались, общий губернский показатель посева и сбора хлебов был весьма приближенным. Но поскольку степень неточности данных по всем губерниям ввиду одинаковых принципов и методов их сбора была примерно одинаковой, то эти данные позволяют получать достаточно точные показатели, во-первых, об изменениях уровня земледелия в каждой губернии во времени, т. е. по годам, и, во-вторых, о различиях в уровне земледелия отдельных губерний и районов. Судить об этом можно, например, по высоте урожайности в «самах» (показатель того, во сколько раз сбор хлебов превышал их высев). Так очень неточные абсолютные данные могут дать надежные сравнительные показатели.

И для проверки точности исходных количественных данных, и для получения из недостаточно достоверных сведений точных показателей могут широко применяться математические методы обработки и анализа этих данных.

Таков основной круг проблем репрезентативности количественных данных, который связан с выявлением достоверности показателей, содержащихся в исторических источниках. На основе данных, содержащихся в источниках, и исходя из исследовательской задачи, историк формализует систему данных, анализ которых будет раскрывать суть изучаемых явлений. Эти данные, естественно, должны быть репрезентативными, представительными для решения поставленной задачи. Эта представительность, так же как и достоверность, имеет два аспекта — качественный и количественный.

* *
*

Качественная, содержательная представительность количественных данных определяется тем, в какой мере показатели, на основе которых изучаются соответствующие явления и процессы исторического развития, отражают именно те черты и свойства, которые характеризуют внутреннюю суть этих явлений и процессов. Поэтому принципиальное значение имеет отбор этих показателей исследователем. Этот отбор должен исходить из исследовательской задачи и основываться на качественном, теоретико-методологическом и конкретно-историческом раскрытии сути этой задачи. Тем самым будет определена та система признаков, учет и измерение которых будут представительно характеризовать изучаемые объекты и явления. Понятно, что практическая возможность получения представительной системы количественных показателей зависит и от состояния источниковой базы исследования, и от того, в какой мере источники содержат непосредственно или позволяют выявить необходимые данные.

Отбор показателей является особенно сложным в тех случаях, когда источники содержат большое число их. Есть немало источников, особенно по новому и новейшему периоду, которые включают не только десятки, но сотни и даже тысячи различных показателей. Следует учитывать, что само по себе количество используемых показателей не определяет успеха исследования. Более того, механическое увеличение числа их может усложнить решение задачи и даже увести от главного в ней. Поэтому в исследованиях, основанных на привлечении источников, содержащих большое число показателей, весьма полезным является выявление наиболее важных из них путем предварительной экспериментальной обработки данных по небольшой выборочной совокупности объектов.

Так, при изучении социально-экономического строя помещичьего хозяйства в России в эпоху капитализма использовались описания имений, составлявшиеся при залоге этих имений в Дворянском земельном банке. Они включают несколько десятков показателей, характеризующих земельные владения, обеспеченность имений инвентарем, тягловым и продуктивным скотом, рабочей силой, структуру доходов и расходов и др. В результате экспериментальной обработки небольших выборок данных было установлено, что характер внутреннего строя помещичьего хозяйства, т. е. степень развития в нем капиталистических отношений и сохранения крепостнических пережитков, а именно отработочной системы, представляется возможным раскрыть на основе анализа взаимосвязей всего нескольких показателей: обеспеченность имений инвентарем и тягловым скотом, затраты на рабочую силу, размеры доходов и расходов. Это позволило экономно и целеустремленно вести дальнейший анализ внутреннего строя этого хозяйства. Следует отметить, что при выявлении наиболее существенных показателей могут успешно применяться и математические методы.

Однако при всей эффективности выделения наиболее важных из совокупности количественных показателей, это неизбежно ведет к потере части информации об изучаемых объектах. Между тем есть много исторических явлений, при изучении которых важно учесть как можно большее число характеризующих их признаков. Чтобы добиться этого и вместе с тем оперировать обозримым количеством признаков, существуют математические методы, позволяющие при сохранении всей исходной информации «сжимать» ее и переходить к меньшому числу наиболее общих, интегральных по своему характеру показателей.

При отборе показателей важно не только избежать включения в их число таких, которые не дают или дают мало полезной информации для решения поставленной задачи, но и обеспечить тот необходимый минимум признаков, без которого нельзя правильно раскрыть суть исследуемых явлений и процессов. В противном случае это может привести к необоснованным и ошибочным выводам.

Так, например, один из советских исследователей, изучая зависимость в XIX в. среднегубернских цен на хлеб от внутренних и внешних факторов, предпринял попытку решить эту важную и сложную проблему, используя лишь следующие показатели: 1) цена хлеба; 2) урожайность хлебов; 3) доля (в %) городских жителей; 4) расстояние от губернского города до ближайшего морского порта; 5) цена хлеба в порту. Ясно, что этих показателей недостаточно для решения поставленной задачи. Необходимы сведения о соотношении местного спроса и предложения на хлеб, объема вывоза хлеба из губерний, доле в этом вывозе хлеба, идущего на экспорт, о спросе и ценах на хлеб на внешних рынках и др. Понятно, что без этого указанная задача не может быть решена, а попытка ее решения на базе содержательно непредставительных показателей не обоснована.

Таким образом, обеспечение содержательной представительности количественных показателей является непременным методологическим требованием при применении количественных методов. Если источники не позволяют обеспечить эту представительность, следует внести коррективы в постановку задачи либо воздержаться от попытки ее решения, несмотря на всю привлекательность такой попытки.

* *
*

Наряду с качественной, содержательной представительностью показатели, используемые при применении количественных методов, должны быть репрезентативными в количественном отношении. Их должно быть достаточно для получения надежных, т. е. имеющих необходимую точность, численных значений признаков, характеризующих изучаемые явления и процессы. В тех случаях, когда исследователь располагает данными, характеризующими все объекты рассматриваемой совокупности, и подвергает их сплошной обработке, проблемы количественной представительности показателей не возникает. Она появляется в тех случаях, когда о свойствах всей изучаемой совокупности объектов приходится судить по сведениям, которые охватывают лишь часть этих объектов, т. е. по выборочным данным. При этом историк может иметь дело с двумя типами выборочных показателей или двумя вариантами выборочного метода.

Первый из них, который и является выборочным методом в собственном смысле, состоит в том, что при наличии сведений, характеризующих все объекты рассматриваемой совокупности (она в статистике называется **генеральной совокупностью**), изучение ведется на основе выборки данных. Эта выборка формируется исследователем. Понятно, что общие заключения, сделанные по выборочным данным, могут быть верными только в том случае, если выборка является количественно представительной, т. е. достаточной для правильного отражения числовых значений признаков в генеральной совокупности исследуемых объектов. Доказано, что такими свойствами обладают **случайные выборки,** т. е. такие выборки, при формировании которых каждый объект генеральной совокупности имеет одинаковый шанс попасть в выборку. В математической статистике разработаны различные способы формирования случайных выборок из генеральной совокупности данных.

Эти методы позволяют получать количественные значения рассматриваемых признаков с любой заданной исследователем точностью, т. е. погрешностью в расчетах, и достоверностью, т. е. вероятностью получения истинного результата [1]. Поэтому возникающие иногда споры о преимуществах сплошной обработки количественных показателей сравнительно с выборочным их анализом и о правомерности применения выборочного метода в исторических исследованиях являются необоснованными. Вопрос о сплошном или выборочном анализе количественных данных должен решаться исходя из практической необходимости или целесообразности. При изучении массовых явлений и процессов, зафиксированных историческими источниками, выборочный метод позволяет получить репрезентативные количественные данные для характеристики всей совокупности объектов и избежать больших затрат времени и средств на сплошную обработку данных. Этим и обусловлено широкое и эффективное его применение при изучении массовых явлений и процессов общественной жизни.

С другим типом выборочных данных историк имеет дело в тех случаях, когда он вынужден изучать те или иные явления и процессы на основе частичных, а то и вовсе разрозненных количественных данных, объем которых не представляется возможным изменить. Такие данные получили название **естественных выборок.** Они могут представлять собой лишь сохранившуюся часть некогда сплошных сведений, характеризующих некую генеральную совокупность, либо данные, которые были в силу тех или иных причин собраны лишь по части этой совокупности либо являются результатом выборочного ее обследования и т. д. Так, первичные материалы всякого рода переписей и обследований, делопроизводственного учета и отчетности, личного учета и другие сохранились лишь частично. Многие переписи и обследования, такие, например, как земско-статистические описания крестьянского и помещичьего хозяйства в России периода капитализма, не охватывали всех описываемых объектов либо были выборочными.

Очевидно, что естественные выборки количественных данных вполне могут быть случайными, т. е. репрезентатив-

---

[1] Этим достоверность и точность анализа выборочных данных отличается от достоверности и точности самих количественных показателей, на основе которых ведется этот анализ.

ными. Но они могут и не быть такими. Следовательно, всегда нужна проверка того, в какой мере естественные выборки являются случайными, а следовательно, и представительными для характеристики соответствующей генеральной совокупности. К сожалению, строго математически здесь можно лишь установить, в какой мере объем естественных выборок обеспечивает необходимую точность оценки средних значений признаков в генеральной совокупности по данным выборки. Для этого могут быть использованы математические методы, предназначенные для собственно выборочного метода.

Возможна математическая проверка и того, насколько различия в числовых значениях того или иного признака «естественной выборки» являются случайными. Это дает некоторые основания для суждения о том, является ли данная выборка случайной.

В целом же пока не существует достаточно эффективных математических способов проверки случайности естественных выборок, и историк должен здесь наряду с математическими опираться и на собственно исторические средства.

Весьма важным является выяснение истории возникновения и дальнейшей судьбы данных, дошедших в составе естественной выборки. Если первоначально эти данные охватывали всю совокупность тех или иных объектов и явлений, то выявление последующей судьбы этих сведений, условий их хранения, причин утраты данных и т. д. может помочь выяснению того, в какой мере данные естественной выборки можно считать случайными вариантами некогда полной совокупности показателей. Если же естественная выборка представляет собой итог неполного или выборочного обследования некоей генеральной совокупности объектов, то изучение задач такого обследования и того, как оно проводилось, может дать информацию о том, насколько объекты, характеризуемые естественной выборкой, можно рассматривать как случайные варианты генеральной совокупности.

При проверке случайности естественных выборок очень существенное значение имеет определение того, насколько данные этой выборки равномерно охватывают исследуемую генеральную совокупность объектов в пространстве и времени. Это помогает установить границы генеральной совокупности, характеризуемой естественной выборкой.

Очевидно, что указанные и подобные им содержательно-исторические способы проверки представительности естественных выборок количественных показателей являются приближенными. Но в сочетании с математическими методами оценки случайности вариации значений признаков в естественных выборках и возможной точности и надежности оценки на их основе средних значений признаков в генеральной совокупности это позволяет решить вопрос о репрезентативности естественных выборок.

Приведем один пример. Так, весьма ценные количественные показатели о крестьянском хозяйстве феодальной и капиталистической эпох содержат подворные описания этих хозяйств. Они обычно включают сведения о количестве населения и работников в крестьянских дворах, об обеспеченности их землей, рабочим и продуктивным скотом, о запасах хлеба, размерах повинностей и т. п. Но подобные переписи в период до появления сплошных описаний (которые в России, например, стали входить в практику лишь с конца XIX в.) охватывали лишь отдельные помещичьи имения или селения других категорий крестьян. Так, по русской крепостной деревне первой половины XIX в. в вотчинных фондах содержатся подворные описи, которые охватывают всего несколько сот помещичьих имений из более чем 50 тыс. имений. В большинстве имений подворные описи не составлялись вообще, а из составленных описей сохранилась лишь небольшая часть. Естественно, что для того, чтобы по имеющимся описям можно было судить о состоянии хозяйства и положении всех помещичьих крестьян в это время, необходимо установить, насколько эта естественная выборка может рассматриваться как случайная, т. е. представительная. Изучение происхождения и дальнейшей судьбы этих описей помогает решить этот вопрос. Составление подворных описей крестьянских хозяйств чаще всего (даже в тех имениях, где они велись более или менее регулярно) было связано с различными случайными обстоятельствами (переход имения к новому владельцу, составление обзоров владений, проверка деятельности вотчинной администрации, изменение форм эксплуатации крестьян и размеров их повинностей и т. п.). Так как подворные описи имели практическую ценность всего в течение нескольких лет, ибо положение крестьян и состояние их хозяйства изменялись, то их дальнейшее хранение и вся последующая судьба зависела от множества случайных причин.

Что касается равномерности охвата сохранившимися описями всей совокупности помещичьих имений, то на каждую губернию, входившую в зону размещения русского крепостного крестьянства, приходится по нескольку описей. По времени же описи в основном охватывают 10—20-е и 40—50-е годы XIX в., т. е. начало и середину века.

Все это позволяет считать, что сохранившиеся описи представляют собой случайную естественную выборку, которая может считаться представительной для характеристики состояния хозяйства и положения русского крестьянства в крупных помещичьих имениях в начале и середине XIX в. Имеющиеся данные с высокой достоверностью (вероятностью, равной 95%) позволяют получить достаточно точные общие показатели о состоянии хозяйства и положении русского крепостного крестьянства в то время. Так, в предреформенную эпоху на мужскую душу населения во всей Среднечерноземной полосе (Черноземный центр и Среднее Поволжье) приходилось всего скота в переводе на крупный у беднейших крестьян 0,70—0,90, а у зажиточных 1,97—2,43 голов, т. е. различия между полярными группами крестьян, несмотря на значительность интервалов, в пределах которых могли находиться средние значения признака по всей совокупности, выступают весьма ярко.

Таким образом, даже небольшие по объему случайные естественные выборки могут быть основой для изучения обширных генеральных совокупностей объектов. Поскольку очень многие явления и процессы исторического развития можно исследовать лишь по данным естественных выборок, необходима дальнейшая разработка как исторических, так и математических методов проверки репрезентативных количественных данных этих выборок.

Таковы основные проблемы, возникающие при проверке репрезентативности количественных показателей, как содержащихся в источниках, так и используемых для изучения исторических явлений и процессов.

§ 3. Методологические аспекты применения количественных методов и моделирования исторических явлений и процессов

Решающим этапом в применении количественных методов в исторических исследованиях является математико-статистическая обработка и анализ количественных показателей, привлекаемых исследователем для решения поставленной задачи. Он состоит в использовании тех или иных математических средств (математических функций, уравнений, неравенств, методов выявления взаимосвязей, многомерного анализа, распознавания образов и т. д.), что также требует решения целого ряда методологических проблем.

Основные принципы применения математических методов. Основная цель применения математических методов в любых, в том числе и исторических, исследованиях состоит в том, чтобы в результате обработки и анализа исходных количественных показателей этими методами получить **новую,** т. е. непосредственно не выраженную в исходных данных, **информацию** об изучаемых явлениях и процессах. Историко-содержательный анализ этой информации будет расширять и углублять **знания** об этих явлениях и процессах. Поэтому с познавательной точки зрения применение математических методов обработки и анализа количественных данных оправдано только тогда, когда это действительно дает новую информацию и углубляет знания, а не просто в иной, т. е. математической, форме выражает уже известное. К сожалению, порой престижные соображения порождают стремление облечь в математическую форму и без того очевидные исторические истины.

Очевидно, что полученная математическими методами новая информация должна характеризовать существенные черты и свойства исследуемых явлений и процессов, а также основные закономерности и особенности их функционирования и развития. Это может быть достигнуто при условии **адекватного** отражения применяемым математическим аппаратом сущности этих явлений и процессов. Поэтому выяснение того, в какой мере этот аппарат адекватно воспроизводит эту суть, является важнейшим и непременным условием применения математических методов.

Для суждения о том, какие из возможных математических методов наиболее приемлемы при решении той или иной задачи, историк должен ясно представлять себе логическую суть этих методов. Соотнесение этой сути с логикой самого явления — основной путь решения вопроса об адекватности, а следовательно, и эффективности применяемых математических средств. Понимание логической сути математических методов необходимо и для правильной сущностно-содержательной интерпретации результатов математического анализа.

Практическое осуществление этой процедуры начинается с теоретически-абстрактного определения содержательной сути и путей решения поставленной исследовательской задачи. Оно состоит в том, что выдвигается идея или совокупность идей, которые объясняют суть изучаемых явлений. Это объяснение имеет гипотетический, т. е. абстрактно-теоретический, характер. Далее, исходя из этого и выявляются те математические методы, которые следует применить для обработки и анализа количественных данных.

Допустим, что историк изучает арендные отношения в среде крестьянства в том или ином районе нашей страны в эпоху капитализма. Исходя из того, что аренда по своему характеру могла быть либо предпринимательской, т. е. направленной на производство избыточной продукции для продажи, либо арендой из нужды, когда земля нанималась для производства недостающего продукта, он может выдвинуть гипотезу о преобладании аренды из нужды. Проверить правильность этой гипотезы можно путем выявления соотношения обеспеченности крестьян собственной и арендуемой землей. Содержательно-логически при преобладании аренды из нужды должна иметь место обратная взаимосвязь между обеспеченностью собственной и арендуемой землей, т. е. чем меньше у крестьян было собственной земли, тем больше они должны были иметь земли арендованной. Для решения этой задачи можно применить такой математический метод выявления взаимосвязей, как корреляционный анализ. Если линейный коэффициент корреляции будет свидетельствовать о наличии обратной взаимосвязи между признаками, то гипотеза будет подтверждена. Если же взаимосвязь будет прямой, то ее следует отвергнуть и признать верной обратную ей гипотезу, т. е. преобладание предпринимательской аренды. Это простейший пример, но он отчетливо иллюстрирует этапы постановки и формализации задачи и выбора адекватного метода ее математического решения.

Таковы общие принципы и пути проверки адекватности отражения математическими методами сути явлений. Но можно указать и некоторые конкретные способы такой проверки.

Одним из них может быть применение разных методов обработки и анализа одних и тех же данных. В случае получения одинаковых по своей содержательной сути результатов это будет свидетельствовать об адекватности отражения этими методами соответствующих свойств изучаемых явлений. Другой путь может состоять в том, что из совокупности данных о тех или иных явлениях и процессах формируются две такие выборки, которые характеризуют объекты или явления с заведомо существенно различными свойствами. Обработка этих выборок одним и тем же математическим методом позволит установить, насколько чутко он улавливает эти различия, а следовательно, адекватно отражает суть явлений.

Указанные «испытания» математических методов на адекватность воспроизведения ими исследуемых явлений и процессов исторического развития могут быть очень полезными в тех случаях, когда сопоставление логики явления и логики метода оказывается затруднительным или не дает ясного результата.

Адекватное отражение математическим аппаратом сути явлений — важнейшее, но не единственное условие успешной математической обработки и анализа количественных данных. Этот успех определяется также **корректностью** применения математических методов. В строгом смысле она является составным компонентом адекватности, но о ней следует сказать особо. Дело в том, что математические методы имеют свой диапазон применения, в котором они дают наиболее достоверный и точный результат. Этот диапазон определяется условиями применения соответствующего метода. Соблюдение этих условий и есть корректность применения метода. Так, многие методы математической статистики требуют, чтобы значения количественного признака имели так называемое нормальное распределение, присущее случайным массовым явлениям. Однако закон больших чисел проявляется в общественной жизни, как и в явлениях природы, лишь при определенных обстоятельствах. Основным из них выступает сходство в условиях функционирования объектов и явлений той или иной их массовой совокупности. Тогда различия в числовых значениях тех или иных признаков будут случайными. Если, например, ремесленники одной профессии в том или ином городе, обладая одинаковым уровнем мастерства и имея равные условия производства, имеют неодинаковые издержки производства продукции, то эти различия будут обусловлены разного рода случайными причинами и их распределение должно быть нормальным. В целом же закон больших чисел в общественной жизни представляет собой по существу проявление диалектического единства закономерного и случайного.

Разумеется, далеко не все даже массовые явления общественной жизни подчинены нормальному распределению. Числовые значения признаков, характеризующие эти явления, могут иметь и иные формы распределения. Во всех случаях, когда тот или иной математико-статистический метод рассчитан на определенное распределение, а исследование ведется на основе выборочных данных, непременно необходима проверка характера распределения значений признаков.

Кроме того или иного распределения, многие математико-статистические методы основаны на каком-либо виде функциональной зависимости. Поэтому очень важным является выяснение того, какой вид функциональной зависимости наилучшим образом описывает зависимость, имеющую место в действительности. При этом следует иметь в виду, что если учет характера распределения требуется при анализе только выборочных данных, то подбор надлежащей функции необходим всегда. В противном случае результаты применения математического метода могут быть ошибочными. Так, при анализе внутреннего строя помещичьего хозяйства по данным сельскохозяйственной переписи 1917 г. наряду с другими взаимосвязями выявлялась зависимость обеспеченности помещичьих имений продуктивным скотом от доли (в %) пашни в общем количестве помещичьих земель.

Так, в Средневолжском районе **линейный** коэффициент корреляции показал отсутствие зависимости между этими признаками (он был равен — 0,04). Между тем, корреляционный коэффициент, показывающий тесную взаимосвязь между признаками при их **криволинейной** зависимости, равнялся 0,66, что свидетельствует о наличии тесной обратной зависимости, соответствующей реальному положению.

Существуют и другие условия, учет которых необходим для корректного применения математических методов. О них подробно будет идти речь при характеристике этих методов в следующем разделе книги.

При применении математических методов естественно ориентироваться на более строго обоснованные и эффективные, дающие больше новой и надежной информации методы. Однако при примерно равных возможностях различных методов следует обращаться к более простым из них. Это облегчает решение проблем адекватности методов и корректности их применения, с одной стороны, и сокращает объем счетных операций, что немаловажно при обработке больших массивов данных, — с другой.

Заключительной стадией в применении количественных методов в исторических исследованиях является интерпретация результатов математической обработки и анализа количественных данных и обобщение полученных итогов. Успех здесь определяется уровнем качественного, сущностно-содержательного исторического анализа и глубиной и конкретностью перевода математических показателей на содержательный исторический уровень. При этом необходимо ясно представлять, на какие из поставленных вопросов проведенный анализ дает ответ, а какие из них ответа не получили. Здесь нельзя поддаваться той иллюзии, что строгость математического анализа будто бы придает ему всеобъемлющий характер и гарантирует решение всех проблем. Математический анализ так же, как и другие методы, может и не объяснить всей сути и всего «поведения» изучаемых явлений и процессов. Он также может подтвердить сущностно-содержательную гипотезу, выдвинутую при обращении к математическому анализу, а может и опровергнуть ее. Допустим, выдвигалась гипотеза о том, что забастовочное движение рабочих было обусловлено повышением стоимости жизни. Если при этом анализ взаимосвязи между соответствующими признаками показал, что такая взаимосвязь была низкой или вовсе отсутствовала, то это значит, что забастовочное движение в слабой степени определялось вздорожанием жизни либо вовсе не зависело от него. Следовательно, исходная гипотеза будет отвергнута, а причины борьбы рабочих останутся не выявленными. Итог анализа в подобном случае будет говорить лишь о том, что его следует начать снова, на более широкой основе, ибо борьба рабочих могла вызываться и другими причинами.

Таким образом, определение того, что мы узнали и что осталось неизвестным, является непременным условием завершающей фазы количественного анализа исследуемых явлений и процессов.

Методологические проблемы моделирования исторических явлений и процессов. Применение математических методов обработки и анализа количественных данных, характеризующих исторические явления и процессы, представляет собой построение их формально-количественных, математических моделей. Оно связано с решением целого ряда методологических вопросов.

Суть и цели моделирования. Типы моделей. Как хорошо известно, математическое моделирование определенных явлений и процессов общественной жизни имеет давнюю традицию. Наиболее широкое применение оно получило в демографии и некоторых областях экономической истории. Однако распространенным методом изучения развития общества оно стало лишь в современную эпоху.

Любая научная модель представляет собой абстрагированное выражение основной сущности исследуемых явлений и процессов объективного мира. В основе моделирования лежит теория подобия, а модель выступает в качестве приближенного аналога этих явлений и процессов.

Очевидно, что никакая количественная модель не может быть построена без модели качественной. Поэтому любое научное моделирование состоит из двух этапов — **сущностно-содержательного** и формально-количественного. С этой точки зрения моделирование включает построение **модели качественной** и **модели количественной.**

Сущностно-содержательная модель является результатом теоретического анализа конкретно-научных представлений об объекте моделирования и в обобщенном виде выражает основные черты, закономерности и особенности функционирования исследуемых явлений и процессов, а также их теоретически допустимые состояния. Она служит основой для построения модели формально-количественной и содержательной интерпретации результатов математического моделирования. Этим обусловлена определяющая роль качественной, сущностно-содержательной стороны в процессе моделирования.

Понятно, что научная объективность и эффективность качественной модели всецело зависят от характера той теории и методологии научного познания, на которых она основана. Именно марксистская теория и методология общественного познания позволяет объективно, всесторонне и глубоко раскрывать сущность общественных процессов и явлений и формулировать их содержательные модели. Хорошо известно, что именно марксистский анализ исторического развития достиг той глубины, которая позволила определить сущностно-содержательные модели многих важных проявлений этого развития. К таким моделям относятся, например, определения основных признаков и черт общественно-экономических формаций, способов общественного производства, классов, революционных ситуаций и многие другие. Подобные сущностно-содержательные модели обрели статут научных теорий, подтвержденных ходом исторического развития, и служат для советских ученых основой для построения формально-количественных моделей общественных явлений и процессов как на макро-, так и на микроуровне.

Формально-количественное моделирование состоит в формализованном выражении сущностно-содержательной модели посредством тех или иных математических средств. Наполнение этой модели конкретными количественными данными и их математическая обработка и анализ дают новую информацию об исследуемых явлениях и процессах. Сущностно-содержательное истолкование этой информации приводит к раскрытию сути этих явлений.

В исследовательской практике обществоведов формально-количественные, математические модели могут использоваться, во-первых, для анализа тех или иных явлений и процессов общественного развития, для раскрытия тех его сторон, закономерностей и особенностей, которые не удается выявить более простыми методами. В этом случае модель **отражает реальные**, фактически имевшие место в действительности, черты и свойства явлений и процессов общественной жизни и выступает как их **измеритель**, т. е. как показатель количественной меры тех или иных свойств и состояний объекта моделирования. Такие модели можно назвать **отражательно-измерительными.**

Во-вторых, математические модели могут использоваться для прогнозирования дальнейшего хода развития или оптимального в тех или иных отношениях функционирования определенных систем. Для этого модель должна не только давать информацию, отражающую основные свойства объекта моделирования, но и позволять **имитировать** его функционирование, т. е. воспроизводить данные о его последующем развитии или возможных или допустимых его состояниях. В этом плане моделирование является мощным средством социального прогнозирования, планирования, оптимизации функционирования различных производственных, социальных, управленческих и других систем. Имитация позволяет установить оптимальные с точки зрения стоящих задач варианты развития или функционирования определенных систем. Для реализации этих вариантов затем создаются необходимые условия.

В отличие от отражательно-измерительных моделей рассматриваемые модели можно назвать **имитационно-прогностическими.** Для решения прогностических задач могут использоваться и многие отражательно-измерительные модели. Однако в целом это два разных типа моделей. Основное требование к измерительным моделям состоит в том, чтобы они адекватно отражали то или иное из реально имевших место состояний объекта моделирования. Имитационные модели, кроме того, должны учитывать характер возможных изменений и правильно их воспроизводить. Для этого необходим выбор такого математического аппарата, который допускает возможность имитации. Это значительно усложняет их построение, требует повышения их «чувствительности» к возможным тенденциям в развитии и функционировании объектов моделирования.

Так как функционирование и развитие общественных систем определяется множеством факторов с изменяющейся интенсивностью их действия и характером взаимосвязей, то добиться должной адаптации имитационной модели к возможным изменениям весьма сложно. Поэтому, несмотря на обилие привлекаемой информации, на разнообразие математических средств, используемых для построения прогностических моделей, и широкие возможности для имитации, обеспечиваемые применением ЭВМ, далеко не всегда высказанные прогнозы оправдываются, особенно если они касаются сложных явлений и процессов общественного развития. Так, например, из множества моделей, прогнозировавших развитие экономики капиталистических стран в 70-х годах, ни одной не удалось предсказать экономический кризис 1974—1975 гг. Существующий в настоящее время уровень построения сложных прогностических моделей дает хорошие результаты лишь применительно к плавно протекающим явлениям и процессам.

Отражательно-измерительное моделирование исторических явлений и процессов. В историческом моделировании усилия историков-марксистов направлены прежде всего на построение отражательно-измерительных моделей различных явлений и процессов. Это обусловлено тем, что главной задачей, которую призвана решать историческая наука, является раскрытие сущности, внутренней обусловленности, движущих сил, закономерностей и особенностей реального хода исторического развития.

Можно выделить несколько подходов к построению содержательно-измерительных исторических моделей.

Один из этих подходов связан с моделированием таких явлений и процессов, предельные состояния которых могут быть четко определены теоретически, т. е. сущностно-содержательная модель указывает критерии для выделения разных качественных состояний этих явлений и процессов. Тогда показатели математической обработки количественных данных будут характеризовать стадию развития данного явления или процесса. Например, при всеобщем характере действия закона стоимости изменение цен на рынке должно иметь тесную сбалансированность. Изучая математическими методами взаимосвязь цен на том или ином рынке, по показателям ее тесноты можно судить о степени всеобщности действия закона стоимости. На таком принципе основывалось изучение посредством моделирования процесса формирования Всероссийского аграрного рынка, о чем будет подробно рассказано в третьем разделе книги.

Другой путь построения отражательно-измерительных моделей исторических явлений состоит в том, что качественная суть этих явлений и ее математическое выражение раскрываются на эталонной модели. Для этого из изучаемой совокупности объектов делаются выборки, которые включают объекты с заведомо существенно различными свойствами. Их формально-количественные модели и будут эталонами. Затем проводится математическая обработка данных по всей совокупности. Соотношение показателей этой обработки с эталонными моделями позволит выявить преобладающую тенденцию.

Такой путь моделирования может быть наиболее эффективным при изучении сложных явлений и процессов, при теоретически содержательном моделировании внутренней сути которых трудно выделить четкие критерии, характеризующие качественно отличные ее состояния. Допустим, например, что на основе количественных признаков изучается социальная структура рабочего класса и ставится задача выяснения ее наиболее существенных общих черт. Эти черты складываются из того облика, который присущ отдельным слоям рабочих. Полярными из них являются кадровые рабочие и молодые пополнения рабочего класса. Черты их социального облика будут различными. Но эти различия трудно выразить в четких и очевидных сущностно-содержательных критериях. По тем самым причинам затруднительно сформулировать и сущностно-содержательную модель общего социального облика рабочих. Построение моделей социального облика кадровых рабочих и новых пополнений рабочего класса позволит выявить основные черты облика полярных прослоек. Эти модели и будут эталонами при оценке сущности общего облика рабочего класса, характеризуемого моделью, основанной на данных обо всех слоях рабочих. Разумеется, для практической реализации такого пути моделирования необходимы данные, позволяющие построить эталонные модели.

Еще одним путем моделирования исторических явлений и процессов может быть построение эмпирически-сравнительных отражательно-измерительных моделей. Он состоит в расчленении совокупности изучаемых объектов на ряд однотипных групп и моделировании их структуры. Такой путь также будет целесообразным при невозможности или затруднительности построения четких сущностно-содержательных моделей исследуемой совокупности объектов и составляющих ее групп объектов. Если, например, изучается внутренний строй различных по своей хозяйственной состоятельности групп крестьянства, то трудно теоретически определить, какими специфически конкретными чертами может характеризоваться этот строй у различных групп. Имея, скажем, в виду степень сбалансированности и сопряженности обеспеченности крестьянского хозяйства различными производственными компонентами (земля, рабочая сила, тягловый и продуктивный скот, инвентарь и т. д.), можно предполагать, что эта сбалансированность повышается от беднейших к состоятельным группам. Но такого положения может и не быть. Значит, только построение моделей внутреннего строя крестьянского хозяйства в отдельных группах и их сравнение может дать основу для выявления направления его изменений в зависимости от хозяйственной состоятельности крестьян. И только имея такие модели, можно понять внутренний строй крестьянского хозяйства в целом, т. е. включая все его группы.

Разумеется, указанные пути и принципы построения отражательно-измерительных исторических моделей не исчерпывают всего их многообразия и никоим образом не должны абсолютизироваться. Они лишь показывают сложность построения моделей исторических явлений и процессов и необходимость всякий раз конкретного, внимательного и творческого подхода к этому построению и ведущую роль в нем качественного, сущностно-содержательного подхода.

Успешное применение моделирования в исторических исследованиях требует дальнейшей разработки целого ряда проблем, связанных как с сущностно-содержательным, так и формально-количественным моделированием. Важнейшими из них являются принципы построения качественных моделей, способы проверки адекватности математических моделей и их адаптации к особенностям объектов моделирования и др. При этом наряду с совершенствованием методов построения структурных отражательно-измерительных моделей необходимо особое внимание обратить на динамическое моделирование, ибо общественная жизнь представляет собой процесс непрерывных изменений.

Отражательно-измерительное моделирование фактических исторических явлений широко применяется и в современной буржуазной историографии. Это связано с распространением структуралистской методологии исторического познания, которая основным объектом этого познания считает общественные структуры, а его единственно научным конкретным методом — моделирование. В области формализации и измерения исторических явлений, в создании банков машинной исторической информации, методике и технике построения моделей структуралисты добились несомненных успехов, которые заслуживают внимания. Однако общая ограниченность структуралистской методологии, ошибочное понимание общественных структур как неких непротиворечивых, неизменных и лишенных внутренних движущих сил образований, отрицание возможности познания прошлого в его развитии как поступательно-прогрессивного процесса и абсолютизация синхронного принципа в историческом познании — все это отразилось и на моделировании. В нем недооценивается качественная, сущностно-содержательная сторона и выдвигаются на первый план формально-количественные аспекты. Это ведет к чрезмерной формализации и схематизации, грозит отрывом модели от объекта моделирования, что не только ограничивает познавательные возможности моделирования, но и чревато субъективизмом и искажением исторической действительности[1].

Имитационное моделирование в исторических исследованиях. Наряду с отражательно-измерительными, фактическими моделями исторических явлений и процессов заметное распространение в исследованиях буржуазных историков получило построение контрфактических, имитационных (или как еще их называют, симуляционных) моделей.

Суть подобного моделирования состоит в том, что историк, исходя из той или иной идеи, имитирует контрфактическую, т. е. нереальную, не имевшую места в действительности, ситуацию, строит ее модель и, сравнивая полученные конструкции с исторической действительностью, приходит к выводу о том, «так» или «не так» шло историческое развитие. Приведем пример.

Американский историк-экономист Р. Фогель поставил своей задачей выяснить, в какой мере экономическое развитие США во второй половине XIX в. могло бы обойтись без железных дорог[2]. Для этого имитировалась контрфактическая ситуация — все перевозки грузов якобы осуществляются водным и гужевым транспортом. Полагая, что основной эффект железных дорог состоял в удешевлении перевозок, Фогель путем сложных расчетов, основанных на широком применении различных математических методов, пытался установить, во что обошлись бы перевозки без железных дорог, и тем самым определить, как велико было то «общественное сбережение», которое они давали. Расчеты

---

[1] См. об этом в работе: *Ковальченко И. Д., Сивачев Н. В.* Структурализм и структурно-количественные методы в современной исторической науке. — История СССР, 1976, № 5.

[2] Подробно см.: *Промахина И. М.* Количественные методы исследования в трудах представителей «новой экономической истории» (США). — В кн.: Математические методы в исследованиях по социально-экономической истории. М., 1975.

показали незначительность такого «сбережения». На основании этого был сделан тот вывод, что железные дороги не оказали сколько-нибудь существенного влияния на экономическое развитие США. Общий итог сводится к тому, что при анализе исторического развития историкам не следует придавать большого значения всякого рода техническим открытиям и усовершенствованиям, т. е., говоря марксистским языком, развитию производительных сил.

Работа Фогеля показывает, что отличительной чертой контрфактического имитационного моделирования, как оно понимается и реализуется в современной буржуазной историографии, является субъективизм, произвольность и необоснованность всех его аспектов.

Начать с того, что произвольно конструируется объект моделирования. Без всяких объективных оснований Фогель исключает из экономического организма США железные дороги. С таким же успехом можно устранить и любой другой компонент (например, какую-либо отрасль промышленности или сельского хозяйства, элемент природной или демографической среды и т. д.).

Таким образом, объективная, реальная историческая действительность перекраивается по усмотрению историка. Это неизбежно сказывается на формировании сущностно-содержательной модели, если позволительно говорить о такой модели применительно к реально не существующему объекту моделирования.

Суть этой «модели» у Фогеля сводится к тому, что строительство железных дорог исторически якобы оправдано тогда, когда они дают значительное «общественное сбережение», т. е. сокращают стоимость перевозок. Но ведь хорошо известно, что их эффективность для общества заключалась не только и скорее даже не столько в этом. Железные дороги сделали перевозки независимыми от природных условий, ускорили эти перевозки, что открыло путь на рынок многим товарам, втянули в орбиту активного социально-экономического развития многие регионы, стимулировали развитие ряда отраслей промышленности и т. д. Все это не учитывается.

Субъективен и необоснован подход и к расчетам величины «общественного сбережения». Фогель исходит из того, что цены перевозок водным и гужевым транспортом при отсутствии железных дорог остаются такими же, как и при их наличии. При быстром росте объема перевозок такое допущение (даже при условии строительства новой сети водных каналов) маловероятно.

Наконец, необоснован и конечный вывод о несущественной роли в экономическом развитии технических открытий и усовершенствований. Он идет вразрез со всей историей экономического развития.

Таковы наиболее существенные черты контрфактического имитационного моделирования исторических явлений и процессов, как оно практикуется в исследованиях буржуазных историков. Произвольность и несостоятельность подобного моделирования исторических явлений и процессов очевидны. Чем же порождено такое моделирование?

Уже отмечавшиеся недооценка структурализмом сущностно-содержательной, качественной стороны моделирования и выдвижение на первый план его формально-количественного аспекта, а также допущение возможности отрыва модели от реальных явлений исторической действительности и породили иллюзию о возможности механического перенесения имитационно-прогностического моделирования текущего хода общественного развития на историческое прошлое. При этом не учитывается принципиально существенное различие между имитационно-прогностическим моделированием текущих явлений и процессов общественной жизни и контрфактическим имитационным моделированием исторического прошлого.

Прогнозирование на основе имитации оптимального хода дальнейшего развития или функционирования определенных общественных систем оправдано тем, что для реализации избранного варианта **могут быть созданы необходимые условия.** В этом и состоит большой познавательный и практический смысл имитационно-прогностического моделирования.

Иное дело историческое прошлое. Здесь **условия и обстоятельства развития** того или иного процесса или явления **«заданы» в самом ходе этого развития и не могут быть изменены**. Следовательно, в познавательном плане контрфактическое имитационное моделирование может иметь смысл только в том случае, если историческая действительность содержала возможность иного исхода, т. е. когда объективно существовала альтернатива реально осуществившемуся варианту развития.

Проблемы возможного и действительного, альтернативности в историческом развитии и исторической науке относятся к числу сложных и важных теоретико-методологических аспектов исторического познания. Они вызывают большой интерес у историков, но, к сожалению, еще не получили должной разработки.

Альтернативный характер многих явлений общественной жизни выражает ту ее особенность, которая состоит как в том, что один и тот же экономический базис «...благодаря бесконечно разнообразным эмпирическим обстоятельствам, естественным условиям, расовым отношениям, действующим извне историческим влияниям и т. д.— может обнаруживать в своем проявлении бесконечные вариации и градации...»[1], так и в том, что при всем объективном характере общественное, историческое развитие вместе с тем представляет собой осознанное и целенаправленное действие людей, есть «...*не что иное,* как деятельность преследующего свои цели человека»[2]. Столкновения и борьба различных социальных, классовых и общественно-политических сил и приводят к тому, что из ряда объективно возможных исходов оказываются реализованными, превращаются в действительность лишь некоторые.

Таким образом, конкретным выражением того, что в исторической действительности имела место альтернатива тому или иному реальному исходу, является наличие, с одной стороны, объективной возможности для иного исхода, а с другой — общественных сил, ведущих борьбу за реализацию этой возможности. При этом оба эти компонента тесно взаимосвязаны в любой альтернативе. Объективная потенция без общественных сил, борющихся за ее претворение в реальность, сама по себе не создает альтернативы. Точно так же те или иные субъективные усилия, направленные на борьбу за отличный от действительного исход развития, без наличия объективной возможности такого исхода не приводят к альтернативной ситуации.

Очевидно, что, изучая историческое развитие, историк не только может, но и должен анализировать наряду с тем, «что произошло», и то, «что могло быть», т. е. представляло альтернативу осуществившемуся, но оказалось нереализованным. Такой анализ позволяет более глубоко охарактеризовать внутреннюю суть исторического развития и его реальный ход. В этом и состоит познавательный смысл изучения исторических альтернатив.

Говоря об альтернативности в историческом развитии и необходимости изучения альтернативных ситуаций, следует подчеркнуть неправомерность и научную несостоятельность искусственного конструирования альтернатив истори-

[1] *Маркс К., Энгельс Ф.* Соч. 2-е изд., с. 25, ч. II, с. 354.
[2] *Маркс К., Энгельс Ф.* Соч. 2-е изд., т. 2, с. 102.

ком, попыток увидеть их там, где в действительности их не было. Подобная практика получила широкое распространение в буржуазной историографии. Охарактеризованная работа Фогеля является в этом отношении типичным примером. В историческом развитии США во второй половине XIX в. альтернативы — «с железными дорогами или без них» — не существовало. Альтернативная ситуация здесь всецело сконструирована исследователем. Поэтому ее анализ в сущности ничего и не дает для углубления характеристики экономического развития страны в то время и лишь уводит историков в сторону от изучения действительного хода этого развития.

Следовательно, контрфактическое **имитационное моделирование** исторических явлений и процессов в научно-познавательном отношении **правомерно лишь как альтернативное**. Поэтому такие модели и можно назвать **имитационно-альтернативными.**

Подобные модели могут быть построены на основе реальной информации, т. е. данных, характеризующих действительные черты и свойства исторических явлений и процессов. Представительность и надежность этой информации может быть проверена теми же способами и средствами, что и при построении отражательно-измерительных моделей. Это же относится и к выявлению адекватности используемого математического аппарата. Иначе говоря, альтернативная модель как таковая будет одновременно и моделью измерительной. Но, кроме того, альтернативная модель должна давать возможность для имитационных экспериментов, т. е. обладать способностью воспроизводить объективно возможные, но реально не существовавшие состояния объекта моделирования и выдавать информацию об этих состояниях. Такая возможность обеспечивается выбором соответствующего математического аппарата.

Что касается имитационных моделей, которые изображают не альтернативные, а сконструированные исследователем, т. е. произвольные, исторические ситуации, то они не только не представляют познавательной ценности, но и чреваты искажением реального хода исторического развития. Коль скоро такие модели все же строятся, то в отличие от имитационно-альтернативных моделей их можно назвать **имитационно-контрфактическими.** Основной порок таких моделей можно выразить словами Г. В. Плеханова, который, критикуя Н. К. Михайловского, подчеркивал, что формула прогресса Михайловского «говорит не о том, как шла история, а о том, как она должна была бы идти, чтобы заслужить одобрение г. Михайловского»[1]. Имитационно-контрфактическое моделирование рисует историческое развитие не таким, каким оно было или объективно могло быть в действительности, а таким, каким его хотят видеть авторы этих моделей. В подобном моделировании ограниченность структуралистской методологии исторического познания находит свое логическое завершение — эта методология возвращается на последовательные позиции субъективизма.

* *
*

Таковы основные типы моделей, применяемых в исторических исследованиях, принципы их построения и познавательная эффективность.

Прежде всего очевидно, что моделирование в форме фактических, **отражательно-измерительных** моделей является мощным и эффективным средством, позволяющим на объективной основе значительно углубить анализ коренной сущности, внутреннего механизма, закономерностей и особенностей функционирования и развития явлений и процессов исторической действительности, а также преодолевать трудности, возникающие в связи с отсутствием в источниках данных, непосредственно характеризующих те или иные существенные черты этих явлений и процессов. Основной познавательный эффект таких моделей состоит в том, что они позволяют установить количественную меру соответствующих качественных свойств и состояний объектов моделирования.

Важным средством углубления анализа хода исторического развития могут быть и **имитационно-альтернативные** модели, характеризующие отличные от реальных, но объективно возможные тенденции и варианты развития. Такие модели должны строиться на основе данных, отражающих реальное состояние объектов моделирования, но давать возможность для имитации альтернативных ситуаций. Поскольку сам процесс имитации основывается хотя и на допустимых, но условных посылках и расчетах, постольку его результаты могут иметь лишь **гипотетический** характер. Но они позволяют углубить понимание реального хода развития.

Что же касается **имитационно-контрфактических** моделей, то и построение их для изображения контрфактичес-

[1] *Плеханов Г. В.* Соч. М. — Л., 1925, т. VII, с. 112—113.

ких ситуаций и для реконструкции реальной исторической действительности является неправомерным. Такие модели в силу субъективистского их характера не только не углубляют анализ исторической действительности, но и чреваты ее искажением.

И, наконец, главное. Успех и отражательно-измерительного и имитационно-альтернативного моделирования определяется прежде всего его качественной, сущностно-содержательной стороной, которая в свою очередь зависит от характера той теории и методологии исторического познания, из которых исходит историк. Именно марксистская теория и методология исторического познания с присущими ей принципами последовательного историзма, всестороннего подхода к явлениям исторического прошлого, признанием правомерности применения различных методов исследования и требованием глубокого истолкования и обобщения полученных результатов обеспечивает основу для наиболее объективного и эффективного применения количественных методов и моделирования в исторических исследованиях.

2

ОСНОВНЫЕ МЕТОДЫ МАТЕМАТИКО-СТАТИСТИЧЕСКОГО АНАЛИЗА

Настоящий раздел работы содержит характеристику основных методов математико-статистического анализа, которые широко применяются в научных исследованиях, в том числе и в исторических. Главной задачей при изложении этих методов является раскрытие их логической сути, что дает возможность решать вопрос о сферах их применения и содержательной интерпретации соответствующих показателей и коэффициентов. Для этого суть рассматриваемых методов и методика их применения иллюстрируются на исторических примерах. Что касается техники вычислений, то она дается развернуто в тех случаях, когда применение статистических методов возможно в «ручном» исполнении, т. е. на основе малой вычислительной техники. Если же эти методы требуют обращения к ЭВМ и использования стандартных или составления новых программ, необходим контакт историков с математиками.

ГЛАВА 4
ВАРИАЦИОННЫЕ РЯДЫ И ИХ ХАРАКТЕРИСТИКИ

§ 1. Вариационный ряд

Понятие вариационного ряда. Первичные статистические данные, с которыми имеет дело историк, часто представлены неупорядоченной последовательностью чисел, характеризующей ту или иную сторону процесса или явления. В этой совокупности чисел бывает трудно разобраться, и первичная обработка материалов сводится к приведению имеющихся данных к виду, удобному для анализа.

Пример 1. При обследовании студентов первого курса по возрасту были зафиксированы следующие данные:
17 18 18 18 19 18 20 20 19 18 18 21 19 22 23 18 19 19 19 21 21 18 18 18 18 22 19 18 20 18 19 18 20 19 21 20 22 18 19 21 19 19 22 23 19 20 21 22 17 19.

Таблица 1. Распределение студентов первого курса по возрасту

Возраст студентов (варианты x_i)	Число студентов с данным возрастом (частоты n_i)	Частости (q_i) относительные	Частости (q_i) в %
17	2	0,04	4
18	15	0,30	30
19	14	0,28	28
20	6	0,12	12
21	6	0,12	12
22	5	0,10	10
23	2	0,04	4
Итого	50	1,0	100

Полученный в результате обследования набор чисел будем в дальнейшем называть **статистической совокупностью**, а сами числа, показывающие изменение (вариацию) подлежащего изучению признака, — **вариантами** (обозначим их x_i, где i — номер варианты).

Если упорядочить совокупность исходных данных в убывающем или возрастающем порядке, то получим так называемый **ранжированный ряд.**

Используем для упорядоченной таким образом совокупности более компактную запись (см. табл. 1). В первой колонке поставим различающиеся по величине варианты, расположив их в возрастающем порядке, во второй — числа, показывающие, сколько раз (или как часто) встречаются отдельные значения вариант (назовем их **частотами** и обозначим n_i).

Полученный ряд называется **вариационным.** Сведение первичных данных в вариационный ряд облегчает анализ совокупности. Так, например, видно, что в обследованной группе чаще встречаются студенты в возрасте 18—19 лет, меньше всего студентов с крайними для данной группы значениями возрастов (17 лет, 23 года). Кроме того, вариационный ряд является исходным материалом для большинства методов математической статистики.

При построении вариационного ряда можно приписывать вариантам не частоты, а рассматривать доли каждой варианты во всей совокупности. Они вычисляются как отношения соответствующих частот к объему всей совокупности и называются **частостями** (обозначим их q_i). Частости могут быть выражены в относительных числах или процентах (см. табл. 1).

Дискретный и интервальный вариационные ряды. Изменение признака, по которому обследуются объекты, может быть дискретным и непрерывным. **Дискретной вариацией** признака называется такая, при которой отдельные значения варианты отличаются на некоторую конечную величину. В приведенном примере вариация признака зафиксирована как дискретная (отдельные значения варианты отличаются на единицу). Вариация называется **непрерывной,** если отдельные значения признака могут отличаться друг от друга на сколь угодно малую величину. Примером непрерывной вариации признака служит распределение посевных площадей по урожайности.

В зависимости от вида вариации различают дискретные и интервальные вариационные ряды. Дискретный признак служит основой для построения **дискретного ряда** (см. табл. 1). В случае непрерывного признака варианты объединяют в интервалы, образуя **интервальный ряд**.

В практике исторических исследований непрерывные вариации признака встречаются сравнительно редко, тем не менее интервальные ряды имеют большое значение в обработке исторических данных. Дело в том, что некоторые признаки, принципиально являясь дискретными, принимают такое большое количество значений, что составленный по ним дискретный ряд является практически необозримым, при этом весьма затрудняется дальнейший его анализ. В такой ситуации прибегают к построению интервального ряда (см. табл. 2).

В интервальном вариационном ряду частоты относятся не к какому-либо отдельному значению признака, а ко всему интервалу. Часто в ходе исследования возникает необходимость интервальный ряд рассматривать как дискретный. В таких случаях за значение признака в интервале берут середину этого интервала (центральное значение).

Принципы построения интервального ряда. Первым шагом при построении интервального вариационного ряда является выбор определенного принципа, который кладется в основу построения интервального ряда. Выбор этого принципа зависит от степени однородности рассматриваемой совокупности.

Если совокупность однородна, то при построении ряда используют **принцип равных интервалов.** При этом вопрос об однородности решается содержательным анализом изучаемых явлений.

Следует отметить, что принцип равных интервалов применяется также в тех случаях, когда признак изменяется значительными скачками, природа которых неясна.

Пример 2. Приведем пример вариационного интервального ряда, построенного по принципу равных интервалов (см. таблицу 2).

Если совокупность не совсем однородна, то при построении ряда используют принцип **неравных интервалов**, при этом стремятся добиться качественной однородности объектов внутри интервалов.

Например, при построении вариационного ряда распределения городов и поселков городского типа по числу жителей, применив принцип равных интервалов, мы вынуждены образовать, скажем, такие интервалы: до 50 тыс. жителей, от 50 тыс. до 100 тыс. и т. д., от

Таблица 2. Распределение душ по размеру прирезки в Бельском уезде Смоленской губернии [1]

Размер прирезки x_i, %	Количество душ, к наделам которых сделаны прирезки, q_i, %
До 10	24,5
11—20	26,7
21—30	19,3
31—40	9,7
41—50	5,8
Свыше 50	14,0
Всего . . .	100,0

[1] *Будаев Д. И.* Крестьянская реформа 1861 года в Смоленской губернии. Смоленск, 1967, с. 167.

Таблица 3. Распределение городов и поселков городского типа СССР по числу жителей (по данным переписи 1926 г.) [1]

Группы городских поселений по числу жителей [1] (x_i)	Число городских поселений (n_i)
Менее 10 тыс.	1446
10—20 тыс.	253
20—50 тыс.	135
50—100 тыс.	60
100—500 тыс.	28
500 и более тыс.	3
Всего городских поселений	1925

[1] Народное хозяйство СССР. М., 1956, с. 27.

450 тыс. до 500 тыс., 500 тыс. и более. Но различия между населенными пунктами, имеющими 3 тыс. жителей и 50 тыс. жителей, безусловно, существеннее, чем такие же по абсолютной величине различия между городами, насчитывающими 453 тыс. и 500 тыс. жителей. Очевидно, эти данные целесообразно свести в вариационный ряд с неравными интервалами (см. табл. 3), которые объединяют схожие по размерам города и поселки.

Но не всегда удается получить удовлетворительные результаты и с помощью неравных интервалов. Тогда в основу построения интервального ряда кладется социально-экономический критерий, который призван определить типы, однородные в социально-экономическом отношении. Социально-экономический анализ направлен на то, чтобы определить границы интервалов там, где количественное изменение признака приводит к появлению нового качества. Подобный принцип носит название **типологического.**

Широко использовал типологический принцип в своих статистических исследованиях В. И. Ленин. В частности, анализируя данные германской сельскохозяйственной переписи 1907 г., Ленин вместо 18 групп-интервалов по обеспеченности землей, построенных официальной статистикой, выделил три социально отличные группы хозяйств: пролетарские, крестьянские и капиталистические. Такое выделение позволило выявить степень развития капитализма в сельском хозяйстве Германии[1].

Наконец, еще более тонким способом группировки является способ **специализированного интервала**, идея которого принадлежит В. И. Ленину. Суть этого метода заключается в том, что совокупность разбивается на однородные группы (отрасли, типы хозяйства и т. п.) и для каждой группы строится своя шкала интервалов.

Обратимся к работам В. И. Ленина, в которых он неоднократно указывал на необходимость использования этого способа группировки материала. Так, рассматривая аграрный вопрос в России к концу XIX в., Ленин писал: «Нельзя брать одинаковую мерку крупного и мелкого хозяйства для степного посевщика, для огородника, для табаковода, для «молочного фермера»...»[2]. В работе «Развитие капитализма в России» Ленин дает пример конструктивного подхода к проблеме группировки. Анализируя данные подворных переписей кустарей в Московской губернии, В. И. Ленин выделяет три типа («разряда») кустарных заведений: низший, средний и высший. При этом, утверждает Ленин, «...необходимо было в различных промыслах брать различные основания для разделения кустарей на разряды, напр., в очень мелких промыслах относить к низшему разряду заведения с 1 рабочим, к среднему — с 2-мя, к высшему — с 3-мя и более, а в более крупных промыслах к низшему — заведения с 1—5 рабочими, к среднему с 6—10 и т. д. Без применения различных приемов группировки мы не могли бы представить по *каждому* промыслу данных о заведениях различной величины»[3].

Для того чтобы построить интервальный ряд, после выбора принципа построения нужно определить **величину интервала.** Величина интервала должна быть такой, чтобы, с одной стороны, ряд не оказался слишком громоздким и, с другой стороны, в нем не исчезали бы особенности изучаемого явления. Величина интервала для ряда с равными интервалами определяется соотношением

$$h = R/k, \qquad (4.1)$$

где R — размах **вариации;** k — количество интервалов.

Для ранжированного ряда легко посчитать размах вариации, т. е. разность между наибольшим и наименьшим

---

[1] См.: *Ленин В. И.* Полн. собр. соч., т. 19, с. 326—330.
[2] *Ленин В. И.* Полн. собр. соч., т. 17, с. 121.
[3] *Ленин В. И.* Полн. собр. соч., т. 3, с. 342—343.

значениями признака: $R = x_{max} - x_{min}$, где x_{max} — наибольшее в ряду значение варианты; x_{min} — наименьшее значение варианты.

Тогда для подсчета величины интервала достаточно определить количество интервалов. Вопрос о количестве интервалов решается исследователем в каждом конкретном случае в зависимости от поставленной задачи и особенностей исходных данных.

Величину интервала можно подсчитать и непосредственно. Для ряда с равными интервалами может быть предложена следующая приближенная формула для «оптимальной» (наилучшей) величины интервала:

$$h \approx \frac{x_{max} - x_{min}}{1 + 3{,}2 \lg n}, \qquad (4.2)$$

где n — объем совокупности (число элементов совокупности); $\lg n$ — десятичный логарифм числа n.

Пример 3. Пусть статистическая совокупность состоит из 400 элементов, наибольшее значение варианты равно 65, наименьшее — 5, т. е. в наших обозначениях $n=400$, $x_{max}=65$, $x_{min}=5$. Определить величину интервала для этих данных.

Воспользовавшись таблицей логарифмов и подставив исходные значения в формулу для оптимальной величины интервала — формулу (4.2), получим

$$h \approx \frac{65 - 5}{1 + 3{,}2 \cdot 2{,}6} = 6{,}4.$$

Определение величины интервала для ряда, в основу построения которого положен не принцип равных интервалов, должно базироваться на знании исходного материала, универсальных рекомендаций в этом случае дать невозможно.

Закономерность распределения признака. Анализ вариационного ряда начинается с выявления зависимости между вариантами и частотами (частостями).

В случае неравных интервалов закономерность соотношения между вариантами и частотами (частостями) может не проявиться или же иметь искаженный вид. Поэтому для рядов с неравными интервалами необходимо обеспечить сравнимость частот (частостей), что достигается вычислением **плотности распределения.**

Плотность распределения рассчитывается как отношение частоты (n_i) или частости (a_i) к величине соответствующего интервала (h_i). В зависимости от того, какое берется соотношение, различают **абсолютную** и **относительную** плотно-

Таблица 4. Распределение хозяйств Актюбинского уезда по величине посева [1]

Размер посева, дес.	Частость (q_i)	Величина интервала (h_i)	Плотность распределения
До 3	6,1	3	2,03
3—5	23,4	2	11,70
5—10	36,0	5	7,20
10—15	13,2	5	2,64
15—25	15,5	10	1,55
25—50	4,6	25	0,18
Более 50	1,2	—	—

[1] Исходные данные (столбцы 1 и 2) см.: *Макаров И. Ф.* Казахское земледелие в конце XIX — начале XX века (по материалам экспедиционных обследований 1896—1913 гг.). — В кн.: Материалы по истории сельского хозяйства и крестьянства СССР. М., 1959, сб. III, с. 405.

сти распределения:

$$f_i^a = n_i/h_i, \quad f_i^o = q_i/h_i, \tag{4.3}$$

где f_i^a — абсолютная плотность распределения; f_i^o — относительная плотность распределения.

Пример 4. В табл. 4 дано распределение крестьянских хозяйств Актюбинского уезда по величине посева. Анализируя изменение частостей, обнаруживаем, что самой многочисленной является группа хозяйств, имеющих размер посева от 5 до 10 дес. Примерно в полтора раза меньше хозяйств с посевом от 3 до 5 дес. Группа хозяйств с размером посева от 15 до 25 дес. превосходит группу хозяйств, засевающих от 10 до 15 дес. Эти выводы верны для групп, но не могут дать верного представления о фактическом распределении признака. Дело в том, что группы (интервалы), которые мы рассматриваем, неодинаковы, более крупные из них уже в силу своей величины могут содержать в себе большее число хозяйств, чем менее крупные. Чтобы избавиться от искажающего влияния неравных интервалов и сделать частоты сопоставимыми, рассчитаем плотности распределения, т. е. вычислим, сколько хозяйств приходится на единицу интервала (столбец 4-й табл. 4).

После обеспечения сравнимости частот видим несколько иную картину. Плотность, возрастая, достигает максимального значения на интервале 3—5 и затем постепенно убывает. Значит, самой многочисленной в переводе на единицу группировочного признака является группа хозяйств с посевами от 3 до 5 дес.

Однако и в случае, когда сравнимость обеспечена, закономерность ряда, если даже таковая имеется, не всегда выступает как очевидная. Так, при небольшом числе наблюдений часто не удается получить явно выраженную закономерность. Иногда помочь ее выявлению может **укрупнение интервалов.**

Представляет интерес и обратное укрупнению интервалов преобразование — **расщепление интервалов.** Потребность в расщеплении интервалов возникает, например, тогда, когда сопоставляются два вариационных ряда, построенных для одинаковых признаков, но с разными интервалами либо построенных для разных признаков.

Расщепление интервалов. Существуют способы расщепления интервалов, используемые при различных предположениях относительно распределения признака. Мы ограничимся рассмотрением простейшего случая, когда признак в интервалах распределен равномерно[1].

По существу, из этого исходил и В. И. Ленин при обработке материалов земской статистики. Эти материалы имели тот недостаток, что были сгруппированы по разным признакам: в одних уездах данные группировались по количеству обработанной земли, в других — по посевной площади, в третьих — по рабочему скоту и т. д.

«Для того, чтобы сравнить между собою и свести воедино вышеприведенные данные о разложении крестьянства, мы не можем, очевидно, брать абсолютные цифры и складывать их по группам: для этого требовались бы полные данные по целой группе районов и одинаковость приемов группировки. Мы можем сравнивать и сопоставлять только *отношения между группами высшими и низшими* (по владению землей, скотом, орудиями и т. д.). Отношение, выраженное, например, тем, что 10% дворов имеют 30% посева, абстрагирует различие абсолютных цифр и потому годно для сравнения со всяким подобным отношением любой местности. Но для такого сравнения надо выделить в другой местности тоже 10% дворов, не больше и не меньше. Между тем размеры групп в разных уездах и губерниях не равны. Значит, приходится *дробить* эти группы, чтобы взять по каждой местности *одинаковую процентную долю дворов*. Условимся брать 20% дворов для зажиточного крестьянства и 50% — для несостоятельного, т. е. будем составлять из высших групп группу в 20% дворов, а из низших групп — группу в 50% дворов»[2].

В. И. Ленин поясняет используемый им прием дробления (расщепления) интервалов на условном примере. Пусть имеется пять групп такого размера от низшей к высшей:

---

[1] Способы расщепления интервалов в предположении, что признак изменяется по параболе 2-го порядка, изложены в кн.: *Венецкий И. Г., Кильдишев Г. С.* Основы теории вероятностей и математической статистики. М., 1968.

[2] *Ленин В. И.* Полн. собр. соч., т. 3, с. 119—120.

30, 25, 20, 15 и 10% дворов, им соответствуют такие процентные доли посева: 15, 20, 20, 21 и 24%. Для того чтобы в низшей группе было 50% дворов, необходимо к первой из пяти групп прибавить 20% из второй группы, т. е. $^4/_5$ второй группы: $30+25\cdot{}^4/_5=50\%$.

Чтобы соотношение между группами не изменилось (здесь мы пользуемся предположением о том, что признак распределен равномерно), нужно из процентной доли посева второй группы выделить также $^4/_5$ ее и присоединить к первой группе: $15+20\cdot{}^4/_5=31\%$.

«Подобный прием, — отмечает В. И. Ленин, — допускает небольшую ошибку, вследствие которой разложение представляется *более слабым,* чем оно есть на самом деле. Именно: к высшей группе прибавляются средние, а не высшие представители следующей группы; к низшей группе прибавляются средние, а не низшие представители следующей группы. Ясно, что эта ошибка тем больше, чем крупнее группы, чем меньше число групп»[1].

§ 2. Основные характеристики вариационного ряда

Построение вариационного ряда является только первым шагом в изучении статистических данных. Для более глубокого исследования материала необходимы обобщающие количественные показатели, вскрывающие общие свойства статистической совокупности. Эти показатели, во-первых, дают общую картину, показывают тенденцию развития процесса или явления, нивелируя случайные индивидуальные отклонения, во-вторых, позволяют сравнивать вариационные ряды и, наконец, используются во всех разделах математической статистики при более полном и сложном математическом анализе статистической совокупности.

Существуют две группы характеристик вариационного ряда: 1) меры уровня, или средние, 2) меры рассеяния.

Меры уровня, или средние. Наиболее употребительными в статистических исследованиях являются три вида средних: средняя арифметическая, мода и медиана.

Выбор типа средней для характеристики вариационного ряда зависит от цели, для которой исчисляется средняя, от особенностей исходного материала и от возможностей той или иной средней.

Прежде чем перейти к характеристике отдельных видов средней, сформулируем некоторые, самые общие требования к средней.

[1] *Ленин В. И.* Полн. собр. соч., т. 3, с. 120 (сноска *).

Средняя представляет собой количественную характеристику качественно однородной совокупности. Нарушение этого требования приводит к неверным выводам, искажает суть явления.

Приведем пример, рассмотренный В. И. Лениным. Исследуя арендные отношения в крестьянских хозяйствах Таврической губернии, В. И. Ленин установил, что среди бедняцких элементов деревни в среднем на одно хозяйство арендовалось 2,4 дес. земли, а среди кулацких элементов — 48,6 дес. Формальный подход дает математически верную среднюю для всех крестьянских хозяйств — 12,4 дес. Но подобная средняя является совершенно фиктивной. «Не смешно ли брать «средний» размер аренды, ... складывая вместе крестьян, из которых один берет *2 десятины* за безумную цену (15 руб.), очевидно, из крайней нужды, на разорительных условиях, а другой берет *48 десятин*, сверх достаточного количества своей земли, «покупая» землю *оптом* несравненно дешевле, по 3,55 руб. за десятину?»[1]

Кроме того, необходимо, чтобы **средняя не была слишком абстрактной, а имела ясный смысл в решении задачи.**

Далее, желательно, чтобы процедура вычисления средней была проста. При прочих равных условиях предпочтение отдается той средней, которая проще вычисляется.

И наконец, при выборе средней желательно **свести к минимуму влияние случайных колебаний выборки.** Так, если из одной и той же совокупности взять несколько групп элементов, то средние, им соответствующие, будут, как правило, различаться по величине. Рекомендуется использовать тот вид средней, у которой эти различия минимальны (подробнее о выборке и выборочной средней см. гл. 5 данного раздела).

Наиболее распространенной мерой уровня является **средняя арифметическая:**

$$\bar{x}=\frac{\sum_{i=1}^{k} x_i n_i}{\sum_{i=1}^{k} n_i}=\frac{x_1n_1+x_2n_2+\ldots+x_kn_k}{n_1+n_2+\ldots+n_k}, \qquad (4.4)$$

где $\sum_{i=1}^{k}$ — знак суммирования от 1 до k; x_i — варианты

порядковым номером i; $\sum_{i=1}^{k} n_i=n$ — объем совокупности

[1] *Ленин В. И.* Полн. собр. соч., т. 3, с. 73.

(число элементов совокупности); n_i — частота варианты x_i, k — число вариант. Если вместо частот заданы частости q_i, то формула имеет вид

$$x = \frac{\sum_{i=1}^{k} x_i q_i}{\sum_{i=1}^{k} q_i}, \tag{4.4a}$$

де $\sum_{i=1}^{k} q_i = 1$, или 100%.

Пример 5. Вычислим среднюю арифметическую для данных табл. 1:

$$\bar{x} = \frac{17\cdot2 + 18\cdot15 + 19\cdot14 + 20\cdot6 + 21\cdot6 + 22\cdot5 + 23\cdot2}{50} =$$

$$= \frac{972}{50} \approx 19 \text{ (лет)}.$$

Средняя арифметическая интервального вариационного ряда вычисляется следующим образом: за значение признака принимается середина интервала (центральное значение), которая рассчитывается как арифметическая средняя границ интервала. Часто вычисление средней арифметической для интервального вариационного ряда осложняется тем, что неизвестны или нижняя граница первого интервала, или верхняя граница последнего интервала, или та и другая одновременно. В таких случаях иногда[1] условно предполагают, что интервальная разность неопределенного интервала такая же, что и у рядом стоящего, и на основе этого предположения устанавливают границы крайних интервалов.

Пример 6. Вычислим средние размеры наделов крестьян по данным табл. 5.

Для решения задачи прежде всего необходимо найти середины интервалов. Определенная трудность возникает в связи с тем, что первый и последний интервалы являются открытыми. Нижнюю границу первого интервала естественно принять равной нулю. Тогда середина этого интервала равна $(0+2)/2=1$. Для нахождения центрального значения последнего интервала применим предложенный выше прием. Величина интервала, предшествующего последнему, равна 2. Условно принимаем за величину последнего интервала 2. Тогда верхняя граница

[1] Впрочем, в таких случаях часто отказываются от исчисления средней арифметической, заменяя ее модой или медианой.

Таблица 5. Размеры дореформенного надела у крестьян Симбирской губернии [1]

	Надел x_i, дес.				
	до 2	с 2 до 3	с 3 до 5	с 5 до 7	свыше 7
Середины интервалов	1,0	2,5	4,0	6,0	8,0
Процент барщинных крестьян $q_i^{(1)}$	1,8	18,4	63,5	15,2	1,1
Процент оброчных крестьян $q_i^{(2)}$	12,4	17,5	48,2	13,3	8,6

[1] См.: *Канатов В. И.* Своеобразие проведения реформы 1861 года в Среднем Поволжье (по материалам Симбирской губернии).— Вестник МГУ, серия IX. История, 1964, № 5.

этого интервала — 9 и, следовательно, его середина вычисляется так: $(7+9)/2=8$.

Пользуясь формулой средней арифметической (4.4а) и принимая за значение признака середину интервала (строка 2 табл. 5), рассчитываем средний дореформенный надел у барщинных крестьян:

$$\bar{x}=\frac{1,0\cdot 1,8+2,5\cdot 18,4+4\cdot 63,5+6\cdot 15,2+8\cdot 1,1}{100}=4,016.$$

Аналогично вычисляется средний дореформенный надел у оброчных крестьян: $\bar{x}=3,976$.

Кроме средней арифметической широкое распространение имеет другой вид мер уровня — медиана.

Медианой (обозначим M_e) называется такое значение варьирующего признака, которое приходится на середину вариационного ряда.

При нахождении медианы дискретного вариационного ряда могут возникнуть два случая: 1) число вариант нечетно ($k=2m+1$), 2) число вариант четно ($k=2m$). В первом случае $M_e=x_{m+1}$, т. е. медиана равна центральной (срединной) варианте ряда, во втором случае $M_e=(x_m+x_{m+1})/2$ т. е. медиана принимается равной полусумме находящихся в середине ряда вариант.

Пример 7. Пусть дан ряд с нечетным числом вариант:

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9
8	9	11	12	15	16	18	19	19

Тогда число вариант, равное 9, представимо в виде $2m+1=9$, откуда $2m=8$, $m=4$, т. е. $M_e=x_{4+1}=x_5=15$.

Рассмотрим случай четного числа членов:

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}
8	9	11	12	15	16	18	19	19	23	24	40

Здесь $2m=12$, $m=6$ и $M_e=\frac{x_6+x_7}{2}=\frac{16+18}{2}=17$.

Для интервального вариационного ряда медиана вычисляется по формуле

$$M_e=x_{Me(\min)}+h\frac{0{,}5\sum_{i=1}^{k}q_i-U_{M_e-1}}{q_{Me}}, \tag{4.5}$$

где $x_{M_e(\min)}$ — нижняя граница медианного интервала; h — величина этого интервала, или интервальная разность; q_i — частоты или частости; U_{M_e-1} — накопленная сверху частота (или частость) интервала, предшествующего медианному; q_{M_e} — частота или частость медианного интервала.

Содержание введенных обозначений и процедуру вычислений рассмотрим на конкретном примере.

Пример 8. Вычислим медиану по данным табл. 6.

Вычисление медианы начинается с нахождения интервала, содержащего медиану. Медианному интервалу соответствует первая из накопленных частот или частостей, превышающая половину всего объема совокупности. В нашем случае объем совокупности равен 100%, первая из накопленных частостей, превышающая половину всего объема совокупности, — 60,1 (см. табл. 6). Следовательно, интервал 8—12 будет медианным. Далее, $x_{M_e(\min)}=8$, $h=4, U_{M_e-1}=41$, $q_{M_e}=19{,}1$. Воспользуемся формулой (4.5):

$$M_e=8+4\frac{0{,}5\cdot 100-41}{19{,}1}\approx 9{,}9 \text{ (дес)}.$$

Таким образом, серединный размер посева равен примерно 9,9 дес.

Медиану можно использовать в тех случаях, когда изучаемая совокупность неоднородна, и в такой ситуации она будет иметь вполне конкретный смысл. Так, в рассмотренном примере значение медианы имеет следующий смысл: у одной половины хозяйств размер посева меньше, у другой половины — больше, чем 9,9 дес.

Особо важное значение медиана приобретает при анализе асимметричных рядов, т. е. рядов, у которых нагружены (имеют большие частоты) крайние или близкие к крайним значения вариант. Например, медиана даст более верное представление о среднем уровне личных доходов группы семей в капиталистических странах, чем средняя арифметическая, так как медиана не столь чувствительна к край-

Таблица 6. Распределение хозяйств русских переселенцев Чимкентского уезда по размеру посева (1908 г.) [1]

Размер посева x_i, дес.	Всего хозяйств q_i, %	Накопленные частости[2] U_i	Плотность распределения[3] f_i
0—4	16,6	16,6	4,15
4—8	24,4	41,0	6,10
8—12	19,1	60,1	4,78
12—20	23,9	84,0	2,99
20—30	9,7	93,7	0,97
Более 30	6,3	100,0	

[1] Исходные данные (два первых столбца) см.: *Макаров И. Ф.* Указ. соч., с. 430.

[2] Накопленные сверху частости (или частоты) рассчитываются следующим способом. Первая частость (или частота) остается без изменений и заносится в столбец накопленных частотей (или частот), т. е. $U_1=q_1$. Вторая накопленная частость (или частота) получается так: $U_2=U_1+q_2$, третья — $U_3=U_2+q_3=q_1+q_2+q_3$ и т. д.

[3] Ясно, что последний интервал не может быть модальным, т. е. иметь наибольшую в ряде плотность, поэтому мы и не пытались вычислить плотность для этого интервала.

ним (нетипичным в плане постановки задачи) значениям (семьи с большим доходом), как средняя арифметическая.

Медиану следует применять, если вычисление средней арифметической неправомерно вследствие неопределенности интервалов (первого или последнего, или того и другого вместе).

К достоинствам медианы следует отнести также то, что она менее подвержена случайностям выборки, чем средняя арифметическая.

Медиану не следует использовать, когда число наблюдений невелико.

Наряду со средней арифметической и медианой важное значение как мера уровня имеет мода.

Модой (обозначим M_o) называется варианта, наиболее часто встречающаяся в данном вариационном ряду.

Для дискретного ряда мода равна варианте с наибольшей частотой или частостью.

Для интервального вариационного ряда модальный интервал, т. е. интервал, содержащий моду, определяется по наибольшей частоте (частости) в случае равных интервалов и по наибольшей плотности в случае неравных интервалов. Значение варианты, равное моде, отыскивается приближенными методами.

Довольно грубое приближение можно получить, взяв за моду центральное значение модального интервала, т. е. среднее арифметическое границ интервала.

Пример 9. Вычислим моду по данным табл. 6. В последнем столбце табл. 6 вычислены плотности распределения.

Наибольшая плотность соответствует интервалу 4—8. Это и есть модальный интервал.

Рассчитываем моду:

$$M_o = (4 + 8)/2 = 6 \text{ (дес.)}.$$

Таким образом, получаем, что наиболее типичным по размеру посева хозяйством русских переселенцев Чимкентского уезда в 1908 г. было хозяйство, засевавшее 6 дес. земли.

Моду можно вычислить также как взвешенную среднюю арифметическую из нижней и верхней границ модального интервала (весами в расчете будут служить частоты или частости интервалов предмодального и послемодального). При этом, если ряд построен правильно (см. принципы построения вариационного ряда) и интервалы, соседние с модальными, мало отличаются друг от друга, т. е. распределение близко к симметричному, то этот способ дает хорошие результаты.

Воспользовавшись вторым методом исчисления моды, рассчитаем наиболее типичный размер посева по данным табл. 6:

$$M_o = \frac{4 \cdot 16{,}6 + 8 \cdot 19{,}1}{16{,}6 + 19{,}1} = 6{,}1 \text{ (дес.)}.$$

Мода имеет те же достоинства, что и медиана. Мода и медиана эффективно используются в качестве мер уровня, но сравнительно со средней арифметической реже употребляются как исходный материал для более сложных методов математической статистики.

Меры рассеяния. Рассмотренные выше средние показывают уровень вариационного ряда, другими словами, позволяют ряд чисел охарактеризовать одним числом. Однако средние не содержат в себе информации о том, насколько хорошо они представляют всю совокупность. Одинаковые или близкие по величине средние могут относиться к весьма различным рядам. Для пояснения этого положения рассмотрим условный пример.

Пример 10. В табл. 7 приведены данные о возрасте (для простоты число их невелико).

Рассчитав, получаем, что средний возраст в 1-й и 2-й группах одинаков и равен 36. Но простейшее сравнение этих двух рядов показывает, что одинаковые средние представляют две совершенно различные

Таблица 7. Распределение двух групп по возрасту

	1	2	3	4	5	Всего
1-я группа	31	32	36	40	41	180
2-я группа	14	15	15	66	70	180

по возрастному составу группы, а именно: в 1-ю группу входят люди в зрелом возрасте, тогда как во 2-ю — старики и дети. Иначе говоря, варианты первого ряда довольно тесно группируются вокруг своей средней, т. е. средняя представительна, тогда как во втором ряду обнаруживается сильный разброс (рассеяние) вариант. Чтобы отметить подобные различия, в статистике прибегают к расчету показателей, характеризующих рассеяние признака (мер рассеяния).

Рассмотрим основные меры рассеяния: размах вариации, дисперсию и среднее квадратическое отклонение.

Размах вариации показывает разность между наибольшим и наименьшим значениями признака ($R = x_{max} - x_{min}$). Достоинством этого показателя является простота расчета. Однако возможности его применения ограничены, так как эта характеристика является наиболее грубой из всех мер рассеяния.

Во-первых, при расчете этого показателя рассеяния признака используются только крайние значения признака, остальные же во внимание не принимаются. Во-вторых, размах вариации существенно зависит от случайных колебаний выборки.

Более ценными для характеристики рассеяния признака являются показатели, при расчете которых используются отклонения **всех** вариант от некоторой средней (например, средней арифметической, медианы). К таким мерам рассеяния, в частности, относятся дисперсия и среднее квадратическое отклонение. Последние меры рассеяния меньше любой другой меры подвержены случайным колебаниям выборки. Среднее квадратическое отклонение и дисперсия нашли широкое применение почти во всех разделах математической статистики.

Дисперсия, или **средний квадрат отклонения** (обозначим σ^2) есть средняя арифметическая из квадратов отклонений вариант от их средней арифметической, т. е. в математической записи

$$\sigma^2 = \frac{\sum_{i=1}^{k} (x_i - \bar{x})^2 q_i}{\sum_{i=1}^{k} q_i}, \tag{4.6}$$

где x_i — варианта с порядковым номером i; $\bar{x}$ — средняя арифметическая; k — число вариант; q_i — частота или частость с порядковым номером i.

Часто для исследования удобно представлять меру рассеяния в тех же единицах измерения, что и варианты. Тогда вместо дисперсии используют **среднее квадратическое отклонение**, которое является квадратным корнем из дисперсии, т. е. среднее квадратическое отклонение вычисляется по формуле

$$\sigma = \sqrt{\frac{\sum_{i=1}^{k}(x_i - \bar{x})^2 q_i}{\sum_{i=1}^{k} q_i}}. \qquad (4.7)$$

Пример 11. Рассмотрим распределение дореформенного надела у крестьян Симбирской губернии отдельно для группы барщинных и группы оброчных крестьян (пример 6). Средние величины дореформенных наделов для обеих групп крестьян оказались практически равными (4,018 дес. у барщинных и 3,976 у оброчных). Выясним, насколько одинаковой была вариация показателей в этих группах. С этой целью вычислим средние квадратические отклонения по совокупности барщинных и по совокупности оброчных крестьян.

Таблица 8. Данные для вычисления среднего квадратического отклонения

Размеры наделов, дес.	Середины интервалов x_i	Барщинные крестьяне				Оброчные крестьяне	
		q_i	$\lvert x_i-\bar{x}\rvert$	$(x_i-\bar{x})^2$	$(x_i-\bar{x})^2 q_i$	q_i	$(x_i-\bar{x})^2 q_i$
До 2	1,0	1,8	3,0	9,0	16,2	12,4	111,6
2—3	2,5	18,4	1,5	2,2	40,5	17,5	38,5
3—5	4,0	63,5	0	0	0	48,2	0
5—7	6,0	15,2	2,0	4,0	60,8	13,3	53,2
Свыше 7	8,0	1,1	4,0	16,0	17,6	8,6	137,6
Итого:		100,0			135,1		340,9

Для вычисления средних квадратических отклонений удобно составить вспомогательную таблицу (табл. 8). В ней зафиксированы все промежуточные расчеты. Подставляя результаты этих расчетов в формулу (4.7), получим среднее квадратическое отклонение для барщинных крестьян:

$$\sigma = \sqrt{135{,}1/100} = \sqrt{1{,}351} \approx 1{,}16$$

и среднее квадратическое отклонение для оброчных крестьян:

$$\sigma = \sqrt{340{,}9/100} = \sqrt{3{,}409} \approx 1{,}85,$$

т. е. колебание признака у оброчных крестьян примерно в полтора раза больше, чем у барщинных.

Таким образом, средняя величина дореформенного надела у барщинных и оброчных крестьян Симбирской губернии почти одинакова, т. е.

в среднем эти группы крестьян по обеспеченности землей практически не отличаются. Но в среде оброчных крестьян различия в размере наделов больше, чем среди барщинных крестьян.

Рассмотренные выше меры рассеяния (размах вариации, дисперсия, среднее квадратическое отклонение) являются абсолютными величинами, судить по ним о степени колеблемости признака не всегда можно, в некоторых задачах необходимо использовать относительные показатели рассеяния. Таким показателем является коэффициент вариации.

Коэффициент вариации (обозначим V) представляет собой отношение среднего квадратического отклонения к средней арифметической, выраженное в процентах, т. е.

$$V = \frac{\sigma}{\bar{x}} 100\%. \tag{4.8}$$

Коэффициент вариации позволяет: 1) сравнивать вариацию одного и того же признака у разных групп объектов, 2) выявить степень различия одного и того же признака у одной и той же группы объектов в разное время, 3) сопоставить вариацию разных признаков у одних и тех же групп объектов.

Пример 12. Проведем сравнительный анализ затрат труда и различий в затратах труда в совхозах и колхозах по данным табл. 9. Она содержит исходные данные и промежуточные вычисления.

Таблица 9. Затраты труда на производство зерна в 1958 г. по районам РСФСР (на 1 ц. в чел.-днях) [1]

Район РСФСР	В совхозах			В колхозах		
	x_i	$x_i-\bar{x}$	$(x_i-\bar{x})^2$	x_i	$x_i-\bar{x}$	$(x_i-\bar{x})^2$
Северный	1,8	1,2	1,44	5,1	3,5	12,25
Северо-Западный	1,2	0,6	0,36	3,9	2,3	5,29
Центральный	0,8	0,2	0,04	2,0	0,4	0,16
Волго-Вятский	0,5	—0,1	0,01	2,1	0,5	0,25
Центрально-Черноземный	0,4	—0,2	0,04	0,9	0,7	0,49
Поволжский	0,2	—0,4	0,16	0,6	1,0	1,00
Северо-Кавказский	0,2	—0,4	0,16	0,4	1,2	1,44
Уральский	0,3	—0,3	0,09	1,1	0,5	0,25
Западно-Сибирский	0,3	—0,3	0,09	0,4	1,2	1,44
Восточно-Сибирский	0,3	—0,3	0,09	0,6	1,0	1,00
Дальневосточный	0,4	—0,2	0,04	0,7	0,9	0,81
	6,4		2,52	17,8		24,38

[1] Народное хозяйство СССР. М., 1960.

Используя формулы средней арифметической, среднего квадратического отклонения и коэффициента вариации, получим: для совхозов — $\bar{x} = 0{,}6$; $\sigma = 0{,}4786$; $V = 79{,}8\%$; для колхозов — $\bar{x} = 1{,}6$; $\sigma = 1{,}489$; $V = 93{,}1\%$.

Следовательно, в среднем по РСФСР затраты труда в совхозах примерно в 2,7 раза ниже, чем в колхозах (0,6/1,6), а различия между районами РСФСР в затратах труда в зерновом производстве в колхозах выше, чем в совхозах, т. е. совхозы составляют более однородную совокупность по затратам труда, чем колхозы.

§ 3. Графическое представление вариационных рядов

Графическое представление играет важную роль в изучении вариационных рядов, так как позволяет в простой и наглядной форме проводить анализ статистических данных.

Существует несколько способов графического изображения рядов (гистограмма, полигон, кумулята, огива), выбор которых зависит от цели исследования и от вида вариационного ряда.

Полигон распределения в основном используется для изображения дискретного ряда, но можно построить полигон и для интервального ряда, если предварительно привести его к дискретному. Полигон распределения представляет собой замкнутую ломаную линию в прямоугольной системе координат с координатами (x_i, q_i), где x_i — значение i-го признака, q_i — частота или частость i-го признака.

Пример 13. Построим полигон распределения по данным табл. 1. В прямоугольной системе координат на горизонтальной оси откладываем значения признака (возраст студентов), а на вертикальной оси — частоты (число студентов с данным возрастом). Полученные точки соединим отрезками прямой. Для того чтобы фигура была замкнутой, введем дополнительно новые значения признака (16 лет, 24 года); соответствующие им частоты, естественно, равны нулю. В результате получим полигон распределения студентов по возрасту (рис. 1).

Гистограмма распределения применяется для изображения интервального ряда. Для построения гистограммы на горизонтальной оси откладывают последовательно отрезки,

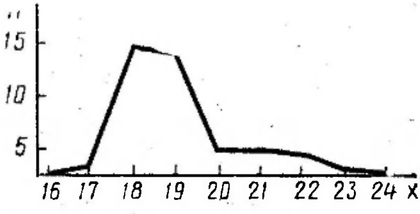

Рис. 1. Полигон распределения студентов по возрасту

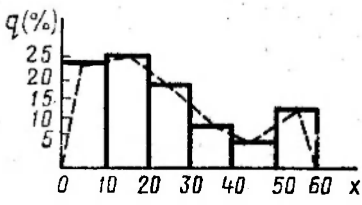

Рис. 2. Гистограмма распределения душ по размеру прирезки

Таблица 10. Расчет накопленных частостей по данным табл. 2

Размер прирезки	Частости	Накопленные частости	Размер прирезки	Частости	Накопленные частости
До 10	24,5	24,5	31—40	9,7	80,2
11—20	26,7	51,2	41—50	5,8	86,0
21—30	19,3	70,5	51—60	14,0	100,0

равные интервалам признака, и на этих отрезках, как на основаниях, строят прямоугольники, высоты которых равны частотам или частостям для ряда с равными интервалами, плотностям — для ряда с неравными интервалами.

Пример 14. Построим гистограмму распределения душ по размеру прирезки в Бельском уезде Смоленской губернии по данным табл. 2 (рис. 2) [1].

Как уже отмечалось, для интервального ряда также можно построить полигон распределения. Для этого за значения признака принимают середины интервалов и для полученного дискретного ряда обычным способом строят полигон. Полигон распределения можно получить и по готовой гистограмме. Достаточно соединить отрезками прямых середины верхних оснований прямоугольников и замкнуть фигуру описанным способом. Результаты такого построения изображены на рис. 2 пунктирной линией.

Кумулята есть графическое изображение вариационного ряда, когда на вертикальной оси откладываются накопленные частоты или частости, а на горизонтальной — значения признака. Кумулята служит для графического представления как дискретных, так и интервальных вариационных рядов.

Пример 15. Построим кумуляту по данным интервального ряда табл. 2. Предварительно рассчитаем накопленные частости.

Обозначим на горизонтальной оси интервалы (рис. 3). Нижней границе первого интервала соответствует частость, равная нулю, а верхней границе — вся частость этого интервала (24,5). Верхней границе второго интервала соответствует накопленная частость первых двух интервалов (51,2) и т. д.

Возможности графического изображения статистических данных не ограничиваются воспроизведением материала в наглядном, легко воспринимаемом виде. Представление данных в виде графика позволяет просто и быстро получить приблизительные значения таких средних характеристик ряда, как мода и медиана.

---

[1] За неимением дополнительных данных при построении графика воспользуемся предположением, что величина последнего открытого интервала равна величине предыдущего.

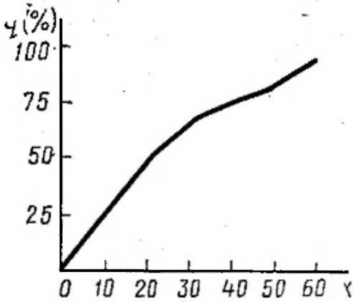

Рис. 3. Кумулята

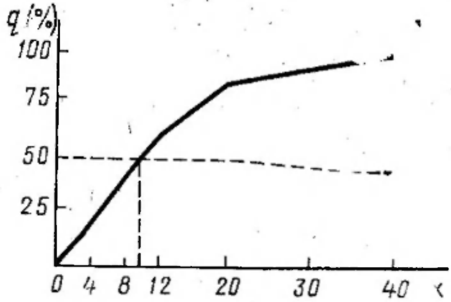

Рис. 4. Графическое определение медианы

Используя определенные виды графического изображения вариационного ряда, можно приближенно оценить моду и медиану. Покажем способы отыскания этих характеристик на конкретных примерах.

Пример 16. Определим приближенно медиану по данным табл. 6. Для этого построим кумуляту и на уровне 50% (середина ряда) проведем прямую линию, параллельную горизонтальной оси. Далее, из точки пересечения этой линии с кумулятой опустим перпендикуляр на горизонтальную ось. Точка пересечения перпендикуляра с осью и показывает приближенное значение медианы (рис. 4). Искомая медиана в нашем примере равна 10 дес. а вычисленная в примере 8—9,9 дес.

Пример 17. Определим моду по тем же данным, что и в предыдущем примере. Напомним, что интервал, которому соответствует наибольшая частота или частость для ряда с равными интервалами, наибольшая плотность — для ряда с неравными интервалами, является модальным. Выпишем из табл. 6 предмодальный, модальный и послемодальный интервалы с соответствующими плотностями.

	Интервалы	Плотность распределения
Предмодальный	0—4	4,02
Модальный	4—8	6,10
Послемодальный	8—12	4,78

Построим для этих интервалов столбики гистограммы и соединим отрезками вершины прямоугольников (рис. 5). Опустив перпендикуляр из пересечения отрезков на горизонтальную ось, получим приблизительное значение моды. Оно находится на пересечении перпендикуляра с горизонтальной осью и равно 6,3 дес., а расчетное значение равно 6,1 (пример 9).

Нормальное распределение. Если уменьшать интервалы и одновременно увеличивать число наблюдений в них, то гистограмма распределения будет все более приближаться к плавной линии. Кривая, к которой стремится график при

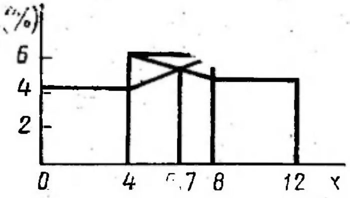

Рис. 5. Графическое определение моды

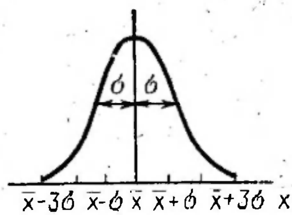

Рис. 6. Кривая нормального распределения

указанном преобразовании, называется **кривой распределения.**

Формы кривых распределения разнообразны. Мы ограничимся рассмотрением одного важного в теоретико-прикладном плане распределения, так называемого **нормального распределения.**

График нормального распределения (рис. 6) представляет собой симметричную одновершинную кривую, напоминающую по форме колокол. Форма нормальной кривой и положение ее на оси абсцисс полностью определяются двумя параметрами — средним арифметическим значением $\bar{x}$ и средним квадратическим отклонением σ. На рисунке видно, что наиболее часто встречаются варианты, близкие к $\bar{x}$, а по мере удаления от $\bar{x}$ варианты встречаются все реже. (Ординаты точек графика на рис. 6 обобщают введенное ранее понятие плотности распределения.)

Каждому значению признака x соответствует при этом определенное значение так называемой функции распределения $F(x)$, показывающее, какова вероятность существования вариант, меньших данного значения x. Геометрически вероятность вариант, меньших x, изображается площадью под кривой слева от точки x. Площадь под всей кривой равна 1, что соответствует полной достоверности (т. е. вероятности того, что признак примет вообще какое-то значение). Таким образом, видно, что функция распределения $F(x)$ обобщает понятие накопленной частоты вариационного ряда.

Ввиду своей важности для практических приложений функция нормального распределения табулирована, т. е. имеются таблицы, где каждому значению x ставится в соответствие вероятность $F(x)$ существования значений, меньших x. Для удобства табулирования в качестве значений признака берутся не сами величины x, а так называемые нормированные отклонения их от среднего значения t, где $t = (x - \bar{x})/\sigma$.

При замене x на t центр распределения смещается в точку 0, а единицей измерения становится величина среднего квадратического отклонения σ, но вид кривой распределения не изменяется. Среднее значение нормированного отклонения t равно нулю, а его среднее квадратическое отклонение равно единице (рис. 7). Нормированная функция нормального распределения обладает следующими свойствами:

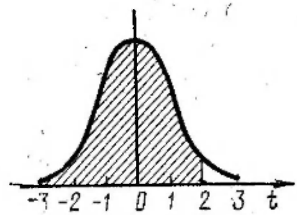

Рис. 7. Нормированная кривая нормального распределения

$F(-\infty)=0$; $F(\infty)=1$; $F(0)=1/2$; $F(-t)=1-F(t)$.

В табл. 7 приложения приведены значения $F(t)$ для положительных значений t. Так, для $t=2$ $F(t)=0{,}97725$. На рис. 7 площадь, соответствующая этой вероятности, заштрихована.

Во многих задачах приходится определять вероятность того, что нормированное отклонение не превысит по модулю некоторой величины t, т. е. значения признака x отклоняются от своего среднего не более чем на $t\sigma$. Это вероятность обозначается $\Phi(t)$ и равна $F(t)-F(-t)=2F(t)-1$. Чаще всего на практике используется именно вероятность $\Phi(t)$, поэтому в табл. 1 приложения табулированы значения $\Phi(t)$. Найдем, например, вероятность того, что нормированное отклонение по модулю не превышает 2, другими словами, значения признака x отличаются от своего среднего по модулю не более чем на 2σ ($|t|\leqslant 2$). По табл. 1 приложения величине $t=2$ соответствует $\Phi(t)=0{,}9545$, т. е. примерно в 95 случаях из 100 значения признака отклоняются от своего среднего не более чем на 2σ.

При использовании статистических методов часто возникает задача проверки нормальности распределения (см. гл. 9), поскольку нормальность является существенным условием их корректного применения.

ГЛАВА 5
ВЫБОРОЧНЫЙ МЕТОД

§ 1. Общая характеристика выборочного метода

Общее понятие о выборочном методе. Множество всех единиц совокупности, обладающих определенным признаком и подлежащих изучению, носит в статистике название **генеральной совокупности.**

На практике по тем или иным причинам не всегда возможно или же нецелесообразно рассматривать всю генеральную совокупность. Тогда ограничиваются изучением лишь некоторой части ее, конечной целью которого является распространение полученных результатов на всю генеральную совокупность, т. е. применяют выборочный метод.

Для этого из генеральной совокупности особым образом отбирается часть элементов, так называемая **выборка,** и результаты обработки выборочных данных (например, средние арифметические значения) обобщаются на всю совокупность.

Теоретической основой выборочного метода является закон больших чисел. В силу этого закона при ограниченном рассеивании признака в генеральной совокупности и достаточно большой выборке с вероятностью, близкой к полной достоверности, выборочная средняя может быть сколь угодно близка к генеральной средней. Закон этот, включающий в себя группу теорем, доказан строго математически. Таким образом, средняя арифметическая, рассчитанная по выборке, может с достаточным основанием рассматриваться как показатель, характеризующий генеральную совокупность в целом.

Разумеется, не всякая выборка может быть основой для характеристики всей совокупности, к которой она принадлежит. Таким свойством обладают лишь **репрезентативные** (представительные) **выборки,** т. е. выборки, которые правильно отражают свойства генеральной совокупности. Существуют способы, позволяющие гарантировать достаточную репрезентативность выборки. Как доказано в ряде теорем математической статистики, таким способом при условии достаточно большой выборки является метод случайного отбора элементов генеральной совокупности, такого отбора, когда каждый элемент генеральной совокупности имеет равный с другими элементами шанс попасть в выборку. Выборки, полученные таким способом, называются **случайными выборками.** Случайность выборки является, таким образом, существенным условием применения выборочного метода.

Области применения выборочного метода в исторических исследованиях. Сфера приложения этого метода в изучении истории обширна. Во-первых, историки могут применять выборочный метод при проведении всякого рода обследований с целью изучения различных явлений и процессов современности. Правда, сейчас такими исследованиями больше занимаются социологи, чем историки, хотя именно историки могут проводить конкретно-социологические обследования, опираясь на исторические данные, и добиваться наибольшего эффекта таких исследований.

Во-вторых, историки нередко имеют дело с сохранившимися данными ранее проведенных собственно выборочных обследований. Такие обследования стали все более широко применяться с конца XIX в. Так, при проведении ряда сплошных обследований и переписей выборочно собирались и собираются сведения по более широкой программе. Многие данные собирались только выборочно. Наиболее интересными среди них для историков являются описания разного рода хозяйственных комплексов (крестьянских хозяйств, промышленных предприятий, колхозов, совхозов и т. д.), а также бюджетные и другого рода обследования различных слоев населения.

В-третьих, в распоряжении историков имеется значительное число разнообразных первичных сплошных массовых данных, полная обработка которых весьма затруднительна даже при применении современной вычислительной техники. При изучении их может быть применен выборочный метод. Такие материалы имеются по всем периодам истории, но особенно много их по истории XIX—XX вв.

Наконец, историкам очень часто приходится иметь дело с частичными данными, так называемыми естественными выборками. При обработке этих данных также может быть применен выборочный метод. Характер естественных выборок бывает различным. Прежде всего они могут представлять собой сохранившийся остаток некогда существовавшей более или менее полной совокупности данных. Так, многие актовые материалы, документы текущего делопроизводства и отчетности представляют остатки в прошлом обширных и систематических массивов данных. Далее, при систематическом сборе тех или иных сведений отдельные показатели могли учитываться лишь частично (именно частично, а не выборочно). Так, при составлении «Экономических примечаний» к Генеральному межеванию второй половины XVIII в., которое охватило большую часть территории страны, ряд показателей (количество населения, площадь земельных угодий и др.) учитывался повсеместно, а некоторые важные данные (о величине барских запашек, размерах оброка) были собраны в силу целого ряда причин лишь частично. Многие сведения вообще собирались только частично. Это прежде всего относится к тем из них, которые не являлись нормативными и сбором которых занимались различные местные органы, научные и общественные организации и отдельные лица.

Итак, области выборочного метода в исторических исследованиях весьма обширны, а задачи, которые следует при этом решать, различны.

Так, при организации выборочного обследования и формировании выборки из имеющихся сплошных данных исследователь располагает определенной свободой маневра для обеспечения репрезентативности выборок. При этом он может опираться на хорошо разработанную в математической статистике теорию, методику и технику получения таких выборок.

При оперировании же данными ранее проведенных выборочных обследований следует проверить, в какой мере они были выполнены в соответствии с требованиями, предъявляемыми к выборочному методу. Для этого надо знать, как было проведено это обследование. Чаще всего это вполне можно сделать.

И совсем иное дело — естественные выборки данных, с которыми очень часто имеет дело историк. Прежде всего необходимо доказать их репрезентативность. Без этого экстраполяция показателей выборок на всю изучаемую совокупность будет необоснованной. Поскольку пока еще нет достаточно надежных методов математической проверки репрезентативности естественных выборок, то решающую роль здесь играет выяснение истории их возникновения и содержательный анализ имеющихся данных.

Виды выборочного изучения. В зависимости от того, как осуществляется отбор элементов совокупности в выборку, различают несколько видов выборочного обследования. Отбор может быть случайным, механическим, типическим и серийным.

Случайным является такой отбор, при котором все элементы генеральной совокупности имеют равную возможность быть отобранными. Другими словами, для каждого элемента генеральной совокупности обеспечена равная вероятность попасть в выборку.

Требование случайности отбора достигается на практике с помощью жребия или таблицы случайных чисел.

При отборе способом жеребьевки все элементы генеральной совокупности предварительно нумеруются и номера их наносятся на карточки. После тщательной перетасовки из пачки любым способом (подряд или в любом другом порядке) выбирается нужное число карточек, соответствующее объему выборки. При этом можно либо откладывать отобранные карточки в сторону (тем самым осуществляется так называемый **бесповторный отбор**), либо, вытащив карточку, записать ее номер и возвратить в пачку, тем самым давая ей возможность появиться в выборке еще раз (**повторный отбор**). При повторном отборе всякий раз после возвращения карточки пачка должна быть тщательно перетасована.

Способ жеребьевки применяется в тех случаях, когда число элементов всей изучаемой совокупности невелико. При большом объеме генеральной совокупности осуществление случайного отбора методом жеребьевки становится сложным. Более надежным и менее трудоемким в случае большого объема обрабатываемых данных является метод использования таблицы случайных чисел.

Таблиц случайных чисел существует несколько, одна из них приведена в приложении (табл. 9). Способ отбора с помощью таблицы случайных чисел рассмотрим на примере.

Пример 1. Пусть совокупность состоит из 900 элементов, а намеченный объем выборки равен 20 единицам.

Из таблицы случайных чисел (см. табл. 9 приложения) отбираем[1] числа, не превосходящие 900, до тех пор, пока не наберем нужных 20 чисел. Получаем:

146	867	505	438	139	653	480	426	765	478	807
	47	220	522	221	835	368	275	424	703	

Выписанные числа будем считать порядковыми номерами тех элементов генеральной совокупности, которые попали в выборку.

Для очень больших совокупностей отбор с помощью таблицы случайных чисел становится трудно осуществимым, так как сложно перенумеровать всю совокупность. Здесь лучше применить механический отбор.

Механический отбор производится следующим образом. Если формируется 10%-ная выборка, т. е. из каждых десяти элементов должен быть отобран один, то вся совокупность условно разбивается на равные части по 10 элементов. Затем из первой десятки выбирается случайным образом элемент. Например, жеребьевка указала девятый номер. Отбор остальных элементов выборки полностью определяется указанной пропорцией отбора и номером первого отобранного элемента. В рассматриваемом случае выборка будет состоять из элементов 9, 19, 29 и т. д.

Механическим отбором следует пользоваться осторож-

[1] Здесь и далее просмотр таблицы начат с первого числа первого столбца, хотя начинать можно с любого числа таблицы.

но, так как существует реальная опасность возникновения так называемых систематических ошибок (см. § 2). Поэтому прежде чем делать механическую выборку, необходимо проанализировать изучаемую совокупность. Если ее элементы расположены случайным образом, то выборка, полученная механическим способом, будет случайной. Однако нередко элементы исходной совокупности бывают частично или даже полностью упорядочены. Весьма нежелательным для механического отбора является порядок элементов, имеющий правильную повторяемость, период которой может совпасть с периодом механической выборки.

Нередко элементы совокупности бывают упорядочены по величине изучаемого признака в убывающем или возрастающем порядке и не имеют периодичности. Механический отбор из такой совокупности приобретает характер **направленного отбора**, так как отдельные части совокупности оказываются представленными в выборке пропорционально их численности во всей совокупности, т. е. отбор **направлен** на то, чтобы сделать выборку представительной.

Механический отбор, как никакой другой, широко использовался в русской и советской статистике.

Большую ценность представляют обследования земских статистиков, которые наряду со сплошным подворным обследованием крестьянских хозяйств по сокращенной «похозяйственной карточке» изучали по расширенной программе определенную часть хозяйств, отобранных механическим способом.

Механический отбор использовался советскими статистиками для учета посевных площадей, численности скота, размеров урожая и многого другого накануне сплошной коллективизации, когда в сельском хозяйстве насчитывалось 25 млн. мелких крестьянских хозяйств (так называемый 10%-ный весенний опрос крестьянских хозяйств и 5%-ный осенний опрос).

Другим видом направленного отбора является **типический отбор.** Следует отличать типический отбор от отбора типичных объектов. Отбор типичных объектов применялся в земской статистике, а также при бюджетных обследованиях. При этом отбор «типичных селений» или «типичных хозяйств» производился по некоторым экономическим признакам, например по размерам землевладения на двор, по роду занятий жителей и т. п. Отбор такого рода не может быть основой для применения выборочного метода, так как здесь не выполнено основное его требование — случайность отбора.

При собственно типическом отборе в выборочном методе совокупность разбивается на группы, однородные в качественном отношении, а затем уже внутри каждой группы производится случайный отбор. Типический отбор организовать сложнее, чем собственно случайный, так как необходимы определенные знания о составе и свойствах генеральной совокупности, но зато он дает более точные результаты.

При серийном отборе вся совокупность разбивается на группы (серии). Затем путем случайного или механического отбора выделяют определенную часть этих серий и производят их сплошную обработку. По сути дела, серийный отбор представляет собой случайный или механический отбор, осуществленный для укрупненных элементов исходной совокупности.

В теоретическом плане серийная выборка является самой несовершенной из рассмотренных. Для обработки материала она, как правило, не используется, но представляет определенные удобства при организации обследования, особенно в изучении сельского хозяйства. Например, ежегодные выборочные обследования крестьянских хозяйств в годы, предшествовавшие коллективизации, проводились способом серийного отбора. Историку полезно знать о серийной выборке, поскольку он может встретиться с результатами таких обследований.

Кроме описанных выше классических способов отбора в практике выборочного метода используются и другие способы. Рассмотрим два из них.

Изучаемая совокупность может иметь многоступенчатую структуру, она может состоять из единиц первой ступени, которые, в свою очередь, состоят из единиц второй ступени, и т. д. Например, губернии включают в себя уезды, уезды можно рассматривать как совокупность волостей, волости состоят из сел, а села — из дворов.

К таким совокупностям можно применять **многоступенчатый отбор,** т. е. последовательно осуществлять отбор на каждой ступени. Так, из совокупности губерний механическим, типическим или случайным способом можно отобрать уезды (первая ступень), затем одним из указанных способов выбрать волости (вторая ступень), далее провести отбор сел (третья ступень) и, наконец, дворов (четвертая ступень).

Примером двухступенчатого механического отбора может служить давно практикуемый отбор бюджетов рабочих.

На первой ступени механически выбираются предприятия, на второй — рабочие, бюджет которых обследуется.

Изменчивость признаков исследуемых объектов может быть различной. Например, обеспеченность крестьянских хозяйств собственной рабочей силой колеблется меньше, чем, скажем, размеры их посевов. В связи с этим меньшая по объему выборка по обеспеченности рабочей силой будет столь же представительной, как и бо́льшая по числу элементов выборка данных о размерах посевов. В этом случае из выборки, по которой определяются размеры посевов, можно сделать подвыборку, достаточно репрезентативную для определения обеспеченности рабочей силой, осуществив тем самым **двухфазный отбор.** В общем случае можно добавить и следующие фазы, т. е. из полученной подвыборки сделать еще подвыборку и т. д. Этот же способ отбора применяется в тех случаях, когда цели исследования требуют различной точности при исчислении разных показателей.

Потребность в многофазном отборе возникла при выборочной обработке материалов профессиональной переписи 1918 года. Как показали исследования, для выявления доли рабочих Ярославской губернии, уходящих на полевые работы, требовалась выборка одного объема, тогда как для изучения общей связи рабочих с землей можно было ограничиться выборкой меньшего объема. Разные объемы выборок потребовались и при изучении групп рабочих различных отраслей промышленности Ярославской губернии. Так, предварительные расчеты показали, что для достаточно надежных выводов по группе рабочих полиграфической промышленности требовалась, по крайней мере, 5%-ная выборка, а для исследования рабочих текстильной, пищевой, металлообрабатывающей и машиностроительной промышленности достаточной оказалась 1%-ная выборка [1].

Изложенные выше способы формирования выборок не исчерпывают собой всех типов отбора, применяемых на практике [2].

§ 2. Стандартные ошибки выборок

Как уже отмечалось, выборочный метод позволяет результаты выборочной обработки материалов переносить на

---

[1] См.: *Соколов А. К.* Методика выборочной обработки первичных материалов профессиональной переписи 1918 г. — История СССР, 1971, № 4.

[2] Наиболее полное описание видов отбора дано в кн.: *Йейтс Ф.* Выборочный метод в переписях и обследованиях. М., 1965.

всю генеральную совокупность. При этом, естественно, имеет место некоторая ошибка, и эффективность выборочного метода заключается в том, что он позволяет оценить эту ошибку.

Ошибки, возникающие при использовании выборочных данных для суждения о всей совокупности, показывают, насколько хорошо характеристики выборки представляют соответствующие характеристики генеральной совокупности, и называются поэтому **ошибками представительности (репрезентативности)**. Различают ошибки представительности двоякого рода: систематические и случайные.

Систематические ошибки возникают в том случае, если не выполнены условия случайности отбора.

Систематическая ошибка может возникнуть и в случае, когда формально отбор произведен случайным образом, но исходная совокупность не является полной и представительной для решения поставленной задачи.

В теории выборочного метода не рассматриваются систематические ошибки, но исследователь должен помнить о возможности их появления и принять меры, обеспечивающие их исключение. С помощью выборочного метода определяются величины ошибок второго рода, т. е. величины случайных ошибок.

Случайные ошибки выборок возникают за счет того, что для анализа всей совокупности используется только часть ее.

Хотя выборочный метод и позволяет обоснованно судить о средней арифметической некоторого количественного признака генеральной совокупности по средней арифметической, исчисленной по выборке, это, однако, не означает, что выборочная средняя совпадает с генеральной средней. Она, как правило, в той или иной степени от нее отличается.

Величина **ошибки выборки** представляет собой разность между генеральной и выборочной средними. Ошибки выборки различны для каждой конкретной выборки и в принципе могут быть обобщенно охарактеризованы с помощью средней из всех таких отдельных ошибок.

В математической статистике получены формулы, которые позволяют приближенно вычислить среднюю ошибку выборки, основываясь на данных только той выборки, которая имеется в распоряжении исследователя. Вычисление средней ошибки выборки зависит от способа отбора элементов из совокупности в выборку.

Средняя ошибка выборки при собственно случайном повторном методе отбора определяется формулой

$$\mu=\sigma/\sqrt{n}, \tag{5.1}$$

где σ — оценка среднего квдаратического отклонения в генеральной совокупности по выборке; n — число элементов в выборке (ее объем) [1].

Как видим, средняя ошибка выборки (ее называют иногда **стандартной ошибкой выборки**) существенно зависит от среднего квадратического отклонения отдельных значений признака от выборочной средней: чем больше среднее квадратическое отклонение, т. е. чем больше разброс значений признака, тем, при прочих равных условиях, больше средняя ошибка выборки. Объем выборки воздействует на среднюю ошибку выборки в обратном направлении: чем больше численность выборки, тем меньше средняя ошибка выборки, что вполне объяснимо, так как большая выборка лучше представляет всю совокупность.

Средняя ошибка выборки при случайном бесповторном отборе находится по формуле

$$\mu=\frac{\sigma}{\sqrt{n}}\sqrt{1-\frac{n}{N}}, \tag{5.2}$$

где N — объем генеральной совокупности.

Формула (5.2) отличается от формулы (5.1) только множителем $\sqrt{1-n/N}$. Множитель всегда меньше единицы, в связи с чем средняя ошибка выборки при бесповторном способе отбора, как правило, бывает меньше средней ошибки повторной выборки того же объема. Это различие становится тем существеннее, чем большую долю генеральной совокупности составляет выборка. Если же отношение n/N мало, то множитель близок к единице и при расчете средней ошибки бесповторной выборки им можно пренебречь. Таким же образом следует поступать и в том случае, когда объем генеральной совокупности неизвестен, с чем историк может нередко столкнуться. Правда, при этом необходимо иметь хотя бы примерное представление о соотношении n и N.

---

[1] На практике величину σ заменяют на среднее квадратическое отклонение выборки по формуле (4.7), но пользоваться этой формулой можно лишь при достаточно большом объеме выборки ($n>30$). Методы расчета средней ошибки для малых выборок изложены в § 4 этой главы.

Рассмотрим расчет средней (стандартной) ошибки выборки на конкретных примерах.

Пример 2. Из 2689 уставных грамот Тамбовской губернии необходимо сделать случайную 10%-ную выборку бесповторным способом и определить средние размеры дореформенного и пореформенного наделов на душу и соответствующие им средние ошибки выборки [1].

Формирование выборки осуществим с помощью таблицы случайных чисел (табл. 9 приложения). Воспользуемся следующим способом, позволяющим рациональнее использовать таблицу случайных чисел. Из чисел от 3001 до 6000 будем вычитать 3000, а из чисел от 6001 до 9000 будем вычитать 6000. Из полученных чисел будем, как указывалось, отбирать те, которые не превосходят 2689. Так, первое число таблицы 5489 дает нам 2489, второе — 3522 дает 522 и т. д. В итоге получаем номера единиц совокупности, попавших в выборку.

Для дальнейшей работы полезно полученные числа расположить в возрастающем порядке. Во-первых, это облегчит отбор уставных грамот с соответствующими порядковыми номерами, во-вторых, выявит повторения, от которых нам нужно избавиться, так как выборка делается бесповторным способом. Исключение повторяющихся чисел приводит к тому, что количество отобранных чисел уменьшается. Обращаясь снова к таблице случайных чисел, доводим объем выборки до нужного размера.

Отобрав соответствующие уставные грамоты (их оказалось 264), переходим к расчету средних арифметических и соответствующих им средних ошибок выборки [2].

Средний дореформенный надел на душу оказался равным 3,16 дес. (суммируем все наделы на душу и делим на число слагаемых — количество грамот):

$$\bar{x}_д = \frac{835,24}{264} = 3,16.$$

Средний пореформенный надел на душу равен 2,71 дес. ($x_п = 2,71$).

Чтобы воспользоваться формулой (5.2) для расчета средней ошибки выборки, необходимо предварительно вычислить средние квадратические отклонения по формуле (4.7):

$$\sigma_д = \sqrt{486,08/264} = \sqrt{1,84} - 1,36, \qquad \sigma_п = 0,56.$$

Пользуясь полученными результатами и учитывая, что $N = 2689$, имеем

[1] Занесенные на специальные бланки материалы уставных грамот были любезно предоставлены авторам Б. Г. Литваком. Комплекс этих материалов, включающих данные о размерах дореформенного и пореформенного наделов, о форме эксплуатации, о величине высшего душевого надела и некоторые другие, возник в связи с отменой крепостного права и определял поземельные отношения крестьян и помещиков.

[2] В этом примере и во всех остальных примерах этой главы, базирующихся на материалах уставных грамот, мы из-за недостатка места не будем давать исходные данные, служащие для расчета выборочных характеристик, и ограничимся приведением результатов проделанных на их основе вычислений.

$$\mu_д = \frac{1,36}{\sqrt{264}} \sqrt{1 - \frac{264}{2689}} = 0,0798, \qquad \mu_п = 0,0328.$$

Поставленная задача полностью решена.

Пример 3. Из тех же 2689 уставных грамот Тамбовской губернии необходимо сделать случайную 10%-ную выборку повторным способом, определить средний размер дореформенного надела на душу по выборке и среднюю ошибку выборки.

Техника подготовительной работы та же, что и в предыдущем примере, только повторно попавшие в выборку грамоты не исключаются. Результаты расчетов среднего размера дореформенного надела и среднего квадратического отклонения выборки по сформированной указанным способом выборке следующие:

$$\bar{x}_д = 3,15, \qquad \sigma_д = 1,37.$$

Для расчета средней ошибки выборки воспользуемся формулой (5.1):

$$\mu_д = 1,37/\sqrt{264} = 0,0846.$$

Итак, средняя ошибка выборки при повторном способе отбора оказалась большей (0,0846), чем при бесповторном (0,0798). Но разница между ними небольшая, так как отношение n к N невелико.

Средняя ошибка выборки при механическом способе отбора вычисляется по формуле случайной бесповторной выборки (5.2) или в случае, когда множителем $\sqrt{1-n/N}$ можно пренебречь, по формуле случайной повторной выборки (5.1).

Пример 4. Генеральная совокупность та же, что и в предыдущих примерах. Необходимо сделать 10%-ную механическую выборку, вычислить средний надел земли на душу до реформы и определить среднюю ошибку выборки.

Случайным образом отбираем в выборку одну уставную грамоту из первых десяти. По жребию выпало число 10. Следовательно, в выборку попадут грамоты с порядковыми номерами 10, 20, 30 и т. д.

Для этой выборки, включающей 263 элемента, средний размер дореформенного надела на душу ($x_д$) равен 2,97 дес., а среднее квадратическое отклонение выборочных данных $\sigma = 1,48$. Воспользовавшись формулой (5.2), определяем среднюю ошибку выборки:

$$\mu_д = \frac{1,48}{\sqrt{263}} \sqrt{1 - \frac{263}{2689}} = 0,0715.$$

Как правило, средняя ошибка выборки при механическом отборе оказывается меньше средней ошибки выборки при собственно случайном отборе.

Средняя ошибка выборки при типическом отборе определяется следующими формулами:

$$\mu=\frac{1}{N}\sqrt{\sum_{i=1}^{k}\frac{\sigma_i^2}{n_i}N_i^2} \qquad (5.3)$$

для повторной выборки и

$$\mu=\frac{1}{N}\sqrt{\sum_{i=1}^{k}\frac{\sigma_i^2}{n_i}N_i^2\left(1-\frac{n_i}{N_i}\right)} \qquad (5.4)$$

для бесповторной выборки, где N — объем генеральной совокупности; N_i — объем i-й типической группы; n_i — объем выборки из i-й типической группы; σ_i — среднее квадратическое отклонение i-й типической группы; k — число типических групп.

Средняя арифметическая типической выборки рассчитывается по формуле

$$\bar{x}=\frac{\sum_{i=1}^{k}\bar{x}_i N_i}{N}, \qquad (5.5)$$

где $\bar{x}_i$ — средняя арифметическая выборки из i-й типической группы; N_i — объем i-й типической группы; N — объем генеральной совокупности.

Для того чтобы сделать типическую выборку, нужно прежде всего решить вопрос о том, каковы должны быть объемы выборки по каждой из выделенных типических групп. В зависимости от исследовательских задач и характера изучаемой совокупности, можно воспользоваться одним из следующих приемов.

Самый простой способ определения объема выборки из каждой типической группы n_i состоит в том, что объем всей намеченной выборки n делят на число типических групп k, т. е.

$$n_i=n/k. \qquad (5.6)$$

Второй, наиболее широко применяемый способ заключается в том, что объемы выборок из групп устанавливаются пропорционально объемам соответствующих типических групп, т. е.

$$\frac{n_1}{N_1}=\frac{n_2}{N_2}=\ldots=\frac{n_k}{N_k}.$$

В итоге для расчетов получается такая формула:

$$n_i = nN_i/N, \tag{5.7}$$

где n_i — объем выборки из i-й типической группы; n — общий объем выборки из генеральной совокупности; N_i — объем i-й типической группы; N — объем генеральной совокупности.

Третий способ состоит в том, что число элементов в выборке для каждой типической группы определяется пропорционально средним квадратическим отклонениям соответствующих типических групп (σ_i), т. е. при определении n_i руководствуются следующим соотношением:

$$n_1/\sigma_1 = n_2/\sigma_2 = \ldots = n_k \sigma_k.$$

Такой прием часто дает ощутимый выигрыш в точности. Сложность его использования состоит в том, что необходимо предварительно знать средние квадратические отклонения признака в типических группах, из которых будет извлекаться выборка. Для этого используются результаты расчетов по аналогичным данным либо делают пробные выборки из каждой группы и их средние квадратические отклонения кладут в основу расчета. Формула для расчета n_i будет такой:

$$n_i = \sigma_i n \Big/ \sum_{i=1}^{k} \sigma_i, \tag{5.8}$$

где σ_i — среднее квадратическое отклонение i-й группы; $\sum_{i=1}^{k} \sigma_i$ — сумма средних квадратических отклонений всех групп; n — объем выборки.

Наконец, четвертый способ образования типической выборки учитывает и размеры типических групп (N_i) и колеблемость признака в этих группах (σ_i); при формировании выборки исходят из того, что

$$\frac{n_1/N_1}{\sigma_1} = \frac{n_2/N_2}{\sigma_2} = \ldots = \frac{n_k/N_k}{\sigma_k}.$$

Формула для расчета n_i четвертым способом такова:

$$n_i = \frac{nN_i\sigma_i}{\sum_{i=1}^{k} N_i\sigma_i}, \tag{5.9}$$

где N_i — объем i-й типической группы; σ_i — среднее квадратическое отклонение i-й группы; n — общий объем выборки из генеральной совокупности; k — число типических групп.

Из указанных четырех способов определения численности выборок из типических групп самым простым, но и самым несовершенным является первый. Несложен для расчетов второй способ. Его целесообразно применять в тех случаях, когда типические группы резко отличаются по объему. Если типические группы имеют примерно одинаковый объем, то лучше формировать выборки с учетом рассеивания признака, т. е. третьим способом. Если, наконец, объемы типических групп различны и заметно отличны их средние квадратические отклонения, то наилучшие результаты достигаются при применении четвертого способа.

Рассмотрим теперь на примерах методику вычисления средних арифметических типических выборок и возникающих при этом стандартных ошибок.

Случайный отбор элементов из типических групп может проводиться двумя способами. Если типические группы в исходных данных разделены и каждая имеет собственную нумерацию, то случайный отбор элементов до нужного объема производится из каждой группы отдельно. Если же элементы типических групп расположены в генеральной совокупности вперемешку, как в нашем случае, то отбор осуществляется из всей совокупности, при этом следят, чтобы объемы отдельных групп не были превышены. Случайные числа, соответствующие элементам тех групп, объемы выборок по которым достигнуты, отбрасываются.

Пример 5. Из совокупности уставных грамот Тамбовской губернии сделать 10%-ную типическую выборку с учетом численности групп. Вычислить средний пореформенный надел на душу и среднюю ошибку выборки.

При знакомстве с уставными грамотами обращает на себя внимание тот факт, что надел земли на душу после реформы тяготеет к высшему душевному наделу. Естественно предположить, что типические группы, образованные с учетом размера высшего душевого надела, будут более однородными, чем вся совокупность в целом.

Разобьем всю совокупность на три группы. К первой группе отнесем селения с размером высшего душевого надела, равным 3,00 дес., ко второй — 3,25 дес., к третьей — 3,50 дес. Объемы групп будут равны соответственно 1717, 445 и 525 [1].

Получены следующие результаты расчетов средних характеристик по каждой из трех групп выборки:

[1] Две грамоты мы не учитываем, так как в одной из них указан высший размер душевого надела, равный 2,0 дес., в другой — 2,75 дес., в результате чего общий объем совокупности составил $N_1+N_2+N_3=$ $=2687$ грамот.

для первой группы (высший душевой надел — 3,00 дес.)

$$\bar{x}_{3,00} = 2,59, \qquad \sigma_{3,00} = \sqrt{0,268} = 0,52;$$

для второй типической группы (высший душевой надел — 3,25 дес.).

$$\bar{x}_{3,25} = 2,82, \qquad \sigma_{3,25} = \sqrt{0,241} = 0,49;$$

для третьей типической группы (высший душевой надел — 3,50 дес.)

$$\bar{x}_{3,50} = 2,93, \qquad \sigma_{3,50} = \sqrt{0,378} = 0,61.$$

Пользуясь соответствующими формулами табл. 2, имеем окончательно:

$$\bar{x} = \frac{2,59 \cdot 172 + 2,82 \cdot 45 + 2,93 \cdot 52}{269} = 2,69;$$

$$\bar{\sigma}^2 = \frac{0,268 \cdot 1717 + 0,241 \cdot 445 + 0,378 \cdot 525}{2687} = 0,28;$$

$$\mu = \sqrt{\frac{0,28}{269} 0,90} = 0,0306.$$

Средняя ошибка выборки, полученная таким способом, оказалась несколько меньше средней ошибки выборки, полученной при случайном отборе. В данном случае различие типических групп невелико. При больших различиях групп выигрыш в точности, даваемый типическим отбором, бывает более существенным.

Пример 6. Определить объемы выборок каждой типической группы так, чтобы они оказались пропорциональными средним квадратическим отклонениям соответствующих групп. Совокупность и общий объем выборки те же, что и в предыдущем примере.

Воспользуемся промежуточными результатами примера 5:

$$\sigma_1 = 0,52; \quad \sigma_2 = 0,49; \quad \sigma_3 = 0,61.$$

Тогда по формуле (5.8) объемы выборок типических групп будут такими:

$$n_1 = \frac{0,52 \cdot 269}{1,62} = 86, \quad n_2 = \frac{0,49 \cdot 269}{1,62} = 81,$$

$$n^3 = 269 - 86 - 81 = 102,$$

т. е. из первой типической группы (высший размер душевого надела равен 3,00 дес.) следует отобрать 86 грамот, из второй типической группы (высший размер душевого надела — 3,25 дес.) — 81 грамоту, из третьей типической группы (высший размер душевого надела — 3,50 дес.) — 102 грамоты.

Пример 7. Генеральная совокупность и критерий, по которому происходит деление на типические группы, те же, что и в предыдущих двух примерах. Сделать типическую 10%-ную выборку, отбирая количество элементов в типических группах пропорционально численности этих групп и средним квадратическим отклонениям.

Рассчитать средний пореформенный надел на душу и среднюю ошибку выборки.

По формуле (5.9) численность выборок из типических групп будет следующей:

$$n_1 = \frac{nN_1\sigma_1}{\sum_{i=1}^{k} N_i\sigma_i} = \frac{269\cdot 1717\cdot 0,52}{0,52\cdot 1717+0,49\cdot 445+0,61\cdot 525} = 168.$$

Аналогично рассчитываются n_2 и n_3: $n_2=41$, $n_3=60$.

Следовательно, из первой типической группы нужно взять 168 грамот, из второй — 41 грамоту, из третьей — 60. Отобрав требуемое количество грамот (техника отбора была изложена выше), переходим к вычислению интересующих нас характеристик.

Результаты расчета средних по группам следующие:

$$\bar{x}_{3,00}=2,56,\quad \bar{x}_{3.25}=2,80,\quad \bar{x}_{3,50}=2,90.$$

Соответствующие им средние квадратические отклонения равны:

$$\sigma_{3,00}=0,52,\quad \sigma_{3,25}=0,50,\quad \sigma_{3,50}=0,61.$$

Средний по всей выборке пореформенный надел на душу равен (по формуле (5.5)):

$$\bar{x} = \frac{2,56\cdot 1717+2,80\cdot 445+2,90\cdot 525}{2687} = 2,67.$$

Для расчета средней ошибки выборки воспользуемся соответствующей формулой из сводной табл. 2:

$$\mu = \frac{1}{N}\sum_{i=1}^{k}\sigma_i N_i \sqrt{\frac{1-n/N}{n}} = \frac{1}{2687}\times$$

$$\times\frac{0,52\cdot 1717+0,50\cdot 445+0,61\cdot 525}{\sqrt{269}}\sqrt{1-\frac{269}{2687}} = 0,0309.$$

Средняя ошибка выборки получилась меньше, чем при случайном методе отбора, но несколько больше соответствующей характеристики, полученной для типической выборки, образованной пропорционально численности типических групп. Последнее произошло, надо полагать, потому, что типические группы по размеру высшего душевого надела отличаются, в основном, по численности и значительно меньше — по разбросу признака.

Сведем воедино итоги рассмотренных примеров, чтобы еще раз сравнить полученные результаты (см. табл. 1).

В целом приведенные примеры подтверждают установленные в статистике общие положения. Важнейшим для применения выборочного метода в исторических исследованиях является то, что наиболее точные результаты дает типический отбор. Стандартная ошибка средней при этом ме-

Таблица 1. Итоги выборочной обработки уставных грамот Тамбовской губернии (10%-ная выборка)

Виды выборок	Крестьянский надел, дес. на мужскую душу					
	дореформенный			пореформенный		
	$\bar{x}$	μ	$\frac{\mu}{\bar{x}}$ 100%	$\bar{x}$	μ	$\frac{\mu}{\bar{x}}$ 100%
Случайная бесповторная	3,16	0,080	2,53	2,71	0,033	1,21
Случайная повторная	3,15	0,085	2,70			
Механическая	2,97	0,072	2,42			
Типическая бесповторная:						
а) с учетом численности групп по размерам высшего надела				2,69	0,031	1,15
б) с учетом численности групп и рассеивания по размерам высшего надела				2,67	0,031	1,16

тоде отбора получается меньшей, чем при случайном и механическом отборе (сравним процентные отношения ошибок к средним арифметическим). При этом следует иметь в виду, что размеры наделов крестьян являются признаком, рассеивание которого является небольшим. При большей неоднородности изучаемых совокупностей данных преимущества типического отбора будут еще очевиднее. Что касается собственно случайного и механического отбора, то они в общем дают близкие результаты. Надо лишь всегда проверять, насколько механический отбор является близким к случайному. Принципиальных различий между бесповторным и повторным случайным отбором нет.

Для удобства пользования формулы выборочного метода, применяемые для вычисления выборочных средних арифметических и их стандартных ошибок при разных видах отбора, сведены в табл. 2. В эту таблицу не вошли формулы для расчета средних ошибок выборок при многоступенчатом способе отбора[1]. Что касается многофазного отбора, то он равносилен взятию выборок различных объемов для разных признаков и ничего нового в вычислительные процедуры не вносит.

---

[1] Эти сведения можно найти в кн.: *Йейтс Ф.* Выборочный метод в переписях и обследованиях.

Таблица 2. Формулы выборочного метода для средней арифметической при различных видах отбора

	Выборочная средняя	Объем выборки из типических групп	Средняя ошибка выборки μ	
			при повторном отборе	при бесповторном отборе
Собственно случайный отбор и механический отбор[1]	$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$		$\frac{\sigma}{\sqrt{n}}$	$\frac{\sigma}{\sqrt{n}} \sqrt{1 - \frac{n}{N}}$
Типический отбор: а) при равных объемах выборки из всех типических групп		$n_i = \frac{n}{k}$		
б) при объемах выборок, пропорциональных средним квадратическим отклонениям типических групп		$n_i = \frac{\sigma_i n}{\sum_{i=1}^{k} \sigma_i}$	$\frac{1}{N} \sqrt{\sum_{i=1}^{k} \frac{\sigma_i^2}{n_i} N_i^2}$	$\frac{1}{N} \sqrt{\sum_{i=1}^{k} \frac{\sigma_i^2}{n_i} N_i^2 \left(1 - \frac{n_i}{N_i}\right)}$
			Эти формулы являются одновременно и общими для всех случаев типического отбора	
в) при объемах выборок, пропорциональных объемам типических групп		$n_i = \frac{n N_i}{N}$	$\sqrt{\frac{\sum_{i=1}^{k} \sigma_i^2 N_i}{nN}}$	$\sqrt{\frac{\left(\sum_{i=1}^{k} \sigma_i^2 N_i\right)\left(1 - \frac{n}{N}\right)}{nN}}$

Продолжение табл. 2

	Выборочная средняя	Объем выборки из типических групп	Средняя ошибка выборки μ	
			при повторном отборе	при бесповторном отборе
г) при объемах выборок, пропорциональных объемам типических групп и их средним квадратическим отклонениям		$n_i = \frac{n N_i \sigma_i}{\sum\limits_{i=1}^{k} N_i \sigma_i}$	$\frac{1}{N} \frac{\sum\limits_{i=1}^{k} \sigma_i N_i}{\sqrt{n}}$	$\frac{1}{N} \sum\limits_{i=1}^{k} \sigma_i N_i \sqrt{\frac{1-\frac{n}{N}}{n}}$

[1] При механическом отборе применяется формула бесповторной выборки, за исключением тех случаев, когда множителем $\sqrt{1-n/N}$ можно пренебречь.

Таблица 3. Формулы выборочного метода для доли признака при различных видах отбора

	Выборочная доля признака	Объем выборки из типических групп	Средняя ошибка выборки μ	
			при повторном отборе	при бесповторном отборе
Собственно случайный отбор и механический отбор [1]	$q = \frac{n_0}{n}$		$\sqrt{\frac{q(1-q)}{n}}$	$\sqrt{\frac{q(1-q)}{n}\left(1-\frac{n}{N}\right)}$
Типический отбор: а) при равных объемах выборки из всех типических групп б) при объемах выборок, пропорциональных средним квадратическим отклонениям типических групп	$q = \frac{\sum\limits_{i=1}^{k} q_i N_i}{N}$	$n_i = \frac{n}{k}$ $n_i = n\frac{\sqrt{q_i(1-q_i)}}{\sum\limits_{i=1}^{k}\sqrt{q_i(1-q_i)}}$	$\frac{1}{N}\sqrt{\sum\limits_{i=1}^{k}\frac{q_i(1-q_i)}{n_i}N_i^2}$ Эти формулы являются одновременно и общими для всех случаев типического отбора	$\frac{1}{N}\sqrt{\sum\limits_{i=1}^{k}\frac{q_i(1-q_i)}{n_i}N_i^2\left(1-\frac{n_i}{N_i}\right)}$
в) при объемах выборок, пропорциональных объемам типических групп	$q = \frac{\sum\limits_{i=1}^{k} q_i n_i}{n}$	$n_i = \frac{n N_i}{N}$	$\sqrt{\frac{q(1-q)}{n}}$	$\sqrt{\frac{q(1-q)}{n}\left(1-\frac{n}{N}\right)}$

Продолжение табл. 3

	Выборочная доля признака	Объем выборки из типических групп	Средняя ошибка выборки μ	
			при повторном отборе	при бесповторном отборе
г) при объемах выборок, пропорциональных объемам типических групп и их средним квадратическим отклонениям	$q=\frac{\sum_{i=1}^{k} q_i N_i}{N}$	$n_i=\frac{nN_i\sqrt{q_i(1-q_i)}}{\sum_{i=1}^{k} N_i\sqrt{q_i(1-q_i)}}$	$\frac{1}{N}\frac{\sum_{i=1}^{k} N_i\sqrt{q_i(1-q_i)}}{\sqrt{n}}$	$\frac{1}{N}\frac{\sum_{i=1}^{k} N_i\sqrt{q_i(1-q_i)}}{\sqrt{n}}\sqrt{1-\frac{n}{N}}$

[1] При механическом отборе применяется формула бесповторной выборки, за исключением тех случаев, когда множителем $\sqrt{1-n/N}$ можно пренебречь.

Средняя ошибка выборки для доли признака. Выборочный метод позволяет оценить не только среднюю арифметическую генеральной совокупности, но и долю некоторого (качественного или количественного) признака во всей совокупности.

Доля признака во всей совокупности (q) вычисляется как отношение числа элементов, обладающих этим признаком (N_0), к числу элементов всей совокупности (N), т. е. $q=N_0/N$.

Отметим, что рассмотренная выше теория и методика применения выборочного метода для расчета средней может быть применена и для расчета доли без каких-либо принципиальных изменений.

Сводка всех формул выборочного метода для доли признака дана в табл. 3.

Пример 8. На основе 10%-ной случайной бесповторной выборки из совокупности уставных грамот Тамбовской губернии вычислить доли селений с системой эксплуатации крестьян; а) оброчной, б) барщинной и в) смешанной, а также соответствующие им средние ошибки выборки.

Из 264 грамот, составивших 10%-ную случайную бесповторную выборку, грамот, описывающих селения с оброчной, барщинной и смешанной системами эксплуатации, оказалось соответственно 51, 197 и 16. Тогда выборочная доля селений с оброчной системой эксплуатации равна $q_{об}=51:264=0{,}19$, выборочные доли селений с барщинной и смешанной системами эксплуатации равны соответственно 0,75 и 0,06.

Воспользовавшись формулой для собственно случайной бесповторной выборки из табл. 3, рассчитаем средние ошибки выборки для доли:

$$\mu_{об}=\sqrt{\frac{q(1-q)}{n}\left(1-\frac{n}{N}\right)}=\sqrt{\frac{0{,}19\cdot 0{,}81}{264}\left(1-\frac{264}{2672}\right)}=$$

$$=0{,}02;$$

$$\mu_{б}=0{,}03;\quad \mu_{см}=0{,}01.$$

§ 3. Точность и надежность выборочного метода: предельные ошибки. Определение объема выборки

Предельная ошибка выборки и доверительный интервал. Средняя ошибка выборки дает некоторое представление об ошибке репрезентативности, т. е. об ошибке, с которой выборочная средняя представляет действительное значение генеральной средней. Именно она показывает, какова будет ошибка в среднем, если из одной и той же генеральной совокупности сделать много выборок одинакового объема. Однако в каждой конкретной выборке ошибка может существенно отличаться от средней ошибки, т. е. нет гарантии, что ошибка, которая действительно была допущена в конкретном выборочном исследовании, не превышает средней ошибки.

Поэтому гораздо полезнее было бы знать те границы, в которых «практически наверняка» находится действительная ошибка, допущенная в данной конкретной выборке. Эти границы (пределы) указываются **предельной ошибкой выборки** (обозначим ее Δ). Предельная ошибка выборки показывает тот предел, которого практически наверняка не превосходит действительная ошибка. Иначе говоря, предельная ошибка Δ показывает действительно допущенную ошибку с избытком, с превышением (возможно, очень значительным) и тем самым гарантирует, что действительная ошибка не превосходит Δ.

Предельная ошибка Δ вычисляется на основе знания средней ошибки μ по формуле

$$\Delta = t\mu, \tag{5.10}$$

где t — величина, вычисляемая по специальной таблице. Обратим внимание на то, что в определении предельной ошибки постоянно употреблялись слова «практически наверняка». Необходимо пояснить смысл понятия «практическая уверенность».

Установленный предел Δ для ошибки выборки лишь указывает, что если из генеральной совокупности сделать много выборок, то для подавляющего большинства из них ошибка выборки не превысит вычисленного нами предела Δ. При этом, правда, могут быть все-таки и такие выборки, у которых ошибка выборки больше Δ, и не исключено, что конкретная выборка входит в их число. Однако можно точно измерить степень уверенности в том, что ошибка конкретной выборки не превысит Δ. Для этого нужно указать долю выборок, у которых ошибка выборки не превосходит Δ. Обозначим эту долю выборок через P, где $0 \leqslant P \leqslant 1$. Чем ближе P к единице, тем больше будет уверенность в том, что ошибка конкретной выборки не превышает Δ.[1] На практике ис-

[1] Читатель, знакомый с понятием вероятности, заметит, что вместо слов «степень уверенности» можно использовать термин «вероятность».

пользуются, например, значения, равные 0,68; 0,95; 0,99 и некоторые другие.

Значением P фактически измеряется надежность результатов выборочного исследования: для значений P, достаточно близких к единице, практически исключается возможность того, что генеральная средняя будет отличаться от вычисленной выборочной средней больше чем на Δ. Со своей стороны Δ указывает точность, гарантируемую заданным уровнем надежности P. Таким образом, предельная ошибка выборки позволяет одновременно и взаимосвязанно указать точность и надежность результатов выборочного исследования.

В математической статистике доказано, что распределение выборочных средних при достаточно больших n подчиняется нормальному закону (см. § 3, гл. 4) со средним значением, равным генеральной средней $\bar{x}_{\text{ген}}$, и средним квадратическим отклонением, равным средней ошибке выборки μ. Значит, для достаточно больших выборок, вероятность P того, что отклонение выборочной средней от генеральной средней не превысит по модулю предельной ошибки, т. е. $|\bar{x}_{\text{ген}}-\bar{x}_{\text{выб}}|\leqslant\Delta$ или $\frac{|\bar{x}_{\text{ген}}-\bar{x}_{\text{выб}}|}{\mu}\leqslant t$, можно найти по табл. 1 приложения (где $\Phi(t)$ соответствует P).

Эта же таблица позволяет решать и обратную задачу: по заданной вероятности P найти величину предельной ошибки Δ, которая соответствует P, другими словами, найти точность, соответствующую данному уровню надежности. Какова, например, предельная ошибка, соответствующая надежности 0,9545? По табл. 1 приложения найдем значение t, соответствующее вероятности $\Phi(t)=0{,}9545$. Оказывается, $t=2$. С вероятностью 0,9545 отклонение выборочной средней от генеральной по модулю не превосходит $\Delta=2\mu$, т. е. не выше двукратной средней ошибки выборки.

Разумеется, всегда желательно обеспечить большую надежность результатов, поэтому надо стараться выбрать P возможно ближе к 1. Однако необходимо учитывать, что с возрастанием надежности увеличивается и t, а значит, и предельная ошибка $\Delta=t\mu$, т. е. падает точность результатов, что может оказаться по тем или иным соображениям недопустимым. Поэтому на практике приходится довольствоваться некоторым компромиссом между противоречивыми требованиями максимальной надежности и максимальной точности. Если такого компромисса достичь не удается и надежность и точность неудовлетворительны, следует сделать вывод, что объем выборки недостаточен и необходимо произвести новую выборку большего объема или же дополнить старую.

Знание предельной ошибки выборки позволяет указать и пределы для генеральной средней. Действительно, поскольку выборочная средняя $\bar{x}_{\text{выб}}$ отличается от генеральной средней $\bar{x}_{\text{ген}}$ (практически наверняка) не более чем на Δ, то

$$|\bar{x}_{\text{ген}} - \bar{x}_{\text{выб}}| \leqslant \Delta$$

или, иначе,

$$\bar{x}_{\text{выб}} - \Delta \leqslant \bar{x}_{\text{ген}} \leqslant \bar{x}_{\text{выб}} + \Delta. \tag{5.11}$$

Таким образом, с помощью вычисления выборочной средней и предельной ошибки выборки можно указать интервал, в котором практически наверняка находится генеральная средняя (так называемый доверительный интервал). При этом всегда указывается надежность P этого результата (то значение P, которое использовалось в вычислении Δ).

Пример 9. Вычислить предельные ошибки выборки по результатам примера 2 § 1 и определить пределы для генеральной средней.

Выборочная средняя для дореформенного надела равна 3,16, средняя ошибка выборки — 0,0798.

Пусть $P=0{,}9545$. Этому значению P по табл. 1 приложения соответствует $t=2$. Пользуясь формулой (5.10), имеем $\Delta=2\cdot 0{,}0798=$ $=0{,}1596\approx 0{,}16$, т. е. предельная ошибка выборки равна приблизительно 0,16.

Переходим к определению пределов. Чтобы вычислить нижний предел, нужно из выборочной средней вычесть предельную ошибку выборки:

$$3{,}16-0{,}16=3{,}00.$$

Верхний предел получаем, прибавив к выборочной средней предельную ошибку:

$$3{,}16+0{,}16=3{,}32.$$

Тогда имеем следующие пределы для генеральной средней $\bar{x}_{\text{ген}}$:

$$3{,}00 \leqslant \bar{x}_{\text{ген}} \leqslant 3{,}32.$$

Результаты можно интерпретировать так: с надежностью (вероятностью) 0,95 генеральная средняя будет не меньше 3,00 дес. и не больше 3,32 дес. Или, другими словами, если выборки повторять много раз, то в 95 случаях из 100 получим, что выборочная средняя будет отстоять от генеральной средней не далее, чем на величину вычисленной нами предельной ошибки, равной 0,16 дес.

Возьмем теперь $P=0{,}9876\approx 0{,}99$. Тогда $t=2{,}5$,

$$\Delta=2{,}5\cdot 0{,}0798\approx 0{,}20$$

и генеральная средняя заключена в следующих пределах:

$$2{,}96 \leqslant \bar{x} \leqslant 3{,}36.$$

Пределы для генеральной средней расширились, но зато увеличилась степень доверия к результатам: уже примерно в 99 случаях из 100 мы не ошибаемся, указывая эти границы для средней.

Как правило, в исторических исследованиях рассмотренный в примере уровень надежности ($P=0{,}95$; $P=0{,}99$) оказывается достаточным.

Порядок вычисления предельной ошибки выборки для доли признака ничем не отличается от вычисления предельной ошибки для средней арифметической.

Определение объема выборки. Вопрос об определении объема выборки является в выборочном методе исходным, ибо всякая выборка имеет заданный объем.

Заметим сразу, что зачастую исследователь лишен возможности решать вопрос об объеме выборки либо в силу ограниченности имеющихся в его распоряжении данных (естественные выборки), либо в силу тех или иных технических причин.

В тех же случаях, когда постановка вопроса об определении объема выборки возможна, его решение производится в следующем порядке.

Прежде всего производится пробная выборка произвольного объема. При этом можно пойти по одному из двух различных путей. Во-первых, можно попытаться сразу угадать нужный объем выборки, основываясь на каких-либо соображениях разумности объема выборки (например, можно попробовать 10%- или 20%-ную выборку). В случае если объем этой выборки окажется недостаточным, можно будет впоследствии дополнить эту выборку до нужного объема.

При втором подходе пробная выборка берется совсем небольшой (как правило, 1% и менее от объема генеральной совокупности). При этом практически следует руководствоваться некоторым компромиссом между требованием достаточной репрезентативности выборки и желанием уменьшить объем предварительных расчетов. На основе этой пробной выборки по приведенной ниже формуле (5.12) определяется необходимый объем окончательной выборки. Далее уже можно делать выборку заданного объема и проводить по ней выборочное исследование.

Анализ пробной выборки начинается с вычисления выборочной средней $\bar{x}_{\text{пр}}$[1]. Исходя из знания величины этой

[1] Полезно вычислить и среднее квадратическое отклонение признака в пробной выборке, чтобы получить представление о величине разброса признака генеральной совокупности.

средней, а также учитывая содержание изучаемой проблемы и конкретные особенности исследования, определяется требуемая точность к оценке генеральной средней (требования к точности задаются с помощью предельной ошибки выборки Δ). Кроме того, задается уровень надежности результатов (требования к надежности задаются с помощью P — степени уверенности в том, что отклонения выборочной средней от генеральной средней не превысят заданной предельной ошибки Δ).

Например, если $\bar{x}_{\text{пр}}=10$, то ясно, что примерно такой же величины будет и генеральная средняя (если разброс признака не слишком велик). Задавшись точностью, скажем, в 5%, определим допустимую предельную ошибку:

$$\Delta=10\cdot5/100=0{,}5.$$

Далее, зададимся уровнем надежности результатов. Выберем, например, $P=0{,}95$.

Заметим, что стремясь к большей точности и надежности результатов, не следует излишествовать в этом направлении, так как может оказаться, что для достижения поставленных требований придется брать выборку объемом во всю совокупность. При этом теряет смысл само применение выборочного метода. Как правило, такие повышенные требования к результатам не оправдываются целями исследования и без ущерба для дела можно остановиться на более умеренных ограничениях. В том же случае, когда высокие требования вытекают из целей исследования и вычисленный объем выборки оказывается порядка объема всей совокупности, следует сделать вывод о том, что в данном случае применение выборочного метода нецелесообразно.

Рассчитав характеристики пробной выборки, переходят к оценке результатов этой выборки. Если используется первый путь исследования (относительно большой пробной выборки), то задав предельную ошибку Δ, следует сравнить ее с предельной ошибкой, вычисленной по пробной выборке $\Delta_{\text{пр}}$ (при одном и том же значении P). Если окажется, что $\Delta_{\text{пр}}\leqslant\Delta$, то пробной выборки вообще достаточно, она может рассматриваться в качестве основной и ее результаты служат результатами всего выборочного исследования. Если же $\Delta_{\text{пр}}>\Delta$, что нередко имеет место при втором пути исследования, то определяют необходимый объем выборки по следующей формуле:

$$n=t^2\sigma^2/\Delta^2, \tag{5.12}$$

где σ^2 — дисперсия признака, вычисленная по пробной выборке; Δ — заданная точность результатов выборочного исследования (заданная предельная ошибка выборки); t — величина, которая находится по табл. 1 приложения исходя из заданной надежности P результатов выборочного исследования.

Заметим, что если пробная выборка мала ($n<30$), то для определения t используется табл. 2 приложения. В ней при определении t учитывается также объем пробной выборки (для нахождения табличного значения t берется объем пробной выборки, предварительно уменьшенный на единицу). Кроме того, в том случае и μ вычисляется так, как указано в § 4 этой главы.

Отметим, что приведенная формула дает общий объем выборки приближенно. Поэтому желательно если есть возможность, еще несколько увеличить объем выборки по сравнению с вычисленным.

Сделав окончательную выборку найденного объема, следует обязательно проверить, совпадает ли ее предельная ошибка с заданной, т. е. удовлетворяются ли заданные требования к точности и надежности результатов. В том редком случае, когда окажется, что действительная предельная ошибка существенно больше заданной (это может произойти из-за нерепрезентативности пробной выборки), придется еще раз повторить процедуру определения объема выборки уже на основе полученных более полных и точных данных.

Приведем также формулу для нахождения необходимого объема выборки при определении доли признака:

$$n=t^2q(1-q)/\Delta^2, \qquad (5.13)$$

где t и Δ имеют тот же смысл, что и в предыдущей формуле, а q — доля признака в пробной выборке.

Рассмотрим пример, поясняющий основные моменты решения задачи об определении объема выборки.

Пример 10. Воспользовавшись данными по предприятиям европейской России за 1879 г.[1], определить объемы выборок, необходимые для расчетов средней стоимости произведенной продукции в расчете на одного рабочего на предприятиях: а) с паровыми двигателями, б) без паровых двигателей.

Сделаем сначала пробную 1%-ную выборку (случайным бесповторным способом)[2]. Отметим, что среди предприятий, попавших в вы-

---

[1] См.: Указатель фабрик и заводов европейской России/Сост. П. А. Орлов. Спб., 1881, вып. 1. В «Указателе» содержатся сведения по фабрикам и заводам со стоимостью производимой продукции свыше 2 тыс. руб. (всего около 12 000 предприятий).

[2] Из-за недостатка места выборочные данные не приведены.

борку (128 предприятий), 87 составляют предприятия без паровых двигателей и 41 — с паровыми двигателями.

Пользуясь выборочными данными, вычисляем по каждому типу предприятий среднюю стоимость произведенной на одного рабочего продукции x, среднее квадратическое отклонение σ, среднюю ошибку выборки μ и предельную ошибку выборки Δ при уровне надежности $P=0{,}95$.

Для предприятий без паровых двигателей получаем соответственно $x_1=1{,}4$ (тыс. руб.); $\sigma_1=1{,}26$; $\mu_1=0{,}14$; $\Delta_1=2\cdot 0{,}14\approx 0{,}3$.

Для предприятий с паровыми двигателями получим $x_2=2{,}4$ (тыс. руб.); $\sigma_2=3{,}20$; $\mu_2=0{,}50$; $\Delta_2=1{,}0$. Отсюда для генеральных средних вычисляются следующие пределы:

для предприятий без паровых двигателей

$$1{,}1\leqslant \bar{x}_{\mathrm{I}}\leqslant 1{,}7;$$

для предприятий с паровыми двигателями

$$1{,}4\leqslant \bar{x}_{\mathrm{II}}\leqslant 3{,}4.$$

При сопоставлении полученных результатов напрашиваются следующие выводы: средняя стоимость продукции на одного рабочего на предприятиях без паровых двигателей и на предприятиях с паровыми двигателями различна, причем на предприятиях с паровыми двигателями она заметно выше. Однако, строго говоря, такой вывод пока еще неправомерен и может рассматриваться лишь как гипотеза. Дело в том, что доверительные интервалы для генеральных средних по предприятиям без паровых двигателей (1,1; 1,7) и по предприятиям с паровыми двигателями (1,4; 3,4) пересекаются, так что средние генеральные вполне могут совпадать или даже находиться в соотношении, противоположном высказанной гипотезе.

Нетрудно заметить, что указанная неопределенность результатов получается главным образом в силу того, что предельная ошибка выборки по предприятиям с паровыми двигателями Δ_2 слишком велика. В самом деле, различие между выборочными средними по двум типам предприятий составляет

$$|\bar{x}_1-\bar{x}_2|=|1{,}4-2{,}4|=1{,}0.$$

Поэтому, чтобы попытаться подтвердить и обосновать высказанную выше гипотезу, достаточно, чтобы предельные ошибки выборок для обеих групп предприятий (Δ_1, Δ_2) не превышали половины этой разности, т. е. 0,5, тогда доверительные интервалы не будут пересекаться.

Отметим, что предельная ошибка выборки по предприятиям без паровых двигателей $\Delta=0{,}3$ вполне удовлетвори тельна. Чтобы обеспечить предельную ошибку выборки, равную 0,5, для другой группы предприятий, рассчитаем необходимый объем выборки из совокупности предприятий с паровыми двигателями. Выбирая t по табл. 1 приложения, соответствующие значению $P=0{,}9545$, и пользуясь формулой (5.12), получим

$$n=2^2\,(3{,}20)^2/(0{,}5)^2\approx 164.$$

Дополнив теперь выборку из группы предприятий с паровыми двигателями до рассчитанного объема, получим новые значения средней, среднего квадратического отклонения, средней и предельной ошибок выборки:

$$\overline{x}_2'=2{,}42;\ \sigma_2'=3{,}3;\ \mu_2'=0{,}26;\ \Delta_2'=0{,}52.$$

Сравним интервалы для генеральных средних. Для предприятий без паровых двигателей используем результат пробной выборки (которая оказалась для этой группы предприятий и окончательной):

$$1{,}1\leqslant \overline{x}_{\mathrm{I}}\leqslant 1{,}7.$$

А для предприятий с паровыми двигателями имеем после увеличения объема выборки

$$1{,}9\leqslant \overline{x}_{\mathrm{II}}\leqslant 2{,}94.$$

Как видим, теперь доверительные интервалы действительно не пересекаются и высказанная выше гипотеза о том, что средняя стоимость продукции на одного рабочего существенно больше для предприятий с паровыми двигателями, получает убедительное и надежное подтверждение. Другими словами, данные, использованные в примере 10, свидетельствуют о том, что внедрение машин повышало производительность труда.

Интересно отметить, что для достижения нужной точности и надежности результатов из совокупности предприятий с паровыми двигателями нам пришлось сделать примерно в два раза большую выборку, чем из группы предприятий без паровых двигателей. Это объясняется тем, что для предприятий с паровыми двигателями существенно больше разброс изучаемого признака, что вполне естественно для прогрессивной технологии, применяемой на этих предприятиях.

В заключение отметим еще один поучительный факт, с которым мы столкнулись в рассмотренном примере. По предприятиям с паровыми двигателями первоначальный объем выборки составлял 41 единицу, при этом обеспечивалась точность, определяемая предельной ошибкой выборки, равная единице ($\Delta_2=1$). Такая точность, как оказалось, была недостаточной, потребовалась в два раза большая точность — $\Delta_2'=0{,}5$. Это привело к тому, что объем новой выборки составил 164 единицы, что в четыре раза больше первоначального.

Следовательно, необходимый объем выборки растет пропорционально квадрату требуемой точности, что следует прямо из формулы (5.12). А так как квадраты чисел при возрастании самих чисел возрастают очень быстро, то повышенные требования к точности могут привести к неумеренному росту объема выборки. Поэтому важно, чтобы требования к точности выборочного исследования всегда диктовались целями и содержанием исследования. В рассмотренном примере такой целью было обоснование содержательной научной гипотезы.

§ 4. Малые выборки

Рассмотренные выше приемы расчета ошибок выборки основаны на доказанном в математике факте нормальности распределения выборочных средних. Однако этот факт имеет место только при достаточно большом объеме выборки n. Если пользоваться изложенными приемами при n меньшем 20, могут возникнуть грубые ошибки.

Выборки, объем которых меньше 20—30 единиц совокупности, будем называть **малыми**[1]. Для расчета ошибок таких выборок используется несколько иной математический аппарат.

Средняя ошибка малой выборки вычисляется по формуле

$$\mu=S/\sqrt{n}, \tag{5.14}$$

где S — оценка среднего квадратического отклонения в генеральной совокупности по малой выборке. Она равна:

---

[1] Четкой границы между большой и малой выборками в общем случае указать невозможно. Выборка, сделанная из совокупности с небольшим разбросом признака, может считаться большой, тогда как выборка такого же объема, произведенная из более разнородной совокупности, окажется малой. Вопрос о том, к какой категории отнести выборку, решается в каждом конкретном случае.

$$S=\sqrt{\frac{\sum_{i=1}^{k}(x_i-\bar{x})^2 n_i}{n-1}}=\sigma\sqrt{\frac{n}{n-1}}, \qquad (5.15)$$

где σ вычисляется по формуле (4.7); n — объем выборки; k — число вариант, т. е. S несколько отличается от оценки среднего квадратического отклонения в генеральной совокупности по большой выборке, см. (5.1).

Пример 11. В табл. 4 приведены данные о размерах оброка в конце XVIII в. (в руб. серебром на муж. душу). Первая выборка состоит из 16 уездов нечерноземной полосы, вторая выборка — из 16 уездов черноземной полосы. Перед нами две «естественные выборки», которые можно рассматривать как случайные, т. е. репрезентативные. Требуется рассчитать выборочные средние и средние ошибки выборок.

Вычисляем последовательно средние арифметические, средние квадратические отклонения малых выборок, и, наконец, стандартные ошибки выборок. Получаем:

для нечерноземной полосы

$$n_1=16,\quad \bar{x}_1=3{,}0,\quad S_1=\sqrt{23{,}76/15}=\sqrt{1{,}58}=1{,}26,\quad \mu_1=0{,}3;$$

для черноземной полосы

$$n_2=16,\quad \bar{x}_2=4{,}6,\quad S_2=\sqrt{43{,}49/15}=\sqrt{2{,}90}=1{,}70,\quad \mu_2=0{,}4.$$

Заметим, что в пределах интересующей нас точности вычислений поправка на малую выборку не изменила вели-

Таблица 4. Средние размеры оброка (руб. серебром на муж. душу) в конце XVIII в. по 16 уездам черноземной и 16 уездам нечерноземной полосы

Уезды нечерноземной полосы	Размер оброка	Уезды черноземной полосы	Размер оброка
Валдайский	2,2	Спасский	5,1
Сычевский	2,2	Михайловский	2,6
Гжатский	2,0	Ряжский	5,1
Юхновский	2,0	Епифанский	2,9
Ветлужский	3,3	Ефремовский	5,0
Костромской	4,0	Орловский	4,5
Нерехтский	3,0	Кромский	3,6
Дмитровский	2,0	Трубчевский	5,7
Клинский	2,1	Севский	4,1
Боровский	2,5	Темниковский	3,2
Тарусский	2,6	Шацкий	5,3
Мосальский	2,4	Спасский	7,3
Семеновский	3,0	Моршанский	6,8
Сергачский	3,0	Козловский	2,4
Егорьевский	7,0	Тамбовский	7,3
Касимовский	3,9	Острогожский	2,0

чины стандартной ошибки. Заметное различие появляется при вычислении предельной ошибки выборки.

Предельная ошибка малой выборки вычисляется по формуле

$$\Delta = t(k)\mu, \qquad (5.16)$$

где t рассчитывают исходя из так называемого закона распределения Стьюдента с k степенями свободы (в отличие от больших выборок, где t вычисляется на основе нормального закона распределения).

Связь между t и вероятностью (уровнем надежности) P в распределении Стьюдента сложнее, чем в нормальном распределении и опосредствуется через объем выборки. При возрастании объема выборки распределение Стьюдента приближается к нормальному, практически с ним совпадая при достаточно больших n.

При вычислении предельной ошибки малой выборки значение $t(k)$ определяется по таблице распределения Стьюдента с k степенями свободы (табл. 2 приложения), с учетом заданного уровня надежности P и объема выборки (для подстановки в таблицу фактический объем выборки надо предварительно уменьшить на единицу: $k=n-1$).

Пример 12. Используя данные предыдущего примера, найти предельные ошибки выборки для средних размеров оброка с уровнем надежности $P=0,9$ и $P=0,95$ и определить границы для генеральной средней.

Обращаясь к табл. 2 приложения и учитывая, что при объеме выборки, равном 16, k, используемое для нахождения табличного значения t, равно $16-1=15$, а заданный уровень надежности — 0,9, находим $t\ (15)=1,75$.

Тогда предельная ошибка выборки для среднего размера оброка нечерноземной полосы по формуле (5.16) будет равна

$$\Delta_1 = 1,75 \cdot 0,3 \approx 0,5.$$

Следовательно, границы генеральной средней таковы:

$$2,5 < \overline{x}_I < 3,5,$$

т. е. с вероятностью 0,9 средний размер оброка в нечерноземной полосе не выйдет за указанные границы.

Предельная ошибка второй выборки (для размеров оброка в черноземной полосе) и границы генеральной средней находятся аналогично. Имеем:

$$\Delta_2 = 1,75 \cdot 0,4 \approx 0,7,$$

$$3,9 < \overline{x}_{II} < 5,3.$$

Чтобы получить более достоверные результаты, возьмем большую вероятность (уровень надежности). Пусть $P=0{,}95$, тогда из табл. 2 приложения найдем $t\ (15)=2{,}13$, и для нечерноземной полосы

$$\Delta_1=2{,}13\cdot 0{,}3\approx 0{,}6,\quad 2{,}4<\bar{x}_I<3{,}6,$$

для черноземной полосы

$$\Delta_2=2{,}13\cdot 0{,}4\approx 0{,}9,\quad 3{,}7<\bar{x}_{II}<5{,}5.$$

Итак, в конце XVIII в. средний размер оброка в черноземной полосе выше, чем средний размер оброка в нечерноземной полосе. Важно, что границы, в которых заключены средние, не пересекаются. Это свидетельствует о том, что различие размеров оброка в двух районах имело не случайный, а закономерный характер.

Для более строгих выводов о существенности различия между двумя выборочными средними есть специальные методы, изложенные в гл. 9 (§ 2 — критерии для средних, § 3 — критерии для дисперсий). Так, если имеются две выборочные средние $\bar{x}_1$ и $\bar{x}_2$, относящиеся к двум различным совокупностям, причем $\bar{x}_1\neq\bar{x}_2$, то можно предположить, что и генеральные средние этих совокупностей различны. Специальный критерий, основанный на распределении Стьюдента, позволяет для фиксированного уровня надежности P и числа степеней свободы $k=n_1+n_2-2$ сделать вывод о значимости или незначимости различия между выборочными средними. В § 2 гл. 9 на данных примера 11 выясняется, что полученное различие между средними размерами оброка у крестьян черноземной и нечерноземной полосы в конце XVIII в. является значимым. Заметим, что проверяя гипотезу о существенности различия средних, пользуются предположением о том, что разброс признака в обеих совокупностях примерно одинаков. Это предположение также можно проверить (см. гл. 9, § 3, пример 11).

Отметим, что в тех же разделах гл. 9 рассмотрены аналогичные критерии для больших выборок, которые вместо распределения Стьюдента используют нормальное распределение, поскольку при возрастании объема выборки распределение Стьюдента стремится к нормальному.

В заключение скажем несколько слов о больших и малых выборках. Различать большие и малые выборки необходимо, но точной границы между ними установить нельзя. Важно иметь в виду, что к большим выборкам можно применять аппарат теории малых выборок, тогда как обратное приводит к значительным ошибкам. В сомнительных случаях для получения надежных результатов рекомендуется пользоваться аппаратом малых выборок.

В больших выборках средние теснее группируются около генеральной средней, что позволяет получать более точные и надежные результаты, тогда как в малых выборках приходится довольствоваться более широкими границами для средних или меньшей достоверностью результатов. Тем не менее теория малых выборок нашла в практике широкое распространение и применяется даже в тех случаях, когда во власти исследователя сделать выборку большой[1].

Историку обычно не приходится выбирать между формированием большой или же малой выборки, поскольку он часто имеет дело с естественными малыми выборками, число которых он не может изменить, т. е. он стоит перед альтернативой: либо воспользоваться данными малой выборки для анализа исследуемых явлений, либо отказаться от такого анализа. Обработка этих выборок методами математической статистики позволяет в ряде случаев (когда само использование выборочного метода возможно) обоснованно решить вопрос о правомерности или неправомерности тех или иных выводов и заключений на основе имеющихся материалов. И в том и в другом случае исследование приобретает более объективный и глубокий характер, нежели при традиционных методах.

Для того чтобы применить выборочный метод к естественным выборкам, необходимо доказать тем или иным способом случайность образования имеющейся выборки. В проверке случайности выборки ведущая роль принадлежит традиционным методам содержательного источниковедческого анализа. Отсутствие преднамеренности в порядке сбора и хранения тех сведений, след от которых остался в виде естественной выборки, свидетельствует о случайности последней. Математические методы позволяют дополнить этот анализ (см. гл. 9).

И наконец, последнее замечание. В этой главе мы ограничились оценкой средней арифметической генеральной совокупности с помощью характеристик, вычисленных по выборке. Но выборочный метод позволяет решать и более сложные вопросы анализа совокупностей. В частности, по выборке можно судить о наличии или об отсутствии связи между признаками, о форме связи. К процедурам выборочного метода мы будем обращаться при необходимости в соответствующих главах.

---

[1] См., например: *Дружинин Н. К.* Выборочный метод и его применение в социально-экономических исследованиях. М., 1970, с. 77.

ГЛАВА 6
АНАЛИЗ ВЗАИМОСВЯЗЕЙ

§ 1. Общие понятия об изучении связей

Анализ взаимосвязей, присущих изучаемым процессам и явлениям, является важнейшей задачей исторических, как впрочем и многих других, исследований. В тех случаях, когда речь идет о явлениях и процессах, обладающих сложной структурой и многообразием свойственных им связей, такой анализ представляет собой сложную задачу. Прежде всего необходимо установить наличие взаимосвязей и их характер. Вслед за этим возникает вопрос о тесноте взаимосвязей и степени воздействия различных факторов (причин) на интересующий исследователя результат. Если черты и свойства изучаемых объектов могут быть измерены и выражены количественно, то анализ взаимосвязей может вестись на основе применения математических методов. Использование этих методов позволяет проверить гипотезу о наличии или отсутствии взаимосвязей между теми или иными признаками, выдвигаемую на основе содержательного анализа. Далее, лишь посредством математических методов можно установить тесноту и характер взаимосвязей или выявить силу (степень) воздействия различных факторов на результат.

Наиболее разработанными в математической статистике методами анализа взаимосвязей являются **корреляционный** и **регрессионный анализ.** Но прежде чем переходить к их характеристике, остановимся на вопросе о характере и форме тех взаимосвязей, которые присущи объективным явлениям природы и общества.

Функциональные зависимости. Функциональная зависимость двух количественных признаков или переменных состоит в том, что каждому значению одной переменной всегда соответствует одно определенное значение другой переменной. Например, при строительстве железных дорог на километр пути приходится вполне определенное количество уложенных рельсов. Поэтому рассматривая статистические данные по таким количественным признакам: y — длина уложенного железнодорожного пути (в км), x — количество истраченного на строительстве рельсового проката (в тоннах), мы будем иметь дело с функциональной зависимостью определенного вида. Рассмотрим эту зависимость подробнее на условном примере:

x	0	1	2	5	1,5	3
y	0	100	200	500	150	300

Соотношение между признаками, отображенное в таблице, удобно представить в наглядной графической форме, рассматривая числовые данные как координаты точек в прямоугольной системе координат (рис. 8).

Графическим изображением анализируемой зависимости (полученным путем соединения непрерывной линией точек, соответствующих данным таблицы) служит прямая линия. Такая зависимость называется прямой пропорциональной зависимостью. Ее аналитическим выражением является уравнение $y=kx$, где k — коэффициент пропорциональности (в нашем случае $k=100$). Прямая пропорциональная зависимость представляет собой частный случай линейной зависимости, которая характеризуется уравнением

$$y=kx+b. \tag{6.1}$$

Графическим изображением **линейной зависимости** также служит прямая линия (рис. 9).

Линейная зависимость является наиболее простой и в определенном смысле универсальной формой связи многих явлений. Ее универсальность состоит в том, что более сложные зависимости часто можно рассматривать «в первом приближении» как линейные. Здесь мы подходим к выяснению роли функциональных зависимостей в анализе стати-

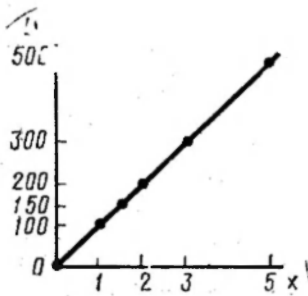

Рис. 8. График зависимости длины уложенного железнодорожного пути (y) и количества используемого рельсового проката (x)

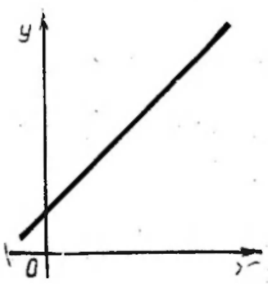

Рис. 9. Вид линейной зависимости

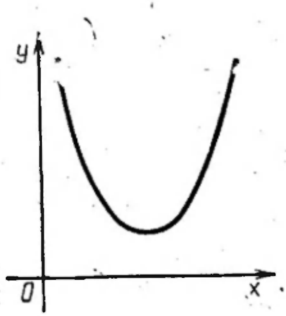

Рис. 10. Вид параболической зависимости

стических связей. Непосредственно функциональные зависимости в чистом виде редко встречаются в общественных явлениях. Связи обычно носят гораздо более сложный характер. Однако их описание во всей сложности часто затруднительно, да и нецелесообразно. Поэтому их рассматривают как соответствующие тем или иным видам функциональной зависимости. Простейшей формой функциональной связи является линейная зависимость, которая широко используется в регрессионном и особенно в корреляционном анализе. Гипотеза о линейной связи между исследуемыми признаками получила широкое распространение в анализе взаимосвязей. Лишь в том случае, если результаты применения гипотезы о линейной зависимости оказываются неудачными или имеются веские основания против линейной связи, используют более сложные функциональные зависимости.

Отметим наиболее употребительные формы функциональной зависимости, применяемые в статистическом анализе.

В случае если прямая линия не соответствует характеру используемых данных, можно использовать **параболу** (рис. 10). Аналитическое выражение ее имеет вид:

$$y=a_0+a_1x+a_2x^2. \qquad (6.2)$$

Наличие в этом уравнении члена a_2x^2 является простейшей формой учета нелинейности.

В том случае, когда мы имеем дело с затуханием роста или падения, удобно использовать гиперболические либо логарифмические зависимости (рис. 11 и 12). Математические выражения для гиперболической и логарифмической зависимостей выглядят так:

$$y=k/x; \quad y=a\lg x.$$

Процессы демографического и экономического роста описываются экспоненциальными зависимостями вида:

$$y=ke^{\lambda x}.$$

Графическое изображение этой зависимости $(k>0,\ \lambda>0)$ иллюстрируется рис. 13.

Подбор подходящей функциональной зависимости на основе графического и логического анализа является важным этапом исследования взаимосвязей, особенно в тех случаях, когда линейная связь оказалась неприемлемой.

Статистические (корреляционные) зависимости. Функциональная зависимость между признаками предполагает их изолированность, она действует, так сказать, «при про-

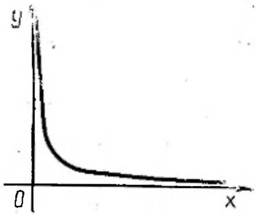

Рис. 11. Вид гиперболической зависимости

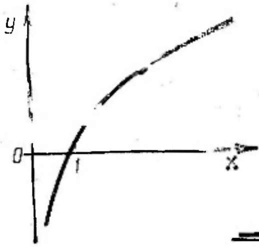

Рис. 12. Вид логарифмической зависимости

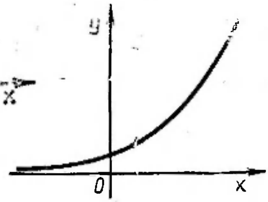

Рис. 13. Вид экспоненциальной зависимости

чих равных условиях». В общественной жизни такие ситуации бывают крайне редко. Как правило, воздействие одной переменной (причины) на другую не изолировано от остальных факторов, а происходит таким образом, что на изучаемую связь прямо или косвенно влияют многие другие факторы. Здесь налицо зависимость особого вида. Для описания и изучения такого рода зависимостей в науке используется понятие **статистический**, или **корреляционной, связи.**

В отличие от функциональной зависимости, когда каждому значению одного признака всегда соответствует определенное значение другого, при статистической зависимости одному и тому же значению одного признака могут соответствовать различные значения другого. Это происходит в силу того, что при статистической зависимости связь устанавливается между признаками (двумя, тремя и т. д.), которые изменяются не только в силу взаимодействия между собой, но и под воздействием множества различных неучтенных факторов. В результате множественного воздействия взаимно переплетающихся факторов связь между признаками существует и проявляется не в каждом отдельном случае, как при функциональной связи, а только в тенденции, «в среднем». Поэтому здесь установить наличие взаимосвязи и определить ее количественную меру можно не на основе единичных наблюдений, а лишь применительно к определенной совокупности объектов, т. е. в среднем по отношению к тем или иным **массовым** объектам или явлениям. Характеризующие эти объекты количественные показатели в источниковедении и в статистике называются **массовыми данными.**

Задачи анализа статистических связей. Анализ статистической, или корреляционной, связи предполагает выявление формы связи, а также оценку тесноты связи. Первая задача решается методами регрессионного анализа, вторая — методами корреляционного анализа. Регрессионный анализ сводится к описанию статистической связи с помощью подходящей функциональной зависимости. Корреляционный анализ позволяет оценивать тесноту связи посредством специальных показателей, причем выбор их зависит от вида функциональной зависимости, пригодной для адекватного описания рассматриваемой статистической взаимосвязи. Как указывалось, наиболее распространенной в изучении связей является гипотеза о линейной зависимости. Соответствующие ей методы корреляционного и регрессионного анализа наиболее полно разработаны в математической статистике. Прежде чем перейти к изложению этих методов, остановимся на двух общих вопросах, относящихся к корреляционному и регрессионному анализу.

Один из важных вопросов, возникающих в изучении связей,— установление «направления» зависимости. Пусть для простоты рассматривается связь между двумя признаками y и x. Какой из этих признаков следует считать подверженным влиянию, или **результативным** (зависимой переменной), какой — оказывающим влияние, или **факторным** (независимой переменной)?

Первостепенное значение в решении этого вопроса имеет содержательный анализ. Положим, мы рассматриваем связь между производительностью труда рабочих и стажем их работы. По-видимому, результативным признаком следует признать производительность труда, а факторным — стаж рабочего. Не всегда «направление» связи проявляется столь очевидно. Тогда при решении вопроса о выборе результативного признака на первый план выступает постановка содержательной проблемы, для исследования которой используется изучение взаимосвязей. Например, устанавливая «направление» связи между такими признаками, как доходность предприятий и их энерговооруженность, мы должны исходить из того, что же мы хотим установить в действительности: влияние внедрения новой техники и технологии на доходность предприятий или же потенциальные возможности предприятий в овладении передовой техникой и технологией. В первом случае результативным признаком естественно считать доходность, во втором — энерговооруженность.

Далее, корреляционный и регрессионный анализ могут применяться и дают вполне корректные результаты при соблюдении определенных условий. Это однородность исходных данных, независимость отдельных значений признака друг от друга и нормальность распределения изучаемых признаков. Некоторые способы проверки однородности, случайности и нормальности распределении данных указаны в гл. 9.

Условия применения корреляционного и регрессионного анализа, а также возникающие на практике нарушения условий и следствия этих нарушений рассмотрены в § 2 и 3 данной главы.

§ 2. Линейная корреляция

Одной из основных мер связи в корреляционном анализе является линейный коэффициент корреляции.

Парный линейный коэффициент корреляции. С помощью парного линейного коэффициента корреляции измеряется теснота связи между двумя признаками. Линейный коэффициент корреляции чаще всего рассчитывается по формуле

$$r_{yx}=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{i=1}^{n}(y_i-\bar{y})^2}}, \quad (6.3)$$

где x_i и y_i — значения признаков x и y соответственно для i-го объекта, $i=1, ..., n$; n — число объектов; $\bar{x}$ и $\bar{y}$ — средние арифметические значения признаков x и y соответственно.

Линейный коэффициент корреляции изменяется в пределах от -1 до $+1$. Равенство коэффициента нулю свидетельствует об отсутствии **линейной связи.** Равенство коэффициента -1 или $+1$ показывает наличие функциональной связи. Знак «+» указывает на **связь прямую** (увеличение или уменьшение одного признака сопровождается аналогичным изменением другого признака), знак «—» — на **связь обратную** (увеличение или уменьшение одного признака сопровождается противоположным по направлению изменением другого признака).

Пример 1. Используя данные табл. 1, рассчитаем коэффициент корреляции между доходом крестьянского хозяйства и стоимостью имеющегося в этом хозяйстве скота.

Воспользуемся формулой (6.3), а промежуточные расчеты сделаем с помощью рабочей табл. 1, в которую введем необходимые для вычислений графы:

$$r_{yx} = 3\,823\,355/\sqrt{27\,571\,129 \cdot 1\,393\,057} = 0{,}62.$$

Как оценить полученное значение коэффициента корреляции? Велико оно или мало? О наличии или об отсутствии связи свидетельствует? Для ответа на эти вопросы проводят проверку значимости коэффициента.

Проверка значимости парного линейного коэффициента корреляции. Коэффициенты корреляции, как правило, рассчитываются для выборочных данных. Чтобы распространить полученные частные результаты на генеральную совокупность, приходится допустить некоторую ошибку, которую можно оценить с помощью средней квадратической ошибки. **Средняя квадратическая ошибка** для парного линейного коэффициента корреляции достаточно большой выборки вычисляется по формуле

$$\sigma_r = (1 - \rho^2)/\sqrt{n}, \tag{6.4}$$

где ρ — коэффициент корреляции генеральной совокупности; n — объем выборки.

В математической статистике доказано, что если признаки x и y распределены по нормальному закону, то в достаточно больших выборках коэффициенты корреляции можно считать распределенными нормально со средним значением ρ и средним квадратическим отклонением σ_r (нормальное распределение рассмотрено в гл. 4). Этот факт используется для построения доверительных интервалов коэффициента корреляции в генеральной совокупности, а также для проверки значимости выборочных коэффициентов корреляции, т. е. для проверки того, могло ли данное значение r получиться в выборке из некоррелированной генеральной совокупности ($\rho=0$) в силу простой случайности.

Очевидно, чем больше отклонение r от ρ, тем менее вероятно, что оно случайно. Так, для нормального распределения вероятность того, что выборочное значение r отличается от ρ больше чем на $3\sigma_r$, меньше 0,01, т. е. величина нормированного отклонения $t = \left|\frac{r-\rho}{\sigma_r}\right|$ может превысить значение 3 лишь в одном случае из ста. Таким образом, для проверки значимости r, задаваясь значением вероятности P (например, $P=0{,}99$), находят по табл. 1 приложения крити-

Таблица 1. Расчет коэффициента корреляции между доходом в крестьянском хозяйстве (y) и стоимостью скота в нем (x) [1]

	y	x_1	x_2	$y-\bar{y}$	$x-\bar{x}$	$(y-\bar{y})^2$	$(x-\bar{x})^2$	$(x-\bar{x})\times$ $\times(y-\bar{y})$
	1	2	3	4	5	6	7	8
1	3612	178	2,1	2147	—109	4 609 609	11 881	—234 023
2	648	353	10,0	—817	66	667 489	4 356	—53 922
3	1786	281	22,0	321	—6	103 041	36	—1 926
4	1945	295	33,0	480	8	230 400	64	3 840
5	803	51	5,7	—662	—236	438 244	55 696	156 232
6	513	6	6,7	—952	—281	906 304	78 961	267 512
7	828	189	12,0	—637	—98	405 769	9 604	62 426
8	706	237	12,0	—759	—50	576 081	2 500	37 950
9	1893	413	30,0	428	126	183 184	15 876	53 928
10	820	263	8,8	—645	—24	416 025	576	15 480
11	2035	474	19,0	570	187	324 900	34 969	106 590
12	841	126	5,9	—624	—161	389 376	25 921	100 464
13	1723	131	7,1	258	—156	66 564	24 336	—40 248
14	2531	131	4,4	1066	—156	1 136 356	24 336	—166 296
15	1461	158	5,3	—4	—129	16	16 641	516
16	2329	744	24,0	864	457	746 496	208 849	394 848
17	3423	525	23,0	1958	238	3 833 764	56 644	466 004
18	1309	146	10,0	—156	—141	24 336	19 881	21 996
19	1731	405	13,0	266	118	70 756	13 924	31 388
20	873	161	16,0	—592	—126	350 464	15 876	74 592
21	1437	570	21,0	—28	283	784	80 089	—7 924
22	724	287	5,9	—741	0	549 081	0	0
23	647	6	7,2	—818	—281	669 124	78 961	229 858
24	1027	433	15,0	—438	146	191 844	21 316	—63 948
25	1497	519	22,0	32	232	1 024	53 824	7 424
26	883	222	11,0	—582	—65	338 724	4 225	37 830
27	4243	923	36,0	2778	636	7 717 284	404 496	1 766 808
28	230	22	2,0	—1235	—265	1 525 225	70 225	327 275
29	973	163	11,0	—492	—124	242 064	15 376	61 008
30	541	100	4,5	—924	—187	853 776	34 969	172 788
31	1410	380	16,0	—55	93	3 025	8 649	—5 115
сумма	45422	8892				27 571 129	1 393 057	3 823 355
средн.	1465	287		61	-2			

[1] В столбце 3 приводятся данные о землепользовании крестьянского хозяйства (в дес.) — исходные данные для примера 5.

ческое значение $t_{кр}$ (в данном случае $t_{кр}=3$) и сравнивают с ним фактическое значение $t_ф$, где для $\rho=0$

$$t=\frac{|r|}{1/\sqrt{n}}\,. \tag{6.5}$$

Если $t_ф > t_{кр}$, т. е. происходит маловероятное событие, предположение о некоррелированности признаков необосновано и коэффициент корреляции считается значимым. Если же $t_ф \leqslant t_{кр}$, коэффициент корреляции считается незначимым.

Пример 2. Выясним, является ли значимым коэффициент корреляции между доходом крестьянского хозяйства и стоимостью скота в хозяйстве.

В нашем примере, задавшись степенью достоверности P, равной 0,99 и найдя соответствующее ей в таблице 1 приложения значение $t_{кр}=2{,}58$, проводим сравнение. Поскольку $t_ф = \frac{0{,}62}{0{,}18} = 3{,}44$ заметно превосходит табличное значение 2,58 и вероятность того, что это случайность, мала (0,01), связь между доходом крестьянского хозяйства и стоимостью имеющегося в нем скота следует признать значимой.

Поскольку вычисленный в примере 1 линейный коэффициент корреляции имеет положительный знак, то взаимосвязь между признаками прямая: чем больше скота в хозяйстве, тем больше доход, и чем меньше скота, тем доход меньше.

Заметим, что линейный коэффициент корреляции является показателем взаимной связи между признаками и не дает представления о том, какой из признаков является факторным, а какой — результативным (в формуле (6.3) признаки x и y совершенно равноправны).

Установив существенность взаимной связи между двумя признаками, можно поставить вопрос о тесноте связи. Напомним, что теснота линейной связи измеряется линейным коэффициентом корреляции, но такая оценка верна только для случая, когда расчеты проведены для генеральной совокупности. Для выборочных данных, что обычно имеет место в практике использования корреляционного анализа, необходимо выявить те границы, в пределах которых находится значение коэффициента корреляции генеральной совокупности, т. е. определить для него доверительные интервалы.

Доверительные интервалы для парного линейного коэффициента корреляции. После расчета выборочного коэффициента корреляции (r) и его средней квадратической ошибки (σ_r) устанавливается уровень достоверности выводов или доверительная вероятность. Для больших выборок соответственно выбранной вероятности P по таблице нормального распределения определяется параметр t. Верхняя граница интервала получается прибавлением к выборочному коэффициенту корреляции $t\sigma_r$, нижняя граница — вычитанием $t\sigma_r$. Формула средней квадратической ошибки (6.4) содержит неизвестный на практике генеральный коэффициент корреляции. Обычно в расчетах его заменяют выборочным коэффициентом корреляции.

Пример 3. Определим доверительные интервалы для коэффициента, вычисленного в примере 1.

2593

1211

Полагая доверительную вероятность P, т. е. вероятность, с которой гарантируются результаты, равной 0,98, находим в табл. 1 приложения соответствующее ей значение t, равное 2,33. Воспользовавшись формулой (6.4), где вместо ρ возьмем выборочный коэффициент корреляции r, равный 0,62, получим значение для средней квадратической ошибки

$$\sigma_r = \frac{1-(0,62)^2}{\sqrt{31}} = \frac{1-0,3844}{5,57} \approx 0.11.$$

Поскольку $t\sigma_r = 2,33 \cdot 0,11 = 0,26$, верхняя и нижняя границы равны соответственно 0,88 и 0,36. Другими словами, с вероятностью 0,98 линейный коэффициент корреляции генеральной совокупности находится в пределах от 0,36 до 0,88.

Важнейшей предпосылкой использования корреляционного анализа является **нормальность распределения** признаков в генеральной совокупности. Нормальность распределения или, по крайней мере, близость распределения к нормальному необходима для корректного проведения проверки значимости связи и вычисления доверительных интервалов. Методы проверки нормальности распределения см. в гл. 9.

Проверка значимости парного линейного коэффициента корреляции для случая малой выборки. При пользовании формулами (6.4) — (6.5) необходимо учитывать два ограничения. Во-первых, они верны для коэффициента корреляции генеральной совокупности и замена последнего выборочным коэффициентом корреляции является искусственным приемом. Риск заменить генеральный коэффициент корреляции существенно отличным от него выборочным коэффициентом увеличивается с уменьшением объема выборки n. Во-вторых, при выборках небольшого объема распределение выборочного коэффициента корреляции может значительно отличаться от нормального, а нормальность распределения является важнейшим условием корректного использования доверительных интервалов и проверки значимости коэффициентов.

Таким образом, при малых объемах выборок (практически при $n<50$) изложенная выше методика определения доверительных интервалов и проверки значимости коэффициентов корреляции неприменима (поэтому примеры 2—3 являются в какой-то мере условными). Особенно возрастают требования к объему выборки при высоких, близких к 1 значениях показателя связи генеральной совокупности.

Таблица 2. Данные и промежуточные расчеты для вычисления коэффициента корреляции между различием в размерах отрезков и соотношением форм эксплуатации

N	y_i	x_i	$y_i-\bar{y}$	$x_i-\bar{x}$	$(y_i-\bar{y})\times$ $\times(x_i-\bar{x})$	$(y_i-\bar{y})^2$	$(x_i-\bar{x})^2$
1	26	0,8	10,2	0,4	4,08	104,04	0,16
2	10	0,3	—5,8	—0,1	0,58	33,64	0,01
3	17	0,2	1,2	—0,2	—0,24	1,44	0,04
4	17	0,6	1,2	0,2	0,24	1,44	0,04
5	13	0,3	—2,8	—0,1	0,28	7,84	0,01
6	12	0,3	—3,8	—0,1	0,38	14,44	0,01
Σ	95	2,5			5,32	162,84	0,27

Для малых выборок проверка значимости коэффициента корреляции осуществляется так, как показано в гл. 9, § 4[1], т. е. вычисляется фактическое значение $t_ф$ величины t, подчиняющейся распределению Стьюдента:

$$t_ф=|r|\sqrt{n-2}/\sqrt{1-r^2} \qquad (6.6)$$

и сравнивается с табличной или критической величиной $t_{кр}$, зависящей от числа $k=n-2$, где n — объем выборки, и от выбранной степени достоверности P. Если $t_ф \geqslant t_{кр}(k)$, то приходят к выводу о наличии связи; если $t_ф < t_{кр}(k)$, то гипотеза об отсутствии связи не отклоняется.

Пример 4. В трех первых столбцах табл. 2 приведены данные, позволяющие анализировать зависимость различий в размерах отрезков, установленных реформой 1861 г. и соотношением форм эксплуатации по шести губерниям Черноземного центра. Вычислим линейный коэффициент корреляции между этими признаками и проверим его значимость (y — % отрезков, x — отношение числа оброчных крестьян к числу барщинных).

Рассчитав средние значения признаков ($\bar{y}=15{,}8$, $\bar{x}=0{,}4$) и сделав промежуточные расчеты в рабочей таблице, вычислим коэффициент корреляции по формуле (6.3):

$$r=\frac{5{,}32}{\sqrt{162{,}84\cdot 0{,}27}}=\frac{5{,}32}{6{,}63}=0{,}80.$$

Для проверки значимости коэффициента воспользуемся формулой (6.6):

$$t_ф=\frac{0{,}8\sqrt{6-2}}{\sqrt{1-(0{,}8)^2}}=\frac{0{,}8\cdot 2}{0{,}6}=2{,}7.$$

---

[1] Там же рассмотрены вопросы сравнения двух выборочных коэффициентов корреляции и оценки корреляции в генеральной совокупности.

Табличное значение $t_{кр}(k)$ находим в табл. 2 приложения. Оно расположено на пересечении строки с номером $k=n-2$ и столбца с удовлетворяющей исследователя степенью достоверности результата. Для нашего примера число степеней свободы $k=6-2=4$. Положив степень достоверности результата P равной 0,90, на пересечении 4-й строки и соответствующего 0,90 столбца находим $t_{кр}(k)=2{,}13$.

Поскольку $t_ф=2{,}7>t(k)=2{,}13$, с риском ошибиться в 10 случаях из 100 отбрасываем гипотезу об отсутствии линейной связи, т. е. считаем, что связь между размером отрезков и формой эксплуатации крестьян существует.

Понятие о частной и множественной корреляции. С помощью парного линейного коэффициента корреляции выявляется связь между двумя признаками, один из которых можно рассматривать как результативный, другой — как факторный. Но в действительности на результат воздействуют несколько факторов. В связи с этим возникают два типа задач: задачи измерения комплексного влияния на результативную переменную нескольких переменных и задачи определения тесноты связи между двумя переменными при фиксированных значениях остальных переменных. Задачи первого типа решаются с помощью множественных коэффициентов корреляции, задачи второго типа — с помощью частных коэффициентов корреляции.

Частный, или чистый, коэффициент корреляции между двумя признаками при исключении влияния третьего признака (обозначим его символом $r_{12.3}$) рассчитывается по формуле[1]

$$r_{12,3}=(r_{12}-r_{13}r_{23})\Big/\sqrt{(1-r_{13}^2)(1-r_{23}^2)}\,, \qquad (6.7)$$

где индексы при r показывают номера признаков, связь между которыми оценивается.

Частный коэффициент корреляции первого и второго признаков при исключении влияния третьего оценивает тесноту линейной корреляционной связи между первым и вторым признаками при фиксированном значении третьего признака. Другими словами, он оценивает влияние на результативный (первый) признак изменения лишь второго признака.

Оценка влияния на результативный признак изменений третьего признака при постоянных значениях второго признака определяется по формуле

[1] Существует алгоритм, позволяющий вычислить частные коэффициенты корреляции при исключенном влиянии двух, трех и т. д. факторов. Этот рекуррентный алгоритм используется в машинном варианте (есть стандартные программы для ЭВМ).

$$r_{13.2}=(r_{13}-r_{12}r_{32}) \Big/ \sqrt{(1-r_{12}^2)(1-r_{32}^2)}. \qquad (6.8)$$

Значения частных коэффициентов корреляции заключаются в тех же пределах от -1 до $+1$, что и значения парных коэффициентов корреляции, и так же интерпретируются.

Пример 5. На основе данных табл. 1 оценить, какое из направлений хозяйства — скотоводческое или земледельческое — определяет доход в зажиточной группе крестьян.

Для решения этого вопроса в дополнение к рассчитанному в примере 1 коэффициенту r_{12}, показывающему тесноту взаимосвязи между доходом хозяйства и стоимостью скота в нем, вычислим еще два парных коэффициента корреляции: r_{13} — коэффициент связи между доходом крестьянского хозяйства и уровнем землепользования в нем, r_{23} — коэффициент связи между землепользованием крестьянского двора и стоимостью всего скота.

Получим $r_{12}=0{,}62$, $r_{13}=0{,}52$, $r_{23}=0{,}74$. Парные коэффициенты корреляции, показывающие тесноту связи между доходом и уровнем землепользования, между доходом и стоимостью скота, почти равны по величине. Отсюда можно сделать вывод о том, что ни та, ни другая формы хозяйствования не имеет преимущественного влияния на доход. Однако взаимосвязь между самими факторными признаками, т. е. между землепользованием и стоимостью скота (r_{23}), оказалась довольно тесной. Наличие же связи между факторными признаками может искажать действительную тесноту связи между результативным и факторными признаками. Расчет частных коэффициентов корреляции позволяет избавляться от этих искажений. Воспользовавшись формулами (6.7) и (6.8), вычислим коэффициенты $r_{12.3}$ и $r_{13.2}$:

$$r_{12.3}=\frac{0{,}62-0{,}52\cdot 0{,}74}{\sqrt{(1-0{,}52^2)(1-0{,}74^2)}}=\frac{0{,}2352}{\sqrt{0{,}3301}}=0{,}41,$$

$$r_{13.2}=\frac{0{,}52-0{,}62\cdot 0{,}74}{\sqrt{(1-0{,}62^2)(1-0{,}74^2)}}=\frac{0{,}0612}{\sqrt{0{,}2785}}=0{,}12.$$

Частный коэффициент корреляции между доходом крестьянского двора и стоимостью скота при постоянном показателе землепользования меньше, чем соответствующий парный коэффициент корреляции (0,41 против 0,62). По-видимому, взаимная зависимость между величиной дохода и стоимостью скота частично опосредствовалась через воздействие уровня землепользования.

Частная корреляция между доходом и землепользованием совсем мала (0,12) при довольно большом значении парного коэффициента (0,52). Это можно объяснить тем, что взаимная зависимость дохода и землепользования в значительной степени были усилена влиянием исключенного здесь фактора — стоимости скота.

Таким образом, частные коэффициенты корреляции позволяют установить тесноту взаимосвязи между двумя признаками при исключении влияния других переменных, но для окончательных выводов необходима проверка уровня значимости частных коэффициентов корреляции.

Значимость частных коэффициентов корреляции зависит не только от величины выборочного коэффициента и объема выборки, но также и от числа введенных в исследование переменных.

Для проверки значимости частных коэффициентов корреляции вычисляют величину $t_ф$ по формуле

$$t_ф=|r_{ij.k}|\sqrt{n-2-m}\,/\sqrt{1-r_{ij.k}^2}, \qquad (6.9)$$

где $r_{ij.k}$ — коэффициент корреляции между i и j признаками при исключенном влиянии k-го признака; m — количество исключенных признаков, и сравнивают ее с табличным или критическим значением, соответствующим принятому уровню значимости.

Наличие связи подтверждается, если вычисленное значение $t_ф$ превосходит значение $t_{кр}$.

Пример 5 (продолжение). Проверить значимость частных коэффициентов корреляции, рассчитанных выше.

Воспользовавшись формулой (6.9), вычислим величины $t_{1ф}$ для $r_{12.3}$ и $t_{2ф}$ для $r_{13.2}$:

$$t_{1ф}=\frac{0{,}41\sqrt{31-2-1}}{\sqrt{1-0{,}41^2}}=2{,}61.$$

$$t_{2ф}=\frac{0{,}12\sqrt{31-2-1}}{\sqrt{1-0{,}12^2}}=0{,}64.$$

Для того чтобы найти табличные значения $t_{кр}$, нужно задаться степенью достоверности выводов и учесть число $n-2-m$, фигурирующее в формуле (6.9).

Табличное значение $t_{кр}$ находится по табл. 2 приложения на пересечении строки с номером $(n-2-m)$ и столбца, соответствующего заданной степени достоверности выводов.

В нашем примере в каждом коэффициенте число признаков, от которых мы абстрагируемся, равно 1 и, следовательно, $n-2-m=31-2-1=28$. Задавшись степенью достоверности, равной 0,90, по табл. 2 приложения находим значение $t_{кр}$, равное 1,7.

Поскольку $t_{1ф}>t_{кр}$, подтверждается наличие взаимосвязи между доходом и стоимостью скота. Для второго из рассматриваемых коэффициентов $t_{2ф}<t_{кр}$, что свидетельствует, по-видимому, об отсутствии непосредственной связи между доходом и землепользованием.

Множественный, или **совокупный,** коэффициент корреляции для случая трех признаков, один из которых — результативный (с номером 1) и два — факторных (с порядковыми номерами 2 и 3) рассчитывается по формуле[1]

[1] Для расчетов используется также формула

$$R_{1(23\ldots p)}=\sqrt{1-(1-r_{12}^2)(1-r_{13.2}^2)(1-r_{14.23}^2)\cdots(1-r_{1p.23\ldots p-1}^2)},$$

пригодная для случаев, когда число признаков, совокупное влияние которых исследуется, превосходит два. Существуют стандартные программы, вычисляющие $R_{1(23\ldots p)}$.

$$R_{1.23}=\sqrt{(r_{12}^2+r_{13}^2-2r_{23}r_{12}r_{13})/(1-r_{23}^2)}, \qquad (6.10)$$

где подстрочные индексы при r показывают номера признаков, связь между которыми оценивается этим коэффициентом корреляции.

Множественный коэффициент корреляции является показателем тесноты линейной связи между результативным признаком и совокупностью факторных признаков.

Множественный коэффициент корреляции изменяется в пределах от 0 до 1. Равенство его нулю говорит об отсутствии линейной связи, равенство единице — о функциональной связи. Указаний на то, является ли связь прямой или обратной, коэффициент не дает.

Пример 6. Рассчитать множественный коэффициент корреляции между величиной дохода крестьянского хозяйства и совокупным влиянием на него уровня землепользования и стоимости скота в хозяйстве.

Воспользуемся результатами предыдущих примеров: $r_{12}=0{,}62$; $r_{13}=0{,}52$; $r_{23}=0{,}74$. По формуле (6.10):

$$R_{1.23}=\sqrt{\frac{0{,}52^2+0{,}62^2-2\cdot 0{,}74\cdot 0{,}52\cdot 0{,}62}{1-0{,}74^2}}=0{,}63,$$

т. е. степень тесноты связи между доходом крестьянского хозяйства и совокупностью двух факторных признаков составляет 0,63.

Вопрос о значимости множественного коэффициента корреляции решается с помощью так называемого F-критерия. Величина $F_ф$, рассчитанная по специальной формуле, сравнивается с табличным (критическим) значением $F_{кр}$ — распределения Фишера, соответствующим выбранной степени достоверности P (или уровню значимости $\alpha=1-P$). Если $F_ф>F_{кр}$, то коэффициент признается значимым на уровне α, так как вероятность того, что $F_ф$ для выборки случайно получилось больше $F_{кр}$, мала (меньше α). Для случая двух факторных признаков величина вычисляется по формуле[1]

$$F_ф[2,\ n-3]=\frac{(n-3)\,R_{1.23}^2}{2\left(1-R_{1.23}^2\right)}. \qquad (6.11)$$

[1] Формула расчетной величины F для общего случая m факторных признаков имеет вид

$$F\,[m,\ n-(m+1)]\doteq\frac{[n-(m+1)]\,R_{1.23\ldots m+1}^2}{m\left(1-R_{1.23\ldots m+1}^2\right)},$$

где $R_{1.23\ldots m+1}$ — множественный коэффициент корреляции; n — число объектов, для которых приводятся значения признаков; m — число факторных признаков. Этот же критерий используется для проверки значимости регрессии, см. (6.20).

Пример 6 (продолжение). Проверить значимость вычисленного в примере 6 множественного коэффициента корреляции для уровня значимости $\alpha=0{,}05$ ($P=0{,}95$).

Воспользуемся формулой (6.11):

$$F_{\text{ф}}[2,28]=\frac{28\cdot 0{,}63^2}{2(1-0{,}63^2)}=9{,}21.$$

В табл. 4А приложения (соответствующей $\alpha=0{,}05$) во 2-м столбце и 28-й строке (номер столбца определяется по первому числу в квадратной скобке при F, а номер строки — по второму числу) находим критическое значение F-распределения. Поскольку при достоверности вывода, равной 0,95, $F_{\text{кр}}=3{,}34$, т. е. $F_{\text{ф}}>F_{\text{кр}}$, совместное влияние на величину дохода уровня землепользования и стоимости скота имеет место.

Коэффициент детерминации. Линейный коэффициент корреляции оценивает тесноту взаимосвязи между признаками и показывает, является ли связь прямой или обратной. Но понятие тесноты взаимосвязи часто может быть недостаточным при содержательном анализе взаимосвязей. В частности, коэффициент корреляции не показывает степень воздействия факторного признака на результативный. Таким показателем является коэффициент детерминации (обозначим его D), для случая линейной связи представляющий собой квадрат парного линейного коэффициента корреляции ($D=r^2$) или квадрат множественного коэффициента корреляции. Его значение определяет долю (в процентах) изменений, обусловленных влиянием факторного признака, в общей изменчивости результативного признака.

Пример 7. Рассчитать по данным примера 4 (табл. 2) коэффициент детерминации.

Коэффициент корреляции равен 0,80. Коэффициент детерминации составляет: $D=0{,}80^2\cdot 100\%=64\%$.

Таким образом, различия в размерах отрезков в губерниях Черноземного центра на 64% определялись различиями в соотношении форм эксплуатации.

Корреляционная таблица. В предыдущем изложении каждый признак был представлен рядом значений. В некоторых случаях удобна другая форма записи исходных данных, в виде корреляционной таблицы.

Корреляционная таблица имеет такое устройство: по строкам располагаются значения одного признака, по столбцам — другого признака. Число, стоящее в клетке на пересечении i-й строки и j-го столбца, показывает, как часто i-е значение первого признака встречается совместно с j-м значением второго признака.

В табл. 3 представлены сведения об уходящих в отставку дворянах (60-е годы XVIII в.). В строках дан возраст ухода в отставку, в столбцах — имущественное положение уходящих в отставку, характеризуемое числом принадлежащих им крестьян. В клетках представлены числа, показывающие, как часто встречаются соответствующие сочетания количественных значений этих двух признаков.

Корреляционная таблица является удобной формой представления данных. Уже беглое знакомство с ней нередко позволяет судить о тесноте связи (не обязательно линейной) между признаками. Если большие значения одного признака в основном сочетаются с большими значениями другого признака, а малые — с малыми, т. е. числа концентрируются вдоль диагонали таблицы, то можно говорить о наличии прямой связи. Если малые значения одного признака в большинстве случаев встречаются совместно с большими значениями другого признака, т. е. числа сосредоточены вдоль другой диагонали, то имеется обратная связь. При этом, чем больше концентрация частот вдоль одной из диагоналей, тем теснее связь.

Корреляционное отношение. Когда изучаемая совокупность (в виде корреляционной таблицы) разбивается на группы по одному (факторному) признаку x, то для каждой из этих групп можно вычислить соответствующие групповые средние результативного признака $\bar{y}_i$. Изменение групповых средних от группы к группе свидетельствует о наличии связи результативного признака с факторным, а примерное равенство групповых средних — об отсутствии связи. Следовательно, чем большую роль в общем изменении результативного признака $\left(\sum_{i=1}^{n}(y_i-\bar{y})^2\right)$ играет изменение групповых средних (за счет влияния факторного признака), тем сильнее влияние этого признака. Величина $S_y^2=\sum_{i=1}^{n}(y_i-\bar{y})^2$ (вариация признака y) складывается из изменения групповых средних, обусловленного влиянием факторного признака, т. е. из межгрупповой или факторной вариации ($S^2_{\text{факт}}$) и из вариации признака внутри групп, обусловленной другими причинами, — внутригрупповой или остаточной ($S^2_{\text{ост}}$). Поэтому отношение

$$S^2_{\text{факт}}/S^2_y=S^2_{\text{факт}}/(S^2_{\text{факт}}+S^2_{\text{ост}})$$

является мерой степени влияния факторного признака на результативный. Значение квадратного корня из этого отношения измеряет «направленную связь» ($x\rightarrow y$) в общем виде и носит название корреляционного отношения ($\eta_{y/x}$).

Таблица 3. Связь между возрастом уходящих в отставку дворян и их имущественным положением (количеством крепостных крестьян)

Возраст ухода в отставку	Середина интервала	Количество крепостных крестьян (x)				Σ
		от 0 до 29	от 30 до 99	от 100 до 799	от 800 и выше	
25—34	30	5	1	4	2	12
35—44	40	6	10	3	1	20
45—54	50	5	0	2	0	7
55—65	60	2	1	0	0	3
Σ		18	12	9	3	42

Корреляционное отношение вычисляется по формуле

$$\eta_{y/x} = \sqrt{S^2_{\text{факт}}/S^2_y}, \tag{6.12}$$

Величины $S^2_{\text{факт}}$ и $S_y{}^2$ вычисляются по формулам:

$$S^2_{\text{факт}} = \sum_{i=1}^{k} (\bar{y}_i - \bar{y})^2 n_i, \tag{6.13}$$

$$S^2_y = \sum_{i=1}^{n} (y_i - \bar{y})^2, \tag{6.14}$$

где $\bar{y}_i$ — среднее значение y в i-й группе; $\bar{y}$ — общее среднее; n_i — объем i-й группы; k — количество групп; n — объем совокупности $\left(n = \sum_{i=1}^{k} n_i\right)$.

Рассмотрим вычисление корреляционного отношения, воспользовавшись данными табл. 3.

Пример 8. В источнике, содержащем данные об уходящих в отставку дворянах, имеются сведения о причинах ухода в отставку. Как правило, в качестве причины указывалась болезнь. Являлась ли эта причина единственной? Есть основания предположить, что возраст ухода в отставку зависел также от имущественного положения находящихся на военной службе дворян: чем более они были обеспечены, тем с большей легкостью могли оставить службу. Для проверки этого предположения измерим тесноту связи между возрастом уходящих в отставку дворян и их имущественным положением по данным табл. 3.

По признаку имущественного положения совокупность разбита на 4 группы. Рассчитаем средний возраст ухода в отставку для каждой группы. Для первой группы

$$\bar{y}_1 = \frac{30 \cdot 5 + 40 \cdot 6 + 50 \cdot 5 + 60 \cdot 2}{18} = 42,2.$$

Аналогично вычислим $\bar{y}_2 = 40,8$, $\bar{y}_3 = 37,7$, $\bar{y}_4 = 33,3$. Средние, как мы видим, различны и убывают при переходе из первой во вторую, из второй в третью, из третьей в четвертую группы. Таким образом, средний возраст ухода в отставку уменьшается с ростом материальной обеспеченности, что вполне согласуется с нашей гипотезой о наличии связи между этими признаками.

Наконец, вычислим общую среднюю арифметическую (средний возраст ухода в отставку по всем данным табл. 3). Для расчета этой средней в качестве весов возьмем числа, стоящие в последнем суммарном столбце таблицы:

$$\bar{y} = \frac{30 \cdot 12 + 40 \cdot 20 + 50 \cdot 7 + 60 \cdot 3}{42} = 40,2.$$

Далее, воспользовавшись последовательно формулами (6.13), (6.14) и (6.12), рассчитаем корреляционное отношение:

$$S^2_{\text{факт}} = (42,2 - 40,2)^2 \cdot 18 + (40,8 - 40,2)^2 \cdot 12 + (37,7 - 40,2)^2 \cdot 9 +$$

$$+ (33,3 - 40,2)^2 \cdot 3 = 294,$$

$$S^2_y = (30 - 40,2)^2 \cdot 12 + (40 - 40,2)^2 \cdot 20 + (50 - 40,2)^2 \cdot 7 +$$

$$+ (60 - 40,2)^2 \cdot 3 = 3098,$$

$$\eta = \sqrt{294/3098} = \sqrt{0,0983} \approx 0,31.$$

Величина корреляционного отношения изменяется в пределах от 0 до 1. Близость ее к нулю говорит об отсутствии связи, близость к единице — о тесной связи.

Корреляционное отношение в отличие от линейного коэффициента корреляции не указывает, является ли связь прямой или обратной. Однако нередко уже по виду исходной таблицы можно решить этот вопрос. Промежуточные расчеты также помогают определить, является связь прямой или обратной: если с ростом факторного группировочного признака растут групповые средние результативного признака, то связь — прямая, если же с увеличением факторного признака значения групповых средних уменьшаются, то связь — обратная.

Как показатель тесноты связи корреляционное отношение имеет более универсальный характер, чем линейный коэффициент корреляции, поскольку его использование не ограничивается случаями линейной связи, а факторный признак может быть не количественным, а ранговым и даже номинальным.

Квадрат корреляционного отношения, выраженный в процентах, и коэффициент детерминации, как он определен выше, имеют одинаковый смысл, только при расчете коэффициента детерминации используется предположение о линейной связи между факторным и результативным признаками, тогда как при вычислении корреляционного отношения вопрос о форме связи не ставится.

Возвращаясь к примеру 8, оценим полученные в нем результаты. Корреляционное отношение, выражающее степень тесноты связи, оказалось равным 0,31, а квадрат корреляционного отношения в процентах (коэффициент детерминации) — 9,9%.

Таким образом, возраст ухода в отставку зависит от имущественного положения, но в целом в небольшой степени: колебания имущественного положения только на 9,9% объясняют колебания возраста ухода в отставку. Причем, поскольку с увеличением количества крестьян у владельцев средний возраст ухода в отставку уменьшается, то связь между этими признаками — обратная [1].

Линейный коэффициент корреляции, как уже отмечалось, является показателем взаимной связи между признаками, но не указывает, какой из них результативный, а какой — факторный. В этом отношении от него выгодно отличается корреляционное отношение, которое позволяет выявить это соотношение. Для этого вычисляются два корреляционных отношения, сравнение которых и помогает определить правильное распределение «ролей» между признаками.

Коэффициент корреляции и корреляционное отношение являются эффективным средством выявления связи между различными количественными признаками и определения ее тесноты. Коэффициент детерминации дает представление и о степени воздействия одних факторов на другие. Однако решение вопроса о тесноте связи нередко упирается в вопрос о форме связи. Кроме того, знание формы связи дает развернутое представление о влиянии различных факторов на результативный признак, а также возможность прогнозирования изменений результата при тех или иных комбинациях значений факторов.

Выявление формы связи осуществляется с помощью методов регрессионного анализа.

§ 3. Линейная регрессия

Регрессионный анализ позволяет приближенно определить форму связи между результативным и факторными

[1] Вопрос оценки значимости корреляционного отношения в выборке рассмотрен в § 4 гл. 9.

признаками, а также решить вопрос о том, значима ли эта связь. Вид функции, с помощью которой приближенно выражается форма связи, выбирают заранее, исходя из содержательных соображений или визуального анализа данных. Математическое решение задачи основано на методе наименьших квадратов.

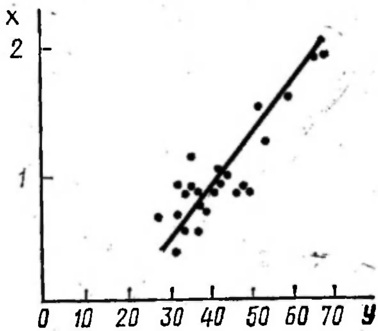

Рис. 14. График зависимости величины сбора хлеба и размера посева на душу населения

Суть метода наименьших квадратов. Рассмотрим содержание метода на конкретном примере. Пусть имеются данные о сборе хлеба на душу населения по совокупности черноземных губерний. От каких факторов зависит величина этого сбора? Вероятно, определяющее влияние на величину сбора хлеба оказывает величина посева и уровень урожайности. Рассмотрим сначала зависимость величины сбора хлеба на душу населения от размера посева на душу (столбцы 1 и 2 табл .4) Попытаемся представить интересующую нас зависимость с помощью прямой линии. Разумеется, такая линия может дать только приближенное представление о форме реальной статистической связи. Постараемся сделать это приближение наилучшим. Оно будет тем лучше, чем меньше исходные данные будут отличаться от соответствующих точек, лежащих на линии. Степень близости может быть выражена величиной суммы квадратов отклонений реальных значений от расположенных на прямой. Использование именно квадратов отклонений (не просто отклонений) позволяет суммировать отклонения различных знаков без их взаимного погашения и дополнительно обеспечивает сравнительно большее внимание, уделяемое большим отклонениям. Именно этот критерий (минимизация суммы квадратов отклонений) положен в основу метода наименьших квадратов.

В вычислительном аспекте метод наименьших квадратов сводится к составлению и решению системы так называемых **нормальных уравнений.** Исходным этапом для этого является подбор вида функции, отображающей статистическую связь.

Тип функции в каждом конкретном случае можно подобрать путем прикидки на графике исходных данных подходящей, т. е. достаточно хорошо приближающей эти данные, линии. В нашем случае связь между сбором хлеба на душу и величиной посева на душу может быть изображена с помощью прямой линии (рис. 14) и записана в виде

$$y = a_0 + a_1 x, \tag{6.15}$$

где y — величина сбора хлеба на душу (результативный признак или зависимая переменная); x — величина посева на душу (факторный признак или независимая переменная); a_0 и a_1 — параметры уравнения, которые могут быть найдены методом наименьших квадратов.

Для нахождения искомых параметров нужно составить систему уравнений, которая в данном случае будет иметь вид

$$\begin{cases} na_0 + a_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i, \\ a_0 \sum_{i=1}^{n} x_i + a_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i. \end{cases} \tag{6.16}$$

Полученная система может быть решена известным из школьного курса методом Гаусса. Искомые параметры системы из двух нормальных уравнений можно вычислить и непосредственно с помощью последовательного использования нижеприведенных формул:

$$a_1 = \frac{n \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i \right)^2}, \quad a_0 = \bar{y} - a_1 \bar{x}, \tag{6.17}$$

где y_i — i-е значение результативного признака; x_i — i-е значение факторного признака; $\bar{y}$ и $\bar{x}$ — средние арифметические результативного и факторного признаков соответственно; n — число значений признака y_i, или, что то же самое, число значений признака x_i.

Пример 9. Найдем уравнение линейной связи между величиной сбора хлеба (y) и размером посева (x) по данным табл. 4. Проделав необходимые вычисления, получим из (6.17):

$$a_1 = \frac{23 \cdot 1001{,}22 - 22{,}5 \cdot 959{,}43}{23 \cdot 24{,}56 - (22{,}5)^2} = 24{,}5759.$$

$$a_0 = \frac{959{,}43}{23} - 24{,}5759 \frac{22{,}5}{23} = 17{,}6727.$$

Таблица 4. Исходные данные для решения уравнений линейной регрессии

№	y	x_1	x_2	№	y	x_1	x_2
1	48,01	0,91	46,08	13	36,26	0,90	40,06
2	38,18	0,76	45,18	14	32,07	0,52	57,91
3	38,70	0,82	41,76	15	32,83	0,66	43,86
4	46,72	0,88	50,94	16	35,16	0,58	58,62
5	41,58	0,88	43,54	17	44,56	0,99	44,39
6	36,89	0,89	38,80	18	59,16	1,63	35,77
7	34,54	0,87	39,22	19	67,99	1,95	35,96
8	42,86	0,94	42,74	20	53,73	1,27	40,99
9	38,97	0,91	41,20	21	52,39	1,55	33,05
10	43,22	1,07	39,35	22	36,10	1,15	30,68
11	28,19	0,69	34,38	23	32,67	0,94	34,26
12	38,65	0,74	48,98	Σ	959,43	22,50	967,72

Примечание. y — сбор хлеба (пуд.) на душу населения, x_1 — посев (дес.) на душу, x_2 — урожай (пуд./дес.).

Таким образом, уравнение связи, или, как принято говорить, **уравнение регрессии**, выглядит следующим образом:

$$y = 17{,}6727 + 24{,}5759x.$$

Интерпретация коэффициента регрессии. Уравнение регрессии не только определяет форму анализируемой связи, но и показывает, в какой степени изменение одного признака сопровождается изменением другого признака.

Коэффициент при x, называемый **коэффициентом регрессии**, показывает, на какую величину в среднем изменяется результативный признак y при изменении факторного признака x на единицу.

В примере 9 коэффициент регрессии получился равным 24,58. Следовательно, с увеличением посева, приходящегося на душу, на одну десятину сбор хлеба на душу населения в среднем увеличивается на 24,58 пуда.

Средняя и предельная ошибки коэффициента регрессии. Поскольку уравнения регрессии рассчитываются, как правило, для выборочных данных, обязательно встают вопросы точности и надежности полученных результатов. Вычисленный коэффициент регрессии, будучи выборочным, с некоторой точностью оценивает соответствующий коэффициент регрессии генеральной совокупности. Представление об этой точности дает средняя ошибка коэффициента регрессии (μ_{a_1}), рассчитываемая по формуле

$$\mu_{a_1} = \sigma_{y.x}/(\sigma_x \sqrt{n}), \qquad (6.18)$$

где

$$\sigma_{y.x}=\sqrt{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2/[n-(m+1)]}, \qquad (6.19)$$

y_i — i-е значение результативного признака; $\hat{y}_i$ — i-е выравненное значение, полученное из уравнения (6.15); x_i — i-е значение факторного признака; σ_x — среднее квадратическое отклонение x; n — число значений x или, что то же самое, значений y; m — число факторных признаков (независимых переменных).

В формуле (6.18), в частности, формализовано очевидное положение: чем больше фактические значения отклоняются от выравненных, тем большую ошибку следует ожидать; чем меньше число наблюдений, на основе которых строится уравнение, тем больше будет ошибка.

Средняя ошибка коэффициента регрессии является основой для расчета предельной ошибки. Последняя показывает, в каких пределах находится истинное значение коэффициента регрессии при заданной надежности результатов. Предельная ошибка коэффициента регрессии вычисляется аналогично предельной ошибке средней арифметической (см. гл. 5), т. е. как $t\,\mu_{a_1}$, где t — величина, числовое значение которой определяется в зависимости от принятого уровня надежности.

Пример 10. Найти среднюю и предельную ошибки коэффициента регрессии, полученного в примере 9.

Для расчета μ_{a1} прежде всего подсчитаем выравненные значения y_i, для чего в уравнение регрессии, полученное в примере 9, подставим конкретные значения x_i:

$$\hat{y}_1=17,6681+24,5762\cdot 0,91=40,04 \text{ и т. д.}$$

Затем вычислим отклонения фактических значений y_i от выравненных и их квадраты.

Далее, подсчитав средний по черноземным губерниям посев на душу ($\bar{x}=0,98$), отклонения фактических значений x_i от этой средней квадраты отклонений и среднее квадратическое отклонение $\sigma_x=\sqrt{2,5588/23}=0,333$, получим все необходимые составляющие формул (6.18) и (6.19):

$$\sigma_{y.x}=\sqrt{\frac{448,7781}{21}}=\sqrt{21,3704}=4,6228;$$

$$\mu_{a1}=\frac{4,6228}{\sqrt{23}\cdot 0,3335}=2,89.$$

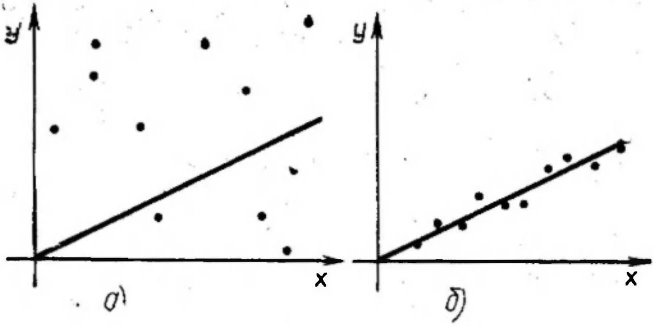

Рис. 15. Два случая взаимоотношения между двумя признаками

Таким образом, средняя ошибка коэффициента регрессии равна 2,89, что составляет 12% от вычисленного коэффициента.

Задавшись уровнем надежности, равным 0,95, найдем по табл. 1 приложения соответствующее ему значение $t=1,96$, рассчитаем предельную ошибку $1,96 \cdot 2,89=5,66$ и пределы коэффициента регрессии для принятого уровня надежности [1]. Нижняя граница коэффициента регрессии равна $24,58-5,66=18,92$, а верхняя граница $24,58+5,66=$ $=30,24$.

Средняя квадратическая ошибка линии регрессии. Уравнение регрессии представляет собой функциональную связь, при которой по любому значению x можно однозначно определить значение y. Функциональная связь лишь приближенно отражает связь реальную, причем степень этого приближения может быть различной и зависит она как от свойств исходных данных, так и от выбора вида функции, по которой производится выравнивание.

На рис. 15 представлены два различных случая взаимоотношения между двумя признаками. В обоих случаях предполагаемая связь описывается одним и тем же уравнением, но во втором случае соотношение между признаками x и y достаточно четко выражено и уравнение, по-видимому, довольно хорошо описывает это соотношение, тогда как в первом случае сомнительно само наличие сколько-нибудь закономерного соотношения между признаками. И в том, и в другом случаях, несмотря на их существенное различие, метод наименьших квадратов дает одинаковое уравнение, поскольку этот метод нечувствителен к потенциальным возможностям исходного материала вписаться в ту или иную схему. Кроме того, метод наименьших квадра-

[1] В случае малых выборок величина t находится из табл. 2 приложения.

тов применяется для расчета неизвестных параметров заранее **выбранного** вида функции, и вопрос о выборе наиболее подходящего для конкретных данных вида функции в рамках этого метода не ставится и не решается. Таким образом, при пользовании методом наименьших квадратов открытыми остаются два важных вопроса, а именно: существует ли связь и верен ли выбор вида функции, с помощью которой делается попытка описать форму связи.

Чтобы оценить, насколько точно уравнение регрессии описывает реальные соотношения между переменными, нужно ввести меру рассеяния фактических значений относительно вычисленных с помощью уравнения. Такой мерой служит **средняя квадратическая ошибка регрессионного уравнения**, вычисляемая по приведенной выше формуле (6.19).

Пример 11. Определить среднюю квадратическую ошибку уравнения, полученного в примере 9.

Промежуточные расчеты примера 10 дают нам среднюю квадратическую ошибку уравнения. Она равна 4,6 пуда.

Этот показатель аналогичен среднему квадратическому отклонению для средней. Подобно тому, как по величине среднего квадратического отклонения можно судить о представительности средней арифметической (см. гл. 5), по величине средней квадратической ошибки регрессионного уравнения можно сделать вывод о том, насколько показательна для соотношения между признаками та связь, которая выявлена уравнением. В каждом конкретном случае фактическая ошибка может оказаться либо больше, либо меньше средней. Средняя квадратическая ошибка уравнения показывает, насколько в среднем мы ошибемся, если будем пользоваться уравнением, и тем самым дает представление о точности уравнения. Чем меньше $\sigma_{y.x}$, тем точнее предсказание линии регрессии, тем лучше уравнение регрессии описывает существующую связь. Показатель $\sigma_{y.x}$ позволяет различать случаи, представленные на рис. 15. В случае б) он окажется значительно меньше, чем в случае а). Величина $\sigma_{y.x}$ зависит как от выбора функции, так и от степени описываемой связи.

Варьируя виды функций для выравнивания и оценивая результаты с помощью средней квадратической ошибки, можно среди рассматриваемых выбрать лучшую функцию, функцию с наименьшей средней ошибкой. Но существует ли связь? Значимо ли уравнение регрессии, используемое для отображения предполагаемой связи? На эти вопросы отвечает определяемый ниже критерий значимости регрессии.

Мерой значимости линии регрессии может служить следующее соотношение:

$$F[m, n-(m+1)]=\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2/(m\sigma^2_{y.x}), \qquad (6.20)$$

где $\hat{y}_i$ — i-е выравненное значение; $\bar{y}$ — средняя арифметическая значений y_i, $\sigma_{y.x}$ — средняя квадратическая ошибка регрессионного уравнения, вычисляемая по формуле (6.19); n — число сравниваемых пар значений признаков; m — число факторных признаков.

Действительно, связь тем больше, чем значительнее мера рассеяния признака, обусловленная регрессией, превосходит меру рассеяния отклонений фактических значений от выравненных.

Соотношение (6.20) позволяет решить вопрос о значимости регрессии. Регрессия значима, т. е. между признаками существует линейная связь, если для данного уровня значимости вычисленное значение $F_ф[m, n-(m+1)]$ превышает критическое значение $F_{кр}[m, n-(m+1)]$, стоящее на пересечении m-го столбца и $[n-(m+1)]$-й строки специальной таблицы (см. табл. 4 приложения).

Пример 12. Выясним, связаны ли сбор хлеба на душу населения и посев на душу населения линейной зависимостью.

Воспользуемся F-критерием значимости регрессии. Подставив в формулу (6.20) данные табл. 4 и результат примера 10, получим

$$F[1,21]=\frac{1545,1331}{21,3704}=72,30.$$

Обращаясь к таблице F-распределения для $P=0,95$ ($\alpha=1-P=0,5$) и учитывая, что $n=23$, $m=1$, в табл. 4А приложения на пересечения 1-го столбца и 21-й строки находим критическое значение $F_{кр}$, равное 4,32 при степени надежности $P=0,95$. Поскольку вычисленное значение $F_ф$ существенно превосходит по величине $F_{кр}$, то обнаруженная линейная связь существенна, т. е. априорная гипотеза о наличии линейной связи подтвердилась. Вывод сделан при степени надежности $P=0,95$. Между прочим, вывод в данном случае останется прежним, если надежность повысить до $P=0,99$ (соответствующее значение $F_{кр}=8,02$ по табл. 4Б приложения для уровня значимости $\alpha=0,01$).

Коэффициент детерминации. С помощью F-критерия мы установили, что существует линейная зависимость между величиной сбора хлеба и величиной посева на душу. Следовательно, можно утверждать, что величина сбора хлеба, приходящегося на душу, линейно зависит от величины посева на душу. Теперь уместно поставить уточняющий вопрос — в какой степени величина посева на душу определяет величину сбора хлеба на душу? На этот вопрос можно ответить, рассчитав, какая часть вариации результативного признака может быть объяснена влиянием факторного признака.

Рассмотрим отношение

$$R^2=\frac{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}. \tag{6.21}$$

Оно показывает долю разброса, учитываемого регрессией, в общем разбросе результативного признака и носит название **коэффициента детерминации.** Этот показатель, равный отношению факторной вариации к полной вариации признака, позволяет судить о том, насколько «удачно» выбран вид функции[1]. Проведя расчеты, основанные на одних и тех же исходных данных, для нескольких типов функций, мы можем из них выбрать такую, которая дает наибольшее значение R^2 и, следовательно, в большей степени, чем другие функции, объясняет вариацию результативного признака. Действительно, при расчете R^2 для одних и тех же данных, но разных функций знаменатель выражения (6.21) остается неизменным, а числитель показывает ту часть вариации результативного признака, которая учитывается выбранной функцией. Чем больше R^2, т. е. чем больше числитель, тем больше изменение факторного признака объясняет изменение результативного признака и тем, следовательно, лучше уравнение регрессии, лучше выбор функции.

Наконец, отметим, что введенный ранее, при изложении методов корреляционного анализа, коэффициент детерминации совпадает с определенным здесь показателем, если выравнивание производится по прямой линии. Но последний показатель (R^2) имеет более широкий спектр применения и может использоваться в случае связи, отличной от линейной (см. § 4 данной главы).

Пример 13. Рассчитать коэффициент детерминации для уравнения, полученного в примере 9.

---

[1] Отметим, что по смыслу коэффициент детерминации в регрессионном анализе соответствует квадрату корреляционного отношения для корреляционной таблицы (см. § 2).

Вычислим R^2, воспользовавшись формулой (6.21) и данными табл. 4:

$$R^2=\frac{1545,1331}{1983,6224}\,100\%=77,89\%.$$

Итак, уравнение регрессии почти на 78% объясняет колебания сбора хлеба на душу. Это немало, но, по-видимому, можно улучшить модель введением в нее еще одного фактора.

Случай двух независимых переменных. Простейший случай множественной регрессии. В предыдущем изложении регрессионного анализа мы имели дело с двумя признаками — результативным и факторным. Но на результат действует обычно не один фактор, а несколько, что необходимо учитывать для достаточно полного анализа связей.

В математической статистике разработаны **методы множественной регрессии**[1], позволяющие анализировать влияние на результативный признак нескольких факторных. К рассмотрению этих методов мы и переходим.

Возвратимся к примеру 9. В нем была определена форма связи между величиной сбора хлеба на душу и размером посева на душу. Введем в анализ еще один фактор — уровень урожайности (см. столбец 3 табл. 4). Без сомнения, эта переменная влияет на сбор хлеба на душу. Но в какой степени влияет? Насколько обе независимые переменные определяют сбор хлеба на душу в черноземных губерниях? Какая из переменных — посев на душу или урожайность — оказывает определяющее влияние на сбор хлеба? Попытаемся ответить на эти вопросы.

После добавления второй независимой переменной уравнение регрессии будет выглядеть так:

$$y=a_0+a_1x_1+a_2x_2, \tag{6.22}$$

где y — сбор хлеба на душу; x_1 — размер посева на душу; x_2 — урожай с десятины (в пудах); a_0, a_1, a_2 — параметры, подлежащие определению.

Для нахождения числовых значений искомых параметров, как и в случае одной независимой переменной, пользуются методом наименьших квадратов. Он сводится к составлению и решению системы нормальных уравнений, которая имеет вид

---

[1] Регрессия называется **множественной, если число независимых переменных, учтенных в ней, больше или равно двум.**

$$\left.\begin{aligned} na_0+a_1\sum_{i=1}^{n}x_{1i}+a_2\sum_{i=1}^{n}x_{2i}&=\sum_{i=1}^{n}y_i,\\ a_0\sum_{i=1}^{n}x_{1i}+a_1\sum_{i=1}^{n}x_{1i}^2+a_2\sum_{i=1}^{n}x_{1i}x_{2i}&=\sum_{i=1}^{n}y_ix_{1i},\\ a_0\sum_{i=1}^{n}x_{2i}+a_1\sum_{i=1}^{n}x_{1i}x_{2i}+a_2\sum_{i=1}^{n}x_{2i}^2&=\sum_{i=1}^{n}y_ix_{2i}. \end{aligned}\right\} \quad (6.23)$$

Когда система состоит из трех и более нормальных уравнений, решение ее усложняется. Существуют стандартные программы расчета неизвестных параметров регрессионного уравнения на ЭВМ. При ручном счете можно воспользоваться известным из школьного курса методом Гаусса.

Пример 14. По данным табл. 4 описанным способом найдем параметры a_0, a_1, a_2 уравнения (6.22). Получены следующие результаты: $a_2=0{,}3288$, $a_1=28{,}7536$, $a_0=-0{,}2495$.

Таким образом, уравнение множественной регрессии между величиной сбора хлеба на душу населения (y), размером посева на душу (x_1) и уровнем урожайности (x_2) имеет вид:

$$y=-0{,}2495+28{,}7536x_1+0{,}3288x_2.$$

Интерпретация коэффициентов уравнения множественной регрессии. Коэффициент при x_1 в полученном уравнении отличается от аналогичного коэффициента в уравнении примера 9.

Коэффициент при независимой переменной в уравнении простой регрессии отличается от коэффициента при соответствующей переменной в уравнении множественной регрессии тем, что в последнем элиминировано влияние всех учтенных в данном уравнении признаков.

Коэффициенты уравнения множественной регрессии поэтому называются **частными** или **чистыми** коэффициентами регрессии.

Частный коэффициент множественной регрессии при x_1 показывает, что с увеличением посева на душу на 1 дес. и **при фиксированной урожайности** сбор хлеба на душу населения вырастает в среднем на 28,8 пуда. Частный коэффициент при x_2 показывает, что **при фиксированном посеве на душу** увеличение урожая на единицу, т. е. на 1 пуд с десятины, вызывает в среднем увеличение сбора хлеба на душу на 0,33 пуда. Отсюда можно сделать вывод, что увеличение сбора хлеба в черноземных губерниях России идет, в основном, за счет расширения посева и в значительно меньшей степени — за счет повышения урожайности, т. е. экстенсивная форма развития зернового хозяйства является господствующей.

Введение переменной x_2 в уравнение позволяет уточнить коэффициент при x_1. Конкретно, коэффициент оказался выше (28,8 против 24,6), когда в изучаемой связи вычленилось влияние урожайности на сбор хлеба.

Однако выводы, полученные в результате анализа коэффициентов регрессии, не являются пока корректными, поскольку, во-первых, не учтена разная масштабность факторов, во-вторых, не выяснен вопрос о значимости коэффициента a_2.

Величина коэффициентов регрессии изменяется в зависимости от единиц измерения, в которых представлены переменные. Если переменные выражены в разном масштабе измерения, то соответствующие им коэффициенты становятся несравнимыми. Для достижения сопоставимости коэффициенты регрессии исходного уравнения стандартизуют, взяв вместо исходных переменных их отношения к собственным средним квадратическим отклонениям. Тогда уравнение (6.22) приобретает вид

$$\frac{y}{\sigma_y} = a_0' + \beta_1 \frac{x_1}{\sigma_{x_1}} + \beta_2 \frac{x_2}{\sigma_{x_2}} .$$

Сравнивая полученное уравнение с уравнением (6.22), определяем стандартизованные частные коэффициенты уравнения, так называемые **бета-коэффициенты**, по формулам:

$$\beta_1 = a_1 \frac{\sigma_{x_1}}{\sigma_y}, \quad \beta_2 = a_2 \frac{\sigma_{x_2}}{\sigma_y}, \tag{6.24}$$

где β_1 и β_2 — бета-коэффициенты; a_1 и a_2 — коэффициенты регрессии исходного уравнения; σ_y, σ_{x_1} и σ_{x_2} — средние квадратические отклонения переменных y, x_1 и x_2 соответственно.

Вычислив бета-коэффициенты для уравнения, полученного в примере 14:

$$\beta_1 = \frac{28{,}7536 \cdot 0{,}3335}{9{,}2867} = 1{,}03, \quad \beta_2 = \frac{0{,}3288 \cdot 7{,}0053}{9{,}2867} = 0{,}2480,$$

видим, что вывод о преобладании в черноземной полосе России экстенсивной формы развития хозяйства над интенсивной остается в силе, так как β_1 значительно больше, чем β_2.

Оценка точности уравнения множественной регрессии. Точность уравнения множественной регрессии, как и в случае уравнения с одной независимой переменной, оценивается средней квадратической ошибкой уравнения. Обозначим ее $\sigma_{y.x_1x_2}$, где подстрочные индексы указывают, что результативным признаком в уравнении является y, а факторными признаками x_1 и x_2. Для расчета средней квадратической ошибки уравнения множественной регрессии применяется приведенная выше формула (6.19).

Пример 15. Оценим точность полученного в примере 14 уравнения регрессии.

Воспользовавшись формулой (6.19) и данными табл. 4, вычислим среднюю квадратическую ошибку уравнения:

$$\sigma_{y.x_1x_2} = \sqrt{\frac{219{,}2968}{20}} = \sqrt{10{,}9948} = 3{,}3158 \text{ (пуда)}.$$

Оценка полезности введения дополнительной переменной. Точность уравнения регрессии тесно связана с вопросом ценности включения дополнительных членов в это уравнение.

Сравним средние квадратические ошибки, рассчитанные для уравнения с одной переменной x_1 (пример 11) и для уравнения с двумя независимыми переменными x_1 и x_2. Включение в уравнение новой переменной (урожайности) уменьшило среднюю квадратическую ошибку почти вдвое.

Можно провести сравнение ошибок с помощью коэффициентов вариации

$$V_\sigma = \frac{\sigma_f}{\bar{y}} 100\%, \qquad (6.25)$$

где σ_f — средняя квадратическая ошибка регрессионного уравнения; $\bar{y}$ — средняя арифметическая результативного признака.

Для уравнения, содержащего одну независимую переменную:

$$V_\sigma = \frac{4{,}6228}{41{,}71} 100\% = 11{,}08\%.$$

Для уравнения, содержащего две независимые переменные:

$$V_\sigma = \frac{3{,}3158}{41{,}71} 100\% = 7{,}95\%.$$

Итак, введение независимой переменной «урожайность» уменьшило среднюю квадратическую ошибку до величины порядка 7,95% среднего значения зависимой переменной.

Наконец, по формуле (6.21) рассчитаем коэффициент детерминации

$$R^2 = \frac{1623,8815}{1983,6183} 100\% = 81,86\%.$$

Он показывает, что уравнение регрессии на 81,9% объясняет колебания сбора хлеба на душу населения. Сравнивая полученный результат (81,9%) с величиной R^2 для однофакторного уравнения (77,9%), видим, что включение переменной «урожайность» заметно увеличило точность уравнения.

Таким образом, сравнение средних квадратических ошибок уравнения, коэффициентов вариации, коэффициентов детерминации, рассчитанных до и после введения независимой переменной, позволяет судить о полезности включения этой переменной в уравнение. Однако следует быть осторожными в выводах при подобных сравнениях, поскольку увеличение R^2 или уменьшение σ и V_σ не всегда имеют приписываемый им здесь смысл. Так, увеличение R^2 может объясняться тем фактом, что число рассматриваемых параметров в уравнении приближается к числу объектов наблюдения. Скажем, весьма сомнительными будут ссылки на увеличение R^2 или уменьшение σ, если в уравнение вводится третья или четвертая независимая переменная и уравнение строится на данных по шести, семи объектам.

Полезность включения дополнительного фактора можно оценить с помощью F-критерия.

Частный F-критерий показывает степень влияния дополнительной независимой переменной на результативный признак и может использоваться при решении вопроса о добавлении в уравнение или исключении из него этой независимой переменной.

Разброс признака, объясняемый уравнением регрессии (6.22), можно разложить на два вида: 1) разброс признака, обусловленный независимой переменной x_1, и 2) разброс признака, обусловленный независимой переменной x_2, когда x_1 уже включена в уравнение. Первой составляющей соответствует разброс признака, объясняемый уравнением (6.15), включающим только переменную x_1. Разность между разбросом признака, обусловленным уравнением (6.22), и разбросом признака, обусловленным уравнением (6.15), определит ту часть разброса, которая объясняется дополнительной независимой переменной x_2. Отношение указанной разности к разбросу признака, регрессией не объясняемому, представляет собой значение **частного F-критерия.** Частный F-критерий называется также последовательным, если статистические характеристики строятся при последовательном добавлении переменных в регрессионное уравнение.

Пример 16. Оценить полезность включения в уравнение регрессии дополнительной переменной «урожайность» (по данным и результатам примеров 12 и 15).

Разброс признака, объясняемый уравнением множественной регрессии и рассчитываемый как сумма квадратов разностей выравненных значений и их средней, равен 1623,8815. Разброс признака, объясняемый уравнением простой регрессии, составляет 1545,1331.

Разброс признака, регрессией не объясняемый, определяется квадратом средней квадратической ошибки уравнения и равен 10,9948 (см. пример 15).

Воспользовавшись этими характеристиками, рассчитаем частный F-критерий

$$F_{\text{ф}}[1,\ n-3]=\frac{1623{,}8815-1545{,}1331}{10{,}9948}=7{,}1623.$$

С уровнем надежности 0,95 ($\alpha=0{,}05$) табличное значение F (1,20), т. е. значение, стоящее на пересечении 1-го столбца и 20-й строки табл. 4А приложения, равно 4,35. Рассчитанное значение $F_{\text{ф}}$ значительно превосходит табличное, и, следовательно, включение в уравнение переменной «урожайность» имеет смысл.

Таким образом, выводы, сделанные ранее относительно коэффициентов регрессии, вполне правомерны.

Важным условием применения к обработке данных метода множественной регрессии является отсутствие сколько-нибудь значительной взаимосвязи между факторными признаками. При практическом использовании метода множественной регрессии, прежде чем включать факторы в уравнение, необходимо убедиться в том, что они независимы.

Если один из факторов зависит линейно от другого, то система нормальных уравнений, используемая для нахождения параметров уравнения, не разрешима. Содержательно этот факт можно толковать так: если факторы x_1 и x_2 связаны между собой, то они действуют на результативный признак y практически как один фактор, т. е. сливаются воедино и их влияние на изменение y разделить невозможно. Когда между независимыми переменными уравнения множественной регрессии имеется линейная связь, следствием которой является неразрешимость системы нормаль-

ных уравнений, то говорят о наличии мультиколлинеарности.

На практике вопрос о наличии или об отсутствии мультиколлинеарности решается с помощью показателей взаимосвязи. В случае двух факторных признаков используется парный коэффициент корреляции между ними: если этот коэффициент по абсолютной величине превышает 0,8, то признаки относят к числу мультиколлинеарных. Если число факторных признаков больше двух, то рассчитываются множественные коэффициенты корреляции. Фактор признается мультиколлинеарным, если множественный коэффициент корреляции, характеризующий совместное влияние на этот фактор остальных факторных признаков, превзойдет по величине коэффициент множественной корреляции между результативным признаком и совокупностью всех независимых переменных.

Самый естественный способ устранения мультиколлинеарности — исключение одного из двух линейно связанных факторных признаков. Этот способ прост, но не всегда приемлем, так как подлежащий исключению фактор может оказывать на зависимую переменную особое влияние. В такой ситуации применяются более сложные методы избавления от мультиколлинеарности [1].

Выбор «наилучшего» уравнения регрессии. Эта проблема связана с двойственным отношением к вопросу о включении в регрессионное уравнение независимых переменных. С одной стороны, естественно стремление учесть все возможные влияния на результативный признак и, следовательно, включить в модель полный набор выявленных переменных. С другой стороны, возрастает сложность расчетов и затраты, связанные с получением максимума информации, могут оказаться неоправданными. Нельзя забывать и о том, что для построения уравнения регрессии число объектов должно в несколько раз превышать число независимых переменных. Эти противоречивые требования приводят к необходимости компромисса, результатом которого и является «наилучшее» уравнение регрессии. Существует несколько методов, приводящих к цели: метод всех возможных регрессий, метод исключения, метод включения, шаговый регрессионный и ступенчатый регрессионный методы.

---

[1] См.: *Мот Ж.* Статистические предвидения и решения на предприятии. М., 1966; *Ковалева Л. Н.* Многофакторное прогнозирование на основе рядов динамики. М., 1980.

Метод всех возможных регрессий заключается в переборе и сравнении всех потенциально возможных уравнений. В качестве критерия сравнения используется коэффициент детерминации R^2. «Наилучшим» признается уравнение с наибольшей величиной R^2. Метод весьма трудоемок и предполагает использование вычислительных машин.

Методы исключения и включения являются усовершенствованными вариантами предыдущего метода. **В методе исключения** в качестве исходного рассматривается регрессионное уравнение, включающее все возможные переменные. Рассчитывается частный F-критерий для каждой из переменных, как будто бы она была последней переменной, введенной в регрессионное уравнение. Минимальная величина частного F-критерия ($F_{\min}$) сравнивается с критической величиной ($F_{кр}$), основанной на заданном исследователем уровне значимости. Если $F_{\min} > F_{кр}$, то уравнение остается без изменения. Если $F_{\min} < F_{кр}$, то переменная, для которой рассчитывался этот частный F-критерий, исключается. Производится перерасчет уравнения регрессии для оставшихся переменных, и процедура повторяется для нового уравнения регрессии. Исключение из рассмотрения уравнений с незначимыми переменными уменьшает объем вычислений, что является достоинством этого метода по сравнению с предыдущим.

Метод включения состоит в том, что в уравнение включаются переменные по степени их важности до тех пор, пока уравнение не станет достаточно «хорошим». Степень важности определяется линейным коэффициентом корреляции, показывающим тесноту связи между анализируемой независимой переменной и результативным признаком: чем теснее связь, тем больше информации о результирующем признаке содержит данный факторный признак и тем важнее, следовательно, введение этого признака в уравнение.

Процедура начинается с отбора факторного признака, наиболее тесно связанного с результативным признаком, т. е. такого факторного признака, которому соответствует максимальный по величине парный линейный коэффициент корреляции. Далее строится линейное уравнение регрессии, содержащее отобранную независимую переменную. Выбор следующих переменных осуществляется с помощью частных коэффициентов корреляции, в которых исключается влияние вошедших в модель факторов. Для каждой введенной переменной рассчитывается частный F-критерий, по величине которого судят о том, значим ли вклад этой переменной. Как только величина частного F-критерия, относящаяся к очередной переменной, оказывается незначимой, т. е. эффект от введения этой переменной становится малозаметным, процесс включения переменных заканчивается. Метод включения связан с меньшим объемом вычислений, чем предыдущие методы. Но при введении новой переменной нередко значимость включенных ранее переменных изменяется. Метод включения этого не учитывает, что является его недостатком. Модификацией метода включения, исправляющей этот недостаток, является шаговый регрессионный метод.

Шаговый регрессионный метод кроме процедуры метода включения содержит анализ переменных, включенных в уравнение на предыдущей стадии. Потребность в таком анализе возникает в связи с тем, что переменная, обоснованно введенная в уравнение на ранней стадии, может оказаться лишней из-за взаимосвязи ее с переменными, позднее включенными в уравнение. Анализ заключается в расчете на каждом этапе частных F-критериев **для каждой переменной** уравнения и сравнении их с величиной $F_{кр}$, точкой F-распределения, соответствующей заданному исследователем уровню значимости. Частный F-критерий показывает вклад переменной в вариацию результативного признака в предположении, что она вошла в модель последней, а сравнение его с $F_{кр}$ позволяет судить о значимости рассматриваемой переменной с учетом влияния позднее включенных факторов. Незначимые переменные из уравнения исключаются.

Рассмотренные методы предполагают довольно большой объем вычислений и практически неосуществимы без ЭВМ. Для реализации ступенчатого регрессионного метода вполне достаточно малой вычислительной техники.

Ступенчатый регрессионный метод включает в себя такую последовательность действий. Сначала выбирается наиболее тесно связанная с результативным признаком переменная и составляется уравнение регрессии. Затем находят разности фактических и выравненных значений и эти разности (остатки) рассматриваются как значения результативной переменной. Для остатков подбирается одна из оставшихся независимых переменных и т. д. На каждой стадии проверяется значимость регрессии. Как только обнаружится незначимость, процесс прекращается и окончательное уравнение получается суммированием уравнений, полученных на каждой стадии за исключением последней.

Ступенчатый регрессионный метод менее точен, чем предыдущие, но не столь громоздок. Он оказывается полезным в случаях, когда необходимо внести содержательные правки в уравнение. Так, для изучения факторов, влияющих на цены угля в Санкт-Петербурге в конце XIX — начале XX в., было получено уравнение множественной регрессии. В него вошли следующие переменные: цены угля в Лондоне, добыча угля в России и экспорт из России. Здесь не обосновано появление в модели такого фактора, как добыча угля, поскольку Санкт-Петербург работал исключительно на импортном угле. Модели легко придать экономический смысл, если независимую переменную «добыча» заменить независимой переменной «импорт». Формально такая замена возможна, поскольку между импортом и добычей существует тесная связь. Пользуясь ступенчатым методом, исследователь может совершить эту замену, если предпочтет содержательно интерпретируемый фактор.

§ 4. Нелинейная регрессия и нелинейная корреляция

Построение уравнений нелинейной регрессии. До сих пор мы, в основном, изучали связи, предполагая их линейность. Но не всегда связь между признаками может быть достаточно хорошо представлена линейной функцией. Иногда для описания существующей связи более пригодными, а порой и единственно возможными являются более сложные нелинейные функции. Ограничимся рассмотрением наиболее простых из них.

Одним из простейших видов нелинейной зависимости является парабола, которая в общем виде может быть представлена функцией (6.2):

$$y=a_0+a_1x+a_2x^2.$$

Неизвестные параметры a_0, a_1, a_2 находятся в результате решения следующей системы уравнений:

$$\begin{cases} na_0+a_1\sum_{i=1}^{n}x_i+a_2\sum_{i=1}^{n}x_i^2=\sum_{i=1}^{n}y_i, \\ a_0\sum_{i=1}^{n}x_i+a_1\sum_{i=1}^{n}x_i^2+a_2\sum_{i=1}^{n}x_i^3=\sum_{i=1}^{n}y_ix_i, \\ a_0\sum_{i=1}^{n}x_i^2+a_1\sum_{i=1}^{n}x_i^3+a_2\sum_{i=1}^{n}x_i^4=\sum_{i=1}^{n}y_ix_i^2. \end{cases} \tag{6.26}$$

Дает ли преимущества описание связи с помощью параболы по сравнению с описанием, построенным по гипотезе линейности? Ответ на этот вопрос можно получить, рассчитав последовательный F-критерий, как это делалось в случае множественной регрессии (см. пример 16).

На практике для изучения связей используются полиномы более высоких порядков (3-го и 4-го порядков). Составление системы, ее решение, а также решение вопроса о полезности повышения порядка функции для этих случаев аналогичны описанным. При этом никаких принципиально новых моментов не возникает, но существенно увеличивается объем расчетов.

Кроме класса парабол для анализа нелинейных связей можно применять и другие виды функций. Для расчета неизвестных параметров этих функций рекомендуется использовать метод наименьших квадратов, как наиболее мощный и широко применяемый.

Однако метод наименьших квадратов не универсален, поскольку он может использоваться только при условии, что выбранные для выравнивания функции линейны по отношению к своим параметрам. Не все функции удовлетворяют этому условию, но большинство применяемых на практике с помощью специальных преобразований могут быть приведены к стандартной форме функции с линейными параметрами.

Рассмотрим некоторые простейшие способы приведения функций с нелинейными параметрами к виду, который позволяет применять к ним метод наименьших квадратов.

Функция $y = a_0 a_1^{x_1} a_2^{x_2}$ не является линейной относительно своих параметров.

Прологарифмировав обе части приведенного равенства

$$\log y = \log a_0 + x_1 \log a_1 + x_2 \log a_2$$

и переобозначив

$$\log y = y', \ \log a_0 = a_0', \ \log a_1 = a_1', \ \log a_2 = a_2',$$

получим функцию, линейную относительно своих новых параметров:

$$y' = a_0' + a_1' x_1 + a_2' x_2.$$

Кроме логарифмирования для приведения функций к нужному виду используют обратные величины.

Например, функция

$$y = a_0 x / (x + a_1)$$

с помощью следующих переобозначений:

$$y'=1/y,\ x'=1/x,\ 1/a_0=a_0';\ a_1/a_0=a_1'$$

может быть приведена к виду

$$y'=a_0'+a_1'x'.$$

Подобные преобразования расширяют возможности использования метода наименьших квадратов, увеличивая число функций, к которым этот метод применим.

Измерение тесноты связи при криволинейной зависимости. Рассмотренные ранее линейные коэффициенты корреляции оценивают тесноту взаимосвязи при линейной связи между признаками. При наличии криволинейной связи указанные меры связи не всегда приемлемы. Разберем подобную ситуацию на примере.

Пример 17. В 1-м и 2-м столбцах табл. 5 приведены значения результативного признака y и факторного признака x (данные условные). Поставив вопрос о тесноте связи между ними, рассчитаем парный линейный коэффициент корреляции по формуле (6.3). Он оказался равным нулю, что свидетельствует об отсутствии линейной связи. Тем не менее связь между признаками существует, более того, она является функциональной и имеет вид

$$y=-6+7x-x^2.$$

Для измерения тесноты связи при криволинейной зависимости используется **индекс корреляции,** вычисляемый по формуле

$$i=\sqrt{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2\Big/\sum_{i=1}^{n}(y_i-\bar{y})^2}, \tag{6.27}$$

Таблица 5. Расчет линейного коэффициента корреляции

N	y_i	x_i	$y_i-\bar{y}$	$x_i-\bar{x}$	$(x_i-\bar{x})(y_i-\bar{y})$
	1	2	3	4	5
1	0	1	—3,3	—2,5	8,25
2	4	2	0,7	—1,5	—1,05
3	6	3	2,7	—0,5	—1,35
4	6	4	2,7	0,5	1,35
5	4	5	0,7	1,5	1,05
6	0	6	—3,3	2,5	—8,25
Σ	20	21			0

где y_i — i-е значение результативного признака; $\hat{y}_i$ — i-е выравненное значение этого признака; $\bar{y}$ — среднее арифметическое значение результативного признака.

Числитель формулы (6.27) характеризует разброс выравненных значений результативного признака. Поскольку изменения выравненных, т. е. вычисленных по уравнению регрессии, значений признака происходят только в результате изменения факторного признака x, то числитель измеряет разброс результативного признака, обусловленный влиянием на него факторного признака. Знаменатель же измеряет разброс признака-результата, который определен влиянием на него всех факторов, в том числе и учтенного. Таким образом, индекс корреляции оценивает участие данного факторного признака в общем действии всего комплекса факторов, вызывающих колеблемость результативного признака, тем самым определяя тесноту зависимости признака y от признака x. При этом, если признак x не вызывает никаких изменений признака y, то числитель и, следовательно, индекс корреляции равны 0. Если же линия регрессии полностью совпадает с фактическими данными, т. е. признаки связаны функционально, как в примере 17, то индекс корреляции равен 1. В случае линейной зависимости между x и y индекс корреляции численно равен линейному коэффициенту корреляции r. Квадрат индекса корреляции совпадает с введенным ранее (6.21) коэффициентом детерминации. Если же вопрос о форме связи не ставится, то роль коэффициента детерминации играет квадрат корреляционного отношения $\eta^2_{y/x}$ (6.12).

Таковы основные принципы и условия, методика и техника применения корреляционного и регрессионного анализа. Их подробное рассмотрение обусловлено тем, что они являются высокоэффективными и потому очень широко применяемыми методами анализа взаимосвязей в объективном мире природы и общества. Корреляционный и регрессионный анализ широко и успешно применяются и в исторических исследованиях (см. раздел III).

ГЛАВА 7
СТАТИСТИЧЕСКИЙ АНАЛИЗ ДИНАМИЧЕСКИХ РЯДОВ

§ 1. Временной ряд. Построение и средние характеристики

Изучение процессов и явлений во времени занимает важное место в исторических исследованиях и в плане математического обеспечения этих исследований имеет свою специфику. В настоящей главе мы остановимся специально на изучении так называемых временных рядов.

Временным рядом называется последовательность числовых значений, характеризующих изменение некоторого признака во времени. Отдельные значения признака, относящиеся к определенным промежуткам или моментам времени, принято называть **уровнями ряда.** Немало имеется статистических данных, из которых могут быть построены временные ряды, характеризующие те или иные процессы в истории. Как и в случае вариационных рядов, первый этап анализа данных, распределенных во времени, — построение временного ряда. При построении временных рядов могут возникнуть сложности, связанные с дефектами исходного материала, но некоторые трудности, являющиеся следствием несопоставимости данных или неполноты их, вполне преодолимы.

Требование сопоставимости при построении временных рядов. Условием правильно построенного временного ряда является сопоставимость составляющих его уровней. Рассмотрим наиболее распространенные причины несопоставимости уровней. Количественные данные могут относиться к территории, границы которой изменяются во времени. Например, при построении временных рядов на основе данных, охватывающих сведения по отдельным губерниям в досоветский период или областям и краям в советский период, следует иметь в виду изменения в их границах. Несопоставимость подобного рода устраняется пересчетом имеющихся данных с учетом изменившихся границ.

Несопоставимость может появиться в связи с различным пониманием единицы изучаемого объекта. Такие единицы изучаемого объекта, как «бедное хозяйство», «зажиточное хозяйство», «мелкое предприятие», «среднее предприятие», «крупное предприятие» и т. п., могут быть во времени по разному количественно определены. Для достижения сопоставимости в подобных случаях необходимо условиться об едином цензе и произвести соответствующий пересчет. Несопоставимость может возникнуть, если данные относятся к разным датам или разным периодам времени в течение года. Нарушения сопоставимости могут оказаться значительными, если рассматриваемые показатели подвержены сезонным или другим правильным периодическим колебаниям. Так, цены на сельскохозяйственные продукты, относящиеся к осенне-зимнему периоду, несопоставимы, например, с ценами, относящимися к весенне-летнему периоду

Несопоставимы уровни, представленные в различном масштабе измерения. Устранение этого недостатка в практике исторических исследований не всегда является формальной процедурой, что затрудняет построение временных рядов.

Построение временных рядов по неполным данным. При формировании временных рядов нередко возникают сложности, причиной которых является отсутствие нужных сведений. Некоторые незафиксированные или утерянные сведения могут быть «восстановлены». Разумеется, реконструированные тем или иным способом данные не могут всесторонне заменить реальные данные, но для решения многих вопросов вполне пригодны. «Восстановление» данных — задача трудная. Для ее решения требуется всесторонний анализ рядов. По имеющимся данным выявляются закономерности, которым подчиняется ряд, и, если таковые имеются и установлены, экстраполируют или интерполируют неизвестный уровень ряда. Определенные результаты можно получить и простыми средствами. Рассмотрим некоторые элементарные приемы построения временных рядов по неполным данным.

Иногда фиксирование показателей во времени прерывается по каким-либо причинам и продолжением ряда служат значения другого, несколько измененного показателя. Если приведенные ряды относятся к близким явлениям (например, цены на разные, но взаимозаменяемые сорта продукции, данные по группам с разным количеством объектов, данные, исчисленные по разным территориям и т. п.), естественно ожидать, что и динамика у них примерно одинакова. Тогда ряды можно сомкнуть. Два простых способа, которые можно использовать для смыкания рядов, покажем на примере.

Пример 1. В столбцах 1 и 2 табл. 1 приведены цены на пшеницу в г. Ельце за 1890—1898 гг. До 1895 г. включительно даны цены на гирку, а с 1895 г. фиксировались только цены местного сорта озимой пшеницы. Ряды по отдельности коротки для использования. Проведем смыкание рядов.

Принимаем уровень 1895 г. за 100% для обоих рядов и рассчитываем процентные отношения для первого (колонка 3) и второго (колонка 4) рядов. В колонке 5 приводится сомкнутый ряд в процентах.

Берем за основание первый ряд, продолжение его рассчитываем, применяя проценты, вычисленные для второго ряда (колонка 4). Результаты приведены в колонке 6.

Частым дефектом временных рядов является отсутствие некоторых уровней ряда. Существуют простые приемы оценки недостающих данных при наличии дополнительной

Таблица 1. Цены на пшеницу (коп. за пуд) в Ельце[1]

Год	1	2	3	4	5	6
1890	81		135,0		135	81
1891	108		180,0		180	108
1892	106		176,7		177	106
1893	87		145,0		145	87
1894	64		106,7		107	64
1895	60	62	100,0	100,0	100	60
1896		64		103,2	103	62
1897		84		135,5	136	81
1898		107		172,6	173	104

[1] Своды товарных цен на главных русских и иностранных рынках. СПб., 1900, с. 2.

информации. Если вместе с дефектным рядом имеются полноценные ряды однотипных показателей, то, пользуясь определенными соотношениями этих рядов, можно оценить отсутствующие данные. Покажем один из способов на конкретном примере.

Пример 2. В табл. 2 представлены цены на пшеницу в Саратове. Цена 1894 г. отсутствует, но есть цены за все годы на пшеничную муку. Между ценами пшеницы и ценами муки существует довольно устойчивое соотношение. Рассчитаем средние за 1890—1893 гг. цены на пшеницу и на муку.

Цены равны соответственно 91,9 и 153,8. Разница между ними составляет 153,8 коп. — 91,9 коп. = 61,9 коп., т. е. в среднем за эти годы цена муки превышала цену пшеницы на 61,9 коп. за пуд. Предполагая, что это соотношение приблизительно верно и для 1894 г., получим примерную цену пшеницы: 120,2—61,9 = 58,3 коп.

Экстраполяция в этом случае оказалась вполне удачной: действительная цена пшеницы в Саратове в 1894 г.

Таблица 2. Цены на пшеницу и пшеничную муку (коп. за пуд) в Саратове[1]

	1890 г.	1891 г.	1892 г.	1893 г.	1894 г.
Цена на пшеницу	80,8	105,4	100,3	81,1	—
Цена на муку	145,7	166,2	156,1	147,1	120,2

[1] Своды товарных цен на главных русских и иностранных рынках. СПб., 1900, с. 2.

была равна 59,1 коп. за пуд, т. е. расхождение между фактической и восстановленной ценами не превышает одной копейки за пуд.

Аналогичным способом можно интерполировать отсутствующие данные.

Средние характеристики временного ряда. Обобщенными характеристиками временного ряда являются средняя хронологическая, средний абсолютный прирост, средний темп роста и средний темп прироста.

Средняя хронологическая показывает, каким уровнем в среднем характеризуется временной ряд, и рассчитывается по формуле

$$\bar{y}=\sum_{i=1}^{n} y_i/n, \tag{7.1}$$

где y_i — i-й уровень ряда; n — число членов или уровней ряда.

Средние хронологические полезны для сравнительного анализа двух или нескольких временных рядов. Скажем, сравнение средних уровней урожайности за ряд лет для различных районов или стран более показательно для характеристики этих районов или стран, чем сравнение уровней за отдельные годы. Для характеристики изменения временного ряда и скорости изменения в среднем служат такие показатели, как средний абсолютный прирост и средний темп роста (прироста).

Средний абсолютный прирост показывает, насколько быстро изменяется конечный уровень ряда относительно первоначального, и вычисляется по формуле

$$\overline{\Delta y}=(y_n-y_1)/(n-1), \tag{7.2}$$

где y_n и y_1 — последний и первый уровни ряда соответственно.

Пример 3. Рассчитать средний абсолютный прирост добычи каменного угля в России за период 1890—1914 гг. (исходные данные приведены в табл. 3).

Пользуясь формулой (7.2), имеем:

$$\overline{\Delta y}=\frac{2176-367}{24}=\frac{1809}{24}=\mathbf{75{,}4},$$

т. е. в среднем прирост добычи каменного угля в России составлял 75,4 млн. пуд. в год.

Таблица 3. Выравнивание временного ряда различными способами[1]

Год	Добыча каменного угля (тыс. пуд.)	Выравнивание с использованием				
		средней абсолютного прироста	средней темпа роста	тройной скользящей средней	скользящей 4-членной средней	центрированной 4-членной средней
1890	367	367	367			
1891	380	442	395	390	409,0	
1892	424	518	425	423	449,0	429
1893	465	593	458	472	492,8	471
1894	527	669	493	516	529,8	511
1895	555	744	531	551	584,5	557
1896	572	819	572	604	640,5	612
1897	684	895	616	669	715,0	678
1898	751	970	663	763	818,5	767
1899	853	1046	714	863	899,8	859
1900	986	1121	769	949	963,2	932
1901	1009	1196	828	1000	1022,8	993
1902	1005	1272	892	1035	1075,5	1049
1903	1091	1347	961	1098	1108,2	1092
1904	1197	1423	1035	1143	1188,5	1148
1905	1140	1498	1115	1221	1312,5	1250
1906	1326	1573	1201	1351	1362,5	1338
1907	1587	1649	1293	1437	1486,8	1425
1908	1397	1724	1393	1540	1536,5	1512
1909	1637	1800	1500	1520	1573,5	1555
1910	1525	1875	1616	1632	1700,2	1637
1911	1735	1950	1740	1721	1840,0	1770
1912	1904	2026	1874	1945	2002,8	1921
1913	2196	2101	2018	2092		
1914	2176	2177	2173			

[1] Исходные данные — добыча каменного угля в России (столбец 1) — взяты в кн.: Обзор внешней торговли России по европейской и азиатской границам. Пб., 1891—1915.

Средний коэффициент роста[1] характеризует среднюю скорость изменения явления или процесса и определяется по формуле

$$\overline{K}_{р}=\sqrt[n-1]{y_n/y_1}. \tag{7.3}$$

Средний коэффициент прироста отличается от среднего темпа роста на единицу, т. е.

$$\overline{K}_{пр}=\sqrt[n-1]{y_n/y_1}-1. \tag{7.4}$$

[1] Средний коэффициент роста и средний коэффициент прироста, выраженные в процентах, называются средним темпом роста ($T_р$) и средним темпом прироста ($T_{пр}$) соответственно.

Пример 4. Вычислить средний коэффициент роста и средний коэффициент прироста добычи каменного угля в России за период 1890—1914 гг.

По данным табл. 3

$$\overline{K}_{р} = \sqrt[24]{2176/367} = \sqrt[24]{5,9292}\ .$$

Чтобы вычислить корень 24-й степени из 5,9292, прологарифмируем обе части последнего равенства:

$$\lg \overline{K}_{р} = \frac{1}{24} \lg 5,9292$$

Воспользовавшись таблицей логарифмов, будем иметь

$$\lg \overline{K}_{р} = \frac{1}{24} \cdot 0,722996 = 0,03221\ .$$

Обратившись к таблице антилогарифмов, находим средний коэффициент роста $\overline{K}_{р} = 1,077$. Тогда $\overline{K}_{пр} = 0,077$, т. е. в среднем за год прирост добычи каменного угля в России составлял 7,7%.

§ 2. Составляющие временного ряда и их анализ

Составляющие временного ряда. Временные ряды, составленные из показателей, характеризующих социально-экономические процессы, имеют свою специфику. Прежде всего эти показатели имеют более или менее выраженную общую для достаточно длительного периода тенденцию к возрастанию или убыванию во времени. Выявление этой составляющей, так называемого **временного тренда,** достигается механическим сглаживанием или аналитическим выравниванием рядов.

Составляющей многих временных рядов являются **сезонные колебания.** Сезонные колебания есть результат влияния смен времен года на те или иные явления и их показатели. Например, цены на сельскохозяйственные товары, объем перевозок некоторых товаров и многие другие показатели содержат сезонные колебания. Реже временные ряды имеют более или менее правильную периодичность, подобную сезонной, но с большим периодом, растянутым на несколько лет. Эту составляющую часть связывают с цикличностью капиталистической экономики и называют **циклическими колебаниями.** Кроме перечисленных составляющих, характеризующих устойчивые изменения показателя, временные ряды содержат изменения частные, случайные или так называемую **случайную компоненту.**

Временной тренд. Временной тренд является результатом длительного воздействия на изучаемый процесс комплекса главных факторов, он отражает существенные типические черты данного процесса. Выявление и выделение временного тренда — важный момент анализа временного ряда.

Обратимся к данным по добыче угля в России за 25 лет (см. табл. 3). Движение этого показателя изобразим ломаной линией (см. рис. 16). В конце XIX в. добыча угля в России уверенно растет, в начале XX в. рост ослабевает, добыча даже несколько падает, но затем рост возобновляется. В 1905, 1908, 1910 и 1914 гг. наблюдаем уменьшение добычи, но, несмотря на временные спады, ряд имеет довольно четко выраженную тенденцию к возрастанию. Существует несколько способов, позволяющих определить основную тенденцию временного ряда, выделить временной тренд.

Выявить общую тенденцию временного ряда можно, воспользовавшись средним абсолютным приростом или средним темпом прироста. Эти способы можно использовать в случае более или менее равномерного изменения явления. Неплохие результаты в выделении тренда достигаются механическим сглаживанием. Этот способ приводит к укорачиванию временного ряда, что, в частности, ограничивает его применение. Эффективным способом определения общей тенденции ряда является аналитическое выравнивание.

Использование среднего абсолютного прироста и среднего коэффициента роста для выявления тренда. Если уровни временного ряда включают в себя две компоненты — тренд и случайные от него отклонения, то результативными и простыми в исполнении являются приемы обработки ряда, использующие средний абсолютный прирост или средний коэффициент роста. Выбор показателя для выравнивания зависит от типа ряда. Если ряд изменяется приблизительно по закону арифметической прогрессии, то используют средний абсолютный прирост. Если нарастание или убывание ряда близко к закону геометрической прогрессии, то применяют средний коэффициент роста.

Если для нахождения тренда используют средний абсолютный прирост, то за первый уровень ряда принимается фактический уровень, второй уровень получается прибавлением к первому среднего абсолютного прироста. Для получения третьего уровня средний абсолютный прирост суммируется с предыдущим преобразованным уровнем или, что равносильно, к первому фактическому уровню прибавляется $2\overline{\Delta y}$ и т. д., для получения n-го уровня к первому уровню прибавляется $(n-1)\overline{\Delta y}$.

При использовании среднего коэффициента роста для выявления тренда за первый уровень, как и в предыдущем способе, принимается фактический уровень, но для получения второго уровня первый уровень умножается на средний коэффициент роста. Для получения третьего уровня первый уровень умножается на средний коэффициент роста во второй степени и т. д., n-й уровень вычисляется умножением $(K_р)^{n-1}$ на первый уровень. Эта процедура равносильна следующей, более удобной для расчетов: первый и второй уровни получаются, как указано; для нахождения третьего уровня второй преобразованный уровень умножается на $K_р$ и т. д.; наконец, для определения n-го уровня $(n-1)$-й преобразованный уровень умножается на $K_р$.

Пример 5. Выравнять ряд, представляющий добычу каменного угля в России, используя: а) средний абсолютный прирост, б) средний коэффициент роста.

Показатели, нужные для нахождения тренда, найдены в примерах 3 и 4: средний абсолютный прирост $\overline{\Delta y}=75{,}4$, средний коэффициент роста $\bar{K}_р=1{,}077$.

а) Воспользуемся средним абсолютным приростом для нахождения плавных выравненных уровней ряда: первый уровень = первому фактическому уровню = 367, второй уровень = $367+75{,}4\approx 380$, третий уровень = $367+75{,}4\cdot 2\approx 424$.

Результаты расчетов для всего ряда приведены в 3-м столбце табл. 3.

б) Возьмем для выравнивания средний коэффициент роста: первый уровень = первому фактическому уровню = 367, второй уровень = $=367\cdot 1{,}077\approx 395$, третий уровень $=395\cdot 1{,}077\approx 425$, и т. д.

Преобразованные уровни для всего ряда записаны в 4-м столбце табл. 3.

Сравнивая выравненные тем и другим способом значения с соответствующими фактическими, видим, что первый способ неэффективен для наших данных, второй дает более точные результаты: выравненные значения ближе к фактическим.

Механическое сглаживание или использование скользящей средней для выравнивания является одним из наиболее известных и давно используемых методов. При определении средних значений случайные отклонения погашаются — такая идея используется в методе скользящей средней. Фактические значения ряда заменяются средними значениями, в которых в той или иной степени нейтрализуются случайные колебания.

Процедура выравнивания начинается с выбора интервала сглаживания. Чем больше интервал, тем более плавным получается тренд. Но с увеличением интервала сглаживания усложняются расчеты и укорачивается преобразованный ряд. В поисках компромисса следует также учитывать цикличность рядов. Если ряд имеет более или менее выраженную цикличность, то длина цикла и принимается за интервал сглаживания.

От величины интервала сглаживания зависит техника расчета. Если число уровней интервала сглаживания нечетное, то рассчитанное значение средней однозначно приписывается срединному уровню интервала. Затем интервал смещается на один уровень (скользит — отсюда и название метода), снова рассчитывается средняя арифметическая, которая приписывается среднему члену смещенного интервала, и т. д. Если число уровней интервала сглаживания четное, то серединой этого интервала будет промежуток между уровнями, и рассчитанное значение средней нельзя отнести ни к одному уровню ряда. Эту сложность обходят с помощью центрирования. Интервал сдвигается еще на один уровень, причем середина его приходится на соседний промежуток. Полусумма из средних арифметических соседних интервалов приписывается уровню, расположенному между серединами этих интервалов.

Пример 6. Сгладить ряд, представляющий добычу каменного угля в России, методом скользящей средней с интервалом скольжения: а) в три года, б) в четыре года. Исходные данные приведены во 2-м столбце табл. 3.

а) Для получения первого уровня сглаженного ряда рассчитываем среднюю арифметическую для трех первых уровней исходного ряда:

$$\bar{x}_1 = \frac{367 + 380 + 424}{3} = 390$$

и приписываем это значение срединной точке интервала сглаживания, т. е. уровню, соответствующему 1891 г.

Для получения второго уровня сглаженного ряда сдвинем интервал сглаживания на один уровень, а среднюю арифметическую

$$\bar{x}_2 = \frac{380 + 424 + 465}{3} = 423$$

отнесем к середине сдвинутого интервала, т. е. к 1892 г.

Аналогично получаем последующие значения выравненного ряда (см. столбец 5 табл. 3).

б) Среднюю арифметическую первых четырех уровней

$$\bar{x}_{11} = \frac{367 + 380 + 424 + 465}{4} = 409$$

отнесем к середине интервала сглаживания, т. е. к промежутку между 1891 и 1892 гг. Сдвинутый на уровень интервал будет иметь серединой промежуток между 1892 и 1893 гг. и среднюю арифметическую:

$$\bar{x}_{12} = \frac{380 + 424 + 465 + 527}{4} = 449.$$

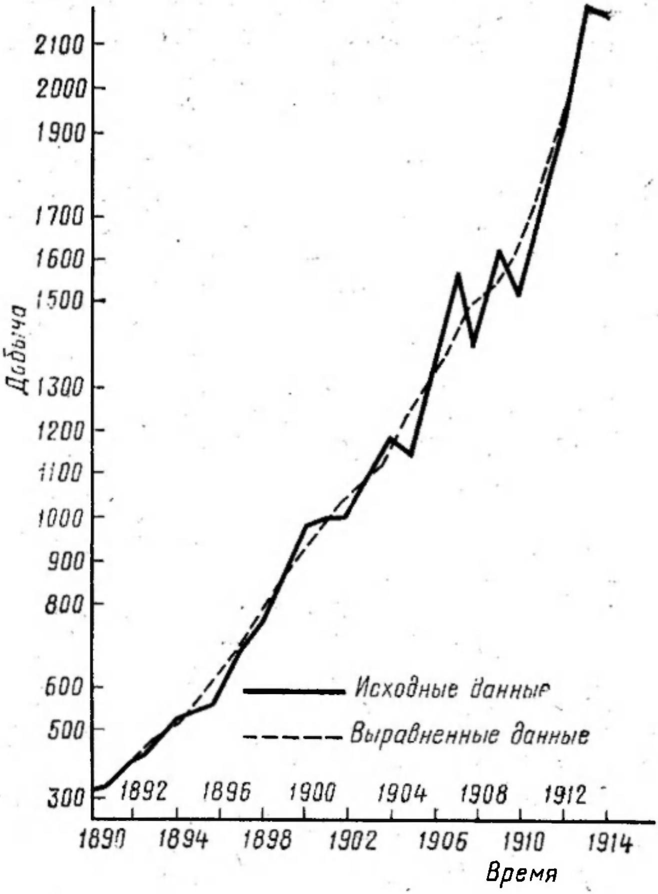

Рис. 16. График добычи угля в России в 1890—1914 гг.

Чтобы получить скользящую среднюю, относящуюся к определенному году, применяем центрирование, т. е. находим «центр» рядом расположенных интервалов (1891—1892 гг. и 1892—1893 гг.), а также находим «центр» или среднее значение пары найденных скользящих средних, относящихся к этим интервалам. «Центром» интервалов является 1892 г., а среднее значение пары скользящих средних вычислим как среднюю арифметическую этой пары:

$$\bar{x}_1 = \frac{409 + 449}{2} = 429.$$

Таким образом, первый уровень сглаженного ряда относится к 1892 г. и равен 429. Аналогично получаем последующие значения.

Результаты расчетов по сглаживанию ряда с помощью четырехчленной скользящей средней сведены в столбце 6 табл. 3 и нанесены на график (рис. 16).

К достоинствам метода скользящей средней относится прежде всего его простота. Этот метод имеет более широкий спектр действия, чем предыдущие способы, которые могут использоваться только в случае более или менее монотонного нарастания или убывания показателей ряда.

Основным недостатком метода является потеря крайних (начальных и конечных) уровней ряда. Увеличение интервала сглаживания обычно приводит к лучшим результатам выравнивания, но с увеличением этого интервала растут потери уровней ряда, что особенно существенно для коротких рядов. Кроме того, тренд, полученный способом скользящей средней, не имеет математического выражения, что делает значительно более сложным анализ тренда, в особенности сравнительный анализ нескольких трендов.

Аналитическое выравнивание. Основным методом аналитического выравнивания является метод наименьших квадратов. В приложении к временным рядам метод наименьших квадратов дает возможность найти функцию от времени $y(t)$, описывающую с большей или меньшей точностью тренд изучаемого процесса. Для определения параметров функции необходимо решить соответствующую систему нормальных уравнений.

Существует прием[1], позволяющий упрощать эти системы для временных рядов. Суть его заключается в следующем: за значения t_i принимаются такие симметричные по абсолютному значению относительно середины ряда и разнящиеся знаком величины, что сумма $\sum_{i=1}^{n} t_i' = 0$, где t_i' — преобразованные значения t_i. Это равенство, а также следующее из него равенство $\sum_{i=1}^{n} (t_i')^3 = 0$ позволяют существенно упрощать системы.

Подбор соответствующих t_i' зависит от числа уровней в ряду. Если число уровней временного ряда нечетное, то из всех значений t_i вычитается значение t, приходящееся на середину ряда, и разности делятся на Δt. Тогда срединное преобразованное значение t' становится равным нулю, значения t_i', соответствующие ранним датам или периодам, приобретают отрицательные значения (—1, —2, —3

[1] Этот прием пригоден для рядов, у которых значения t подчиняются закону арифметической прогрессии, т. е. нарастают (или убывают) на одну и ту же величину Δt. Временные ряды чаще всего имеют именно такой вид.

и т. д.), а значения t_i'. соответствующие поздним датам или периодам, — положительные значения (1, 2, 3 и т. д.).

При четном числе уровней ряда нужный эффект достигается следующим образом: срединной паре приписываются значения, равные -1 и $+1$, далее, вверх от этой пары идут отрицательные, а вниз — положительные числа: 3, 5, 7 и т. д.

Так, неизвестные параметры прямой

$$y(t) = a_0 + a_1 t' \qquad (7.5)$$

находят из системы уравнений:

$$\begin{cases} na_0 = \sum_{i=1}^{n} y_i, \\ a_1 \sum_{i=1}^{n} t_i^2 = \sum_{i=1}^{n} y_i t_i, \end{cases} \qquad (7.6)$$

откуда

$$a_0 = \frac{\sum_{i=1}^{n} y_i}{n}, \quad a_1 = \frac{\sum_{i=1}^{n} y_i t_i}{\sum_{i=1}^{n} t_i^2}. \qquad (7.7)$$

Пример 7. Данные по добыче угля в России выравнять по прямой линии методом наименьших квадратов.

Для решения поставленной задачи составим расчетную таблицу (табл. 4, во 2-м столбце которой приведены исходные данные).

По формулам (7.7) рассчитав необходимые параметры, получим: $a_0 = 1099{,}6$; $a_1 = 74{,}3$.

Таким образом, добыча каменного угля в России может быть приближенно представлена как линейная функция от времени

$$y = 1099{,}6 + 74{,}3t'.$$

Если число n уровней ряда нечетно, то выравнивание упрощенным способом не изменяет величины и смысла коэффициентов регрессии по сравнению с обычным выравниванием (см. § 3 гл. 6). Коэффициент регрессии a_1 показывает, на какую величину в среднем изменяется анализируемый показатель при изменении независимой переменной, т. е. времени, на единицу.

Если число n уровней ряда четно, то при выравнивании упрощенным методом длина интервала, т. е. расстояние между соседними преобразованными t', удваивается и величина параметра a_1 соответственно уменьшается в два раза.

Таблица 4. Выравнивание временного ряда по прямой линии

Год	y	t'	$t'y$	$(t')^2$	Выравненные по прямой	Отклонения фактич. значений от выравн.	Квадраты отклонений
1890	367	—12	—4404	144	208	159	25281
1891	380	—11	—4180	121	282	98	9604
1892	424	—10	—4240	100	357	67	4489
1893	465	—9	—4185	81	431	34	1156
1894	527	—8	—4216	64	505	22	484
1895	555	—7	—3885	49	580	—25	625
1896	572	—6	—3432	36	654	—82	6724
1897	684	—5	—3420	25	728	—44	1936
1898	751	—4	—3004	16	802	—51	2601
1899	853	—3	—2559	9	877	—24	576
1900	986	—2	—1972	4	951	35	1225
1901	1009	—1	—1009	1	1025	—16	256
1902	1005	0	0	0	1100	—95	9025
1903	1091	1	1091	1	1174	—83	6889
1904	1197	2	2394	4	1248	—51	2601
1905	1140	3	3420	9	1322	—182	33124
1906	1326	4	5304	16	1397	—71	5041
1907	1587	5	7935	25	1471	116	13456
1908	1397	6	8382	36	1545	—148	21904
1909	1637	7	11 459	49	1620	17	289
1910	1525	8	12 220	64	1694	—169	28561
1911	1735	9	15 615	81	1768	—33	1089
1912	1904	10	19 040	100	1843	61	3721
1913	2196	11	24 156	121	1917	279	77841
1914	2176	12	26 112	144	1991	185	34225
Σ	27 489		96 602	1300			292 723

Параметр a_0 независимо от числа уровней в ряду интерпретируется как средний уровень ряда.

Возвращаясь к примеру 7, отметим, что добыча каменного угля в России в 1890—1914 гг. нарастала в среднем за год на 74,3 тыс. пудов. Средний уровень добычи за эти годы составлял примерно 1100 тыс. пудов.

Аналитическое выравнивание по параболе сводится к нахождению параметров функции

$$y(t)=a_0+a_1t+a_2t^2. \tag{7.8}$$

При использовании упрощенного метода неизвестные параметры a_0, a_1, a_2 находятся из системы:

$$\begin{cases} na_0+a_2\sum\limits_{i=1}^{n}(t_i')^2=\sum\limits_{i=1}^{n}y_i, \\ a_1\sum\limits_{i=1}^{n}(t_i')^2=\sum\limits_{i=1}^{n}y_it_i', \\ a_0\sum\limits_{i=1}^{n}(t_i')^2+a_2\sum\limits_{i=1}^{n}(t_i')^4=\sum\limits_{i=1}^{n}y_i(t_i')^2. \end{cases} \tag{7.9}$$

Пример 8. Выравнивание по параболе данных о добыче угля в России дает такой результат: $a_1=74{,}3$; $a_0=1014{,}7$; $a_2=1{,}6$.

Уравнение, следовательно, имеет вид $y=1014{,}7+74{,}3t'+1{,}6(t')^2$.

Показатели динамики процесса. Общеупотребительным показателем динамики временного ряда является темп роста, величина которого показывает, как быстро изменяются уровни ряда во времени. В качестве обобщенной характеристики временного ряда используются средний темп роста, а также отличающийся от него на константу средний темп прироста.

Эти показатели широко применяются в практике исследований, но имеют некоторые недостатки. Прежде всего следует отметить заметно выраженную зависимость их от случайных резких колебаний крайних уровней ряда. Кроме того, эти показатели не учитывают промежуточных значений временных рядов.

Чтобы избавиться от указанных недостатков, используют показатель динамики, построенный на выравненных значениях:

$$\overline{K}_{\text{в}}=\sqrt[n-1]{\hat{y}_n/\hat{y}_1}, \tag{7.10}$$

где $\hat{y}_n$ и $\hat{y}_1$ — конечный и начальный уровни выравненного ряда соответственно.

Действительно, показатель устойчив, мало зависит от начального и конечного (фактических) уровней временного ряда и учитывает тенденцию ряда. Рассмотренный показатель может быть использован для исследования динамики сильно колеблющихся рядов.

Однако он является недостаточно чутким инструментом при сравнительном анализе рядов с близкими тенденциями развития.

В некоторых случаях в качестве показателей динамики можно использовать первые производные функций, описывающих ряд.

Для параболы $y=a_0+a_1t+a_2t^2$ первая производная вычисляется по формуле

$$\frac{dy}{dt}=a_1+2a_2t; \tag{7.11}$$

для прямой линии $y=a_0+a_1t$ первая производная совпадает с коэффициентом регрессии a_1.

Первая производная функции показывает скорость изменения этой функции и в этом смысле служит показателем динамики. Этот показатель устойчив, учитывает промежуточные значения ряда и дает различимые характеристики рядов с близкими тенденциями развития. Вычисление первой производной любой функции не представляет технических трудностей, но только для прямой линии и особенно для параболы использование первой производной в качестве показателя динамики оказывается полезным.

Пример 9. Рассчитать средний коэффициент роста добычи угля в России, пользуясь выравненными по параболе данными.

Используя результаты примера 8 и формулу (7.10), имеем

$$\overline{K}_в=\sqrt[24]{2137/354}=\sqrt[24]{6{,}0367}=1{,}078.$$

В данном случае средний коэффициент роста, вычисленный по выравненным данным, мало отличается от среднего коэффициента роста, рассчитанного по фактическим данным.

Средний коэффициент роста, вычисленный как тем, так и другим способом, показывает среднюю за изучаемый период скорость изменения показателя, но не дает представления о том, как во времени изменяется сама скорость. Первая производная параболы дает такую возможность.

Пример 10. Проанализировать динамику роста добычи угля в России методом производных.

Воспользовавшись результатами примера 8 и формулой (7.11), получим

$$\frac{dy}{dt}=74{,}3+3{,}2t'$$

Подставив в полученное выражение значения $t=-12$ и $t'=+12$, построим на графике производную (прямая *I* на рис. 17).

Легко видеть, что скорость добычи угля в России не остается постоянной, она растет. Действительно, с течением времени (при движении по оси времени вправо) значения скорости роста добычи увеличиваются.

Использование производных эффективно при сравнительном анализе скоростей нескольких рядов показателей. Сравним скорости роста добычи, ввоза из-за границы и потребления угля в России. Опуская расчеты, аналогичные тем, что были проведены для добычи угля, приведем их результаты для импорта угля и потребления (прямые *II* и *III* на рис. 17). На графике ясно видно, что скорость роста добычи угля существенно выше скорости ввоза — в любой точке времени значение скорости роста добычи превышает соответствующее той же точке значение скорости роста импорта. Кроме того, близость прямых, одна из которых показывает скорость роста потребления, а другая — скорость роста одного из обеспечивающих потребление факторов (скорость роста добычи), можно истолковать так: скорость роста добычи прежде всего и определяет скорость роста потребления.

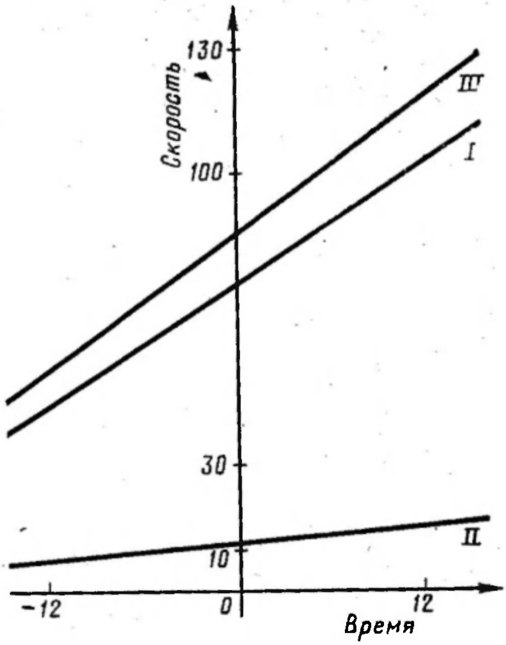

Рис. 17. Динамика роста добычи угля в России: *I* — скорость роста добычи; *II* — скорость роста импорта; *III* — скорость роста потребления

О выборе вида выравнивания. Выше были изложены способы выравнивания временных рядов. Какой из них предпочесть?

Если ряд возрастает или убывает равномерно, то разумно воспользоваться для выравнивания средним абсолютным приростом или средним коэффициентом роста. Причем, если ряд изменяется приблизительно по арифметической прогрессии, т. е. каждый уровень ряда отличается от соседнего примерно на постоянную величину, то применяют первый из этих способов. Если ряд изменяется приблизительно по геометрической прогрессии, т. е. каждый уровень ряда отличается от соседнего примерно в одно и то же число раз, то для выравнивания берут второй способ.

Если условие равномерного изменения ряда не выполняется, то пользуются механическим сглаживанием или аналитическим выравниванием. Механическое сглаживание дает хорошие результаты для достаточно длинных рядов. Основным недостатком метода является отсутствие математического описания тренда. Этот недостаток становится несущественным, если сглаживание проводится не с целью анализа трендов, а с целью исключить тренд. Потребность в исключении тренда появляется, например, при выявлении сезонной волны, при расчете коэффициентов корреляции и т. п.

Универсальным, но более сложным в исполнении является аналитическое выравнивание. Главный вопрос, требующий решения перед аналитическим выравниванием,— выбор функции для выравнивания.

Тип функции может быть подсказан графиком, изображающим анализируемый временной ряд. Вид функции для временных рядов можно определить также с помощью **метода разностей.**

Рассмотрим процедуру использования метода разностей для выбора вида функции.

Разности между последовательными уровнями временного ряда называются **первыми разностями** и обозначаются Δy. Разности между последовательными значениями первых разностей называются **вторыми разностями** и обозначаются $\Delta^2 y$. Аналогично образуются третьи, четвертые и т. д. разности.

Если k-е разности постоянны по величине, а $(k+1)$-е разности близки к нулю, то зависимость показателя от времени может быть представлена уравнением многочлена k-й степени:

$$y = a_0 + a_1 t + a_2 t^2 + \ldots + a_k t^k. \tag{7.12}$$

Пример 11. В табл. 5 приводятся данные о численности населения за 1893—1900 гг. Какова закономерность изменения численности населения России в эти годы?

Вычисляем последовательно первые и вторые разности (Δy и $\Delta^2 y$).

Результаты расчетов приведены в столбцах 4 и 5 расчетной табл. 5. Поскольку первые разности изменяются незначительно, почти постоянны, а вторые разности близки к нулю, можно говорить о том, что численность населения в эти годы линейно зависит от времени.

Данные таковы, что несложно вывести и формулу этой зависимости. Для этого прежде всего преобразуем t: 1893 году припишем условно 0, 1894 году — 1 и т. д. (см. столбец 1).

Таблица 5. Численность населения России (млн. человек) [1]

Годы	t_i'	y_i	Δy	$\Delta^2 y$	y_i'
1893	0	120,1			$120{,}1 = 120{,}1 + 1{,}6 \cdot 0$
1894	1	121,6	1,5		$121{,}7 = 120{,}1 + 1{,}6 \cdot 1$
1895	2	123,2	1,6	0,1	$123{,}3 = 120{,}1 + 1{,}6 \cdot 2$
1896	3	124,8	1,6	0,0	$124{,}9 = 120{,}1 + 1{,}6 \cdot 3$
1897	4	126,4	1,6	0,0	$126{,}5 = 120{,}1 + 1{,}6 \cdot 4$
1898	5	128,1	1,7	0,1	$128{,}1 = 120{,}1 + 1{,}6 \cdot 5$
1899	6	129,9	1,8	0,1	$129{,}7 = 120{,}1 + 1{,}6 \cdot 6$
1900	7	131,7	1,8	0,0	$131{,}3 = 120{,}1 + 1{,}6 \cdot 7$

[1] Сборник сведений по статистике внешней торговли России /Под ред. В. И. Покровского. СПб., 1902, т. I, с. XXXIV.

Первые разности почти постоянны и в среднем равны 1,6, т. е. при переходе от одного уровня к другому происходит нарастание на 1,6 млн. человек. Тогда второй уровень можно представить как сумму первого уровня и среднего прироста уровней за год. Рассуждая аналогично, запишем все уровни ряда как сумму первого уровня и соответствующего числа приростов (последний столбец табл. 5). Все эти записи можно объединить в одной $y' = 120{,}1 + 1{,}6t_i'$, где t_i' принимает значения от 0 до 7.

Следует отметить, что на практике редко встречаются временные ряды, так легко укладывающиеся в определенную схему, как этот. Обычно метод разностей применяется только для оценки порядка уравнения, а для нахождения параметров уравнения используется метод наименьших квадратов.

Сезонные колебания. Многие временные ряды экономических показателей содержат в себе сезонную волну, что необходимо учитывать при анализе этих рядов. Скажем, перевозки хлеба по железной дороге возрастают осенью и в начале зимы в связи с реализацией урожая и уменьшаются весной и летом. Если на первый план выдвигается элиминирование сезонных колебаний с целью изучения остающихся составляющих, то можно использовать приемы построения ряда, исключающие волну. Так, можно составить ряд, уровни которого относятся к определенной фиксированной дате каждого года, или в качестве уровней брать среднегодовые показатели, которые гасят сезонную волну. Эти простые приемы элиминируют сезонные колебания, но приводят к значительной потере информации. Потери тем чувствительнее, чем короче исходные ряды. Если предметом изучения являются сезонные колебания или агрегирование показателей нецелесообразно, то ставится задача выделения сезонных колебаний.

Использование метода скользящей средней для выделения сезонных колебаний. Если исходные данные представлены месячными показателями, то сезонная волна, по длине равная году, содержится в 12-членном ряду (по числу месяцев в году). Если же исходный материал дан в виде поквартальных показателей, то сезонная волна помещается в 4-членном ряду. Мы ограничимся рассмотрением метода скользящей средней для выделения сезонных колебаний на базе месячных данных. Для решения поставленной задачи рассчитывают 12-членные скользящие средние, совокупность которых можно рассматривать как тенденцию данного ряда, свободную от сезонных колебаний. Затем фактические значения ряда относят к соответствующим значениям вычисленной тенденции и выражают результаты в процентах. Рассчитанные таким образом величины показывают отклонения фактических значений от тренда. Избавившись от тренда и полагая, что оставшиеся изменения показателя связаны с сезонными колебаниями, рассчитывают величину изменений для каждого месяца в среднем за все годы.

Пример 12. В табл. 6 приведены ежемесячные цены на озимую пшеницу в Ростове-на-Дону за 1902—1909 гг. Цены колеблются, но в их изменении можно легко усмотреть определенную направленность. Во-первых, цены совершенно явно растут, во-вторых, цены испытывают сезонные колебания: летом и осенью цены ниже, чем зимой и весной. Последняя закономерность искажается общей тенденцией цен к повышению. Скажем, осенние цены 1904 г. выше зимних цен 1903 г. Чтобы выделить изменения цен, связанные с сезонными колебаниями, необходимо избавиться от тренда. Тренд определим методом скользящей 12-членной центрированной средней.

Просуммировав месячные цены за январь — декабрь 1902 г. и поделив результат на число суммируемых уровней, т. е. на 12, получим годовую среднюю цену, относящуюся к концу июня (середине рассматриваемого периода):

$$\bar{y}_1 = \frac{1}{12}(83,6 + 83,7 + 83,6 + 82,2 + 82,9 + 81,4 + 79,8 + 71,9 +$$
$$+ 71,8 + 75,0 + 76,5 + 79,3) = 79,3.$$

Рассчитав подобным образом среднюю хронологическую за фе[в]раль 1902 — январь 1903 г., получим среднегодовую цену, приуроче[н]ную к концу июля, $y_2 = 79,2$.

Среднюю из рассчитанных на конец июня и на конец июля вели[-]чин $y_1' = (79,3+79,2):2 = 79,2$ можно приписать середине июля и пр[и]

Таблица 6. Цены на озимую пшеницу (коп. за пуд) в Ростове-на-Дону[1]

Год	Месяц											
	I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII
1902	83,6	83,7	83,6	82,2	82,9	81,4	79,8	71,9	71,8	75,0	76,5	79,3
1903	82,3	83,8	85,2	87,2	87,6	84,2	82,9	83,8	82,9	85,9	87,4	86,3
1904	87,2	91,7	92,3	92,8	86,7	89,1	91,5	92,3	90,5	93,2	95,5	96,0
1905	97,1	97,5	97,5	97,5	95,7	97,4	95,8	92,1	95,7	99,2	96,0	92,5
1906	96,6	97,7	96,1	92,6	90,4	92,3	93,5	91,8	94,5	96,6	95,8	98,0
1907	104,7	110,6	108,3	105,9	110,7	109,1	103,6	109,4	125,6	138,0	141,6	141,1
1908	143,1	137,0	143,3	130,8	132,3	135,0	137,2	136,7	141,0	138,1	136,0	126,2
1909	128,4	133,0	134,1	135,8	136,9	142,2	125,3	110,7	114,2	117,0	117,6	118,4

[1] Своды товарных цен на главных русских и иностранных рынках. Спб., 1910.

Таблица 7. Расчет 12-членных средних

Год	Месяц											
	I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII
1902							79,2	79,2	79,2	79,6	80,0	80,3
1903	80,5	81,1	82,1	83,0	84,0	84,7	85,2	85,8	86,4	86,8	87,0	87,2
1904	87,8	88,6	89,2	89,8	90,4	91,2	92,0	92,6	93,1	93,5	94,0	94,8
1905	95,3	95,5	95,7	96,2	96,4	96,3	96,2	96,1	96,0	95,8	95,4	95,0
1906	94,7	94,7	94,8	94,6	94,4	94,6	95,2	96,1	97,2	98,2	99,6	101,2
1907	102,3	103,3	105,2	108,2	111,9	115,9	119,0	121,7	124,2	126,8	128,7	130,6
1908	133,1	135,6	137,4	138,1	137,8	137,0	135,8	135,0	134,4	134,2	134,6	135,1
1909	134,9	133,4	131,2	129,2	127,6	126,4						

Таблица 8. Отношения фактических цен (данные табл. 6) к 12-членным средним (данные табл. 7)

Год	Индексы сезонности по месяцам											
	I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII
1902							100,7	90,8	90,6	94,2	95,6	98,8
1903	102,2	103,3	103,8	105,1	104,3	99,4	97,3	97,7	95,9	99,0	100,4	99,5
1904	99,3	103,5	103,5	103,3	95,9	97,7	99,4	99,7	97,2	99,7	101,6	101,3
1905	101,9	102,1	101,9	101,4	99,3	101,1	99,6	95,8	99,7	103,5	100,6	97,4
1906	102,0	103,2	101,4	97,9	95,8	97,6	98,2	98,6	97,2	98,4	96,2	96,8
1907	102,3	107,1	102,9	97,9	98,9	94,4	87,0	89,9	101,1	108,9	110,0	108,0
1908	107,5	101,0	104,3	94,7	96,0	98,5	101,0	101,2	104,9	102,9	101,0	93,4
1909	95,2	99,7	102,2	105,1	107,3	112,5						

Таблица 9. Средние индексы сезонности, рассчитанные разными способами на основе табл. 8[1]

№ строки	Месяц											
	I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII
1	101,5	102,8	102,9	100,8	96,6	100,2	97,6	96,2	98,1	100,9	100,8	99,3
2	102,0	103,2	102,9	101,4	98,9	98,5	99,4	97,7	97,2	99,7	100,6	98,8
3	102,1	102,9	102,9	100,9	98,1	98,5	99,1	97,4	98,0	100,5	100,7	98,6

[1] В строке 1 — индексы сезонности, вычисленные как средние арифметические, в строке 2 — индексы сезонности, рассчитанные как медианы, в строке 3 — индексы сезонности, подсчитанные как средние арифметические по расположению.

нять её за значение тенденций изменения цен в июле. Аналогичным образом вычисляются значения тенденции на последующие месяцы. Результаты расчетов приведены в табл. 7.

Фактические месячные цены отличаются от фиксированной в этом периоде тенденции, причем эти отличия в значительной степени объясняются сезонностью цен. Рассчитывая отношение фактических цен (табл. 6) к соответствующим значениям тенденции (табл. 7), выделим ту часть изменения цен, которая определяется в основном сезонными колебаниями. Полученные индексы цен в процентах представлены в табл. 8.

Размах сезонных колебаний может варьироваться из года в год, на изменения цен влияют всевозможные случайные причины, от которых несвободны и выделенные отношения. Поэтому индексы, относящиеся к одному и тому же месяцу, но к разным годам, колеблются и тем самым только приближенно представляют сезонные колебания.

Чтобы конкретизировать эти наблюдения и уловить типичные черты сезонности, рассчитывают средние за ряд лет индексы для каждого месяца. Эти индексы (**индексы сезонных колебаний**) показывают высоту сезонной волны для каждого месяца года. Для расчета индексов сезонных колебаний используют среднюю арифметическую, медиану и среднюю арифметическую по расположению.

На практике чаще всего применяют среднюю арифметическую. Но средняя арифметическая может оказаться непредставительной, искаженной, если в изучаемом периоде произошли какие-нибудь необычные события, повлекшие за собой нетипичные изменения показателей. Скажем, сильный неурожай, ввод в действие железной дороги и т. п. могут совершенно исказить реально существующие сезонные колебания хлебных цен, если для расчетов индексов сезонных колебаний используется средняя арифметическая. В таких случаях больше годится медиана. В подобных же ситуациях используют для вычисления средних индексов среднюю арифметическую по положению. В ранжированном ряду показателей каждого месяца отбрасывают крайние (самые низкие и самые высокие) значения, по оставшимся рассчитывают среднюю арифметическую. Средний индекс, рассчитанный таким образом, более устойчив, чем индекс, вычисленный как медиана, и не настолько подвержен влиянию крайних, часто нетипичных значений, насколько индекс, рассчитанный как средняя арифметическая.

Вернемся к примеру 12. Сначала подсчитаем средний за имеющиеся годы индекс цен для каждого месяца с помощью средней арифметической. Так, августовский индекс равен

$$i_{VIII}=\frac{1}{7}(90{,}8+97{,}7+99{,}7+95{,}8+98{,}6+89{,}9+101{,}2)=$$

$$=\frac{673{,}7}{7}=96{,}2.$$

Результаты расчетов приведены в первой строке табл. 9. Индексы показывают степень и направление изменения цен в каждом месяце за счет сезонности. Например, рассчитанный выше августовский индекс показывает, что сезонное понижение цен в августе в среднем за семь лет составило 3,8%.

Воспользуемся медианой для расчета средних индексов цен. Для этого ранжируем цены каждого месяца и находим средний член (медиану). Так, январские индексы выпишем в возрастающем порядке: 95,2; 99,3; 101,9; 102,0; 102,2; 102,3; 107,5; найдем медиану, равную 102,0. Аналогичным образом находятся средние индексы оставшихся месяцев. Результаты сведены во вторую строку табл. 9.

Опробуем и третий способ расчета средних индексов. В ранжированном ряду январских индексов отбросим два максимальных и два минимальных члена, для трех оставшихся уровней рассчитаем среднюю арифметическую $i_I=$ $=(101{,}9+102{,}0+102{,}3):3=102{,}1$. Воспользуемся этим приемом для расчета остальных средних индексов и запишем результаты в третью строку табл. 9. Расхождения при использовании для подсчета средних индексов средней арифметической, медианы и средней арифметической по расположению в данном случае незначительны.

Для того чтобы избавиться от сезонной волны, фактические месячные цены делят на соответствующие им индексы и умножают на 100. В нашем примере для получения скорректированных на сезонность цен нужно январскую цену поделить на 101,5 и умножить на 100, февральскую поделить на 102,8 и умножить на 100 и т. д. (см. первую строку табл. 9). Так, для 1902 г. цены, лишенные сезонной составляющей, будут такими:

I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII
82,4	81,4	81,2	81,5	83,3	81,2	81,8	74,7	73,2	74,3	75,9	79,8

Оценка влияния сезонного фактора на временной ряд. В качестве показателя колеблемости временного ряда за счет сезонности используют среднее квадратическое отклонение, определяемое по формуле

$$\sigma=\sqrt{\frac{\sum\limits_{k=1}^{12}(i_k-100)^2}{12}}, \qquad (7.13)$$

где i_k — k-й индекс сезонности.

В нашем примере действие сезонного фактора незначительно:

$$\sigma=\sqrt{\frac{(101{,}5-100)^2+(102{,}8-100)^2+\ldots+(99{,}3-100)^2}{12}}=$$

$$=\sqrt{3{,}7575}=1{,}94.$$

Действительно, показатель разброса месячных цен в 1902—1909 гг. (среднее квадратическое отклонение, подсчитанное для всей совокупности, представленной в табл. 6) равен 20,4. Таким образом, сезонный фактор объясняет только $(1{,}94/20{,}4)\,100\%=9{,}3\%$ всего разброса месячных цен за этот период.

Изложенные выше способы выделения сезонных колебаний могут быть использованы и для выявления циклических колебаний. Применение этих способов требует предварительного решения вопроса о длине цикла.

Кроме того, существуют процедурно сложные и более эффективные во многих конкретных задачах методы исследования периодических (сезонных и циклических) процессов. Наиболее разработанным является так называемый спектральный анализ [1].

§ 3. Изучение связей между временными рядами

В сравнительном анализе временных рядов важное место занимает исследование связей между ними. Для определения тесноты связей используется изложенный в гл. 6 этого раздела корреляционный анализ. Особенности временных рядов как исходной базы для применения методов корреляционного анализа диктуют определенные усложнения этих методов.

Автокорреляция. Временные ряды имеют такую особенность, которая значительно затрудняет применение к ним корреляционного и регрессионного анализа. Дело в том,

---

[1] См.: *Дженкинс Г., Ваттс Д.* Спектральный анализ и его приложения. М., 1971, вып. 1; М., 1972, вып. 2; *Гренджер К., Хатанака М.* Спектральный анализ временных рядов в экономике. М., 1972.

что корректное использование этих методов изучения связей предполагает независимость отдельных членов ряда. Известно, что подавляющая часть временных рядов, составленных из экономических показателей, не удовлетворяет этому требованию. Компоненты таких временных рядов связаны между собой. Скажем, валовой доход промышленности или сельского хозяйства, разумеется, зависит от уровня, достигнутого в предыдущие годы.

Связь уровней внутри ряда можно обнаружить и оценить с помощью коэффициента корреляции, рассчитанного для двух рядов, один из которых — исходный (y_1, y_2, ..., ..., y_{n-1}), а другой — сдвинутый на единицу (y_2, y_3, ...,y_n). Этот коэффициент, измеряющий связь ряда с его же собственными сдвинутыми на единицу времени уровнями, называется **первым коэффициентом автокорреляции.**

Отсутствие автокорреляции при условии нормального распределения означает независимость отдельных членов временного ряда. Поскольку временные ряды, являющиеся объектом исторических исследований, как правило, автокоррелированы, чтобы обеспечить независимость членов ряда, предпринимают попытки избавиться от автокорреляции. Существуют способы, позволяющие избавиться от автокорреляции или, по крайней мере, существенно снизить ее. Довольно прост и эффективен способ, сводящийся к выявлению тренда и исключению его из временного ряда. Широко используются для уменьшения автокорреляции первые разности.

Если попытки избавиться от автокорреляции не привели к успеху, используются специальные методы, позволяющие обойти трудности, возникшие в связи с тем, что анализируемые ряды автокоррелированы. В основе таких методов лежит следующее соображение: внутренне взаимосвязанные ряды несут в себе меньше информации, чем ряды с независимыми элементами. Существуют способы, позволяющие оценивать длину n' такого ряда с независимыми членами, информативность которого соответствует данному автокоррелированному ряду длиной n.

Один из таких способов принадлежит Бартлету, согласно которому длина ряда n' рассчитывается по формуле

$$n' = n\,(1 - r_1' r_2')/(1 + r_1' r_2'), \qquad (7.14)$$

где n — число уровней исходного ряда; r_1' и r_2' — коэффициенты автокорреляции для 1-го и 2-го рядов соответственно.

Для проверки гипотезы об отсутствии связи между рядами (см. формулу (6.6) в § 2 гл. 6) число степеней свободы полагают равным $n'-2$ и, учитывая это, величину t вычисляют по формуле:

$$t = r_{12}\sqrt{n'-2} \Big/ \sqrt{1-r_{12}^2}. \qquad (7.15)$$

Величина t имеет распределение Стьюдента с $(n'-2)$ степенями свободы.

Пример 13. Известно, что в конце XIX — начале XX столетия основными поставщиками пшеницы в мировой торговле были США и Россия. Возникает вопрос, существовала ли связь между ценами на пшеницу в этих странах. Конкретно, поставим вопрос о наличии связи между рядами цен: а) Нью-Йорка и Самары и б) Нью-Йорка и Николаева.

Для уменьшения автокорреляции проведено аналитическое выравнивание по параболе всех трех рядов. Получены отклонения фактических цен от выравненных. Далее, рассчитаны коэффициенты автокорреляции для всех рядов отклонений и коэффициенты корреляции между рядами отклонений цен Нью-Йорка и Самары, Нью-Йорка и Николаева.

Результаты получены следующие. Коэффициенты корреляции между рядами отклонений фактических значений цен от выравненных равны для Нью-Йорка и Самары 0,42, для Нью-Йорка и Николаева — 0,86. Коэффициенты автокорреляции для отклонений цен Нью-Йорка, Николаева и Самары соответственно равны 0,48, 0,60 и 0,59.

Коэффициенты автокорреляции рассмотренных рядов довольно высоки и, следовательно, их необходимо учитывать при оценке связей между рядами. Воспользуемся формулой (7.14), чтобы внести поправки к длинам рассматриваемых рядов:

$$n_1' = \frac{25\,(1-0{,}48\cdot 0{,}59)}{1+0{,}48\cdot 0{,}59} = 13{,}96, \qquad n_2' = \frac{25\,(1-0{,}48\cdot 0{,}60)}{1+0{,}48\cdot 0{,}60} = 13{,}82.$$

Округляя полученные результаты, имеем $n_1'=14$ и $n_2'=14$, т. е. 25-членные автокоррелированные ряды равноценны рядам в 14 независимых уровней.

Для расчета величин t используем формулу (7.15):

$$t_1 = \frac{0{,}42\sqrt{14-2}}{\sqrt{1-(0{,}42)^2}} = 1{,}60, \quad t_2 = \frac{0{,}86\sqrt{14-2}}{\sqrt{1-(0{,}86)^2}} = 5{,}84.$$

При $n_1'-2=n_2'-2=12$ степенях свободы с уровнем значимости (риском ошибиться) $\alpha=0{,}05$ значения t могут достигать $t_{кр}=2{,}18$, если линейная связь между исследуемыми рядами отсутствует (см. табл. 2 приложения). Поскольку вычисленное значение t_1 не превосходит критического, то можно сделать вывод об отсутствии линейной связи между отклонениями фактических значений цен от выравненных для Нью-Йорка и Самары. Что касается связи между отклонениями цен для Нью-Йорка и Николаева, то в этом случае гипотеза об отсутствии связи не подтверждается, так как рассчитанное значение t_2 больше критического.

В заключение главы подчеркнем, что существующие в математической статистике методы позволяют не только анализировать процессы и явления во времени, выявлять присущие им закономерности, но и дают возможность моделировать прошлое.

ГЛАВА 8
МЕТОДЫ АНАЛИЗА КАЧЕСТВЕННЫХ ПРИЗНАКОВ

Анализ содержания исторических источников нередко приводит к необходимости обработки качественной информации. Такая ситуация является типичной при изучении различных анкет, материалов переписей, личных карточек рабочих и т. д., а также при анализе нарративных источников. Формализация данных, содержащихся в источниках такого типа, характеризуется тем, что свойства (признаки) изучаемых объектов или явлений могут быть описаны лишь на качественном уровне, т. е. измерение этих признаков с помощью количественных шкал не представляется возможным.

Общая классификация типов признаков вводилась в гл. 2. Рассмотрим подробнее типы **качественных** признаков, которые разделяются на качественные признаки рангового (порядкового) и номинального (классификационного) типа (см. рис. 18).

Измерение по **номинальной** (простейшей из шкал) эквивалентно отнесению объекта к одному из классов по данному признаку. Частным случаем номинальной является шкала измерения **альтернативных** (дихотомических) признаков, т. е. шкала, включающая два класса (градации), соответствующих, например, наличию и отсутствию у объекта того или иного качества.

Порядковые шкалы — это более высокий уровень измерения качественных признаков. Градации рангового признака, измеренного в такой шкале, упорядочены по степени проявления соответствующего свойства. Ранговый признак задает не простую группировку объектов, как номинальный, а такую, при которой группы являются упорядоченными.

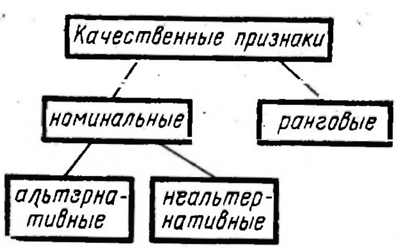

Рис. 18. Классификация качественных признаков

Рассмотрим несколько примеров качественных признаков различных типов. Номинальные признаки: наименование продукции, тип хозяйства, профессия, национальность, пол (последний признак является альтернативным). Ранговые признаки: сословие, образование, квалификация.

Основное внимание в данной главе уделяется важной проблеме измерения связи качественных признаков. Наиболее разработанными являются методы анализа взаимосвязи качественных признаков с одинаковыми уровнями измерения.

Таблица 1. Матрица сопряженности альтернативных признаков

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

§ 1. Анализ связи номинальных признаков

Традиционные меры связи. Традиционный подход к построению мер связи нормальных признаков основан на проверке предположения о статистической независимости рассматриваемых признаков.

Рассмотрим вначале вопрос о взаимосвязи двух **альтернативных** признаков, таблица сопряженности которых имеет следующий вид[1] (см. табл. 1).

Здесь буквами A и A обозначены наименования двух классов по первому признаку, а B и $\bar{B}$ — по второму. Числа в клетках табл. 1 определяют численности объектов, обладающих сочетанием соответствующих свойств. Так, c равно числу объектов, обладающих свойством A и не обладающих свойством B.

Пример 1. По выборочным данным, относящимся к концу XIX в., изучим зависимость между размером имения и типом хозяйства.

Таблица 2. Сопряженность между размером имения и типом хозяйства

Размер имений	Тип хозяйства		
	отработочное (A)	капиталистическое ($\bar{A}$)	
Крупные (B)	40 a	10 b	50
Мелкие ($\bar{B}$)	25 c	45 d	70
	65	55	120

[1] Эту таблицу называют обычно 4-клеточной или таблицей 2×2.

Как следует из табл. 2, число, например, крупных хозяйств отработочного типа равно 40; общее число крупных хозяйств $(a+b)$ равно 50, а общее число хозяйств отработочного типа $(a+c)$—65.

Условие статистической независимости признаков в данном случае эквивалентно требованию, чтобы, например, доля крупных хозяйств среди имений отработочного типа была равна доле крупных хозяйств во всей изучаемой совокупности, т. е. чтобы

$$\frac{a}{a+c}=\frac{a+b}{a+b+c+d} \quad \text{или } ad=bc. \text{ Если же } ad-bc\neq 0,$$

то рассматриваемые признаки являются взаимосвязанными. Данные табл. 2 показывают, что $ad=40\cdot 45$, $bc=10\cdot 25$, т. е. $ad-bc>0$, что указывает на наличие связи типа хозяйства с размером имения.

Для оценки степени связи альтернативных признаков используют обычно следующие два коэффициента:

$$Q=\frac{ad-bc}{ad+bc} \text{ — коэффициент ассоциации (связи);} \qquad (8.1)$$

$$\Phi=\frac{ad-bc}{\sqrt{(a+b)(a+c)(b+d)(c+d)}} \text{ — коэффициент контингенции (сопряженности).} \qquad (8.2)$$

Оба коэффициента Q и Φ принимают значения от —1 до +1 и равны 0, если признаки статистически независимы.

Коэффициент Q равен +1 в случае полной связанности, т. е. либо все объекты со свойством A обладают одновременно и свойством B $(c=0)$, либо все объекты со свойством B обладают одновременно и свойством A $(b=0)$. Значение —1 коэффициент Q принимает в случае полной отрицательной связанности $(a=0$ или $d=0)$.

Коэффициент Φ равен +1, если все объекты со свойством A обладают и свойством B, а все объекты со свойством B обладают и свойством A $(b=0,\ c=0)$. Значение —1 коэффициент Φ принимает в случае, когда все объекты со свойством A не обладают свойством B, а все объекты со свойством B не обладают свойством A $(a=0$ и $d=0)$.

Таким образом, числовые значения коэффициентов Q и Φ для одних и тех же данных могут существенно отличаться друг от друга, поскольку эти коэффициенты измеряют различные аспекты взаимосвязи в 4-клеточной таблице. Φ предназначен для измерения степени двусторонней взаимосвязи между рассматриваемыми альтернативными признаками, в то время как Q отражает лишь односторонне направленную связь. Значения этих коэффициентов совпадают лишь при наличии полной двусторонней взаимосвязи: $Q \geqslant \Phi$. Так, определение величины коэффициентов Q и Φ для данных табл. 2 приводит к следующим результатам:

$$Q = \frac{40 \cdot 45 - 10 \cdot 25}{40 \cdot 45 - 10 \cdot 25} = 0{,}76; \qquad \Phi = \frac{40 \cdot 45 - 10 \cdot 25}{\sqrt{50 \cdot 65 \cdot 55 \cdot 70}} = 0{,}44.$$

Оба коэффициента Q и Φ в данном примере принимают положительные значения, т. е. связь между рассматриваемыми признаками прямая (наличие свойства A связано, как правило, с наличием свойства B).

Довольно высокое значение коэффициента связи Q (0,76) определяется тем, что абсолютное большинство крупных хозяйств (40 из 50, т. е. 80%) относится к отработочному типу. Величина коэффициента сопряженности Φ (0,44) значительно уступает величине Q. Это связано с тем, что связь данных признаков носит преимущественно односторонний характер (среди крупных хозяйств отработочные составляют 80%, в то время как среди хозяйств отработочного типа около 40% (25 из 65) составляют мелкие хозяйства).

Перейдем теперь к вопросу о построении стандартных мер связи **номинальных признаков общего вида.**

Пусть имеются два номинальных признака, число градаций (классов) которых равно l и m. Данные об их взаимосвязи могут быть представлены в виде следующей таблицы сопряженности (табл. 3).

Здесь n_{ij} обозначает число объектов, относящихся к i-му классу по первому признаку и к j-му — по второму признаку. Так, например, если $n_{3,5}=7$, то число объектов, относящихся к 3-му классу по первому признаку и к 5-му классу по второму, равно 7.

Таблица 3

n_{11}	$n_{12} \ldots$	n_{1m}	$n_{1\cdot}$
n_{21}	$n_{22} \ldots$	n_{2m}	$n_{2\cdot}$
.	.	.	.
.	.	.	.
.	.	.	.
n_{e_1}	$n_{e_2} \ldots$		$n_{e\cdot}$
$n_{\cdot 1}$	$n_{\cdot 2}$	$n_{\cdot m}$	N

Через $n_{i\cdot}$ в табл. 3 обозначено общее число объектов i-го класса по первому признаку (т. е. сумма чисел в i-й строке), а через $n_{\cdot j}$ — число объектов j-го класса по второму признаку (сумма чисел в j-м столбце). N — общее число объектов в изучаемой совокупности; $N = n_{1\cdot} + n_{2\cdot} + \ldots + n_{m\cdot} = n_{\cdot 1} + n_{\cdot 2} + \ldots + n_{\cdot l}$.

Для оценки существенности связи двух номинальных признаков на основе принципа статистической независимости вначале ответим на вопрос — какими были бы частоты в клетках табл. 3, если бы рассматриваемые признаки были независимыми?

Обозначим n_{ij} — число в клетке (i, j) табл. 3, которое получилось бы в соответствии с гипотезой о статистической независимости при заданных численностях классов $n_{.1}$, $n_{.2}$, ..., $n_{.e}$ и $n_{1.}$, ..., n_{m}. Соответствующая частость была бы равна $f_{ij}=\tilde{n}_{ij}/N$. Исходя из принципа статистической независимости, ожидаемая теоретическая частость (доля) $\tilde{f}_{ij}$ рассчитывается по формуле

$$\tilde{f}_{ij}=\tilde{n}_{ij}/N=\frac{n_{i.}}{N}\cdot\frac{n_{.j}}{N},$$

отсюда

$$\tilde{n}_{ij}=n_{i.}n_{.j}/N. \tag{8.3}$$

Для оценки существенности различий между реальными данными в табл. 3 и «сконструированными» в соответствии с гипотезой о независимости признаков применяется **коэффициент квадратичной сопряженности** X^2, основанный на так называемом критерии согласия X^2 (хи-квадрат). Вычисление этого коэффициента производится путем суммирования относительных различий между числами n_{ij} и $\tilde{n}_{ij}$ по всем клеткам таблицы сопряженности:

$$X^2=\sum_{i=1}^{l}\sum_{j=1}^{m}\frac{(n_{ij}-\tilde{n}_{ij})^2}{\tilde{n}_{ij}}. \tag{8.4}$$

Как следует из формулы (8.4), значение коэффициента X^2 тем меньше, чем меньше различия между числами n_{ij} и $\tilde{n}_{ij}$; $X^2=0$, когда $n_{ij}=\tilde{n}_{ij}$ для всех клеток табл. 3.

Формула (8.4) легко преобразуется к следующему виду:

$$X^2=N\left(\sum_{i=1}^{l}\sum_{j=1}^{m}\frac{n_{ij}^2}{n_i n_j}-1\right), \tag{8.5}$$

в котором она обычно и используется на практике.

Исходя из этой формулы признаки должны быть признаны зависимыми, если $X^2=0$. Однако это утверждение будет верным лишь в том случае, когда данные, содержащиеся в таблице сопряженности, относятся к генеральной сово-

Таблица 4. Сопряженность между типом хозяйства и сословием владельца

Сословие владельца	Тип хозяйства		Σ
	отработочные	капиталистические	
Крестьяне	5	25	30
Купцы, мещане	10	40	50
Дворяне	50	10	60
Σ	65	75	140

купности. Если же эти данные получены в результате случайной выборки, то это утверждение может оказаться ошибочным (вследствие возможных ошибок репрезентативности). Поэтому основной целью применения критерия X^2 в выборочных исследованиях является установление критического значения $X^2_{кр}$, такого, что вероятность получения значений X^2, больших критического, за счет случайностей выборки весьма мала. Таким образом, если X^2 окажется больше $X^2_{кр}$, гипотеза об отсутствии взаимосвязи между признаками должна быть отклонена для выбранного уровня значимости.

Уровень значимости (α) означает вероятность риска ошибиться, отвергая гипотезу о статистической независимости (см. гл. 9, § 4).

Удобство использования критерия X^2 на практике определяется наличием таблиц, содержащих критические значения этого критерия для различных уровней значимости и размерностей задачи (такая таблица содержится и в данном учебном пособии — табл. 3 приложения). В этой таблице слева указано число степеней свободы k (размерность задачи), а сверху — уровень значимости α. В нашей задаче анализа таблицы сопряженности номинальных признаков число степеней свободы определяется по формуле

$$k=(l-1)(m-1), \tag{8.6}$$

где l и m, как и ранее, обозначают число градаций рассматриваемых признаков. Уровень значимости α обычно выбирают равным 0,01; 0,05 или 0,10.

Критическое значение $X^2_{кр}$ определяется по табл. 3 приложения на пересечении строки, соответствующей данной величине k, и столбца, соответствующего выбранному уровню значимости α.

Пример 2. Поясним методику применения критерия X^2 на следующем иллюстративном примере.

Определим значение коэффициента X^2 для данных табл. 4 по формуле (8.5):

$$X^2 = 140\left(\frac{5^2}{30\cdot 65} + \frac{25^2}{30\cdot 75} + \frac{10^2}{50\cdot 65} + \frac{40^2}{50\cdot 75} + \frac{50^2}{60\cdot 65} + \frac{10^2}{60\cdot 75} - 1\right) = 57{,}56.$$

Число степеней свободы в данном случае равно $k=(3-1)(2-1)=2$. Выберем величину уровня значимости равной $\alpha=0{,}01$. Как следует из табл. 3 приложения, критическое значение X^2 в этом случае равно 9,21. Это означает, что значение, равное или большее этой величины, может встретиться только один раз из ста (так как $\alpha=1/100$) при условии, что гипотеза о статистической независимости верна. Поскольку в нашем примере значение $X^2=57{,}56>9{,}21$, то связь между типом хозяйства и сословием владельца следует считать существенной. Нет и одного шанса из ста, что этот вывод получен вследствие случайных факторов.

Если бы значение X^2 в данном примере получилось меньшим чем 9,21, то это означало бы, что анализируемые данные (табл. 4) согласуются с гипотезой о статистической независимости признаков и не дают оснований отвергнуть эту гипотезу (при данном уровне значимости α).

Итак, с помощью критерия X^2 можно оценить величину риска в принятии предположения о существовании связи. Однако установив факт наличия связи между признаками, исследователь должен измерить ее силу, чтобы иметь возможность сравнивать степень взаимосвязи между различными признаками, сопоставлять результаты, полученные в различных исследованиях.

Существует целый ряд стандартных коэффициентов связи номинальных признаков, основанных на использовании коэффициента X^2. Наиболее распространенными из них являются коэффициенты связи, предложенные в начале XX в. Чупровым и Крамером:

$$T^2 = \frac{X^2}{N\sqrt{(l-1)(m-1)}} \text{ — коэффициент Чупрова;} \qquad (8.7)$$

$$K^2 = \frac{X^2}{N\min(l-1,\, m-1)} \text{ — коэффициент Крамера,} \qquad (8.8)$$

где через $\min(l-1,\, m-1)$ обозначено минимальное из чисел $(l-1,\, m-1)$.

Коэффициенты T^2 и K^2 принимают значение в интервале [0, 1]. Минимальное — нулевое — значение они имеют тогда и только тогда, когда признаки статистически независимы.

Значение 1 коэффициент T^2 принимает при полной связи между признаками только в том случае, если число градаций обоих признаков одинаково ($l=m$); во всех других случаях даже при полной связи коэффициент Чупрова меньше единицы. Наибольшее значение, равное единице, коэффициент Крамера принимает при полной связи признаков независимо от того, равны ли между собой число строк (l) и число столбцов (m) таблицы сопряженности.

Заметим, что коэффициенты T и K эквивалентны, когда $l=m$; в остальных случаях $T<K$. (Так, вычисление по формулам (8.7), (8.8) коэффициентов T и K по данным табл. 4 приводит к значениям $T=0{,}54$, $K=0{,}64$). При анализе таблиц сопряженности 2×2 оба коэффициента равны коэффициенту сопряженности Φ.

Числовое значение рассмотренных коэффициентов связи номинальных признаков остается неизменным при перестановке местами строк или столбцов таблицы сопряженности. Эти коэффициенты всегда выражаются неотрицательными числами (заключение о знаке связи здесь лишено смысла); поэтому выяснение характера зависимости, ее специфических черт должно определяться по таблице сопряженности.

Действительно, при измерении связи количественных признаков (см. гл. 6, § 2) знак коэффициента корреляции характеризует направление связи: если с увеличением значений признака X в среднем увеличиваются и значения признака Y, то знак r_{xy} положителен; если же тенденции к изменению значений этих признаков противоположны, то $r_{xy}<0$. Далее, при измерении связи альтернативных признаков используемые коэффициенты Q и Φ также характеризуются не только числовым значением, но и знаком, который показывает, имеется ли тенденция к наличию у объектов одновременно свойств A и B (в этом случае связь положительна) или одно из этих свойств связано с отсутствием другого (связь отрицательна). Что же касается измерения связи номинальных признаков общего вида, значения которых выражаются лишь наименованиями градаций (классов) и поэтому не могут быть упорядочены, то здесь бессмысленно говорить о знаке связи, можно лишь измерить ее силу, а также указать наиболее характерные сочетания значений двух изучаемых признаков.

Для корректного использования мер связи, основанных на X^2, необходимо обеспечить выполнение следующей практической рекомендации: ни одна из ожидаемых теоретических частостей $\tilde{n}_{ij}$ не должна быть слишком мала ($\tilde{n}_{ij}\geqslant5$).

При нарушении этого условия следует либо уменьшить степень дробности группировки признаков (число градаций), либо обратиться к другому критерию.

Теоретико-информационные меры связи. Развитие математико-статистических методов в последние десятилетия привело к появлению нового подхода к проблеме оценки связи качественных признаков. В основе этого подхода лежит представление о том, что мера связи признаков должна не столько оценивать степень их статистической независимости, сколько характеризовать возможность прогноза значений одного из признаков по значениям другого.

Самый точный прогноз достигается в ситуации, когда для каждого из значений одного признака можно однозначно указать соответствующее значение второго. Мера связи в этой ситуации, соответствующей «полной связи» между признаками, должна принимать максимальное значение, равное 1. Критерием отсутствия связи в рамках данного подхода обычно также считается статистическая независимость признаков; в этом случае мера связи должна принимать минимальное, нулевое, значение.

Таким образом, современная трактовка понятия «полная связь» между признаками X и Y означает, что знание значения признака X устраняет всякую **неопределенность** в знании значения признака Y. Для уменьшения неопределенности необходимо получить некоторое **количество информации.** Измерение статистической связи признаков сводится здесь к оценке относительно уменьшения неопределенности Y при получении знания об X.

Пример 3. Поясним сказанное на примере данных табл. 4, содержащей иллюстративные сведения о сопряженности признаков «сословие» (X) и «тип хозяйства» (Y). Знание значения признака X в данном примере существенно снижает неопределенность в знании значений признака Y. Так, если X принимает значение x_3 («дворяне»), то Y почти наверняка (с вероятностью $0{,}93=50/60$) принимает значение y_1 («отработочный тип»). Если $X=x_2$ («крестьяне»), то неопределенность в знании Y также существенно снижается: с вероятностью $0{,}80=40/50$ $Y=y_2$ («капиталистический тип») и т. д. Отметим, что неопределенность в знании значений признака Y, взятого отдельно, высока — вероятности значений y_1 и y_2 равны $0{,}46=65/140$ и $0{,}54=75/140$ соответственно.

Количественный анализ неопределенности и информации осуществляется на основе результатов **теории информации,** начало которой было положено в конце 1940-х годов. Основным, фундаментальным понятием теории информации является энтропия — мера неопределенности. Применительно к нашей задаче анализа качественных признаков энтропия может быть определена следующим образом.

Пусть изучаемое явление (признак x) характеризуется m состояниями (классами) $x_1, x_2, \ldots, x_m$ и имеет статистическую природу. Обозначим через p_i вероятность того, что признак X принимает i-е значение x_i (при этом $p_1+p_2+\ldots+p_m=1$). Тогда энтропия, т. е. мера неопределенности признака x, вычисляется по формуле

$$H(X)=-(p_1\log p_1+p_2\log p_2+\ldots+p_m\log p_m)=$$
$$=-\sum_{i=1}^{m} p_i \log p_i, \tag{8.9}$$

где $\log p_i$ — логарифм при основании 2 от величины p_i.

При анализе эмпирических данных роль вероятностей p_i играют их выборочные оценки (частости, доли) n_i/N, где N — общее число объектов в выборке; n_i — число объектов i-го класса x_i.

Нетрудно показать, что энтропия H действительно соответствует интуитивным представлениям о степени неопределенности. Минимальное значение энтропии равно нулю, и достигается оно при условии, что все вероятности равны нулю, кроме одной (например, p_1), которая равна 1. Содержательно это означает, что все объекты принадлежат одному классу (x_1), т. е. какая-либо неопределенность в знании значений признака X отсутствует. Максимальное значение энтропии равно $\log m$, и достигается оно при таком распределении вероятностей, когда все они одинаковы и равны $p_i=1/m\,(1\leqslant i\leqslant m)$. Это свойство энтропии как меры неопределенности также согласуется с представлением о том, что максимальная неопределенность, разнообразие изучаемой системы соответствуют равновероятному распределению объектов по классам $x_1, \ldots, x_m$.

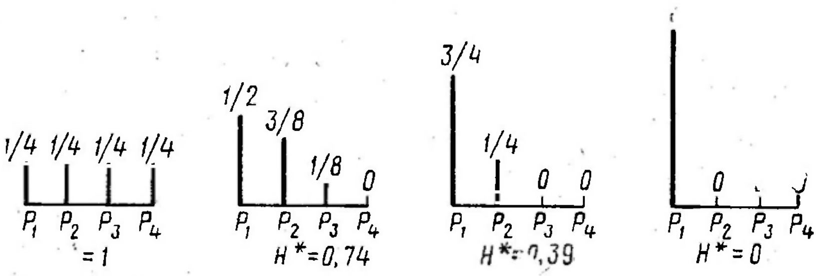

Рис. 19. Зависимость энтропии от распределения вероятностей

Поскольку максимальное значение энтропии ($\log m$) зависит от числа классов m, то на практике часто используют **нормированную энтропию** $H^* = H/\log m$, , которая принимает значения в интервале [0, 1]. Зависимость величины H^* от распределения вероятностей иллюстрируется графически на рис. 19 ($m=4$; число над каждым столбиком обозначает величину соответствующей вероятности p_i).

Вычисление энтропии H^* для признаков X и Y из табл. 4 приводит к следующим результатам. Для признака X («сословие») частоты трех классов равны соответственно 30/140, 50/140 и 60/140, а для признака Y («тип хозяйства») — 65/140, 75/140. Отсюда получаем:

$$H^*(X) = -[(30/140)\cdot\log(30/140) + (50/140)\cdot\log(50/140) + (60/140)\cdot\log(60/140)]/\log 3 = 0{,}89;$$

$$H^*(Y) = -[(65/140)\cdot\log(65/140) + (75/140)\cdot\log(75/140)]/\log 2 = 0{,}99.$$

Таким образом, мера неопределенности обоих признаков оказывается (по данным табл. 4) весьма высокой.

Отметим, что чувствительность величины энтропии к отклонениям от равномерного распределения частот послужила причиной активного использования энтропийного анализа в исторических исследованиях для оценки степени социально-экономического неравенства, имущественной дифференциации.

Перейдем теперь к определению понятия **«количество информации»**, особенно важного для построения мер связи качественных признаков. Пусть знание значений признака X уменьшает неопределенность в знании значений признака Y. Оставшуюся меру неопределенности обозначают $H_x(Y)$ и называют условной энтропией. Количество информации $I(XY)$ определяется как уменьшение неопределенности признака X за счет информации, полученной о связи X и Y:

$$I(XY) = H(Y) - H_x(Y). \tag{8.10}$$

В теории информации показано, что количество информации $I(XY)$ может быть вычислено по формуле

$$I(XY) = \sum_{i=1}^{l}\sum_{j=1}^{m} p_{ij}\log\frac{p_{ij}}{p_{i.}p_{.j}}, \tag{8.11}$$

где p_{ij} — вероятность того, что признак X характеризуется i-м классом, а признак Y — j-м классом; $p_{i.}$ — вероятность i-го значения признака X; $p_{.j}$ — вероятность j-го значения признака Y.

При анализе эмпирических таблиц сопряженности (вида табл. 3) роль вероятностей $p_{i.}$, $p_{.j}$ и p_{ij} играют соответствующие доли $n_{i.}/N$, $n_{.j}/N$ и n_{ij}/N.

Формула (8.11) для количества информации симметрична относительно X и Y, т. е. $I(XY) = I(YX)$, поэтому $I(XY)$ называют также **взаимной информацией** X и Y.

Взаимная информация обращается в нуль тогда и только тогда, когда признаки X и Y статистически независимы. Максимальное значение взаимной информации, равное $H(X)$ или $H(Y)$, соответствует функциональной зависимости (полной связи) признаков X и Y, когда каждому значению x_i признака X соответствует единственное значение y_j признака Y.

При анализе взаимосвязи признаков на основе взаимной информации $I(XY)$, когда таблица сопряженности строится на основе выборочных данных, возникает задача о значимости статистической зависимости, если $I(XY) \neq 0$. Исходя из доказанных в теории информации свойств $I(XY)$, в данном случае пользуются той же методикой, что и изложенная выше для определения уровня значимости коэффициента X^2 с помощью критерия X^2. По таблице X^2 для заданных уровней значимости (например, $\alpha = 0{,}01$), числа степеней свободы $k = (l-1)(m-1)$ и объема выборки (N) определяется критическое значение $I_{\text{кр}} = X^2/2N$, которое далее сравнивается с фактическим $I_{\text{ф}}(XY)$. Если $I_{\text{кр}} \geqslant I_{\text{ф}}(XY)$, то с заданной степенью уверенности принимается гипотеза о независимости X и Y, если же $I_{\text{ф}}(XY) > I_{\text{кр}}$, то утверждается, что на данном уровне значимости гипотеза о независимости X и Y неверна.

Пример 4. Обратимся вновь к данным табл. 4. Как уже было определено выше, при уровне значимости $\alpha = 0{,}01$ и числе степеней свободы $k=2$ критическое значение $X^2_{\text{кр}} = 9{,}21$. Следовательно, в данном случае $I_{\text{кр}} = X^2/2N = 9{,}21/(2 \cdot 140) = 0{,}033$. Вычисление по формуле (8.11) взаимной информации для признаков X и Y по данным табл. 4 приводит к величине $I_{\text{ф}}(XY) = 0{,}321$. Имеем $I_{\text{ф}}(XY) > I_{\text{кр}}$, следовательно, гипотеза о независимости X и Y на данном уровне значимости $\alpha = 0{,}01$ неверна (нет и одного шанса из ста, что признаки X и Y независимы).

Установив факт существенности связи признаков, далее следует измерить ее силу. Наибольшее распространение получила мера связи, для построения которой необходимо пронормировать взаимную информацию $I(XY)$ на $H(XY)$:

$$R(XY)=I(XY)/H(XY), \qquad (8.12)$$

где $H(XY)$ — энтропия совместного распределения признаков X и Y, вычисляемая по формуле

$$H(XY)=-\sum_{i=1}^{l}\sum_{j=1}^{m}p_{ij}\log p_{ij}. \qquad (8.13)$$

Мера зависимости $R(XY)$ (называемая также **симметричным информационным коэффициентом связи**) обладает следующими свойствами:

1) $0\leqslant R(XY)\leqslant 1$; 2) $R(XY)=0$ тогда и только тогда, когда X и Y независимы; 3) $R(XY)=1$ тогда и только тогда, когда X и Y функционально зависимы.

Пример 5. Вычисление коэффициента $R(XY)$ по данным табл. 4 приводит к значению $R(XY)=0{,}321/1{,}977=0{,}16$, что значительно ниже значений коэффициентов T и K, подсчитанных выше для тех же данных. $R(XY)$ дает «осторожную» оценку силы связи.

Отметим, что наряду с симметричной мерой зависимости $R(XY)$ получили распространение и асимметричные (направленные) меры зависимости, например:

$$R(X\to Y)=I(XY)/H(Y);\quad R(Y\to X)=I(XY)/H(X). \qquad (8.14)$$

Однако использование этих коэффициентов для разделения признаков на «зависимые» и «определяющие» не имеет достаточного обоснования.

Завершая рассмотрение теоретико-информационных мер связи, укажем, что их построение может осуществляться не только на основе энтропии. Важным примером такой меры служит **коэффициент Валлиса**, реализующий принцип «пропорциональной предикции», согласно которому мерой связи должно служить относительное уменьшение вероятности ошибки предсказания признака Y при знании признака X в сравнении с вероятностью ошибки прогноза Y без знания X. Коэффициент Валлиса вычисляется по формуле

$$W_{Y/X}=\frac{\displaystyle\sum_{i=1}^{l}\sum_{j=1}^{m}\frac{(p_{ij}-p_{i.}p_{.j})^2}{p_{.i}}}{1-\displaystyle\sum_{j=1}^{m}p_{.j}^2}. \qquad (8.15)$$

Свойства коэффициента Валлиса аналогичны свойствам коэффициента $R(XY)$, за тем лишь исключением, что $W_{y/x}=1$, когда возможен однозначный прогноз Y по X (но не наоборот). Интерпретация коэффициента Валлиса весьма проста: если, например, $W_{y/x}=0{,}50$, то знание X уменьшает число ошибок прогноза вдвое.

Для данных табл. 4 этот коэффициент равен $W_{y/x}=0{,}41$.

Отметим, что в выборочных исследованиях предпочтительнее пользоваться не коэффициентом Валлиса, а информационными мерами связи, значимость которых может быть установлена в соответствии с изложенной выше методикой. Информационные меры связи получили применение в исторических исследованиях, в частности при обработке массовых источников по социально-экономической истории советского общества.

§ 2. Анализ связи ранговых признаков

Ряд объектов, упорядоченных в соответствии со степенью проявления некоторого свойства, называют **ранжированным**; каждому числу такого ряда присваивается **ранг.** Будем обозначать ранги порядковыми числительными 1, 2, ..., n, где n — количество объектов. Таким образом, если какой-либо объект после ранжирования занимает третье место в ряду, ему присваивается ранг 3.

Меры взаимосвязи между парой признаков, каждый из которых ранжирует изучаемую совокупность объектов, называются в статистике коэффициентами **ранговой корреляции**. Эти коэффициенты строятся на основе следующих трех свойств:

а) если ранжированные ряды по обоим признакам полностью совпадают (т. е. каждый объект занимает одно и то же место в обоих рядах), то коэффициент ранговой корреляции должен быть равен +1, что означает полную положительную корреляцию;

б) если объекты в одном ряду расположены в обратном порядке по сравнению со вторым, коэффициент равен —1, что означает полную отрицательную корреляцию;

в) в остальных ситуациях значения коэффициента заключены в интервале [—1, 1]; возрастание модуля коэффициента от 0 до +1 характеризует увеличение соответствия между двумя ранжированными рядами.

Указанными свойствами обладают коэффициенты ранговой корреляции Спирмена ρ и Кендалла τ.

Таблица 5. Зависимость между размером имений и сословным рангом владельца (по выборочным данным конца XIX в.)

Сословие владельца	Средний размер имения, дес.	Ранги		d_i	d_i^2
		I	II		
Дворяне	400	1	1	0	0
Купцы	70	2	3	—1	1
Мещане	150	3	2	1	1
Крестьяне	40	4	4	0	0
				$\sum_{i=1}^{4} d_i^2 = 2$	

Коэффициент ранговой корреляции Спирмена вычисляется по формуле

$$\rho = 1 - \frac{6 \sum_{i=1}^{n} d_i^2}{n(n^2-1)}, \qquad (8.16)$$

где d_i — разность между парами рангов для i-го объекта; n — число сопоставляемых пар рангов (объектов).

Пример 6. Поясним технику вычисления коэффициента ρ на следующем иллюстративном примере.

В 3-м столбце табл. 5 приводятся значения рангов по признаку «сословие владельца», а в четвертом — значения рангов, полученных при ранжировке средних размеров имений. Подставляя промежуточные величины, вычисленные в табл. 5, в формулу (8.16), получим

$$\rho = 1 - \frac{6 \cdot 2}{4(4^2-1)} = 0{,}8,$$

т. е. связь между данными признаками прямая и довольно высокая.

Коэффициент ранговой корреляции Кендалла вычисляется по формуле

$$\tau = \frac{S^+ - S^-}{\frac{1}{2} n(n-1)} = \frac{S}{\frac{1}{2} n(n-1)}, \qquad (8.17)$$

где S определяется таким образом, как показано в примере 7.

Пример 7. Обратимся вновь к данным табл. 5. Таблица упорядочена так, что в столбце «Ранг I» ранги расположились в порядке воз-

растания их значений (это существенно для вычисления S). Первые четыре столбца табл. 5 при вычислении τ такие же, как и при вычислении ρ, а последние два столбца здесь заменяются следующими:

s_i^+	s_i^-
3	0
1	1
1	0
0	0
5	1

Эти два столбца заполняются на основе расположения рангов в столбце «Ранг II». Берем значение ранга, стоящего в столбце «Ранг II» на первом месте, 1; все три расположенных ниже данного ранга значения его превышают, поэтому в первую строку столбца S_i^+ заносим число 3. Для второго ранга (со значением 3) аналогичный подсчет дает число 1, и т. д. Для заполнения столбца 5 определяем для каждого ранга, сколько из расположенных ниже его рангов имеют значение меньше данного. Так, для первого ранга (1) ни один из нижерасположенных рангов не имеет значения меньше 1, поэтому в первую строку столбца S_i^- заносится 0. Аналогичный подсчет для второго ранга со значением 3 дает число 1 и т. д. Таким образом, суммы по этим столбцам равны $S^+=5$ и $S^-=1$, а число $S=S^+-S^-=5-1=4$. Подставив это значение в формулу (8.17) для коэффициента Кендалла, получим

$$\tau=\frac{4}{1/2\cdot 4\,(4-1)}=0{,}66.$$

Коэффициент Кендалла τ дает более осторожную оценку корреляции, чем коэффициент Спирмена ρ (числовое значение τ всегда меньше, чем ρ). Хотя вычисление коэффициента ρ менее трудоемко, чем вычисление коэффициента τ, последний легче пересчитать, если к ряду добавляется новый член.

Важное достоинство коэффициента τ состоит в том, что с его помощью можно определить коэффициент **частной** ранговой корреляции, позволяющий оценить степень «чистой» взаимосвязи двух ранговых признаков, устранив влияние третьего:

$$\tau_{12\cdot 3}=\frac{\tau_{12}-\tau_{13}\tau_{23}}{\sqrt{(1-\tau_{13}^2)(1-\tau_{23}^2)}}. \qquad (8.18)$$

При ранжировании объектов нередко возникает ситуация, когда два (или большее число) объектов получают одинаковые ранги (такие объекты называют **связанными**). Расположение студентов в соответствии с их экзаменационными оценками является известным примером такого рода связей. В этом случае значение ранга связанных объектов берется равным среднему значению тех рангов, которые имели бы эти объекты, если они были бы различны.

Таблица 6. Взаимосвязь между размером имения и сословием владельца

Номер имения	Размер имения, дес.	Сословие владельца	Ранги I	Ранги II	S_i^+	S_i^-
1	465	Дворянин	1	1,5	5	0
2	310	»	2	1,5	5	0
3	150	Мещанин	3	4,5	2	1
4	72	Купец	4	3	3	0
5	65	Мещанин	5	4,5	2	0
6	28	Крестьянин	6	6,5	0	0
7	20	»	7	6,5	0	0
					$\Sigma = 17$	$\Sigma = 1$

Например, если связанными оказались 3-й и 4-й объекты в ранжированном ряду, то каждому из них приписывается ранг $3\frac{1}{2}$, а если связываются все объекты от 2-го до 6-го, то каждый получает ранг $(2+3+4+5+6)/5=4$. Если число связанных рангов невелико, то при вычислении ранговой корреляции можно пользоваться введенными здесь формулами для коэффициентов ρ и τ; в противном случае эти формулы несколько усложняются.

Пример 8. Пусть по небольшой выборке, включающей 7 хозяйств, изучается взаимосвязь между размером имения и сословием владельца, приведенными в табл. 6.

В этой таблице имения упорядочены по размеру, а среди рангов сословия владельцев имеются связанные ранги. Вычислив значения S_i^+ и S_i^-, определим коэффициент ранговой корреляции:

$$\tau = \frac{17-1}{1/2\cdot 7\,(7-1)} = 0{,}76.$$

Значимость коэффициентов ранговой корреляции. При определении силы ранговой корреляции на основе выборочных данных необходимо рассмотреть следующий вопрос: с какой степенью надежности можно полагаться на заключение о том, что в генеральной совокупности существует корреляция, если получен некоторый **выборочный** коэффициент ранговой корреляции. Другими словами, следует проверить значимость наблюдавшихся корреляций рангов исходя из гипотезы о статистической независимости двух рассматриваемых ранжировок.

При сравнительно большом объеме n выборки проверка значимости коэффициентов ранговой корреляции может осуществляться с помощью таблицы нормального распределения (табл. 1 приложения). Для проверки значимости коэффициента Спирмена ρ (при $n>20$) вычисляют значение

$$t_S = \rho/\sqrt{n-1}, \qquad (8.19)$$

а для проверки значимости коэффициента Кендалла τ (при $n>10$) вычисляют значение

$$t_K = \frac{S}{\sqrt{n(n-1)(2n+5)/18}}, \qquad (8.20)$$

где $S = S^+ - S^-$; n — объем выборки.

Далее задаются уровнем значимости α, определяют по табл. 1 приложения критическое значение $t_{кр}$ и сравнивают с ним вычисленное значение (t_S) или (t_K).

Пример 9. Проверим значимость коэффициента τ для данных табл. 6 (вычисления проделаем лишь для иллюстрации, так как объем выборки здесь меньше 10, $n=7$). По формуле (8.20):

$$t_K = \frac{16}{\sqrt{7(7-1)(2\times 7+5)/18}} = 2{,}41.$$

По табл. 1 приложения находим, что $t_{кр}=1{,}96$, если $\alpha=0{,}05$. Поскольку $t_K > t_{кр}$, то τ — значим.

При малых выборках проверка значимости коэффициентов ранговой корреляции проводится с помощью специальных таблиц, построенных на основе более сложных критериев.

Множественный коэффициент ранговой корреляции W. Этот коэффициент (называемый также коэффициентом конкордации) предназначен для измерения связи произвольного числа ранговых признаков. Поясним способ вычисления W на следующем примере.

Пример 10. Пусть имеются $n=7$ хозяйств, характеризующихся набором из $m=3$ ранговых признаков (табл. 7). В пятом столбце табл. 7 приводятся суммы рангов, полученных каждым объектом.

Если ранжировки объектов по разным признакам совпадают (или близки), то суммарные ранги объектов будут сильно различаться. Если же все m ранжировок слабо согласованы, то суммарные ранги объектов будут почти одинаковыми и близкими к их средней сумме, равной $\bar{a}=m(n+1)/2$.

Для построения коэффициента конкордации W вычисляют сумму S' квадратов разностей Δ_i между фактическими суммарными рангами объектов и их средним значением $\bar{a}$. Полученную сумму S' нормируют на максимально возможное ее значение, равное $m^2(n^3-n)/12$. Таким

Таблица 7. Данные для вычисления множественного коэффициента ранговой корреляции

Номер хозяйств	Ранги I	Ранги II	Ранги III	Сумма рангов	Δ_i	Δ_i^2
1	1	5	7	13	1	1
2	2	3	6	11	—1	1
3	3	4	5	12	0	0
4	4	6	4	14	2	4
5	5	1	3	9	—3	9
6	6	7	2	15	3	9
7	7	2	1	10	—2	4
$n=7$				$\bar{a}=12$		$S'=28$

образом, формула для коэффициента конкордации имеет следующий вид:

$$W = 12S'/[m^2(n^3 - n)]. \qquad (8.21)$$

Значения W заключены в интервале [0, 1]. Равенство W нулю означает полную несогласованность m ранжировок; если же $W=1$, то все m ранжировок совпадают.

Определим значение W по данным табл. 7. Здесь $a=\frac{1}{2}\cdot 3(7+1)=$ $=12$; $S'=28$,

$$W = 12\cdot 28/[3^2(7^3 - 7)] = 0{,}11,$$

т. е. данные три ранжировки согласованы весьма слабо.

Значимость полученной величины W может быть проверена по критерию X^2:

$$X^2 = 12S'/[mn(n+1)] \qquad (8.22)$$

с числом степеней свободы $k=n-1$. В данном примере $X^2_ф=2{,}00$; $k-6$. Для уровня значимости $\alpha=0{,}01$ из табл. 3 приложения находим критическое значение $X^2_{кр}=16{,}81$. Поскольку фактическое значение $X^2_ф$ меньше критического, гипотеза об отсутствии связи между рассматриваемыми ранговыми признаками не отклоняется, т. е. коэффициент W в данном случае не является значимым.

Прикладные аспекты ранговой корреляции. Как уже отмечалось, коэффициенты ранговой корреляции могут использоваться не только для анализа взаимосвязи двух ранговых признаков, но и при определении силы связи между ранговым и количественным признаками. В этом случае значения количественного признака упорядочиваются и им приписываются соответствующие ранги.

Существует ряд ситуаций, когда вычисление коэффициентов ранговой корреляции целесообразно и при определении силы связи двух количественных признаков. Так, при существенном отклонении распределения одного из них (или обоих) от нормального распределения определение уровня значимости выборочного коэффициента корреляции r становится некорректным, в то время как ранговые коэффициенты ρ и τ не сопряжены с такими ограничениями при определении уровня значимости.

Другая ситуация такого рода возникает, когда связь двух количественных признаков имеет нелинейный (но монотонный) характер. Если количество объектов в выборке невелико или если для исследователя существен знак связи, то использование корреляционного отношения η может оказаться здесь неадекватным. Вычисление же коэффициента ранговой корреляции позволяет обойти указанные трудности.

* * *

Завершая рассмотрение проблем анализа взаимосвязи, отметим следующие существенные положения.

1) Величины мер связи признаков различной природы не сравнимы между собой. Например, если величина коэффициента ранговой корреляции τ_{XY} оказалась выше величины коэффициента ассоциации Q_{ZV}, то это не означает, что связь ранговых признаков X и Y «сильнее», чем связь альтернативных признаков Z и V.

2) Если рассматриваемый набор признаков содержит показатели различной природы (номинальные, ранговые, количественные), то для сопоставления силы связи между любой парой признаков обычно используют меры зависимости, пригодные для номинального уровня измерения (информационные коэффициенты, T, K и др.). Такой подход позволяет анализировать в комплексе все связи. При этом, однако, следует учитывать, что возникают определенные потери исходной информации, ее «огрубление». Так, для ранговых признаков теряется информация о соответствующем упорядочении объектов, а значения количественных признаков группируются в интервалы, которые при переводе на номинальный уровень измерения также оказываются неупорядоченными. Иногда такое огрубление полезно, поскольку позволяет количественные данные с грубыми ошибками трактовать как ранговые или даже номинальные. При этом

уменьшение точности компенсируется повышением надежности данных.

С этими проблемами исследователь не сталкивается в том случае, когда все анализируемые признаки характеризуются одинаковым уровнем измерения.

ГЛАВА 9
СТАТИСТИЧЕСКАЯ ПРОВЕРКА ГИПОТЕЗ

Одним из важных инструментов научного исследования является построение и проверка тех или иных гипотез, касающихся объекта изучения, его состава, структуры, связей. Иногда такие гипотезы не формируются в ходе исследования явным образом, однако все же присутствуют в неявной форме.

В исторической науке так же, как и в других областях знания, исследователь либо строит и пытается обосновать некую гипотезу, либо стремится опровергнуть другую гипотезу. Так, примерами могут служить гипотезы о времени формирования всероссийского аграрного рынка, гипотеза о снижении экономического уровня крестьянского хозяйства в России к концу периода крепостничества и т. п. Это примеры научных гипотез, которые выдвигают историки. Теория же статистической проверки гипотез суживает и формализует общенаучное понятие гипотезы, что делает возможным применение этой теории в самых различных областях знания.

§ 1. Основные понятия

Статистическая гипотеза. Понятие статистической гипотезы гораздо уже, чем просто научной гипотезы.

Статистическими гипотезами называют различного рода предположения о свойствах генеральной совокупности (совокупностей), подтверждаемые или отвергаемые методами математической статистики на основе выборочных данных.

Таким образом, ясно, что статистические гипотезы, как и выборочное исследование, связаны с необходимостью делать выводы обо всем явлении, процессе на основе имеющихся данных об его части. Поскольку при статистической проверке гипотез мы не располагаем данными обо всей генеральной совокупности, всегда есть риск совершить ошибку: отклонить верную гипотезу или принять неверную. Поэтому статистическая проверка гипотез является по своей природе вероятностной. Однако на практике чрезвычайно важно, что теория статистической проверки гипотез позволяет оценить вероятность совершить ошибку и дает, таким образом, оценку надежности получаемых выводов.

Хотя из приведенного определения видно, что не всякая гипотеза в обычном смысле является статистической, многие научные гипотезы можно свести к целому ряду статистических гипотез (например, гипотезу о формировании аграрного рынка свести к гипотезе о достижении высоких значений коэффициентами корреляции между ценами на основные сельскохозяйственные продукты по всем губерниям, образующим этот рынок).

Итак, проверка любой статистической гипотезы начинается, естественно, с ее формулирования в соответствии с приведенным определением.

Пример 1. Пусть имеются данные уставных грамот Тамбовской губернии в виде 10%-ной случайной выборки (см. прим. 2, гл. 5), всего 264 грамоты. По этим данным установлено, что средний размер дореформенного надела равен 3,16 дес. Это дает основание считать, что в качестве среднего размера надела в Тамбовской губернии можно принять 3 дес. Нашей гипотезой является гипотеза о равенстве среднего размера дореформенного надела 3 дес. Эта гипотеза может быть проверена и называется она испытуемой или нулевой гипотезой и обозначается в статистике H_0. Сформулируем теперь математически данное определение испытуемой гипотезы: $H_0: X_{\text{г.с}}=3$, где $X_{\text{г.с}}$ обозначает неизвестное нам среднее значение размера надела в генеральной совокупности (Тамбовская губерния). Как же проверить эту гипотезу?

Статистический критерий и статистическая характеристика. Центральным понятием в теории статистической проверки гипотез является понятие **статистического критерия.**

Статистическим критерием называется совокупность строго определенных правил, указывающих, при каких результатах выборочного исследования испытуемая гипотеза отклоняется, а при каких — считается допустимой.

Результаты выборочного исследования конкретизируются в форме статистической характеристики, в соответствии со значениями которой гипотеза отклоняется или признается допустимой. Функция, на основании которой статистический критерий отклоняет или не отклоняет испытуемую гипотезу, называется статистической характеристикой гипотезы.

Вернемся к нашему примеру и выясним смысл введенных понятий.

Как известно, выборочные средние отличаются от генеральной средней, но эти отличия характеризуются тем, что чаще встречаются выборочные средние, близкие к генеральной. Таким образом, наличие выборочной средней, которая сильно отклоняется от предполагаемой генеральной, свидетельствует, скорее всего, о неверности испытуемой гипотезы. Действительно, чем дальше от 3 дес. лежит выборочное значение $\bar{x}$, тем больше у нас оснований сомневаться в справедливости испытуемой гипотезы.

Значит, желательно, чтобы при больших отклонениях выборочной средней от значения 3 ($X_{г.с}=3$) критерий отвергал испытуемую гипотезу, а при небольших отклонениях признавал ее допустимой. Остается решить, какие же отклонения считать большими, а какие — небольшими.

Вспомним (см. гл. 5), что для больших выборок известно распределение выборочных средних, т. е. вероятность появления каждого конкретного значения $\bar{x}$ (выборочного среднего), если известно значение $X_{г.с}$. Оказывается, распределение выборочных средних нормально, и, следовательно, чем больше значение отличается от 3, тем меньше вероятность его появления в конкретной выборке, причем эта вероятность точно известна. Значит, при получении маловероятного значения $\bar{x}$ критерий будет отклонять испытуемую гипотезу.

В основе такого вывода (и в основе всей теории проверки статистических гипотез) лежит так называемый *принцип практической невозможности,* который гласит, что маловероятное событие практически невозможно в единичном случае, каким является наша выборка. К сожалению, нельзя указать годную для всех случаев границу, такую, что событиями с вероятностью, меньшей этой границы, мы пренебрегаем, считая их невозможными. Дозволенная степень риска, который связан с пренебрежением событиями с малой вероятностью, зависит от различного рода обстоятельств и связана с практической важностью следствий, вытекающих из наступления таких событий.

Итак, в соответствии с принципом практической невозможности статистической характеристикой гипотезы H_0 может служить величина отклонения выборочной средней от предполагаемого генерального значения. По табл. 1 приложения можно определить вероятность появления отклонений, превышающих по абсолютному значению данное отклонение ($\bar{x}$—3). Для удобства пользования этой таблицей в качестве статистической характеристики берется не само отклонение $\bar{x}$—3, а так называемое нормированное отклонение t:

$$t=(\bar{x}-\bar{X}_{г.с})/\mu, \qquad (9.1)$$

где $\mu=\sigma/\sqrt{n}$ — стандартная ошибка выборки (см. гл. 5).

В данном случае известно, что $\mu=0{,}08$ (напомним, что пример 1 взят из гл. 5). Получим $t=(\bar{x}-3)/0{,}08$ — это и есть статистическая характеристика. Ее фактическое значение в нашем примере равно

$$t_ф=(3,16-3)/0,08=2.$$

Пользуясь табл. 1 приложения, найдем вероятность $\Phi(t_ф)=$ $=\Phi(2)=0{,}9545$. Напомним, как следует интерпретировать полученное значение. Вероятность 0,9545 означает, что если бы мы произвели 10 000 выборок из нашей генеральной совокупности, то в 9545 из них отклонение $\bar{x}$ до 3 дес. не превысило бы величину 0,16 и в 455 (455=1000—9545) $|\bar{x}-3|$ могло бы превзойти 0,16, разумеется, **при условии,** что средний размер надела в генеральной совокупности действительно равен 3. (Графически это распределение см. на рис. 20). Таким образом, площадь под нормальной кривой между значениями 2,84 и 3,16 составляет 0,9545 всей площади под нормальной кривой.

Осталось выяснить, можно ли считать вероятность 0,0455 настолько малой, чтобы пренебречь площадью справа от точки 3,16 и слева от точки 2,84 по сравнению с площадью под всей кривой (как известно, она равна 1). Если считать вероятность 0,0455 достаточно малой, то значения выборочной средней, равные 3,16 или большие, являются практически невозможными, и гипотезу H_0 придется отклонить. Однако, если считать, что эта вероятность не слишком мала, значение 3,16 вполне возможно и не противоречит нашей гипотезе.

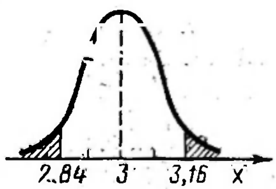

Рис. 20. Нормальное распределение выборочных средних при $\bar{X}_{г.с}=3$

Тем самым очевидно, что наши выводы существенно зависят от того значения вероятности, при котором x считается практически невозможным для данной испытуемой гипотезы.

Уровень значимости и критическая область. Значение вероятности, начиная с которого событие считается практически невозможным, называется **уровнем значимости** критерия.

Уровень значимости обозначается α и обычно полагается равным 0,1; 0,05; 0,025 или 0,01. (Например, уровень значимости 0,01 или 1% означает, что мы считаем практически невозможными события, вероятность наступления которых не более 1%, т. е. события, которые могут произойти не более чем в 1 случае из 100.)

Рассмотрим на нашем примере, какие значения статистической характеристики t соответствуют уровню значимости 0,025. По табл. 1 приложения для $\Phi(t)=1-\alpha=$ $=1-0{,}025=0{,}975$ находим $t=2{,}24$. Это значит, что с вероятностью 0,025 значение выборочного среднего может отклониться от предполагаемого генерального значения больше чем на $t\mu=2{,}24\cdot 0{,}08=0{,}18$, т. е. практически невозможными считаются значения $\bar{x}$, большие 3,18 $(3+$ $+0{,}18)$ и меньшие 2,82 $(3-0{,}18)$.

Вспомним, что в данном случае было получено $\bar{x}_{\Phi}=3{,}16$, которое, следовательно, не является практически невозможным, и при таком уровне значимости (0,025) испытуемая гипотеза не отклоняется.

Следует подчеркнуть, что уровень значимости выбирается исследователем в зависимости от конкретной задачи.

Каждому уровню значимости соответствует **критическое значение** статистической характеристики, которое делит все множество значений характеристики на две области: **допустимых значений** и **критическую.**

Критической областью испытуемой гипотезы являются все значения статистической характеристики, вероятность появления которых меньше выбранного уровня значимости. Все остальные значения статистической характеристики образуют область допустимых значений.

В нашем примере критической областью являются те значения t, для которых верно $|t|\geq t_{\text{кр}}$, а областью допустимых значений — те, для которых $|t|<t_{\text{кр}}$.

Между статистической характеристикой, выбранным уровнем значимости и критической областью существует следующее соотношение.

Вероятность того, что статистическая характеристика попадет в критическую область, если верна испытуемая гипотеза, равна выбранному уровню значимости α, т. е. критическая область содержит именно те значения статистической характеристики, которые мы считаем практически невозможными.

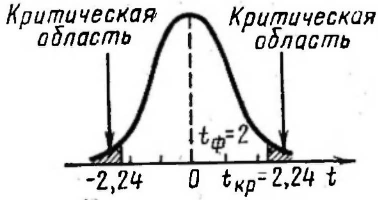

Рис. 21. Критическая область для уровня значимости $\alpha=0{,}025$

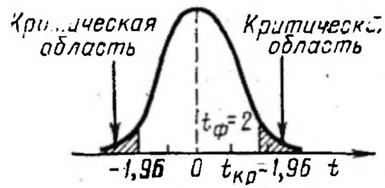

Рис. 22. Критическая область для уровня значимости $\alpha=0{,}05$

Поясним изложенное на графике (рис. 21). Это график нормального распределения для $t=\frac{x-\overline{X}_{г.с}}{\mu}$ $(\overline{X}_{г.с}=3)$. Заштрихована критическая область, соответствующая уровню значимости $\alpha=0{,}025$. Видно, что значение $t_ф=2$ попадает в область допустимых значений, т. е. не является практически невозможным, если $\overline{X}_{г.с}=3$; значит, испытуемая гипотеза не отклоняется.

Из приведенного примера видно, что статистическая гипотеза проверяется в такой последовательности:

1) формулируется испытуемая гипотеза (в нашем примере $H_0:\overline{X}_{г.с}=3$); 2) определяется статистическая характеристика этой гипотезы $\left(t=\frac{\overline{x}-X_{г.с}}{\mu}\right)$; 3) рассчитывается ее фактическое значение $(t_ф=2)$; 4) выбирается уровень значимости $\alpha(0{,}025)$; 5) определяется соответствующая ему критическая область $(|t|\geqslant 2{,}24)$. Если фактическое значение статистической характеристики попадает в критическую область, критерий отклоняет испытуемую гипотезу, если в область допустимых значений — гипотеза признается допустимой (в нашем примере $t_ф<t_{кр}$, и гипотеза не отклоняется на уровне значимости 0,025).

Если гипотеза не отклоняется, это еще не значит, что она верна: дальнейшие исследования могут привести к отклонению гипотезы, но наш критерий не дает оснований отклонить ее.

Очевидно, попадание или непопадание точки $t_ф$ в критическую область зависит от размеров этой области, а они, в свою очередь, зависят от величины α. Например, если увеличить α и вместо 0,025 взять 0,05, получим (табл. 1 приложения) $t_{кр}=1{,}96$, при этом критическая область расширится и $t_ф$ попадет в нее, что означает отклонение испытуемой гипотезы (см. рис. 22). Какой же результат испытания гипотезы считать более обоснованным и не противоречат ли они друг другу?

Ошибка первого рода. Выясним смысл уровня значимости α. Если $\alpha=0{,}05$, это говорит о том, что в 5 выборках из 100 **могут** все же получиться отклонения выборочного среднего от 3 дес., превышающие критическое значение, и **при справедливости** испытуемой гипотезы (за счет случайностей выборки). Однако при этом $t_ф$ попадет в критическую область, и критерий **отклонит** гипотезу, тогда как она **верна.** Таким образом, попадание в критическую область не обязательно связано с отклонением действительно неверной гипотезы — в 5 случаях из 100 это означает ошибочное отклонение верной гипотезы, и, значит, величина α — это риск совершить так называемую **ошибку первого рода** (величина $P=1-\alpha$ называется уровнем доверия, эта вероятность правильного результата проверки гипотезы).

Ошибкой первого рода называется вероятность отклонить испытуемую гипотезу, когда она верна.

Очевидно, для $\alpha=0{,}025$ риск совершить ошибку первого рода меньше, чем для $\alpha=0{,}05$. На первый взгляд, надо стремиться уменьшить α, уменьшая тем самым вероятность ошибки первого рода. Но из сравнения рис. 21 и рис. 22 видно, что с уменьшением α критическая область сужается и, следовательно, расширяется область допустимых значений, т. е. становятся вполне допустимыми значения $\bar{x}$, даже далекие от 3 дес., а такие значения, возможно, лучше соответствуют каким-то другим гипотезам.

Следовательно, слишком широкая область допустимых значений способствует неотклонению испытуемой гипотезы, когда она неверна, т. е. все реже неотклонение будет эквивалентно принятию гипотезы. Например, если вместо $H_0: X_{г.с}=3$ испытать гипотезу $H_0: X_{г.с}=3{,}2$, по формуле (9.1) получим $t_ф=\dfrac{3{,}16-3{,}2}{0{,}08}=-0{,}5$, причем снова $|t_ф|<t_{кр}=2{,}24$ $(\alpha=0{,}025)$. Таким образом, на уровне значимости $\alpha=0{,}025$ эта гипотеза тоже не отклоняется, значит, наши данные не противоречат не только гипотезе $\bar{X}_{г.с}=3$, но и гипотезе $X_{г.с}=3{,}2$, а также и любым гипотезам, дающим значения $\bar{X}_{г.с}$ в интервале 3—3,2 дес.

Ошибка второго рода. Оказывается, наш критерий имеет еще одну слабость: если истинным значением $\bar{X}_{г.с}$ является 3,2 дес., а не 3 дес., критерий не сможет этого уловить, т. е. не сможет отклонить испытуемую гипотезу $(H_0: \bar{X}_{г.с}=3)$, хотя она и неверна. И чем меньше уровень значимости α и уже соответствующая ему критическая область, тем чаще возникают подобные ситуации. Следовательно, с уменьшением α повышается вероятность так называемой **ошибки второго рода** — ошибочного принятия неверной гипотезы.

Ошибкой второго рода называется неотклонение испытуемой гипотезы H, когда на самом деле она неверна. Вероятность ошибки второго рода обозначается β.

Теперь становится ясно, почему неотклонение гипотезы еще не равносильно ее принятию. Ведь если в такой ситуации велика вероятность ошибки второго рода, мы рискуем принять H_0 тогда, когда верна какая-то другая гипотеза. При испытании гипотезы H_0 можно получить четыре возможных результата:

I. Гипотеза H_0 верна: 1) она не отклоняется (правильный результат), 2) она отклоняется (ошибка первого рода).

II. Гипотеза H_0 неверна: 3) она отклоняется (правильный результат), 4) она не отклоняется (ошибка второго рода).

Итак, вероятность ошибки второго рода возникает в тех случаях, где нулевая гипотеза не отклоняется.

В таких случаях требуется проверить, не соответствуют ли данные выборки другой, конкурирующей с H_0, гипотезе, т. е. требуется выдвинуть **альтернативную** гипотезу, обозначаемую H_a.

Однако альтернативная гипотеза может быть сформулирована по-разному. Например, в качестве гипотезы H_a можно взять следующие гипотезы:

а) средний размер надела не равен 3 дес.:

$$H_a : \overline{X}_{г.с} \neq 3;$$

б) средний размер надела больше 3 дес.:

$$H_a : \overline{X}_{г.с} > 3;$$

в) средний размер надела меньше 3 дес.:

$$H_a : \overline{X}_{г.с} < 3;$$

г) средний размер надела равен 3,2 дес.:

$$H_a : \overline{X}_{г.с} = 3{,}2;$$

д) средний размер надела равен 3,5 дес. (3,1 дес., 3,8 дес. и т. п.):

$$H_a : \overline{X}_{г.с} = 3{,}5 (3{,}1;\ 3{,}8 \text{ и т. п.}).$$

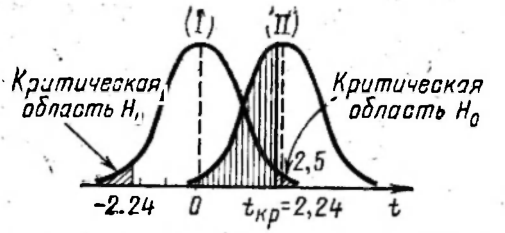

Рис. 23. Нормальное распределение статистической характеристики для испытуемой (I) и альтернативной (II) гипотез

Все эти гипотезы являются альтернативными (противоположными) для H_0, но имеют, очевидно, различный смысл. Заметим, что гипотезы г) и д) (так же, как испытуемая гипотеза H_0) утверждают равенство $\bar{X}_{г.с}$ одному, вполне определенному числу. Это так называемые **простые гипотезы.** Гипотезы а) — в) в отличие от простых относятся не к отдельным значениям $\bar{X}_{г.с}$, а к целым совокупностям значений $\bar{X}_{г.с}$. Так, гипотеза б) включает в себя как частные случаи гипотезы г), д) и множество подобных им гипотез $H_а: \bar{X}_{г.с} = a$, где $a > 3$.

Как правило, на практике чаще всего встречаются сложные гипотезы, особенно в общественных науках, что отражает большую сложность анализируемых явлений. Допустим, что нас интересует альтернативная гипотеза $H_а: \bar{X}_{г.с} > 3$, и будем считать несущественными малые (менее 0,2 дес.) различия между альтернативной и испытуемой гипотезами. Тогда гипотезу $H_а: \bar{X}_{г.с} > 3$ можно заменить гипотезой $H_а: \bar{X}_{г.с} \geqslant 3{,}2$, а ее, в свою очередь, уже простой гипотезой $H_а: \bar{X}_{г.с} = 3{,}2$. Рассмотрим этот случай. Снова изобразим графически нормальное распределение выборочных средних для нулевой гипотезы $H_0: \bar{X}_{г.с} = 3$ (I) и его критическую область (для $\alpha = 0{,}025$) (см. рис. 23). На этом же рисунке покажем, как выглядит нормальное распределение, если верна альтернативная гипотеза: $H_а: \bar{X}_{г.с} = 3{,}2$ (II). Очевидно, эта кривая будет смещена вправо на $(3{,}2 - 3)/0{,}08 = 2{,}5$ единицы.

Посмотрим, как часто значения, попадающие в область допустимых значений испытуемой гипотезы, будут соответствовать альтернативной гипотезе $H_а$, т. е. определим, какая часть площади под нормальной кривой (II) располагается левее точки $t_{кр}$. Эта часть на рисунке заштрихована вертикально. Оказывается, значительная часть кривой (II) попала в область допустимых значений кривой (I), а это и значит, что вероятность ошибки II рода велика (не приводя вычислений, дадим значение этой вероятности: $\beta \approx 0{,}4$). Если же в качестве альтернативной гипотезы выбрать $H_а: \bar{X}_{г.с} = 3{,}1$, получим $\beta = 0{,}84$, т. е. чем ближе друг другу испытуемая и альтернативная гипотезы, тем выше вероятность ошибки II рода.

Мощность критерия. Очевидно, чем меньше величина β, тем лучше критерий «различает» гипотезы, т. е. испытуемая гипотеза редко признается допустимой, если она на самом деле неверна, поэтому величина $1-\beta$ называется **мощностью** критерия. •

Величина $1-\beta$, т. е. вероятность **правильного** отклонения испытуемой гипотезы (когда верна альтернативная гипотеза), называется мощностью критерия.

Как показано в предыдущем разделе, мощность критерия зависит от степени «близости» между испытуемой и альтернативной гипотезами. Как правило, выявление незначительных различий между гипотезами не имеет практического значения, поэтому в качестве альтернативной гипотезы обычно берут гипотезу, значимо отличающуюся от испытуемой. В нашем примере мы считаем несущественными различия менее 0,2 дес. и при этом получим невысокую мощность критерия ($\approx 0{,}6$), которая может нас не удовлетворить. Эта мощность означает, что с вероятностью α ($\alpha=0{,}025$) в 4 выборках из 10 гипотеза H_0 будет **ошибочно** считаться допустимой.

Но если, например, считать значимыми различия в 0,5 дес., то в качестве альтернативной гипотезы для $H_0: X_{г.с}=3$ можно взять уже $H_а: X_{г.с}\geqslant 3{,}5$, что дает мощность критерия, близкую 1, и в такой ситуации можно утверждать, что неотклонение гипотезы равносильно признанию того, что она верна.

Однако если нельзя повысить мощность критерия таким образом (увеличивая «расстояние» между гипотезами) и приходится иметь дело с фиксированной альтернативной гипотезой (например, $H_а: \overline{X}_{г.с}=3{,}2$), то мощность критерия можно повысить за счет увеличения объема выборки до тех пор, пока ошибка второго рода не станет приемлемой для целей исследования.

Односторонняя и двусторонняя проверка. В заключение первого раздела необходимо подчеркнуть, что при изложении основных понятий и этапов проверки гипотезы мы пользовались так называемой **двусторонней проверкой** — нас одинаково интересовали как положительные, так и отрицательные отклонения $\bar{x}$ от $\overline{X}_{г.с}$. Именно поэтому критическая область у нас состояла из двух частей. Однако возможны ситуации, когда нас интересуют только положительные или только отрицательные отклонения, (например, если средний размер надела не может быть меньше 3 дес.). В таких случаях вместо двусторонней проводят **одностороннюю проверку.** При этом критическая область состоит из одного

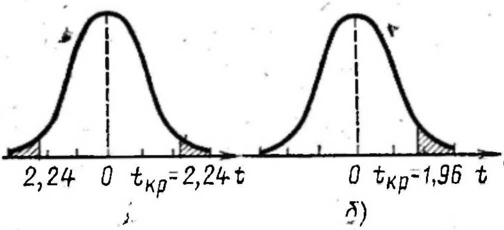

Рис. 24. Односторонняя (*а*) и двусторонняя (*б*) критические области

участка и соответственно имеется только одно критическое значение, несколько меньшее по абсолютной величине, чем $t_{кр}$ для двусторонней проверки (при одном и том же α). Так, на рис. 24, *а*, соответствующем двусторонней проверке, обе заштрихованные области вместе составляют 2,5% площади нормальной кривой, тогда как на рис. 24, *б*, соответствующем односторонней проверке, эти же 2,5% площади содержатся в одном участке критической области.

Значение $t_{кр}$ для односторонней проверки находится по табл. 7 приложения (см. гл. 4). Для уровня 0,025 $t_{кр}=1{,}96$.

Заметим, что если бы вместо двусторонней проверки гипотезы $H_0:\overline{X}_{г.с}=3$ проводилась ее односторонняя проверка, то на уровне 2,5% эта гипотеза была бы отклонена, ибо $t_ф=2>1{,}96$. Эту проверку можно было бы проводить, если бы мы были уверены, что большие отрицательные отклонения невозможны (например, если бы средний размер надела не мог быть меньше 3 дес.). Однако если нет оснований считать направление отклонений вполне определенным, следует использовать, как более строгий, двусторонний критерий.

§ 2. Критерии для средних

Критерий для средней малой выборки. В § 1 рассмотрен критерий для среднего значения большой выборки. Напомним постановку этой задачи. По данным выборки получено среднее значение признака $\bar{x}$ (размер надела).

1) Проверяется гипотеза о том, что данная выборка получена из генеральной совокупности со средним значением $X_{г.с}$. 2) В качестве меры отклонения выборочного среднего от генерального берется нормированное отклонение $t=(\bar{x}-\overline{X}_{г.с})/\mu$ (где $\mu=\sigma/\sqrt{n}$ — средняя ошибка выборки, а σ — среднее квадратическое отклонение, см. формулу (4.7) для большой выборки). Величина t является статистической характеристикой. 3) Подсчитывается ее фактическое значение $t_ф$. 4) Выбирается уровень значимости α (обычно $\alpha=0{,}05$; 0,1; 0,025; 0,01). 5) По заданному уровню значимости α и табл. 1 приложения находится $t_{кр}$ и строится критическая область испытуемой гипотезы. 6) При попадании $t_ф$ в эту область гипотеза отклоняется.

Сохраним эту схему и при дальнейшем изложении основных критериев. В частности, построим критерий для проверки гипотезы о среднем значении малой выборки. Этот критерий будет отличаться от рассмотренного своей статистической характеристикой. Известно, что для малых выборок (см. гл. 5, § 4) в формулу средней ошибки выборки μ подставляется не $\sigma_{\text{выб}}$, а величина S:

$$S = \sqrt{\frac{\sum_{i=1}^{k} (x_i - \bar{x})^2 n_i}{n-1}} = \sigma \sqrt{\frac{n}{n-1}}, \tag{9.2}$$

где n_i — частота значения x_i: $\sum_{i=1}^{k} n_i = n$: k — число вариант признака X; n — объем выборки (ср. с формулой (5.15) гл. 5).

Значение S отличается от σ тем, что вместо n содержит в знаменателе $n-1$.

Следовательно, статистической характеристикой гипотезы о среднем значении $\bar{X}_{\text{г.с}}$ по данным малой выборки является нормированное отклонение t:

$$t = \frac{\bar{x} - \bar{X}_{\text{г.с}}}{S/\sqrt{n}}, \tag{9.3}$$

где S вычисляется по формуле (9.2). Величина t подчиняется уже не нормальному распределению, а распределению Стьюдента с k степенями свободы, где $k=n-1$. С увеличением k это распределение приближается к нормальному, поэтому при $n>30$ проверку гипотезы о среднем значении в генеральной совокупности производят так, как показано в § 1, пользуясь таблицей нормального распределения. При $n \leqslant 30$ надо использовать таблицу распределения Стьюдента (табл. 2 приложения), в которой по заданным уровню значимости α и числу степеней свободы $k=n-1$ находят значение $t_{\text{кр}}$.

Пример 2. Пользуясь схемой проверки гипотез, проверим гипотезу о среднем на примере. Снова воспользуемся примером 11 из гл. 5. По 16 уездам черноземной полосы получен средний размер оброка (руб. сер. на муж. душу) в конце XVIII в., равный 4,6 руб. Величина S подсчитана по формуле (9.2) и равна 1,7 руб.

1) Сначала сформулируем испытуемую гипотезу. Пусть такой гипотезой является гипотеза о том, что средний размер оброка для черноземной полосы равен 4 руб.: $H_0: \bar{X}_{\text{г.с}} = 4$.

2) В качестве статистической характеристики берем t, вычисленное по формуле (9.3).

3) Определим $t_{\phi} = \frac{4,6-4}{1,7/\sqrt{16}} = \frac{0,6 \cdot 4}{1,7} \approx 1,4.$

4) В качестве α выберем 0,05 или 5%. При этом значении α мы согласны на риск ошибочно отклонить правильную гипотезу в 5 случаях из 100.

5) Найдем $t_{кр}$ по табл. 2 приложения, причем сначала допустим, что проводится двусторонняя проверка, т. е. гипотеза H_0 отклоняется, если выборочное среднее сильно отклоняется от $X_{г.с}$ как в положительную, так и в отрицательную сторону. В таком случае по $\alpha=0,05$ и $k=16-1=15$ в табл. 2 найдем $t_{кр}=2,13$. Значит, критической области принадлежат те значения t, которые по абсолютному значению превосходят 2,13.

6) Так как $t_{\phi}=1,4<2,13$, гипотеза не отклоняется.

При неотклонении гипотезы, как указано в § 1, возможна ошибка второго рода, т. е. ситуация, когда неверная испытуемая гипотеза не отклоняется. Вычисление вероятности ошибки второго рода (β) требует построения альтернативной гипотезы и часто весьма трудоемко. Подчеркнем лишь то, что вероятность ошибки второго рода тем меньше, чем «дальше» альтернативная гипотеза от испытуемой и чем больше объем выборки.

Малая вероятность ошибки второго рода и соответственно большая мощность критерия позволяют утверждать, что неотклонение гипотезы равносильно ее справедливости.

Критерий для разности средних значений. Нередко возникает необходимость сравнения двух генеральных совокупностей путем сравнения выборок из этих совокупностей. Допустим, что имеются две выборки объемом n_1 и n_2 со средними значениями $\bar{x}_1$ и $\bar{x}_2$.

Мы утверждаем, что эти выборки принадлежат генеральным совокупностям, в которых средние значения совпадают, а наблюдаемое различие между $\bar{x}_1$ и $\bar{x}_2$ объясняется случайными причинами. При этом обычно предполагается, что генеральные дисперсии этих совокупностей равны ($\sigma_1^2=\sigma_2^2$).

Пример 3. Вернемся к примеру 2. Размер оброка в выборке 16 уездов черноземной полосы равен 4,6 руб. ($S=1,7$ руб.). Допустим, что по 16 уездам нечерноземной полосы размер оброка равен 3 руб. ($S=1,26$ руб.) (см. пример 11, гл. 5). Требуется проверить, равны ли средние размеры оброка в черноземной и нечерноземной зонах в целом.

1) Сформулируем испытуемую гипотезу: различие между выборочными средними случайно, т. е. $H_0: \bar{X}_{1\,г.с}=\bar{X}_{2\,г.с}$.

2) В качестве статистической характеристики снова используется величина t, имеющая распределение Стьюдента. В данном случае

$$t = \frac{\bar{x}_1 - \bar{x}_2}{S\sqrt{1/n_1 + 1/n_2}}, \tag{9.4}$$

где $\bar{x}_1$ и $\bar{x}_2$ — выборочные средние; n_1 и n_2 — объемы выборок; S — оценка неизвестного нам значения σ в генеральных совокупностях по малым выборкам (напомним, что предполагается $\sigma_1 = \sigma_2$);

$$S = \sqrt{\frac{S_1^2 k_1 + S_2^2 k_2}{k_1 + k_2}}, \qquad (9.5)$$

где k_1 и k_2 — числа степеней свободы в выборках ($k_1 = n_1 - 1$, $k_2 = n_2 - 1$); S_1 и S_2 — значения, полученные по формуле (9.2).

Если S_1 и S_2 предварительно не подсчитаны, то для вычисления S удобнее пользоваться следующей формулой:

$$S = \sqrt{\frac{\sum_{i=1}^{n_1} (x_i - \bar{x}_1)^2 + \sum_{i=1}^{n_2} (x_i - \bar{x}_2)^2}{n_1 + n_2 - 2}}. \qquad (9.6)$$

3) Рассчитаем $t_ф$.

Для этого сначала вычислим S по формуле (9.5):

$$S = \sqrt{\frac{(1,7)^2 \cdot 15 + (1,26)^2 \cdot 15}{30}} \approx 1,5.$$

Следовательно,

$$t_ф = \frac{4,6 - 3}{1,5\sqrt{1/16 + 1/16}} \approx 3.$$

4) Выберем $\alpha = 0,05$.

5) По табл. 2 приложения для $\alpha = 0,05$ и $k = k_1 + k_2 = 30$ найдем $t_{кр} = 2,04$.

6) Видно, что $t_ф$ значительно превышает критическое значение, и гипотеза о том, что размеры оброка в черноземной и нечерноземной полосах были одинаковы, отклоняется.

Заметим, что в случае больших выборок в качестве статистической характеристики можно использовать нормированное отклонение t:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sigma/\sqrt{n}}, \qquad (9.7)$$

где $n = n_1 + n_2$; σ — оценка неизвестного значения среднего квадратического отклонения в генеральной совокупности по большим выборкам,

$$\sigma = \sqrt{\frac{\sigma_1^2 n_1 + \sigma_2^2 n_2}{n_1 + n_2}}, \qquad (9.8)$$

σ_1 и σ_2 — средние квадратические отклонения для больших выборок.

Если σ_1 и σ_2 заранее не подсчитаны, то вместо формулы (9.8) можно использовать следующую формулу:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n_1} (x_i - \bar{x}_1)^2 + \sum_{i=1}^{n_2} (x_i - \bar{x}_2)^2}{n_1 + n_2}}. \qquad (9.9)$$

При использовании формул (9.7—9.9) вместо (9.4—9.6), очевидно, берется табл. I приложения.

Замечание. Иногда, хотя довольно редко, значения σ_1 и σ_2 в генеральных совокупностях бывают известны. Тогда формула (9.7) приобретает следующий вид:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}.$$

Ограничения при использовании распределения Стьюдента t. В предыдущих разделах упоминалось, что величина t, используемая при проверке гипотез о средних значениях малых выборок, имеет распределение Стьюдента. Однако это верно не всегда. Для того чтобы использовать это распределение, необходимо, чтобы соблюдались следующие условия: 1) чтобы значения x_i признака были независимы, 2) чтобы распределение этого признака в генеральной совокупности было нормально, 3) кроме того, для проверки гипотезы о равенстве средних — чтобы дисперсии генеральных совокупностей были равны.

Первое из этих требований обычно выполняется, а второе и третье можно проверить с помощью специальных критериев, которые будут рассмотрены в следующих параграфах.

Критерий знаков. Это один из простейших критериев, который часто применяется для предварительного суждения о генеральной совокупности. Обычно критерий знаков используется при проверке гипотезы о том, что два признака в генеральной совокупности имеют одинаковые средние значения. Мы уже знаем, что гипотезу о равенстве средних значений можно проверить с помощью распределения Стьюдента, однако критерий знаков имеет большое преимущество, состоящее в том, что для его применения не нужно ставить никаких ограничений относительно характера распределения признака в генеральной совокупности. Это распределение может быть отличным от нормального и даже вообще неизвестным.

Пример 4. Рассмотрим выборку семей с одинаковым уровнем дохода и одинаковым числом членов семьи (5). Для каждой семьи известно значение признака «расходы на культурные нужды» (в % от общего дохода) и признака «расходы на транспорт» (в %). Обозначим значения первого признака x_i, а значения второго y_i. Если предположить, что для семей из 5 человек с данным уровнем дохода расходы на культурные нужды равны расходам на транспорт, то, подсчитывая знаки попарных разностей $x_i—y_i$ для каждой семьи, мы должны, очевидно, получить примерно одинаковое число плюсов и минусов (т. е. при равенстве этих двух статей расходов число семей, где расходы на

Таблица 1. Распределение расходов в разных семьях

	1	2	3	4	5	6	7	8
Расходы на культнужды, руб.	12,1	6,2	6,0	8,4	4,2	6,8	6,9	20,0
Расходы на транспорт, руб.	5,3	5,7	4,2	4,6	10,8	5,0	5,2	12,7
Знаки разностей	+	−	+	+	−	+	+	+

	9	10	11	12	13	14	15
Расходы на культнужды, руб.	13,5	7,2	8,5	10,7	7,9	1,3	6,9
Расходы на транспорт, руб.	6,3	3,2	5,0	4,8	5,7	8,4	6,7
Знаки разностей	+	+	+	+	+	−	+

транспорт превышают расходы па культурные нужды, будет примерно равно числу семей, где это соотношение обратно). Если же будут заметно преобладать положительные или отрицательные разности, можно будет сделать вывод о том, что расходы по одной статье преобладают над расходами по другой в генеральной совокупности.

На таких простых соображениях и основан критерий знаков. Обозначим d_1 — количество положительных разностей, а d_2 — количество отрицательных. Пусть d — минимальное из этих двух чисел. Распределение d известно и приводится в табл. 5 приложения, где для $\alpha=0{,}01$; $0{,}02$; $0{,}05$ даны границы критической области для d.

Возьмем конкретную выборку (см. табл. 1).

1) Проверим гипотезу о равенстве расходов на транспорт на культурные нужды: $H_0: \overline{X}_{1\,\text{г.с}}=X_{2\,\text{г.с}}$.

2) В качестве статистической характеристики берется d — минимальное из чисел d_1 и d_2 (число плюсов и число минусов).

3) Находим $d_ф=2$.

4) Выберем в качестве α значение 0,01.

5) По табл. 5 приложения для $n=15$ и $\alpha=0{,}1$ получим $d_{кр_1}=3$, $d_{кр_2}=12$. Так как слишком малые и слишком большие значения d противоречат испытуемой гипотезе (ибо при

$X_{1\,г.с} = X_{2\,г.с}$ $a_1 \approx a_2 \approx n/2$, где n — объем выборки), то критическую область образуют значения $d \leqslant 3$ и $d \geqslant 12$.

6) Значение $d_ф = 2$ попадает в критическую область, и гипотеза отклоняется на уровне значимости 0,1.

§ 3. Критерии для дисперсий

Критерий равенства дисперсии генеральной совокупности определенному значению. Часто для выявления характера изменения признака в генеральной совокупности необходимо оценить не только среднее значение этого признака, но и степень его рассеяния или дисперсию на основании выборочных данных.

Пример 5. Вернемся к примеру 2 § 2 о величине оброка в черноземной полосе в конце XVIII в. Допустим, что по результатам других исследователей принято считать $\sigma_{г.с} = 1{,}2$ руб. Верно ли это? В качестве оценки значения σ в генеральной совокупности по выборочным данным получена величина S. Напомним формулу для вычисления S:

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

Известно, что $S = 1{,}7$ руб.

1) Проверим гипотезу о том, что различие между $\sigma_{г.с}$ и S вызвано случайными причинами, т. е. что числовое значение среднего квадратического отклонения признака (оброк) в генеральной совокупности равно 1,2 руб.:

$$H_0: \sigma_{г.с} = 1{,}2.$$

2) В качестве статистической характеристики используется величина

$$X^2 = \frac{S^2 k}{\sigma^2_{г.с}}, \qquad (9.10)$$

где k — число степеней свободы ($k = n - 1$); S — оценка $\sigma_{г.с}$, полученная по формуле (9.2). Известно, что величина $S^2 k / \sigma^2_{г.с}$ подчинена так называемому распределению X^2 с k степенями свободы. Значения вероятностей для X^2 даны в табл. 3 приложения.

3) Найдем фактическое значение статистической характеристики:

$$X^2_ф = \frac{(1{,}7)^2 \cdot 15}{(1{,}2)^2} = 30{,}14.$$

4) Выберем уровень значимости α, например $\alpha=0{,}1$.

5) По табл. 3 приложения для $\alpha=0{,}1$ и $k=15$ найдем $X^2_{кр}=22{,}3$.

Заметим, что в таблице X^2 даны значения, соответствующие односторонней проверке, т. е. противоречащими нашей гипотезе считаются слишком большие значения характеристики (бо́льшие $X^2_{кр}$). Однако слишком маленькие значения X^2 также могут привести к отклонению гипотезы, поэтому проведем двустороннюю проверку нашей гипотезы. Для этого в табл. 3 найдем значения $X^2_{кр}$, соответствующие вероятностям $\alpha/2$ и $1-\alpha/2$ (т. е. 0,05 и 0,95). Получим $X^2_{кр\,1}=7{,}26$ и $X^2_{кр\,2}=25{,}00$. При этом лишь в 1 выборке из 10 ($\alpha=0{,}1$) значение X^2 может лежать вне границ 7,26—25,00. Областью допустимых значений будет интервал 7,26—25,00, а критическая область будет состоять из значений 0—7,26 и значений, больших 25.

6) Как видно, и односторонняя, и двусторонняя проверки приводят к одному результату — испытуемая гипотеза отклоняется ($X_ф^2=30{,}14$), мы не можем считать, что наша выборка получена из генеральной совокупности со значением $\sigma_{г.с}=1{,}2$ руб.

Замечание. В случае больших выборок распределение X^2 приближается к нормальному. Если $k>30$, то в качестве статистической характеристики можно использовать величину

$$t=\sqrt{2X^2}-\sqrt{2k-1}, \tag{9.11}$$

которая подчинена нормальному распределению. Пользуясь табл. 1 приложения, можно получить $t_{кр}$, откуда

$$X^2_{кр}=(1/2)\left(t_{кр}+\sqrt{(2k-1)}\right)^2. \tag{9.12}$$

Критерий равенства двух дисперсий. Пример 6. Вспомним пример 2. Нам были известны значения средних двух выборок, $\bar{x}_1$ и $\bar{x}_2$. Требовалось проверить гипотезу о том, что в генеральных совокупностях средние равны ($\bar{X}_{1\,г.с}=\bar{X}_{2\,г.с}$). При этом использовалось существенное предположение о равенстве генеральных дисперсий ($\sigma_1^2=\sigma_2^2$). Это предположение можно проверить, используя оценки S_1^2 и S_2^2, полученные по формуле (9.2).

Если взять отношение S_1^2/S_2^2, то видно, что для хорошего согласования с гипотезой о равенстве дисперсий это отношение не должно быть очень большим или очень малым. Величина S_1^2/S_2^2 называется **дисперсионным отношением** и обозначается F. Она имеет так называемое F-распределение с k_1 и k_2 степенями свободы ($k_1=n_1-1$, $k_2=n_2-1$).

В табл. 4А, 4Б приложения приведены значения F для вероятностей $\alpha=0{,}05$ и 0,01.

1) Таким образом, испытуемой гипотезой является гипотеза о равенстве генеральных дисперсий: $H_0 : \sigma_1^2=\sigma_2^2$.

2) Статистической характеристикой является дисперсионное отношение F:

$$F=S_1^2/S_2^2, \tag{9.13}$$

где S_1 и S_2 подсчитываются для выборок по формуле (9.2).

3) Находим фактическое значение F:

$$F_\text{ф}=(1{,}7)^2/(1{,}26)^2\approx 1{,}8.$$

4) Выберем в качестве α значение 0,05.

5) Пользуясь табл. 4А для $\alpha=0{,}05$, $k_1=k_2=15$, найдем $F_\text{кр}=2{,}40$.

Заметим, что для использования табл. 4А в качестве S_1 надо брать бо́льшее из двух значений S, тогда F всегда больше 1.

Так как $F=1$ дает наилучшее соответствие испытуемой гипотезе, то гипотеза будет отклоняться, если F сильно превосходит 1, а это соответствует односторонней проверке. Таким образом, в данном случае наш критерий является односторонним, и критическую область образуют значения F, превышающие $F_\text{кр}$.

6) Сравнивая $F_\text{ф}$ с $F_\text{кр}$, видим, что $F_\text{ф}$ попадает в область допустимых значений, и на уровне значимости 0,05 гипотеза не отклоняется.

Ограничения при использовании критериев для дисперсий. При использовании в проверке статистических гипотез распределений χ^2 и F необходимо учитывать следующие требования: 1) независимость выборочных значений X_i, 2) нормальность распределения признака в генеральной совокупности.

§ 4. Критерий для связей

При изучении совокупности объектов, обладающих несколькими признаками, рано или поздно возникает вопрос о связи между этими признаками, о зависимости значений одного признака от значений другого. Для решения таких вопросов разработано много методов, в том числе и в теории проверки гипотез. Рассмотрим сначала вопрос о наличии или отсутствии связи между двумя качественными при-

Таблица 2. Распределение группы рабочих по двум признакам

Образование	Перевыполняют план		
	Да	Нет	Σ
Начальное	5	25	30
Неполное среднее	15	25	40
Среднее и среднее специальное	50	10	30
Σ	40	60	100

Таблица 3. Распределение группы рабочих по двум признакам

Образование	Перевыполняют план		
	Да	Нет	Σ
Начальное	12	18	30
Неполное среднее	16	24	40
Среднее и среднее специальное	12	18	30
Σ	40	60	100

знаками (более подробно качественные признаки рассмотрены в гл. 8).

Критерий независимости. Известно, что распределение двух качественных признаков задается таблицей, содержащей ml клеток, где m — число градаций одного признака, а l — число градаций другого. В каждой клетке таблицы записывается частота n_{ij} (i — номер строки, j — номер столбца), равная количеству объектов совокупности, имеющих i-ю градацию одного признака и j-ю градацию другого (табл. 2).

Как установить по данным этой таблицы наличие или отсутствие связи между уровнем образования и степенью выполнения плана?

Заметим, что доля рабочих, перевыполняющих план, в первой группе (с начальным образованием) составляет 5/30, или $\approx 16\%$, во второй группе (с неполным средним образованием) 15/40, или $\sim 38\%$, а в третьей — 20/30, или $\approx 67\%$.

Однако если бы связи между признаками не было, то процент перевыполняющих план в каждой группе был бы одинаков и совпадал бы со значением, стоящим в итоговой строке (40/100=40% — всего 40% рабочих перевыполняют план). Значит, в каждой группе рабочих, имеющих одинаковое образование, план бы перевыполняли 40% независимо от этого образования (см. табл. 3).

Табл. 3 строится следующим образом. Первое число каждой строки получено как 40% от итога этой строки, т. е.

на первом месте в первой строке стоит $30 \cdot 0{,}4 = 12$, во второй строке $40 \cdot 0{,}4 = 16$, в третьей — $30 \cdot 0{,}4 = 12$. Соответственно на втором месте в каждой строке стоит число, равное 60% итога этой строки. При этом итоги всех строк и столбцов сохраняются.

Выразим эти числа через итоги строк и столбцов. Таким образом, в случае «независимости» признаков в i-й строке и j-м столбце таблицы стоит значение $\tilde{n}_{ij} = \frac{n_{i.}n_{.j}}{N}$, где $n_{i.}$ — итог i-й строки; $n_{.j}$ — итог j-го столбца; N — объем совокупности.

Сравнивая исходную табл. 2 с полученной нами табл. 3, видим, что эти ситуации отличаются.

В качестве меры этого отличия рассматриваются разности между частотами исходной таблицы n_{ij} и частотами, вычисленными для случая независимости, $\tilde{n}_{ij}$. Эти разности возводятся в квадрат, делятся на $\tilde{n}_{ij}$, и полученные значения суммируются по всем клеткам таблицы. В результате получается величина, имеющая уже известное нам распределение X^2 (см. гл. 8, § 2).

1) На примере табл. 2 проведем проверку гипотезы о том, что степень выполнения плана не зависит от уровня образования рабочего.

2) В качестве статистической характеристики берется величина

$$X^2 = \sum_{i=1}^{l} \sum_{j=1}^{m} \frac{(n_{ij} - \tilde{n}_{ij})^2}{\tilde{n}_{ij}}. \tag{9.14}$$

Чем ближе исходные частоты к частотам для случая независимости признаков, тем ближе к нулю значение статистической характеристики. Значит, большие значения этой характеристики должны вызывать отклонение испытуемой гипотезы.

3) Найдем фактическое значение статистической характеристики:

$$X^2_{\text{ф}} = \frac{(5-12)^2}{12} + \frac{(25-18)^2}{18} + \frac{(15-16)^2}{16} + \frac{(25-24)^2}{24} + \frac{(20-12)^2}{12} + \frac{(10-18)^2}{18} \approx 15{,}8.$$

4) Выберем в качестве уровня значимости $\alpha = 0{,}1$.

5) По табл. 3 приложения надо найти $X^2_{кр}$. Заметим, что в данном случае число степеней свободы k подсчитывается по формуле $k=(l-1)(m-1)$.

Следовательно, в данном случае $k=(2-1)(3-1)=2$ и $X^2_{кр}=4{,}61$.

В данном случае проводится односторонняя проверка гипотезы, так как малые значения X^2 не противоречат гипотезе о независимости.

6) Сравнивая $X^2_ф$ и $X^2_{кр}$, видим, что $X^2_ф$ попадает в критическую область, и гипотеза о независимости признаков отвергается на уровне значимости 0,1. Что же касается степени связи признаков, то критерий не дает ответа на этот вопрос. Для измерения степени связи признаков существуют специальные коэффициенты (см. гл. 8).

Критерий дисперсионного анализа. Если хотя бы один из двух признаков является количественным, к решению вопроса о наличии или отсутствии связи между этими признаками можно применять дисперсионный анализ. Суть однофакторного дисперсионного анализа в следующем. Если по одному из признаков совокупность разбита на несколько групп (этот признак называется факторным), то в каждой из этих групп можно подсчитать среднее значение другого признака (он должен быть количественным и называется результативным). Заметим, что если наши признаки независимы, эти групповые средние не должны различаться между собой.

Пример 7. Вернемся к предыдущему примеру и для трех групп, образованных по качественному признаку (образование), подсчитаем процент выполнения плана (т. е. теперь этот признак рассматривается как количественный). Допустим, что на основе исходных данных по первой группе (начальное образование) получим 90% выполнения плана, по второй группе (неполное среднее образование) — 95%, по третьей группе (среднее и среднее специальное образование) — 110%. Видим, что между этими тремя групповыми средними существует расхождение.

Если мы полагаем, что признаки независимы, то это различие должно быть вызвано случайными причинами, а не влиянием признака, положенного в основу группировки (т. е. образования). Подсчитаем средний процент выполнения плана для всей совокупности из 100 рабочих. Обозначим эти общую среднюю $\bar{x}$, групповые средние — $\bar{x}_i$, а численности групп — n_i. Тогда $\bar{x}$ вычисляется по формуле

$$\bar{x}=\frac{\sum\limits_{i=1}^{k}\bar{x}_i n_i}{\sum\limits_{i=1}^{k} n_i}, \qquad (9.15)$$

где k — число групп.

В данном случае $\bar{x} = \frac{90 \cdot 30 + 95 \cdot 40 + 110 \cdot 30}{100} = 98\%$.

Оценим дисперсию признака «процент выполнения плана». Как известно, выборочной оценкой дисперсии σ^2 по выборке из n элементов является величина S^2. Напомним формулу для вычисления S^2:

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

Как известно, эту величину можно разложить на две составляющие: $S^2_{\text{факт}}$ и $S^2_{\text{ост}}$, где $S^2_{\text{ост}}$ учитывает отклонения исходных значений от соответствующих групповых средних; $S^2_{\text{факт}}$ — отклонения этих групповых средних от общей средней. Таким образом, $S^2_{\text{факт}}$ характеризует ту часть общей дисперсии, которая обусловлена влиянием **факторного признака**, т. е. признака, положенного в основу группировки, а $S^2_{\text{ост}}$ характеризует влияние остальных, неучтенных факторов. Поэтому $S^2_{\text{факт}}$ и $S^2_{\text{ост}}$ называются оценками **факторной** и **остаточной** дисперсий.

Очевидно, чем сильнее связь между факторным и результативным признаком, тем больше доля $S^2_{\text{факт}}$ в сумме $S^2_{\text{факт}} + S^2_{\text{ост}}$. Но если мы предполагаем, что результативный признак не зависит от факторного, то обе величины — и $S^2_{\text{факт}}$, характеризующая разброс групповых средних, и $S^2_{\text{ост}}$, характеризующая разброс относительно групповых средних,— не зависят от факторного признака и дают две оценки одной и той же величины — генеральной дисперсии σ^2. Следовательно, возникает уже знакомая задача о проверке равенства дисперсий (см. § 3), и, значит, проверка гипотезы о независимости признаков в данном случае использует дисперсионное отношение $S^2_{\text{факт}}/S^2_{\text{ост}}$, которое подчиняется F-распределению.

1) Таким образом, проверим гипотезу о том, что результативный признак (процент выполнения плана) не зависит от факторного признака (образование).

2) В качестве статистической характеристики берется дисперсионное отношение $S^2_{\text{факт}}/S^2_{\text{ост}}$, которое тем больше, чем сильнее связь между признаками. Значение $S^2_{\text{факт}}$ вычисляется по формуле

$$S^2_{\text{факт}} = \frac{\sum_{i=1}^{k} (\bar{x}_i - \bar{x})\, n_i}{k-1}, \qquad (9.16)$$

где $\bar{x}_i$ — групповые средние; n_i — численность групп; k — число групп; $\bar{x}$ — общее среднее, подсчитанное по формуле (9.15).

Значение $S^2_{ост}$ не вычисляют непосредственно, но получают как $\frac{S^2(n-1)-S^2_{факт}(k-1)}{n-k}$, где S^2 и $S^2_{факт}$ получаются соответственно по формулам (9.2) и (9.15).

3) Найдем фактическое значение статистической характеристики, для чего сначала вычислим $S^2_{факт}$ по формуле (9.16):

$$S^2_{факт}=\frac{(90-98)^2\cdot 30+(95-98)^2\cdot 40+(110-98)^2\cdot 30}{2}=3300.$$

Для подсчета S^2 надо знать индивидуальные значения результативного признака у всех объектов. Допустим, что они известны и на их основе подсчитана величина $S^2=400$.

Значит,

$$S^2_{ост}=\frac{400\cdot 99-3300\cdot 2}{97}\approx 340,\quad F_ф=\frac{S^2_{факт}}{S^2_{ост}}=\frac{3300}{340}\approx 9{,}7.$$

4) В качестве уровня значимости выберем $\alpha=0{,}01$.

5) По табл. 4А приложения найдем $F_{кр}$. Числа степеней свободы в данном случае равны $k_1=k-1$ и $k_2=n-k$, где k — число групп; n — объем выборки.

Значит, для $\alpha=0{,}01$, $k_1=3-1=2$ и $k_2=100-3=97$ получим $F_{кр}\approx 4{,}83$.

6) Так как $F_ф>F_{кр}$, то на уровне значимости 0,01 испытуемая гипотеза отклоняется.

Заметим, что, как и в предыдущем разделе, проведена односторонняя проверка гипотезы.

На вопрос о силе связи между признаками данный критерий ответа не дает. Однако если установлено наличие связи (гипотеза о независимости признаков отклонена), сила этой связи может быть измерена с помощью корреляционного отношения η (см. также § 2, гл. 6):

$$\eta=\sqrt{S^2_{факт}(k-1)/[S^2(n-1)]} \qquad (9.17)$$

При группировке по нескольким факторным признакам используется многофакторный дисперсионный анализ. Принцип его остается тем же, но расчеты сильно усложняются.

Ограничения при использовании распределений X^2 и F при проверке независимости. Необходимо отметить, что при

Таблица 4. Группировка хозяйств по плотности населения

x	n	y
0—10	3	40
10—20	5	46
20—30	5	42
30—40	5	50
40—50	2	55
$\bar{x}=24$	$n=20$	$\bar{y}=46$

использовании F-распределения в дисперсионном анализе следует учитывать те же ограничения, что при его использовании для проверки равенства дисперсий (см. § 3), т. е. значения признака должны быть независимы и распределены в генеральной совокупности нормально. Кроме того, в дисперсионном анализе предполагается равенство групповых дисперсий (влияние прочих факторов, кроме группировочного, во всех группах должно быть примерно одинаковым).

Наконец, подчеркнем, что результаты применения критериев X^2 и F могут существенно зависеть от разбивки совокупности на группы. Поэтому желательно иметь достаточно много групп и достаточное число объектов в каждой из групп (не менее пяти). Желательно также, чтобы численности групп не слишком сильно различались между собой.

Критерий значимости для коэффициента корреляции. Если изучаемые признаки являются количественными, для изучения связи между ними можно применять корреляционный и регрессионный анализ (см. гл. 6). Задача, которая ставится в этом разделе, аналогична предыдущим задачам, т. е. проверяется гипотеза об отсутствии связи. Но в отличие от предыдущих случаев здесь проверяется не наличие или отсутствие связи вообще, а наличие или отсутствие связи определенной формы. Наиболее часто проверяется гипотеза о линейной связи между признаками, которую выражает линейный коэффициент корреляции r. Таким образом, если для выборочных данных получено некоторое значение r, предполагается, что его значение лишь случайно не равно 0, тогда как в генеральной совокупности линейный коэффициент корреляции (ρ) равен нулю и линейной связи между признаками нет (см. также § 2, гл. 6). Следовательно, высокие коэффициенты корреляции должны способствовать отклонению гипотезы об отсутствии линейной связи.

Замечание. Необходимо помнить, что отсутствие линейной связи не означает отсутствия связи вообще, т. е. независимости признаков. Между ними может существовать весьма сильная, хотя и нелинейная по форме связь.

Пример 8. Рассмотрим два признака: средняя урожайность ржи y (в пудах на дес.) и плотность населения x. Количество хозяйств обозначим n. Данные приводятся по центру России, конец XIX в. (табл. 4).

Вычислим по формуле выборочный коэффициент корреляции r для этих 20 хозяйств:

$$r=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\left[\sum_{i=1}^{n}(x_i-\bar{x})^2\right]\left[\sum_{i=1}^{n}(y_i-\bar{y})^2\right]}}, \tag{9.18}$$

где $\bar{x}$ и $\bar{y}$ — средние значения признаков. Не приводя подробно вычислений, дадим полученное из табл. 4 значение $r=0{,}88$.

1) Проверим гипотезу о некоррелированности этих признаков, т. е. гипотезу об отсутствии линейной связи между плотностью населения и урожайностью в генеральной совокупности.

2) Статистической характеристикой служит не сам коэффициент корреляции r, а величина r, деленная на ошибку коэффициента корреляции σ_r, равную $\sqrt{(1-r^2)/(n-2)}$.

Таким образом, статистической характеристикой является величина

$$t=r\sqrt{n-2}/\sqrt{1-r^2}, \tag{9.19}$$

где n — объем выборки.

Величина t распределена по уже известному нам закону Стьюдента.

3) Подсчитаем фактическое значение характеристики:

$$t_{ф}=\frac{0{,}88\sqrt{18}}{\sqrt{1-(0{,}88)^2}}\approx 7{,}8.$$

4) Выберем $\alpha=0{,}01$.

5) Пользуясь табл. 2 приложения найдем для $\alpha=0{,}01$ и числа степеней свободы k, равного для данного критерия n—2, $t_{кр}=2{,}88$.

6) Поскольку $t_{кр}<t_{ф}$, значение характеристики попадает в критическую область и гипотеза об отсутствии линейной связи между признаками в генеральной совокупности

отклоняется, т. е. коэффициент 0,88 является существенно значимым [1].

Оценка коэффициента корреляции в генеральной совокупности. Часто в корреляционном анализе возникает задача оценить, насколько существенным является отклонение выборочного коэффициента корреляции r от предполагаемого генерального значения ρ (случай $\rho=0$ рассмотрен в предыдущем разделе), либо оценить, насколько существенно различие коэффициентов корреляции для двух выборок. Для решения этих задач используется так называемое преобразование Фишера, которое каждому значению r ставит в соответствие величину

$$z=\frac{1}{2}\ln\frac{1+r}{1-r}, \tag{9.20}$$

распределенную приближенно по нормальному закону со средним значением

$$\bar{z}=\frac{1}{2}\ln\frac{1+\rho}{1-\rho}+\frac{\rho}{2(n-1)} \tag{9.21}$$

и дисперсией

$$\sigma_z^2=1/(n-3), \tag{9.22}$$

где ρ — генеральное значение коэффициента корреляции; n — объем выборки.

Поскольку величина z распределена приближенно нормально, то ее нормированное отклонение (как и в случае со средним арифметическим) может служить мерой различия между выборочным коэффициентом корреляции r и предполагаемым генеральным значением ρ.

---

[1] Заметим, что значимость величины r^2 для линейной связи, а также величины η^2 [см. (9.17)] для связи общего вида можно проверять и непосредственно, не обращаясь к вычислению значений t и F, поскольку существуют построенные на основе этих критериев специальные таблицы критических значений r^2 и η^2 для различных уровней значимости. Таблица $r^2_{кр}$ и $\eta^2_{кр}$ для $\alpha=0{,}01$ приведена в приложении (табл. 10). Так, по этой таблице $r^2_{кр}$ ($\alpha=0{,}01$) находится на пересечении столбца с номером $k_1=k-1$ (где k — число параметров уравнения регрессии, т. е. $k=2$) и строки с номером $k_2=n-2$. В нашем примере 9 $r^2_{кр}=0{,}315$ и $r^2_{ф}=0{,}83^2\approx0{,}68>r^2_{кр}$, т. е. коэффициент корреляции является значимым. Для проверки значимости корреляционного отношения $\eta^2_{кр}$ ($\alpha=0{,}01$) находится на пересечении столбца с номером $k_1=k-1$ (где k — число групп) и строки с номером $k_2=n-k$. В примере 8 $\eta^2_{кр}\approx0{,}088$, а $\eta_{ф}^2$ [по формуле (9,17)] равно $\frac{3300\cdot(3-1)}{400\cdot99}\approx$ $\approx0{,}17$, т. е. является значимым.

Пример 9. Проверим гипотезу о том, что коэффициент корреляции $r=0{,}88$, полученный в примере из предыдущего раздела, лишь случайно получился отличным от значения $\rho=0{,}7$. (Это значение, допустим, получено в результате изучения аналогичной генеральной совокупности хозяйств, но по другому региону.)

1) Таким образом, проверим гипотезу о том, что и в данной генеральной совокупности генеральное значение коэффициента корреляции ρ равно 0,7: $H_0 : \rho=0{,}7$.

2) Пользуясь нормальностью величины z, полученной по формуле (9.20), возьмем в качестве статистической характеристики нормированное отклонение

$$(z-\bar{z})/\sigma_z, \tag{9.23}$$

позволяющее пользоваться табл. 1 приложения.

3) Найдем фактическое значение статистической характеристики. Для этого воспользуемся специальной таблицей (табл. 6 приложения), в которой для каждого значения коэффициента корреляции вычислено значение функции z по формуле (9.20). Из этой таблицы найдем: для $r=0{,}88$, $z=1{,}3758$; для $\rho=0{,}7$ $z=0{,}8673$.

Теперь по формуле (9.21) найдем $\bar{z}$:

$$\bar{z}=0{,}8673-\frac{0{,}7}{2\cdot 19}\approx 0{,}8489,$$

по формуле (9.22) найдем σ_z:

$$\sigma_z=1/\sqrt{17}\approx 0{,}2427.$$

Подставляя полученные величины в формулу (9.23), получим

$$\frac{z-\bar{z}}{\sigma_z}=\frac{1{,}3758-0{,}8489}{0{,}2427}\approx 2{,}17.$$

4) Выберем в качестве уровня значимости $\alpha=0{,}05$.

5) По табл. 1 приложения получим критическое значение характеристики, равное 1,96.

6) Поскольку фактическое значение статистической характеристики попадает в критическую область, испытуемая гипотеза отклоняется, и мы не можем считать, что в нашей генеральной совокупности значение коэффициента корреляции равно 0,7.

Сравнение двух выборочных коэффициентов корреляции.

Пример 10. Рассмотрим случай, когда имеются две выборки, коэффициенты корреляции в которых равны r_1 и r_2, и требуется выяснить, существенно различие между ними или оно обусловлено случайными причинами, и выборки получены из двух генеральных совокупностей с одинаковыми коэффициентами корреляции.

Вернемся к предыдущему примеру и вспомним, что $r_1=0{,}88$ — коэффициент корреляции в выборке 20 хозяйств центра России между урожайностью ржи и плотностью населения. Допустим, что имеется еще одна выборка, по югу России, состоящая из 10 хозяйств и дающая $r_2=0{,}8$. Нам хотелось бы выяснить, является ли различие r_1 и r_2 случайным при равенстве генеральных коэффициентов ρ_1 и ρ_2 по Центру и Югу России.

1) Таким образом, проверяется гипотеза об одинаковой силе линейной связи урожайности ржи с плотностью населения в двух регионах:

$$H_0: \rho_1=\rho_2.$$

2) Аналогично формуле (9.23) введем формулу нормированного отклонения двух коэффициентов корреляции:

$$(z_1-z_2)/(\sigma_{z_1-z_2}), \tag{9.24}$$

где z_1 и z_2 получены из r_1 и r_2 по формуле (9.20), а $\sigma_{z_1-z_2}$ — ошибка разности коэффициентов корреляции, имеющая вид:

$$\sigma_{z_1-z_2}=\sqrt{\frac{1}{n_1-3}+\frac{1}{n_2-3}}\,. \tag{9.25}$$

Если гипотеза H_0 верна, нормированное отклонение (9.24) имеет нормальное распределение, позволяющее использовать табл. 1 приложения, и является статистической характеристикой. Чем больше абсолютное значение этой характеристики, тем больше оснований отклонить испытуемую гипотезу.

3) Чтобы подсчитать фактическое значение статистической характеристики, найдем по табл. 6 приложения z_1 и z_2, соответствующие r_1 и r_2: для $r_1=0{,}88$ $z_1=1{,}3758$; для $r_2=0{,}80$ $z_2=1{,}0986$. По формуле (9.25) найдем величину $\sigma_{z_1-z_2}$:

$$\sigma_{z_1-z_2}=\sqrt{1/17+1/7}\approx 0{,}448.$$

Подставляя полученные значения в (9.24), получим

$$\frac{z_1-z_2}{\sigma_{z_1-z_2}}=\frac{1{,}3758-1{,}0986}{0{,}448}\approx 0{,}62.$$

4) Выберем $\alpha=0{,}05$.

5) По табл. 1 приложения для $\alpha=0{,}05$ получим критическое значение характеристики, равное 1,96.

6) Сравнивая фактическое и критическое значения, видим, что данные выборок не дают оснований отклонить испытуемую гипотезу.

Ограничения при проверке гипотез в рамках корреляционного анализа. При решении задач проверки статистических гипотез надо учитывать требования, которые накладывает на данные корреляционный анализ:

1) как обычно, предполагается нормальное распределение признаков в генеральной совокупности и независимость отдельных значений признаков;

2) кроме того, предполагается, что дисперсия одного признака при всех значениях другого постоянна, т. е. связь между признаками влияет на их средние значения, но не на дисперсии;

3) наконец, использование линейного коэффициента корреляции требует действительного наличия линейной связи между признаками в генеральной совокупности (иначе он не имеет практического значения).

Условие 3) можно проверить с помощью специального критерия (см. § 3, гл. 6), но необходимо и хорошее знание изучаемого материала. Условие 2) аналогично условию равенства групповых дисперсий для дисперсионного анализа и может быть проверено теми же средствами. Нормальность распределения проверяется с помощью критериев, описанных в следующих разделах.

Замечание. Критерии, рассмотренные здесь, связаны с линейной зависимостью. Существуют критерии для проверки гипотез для зависимости одного признака от нескольких (множественная, частная корреляция), связанные с коэффициентом детерминации, которые упоминаются в гл. 6, а также критерии для коэффициентов ранговой корреляции (см. § 2, гл. 8).

§ 5. Критерии согласия

Одним из наиболее важных разделов теории статистической проверки гипотез является проверка гипотез о законах распределения изучаемых признаков в генеральных совокупностях по выборочным данным. Это означает, что вариационный ряд (см. гл. 4), соответствующий распределению признака в выборке, отражает некий закон распределения этого признака, справедливый для всей генеральной совокупности.

Часто знание изучаемого материала или графическое изображение вариационного ряда позволяет предположить, что это неизвестное распределение является вполне определенным, например нормальным. Это предполагаемое распределение называется **теоретическим**, тогда как выборочное распределение называется **эмпирическим.** Естественно,

Таблица 5. Эмпирическое распределение

Расход на культурные нужды, руб.	Число хозяйств, % от общего числа
До 2	3
2—3	16
3—4	30
4—5	33
5—6	14
Свыше 6	4
	Итого 100

Таблица 6. Теоретическое распределение

Расход, руб.	$t_i=$ $=(x_i-\bar{x})/\sigma_x$	$F(t_i)$	$\tilde{f}_i$
0—2	—2	0,023	0,023
2—3	—1	0,159	0,136
3—4	0	0,5	0,341
4—5	1	0,841	0,341
5—6	2	0,977	0,136
Свыше 6	—	1	0,023

возможно некоторое расхождение между теоретическим и эмпирическим распределениями. Для того чтобы оценить, связаны эти расхождения со случайностями выборки или же с неверным подбором теоретического закона распределения, и предназначены критерии согласия.

Критерий согласия X^2 между эмпирическим и теоретическим распределениями. Уже известное нам распределение X^2 может служить для проверки гипотезы о соответствии между теоретическим и эмпирическим распределениями.

Пример 11. Допустим, что у нас имеются данные о расходах на культурные нужды (на канцелярские принадлежности, книги, журналы, лекарства) по 200 крестьянским бюджетам 20-х годов XX в. (Екатеринославская губерния), т. е. имеется вариационный ряд, дающий эмпирическое распределение признака (табл. 5).

Мы хотим проверить, согласуются ли эти данные с гипотезой о нормальном распределении признака в генеральной совокупности. Чтобы сравнить имеющиеся данные с теоретическими, требуется определить те частности, которые будут соответствовать каждому интервалу изменения признака в случае нормального распределения. Для этого необходимо найти среднее значение и дисперсию признака по исходным данным. Пользуясь исходными (несгруппированными) данными, получили значения $\bar{x}=4$ руб. и $\sigma_x=1$ руб. (значения $\bar{x}$ и σ_x округлены для удобства вычислений).

Теперь для каждого интервала изменения признака получим значение нормированного отклонения t_i:

$$t_i=(x_i-\bar{x})/\sigma_x, \tag{9.26}$$

где x_i — верхняя граница интервала; i — номер интервала. Приведем их значения в табл. 6.

Третий столбец этой таблицы получается с помощью табл. 7 приложения. В этой таблице для каждого значения t даны вероятности $F(t)$ нормированных отклонений, не превышающих t. Значения $F(t)$ для отрицательных t получаются по формуле $1-F(|t|)$. Так, для $t=-2$ надо взять величину $1-F(2)=1-0{,}977=0{,}023$, для $t=-1$ — величину $1-F(1)=1-0{,}841=0{,}159$, для $t=0$ — величину $F(0)=0{,}5$, для $t=1$ — величину $F(1)=0{,}841$ и т. д. Для последнего интервала берем $F=1$.

Таблица 7. Сопоставление теоретического и эмпирического распределений

Расход, руб.	Частости, %	
	эмпирич.	теорет.
До 2	3	2,30
2—3	16	13,6
3—4	30	34,1
4—5	33	34,1
5—6	14	15,6
Свыше 6	4	2,3
	100	100

Таким образом, мы получили вероятности того, что признак x (расход) будет иметь значения, не превышающие соответственно 2, 3, 4 руб. и т. д. Чтобы получить вероятность попадания этого признака, например, на интервал 2—3, надо из значения вероятности F для верхней границы этого интервала вычесть значение F для его нижней границы, получая $\tilde{f}=0{,}159-0{,}023=0{,}136$. Аналогичным образом получим значения вероятности попадания в каждый интервал (теоретические частости). Эти значения записаны в 4-м столбце табл. 6.

Для того чтобы обеспечить сравнимость с эмпирическими данными, выразим теоретические частости $\tilde{f}$ в процентах, умножая их на 100.

Сведем теоретические и эмпирические значения частостей в одну таблицу (см. табл. 7).

Видно, что между эмпирическими и теоретическими значениями существует расхождение, однако надо измерить степень этого расхождения. Вспомним, что в § 4 нам тоже пришлось сравнивать реальные частоты с ожидаемыми в случае независимости. Прием, использованный в § 4, можно применить и здесь, подсчитав величину

$$X^2=\frac{N}{100}\sum_{i=1}^{n}\frac{(f_i-\tilde{f}_i)^2}{\tilde{f}_i}, \qquad (9.27)$$

где f_i — эмпирические частости; $\tilde{f}_i$ — теоретические частости; n — число интервалов изменения признака.

Замечание. Если использовать частости не в процентах, а в долях, X^2 надо считать по формуле

$$X^2 = N \sum_{i=1}^{n} \frac{(f_i - \tilde{f}_i)^2}{\tilde{f}_i},$$

а если использовать вместо частостей частоты, то

$$X^2 = \sum_{i=1}^{n} \frac{(n_i - \tilde{n}_i)^2}{\tilde{n}_i}.$$

Очевидно, что чем лучше соответствие между теоретическим и эмпирическим распределениями, тем ближе X^2 к нулю, значит, большие значения X^2 должны способствовать отклонению гипотезы о нормальности распределения признака в генеральной совокупности.

1) Таким образом, проверяется гипотеза о том, что выборка получена из генеральной совокупности, в которой распределение хозяйств по расходам на культурные нужды нормально:

$$H_0: f_i = \tilde{f}_i.$$

2) Статистической характеристикой является величина, вычисляемая по формуле (9.27).

3) Вычислим фактическое значение характеристики:

$$X^2_{ф} = 2\left[\frac{(0,7)^2}{2,3} + \frac{(2,4)^2}{13,6} + \frac{(4,1)^2}{34,1} + \frac{(1,1)^2}{34,1} + \frac{(0,4)^2}{13,6} + \frac{(1,7)^2}{2,3}\right] \approx$$

$$\approx 2(1,47 + 0,44 + 0,53) \approx 4,88.$$

4) Выберем уровень значимости $\alpha = 0,1$.

5) По табл. 3 приложения найдем критическое значение X^2. Число степеней свободы k в данном случае считаем по формуле $k = n - l - 1$, где n — число интервалов, т. е. 6, а l — число параметров нормального распределения, вычисляемых с помощью выборочных данных. Поскольку с помощью выборочных данных мы определяли $\bar{x}$ и σ_x, $l = 2$. Следовательно, $k = 6 - 2 - 1 = 3$, $X^2_{кр}$ для $k = 3$ и $\alpha = 0,1$ равно 6,25.

6) Поскольку $X_{ф}^2$ меньше критического значения, гипотеза о нормальности распределения не отклоняется.

Сравнение двух эмпирических распределений.

Пример 12. Допустим, что кроме данных о распределении расходов на культурные нужды по 200 бюджетам Екатеринославской губернии имеются аналогичные данные по 300 бюджетам Саратовской губернии. Требуется проверить предположение о том, что в этих двух губерниях (генеральных совокупностях) распределение доходов одинаково. При этом неважно, какой вид имеет распределение, оно нас не интересует само по себе, важно лишь установить согласие распределений в двух совокупностях. Пусть данные представлены в виде табл. 8.

Таблица 8. Сравнение двух эмпирических распределений

Расход на культурные нужды, руб.	Частоты, %		Теоретические частости, %
	Екатерин. губ.	Саратов. губ.	
До 2	3	5	4,2
2—3	16	14	14,8
3—4	30	28	28,8
4—5	33	32	32,4
5—6	14	17	15,8
Свыше 6	4	4	4
	100	100	100

Для решения поставленной задачи на основе этих двух эмпирических распределений надо построить теоретическое, находя для каждого интервала взвешенное среднее двух эмпирических частостей (частот). В 4-м столбце табл. 8 записаны значения этих теоретических частостей. (Значение 4,2, например, получается как $100\%\left[\frac{0,03\cdot200+0,5\cdot300}{500}\right]$ или $\frac{3\cdot2+5\cdot3}{5}$.)

Согласие эмпирических распределений между собой отражается в величине $X^2=X_1^2+X_2^2$, где X_1^2 и X_2^2 получены по формуле (9.27) для обоих эмпирических распределений. Чем лучше согласие между распределениями, тем ближе X^2 к нулю, следовательно, большие значения X^2 ведут к отклонению испытуемой гипотезы.

1) Таким образом, проверим гипотезу о совпадении распределения крестьянских хозяйств по расходам на культурные нужды в двух губерниях, т. е. $H_0: f_{1i}=f_{2i}$.

2) Статистической характеристикой служит величина X^2:

$$X^2=X_1^2+X_2^2=\frac{N_1}{100}\sum_{i=1}^{n}\frac{(f_{1i}-\tilde{f}_i)^2}{\tilde{f}_i}+\frac{N_2}{100}\sum_{i=1}^{n}\frac{(f_{2i}-\tilde{f}_i)^2}{\tilde{f}_i}. \tag{9.28}$$

где f_{1i} — частости первого эмпирического распределения; f_{2i} — частости второго эмпирического распределения, $\tilde{f}_i$ — теоретические частости, полученные по формуле взвешенного среднего; n — число интервалов изменения признака; N_1, N_2 — объемы выборок.

3) Найдем $X_ф^2$ в нашем примере:

$$X_ф^2=2\left[\frac{(1,2)^2}{4,2}+\frac{(1,2)^2}{14,8}+\frac{(1,2)^2}{28,8}+\frac{(0,6)^2}{32,4}+\frac{(1,8)^2}{15,8}\right]+$$

$$+3\left[\frac{(0,8)^2}{4,2}+\frac{(0,8)^2}{14,8}+\frac{(0,8)^2}{28,8}+\frac{(0,4)^2}{32,4}+\frac{(1,2)^2}{15,8}\right]\approx 2.34.$$

4) Выберем, как и раньше, $\alpha=0,1$.

5) По табл. 3 приложения найдем $X^2_{кр}$, но сначала определим число степеней свободы k. Известно, что для критериев согласия $k=n-l-1$ (см. предыдущий раздел). Поскольку теоретическое распределение само по себе нас не интересует, никакие его параметры и не вычисляются, т. е. $l=0$ и $k=n-1=5$. Для $\alpha=0,1$ и $k=5$ $X^2_{кр}=9,24$.

6) Сравнивая $X^2_ф$ и $X^2_{кр}$, видим, что данные выборок не противоречат испытуемой гипотезе и она не отклоняется.

Напомним, что критерий X^2 желательно применять, если численности групп достаточно велики (практически в каждой группе должно быть не менее 5 объектов). Если это условие нарушено, группы, в которых менее 5 объектов (обычно это крайние группы), объединяют с соседними.

Наконец, результат применения критерия X^2 существенно зависит от того, каким образом область изменения признака разбита на интервалы, и при изменении границ интервалов значение X^2 меняется, что снижает надежность применения этого критерия.

Отметим, что рассмотренные здесь критерии связаны с односторонней проверкой, т. е. испытуемой гипотезе противоречат слишком большие значения их статистических характеристик. Однако если будут получены значения характеристик, близкие к 0, это будет говорить о «слишком хорошем» согласии с испытуемой гипотезой и может вызвать подозрение в том, что выборка не является случайной, а «подобрана» так, чтобы наилучшим образом соответствовать испытуемой гипотезе. В таких случаях желательно проводить двустороннюю проверку гипотезы.

Критерии проверки нормальности распределения с помощью коэффициентов асимметрии и эксцесса. При использовании многих основных критериев в теории статистической проверки гипотез существенной является нормальность распределения изучаемого признака в генеральной совокупности. Как следует из приведенных примеров § 5, проверку нормальности можно осуществлять с помощью критериев согласия. Укажем еще один способ проверки нормальности, который позволяет определить степень со-

Таблица 9. Данные для расчета коэффициента асимметрии

Расход на культурные нужды, руб.	Частости, %	Середины интервалов	$x_i-\bar{x}$	$(x_i-\bar{x})^3$	$(x_i-\bar{x})^3 f_i$
До 2	3	1	−3	−27	−81
2—3	16	2,5	−1,5	−3,375	−54
3—4	30	3,5	−0,5	−0,125	−3,75
4—5	33	4,5	0,5	0,125	4,125
5—6	14	5,5	1,5	3,375	41,25
Свыше 6	4	6,5	2,5	15,625	62,5
	100				−24,875

гласия между эмпирическим и нормальным теоретическим распределением. Это способ измерения коэффициентов асимметрии и эксцесса.

Коэффициент асимметрии A характеризует скошенность распределения в сторону больших или меньших значений признака. Этот коэффициент рассчитывается по формуле

$$A=\frac{\sum_{i=1}^{k}(x_i-\bar{x})^3 f_i}{\left(\sum_{i=1}^{k} f_i\right)\sigma^3}, \qquad (9.29)$$

где f_i — частота или частость i-го интервала изменения признака; x_i — середина интервала; $\bar{x}$ — среднее арифметическое значение; σ — среднее квадратическое отклонение признака; k — число интервалов.

Для нормального распределения $A=0$ (случай симметрии).

Напомним, что значения $\bar{x}$ и σ_x (оцененные по исходным данным) известны и равны соответственно 4 руб. и 1 руб. (Известно также и значение середины последнего, незакрытого интервала.)

$$A=\frac{-24{,}875}{100\cdot 1}\approx -0{,}25.$$

Таким образом, эмпирическое распределение имеет левостороннюю скошенность (см. табл. 9).

Таблица 10. Данные для расчета коэффициента эксцесса

Расход, руб.	Частости, %	$x_i-\bar{x}$	$(x_i-\bar{x})^4$	$(x_i-\bar{x})^4 f_i$
До 2	3	—3	81	243
2—3	16	—1,5	5,0625	81
3—4	30	—0,5	0,0625	1,875
4—5	33	0,5	0,0625	2,0625
5—6	14	1,5	5,0625	70,875
Свыше 6	4	2,5	39,0625	156,25
	100			555,0625

Коэффициент эксцесса E характеризует степень островершинности распределения и рассчитывается по формуле

$$E=\frac{\sum_{i=1}^{k}(x_i-\bar{x})^4 f_i}{\sigma^4\sum_{i=1}^{k} f_i}, \qquad (9.30)$$

где f_i — частота или частость интервала изменения признака; x_i — середина интервала; $\bar{x}$, σ — среднее арифметическое значение и среднее квадратическое отклонение признака; k — число интервалов.

Для нормального распределения $E=3$.

Значением E будет величина $\frac{555,0625}{100\cdot 1}=5,55$. Поскольку $E>3$, это распределение характеризуется островершинностью (см. табл. 10).

Излагаемый метод проверки нормальности связан с сопоставлением полученных в выборке фактических значений показателей асимметрии и эксцесса со значениями, соответствующими нормальному распределению. Чем дальше значение A от нуля, а значение E от 3, тем хуже согласие с нормальным распределением.

Пример 13. 1) Проверим гипотезу о нормальности распределения признака в примере 11 (табл. 5), т. е. предположим, что в генеральной совокупности $A=0$ и $E=3$ и что отличие их фактических значений от теоретических вызвано случайными причинами ($H_0: A=0,\ E=3$).

2) Статистическими характеристиками являются сами коэффициенты A и E, поскольку их распределения известны и табулированы.

3) Вычислим $A_ф=-0,25$ и $E_ф=5,55$.

4) Выберем $\alpha=0,05$.

5) По табл. 8 приложения находим критические значения A и E для $\alpha=0{,}05$ и 0,01. Эти значения зависят от объема выборки N. Для $N=200$ и $\alpha=0{,}05$ $E_{кр}=3{,}57$, $A_{кр}=0{,}28$.

6) Поскольку $E_{ф}>E_{кр}$, гипотеза о нормальности отклоняется.

Упомянем **приближенный** способ оценки близости распределения к нормальному с помощью коэффициентов A и E. При этом способе фактические значения A и E сопоставляются с величинами их средних ошибок в выборке σ_A и σ_E.

Эти ошибки рассчитываются по формулам:

$$\sigma_A=\sqrt{\frac{6(N-1)}{(N+1)(N+3)}}, \tag{9.31}$$

$$\sigma_E=\sqrt{\frac{24N(N-2)(N-3)}{(N-1)^2(N+3)(N+5)}}. \tag{9.32}$$

Для $N=200$ эти значения равны соответственно 0,17 и 0,34.

Значения $|A|$ и $|E-3|$ сравниваются со своими утроенными ошибками, и если окажется, что они превышают утроенные ошибки, гипотеза о нормальности отклоняется.

Пример 14. Возьмем выборку из совокупности бюджетных обследований крестьянских хозяйств Екатеринославской губернии. В эту выборку попали 94 хозяйства, относящиеся к «типичным середняцким». По данным выборки подсчитаны значения $A_ф$ и $E_ф$ и их средние ошибки для признака «расходы на культурные нужды»: $A=2{,}95$, $\sigma_A=0{,}246$; $E=9{,}71$, $\sigma_E=0{,}475$.

Видно, что A и $E-3$ намного больше обеих утроенных ошибок и распределение нельзя считать нормальным. В этой ситуации возникает предположение, что причиной такого сильного расхождения между эмпирическим и теоретическим нормальным распределениями явилось наличие в этой выборке лишь 40% хозяйств, обладающих данным признаком. Была выделена подгруппа хозяйств с наличием данного признака, объем этой выборки равен 25. Для нее были получены значения A, E, σ_A и σ_E, дающие удовлетворительное согласие с нормальным распределением: $A=1{,}13$, $\sigma_A=0{,}445$; $E=1{,}06$, $\sigma_E=0{,}782$. Последний пример демонстрирует, что хорошее знание материала исследования позволяет правильно оценить вид распределения в генеральной совокупности.

Критерий Вилкоксона для малых выборок. Допустим, что имеются две выборки и требуется проверить их однородность, т. е. проверить, что эти два эмпирических распределения соответствуют одному и тому же теоретическому распределению. Эта задача аналогична поставленной ранее. Но в том случае, когда объемы выборок малы, нельзя применять рассмотренные критерии согласия, так как их результаты становятся ненадежными.

Для таких случаев предлагается следующая процедура: данные обеих выборок (обозначим их $x_1, x_2, ..., x_{N_1}$ и $y_1, y_2, ..., y_{N_2}$) располагают в виде одной упорядоченной по возрастанию последовательности, например, в виде

$$x_1 y_1 y_2 x_2 y_3 x_3 x_4 x_5 y_4, \dots \text{ и т. д.}$$

Если совокупности однородны, члены одной выборки должны равномерно располагаться среди членов другой выборки, не скапливаться большими группами. Поэтому для каждого значения x_i подсчитывается количество значений y_i, предшествующих ему, и эти величины суммируются, давая величину U. Оказывается, при N_1 и $N_2 > 10$ величина U распределена приближенно нормально со средним значением $\bar{U} = \frac{N_1 N_2}{2}$ и средним квадратическим отклонением $\sigma_u = \sqrt{N_1 N_2 (N_1 + N_2 + 1)/12}$, что позволяет использовать в качестве статистической характеристики нормированное отклонение:

$$t = (U - \bar{U})/\sigma_u \tag{9.33}$$

и обращаться к табл. 1 приложения.

Таким образом, чем сильнее величина U будет отклоняться от среднего значения, тем больше неравномерность распределения элементов одной выборки среди элементов другой и тем больше оснований отклонить гипотезу об однородности двух генеральных распределений.

Пример 15. Пусть имеются две выборки по признаку «величина надела» (табл. 11).

1) Мы хотим проверить гипотезу об однородности распределений крестьянских хозяйств по размеру надела в обеих губерниях.

2) Статистический характеристикой является величина t из формулы (9.33).

3) Чтобы получить значение статистической характеристики, составим из элементов обеих выборок одну последовательность, упорядоченную по возрастанию. Она будет иметь вид

0,6	0,7	0,8	0,9	1,0	1,3	1,4	1,5	1,6	1,8	2,2
x_1	x_2	y_1	y_2	y_3	x_3	x_4	y_4	x_5	y_5	x_6
2,3	2,5	2,8	3,0	3,2	3,3	3,5	3,7	4,1	4,2	4,5
y_6	x_7	y_7	x_8	y_8	y_9	x_9	y_{10}	x_{10}	y_{11}	y_{12}

Подсчитаем фактические значения величин, входящих в формулу (9.33). $U_ф$ получается следующим образом: сначала подсчитываем, сколько значений y предшествует x_1, это число равно 0; для x_2 это число также равно 0, а для x_3 оно равно 3, поскольку перед x_3 стоят

Таблица 11. Величина надела крестьянских хозяйств в двух губерниях

	1	2	3	4	5	6	7	8	9	10	11	12
Первая губерния (x_i)	0,6	0,7	1,3	1,4	1,6	2,2	2,5	3,0	3,5	4,1		
Вторая губерния (y_i)	0,8	0,9	1,0	1,5	1,8	2,3	2,8	3,2	3,3	3,7	4,2	4,5

три значения: y_1, y_2 и y_3, для x_4 снова получим 3 и т. д. Таким образом, $U_ф=0+0+3+3+4+5+6+7+9+10=47$.

Затем подсчитаем $U=\frac{N_1N_2}{2}=60$,

$$\sigma_u=\sqrt{\frac{N_1N_2(N_1+N_2+1)}{12}}\approx 15{,}2.$$

Следовательно,

$$t_ф=\frac{47-60}{15{,}2}\approx -0{,}86.$$

4) Выберем в качестве уровня значимости $\alpha=0{,}05$.

5) По табл. 1 приложения для $\alpha=0{,}05$ найдем $t_{кр}=1{,}96$.

6) Сравнивая $|t_ф|$ и $t_{кр}$, видим, что $t_ф$ по абсолютной величине не превосходит $t_{кр}$ и, следовательно, лежит в области допустимых значений, что не дает оснований отвергнуть гипотезу об одинаковом для обеих губерний законе распределения хозяйств по величине надела.

§ 6. Критерии случайности

Для применения статистических критериев и правильной интерпретации полученных выводов необходимо, чтобы выборка была случайной. Часто историку приходится иметь дело с естественными выборками, условия формирования которых ему неизвестны, т. е. неизвестно, по каким причинам сохранились и дошли до него именно эти, а не другие данные. В такой ситуации крайне важно проверить, случайна ли имеющаяся выборка. Для такой проверки существуют специальные критерии.

Критерий серий. Этот критерий служит для проверки случайности последовательности расположения элементов в выборке. Для этого определяется среднее значение признака $\tilde{x}$ (или медиана), а затем числа, большие $\tilde{x}$, обозначают единицей, а меньшие x — нулем. Эти значения некоторым образом чередуются. Каждая группа идущих подряд нулей или единиц называется **серией**. Слишком большое или слишком малое число серий заставляет предполагать, что выборка составлена по некоторому закону и не является случайной. Известно, что количество серий R при объеме выборки $n>10$ распределяется приближенно нормально со средним значением $\overline{R}=(n+1)/2$ и средним квадратическим отклонением $\sigma_R=\frac{1}{2}\sqrt{n-1}$.

Таким образом, сильное отклонение числа серий от среднего значения позволяет отклонить гипотезу о случайности выборки.

Пример 16. Рассмотрим выборку 10 крестьянских хозяйств по признаку «количество лошадей»: 0, 3, 2, 1, 0, 5, 1, 2, 8, 0.

1) Испытуемой является гипотеза H_0 о том, что эта выборка является случайной.

2) Статистической характеристикой t является нормированное отклонение числа серий R от среднего значения:

$$t=\frac{R-\overline{R}}{\sigma_R}. \tag{9.34}$$

Найдем значения R и σ_R:

$$\overline{R}=\frac{n+1}{2}=5{,}5, \quad \sigma_R=\frac{1}{2}\sqrt{n-1}=1{,}5.$$

Формула (9.34) приобретает вид

$$t=\frac{R-5{,}5}{1{,}5}.$$

3) Чтобы подсчитать $R_ф$, найдем $\quad \overline{x}=\frac{\sum_{i=1}^{n} x_i}{n}=2{,}2.$

Значения x_i меньшие или равные 2, обозначены нулем, а значения, большие 2, единицей:

0, 1, 0, 0, 0, 1, 0, 0, 1, 0.

В этом ряду 7 серий, следовательно, $R_ф=7$ и

$$t_ф=\frac{7-5{,}5}{1{,}5}=1.$$

4) Выберем $\alpha=0{,}1$.

5) В силу нормальности распределения R по табл. 1 приложения для $\alpha=0{,}1$ найдем $t_{кр}=1{,}65$.

6) Сравнивая $t_ф$ и $t_{кр}$, видим, что данные выборки не дают оснований отклонить гипотезу о случайном характере этой выборки, так как $t_ф<t_{кр}$.

Замечание 1. Этот критерий можно применять для проверки случайности выборки по качественному альтернативному признаку (наличие признака обозначается 1, а отсутствие — 0).

Замечание 2. Критерий серий наряду с критерием Вилкоксона можно использовать для проверки однородности распределений по двум малым выборкам, для чего элементы обеих выборок образуют упорядоченную последовательность, и элементы одной выборки обозначают нулями, а элементы другой — единицами.

Метод последовательных разностей. Этот критерий также служит для проверки случайности выборки, для чего вычисляют разности между последовательными элементами выборки: $a_1 = x_2 - x_1$, $a_2 = x_3 - x_2$, ..., $a_{n-1} = x_n - x_{n-1}$. Затем подсчитывается сумма квадратов этих разностей и вычисляется величина

$$C = \frac{\sum_{i=1}^{n-1} a_i^2}{2S^2(n-1)} \quad (n \text{ — объем выборки}), \tag{9.35}$$

где S^2 — оценка генеральной дисперсии, вычисляемая по формуле

$$S^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}. \tag{9.36}$$

Оказывается, при $n \geqslant 20$ величина C распределена приближенно нормально со средним значением $C = 1$ и средним квадратическим отклонением $\sigma_c = \frac{1}{\sqrt{n+1}}$.

Следовательно, чем больше значение $C_ф$ отличается от 1, тем больше оснований считать выборку неслучайной.

Пример 17. Снова возьмем выборку крестьянских хозяйств по числу лошадей, но добавим в нее новые элементы так, чтобы число их было равно 20:

0, 3, 2, 1, 0, 5, 1, 2, 8, 0, 3, 2, 2, 0, 1, 0, 0, 3, 6, 1.

1) Будем проверять гипотезу H_0 о том, что эта выборка является случайной.

2) В качестве статистической характеристики возьмем нормированное отклонение t:

$$t = \frac{C - \bar{C}}{\sigma_C}. \tag{9.37}$$

В данном случае $\sigma_C = \frac{1}{\sqrt{n+1}} = \frac{1}{\sqrt{21}}$, а $\bar{C} = 1$;

поэтому

$$t=(C-1)\sqrt{21}\approx 4{,}6(C-1).$$

3) Для подсчета $C_ф$ надо определить C по формуле (9.35). Сначала по формуле (9.36) найдем значение $S^2(n-1)$ и $\bar{x}$:

$$\bar{x}=\frac{\sum_{i=1}^{n}x_i}{n}=2,$$

$$S^2(n-1)=\sum_{i=1}^{n}(x_i-2)^2=92.$$

Теперь выпишем ряд последовательных разностей:

$$3,\ -1,\ -1,\ -1,\ 5,\ -4,\ 1,\ 6,\ -8,\ 3,\ -1,\ 0,\ -2,$$
$$1,\ -1,\ 0,\ 3,\ 3,\ -5$$

и найдем сумму их квадратов

$$\sum_{i=1}^{n}a_i^2=213.$$

Подставляя найденные величины в формулу (9.35), получим

$$C_ф=\frac{213}{2\cdot 92}\approx 1{,}16,$$

откуда $t_ф=0{,}16\cdot 4{,}6=0{,}736$.

4) Выберем $\alpha=0{,}1$.

5) По табл. 1 приложения для $\alpha=0{,}1$ найдем $t_{кр}=1{,}65$.

6) Поскольку $t_ф<t_{кр}$, гипотеза о случайности нашей выборки не отклоняется.

§ 7. Особенности применения критериев в исторических исследованиях

1. В историческом исследовании, как правило, трудно свести сложную исследовательскую задачу к формализованной задаче проверки статистической гипотезы о значении некоторого параметра или о виде распределения признака. Ввиду того что формализация задачи упрощает сложные взаимосвязи, типичные для общественных наук, часто статистическая формулировка этой задачи не тождественна испытуемой гипотезе, однако желательно подбирать статистическую гипотезу, которая наилучшим образом отражает задачу конкретно-исторического исследования.

2. Отметим, что в общественных науках существенным является вопрос выявления закономерностей сложных яв-

лений, установления взаимосвязей. При этом важно не установить наличие связи там, где ее на самом деле нет. Поэтому обычно проверяют гипотезы об отсутствии взаимосвязей, закономерностей (см. § 4). Заметим также, что выбор уровня значимости при проверке гипотез, т. е. выбор вероятности отвергнуть правильную гипотезу, всецело зависит от знания объекта исследования и его задач. Иногда важно не столько уменьшать уровень значимости, сколько повышать мощность критерия, с тем чтобы неотклонение испытуемой гипотезы фактически означало ее принятие (в частности, для проверки гипотез о значении среднего или дисперсии в генеральной совокупности), что требует проверки гипотезы с довольно большим уровнем значимости α.

3. Наконец, часто историку приходится иметь дело не с выборкой, а с самой генеральной совокупностью (когда речь идет о сплошном обследовании небольшой по размеру генеральной совокупности — например, данные о губернских хлебных ценах по некоторому региону). В этом случае параметры, вычисленные по таким данным, не требуют оценок с помощью выборочной теории или теории проверки гипотез. Однако проверка гипотез все же имеет смысл для задач установления связей или законов распределения, которые в малых по размеру совокупностях могут затемняться действием случайных причин. При этом гипотеза проверяется не для распространения полученных выводов на некую более обширную генеральную совокупность, а для того, чтобы установить, насколько закономерными или же случайными являются полученные выводы для имеющихся в данной совокупности условий.

ГЛАВА 10

МЕТОДЫ МНОГОМЕРНОГО СТАТИСТИЧЕСКОГО АНАЛИЗА

При изучении массовых исторических источников исходная информация часто может быть представлена в виде набора объектов, каждый из которых характеризуется рядом признаков (показателей). В качестве объектов могут выступать хозяйства, поселения, административно-территориальные единицы и т. д., а в качестве признаков — различные показатели социально-экономической или демографической структуры изучаемых объектов.

Пусть имеется n объектов, каждый из которых характеризуется набором из m признаков. Обозначим через x_{ik} зна-

Номер объекта	Номер признака			
	1	2	. . .	m
1	x_{11}	x_{12}	. . .	x_{1m}
2	x_{21}	x_{22}		x_{2m}
⋮	⋮	⋮		⋮
n	x_{n1}	x_{n2}	. . .	x_{nm}

Рис. 25. Матрица данных

чение k-го признака для i-го объекта. Тогда исходная информация может быть представлена в виде таблицы, которую называют **матрицей данных.** Эта таблица имеет n строк (по числу объектов) и m столбцов (по числу признаков). Таким образом, каждая строка этой таблицы соответствует одному из объектов, а каждый столбец — одному из признаков (рис. 25).

Как показывает опыт анализа массовых источников, число объектов может достигать многих десятков и сотен; число признаков также может исчисляться десятками. Очевидно, непосредственный (визуальный) анализ матрицы данных при большом количестве объектов и признаков практически малоэффективен — можно лишь выявить отдельные особенности изучаемой структуры, извлечь иллюстративные, частные примеры. В этих условиях возникают задачи укрупнения, концентрации исходных данных, т. е. построения обобщенных характеристик множества признаков и множества объектов. Решение этих задач может осуществляться с помощью современных методов многомерного статистического анализа. При этом методы, ориентированные на анализ структуры множества признаков и выявление обобщенных факторов, известны как методы **факторного анализа**, а методы анализа структуры множества объектов образуют совокупность методов **автоматической классификации.**

§ 1. Методы автоматической классификации

Классификация изучаемых явлений, объектов является одной из важнейших задач науки вообще и исторической науки в частности. В историко-социальных исследованиях эта задача возникает, как правило, в связи с необходимостью решения проблем типологии. Традиционный метод построения типологии сводится обычно к группировке изучаемых объектов на основе одного (или двух-трех) признаков, характеризующих эти объекты (в последнем случае группировки называют комбинационными). Важно отметить, что традиционные приемы типологической группировки направлены на выявление качественно однородных групп объектов путем определения границ интервалов на оси одного из группообразующих признаков; эти приемы носят неформальный характер и осуществляются на основе содержательных концепций и опыта предшествующих исследований.

Современный уровень развития методов многомерного количественного анализа и наличие ЭВМ позволяют осуществлять классификацию объектов на более широкой и более объективной основе — с учетом всех существенных структурно-типологических признаков и характера распределения объектов в заданной системе признаков.

В настоящее время существует целый ряд методов построения классификации многомерных объектов с помощью ЭВМ.

При этом традиционно выделяют две группы методов. Методы первой группы связаны с задачей «узнавания», идентификации объектов; они получили название методов **распознавания образов.** Смысл распознавания заключается в том, чтобы любой предъявляемый машине объект с наименьшей вероятностью ошибки был отнесен к одному из заранее сформированных классов. Здесь машине сначала предъявляют «обучающую последовательность» объектов (о каждом из которых известно, к какому классу или «образу» он принадлежит). Затем, «обучившись», машина должна распознать, к каким классам относятся новые объекты из изучаемой совокупности.

Более общий подход к классификации включает не только отнесение объектов к одному из классов, но и одновременное формирование самих «образов», число которых может быть заранее неизвестно. При отсутствии обучающей последовательности такая классификация производится на основе стремления собрать в одну группу в некотором смысле схожие объекты, да еще так, чтобы объекты из разных групп (классов) были по возможности несхожими. Именно такие методы получили название методов **автоматической классификации** (кластерного анализа, таксономии).

Отсутствие априорной информации о характере распределения объектов внутри каждой группы предполагает построение многомерной классификации на основе методов **кластерного анализа**[1]. Рассмотрим на примере кластерного анализа основные этапы построения многомерной классификации.

Кластер-анализ. Будем считать, что все m признаков измерены в количественной шкале. Тогда каждый из n объектов может быть представлен точкой в m-мерном пространстве признаков. Характер распределения этих точек в рассматриваемом пространстве определяет структуру сходства и различия объектов в заданной системе показателей.

О сходстве объектов можно судить по расстоянию между соответствующими точками. Содержательный смысл такого понимания сходства означает, что объекты тем более близки, похожи в рассматриваемом аспекте, чем меньше различий между значениями одноименных показателей.

Для определения близости пары точек в многомерном пространстве обычно используют евклидово расстояние, равное корню квадратному из суммы квадратов разностей значений одноименных показателей, взятых для данной пары объектов:

$$d_{ij}=\sqrt{\sum_{k=1}^{m}(x_{ik}-x_{jk})^2} \quad (i,\ j=1,\ldots,n), \tag{10.1}$$

где d_{ij} — евклидово расстояние между i-м и j-м объектами; x_{ik} — значение k-го показателя для i-го объекта.

Вычислив расстояние между каждой парой объектов, получим квадратную матрицу D, имеющую размеры $n \times n$ (по числу объектов); эта матрица, очевидно, симметрична, т. е. $d_{ij}=d_{ji}$ $(i,\ j=1,\ \ldots,n)$.

Матрица расстояний D служит основой при реализации методов кластерного анализа, в том числе и **агломеративно-иерархического метода**, который часто используется для многомерной классификации объектов в социально-экономических исследованиях. Основная идея этого метода заключается в последовательном объединении группируемых объектов — сначала самых близких, затем все более удаленных друг от друга. Процедура построения классификации состоит из последовательности шагов, на каждом из которых производится объединение двух ближайших групп объектов (кластеров).

---

[1] Cluster (англ.) — скопление, «гроздь», группа объектов, характеризующихся общими свойствами.

Существуют различные способы определения расстояния между группами объектов (различающие методы кластерного анализа). Обычно близость двух кластеров определяется как средний квадрат расстояния между всеми такими парами объектов, где один объект пары принадлежит к одному кластеру, а другой — к другому:

$$D^2_{pq} = \sum_{i \in X_p} \sum_{j \in X_q} d^2_{ij}/n_p n_q, \qquad (10.2)$$

где D^2_{pq} — мера близости между p-м и q-м кластерами; X_p — p-й кластер; X_q — q-й кластер; n_p — число объектов в p-м кластере; n_q — число объектов в q-м кластере.

На первом шаге процедуры агломеративно-иерархического метода кластерного анализа рассматривается начальная матрица расстояний между объектами и по ней определяется минимальное число $a_{i_1 j_1}$; далее, наиболее близкие объекты с номерами i_1 и j_1 объединяются в один кластер, в матрице вычеркиваются строка и столбец с номером j_1, а расстояния от нового кластера (он получает номер i_1) до всех остальных кластеров (на первом шаге — объектов) вычисляются по формуле (10.2) — в данном случае квадраты таких расстояний равны полусуммам квадратов расстояний от i-го и j-го объектов до каждого из остальных. Эти вновь вычисленные значения расстояний заносятся в i_1-ю строку и i_1-й столбец матрицы D.

На втором шаге процедуры по матрице D, содержащей уже n—1 строк и столбцов, определяют минимальное число $a_{i_2 j_2}$ и формируют новый кластер с номером i_2. Этот кластер может быть построен в результате объединения либо двух объектов, либо одного объекта с i_1-м кластером, построенном на первом шаге. Далее, в матрице D вычеркиваются строка и столбец с номером i_2, а строка и столбец с номером i_2 пересчитываются, и т. д.

Таким образом, метод кластерного анализа включает n—1 аналогичных шагов. При этом после выполнения k-го шага ($k \leqslant n-1$) число кластеров равно $n-k$ (некоторые из них могут являться отдельными объектами), а матрица D имеет размеры $(n-k)\times(n-k)$. В конце этой процедуры, на $(n-1)$-м шаге, получится кластер, объединяющий все n объектов.

Результаты классификации, построенной изложенным методом, удобно изобразить в виде дерева иерархической структуры (дендрограммы), содержащего n уровней, каждый из которых соответствует одному из шагов описанного процесса последовательного укрупнения кластеров.

Одним из важных вопросов в кластерном анализе является выбор необходимого числа кластеров. В некоторых случаях число кластеров может быть выбрано из априорных соображений, однако чаще это число определяется в процессе формирования кластеров исходя из значений некоторых показателей их однородности и степени удаленности друг от друга (например, показателей внутригрупповой дисперсии или вариации).

Другой важный вопрос — зависимость результатов классификации от масштабов используемых значений показателей. Из формулы (10.1) следует, что изменение масштаба значений показателей приводит к изменению расстояний между объектами. Так, например, если некоторый показатель, выраженный в рублях, переведен в копейки, то это приведет к увеличению относительного «вклада» этого показателя при вычислении меры близости D^2_{pq} в $(100)^2 = 10\,000$ раз. Поэтому для устранения такой неоднородности исходных данных показатели нормализуют путем вычитания среднего значения и деления на среднее квадратическое отклонение, так что дисперсия каждого показателя оказывается равно 1, а среднее — 0. Тем самым все показатели оказываются равноценными с точки зрения определения сходства рассматриваемых объектов.

Завершая описание агломеративно-иерархического метода кластерного анализа, отметим, что минимизация среднего расстояния между кластерами, которая производится на каждом шаге, эквивалентна минимизации некоторого критерия качества классификации, оценивающего степень однородности формируемых кластеров.

Пример 1. Проиллюстрируем изложенный метод кластерного анализа на примере задачи построения аграрной типологии губерний европейской России на рубеже XIX—XX вв. Для простоты изложения выберем небольшое число губерний и показателей их аграрного развития[1]. Возьмем, например, шесть северо-западных губерний и два признака: 1) наемные сельскохозяйственые рабочие по отношению к местным работникам; 2) доля дворянской земли в удобной земле.

Сформируем матрицу данных, имеющую в данном случае размеры 6×2 (табл. 1). Вычислив средние значения и средние квадратические отклонения для каждого из двух отобранных показателей, нормализуем их (табл. 1). Используя формулу (10.1), вычислим расстояния для каждой пары губерний в пространстве данных двух признаков (табл. 2, верхняя часть над диагональю). В нижней части табл. 2 приводятся нормированные расстояния, полученные путем деления чисел d_{ij} на

[1] Данные приводятся из работы: *Ковальченко И. Д., Бородкин Л. И.* Аграрная типология губерний европейской России на рубеже XIX—XX веков. — История СССР, 1979, № 1.

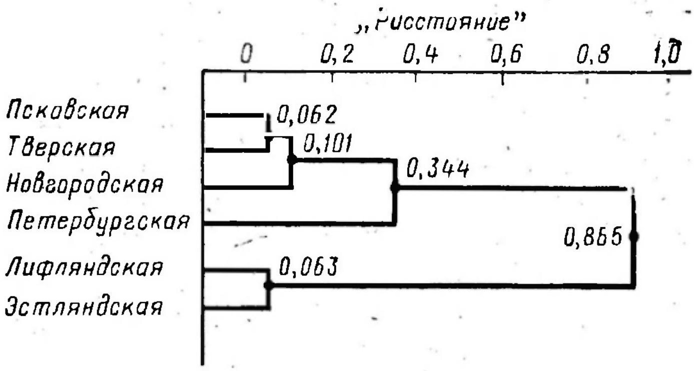

Рис. 26. Дерево иерархической классификации (дендрограмма)

максимальное из этих чисел (в данном случае максимальным является число $d_{3,5}=3{,}363$).

В соответствии с первым шагом агломеративно-иерархического метода кластерного анализа находим минимальное число в нижней части таблицы 2 ($d_{3,1}=0{,}062$) и объединяем в одну группу объекты 3 и 1 (Тверскую и Псковскую губернии). Вычеркнув из матрицы расстояний 3-ю строку и 3-й столбец и определив новые значения 1-й строки и 1-го столбца, переходим ко второму шагу алгоритма. Теперь уже минимальным будет расстояние между 5-м и 6-м объектами ($d_{5,6}=0{,}063$), следовательно, в одну группу объединяются Эстляндская и Лифляндская губернии и т. д. Действуя таким образом, на третьем шаге ($D_{pq}=0{,}101$) объединим в одну группу объекты 1, 3 и 2 (т. е. к Псковской и Тверской губернии присоединится Новгородская); на четвертом шаге в эту группу будет включена и Петербургская губерния ($D_{pq}=0{,}344$), а на последнем, пятом шаге алгоритма все шесть губерний будут объединены в одну общую группу.

Результаты построения данной классификации могут быть компактно представлены в виде дендрограммы — дерева иерархической классификации (рис. 26). Число, стоящее у каждого узла дерева, показывает расстояние между объединяющимися на данном шаге кластерами. Анализируя структуру полученной классификации, можно видеть, что при учете двух выбранных показателей рассматриваемые шесть губерний могут быть разделены на две группы (1, 2, 3, 4) и (5, 6); при этом можно детализировать структуру каждой группы. Так, группа (1, 2, 3, 4) состоит из весьма близких объектов 1, 2 и 3 (максимальное расстояние $d_{2,3}=0{,}126$) и относительно удаленного от них объекта 4 (минимальное расстояние от этого объекта до объектов из группы (1, 2, 3) равно $d_{2,4}=0{,}276$). При более детальной классификации (на три кластера) выделяются группы (1, 2, 3), (4), (5, 6) (см. рис. 26).

Таблица 1. Исходные и нормализованные значения признаков

Губернии	Исходные значения признаков		Нормализованные значения признаков	
	наемные с.-х. рабочие, % к местным работникам	доля дворянских земель в удобной земле	наемные с.-х. рабочие, % к местным работникам	доля дворянских земель в удобной земле
Псковская	0,027	0,140	—0,80	—0,93
Новгородская	0,032	0,168	—0,76	—0,71
Тверская	0,029	0,116	—0,78	—1,13
Петербургская	0,068	0,267	—0,47	0,18
Лифляндская	0,305	0,407	1,44	1,39
Эстляндская	0,298	0,384	1,38	1,19
	$\overline{X}=0,126$ $\sigma=0,124$	$\overline{X}=0,247$ $\sigma=0,115$	$X=0$ $\sigma=1$	$X=0$ $\sigma=1$

Простота рассмотренного примера позволяет дать геометрическую интерпретацию структуры распределения объектов в пространстве двух признаков. На рис. 27 изображены шесть точек, соответствующих шести объектам (губерниям), заданным значениями этих двух признаков. В данном случае классификация объектов может быть легко получена и «на глаз», путем визуального выделения скоплений, групп близко расположенных точек на плоскости двух параметров. Однако при большом числе признаков и объектов такая геометрическая интерпретация структуры не может быть получена принципиально, и это приводит к необходимости использования в типологических задачах большой размерности методов автоматической классификации и ЭВМ. Так, построение с помощью кластерного

Таблица 2. «Расстояния» между губерниями в двухмерном пространстве признаков

№ п/п	Губернии	„Расстояния" между губерниями					
		1	2	3	4	5	6
1	Псковская	0	0,221	0,209	1,152	3,225	3,048
2	Новгородская	0,066	0	0,496	0,931	3,038	2,865
3	Тверская	0,062	0,126	0	1,350	3,363	3,190
4	Петербургская	0,343	0,276	0,401	0	2,265	2,126
5	Лифляндская	0,959	0,903	1,000	0,673	0	0,210
6	Эстляндская	0,906	0,855	0,948	0,632	0,063	0

анализа классификации 50 губерний в 19-мерном пространстве аграрных признаков требует выполнения десятков и сотен тысяч вычислительных операций, и решить эту задачу без ЭВМ невозможно.

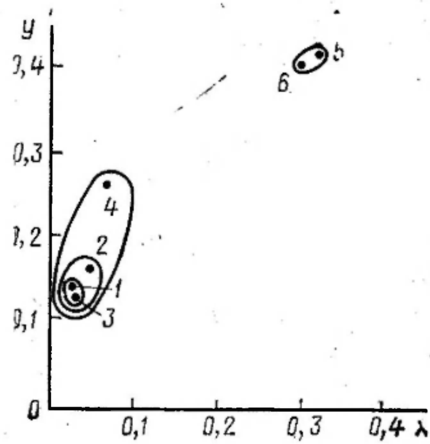

Рис. 27. Диаграмма распределения объектов (губерний) в двухмерном пространстве признаков

Меры близости объектов. Отметим, что степень сходства многомерных объектов может быть охарактеризована не только с помощью евклидова расстояния (10.1), но и с помощью других показателей, выбор которых определяется структурой признакового пространства и целью классификации. Например, если признаки имеют качественную природу (пусть, для определенности, все m признаков — альтернативные, т. е. принимают значения 0 или 1), то степень сходства пары объектов (i, j) может быть выражена целым рядом коэффициентов, из которых приведем здесь следующие два:

а) расстояние по Хеммингу

$$\hat{d}_{ij} = \sum_{k=1}^{m} |x_{ik} - x_{jk}|, \qquad (10.3)$$

б) коэффициент композиционного сходства

$$S_{ij} = P_{ij}/(m + q_{ij}), \qquad (10.4)$$

где p_{ij} и q_{ij} — число признаков, имеющих соответственно одинаковые и различающиеся значения для i-го и j-го объектов.

Как следует из (10.3), расстояние по Хеммингу $\hat{d}_{ij}$ равно количеству признаков, значения которых для обоих объектов не совпадают. Значения $\hat{d}_{ij}$ изменяются от 0 до m; они тем меньше, чем ближе эти объекты в заданной системе признаков.

Что касается коэффициента композиционного сходства S_{ij}, то его значение тем больше, чем ближе данные два объекта; S_{ij} изменяется в пределах от 0 до 1. Как следует из (10.4), $S_{ij}=0$, если значения всех одноименных признаков для обоих объектов различаются, и $S_{ij}=1$, если значения всех признаков для них совпадают.

Подсчитав значения коэффициентов $\hat{a}_{ij}$ или S_{ij} для всех пар объектов, получим квадратную матрицу размеров $n \times n$, аналогичную матрице расстояний D (и также симметричную), которая далее может анализироваться с помощью какого-либо метода автоматической классификации.

Построенная с помощью этих методов многомерная группировка объектов может рассматриваться в типологическом аспекте, если содержательный анализ полученных результатов позволяет указать качественные и количественные особенности выделенных групп — кластеров.

§ 2. Факторный анализ

Важнейшей составной частью многих исторических исследований оказывается задача выявления и анализа структуры взаимосвязей показателей, выбранных для описания того или иного явления или процесса. При наличии количественной информации, записанной в виде матрицы исходных данных (рис. 25), для решения указанной задачи может быть привлечен корреляционный анализ. В этом случае по матрице данных рассчитывают матрицу корреляций:

$$\begin{matrix} 1 & r_{12} & r_{13} & \dots & r_{1j} & \dots & r_{1m} \\ r_{21} & 1 & r_{23} & \dots & r_{2j} & \dots & r_{2m} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ r_{i1} & r_{i2} & r_{i3} & \dots & r_{ij} & \dots & r_{im} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ r_{m1} & r_{m2} & r_{m3} & \dots & r_{mj} & \dots & 1, \end{matrix}$$

где r_{ij} — коэффициент корреляции между показателями, записанными в i-м и j-м столбцах матрицы данных; на диагонали матрицы корреляций стоят единицы, меры связи каждого показателя с самим собой; матрица симметричная, т. е. $r_{ij}=r_{ji}$, $i, j=1, 2, \dots, m$; m — число показателей. Изучение матрицы корреляций позволяет выделить в основном группы взаимосвязанных характеристик, подсчитать среднюю тесноту связи в каждой группе и т. д., т. е. дает самое общее представление о структуре связей. Естественно, перед исследователем встает вопрос: каковы причины, обусловившие именно такую структуру взаимосвязей показателей в данном явлении? Коэффициенты корреляции, которые лишь фиксируют факт наличия связи определенной величины и определенного направления, не дают никакой информации на этом этапе исследования.

В решении поставленного вопроса существенную помощь может оказать метод многомерного статистического анализа — **факторный анализ**. Известно, что взаимосвязь двух или более показателей объясняется либо тем, что они взаимно обусловливают друг друга, либо тем, что на каждый из них воздействует некий скрытый, не входящий в данную группу признак. В факторном анализе исходят из второй ситуации. Основное предположение этого метода состоит в следующем: любое явление или процесс могут быть описаны небольшим числом некоторых скрытых, обобщенных характеристик, которые не поддаются непосредственному наблюдению, но воздействуют на «внешние», наблюдаемые показатели, определяют их изменения и обусловливают тем самым взаимосвязи между ними. Эти скрытые характеристики явления в факторном анализе называются **общими факторами**.

Рассмотрим, например, такие понятия, как уровень развития капитализма, структура населения, его мобильность, уровень жизни или культурный уровень какой-то группы людей, размер хозяйства и т. д. Очевидно, все эти обобщенные показатели характеризуют реально существующие явления и процессы, но в силу многосторонности своего проявления не могут быть измерены непосредственно. С другой стороны, историк, изучающий, скажем, положение крестьян в России конца XIX — начала XX в. и рассматривающий некоторую группу крестьян в некоторый год, включит в свое исследование показатели, характеризующие развитие капитализма в сельском хозяйстве: количество наемных работников, количество лошадей, размер аренды и т. д. Если большинству дворов изучаемой совокупности были присущи в той или иной степени элементы ведения хозяйства по-капиталистически, то перечисленные показатели окажутся тесно взаимосвязанными с некоторыми другими социально-экономическими характеристиками. В этом случае можно сказать, что для данной группы дворов структура взаимосвязей исходных показателей объясняется, в частности, такой обобщенной характеристикой, как капиталистический способ ведения хозяйства. Если же изучается крестьянство тех районов или тех периодов времени, где и когда развитие капитализма еще только начиналось, исследователь указанных выше взаимосвязей не обнаружит. Верно и обратное — наличие взаимосвязей между показателями развития капитализма и рядом других социально-экономических параметров позволяет характеризовать данную группу крестьян как ведущую в основном хозяйство по-капиталистически.

Таким образом, сложную структуру взаимосвязей показателей в каком-либо явлении факторный анализ объясняет наличием у этого явления небольшого числа скрытых, обобщенных характеристик — **общих факторов.** Каждый общий фактор в той или иной мере воздействует на все исходные показатели; те из них, чьи изменения в наибольшей степени определяются этим фактором, оказываются тесно связанными между собой. Исходные показатели рассматриваются в факторном анализе как «несамостоятельные» переменные — их поведение, т. е. изменение от объекта к объекту, обусловливается соответствующими изменениями общих факторов, присутствующих в данном явлении. Однако общие факторы определяют каждый показатель не полностью, а лишь ту его часть, которая согласована, взаимосвязана с изменениями остальных исходных характеристик. Для объяснения же индивидуальных, несогласованных изменений показателя вводятся так называемые **характерные факторы**. Другими словами, предполагается, что поведение каждого исходного показателя определяется действием на него совокупности небольшого числа общих факторов, влияющих на все показатели и обусловливающих взаимосвязи между ними, и характерного фактора, воздействующего только на данный показатель.

Факторный анализ позволяет выявлять общие факторы, дает ключ к их содержательному толкованию, оценивает их воздействие на отдельные показатели и на все изучаемое явление в целом, количественно выражает их значения для каждого из рассматриваемых объектов и на основании всего этого дает возможность решать целый ряд задач, возникающих при обработке массовых исторических источников.

Обозначим через $x_1, x_2, ..., x_m$ исходные показатели, $f_1, f_2, ..., f_k$ — общие, а $e_1, e_2, ..., e_m$ — характерные факторы. Тогда основная модель факторного анализа запишется в виде[1]

$$
\begin{aligned}
x_1 &= a_{11}f_1 + a_{12}f_2 + \dots + a_{1j}f_j + \dots + a_{1k}f_k + u_1e_1,\\
x_2 &= a_{21}f_1 + a_{22}f_2 + \dots + a_{2j}f_j + \dots + a_{2k}f_k + u_2e_2, \qquad (10.5)\\
&\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots
\end{aligned}
$$

[1] Здесь будет рассматриваться только линейная модель факторного анализа с некоррелированными общими факторами.

$$x_i = a_{i1}f_1 + a_{i2}f_2 + \ldots + a_{ij}f_j + \ldots + a_{ik}f_k + u_i e_i,$$

$$\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots$$

$$x_m = a_{m1}f_1 + a_{m2}f_2 + \ldots + a_{mj}f_j + \ldots + a_{mk}f_k + u_m e_m.$$

Здесь $x_1, x_2, \ldots, x_m$ рассматриваются как случайные величины, подчиненные многомерному нормальному распределению; $f_1, \ldots, f_k$ предполагаются независимыми, нормально распределенными случайными величинами с нулевыми средними и единичными дисперсиями; характерные факторы $e_1, \ldots, e_m$ считаются случайными величинами, независимыми как между собой, так и с общими факторами, с нулевыми средними и единичными дисперсиями, параметры u_i, $i=1, \ldots m$, оценивают влияние характерного фактора e_i на x_i; число общих факторов k предполагается значительно меньшим числа исходных показателей m; каждое из чисел a_{ij}, $i=1, \ldots, m$, $j=1, \ldots, k$, показывает степень влияния j-го общего фактора на i-й признак и называется **факторной нагрузкой**. Чтобы на проведение и результаты анализа не влияли различия исходных показателей в значениях их средних и дисперсий, эти показатели нормируют, т. е. преобразуют так, что средние становятся равными нулю, а дисперсии — единице[1]. В модели (10.5) и везде далее исходные показатели считаются нормированными.

Так как общие факторы сами ненаблюдаемы, а обнаруживают себя, влияя на «внешние» переменные, каждый из них в рамках данного исследования можно полностью охарактеризовать теми воздействиями, которые он оказывает на исходные показатели, т. е. соответствующим набором m факторных нагрузок $(a_{1j}, \ldots, a_{mj})$, $j=1, \ldots, k$. Факторные нагрузки рассчитываются по матрице корреляций и представляются в виде основной таблицы факторного анализа (см. табл. 3).

Табл. 3 имеет m строк по числу исходных показателей и k столбцов по числу общих факторов. Кроме того, таблица имеет один дополнительный столбец и одну дополнительную строку, о которых будет сказано ниже.

Рассмотрим основные параметры, входящие в табл. 3.

Факторные нагрузки $(a_{ij}, i=1, 2, \ldots, m, j=1, 2, \ldots, k)$. В сделанных выше предположениях о модели (10.5) каждая факторная нагрузка a_{ij} представляет собой коэффициент корреляции между j-м общим фактором и i-м исходным показателем и обладает всеми свойствами этого коэффици-

---

[1] Способ нормировки указан в предыдущем параграфе.

Таблица 3. Факторные нагрузки

№ признаков	№ факторов							Общности
	1	2	...	j	...	k		
1	a_{11}	a_{12}	...	a_{1j}	...	a_{1k}		h_1^2
2	a_{21}	a_{22}	...	a_{2j}	...	a_{2k}		h_2^2
.	. . .	. . .	. . .	. . .	. . .	. . .		. . .
i	a_{i1}	a_{i2}	...	a_{ij}	...	a_{ik}		h_i^2
.	. . .	. . .	. . .	. . .	. . .	. . .		. . .
m	a_{m1}	a_{m2}	...	a_{mj}	...	a_{mk}		h_m^2
Вклады факторов	V_1^2	V_2^2	...	V_j^2	...	V_k^2		Общий вклад всех факторов D

ента. Это значит, что значения факторных нагрузок заключены в пределах от —1 до 1; абсолютная величина нагрузки указывает на тесноту связи между фактором и признаком: чем больше a_{ij} по модулю, тем в большей степени изменения i-го признака определяются j-м фактором, и, наоборот, близкие к нулевым значения указывают на практическую независимость i-го признака от воздействия j-го фактора; знак нагрузки говорит о направлении связи: если a_{ij} положительна, то объектам, у которых j-й фактор проявляется сильнее, соответствуют в основном бо́льшие значения i-го показателя, если же a_{ij} отрицательна, то бо́льшему проявлению фактора соответствуют меньшие значения показателя. Величина a^2_{ij} (коэффициент детерминации) показывает ту долю дисперсии i-го показателя, которая объясняется влиянием j-го фактора.

Каждый общий фактор в табл. 3 представлен столбцом, а каждый исходный показатель — строкой факторных нагрузок. Поэтому по нагрузкам, стоящим в j-м столбце, можно судить о влиянии j-го фактора на каждый из показателей, а по нагрузкам i-й строки — о влиянии каждого из общих факторов на i-й показатель.

Вопрос о значимости факторных нагрузок во многих прикладных работах решается на основании эвристического подхода: значимыми считают нагрузки, абсолютная величина которых превосходит некое заданное значение — обычно 0,3 или 0,4. Существуют также приближенные формулы и соответствующие им таблицы, позволяющие получать стандартные ошибки нагрузок [1].

Основное свойство факторных нагрузок, вытекающее из модели (10.5), записывается в виде равенства

$$r_{iv}=a_{i1}a_{v1}+a_{i2}a_{v2}+\ldots+a_{ij}a_{vj}+\ldots+a_{ik}a_{vk}, \qquad (10.6)$$
$$i,\ v=1,\ 2,\ldots,m,$$

которое означает, что коэффициент корреляции между любыми двумя исходными показателями может быть представлен как сумма произведений факторных нагрузок, стоящих в соответствующих строках табл. 3. Каждое слагаемое $a_{ij}a_{vj}$ в (10.6) есть та часть корреляции i-го и v-го исходных показателей, которая обусловливается влиянием на них j-го общего фактора ($j=1$, 2, ..., k). Таким образом, любой внедиагональный элемент матрицы корреляций может быть вычислен из таблицы факторных нагрузок. В этом смысле общие факторы чисто формально объясняют взаимосвязи исходных показателей, а табл. 3, имеющая размер $m\times k$, является сжатием информации, представленной в матрице корреляций размера $m\times m$. Для диагональных элементов матрицы корреляций формула (10.6) дает значение $a^2_{i1}+a^2_{i2}+\ldots+a^2_{ik}=h_i^2$, $i=1$, 2, ..., m, которое, вообще говоря, меньше единицы, т. е. элементы r_{ii}, $i=1$, ..., m, воспроизводятся не полностью. Это закономерно, так как общие факторы объясняют лишь ту часть изменений каждого признака, которая взаимосвязана с изменениями других показателей, а коэффициент корреляции признака с самим собой учитывает и его индивидуальное, несогласованное изменение. В связи с этим перед проведением вычислительных процедур или во время их матрица корреляций редуцируется, т. е. единицы на ее диагонали заменяются полученными по некоторым правилам оценками величин h^2_i, $i=1$, 2, ..., m. Равенство (10.6) является основой математических преобразований, переводящих матрицу корреляций в таблицу факторных нагрузок.

Общности показателей (h_i^2, $i=1$, 2, ..., m). Величина h_i^2 оценивает степень воздействия всех общих факторов на i-й показатель. Из (10.5) следует, что дисперсия i-го показателя может быть записана в виде:

$$D^2x_i=1=a_{i1}^2+a_{i2}^2+\ldots+a_{ik}^2+u_i^2=h_i^2+u_i^2, \qquad (10.7)$$
$$i=1,\ 2,\ldots,m.$$

[1] См.: *Харман Г.* Современный факторный анализ. М., 1972, с. 448—450.

Каждое число a^2_{ij} есть доля дисперсии i-го признака, обусловленная влиянием на него j-го фактора. Поэтому сумма $h_i^2 = a^2_{i1} + a^2_{i2} + ... + a^2_{ik}$, называемая общностью, есть доля дисперсии i-го признака, обусловленная влиянием на него всей совокупности выделенных k факторов. Та часть дисперсии u_i^2, которая оценивает воздействие на показатель характерного фактора, называется **характерностью.** Чем больше h_i^2, а значит, меньше u_i^2, тем сильнее взаимосвязан данный показатель с другими, тем в большей степени объясняется его поведение выделенной системой общих факторов. Наоборот, невысокие значения h_i^2 свидетельствуют о слабой коррелированности показателя с другими, а значит, о слабом воздействии на него выделенных общих факторов. Причиной низкой общности показателя может быть отсутствие в исходном наборе таких признаков, с которыми взаимосвязан данный показатель. В таких ситуациях желательно привлечь к исследованию соответствующие характеристики; последующий расчет и анализ новой, расширенной матрицы корреляций приведет, возможно, к выявлению дополнительных общих факторов. Низкие общности могут быть и результатом неточностей, допущенных при измерении показателей. Как правило, общности показателей выражаются в процентах.

Вклады общих факторов в суммарную дисперсию показателей (V_j^2, $j = 1, 2, ..., k$). Параметр V_j^2 рассчитывается по формуле

$$V_j^2 = a_{1j}^2 + a_{2j}^2 + ... + a_{mj}^2, \ j = 1, 2, ..., k. \tag{10.8}$$

Так как значение a^2_{ij} есть часть дисперсии i-го показателя, связанная с воздействием j-го фактора, сумма $\sum_{i=1}^{m} a_{ij}^2$ есть та часть суммарной дисперсии исходных показателей, которая обусловлена влиянием j-го фактора. Поэтому по величине V_j^2 можно судить о влиянии j-го фактора на всю совокупность исходных показателей в целом. Чем больше значение V_j^2, тем в большей степени общее поведение исходных показателей определяется j-м фактором, тем ярче проявляется этот фактор в совокупности данных признаков, тем важнее соответствующая обобщенная характеристика для описания изучаемого явления. Значения V_j^2 позволяют ранжировать факторы по их влиянию на исходные показатели. В большинстве вычислительных процедур факторного анализа общие факторы выделяются последовательно по мере убывания соответствующих им V_j^2, вкладов факторов в суммарную дисперсию показателей.

Вклады факторов можно выражать в процентах от всей суммарной дисперсии показателей (так как показатели нормированы, их суммарная дисперсия равна m), т. е. в виде $\left(\sum_{i=1}^{m} a_{ij}^2/m\right) 100\%$.

Число общих факторов k. Совокупность общих факторов должна объяснять, в смысле равенства (10.6), матрицу корреляций. Если матрица корреляций рассчитана по выборке, то получающиеся при ее анализе факторы, особенно факторы с небольшими вкладами V_j^2, могут не отражать действительных взаимосвязей показателей, а быть результатом ошибок выборки. Существуют различные формальные критерии, позволяющие оценивать истинное число общих факторов в данной задаче. В некоторых из них рассматривается статистическая значимость корреляций, оставшихся необъясненными после выделения какого-то числа факторов, и на основании этого выносится суждение о необходимости выделения следующего фактора; другие ориентированы на сравнение суммы вкладов выделенных факторов с суммарной общностью показателей, оцененной по некоторым правилам: если сумма вкладов выделенных факторов составит 70—80% суммарной общности показателей, а следующий выделяемый фактор будет учитывать менее 5% суммарной общности, то этот фактор отбрасывается и факторизация прекращается (аналогично может рассматриваться и суммарная дисперсия показателей);и т. д. Не все критерии дают одинаковые результаты. Кроме того, некоторые из факторов, значимых по формальным критериям, могут плохо поддаваться осмыслению. Поэтому решение вопроса о числе общих факторов, удовлетворительно объясняющих взаимосвязи исходных показателей, должно сочетать в себе как формальный, так и содержательный подход. Обычно рекомендуется применять целый ряд критериев, а затем, сравнивая полученные результаты и учитывая возможность содержательного толкования факторов, делать окончательный вывод о их числе. Если исследование проводится в уже хорошо изученной области, а историк отчетливо представляет себе особенности обрабатываемого материала, он может сам указать число общих факторов, оптимальное в данной задаче.

Полный вклад всех общих факторов в суммарную дисперсию показателей (D). Величина D рассчитывается по формуле

$$D=\sum_{i=1}^{m} h_i^2=\sum_{j=1}^{k} V_j^2=\sum_{i=1}^{m}\sum_{j=1}^{k} a_{ij}^2 \qquad (10.9)$$

и обычно выражается в процентах от суммарной дисперсии показателей: $(D/m)100\%$. Из формулы видно, что D оценивает совместное воздействие всех общих факторов на все исходные показатели в целом. Значение D позволяет судить о структурности явления, т. е. о степени согласованности в изменениях его характеристик. Чем больше D, тем больше связей существует между исходными показателями, тем теснее эти связи, тем в большей мере поведение показателей объясняется действием на них общих причин.

По табл. 3 проводится важнейший этап в работе историка, применяющего данный статистический метод, — интерпретация общих факторов, т. е. осмысление, содержательное толкование, наименование выделенных факторным анализом внутренних, обобщенных характеристик явления. Эта процедура в наибольшей степени зависит от знаний исследователя, его интуиции. Интерпретация факторов основывается на значениях соответствующих им факторных нагрузок. При толковании конкретного фактора прежде всего выделяют показатели, наиболее тесно с этим фактором связанные, т. е. имеющие по нему наибольшие по абсолютной величине нагрузки (обычно рассматривают нагрузки по модулю большие 0,5—0,7). Анализируя затем полученную совокупность показателей, исследователь, исходя из своих представлений об изучаемом вопросе, должен попытаться определить ту сторону явления, которую эта совокупность может описывать. При этом, чем больше абсолютная величина нагрузки, тем большее значение придается при интерпретации соответствующему показателю. Знак нагрузки также учитывается: плюс говорит о том, что с ростом показателя фактор увеличивается, минус — уменьшается. Дополнительные обоснования предложенному толкованию фактора должен дать анализ показателей, практически с ним не связанных, т. е. имеющих по нему нагрузки, близкие к нулевым.

Как правило, легче всего интерпретируются факторы с большими вкладами в суммарную дисперсию показателей; наоборот, толкование факторов с малыми V_j^2 часто бывает затруднено. В этих случаях не следует стремиться во что

Таблица 4. Факторные веса

№ объектов	№ факторов					
	1	2	...	j	...	k
1	b_{11}	b_{12}	...	b_{1j}	...	b_{1k}
2	b_{21}	b_{22}	...	b_{2j}	...	b_{2k}
.	.	.	.	.	.	.
i	b_{i1}	b_{i2}	...	b_{ij}	...	b_{ik}
.	.	.	.	.	.	.
n	b_{n1}	b_{n2}	...	b_{nj}	...	b_{nk}

бы то ни стало наименовать фактор. Гораздо лучше указать на трудности, не позволившие четко определить соответствующую обобщенную характеристику, и оставить фактор без объяснения с тем, чтобы дальнейшие исследования другими методами или на другом материале подтвердили или опровергли действительное присутствие этого фактора в изучаемом явлении. Для успешного проведения интерпретации помимо хорошего знания исследователем проблемы и материала желательно его знакомство с аналогичными работами, использующими факторный анализ.

Вторая таблица, получаемая в факторном анализе (см. табл. 4), имеет n строк по числу объектов и k столбцов по числу общих факторов. Каждое из чисел b_{ij}, $i=1, 2, ..., n$, $j=1, 2, ..., k$, называется факторным весом i-го объекта по j-му фактору.

Факторные веса представляют собой количественные значения выделенных общих факторов для каждого из n имеющихся объектов. Объектам с бо́льшими значениями факторных весов присуща бо́льшая степень проявления свойств, характеризующих данный фактор, т. е. бо́льшая степень развития в соответствующем фактору аспекте. Предположим, что историк изучает n крестьянских хозяйств по набору каких-то социально-экономических параметров и в результате анализа таблицы факторных нагрузок он пришел к выводу, что какой-то, скажем первый, из выделенных факторов может быть интерпретирован как фактор ведения хозяйства по-капиталистически. Тогда числа b_{11}, b_{21}, ..., b_{n1}, стоящие в 1-м столбце табл. 4, будут представлять собой показатели степени проявления элементов капитализма в каждом хозяйстве; при этом, чем больше факторный вес какого-то хозяйства, тем более капиталистическим оно является. Обычно табл. 4 рассчитывается таким образом, что нулевые или близкие к нулевым факторные веса получают те объекты, у которых степень проявления свойств, характеризующих данный фактор, находится на среднем для данной совокупности объектов уровне. Соответственно положительные веса оказываются у объектов с бо́льшим, а отрицательные — меньшим среднего уровня проявлением данного фактора. Факторные веса позволяют ранжировать объекты по каждому фактору.

Рассмотрим основные типы задач, которые могут быть решены с помощью факторного анализа.

Определение обобщенных характеристик. Это основная задача факторного анализа. Для ее решения необходимо рассчитать и проанализировать таблицу факторных нагрузок и проинтерпретировать полученные общие факторы. Нахождение обобщенных характеристик углубляет представление об изучаемом явлении, дает формальное подтверждение полученным на содержательном уровне концепциям, объясняющим взаимосвязи исходных показателей, либо приводит к новым гипотезам о причинах этих взаимосвязей.

Измерение обобщенных показателей. Выше уже отмечалось, что некоторые показатели, как, например, уровень жизни, уровень развития капитализма, размер хозяйства и т. д., в силу многосторонности своих проявлений не поддаются непосредственному измерению. Для их косвенного оценивания можно использовать факторный анализ. Возможность применения здесь этого метода связана с тем, что отношения между обобщенным показателем и соответствующими ему частными показателями — его внешними проявлениями, допускающими прямое измерение на объектах, — могут быть описаны однофакторной моделью факторного анализа. Действительно, так же, как общий фактор, выделенный по системе каких-то признаков, воздействует на эти признаки, обусловливая их поведение, обобщенный показатель влияет на свои частные характеристики и определяет их изменения от объекта к объекту.

Схема проведения анализа имеет в этом случае следующий вид. Исследователь выбирает группу объектов, для которых должны быть получены количественные значения некоторого обобщенного показателя, и группу признаков, наиболее полно этот показатель описывающих. Из источников или путем непосредственного измерения определяются числовые значения выбранных признаков для выбранных объектов. По полученной матрице данных проводится затем факторный анализ с заданным числом подлежащих выделению общих факторов, равным одному. Выделенный общий фактор и будет представлять собой данный обобщенный показатель. Полученные значения факторных нагрузок позволят определить сравнительную важность каждого из признаков для характеристики обобщенного показателя. Последующее вычисление факторных весов даст количественное выражение этого показателя для каждого объекта.

Сжатие информации. Как уже указывалось, факторный анализ представляет собой метод, позволяющий описывать набор исходных показателей меньшим числом независимых характеристик — общих факторов. В соответствии с этим табл. 4 факторных весов размера $n\times k$ может рассматриваться как сжатие информации, содержащейся в матрице данных размера $n\times m$. Каждый общий фактор представляет собой более емкую характеристику объекта, концентрирующую в себе информацию, содержащуюся в группе коррелированных признаков. Однако общие факторы воспроизводят лишь общую часть изменений показателей, обусловленную их взаимосвязями; информация о тех изменениях каждого показателя, которые объясняются характерными факторами, при переходе к общим факторам теряется. Такая потеря будет тем меньше, чем больше связей образуют между собой исходные показатели и чем сильнее эти связи. Возможность описания объектов меньшим числом концентрированных характеристик полезна сама по себе. Например, гораздо эффективнее сопоставлять губернии по уровню развития сельского хозяйства, степени его интенсификации и качеству земли, страны — по их экономическому потенциалу, предприятия — по размерам, семьи — по уровню жизни и т. д., чем по совокупностям показателей, которые определяют каждую из перечисленных обобщенных характеристик. Кроме того, идея перехода от исходных показателей к небольшому числу агрегированных признаков оказывается плодотворной при решении задач классификации.

Классификация объектов. Главная задача историка, применяющего методы многомерной классификации, состоит в объяснении внутреннего единства полученных групп объектов. Решение этой задачи при большой размерности признакового пространства часто бывает затруднительным. В таких случаях чрезвычайно полезным может оказаться использование факторного анализа, так как этот метод позволяет с небольшой потерей информации перейти от исходного признакового пространства к меньшему числу более емких характеристик — общих факторов. Последующее проведение классификации на основе полученных факторных весов объектов дает обычно легко интерпретируемые результаты.

Если число выделенных факторов не превосходит трех, возможна наглядная, графическая классификация объектов. Например, в случае двух факторов значения факторных весов по одному фактору принимаются за абсциссы объектов, а по другому — за их ординаты. Тогда каждый объект изображается точкой на плоскости. Компактные скопления точек-объектов следует трактовать как группы похожих, сходных объектов. Если общих факторов больше трех, подобная графическая классификация в пространстве каких-то одного, двух или трех из выделенных факторов также может представлять интерес.

Таким образом, факторный анализ, один из методов многомерного статистического анализа, может быть с успехом применен во многих исторических исследованиях, имеющих своей составной частью анализ количественной информации. Работа историка, применяющего этот метод, состоит из двух основных моментов. Первое — подготовка для математической обработки матрицы исходных данных, второе — осмысление, содержательное толкование результатов работы математического аппарата метода, представленных таблицей факторных нагрузок и таблицей факторных весов. Основные задачи, которые могут быть решены с помощью факторного анализа, — нахождение внутренней структуры, обобщенных характеристик изучаемого явления и группировка, классификация объектов наблюдения.

* *
*

Заключая рассмотрение методов многомерного количественного анализа, отметим тенденцию к объединению методов факторного анализа и автоматической классификации при решении задач типологии. При этом можно указать две основные схемы совместного применения указанных методов для построения типологии.

В первом случае сначала с помощью методов автоматической классификации получают группировку объектов в исходном многомерном пространстве признаков, а затем с помощью факторного анализа выявляют небольшое количество основных факторов (как правило, ограничиваются двумя факторами). В результате каждый объект может быть представлен точкой на плоскости двух главных факторов, а каждый кластер — группой точек. Расположение этих групп на плоскости факторов определяет «статус» каждой из них, а также дает наглядную интерпретацию компактности полученных групп и их относительной удаленности друг от друга.

При использовании второй схемы типологического анализа вначале применяют один из методов факторного анализа, а затем исходя из полученных значений факторных весов для каждого из объектов строят автоматическую классификацию объектов в пространстве уже меньшей размерности, определяемой числом факторов. Такой подход используется обычно в ситуации, когда имеется большое число равноправных признаков, многие из которых взаимосвязаны и в значительной мере дублируют друг друга.

ГЛАВА 11

ЭВМ В ИСТОРИЧЕСКИХ ИССЛЕДОВАНИЯХ

В предыдущих главах изложены основные принципы научного анализа с использованием количественных методов, показаны сильные стороны этих методов и их ограничения. Напомним, что неоспоримыми достоинствами количественных методов в изучении исторических явлений и процессов являются возможность построения их моделей, оценка достоверности и точности используемой информации, а также надежности получаемых результатов и, наконец, сопоставимость результатов различных исследований. Однако все эти положительные результаты могут быть получены, как и при использовании других методов, лишь тогда, когда применение этих методов основывается на глубоком качественном, содержательном анализе изучаемых явлений и при корректном оперировании математическим аппаратом. Это должно предостеречь историков от бездумного обращения к математическим методам и вычислительной технике без предварительного серьезного анализа проблемы и материала в надежде немедленно получить существенные результаты. В последнее время былое недоверие к количественным методам и ЭВМ у некоторых исследователей переходит в свою противоположность и приводит их к забвению определяющей роли теоретико-методологического и научно-методического подхода. Без этого невозможна ни математическая постановка соответствующей задачи, ни выбор метода ее решения, ни правильный подход к отбору информации, ни, что наиболее важно, правильная интерпретация получаемых результатов. Иначе говоря, обращаясь к помощи ЭВМ, исследователь уже должен представлять себе, как и какие результаты он может получить и как их следует интерпретировать, т. е. он уже должен иметь четкую содержательную гипотезу (или гипотезы), доказать или опровергнуть которую и помогают математические методы и вычислительная техника. Таким образом, основой исторического исследования остается теория и методология, содержательный анализ, а методика и техника играют хотя и важную, но вспомогательную роль.

* *
*

Исследование, в котором используются современные методы количественного анализа и вычислительная техника, складывается из ряда этапов.

1. Постановка исследовательской задачи и формулировка содержательной гипотезы (или гипотез) относительно разрешения этой проблемы.

2. Построение содержательной модели изучаемого явления или процесса и отбор системы показателей, характеризующих объект изучения в рамках выбранной модели.

3. Кодировка отобранных показателей, т. е. перевод их в такую форму представления, которая требуется для ввода данных в память ЭВМ.

4. Выбор количественного метода, позволяющего формализовать содержательную модель и дать четкую математическую постановку задачи.

5. Составление алгоритма решения задачи, т. е. описание совокупности этапов ее решения за конечное число шагов и создание программы, реализующей этот алгоритм для конкретной вычислительной системы. Программа должна задавать ЭВМ информацию о форме представления исходного материала, о той обработке, которой он подвергается, и о виде представления результатов обработки.

6. Интерпретация полученных результатов (матриц, таблиц распределений, факторов и т. д.), подтверждение или опровержение выдвинутых гипотез и постановка новых проблем, позволяющих продолжить анализ на качественно более высоком уровне.

Только путь такого постепенного усложнения проблем, углубления анализа на основе «диалога» с ЭВМ может быть плодотворным. Результаты, полученные ЭВМ, должны не подавлять обилием малозначимой или неинтерпретируемой информации, а создавать основу для углубления анализа изучаемых явлений.

Рассмотрим более подробно каждый из названных этапов.

1. Построение гипотезы (или гипотез), подтвердить или опровергнуть которые хочет исследователь, — это, по существу, этап, обязательный в любой научной работе. Однако построение гипотез при использовании количественных методов предъявляет к исследователю более высокие требования, поскольку гипотеза определяет направление исследования, обеспечивает наиболее полный охват проблемы, а также служит критерием отбора данных и основой для сопоставления результатов различных работ по данной проблеме. Поэтому при построении гипотезы, на основе которой проводится последующая формализация, должны быть тщательно продуманы все основные понятия, определения, формулировки с тем, чтобы исключить неоднозначность в их истолковании и выявить их взаимосвязь. Так, гипотеза о временных границах формирования всероссийского рынка [1] требует раскрытия таких понятий, как момент появления товарного рынка, итог складывания единого товарного рынка, местные и региональные рынки, и т. д.

Построение гипотезы непосредственно связано с конструированием модели явления или процесса, соответствующей этой гипотезе, а при построении модели, естественно, встает вопрос о том наборе данных (характеристик объекта исследования), которым оперирует эта модель [2].

2. Моделирование позволяет выявить наиболее существенные черты изучаемого объекта, абстрагируясь от случайных влияний множества факторов, искажающих скрытые, но объективные закономерности. При этом для достаточно сложной гипотезы можно строить различные модели, позволяющие изучать отдельные аспекты проблемы.

В соответствии с моделью производится и отбор данных. Так, при изучении единого аграрного рынка моделью его формирования служит процесс нивелировки рыночных цен и увеличение их синхронности; при этом из всего многообразия показателей, характеризующих рынок, отбираются данные о ценах на основные сельскохозяйственные продукты (рожь, овес, рабочую силу, землю и тягловый скот). Как

---

[1] См.: *Ковальченко И. Д., Милов Л. В.* Всероссийский аграрный рынок, XVIII — начало XX в. (Опыт количественного анализа). М., 1974.

[2] Методологические проблемы постановки исследовательской задачи, репрезентативности данных и моделирования исторических явлений и процессов подробно рассмотрены в части первой данного пособия.

правило, каждый историк стремится выделить для изучения стоящей перед ним проблемы некий «идеальный» набор характеристик, выявляющий существо этой проблемы. Однако то, чем он обычно располагает, отличается от «идеала».

Сформулируем некоторые общие требования к данным. Во-первых, данные должны быть однородными, т. е. полученными из одной и той же генеральной совокупности, постоянной в процессе получения этих данных. В частности, единицы измерения каждого признака должны быть постоянны. Затем, используемые данные должны быть точными и достоверными, т. е. надежными в плане адекватного отражения сути изучаемых явлений. Данные не всегда бывают свободны от искажений, и в этом случае их точность желательно оценить, поскольку она влияет на надежность получаемых выводов. Иногда вследствие грубых ошибок приходится количественные признаки считать качественными или ранговыми, предпочитая потерю информации потере достоверности.

После того как исследователь убедится в достоверности, точности и однородности своих данных и выберет показатели в соответствии с принятой им моделью, перед ним встает вопрос об изучении либо выборочной, либо генеральной совокупности объектов. Если объем генеральной совокупности слишком велик, возникает проблема получения репрезентативной выборки.

После того как выборка получена, часто возникает актуальная для историка задача заполнения отсутствующих значений некоторых признаков у ряда объектов. Пропущенные значения не всегда соответствуют нулевым значениям признаков, чаще они означают, что эта информация по каким-либо причинам утрачена. Если информация о некотором признаке отсутствует у более чем половины всех объектов, такой признак обычно исключают из анализа. Если же пропусков не столь много, то их можно заполнять логически или статистически. Под логическим способом понимается восстановление или оценка отсутствующих значений по значениям других признаков, логически связанных с данным. Например, урожайность можно оценить, зная валовой сбор и площадь посева. Разумеется, в большинстве случаев такая оценка бывает гораздо более сложной.

Простейшим способом статистической оценки является заполнение пропусков средним арифметическим значением признака. Более точным способом является оценка с помощью уравнения регрессии для выражения зависимости между интересующим нас признаком и наиболее сильно связанным с ним признаком (или признаками), значение которого для данного объекта известно.

3. После того как полностью сформирована совокупность объектов и определен набор признаков, данные следует перевести в форму, необходимую для работы с ЭВМ. Вид представления данных зависит от конкретной вычислительной системы. При этом, как правило, количественные данные не требуют каких-либо существенных изменений, а качественные признаки обычно кодируют.

Кодировка качественных признаков связана с формализацией исторических явлений и соответственно требует большой тщательности при учете всех возможных категорий признака и их взаимосвязей. Например, такой признак, как профессия, имеет огромное количество категорий, и вводить их можно по-разному. Наиболее универсальной является иерархическая кодировка, т. е. кодировка по уровням, когда профессии классифицируются сначала по отраслям хозяйства, затем по виду продукции и т. д. Сейчас, при развитии стандартизации, можно использовать уже разработанные официальные системы кодов там, где они есть (это, в частности, позволяет непосредственно сопоставлять результаты, полученные разными исследователями и на однотипном материале). С учетом всех возможных вариантов значений признака и составляется так называемый макет кодировки, позволяющий переводить качественные признаки в их числовые коды. Безусловным требованием к макету является полнота и однозначность, т. е. каждый объект должен попадать в одну и только одну категорию по каждому признаку.

Заметим, что современные ЭВМ позволяют работать непосредственно с именами категорий, а не с условными их номерами. Следовательно, в тех задачах, где это возможно, в память машины можно вводить сразу текстовые данные.

Подготовленные и закодированные данные записываются затем на стандартные бланки, и сотрудники вычислительного центра, работающие на специальных устройствах — перфораторах, переносят эти данные с бланков на перфокарты. На рис. 28 изображена перфокарта — один из основных носителей информации для ЭВМ. Комбинации отверстий в 80 колонках перфокарты обозначают любые цифры, буквы и другие символы из предусмотренного в вычислительной системе набора символов. Над пробитыми в колонках отверстиями на перфокарте печатаются соответ-

Рис. 28. Образец перфокарты

ствующие символы, что облегчает проверку правильности перфорации.

Заметим, что на этом этапе возможны ошибки как вследствие неверного перенесения исходных данных на бланки (описки, пропуски, неверная кодировка), так и вследствие неправильной перфорации содержания бланков, поэтому требуется много времени на неоднократную проверку и исправление ошибок. Историку можно научиться перфорировать и самому, это полезно при исправлении ошибок и работе с количественными данными. Однако при большом объеме информации высокая скорость и аккуратность персонала вычислительного центра делают его помощь незаменимой.

Проверенный массив перфокарт (колоду) можно вводить в ЭВМ и обрабатывать с помощью соответствующих программ. Перфокарты являются, однако, не единственным способом записи и хранения иформации для ЭВМ. Данные можно записывать на перфоленты, магнитные барабаны, ленты и магнитные диски, а также вводить непосредственно с дисплея (телевизора). Хранение данных на магнитных лентах и дисках имеет ряд существенных преимуществ: большая скорость ввода и компактность записи (информация, занимающая большую колоду перфокарт, может разместиться на маленьком участке магнитной ленты). Дисплейная техника позволяет облегчить задачи поиска нужной информации и внесение в нее изменений, роль дисплеев повышается и в связи с созданием банков информации, т. е. своеобразных архивов данных, обеспеченных системами организации информации, поиска и обработки (информационно-поисковые системы).

4. Выбор математического метода решения поставленной проблемы и, следовательно, вид обработки данных связан с формализацией содержательной задачи, моделированием. Так, если изучаются и выявляются связи признаков, методом может быть корреляционный, регрессионный, факторный энтропийный анализ, если изучается структура совокупности объектов — факторный, кластерный анализ, если изучаются процессы, развивающиеся во времени, — анализ динамических рядов и т. д.

Правильный выбор метода требует от исследователя знакомства с основными принципами, лежащими в основе конкретных методов, и их ограничениями. Иногда метод, вполне пригодный для изучения данной модели, может дать неверные или бессмысленные результаты потому, что не выполнено какое-либо из условий его применения. Так,

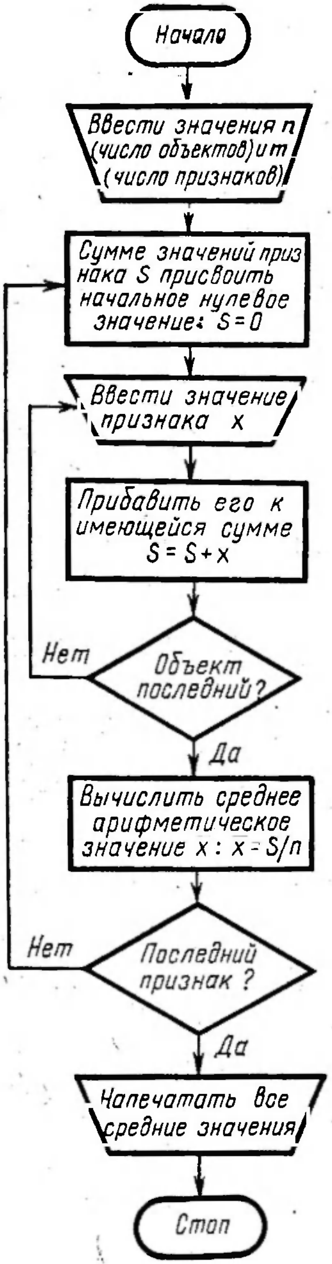

Рис. 29. Пример блок-схемы

линейная корреляция не может дать ответа на вопрос о наличии связи в общем смысле, а «естественная» выборка не обязательно является репрезентативной [1].

5. На следующем этапе исходя из математической постановки задачи создается алгоритм ее решения. Этот алгоритм иногда изображают в виде блок-схемы, графически отражающей ход решения задачи. На рис. 29 дан пример блок-схемы алгоритма вычисления средних арифметических значений признаков некоторой совокупности объектов (разумеется, обычно блок-схемы составляются для гораздо более сложных программ).

Из вида приведенной выше блок-схемы ясно, что все действия ЭВМ могут быть подразделены на четыре типа. Это ввод и вывод информации, ее хранение или запоминание, обработка информации (в частности, арифметические вычисления) и, наконец, управление последовательностью выполнения различных операций. Конкретная вычислительная система физически состоит из трех типов устройств, призванных выполнять все эти операции. Схематически вычислительная система или конфигурация ЭВМ изображена на рис. 30.

Таким образом, вычислительная система представляет собой центральный процессор, т. е блок, содержащий схемы, управ-

[1] Принципы применения математических методов подробно рассмотрены в части первой данного пособия.

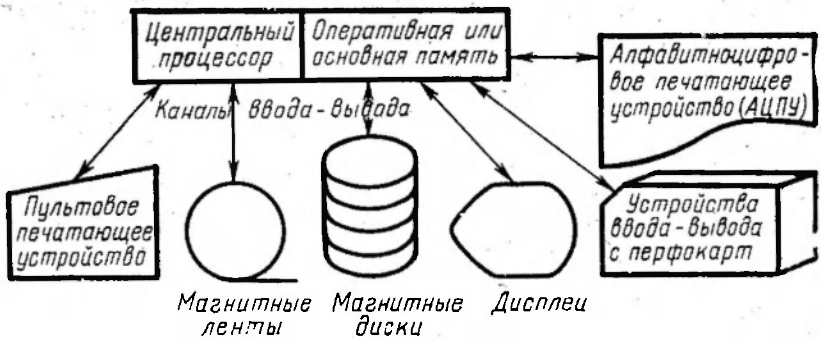

Рис. 30. Конфигурация ЭВМ

ляющие расшифровкой и выполнением команд, с подключенными к нему основной быстродействующей памятью (оперативной), внешними запоминающими устройствами (магнитные ленты, диски) и устройствами ввода-вывода (дисплеи, устройство ввода с перфокарт, устройство построчной печати).

Для того чтобы вычислительная установка работала, необходимо программное обеспечение, называемое операционной системой. Операционная система управляет выполнением программ, вводом-выводом, оценкой затраченных ресурсов для расчетов с пользователями, распределением памяти и времени, организацией данных и др.

Для работы с ЭВМ составленный (например, в виде блок-схемы) алгоритм реализуется на одном из языков программирования, т. е. на одном из специальных языков, на которых создаются программы. Язык программирования состоит из утверждений, напоминающих предложения в обычном языке и описывающих выполнение последовательных шагов алгоритма. Так называемые языки высокого уровня (АЛГОЛ, ФОРТРАН, ПЛ) являются очень гибкими по своим возможностям и пригодны для решения самых различных проблем, в частности для задач количественного анализа исторической информации.

Однако непосредственно по программе, написанной на языке программирования, ЭВМ не работает; операционная система «переводит» (транслирует) эту исходную программу на «внутренний» язык машины — язык двоичных кодов, но это происходит без участия программиста, который всегда работает с исходной программой, внося в нее нужные исправления или изменения.

Написанная программа должна быть отлажена, т. е. в ней надо выявить и устранить возможные ошибки как логического, так и формального характера. Затем отлаженная программа тестируется — проверяется на небольшом массиве данных, и результаты сопоставляются с полученными вручную или с помощью калькулятора результатами обработки тех же данных.

Хотя процесс создания программы на языке высокого уровня мало связан с конкретной машиной, однако он невозможен без знания операционной системы, управляющей работой ЭВМ, «библиотечных» программ, доступных пользователю, и многих других компонентов конкретной вычислительной системы.

На этапах формализации исследовательской гипотезы и математической постановки задачи к работе обычно привлекаются математики и программисты, которые либо создают программу специально для данной задачи, либо могут предложить для ее решения одну или несколько из уже существующих программ. В частности, все более широкое распространение получают пакеты прикладных программ для определенного круга математических методов, например пакеты программ статистической обработки. Это значительно облегчает контакт с ЭВМ для неспециалистов, однако следует учитывать, что пакетные программы нередко накладывают чрезмерно жесткие ограничения на вид обрабатываемой информации (например, не обрабатывают данные с неполной информацией), поэтому помощь программистов остается необходимой.

После того как исходные данные и программа готовы, операционная система вводит и транслирует программу и в соответствии с этой программой вводит, обрабатывает информацию и выводит результаты обработки либо на широкую бумажную ленту, либо на перфокарты, магнитные ленты (диски) или экран дисплея.

6. Наконец, на этапе интерпретации исследователь проверяет свою гипотезу (гипотезы) на основании полученных результатов и строит новые, более сложные гипотезы и модели, вновь требующие помощи ЭВМ. Таким образом, количественный анализ и ЭВМ — лишь один из инструментов исследования, а качество и надежность результатов есть следствие достоверности данных и корректности применения количественных методов. Лишь при этом высокая скорость и точность вычислений, возможность обработки огромного объема информации и реализации сложнейших математических методов окупают трудности и большие затраты времени и сил на применение количественного анализа в истории.

3

ОСНОВНЫЕ НАПРАВЛЕНИЯ ПРИМЕНЕНИЯ КОЛИЧЕСТВЕННЫХ МЕТОДОВ В ИСТОРИЧЕСКИХ ИССЛЕДОВАНИЯХ

В настоящем разделе работы во вводной его части дается сжатый обзор основных направлений применения количественных методов в исследованиях советских историков, а в соответствующих главах рассматриваются конкретные примеры применения этих методов при изучении различных исторических явлений и процессов. Цель этого рассмотрения состоит в том, чтобы, с одной стороны, показать эффективность применения количественных методов при исследовании многих явлений исторического прошлого, а с другой — на конкретных примерах раскрыть весь механизм этого применения, начиная с постановки исследовательской задачи и кончая интерпретацией результатов количественного анализа и формулированием выводов. Тем самым будет выявлена ведущая роль теоретико-методологических принципов при применении количественных методов, сущность системного, целостного подхода к изучаемым явлениям и процессам, методика и техника количественного анализа.

Естественно, число приводимых примеров является ограниченным. Они не исчерпывают всего богатства и многообразия полученных результатов и по ним нельзя судить об общем размахе и итогах применения количественных методов в исследованиях советских историков.

Интерес к возможностям использования математических методов и применения электронно-вычислительной техники в исторических исследованиях проявился на рубеже 50—60-х годов как у историков нашей страны, так и за рубежом — в США, европейских капиталистических странах —

Швеции, ФРГ, Франции, некоторых социалистических — ГДР, ПНР. С этого времени не только стремительно растет число исследований с применением современного математического аппарата и ЭВМ, но возникают специальные научные центры, объединяющие исследователей этого профиля, проводятся международные конференции по вопросам применения количественных методов и ЭВМ в исторических исследованиях, в ряде стран начинают выходить специальные периодические издания. В СССР наиболее активно ведутся исследования с применением математических методов и ЭВМ историками Москвы, Ленинграда, Таллина, Тарту, Новосибирска. Ведущую роль здесь играют Институт истории СССР АН СССР, исторический факультет МГУ, Институт истории АН Эстонской ССР. Работа в этом направлении координируется Комиссией по применению математических методов и электронно-вычислительных машин в исторических исследованиях при Отделении истории АН СССР.

Основным центром исследований по так называемой «количественной истории» в США является межуниверситетский консорциум по политическим исследованиям в Анн-Арборе, организованный в 1962 г., а также Центр политических наук Института социальных исследований, организационно связанный с Американской исторической ассоциацией и Американской ассоциацией политических наук.

Из европейских стран в 1965 г. в Швеции при отделе истории Уппсальского университета была создана исследовательская группа под руководством К. Г. Андре и С. Лундквиста. В ФРГ в Кельнском университете образован центр историко-социальных исследований, на базе которого создана Международная комиссия по применению количественных методов в истории [1].

---

[1] Подробнее о применении количественных методов историками см.: *Палли Х. Э.* Применение математических методов в исследованиях скандинавских историков.— В кн.: Математические методы в исторических исследованиях. М., 1972; *Промахина И. М.* Количественные методы в работах представителей «новой экономической истории» (США). — В кн.: Математические методы в исследованиях по социально-экономической истории; *Селунская Н. Б.* Количественная история в США: итоги, проблемы, дискуссии. — В кн.: Математические методы в историко-экономических и историко-культурных исследованиях. М., 1977; *Соколов А. К.* О применении новых методов в исследованиях историков США. — В кн.: Математические методы в социально-экономических и археологических исследованиях. М., 1981; *Воронцов Г. А.* Некоторые новейшие направления в буржуазной историографии ФРГ. — Вопросы истории, 1974, № 9.

В 60-е годы в нашей стране наметились основные направления применения математико-статистических методов в исторической науке. Это прежде всего работы, в которых ставились общие вопросы о возможностях и пределах применения количественных методов, их месте в исторической науке [1].

Первые опыты применения ЭВМ и математико-статистических методов в конкретно-исторических исследованиях были связаны с обработкой локальных историко-статистических данных по социально-экономической, прежде всего аграрной истории России [2].

70-е годы знаменуют собой новый этап в развитии исторических исследований с применением математико-статистических методов и ЭВМ.

Отличительными чертами современного этапа развития исследований с применением количественных методов в нашей стране является не просто рост их числа, расширение сферы применения, разнообразие и совершенствование методических приемов и средств, но прежде всего постановка таких исследовательских задач, которые трудно или вообще невозможно решить традиционными методами.

В центре внимания советских историков по-прежнему остается изучение социально-экономической истории нашей страны. Ведущее место здесь занимают исследования аграрной истории России. Главной их целью является целостное раскрытие внутренней сути и механизма изучаемых явлений, а не отдельные частные наблюдения о тех или иных сторонах его и чертах. Ведущим методом достижения этой цели стало моделирование. Так, в итоге изучения внутреннего строя крестьянского и помещичьего хозяйства на основе обработки массивов статистических данных (материалов писцовых книг, земско-статистических обследований, сельскохозяйственных и поземельных переписей и др.) с помощью методов корреляционного и регрессионного анализа советскими историками получены модели социально-эконо-

[1] См.: *Ковальченко И. Д.* О применении математико-статистических методов в исторических исследованиях. — В кн.: Источниковедение. Теоретические и методические проблемы. М., 1969; *Кахк Ю. Ю.* Нужна ли новая история. — Вопросы истории, 1969, № 3, и др.

[2] См.: *Кахк Ю. Ю., Лиги Х. М.* К вопросу об экономическом положении и феодальных повинностях крестьян в Эстляндской губернии в XVIII в. (Опыт применения электронно-счетных машин в историческом исследовании). — Ежегодник по аграрной истории Восточной Европы. 1962. Минск, 1964; *Ковальченко И. Д., Устинов В. А.* О применении ЭВМ для обработки историко-статистических материалов. — Вопросы истории, 1964, № 1; *Ковальченко И. Д.* Русское крепостное крестьянство в первой половине XIX века. М., 1967, и др.

мической структуры этих хозяйств России XVII—XX вв., выявлены особенности их функционирования и взаимодействия [1].

При этом использовались не только первичные материалы о крестьянском и помещичьем хозяйстве, которые позволяют выделить различные типы этих хозяйств, но и разного рода сводные данные, которые обычно признавались непригодными для анализа внутреннего, социально-экономического строя крестьянского и помещичьего хозяйства. Например, первичные материалы земских подворных описаний крестьянского хозяйства позволяют проводить группировку крестьянских дворов по тем или иным признакам и анализировать внутренний строй хозяйства этих групп. Однако подавляющее большинство гигантского материала земской статистики конца XIX — начала XX в. опубликовано в виде сводок по селениям (общинам), волостям и уездам. Сводный характер данных имеют и другие ценные источники по истории крестьянства России эпохи капитализма (например, сельскохозяйственные переписи 1916 и 1917 гг.). Поэтому историки стоят перед задачей необходимости разработки методов анализа сводных данных для выявления степени буржуазно-капиталистической эволюции крестьянского хозяйства. Применение количественных методов, в частности корреляционного и регрессионного анализа, позволяет вполне успешно решать эту трудную задачу.

Значительный интерес представляют работы эстонских историков по аграрной истории Эстонии и эстонского крестьянства. В их трудах содержится разработка многих математических методов, начиная от организации достаточно представительных выборок, приемов корреляционного и регрессионного анализа до математических способов классификации крестьянских хозяйств и применения теории решений при анализе мотивов аграрной политики. Математические методы нашли также применение в изучении средневековья [2].

---

[1] См.: Математические методы в исторических исследованиях. Математические методы в исследованиях по социально-экономической истории; Математические методы в историко-экономических и историко-культурных исследованиях; Математические методы в социально-экономических и археологических исследованиях.

[2] См.: *Кахк Ю. Ю.* О работе историко-математической группы Института истории АН ЭССР. — В кн.: Математические методы в историко-экономических и историко-культурных исследованиях. *Хвостова К. В.* Количественный подход в средневековой социально-экономической истории. М., 1980; *Бессмертный Ю. Л.* Математические методы и их применение при исследовании проблем средневековья.— В кн.: Средние века. М., 1971, вып. 34.

В 70-е годы широкое распространение получили математические методы при изучении аграрных отношений и рынка советской деревни 20-х годов[1].

Распространение новейших приемов исследования повысило требовательность историков к достоверности, репрезентативности исторических данных. Применение математических методов позволило повысить информативную отдачу источников, обеспечить высокий уровень их критического анализа. В этой связи все большее внимание уделяется выборочному методу в исторических исследованиях.

Развитие историографической традиции применения математико-статистических методов и ЭВМ в истории проявляется и в том, что в сферу количественного анализа вовлекаются не только статистические, но и нарративные, текстовые материалы. Так, применение математических методов и ЭВМ в текстологии позволило не только проверить надежность оценок, сделанных на основе применения традиционных методов, но и получить ряд новых интересных выводов и наблюдений.

Поиск новых эффективных методов характерен и для исследователей социально-политической истории. Здесь следует отметить попытки применения новых методов к анализу таких сложных явлений в историческом развитии, как классовая и политическая борьба, аграрная политика и др.[2]

Традиционным направлением применения количественных методов являются историко-демографические исследования. Наиболее успешно в этом направлении ведут исследования эстонские историки[3].

---

[1] См.: *Мошков Ю. А.* Опыт применения корреляционного анализа в изучении отчетов о распределении колхозных доходов в начале массовой коллективизации. — В кн.: Источниковедение отечественной истории. М., 1973, вып. 1; *Миняйло Н. Г.* Опыт применения корреляционного анализа при обработке крестьянских бюджетов. — В кн.: Математические методы в исследованиях по социально-экономической истории. *Пушков В. П., Промахина И. М.* Опыт применения факторного анализа для классификации изучения структуры и моделирования социальных типов крестьянских хозяйств... — В кн.: Математические методы в социально-экономических и археологических исследованиях.

[2] См.: *Кахк Ю. Ю., Лиги Х. М.* О связи между антифеодальными выступлениями крестьян и их положением. — История СССР, 1976, № 2; *Лабутина Т. Л.* Применение ЭВМ при исследовании формирования политических партий «тори» и «вигов». — В кн.: Математические методы в историко-экономических и историко-культурных исследованиях; *Деопик Д. В.* Опыт количественного анализа древней восточной летописи «Чуньцю». — Там же.

[3] См.: *Палли Х.* Естественное движение сельского населения Эстонии (1650—1799). Таллин, 1980, вып. 1—3.

Все более широкое распространение получают количественные методы в археологии, что обусловлено наличием здесь массовых комплексов вещественных исторических памятников[1].

Применение количественных методов при изучении конкретно-исторических явлений и процессов, естественно, требовало решения целого ряда теоретико-методологических проблем их применения, в том числе и вопросов о месте и роли количественных методов в исторической науке[2].

Существенной чертой современного этапа является внимание советских историков к изучению зарубежного опыта применения математико-статистических методов в исторических исследованиях как в плане критического анализа теоретико-методологических посылок, так и осмысления опыта зарубежной историографии по разработке методических приемов и средств обработки и анализа исторической информации.

В современной буржуазной историографии, переживающей глубокий идейно-методологический кризис, отчетливо прослеживается тенденция к поискам передовыми учеными выхода из этого кризиса.

Эти поиски ведутся прежде всего в области методологии исторического познания и методов исторического исследования[3]. Средством, которое якобы позволяет сделать историю подлинно научной, буржуазные историографы считают научно-естественные и прежде всего количественные методы исследования. Однако любые методы исторических исследований сами по себе не могут дать выхода из этого кризиса. Успех применения любых, в том числе и количественных, методов, как было показано, определяется возможностями той теории и методологии исторического познания, на которых базируется применение этих методов.

---

[1] См.: *Квирквелия О. Р.* Краткий обзор советской литературы по вопросам применения математико-статистических методов исследования в археологии. — В кн.: Математические методы в социально-экономических и археологических исследованиях.

[2] См.: *Кахк Ю. Ю., Ковальченко И. Д.* Методологические проблемы применения количественных методов в исторических исследованиях. — История СССР, 1974, № 5; *Хвостова К. В.* Методологические проблемы применения математических методов в исторических исследованиях. — Вопросы истории, 1975, № 11; *Ковальченко И. Д.* О моделировании исторических явлений и процессов. — Вопросы истории, 1978, № 8, и др.

[3] См.: *Ковальченко И. Д., Сивачев Н. В.* Структурализм и структурно-количественные методы в современной исторической науке. — История СССР, № 5, с. 59.

Ограниченность и ошибочность тех теории и методологии, на основе которых применяют количественные методы буржуазные историки, препятствуют эффективному их использованию. Поэтому наиболее ценным в этом опыте являются не конкретно-исторические результаты, а методические и технические приемы и банки машинной исторической информации, созданные при этом. Все это хорошо видно на примере американской историографии.

Как признают сами американские исследователи, новое поколение историков-клиометристов сконцентрировало прежде всего и более всего свои усилия на сборе данных, розыске новых и разработке малоизученных источников, создании синтезированных данных, а также разработке единых программ обработки и хранения исторической информации. Убедительным свидетельством этому является богатая коллекция данных, «банк машинной информации» в Анн-Арборе, который содержит около 11 млн. карточек данных по итогам выборов в конгресс, списков голосования и других источников по истории США.

По типу коллекции в Анн-Арборе создаются различные машинные архивы исторических данных по социальной истории, особенно «новой» городской истории, содержащие сведения промышленных переписей, переписей населения второй половины XIX—XX вв., аграрной истории Америки (коллекции данных по Югу и Северу США, содержащие сведения земельных цензов, переписей населения и другие источники). Важным событием американской историографии в этой области является создание «Проекта социальной истории Филадельфии» [1], составной частью которого является архив исторических данных, введенных в ЭВМ и описывающих профессиональную карьеру 2,5 млн. жителей Филадельфии второй половины XIX в. Комплекс программ, обслуживающих банк данных, включает программы по сверке и корректировке исходных данных, преобразование этих данных, программы статистического анализа и выдачи результатов в виде статистических таблиц, карт, диаграмм и т. д.

Методика и техника создания машинных банков исторической информации заслуживает внимания и изучения.

Что касается характера и уровня исторических исследо-

---

[1] Подробнее см.: *Бородкин Л. И., Селунская Н. Б.* Методы изучения социальной истории в американской историографии (по поводу «Проекта социальной истории Филадельфии»). — История СССР, 1978, № 2.

ваний, то здесь показательным является то, что представители «количественной истории» в США не только расщепляют исследование исторического процесса на отдельные элементы, но и стремятся исключить изучение базисных явлений из сферы исторического исследования. Большая же часть исследований по экономической истории также носит характер эмпирических исследований (рост прибыли, распределение доходов, спекуляция землей и др.).

Отдельными исследователями предпринимались попытки более широких концепционных исследований. Но они базировались либо на построении неправомерных для исторических исследований контрфактических моделей экономического развития[1], либо на ограниченном теоретико-методологическом подходе к изучаемым экономическим явлениям[2].

Неубедительность подобных концепционных построений ощущают и сами буржуазные историки. Этим объясняются призывы со стороны самих клиометристов, с одной стороны, к поискам некоей общей, противостоящей марксизму теории исторического познания, а с другой — к сближению «количественной» исторической науки с «традиционной», главная цель которой состоит якобы в поиске морально-этических причин и результатов человеческого поведения.

Таковы основные направления применения количественных методов в исторических исследованиях. В последующих главах дается конкретный разбор наиболее типичных из возникающих задач, решение которых дало эффективные исследовательские результаты.

ГЛАВА 12
КОЛИЧЕСТВЕННЫЕ МЕТОДЫ В ИЗУЧЕНИИ СОЦИАЛЬНО-ЭКОНОМИЧЕСКИХ ЯВЛЕНИЙ И ПРОЦЕССОВ

Как известно, исследования по социально-экономической истории по праву занимают ведущее место в советской историографии, ибо они направлены на изучение базисных

---

[1] Примером построения такой модели является уже рассматривавшаяся работа Р. Фогеля «Железные дороги и экономический рост Америки».

[2] Наиболее ярким примером здесь является работа Р. Фогеля и Ст. Энгермана «Время на кресте», в которой авторы пытаются доказать большую экономическую эффективность в середине XIX в. рабского труда сравнительно со свободным вольнонаемным трудом. См.: *Ковальченко И. Д., Сивачев Н. В.* Структурализм и структурно-количественные методы, с. 79—82.

явлений, составляющих основу системы общественных отношений. Объектом исследований по социально-экономической истории является все то, что выражает собой объективный и внутренне обусловленный характер исторического развития, все то, что характеризует общественную жизнь как совокупность различных систем с присущими им структурами и как процесс, выражающийся в определенной динамике этих систем. Подобные исследования, как правило, базируются на изучении массовых объектов и явлений и непременно включают в себя обобщающие заключения, раскрывающие внутреннюю суть рассматриваемых общественных систем. Массовый характер явлений и процессов социально-экономического развития, естественно, обусловил и накопление массовых данных, в том числе массовых статистических источников по социально-экономической истории. И только на основе широкого и всестороннего изучения этих данных можно раскрыть содержание и ход, внутреннюю сущность, закономерности и особенности социально-экономического развития как в целом, так и в его отдельных аспектах. Подчеркивая значимость использования массовых исторических источников для изучения социально-экономической истории нового и новейшего времени, В. И. Ленин писал, что «целый ряд вопросов и притом самых коренных вопросов, касающихся экономического строя современных государств и его развития, которые решались прежде на основании общих соображений и примерных данных, не может быть разрабатываем сколько-нибудь серьезно в настоящее время без учета массовых данных, собранных относительно всей территории известной страны...» [1]

Таким образом природа объекта социально-экономической истории, характер данных обусловливают возможность и диктуют необходимость применения количественных методов при анализе социально-экономических процессов. Этим определяется тот факт, что количественные методы прежде всего применяются при изучении этой сферы общественного развития.

Исследование социально-экономических объектов, явлений и процессов, как и любое историческое исследование, включает в себя такой важный этап, как формирование его фактической основы — источниковой базы. От того, насколько достоверны, точны и представительны исторические

---

[1] *Ленин В. И.* Полн. собр. соч., т. 19, с. 323.

данные, т. е. насколько они верно и всесторонне отражают свойства, характерные черты изучаемого явления, объекта, процесса, зависят итоговые заключения, конкретно-историческая оценка сущности данного явления, объекта, процесса, а следовательно, и научная и социальная значимость всего исследования. Вот почему одной из важнейших задач является задача дальнейшего совершенствования методики источниковедческой критики привлекаемых исторических данных. Применение количественных методов, в том числе выборочного метода и методов анализа взаимосвязей (корреляционного анализа), позволяет уже теперь сделать значительный шаг вперед в решении источниковедческих проблем социально-экономической истории.

Применение количественных методов на современном этапе развития исторической науки доказало свою эффективность в решении ряда кардинальных аспектов социально-экономической истории. Как известно, целью исследований по социально-экономической истории является выявление конкретно-исторических закономерностей, характера и особенностей эволюции производительных сил и производственных отношений (и их отдельных компонентов) различных стран и регионов в определенные исторические периоды развития общества. Раскрытие этих закономерностей предполагает всесторонний анализ сложных социально-экономических явлений и объектов, исследование взаимосвязей элементов, входящих в их состав, т. е. ставит задачу анализа социально-экономических объектов и явлений. При решении этой задачи весьма эффективным исследовательским аппаратом являются методы количественного анализа, в том числе применение корреляционного, регрессионного и факторного анализа, построение количественных моделей социально-экономических явлений и процессов.

Исследование социально-экономической истории неотъемлемой частью имеет сравнительный анализ многообразия социально-экономических объектов, явлений и процессов, установление общего и особенного, выявление стадий и форм эволюции элементов социально-экономического развития, т. е. требует типологической классификации социально-экономических объектов и явлений. Математико-статистические методы и ЭВМ, прежде всего методы многомерного анализа, позволяют решить эту сложнейшую и кардинальную проблему на принципиально новом уровне.

Наконец, применение методов количественного анализа и ЭВМ доказало свою перспективность в изучении динамики социально-экономических явлений и процессов, раскрытии изменений в конкретно-историческом содержании и формах социально-экономического развития.

Таковы основные аспекты применения количественных методов и ЭВМ в исследованиях по социально-экономической истории. Рассмотрим их подробней на примере решения конкретных исследовательских задач.

§ 1. Количественные методы в источниковедении массовых источников по социально-экономической истории

Источниковедческий анализ массовых источников по социально-экономической истории имеет своей основной целью установление, во-первых, достоверности и точности и, во-вторых, репрезентативности количественных данных, зафиксированных в этих источниках.

Довольно часто исследователь сталкивается с тем, что те или иные количественные сведения содержатся в двух или более источниках. Естественно, возникает вопрос об их сравнительной достоверности и точности. Обычно при решении этого вопроса прибегают к сравнению средних арифметических значений признаков, полученных по данным разных источников. Такое сравнение необходимо и оно должно основываться на математико-статистических методах выявления существенности различий средних значений, а не на субъективной оценке различий, как это обычно бывает. Однако сравнение средних значений недостаточно для выявления степени сходства соответствующих рядов данных, ибо оно не показывает степени сходства вариации признаков в сравниваемых рядах. Выявить сопряженность данных в разных рядах можно путем вычисления коэффициентов корреляции.

Указанными методами еще в начале XX в. проверялось сходство данных об урожайности хлебов в России в конце XIX — начале XX в., которые собирались Центральным статистическим комитетом, Министерством земледелия и земствами. Так, корреляция погубернских данных об урожайности, зафиксированных ЦСК и Министерством земледелия в 1885—1908 гг., была очень высокой (по всем губерниям коэффициент превышал 0,90). Очень тесной была и взаимосвязь данных ЦСК и земств[1]. Итак, данные всех трех независимых источников рисуют одинаковую картину.

---

[1] Подробнее см.: Массовые источники по социально-экономической истории России периода капитализма. М., 1979.

Следовательно, все они отличаются высокой достоверностью.

Однако исследователи нередко сталкиваются с такой ситуацией, когда сопряженность показателей разных источников отсутствует, а также различаются и их средние значения. В этом случае либо достоверным является один из источников, либо оба они недостоверны и возникает проблема проверки этой альтернативы. Такая проблема возникает и в том случае, когда сведения о тех или иных явлениях и процессах содержатся только в одном источнике.

Системный подход и количественные методы позволяют решить эту задачу. Принципы этого решения таковы. Всякое явление общественной жизни представляет собой составной компонент более широкой системы. При нормальном функционировании системы все ее компоненты сбалансированы, т. е. находятся во взаимосвязи. Это дает возможность выявить достоверность данных того или иного источника путем включения этих данных в совокупность других показателей, характеризующих систему, в которую входит и изучаемое явление.

Если данные о том или ином явлении вписываются в совокупность других показателей, т. е. взаимосвязаны с ними, то они являются достоверными. В противном случае они недостоверны. Поясним это на примере. Так, историки располагают погубернскими данными о количестве наемных сельскохозяйственных рабочих в европейской России, зафиксированных переписью 1897 г. и обследованием 1901 г. Между собой эти данные не взаимосвязаны (коэффициент корреляции равен 0,01) и решить вопрос об их достоверности можно лишь методом включения их в систему других показателей об аграрном развитии (см. табл. 1)

Таблица 1. Корреляционная взаимосвязь обеспеченности сельскохозяйственными рабочими с другими факторами социально-экономического развития

Факторы	Перепись 1897 г.	Данные 1901 г.
Хозяйства с наймом, % к общему числу	0,75	—0,01
Грамотных, %	0,81	—0,09
Посевы (дес. на душу сельского населения	0,01	0,57
Продуктивный скот, голов на душу населения	0,33	0,20
Урожайность зерновых, пудов с дес.	0,28	—0,3[illegible]

Как видим, данные о наемном труде переписи 1897 г. совсем не связаны с размерами посевов, т. е. относительные размеры земледельческого производства не определяли степени применения наемного труда. Имела место слабая взаимосвязь применения наемного труда с урожайностью и обеспеченностью продуктивным скотом и очень тесная — с долей хозяйств, применяющих наемный труд (0,75), т. е. с распространением капитализма вширь, с грамотностью населения (0,81), которая отражает общий уровень буржуазно-капиталистического развития. В общем сведения 1897 г. довольно органически вписываются в систему других показателей.

Иное положение с данными 1901 г. Они довольно тесно взаимосвязаны с размерами посевов и слабо с обеспеченностью скотом. Это обусловлено тем, что данные 1901 г. о наемных рабочих исчислялись исходя из учета их потребности в земледелии и животноводстве. С долей хозяйств с наймом и грамотного населения сведения 1901 г. совсем не связаны, а с урожайностью имеют даже обратную связь. В целом данные 1901 г. не вписываются в систему других показателей, с которыми по внутренней природе явления они должны быть взаимосвязаны.

Следовательно, достоверную картину различий между губерниями в обеспеченности сельскохозяйственными наемными рабочими рисуют данные переписи 1897 г. Сведения же 1901 г. являются в этом отношении недостоверными.

Подобным образом может быть проверена достоверность и многих других количественных показателей, т. е. могут решаться важные и трудные источниковедческие задачи.

Другой острой проблемой, стоящей при изучении массовых исторических явлений и процессов, является проблема представительности характеризующих их конкретно-исторических данных.

Представительность конкретных данных имеет, как уже говорилось в первой части работы, два аспекта: качественный (сущностно-содержательный) и количественный.

Качественная представительность данных источника определяется тем, в какой мере они отражают наиболее существенные черты и свойства изучаемых явлений и процессов, т. е. позволяют раскрыть их внутреннюю суть. Она определяется теоретико-содержательным анализом изучаемого явления и тех данных, которые содержат источники, привлекаемые для анализа этого явления.

Проблема количественной репрезентативности возникает в том случае, когда исследователь должен судить об исследуемых явлениях и процессах на основе данных, которые охватывают не всю совокупность составляющих их объектов, а лишь их часть. При этом выборка используемых данных, как известно, либо формируется самим исследователем из сведений генеральной совокупности на основе принципов и требований выборочного метода, либо представляет собой естественную выборку, объем которой не может быть изменен, а ее представительность должна быть проверена.

Соблюдение исследователем требований выборочного метода при формировании выборки из данных генеральной совокупности и проверка представительности используемых естественных выборок — важнейший источниковедческий этап, который обеспечивает создание фундамента из точных и бесспорных фактов, необходимых для анализа любых явлений и процессов.

Применение выборочного метода дает эффективные исследовательские результаты при оперировании большими объемами массовых исторических источников, сплошная обработка которых затруднительна. Серия выборок позволяет здесь получить репрезентативные данные для характеристики всей совокупности изучаемых объектов. Ярким примером успешного решения важной исследовательской задачи на основе выборочного метода является изучение социальной структуры советского рабочего класса по данным профессиональной переписи 1918 г. [1]

Первичные материалы профессиональной переписи 1918 г. составляют около 1 млн. личных карточек рабочих, в которых содержатся десятки различных демографических, социально-экономических и других характеристик рабочего класса, в том числе связанные с внутриклассовой дифференциацией рабочих (стаж, преемственность фабрично-заводского труда, связь с землей, источники пополнения рабочего класса и др.). Очевидно, что работать с таким массивом можно было лишь на основе выборочного метода.

Профессиональная перепись, проведенная ЦСУ осенью — зимой 1918 г., охватывала 31 губернию, на территорию которых в то время распространялась Советская власть, и включала полностью или частично 8 из 14 экономико-гео-

---

[1] См.: *Дробижев В. З., Соколов А. К., Устинов В. А.* Рабочий класс Советской России в первый год диктатуры пролетариата. М., 1975.

графических районов: Центрально-промышленный; Центрально-земледельческий, Средневолжский, Нижневолжский, Приуральский, Северный, Приозерный, Западный. Из каждого района методом направленного отбора выбиралась типическая единица — губерния, рабочие которой и представляли в выборке рабочих всего района. Это не математический, а чисто содержательный способ, когда единицы извлекаются из типического района с помощью определенной системы доводов, что обеспечивает большую репрезентативность выборке. Основой для выбора губерний явился анализ конкретно-исторической специфики развития рабочего класса отдельных губерний района, соотношение общих, характерных для рабочего класса района, черт и специфики по губерниям. Итогом конкретно-исторического анализа условий и особенностей формирования и развития промышленного пролетариата различных районов России был выбор губерний, представляющих эти районы. Например, материалы Ярославской губернии явились основой для характеристики рабочего класса развитого промышленного района, ибо в ней представлено 17 из 22 тогдашних групп производств фабрично-заводской промышленности, причем преобладает профессиональная группа текстильщиков, что было характерно для промышленного района. Материалы Воронежской губернии были признаны характерными для изучения рабочего класса сельскохозяйственного Центрально-земледельческого района.

На следующем этапе методом случайного механического отбора формировались выборки по губерниям. Сочетание направленного отбора губерний из экономических районов и механической выборки анкет из этих губерний явилось наиболее эффективным методом выборочной обработки профессиональной переписи 1918 г. На этом этапе исследования необходимо было решить вопрос о доле механического отбора из материалов отдельных губерний.

Учитывая специфику задачи — необходимость анализа большого числа признаков, раскрывающих внутреннюю структуру рабочего класса, — проводилось специальное экспериментальное исследование, позволяющее решить задачу об объеме выборки применительно к различным районам страны с учетом конкретно-исторической специфики структуры рабочего класса. Для определения необходимой доли отбора, обеспечивающей достаточно точную оценку признаков, в рамках экспериментального исследования было решено сделать несколько выборок различного объема.

Таблица 2. Распределение рабочих по группам производств по переписи 1918 г. в Ярославской губернии (% от общего числа рабочих)

Группы производств	Выборки (средние значения)			Доверительные интервалы для средних значений		
	10%	5%	1%	10%	5%	1%
Горная и горнозаводская промышленность	2,1	2,3	1,5	1,7—2,5	1,5—3,1	0,3—2,4
Металлообработка и машиностроение	5,3	5,0	4,1	4,9—5,7	3,8—6,2	2,1—6,1
Обработка дерева	2,5	2,8	3,8	1,7—2,7	1,9—3,7	1,8—5,8
Пищевая промышленность	13,6	12,9	14,7	12,5—14,7	11,2—14,6	11,1—18,3
Обработка хлопка	37,1	35,4	34,3	35,6—38,6	33,2—37,6	29,4—39,4
Обработка льна	23,6	26,0	25,3	22,3—24,9	23,6—28,4	20,9—29,7

По Ярославской губернии были сформированы 10, 5 и 1%-ные выборки[1]. Исходя из анализа эмпирических распределений признаков в выборках разного объема определялся оптимальный вариант, обеспечивающий представительность анализируемых групп. Методика механической выборки применена была в данном случае не к однородному массиву генеральной совокупности, а к очень сложному, так как рабочий класс в переписи был распределен по многочисленным группам различных производств. Таким образом, встала задача проверить, как при одинаковой доле отбора в выборке представлены все группы производств. Итоги этой работы частично отражены в табл. 2.

Из нее видно, что все три выборки рисуют примерно одинаковую картину распределения рабочих по отраслям производства. Доказательством этого является пересечение доверительных интервалов для средних значений. Значения интервалов одной выборки включают значения интервалов других выборок. Доверительные интервалы меньшей выборки (1%-ной) вообще включают целиком доверительные интервалы более точной, т. е. более крупной выборки (10 и

[1] См.: *Соколов А. К.* Методика выборочной обработки первичных материалов профессиональной переписи 1918 г. — История СССР 1971, № 4.

5%-ной). Пересечение доверительных интервалов по всем или по подавляющему большинству групп производства свидетельствует о репрезентативности выборок не только в целом по губернии, но и по каждой группе производств. Кроме того, на основе сопоставления выборочных средних показателей по различным признакам (распределение по стажу, группам производств, профессиональным группам, возрасту) с опубликованными среднегубернскими значениями было сделано заключение о достаточном объеме выборок для отдельных губерний. Так, было установлено, что для губерний со слабым развитием промышленности и преобладанием в ее структуре нескольких отраслей оптимальной является выборка в 400—500 единиц. Для промышленно развитых губерний представительной является выборка в 4—5% от общего числа рабочих.

В итоге общий объем анкет, которые необходимо было обработать, составил 11 000 личных карточек рабочих, т. е. 1% их общего числа.

Изучение на этой основе социальной структуры рабочего класса России по отдельным районам страны показало органическое единство, близость различных прослоек между собой по основным экономическим, психологическим и политическим признакам, единство общих закономерностей формирования и развития рабочего класса.

Установление репрезентативности естественных выборок сводится к выявлению того, в какой мере эти выборки являются случайными. К сожалению, здесь историк не может опираться на такие же строгие математические методы установления их представительности, которые применяются при формировании выборки из генеральной совокупности. При установлении представительности естественных выборок определяющее значение имеет историко-содержательный анализ. Выявление истории возникновения источника и его последующей судьбы может прояснить, в какой мере сохранившиеся сведения являются случайной выборкой. Но некоторые важные аспекты проверки репрезентативности естественных выборок могут основываться и на количественных методах. Так, возможно применение количественных методов проверки случайности варьирования признака в выборках. Поскольку такое варьирование имеет место в случайных выборках, постольку выявление его может помочь решить вопрос о случайности естественной выборки, а следовательно, и ее представительности.

Возможность и эффективность выявления представительности данных естественных выборок с помощью математико-статистических методов проиллюстрируем некоторыми примерами.

Так, при изучении аграрного развития европейской России во второй половине XVIII — первой половине XIX в. важнейшим моментом является выяснение обеспеченности барщинного крестьянства пашенными угодьями. Сведения о наделе пашни у барщинных крестьян в середине XIX в. имеются по всем уездам. Для конца же XVIII в. такие сведения, содержащиеся в «Полных экономических примечаниях», относятся лишь к 18 уездам Центрально-Черноземного района и 13 уездам нечерноземной полосы. Выяснить, насколько данные указанных уездов отражают обеспеченность барщинных крестьян пашней в рассматриваемых районах, можно путем определения по имеющимся данным возможных предельных границ (доверительных интервалов), в которых могла колебаться действительная средняя обеспеченность крестьян пашней в каждом из районов. Однако чтобы получить интервальную оценку среднего значения признака для всей генеральной совокупности, необходимо проверить случайность имеющихся естественных выборок. Она подтверждается следующими моментами. Во-первых, сведения «Полных экономических примечаний» являются частью общего сплошного описания всех имений названных районов, т. е. генеральной совокупности. История сохранности материалов лишь по 31 уезду показывает чисто случайный характер «формирования» этой выборки. Все элементы генеральной совокупности имели равную возможность попасть в эту выборку. Во-вторых, «естественная выборка» данных о наделах крестьян равномерно охватывает генеральную совокупность, т. е. имеются сведения по нескольким уездам почти всех губерний, входящих в Центрально-Черноземный и Центрально-Нечерноземный районы. В-третьих, для определения правильности отражения в выборочных данных основных свойств генеральной совокупности проверялось, насколько различия между отдельными показателями выборки являлись случайными. Для этого использовался критерий знаков (см. табл. 3). Итоги проверки по Центрально-промышленному району показывают, что различия между отдельными показателями выборки в большей степени являются случайными. Все указанное свидетельствует о том, что сохранившиеся выборки являются случайными, т. е. представительными для выявления размеров надела крестьян в каждом из районов в целом. Сравнение наделов конца XVIII в. (а их размеры могли варьи-

Таблица 3. Проверка случайности выборки методом «критерия знаков»

Уезды	Надел пашни в XVIII в., дес.	Знаки разности	Уезды	Надел пашни в XVIII в., дес.	Знаки разности
Костромской	3,0		Смоленский	3,00	+
Нерехтский	3,10	—	Боровский	2,10	+
Кинешемский	2,30	+	Малоярославский	2,20	—
Дмитровский	1,40	+	Тарусский	2,00	+
Клиниский	2,50	—	Мосальский	2,20	—
Гжатский	3,80	—	Сергачский	2,80	—
Вяземский	3,40	+			

Число плюсов в выборке	6
Критические границы для числа плюсов [1]	3—10

[1] Вариация значений признака является случайной, если число плюсов не выходит за пределы критических границ.

роваться в определенном интервале) с наделами середины XIX в. позволило выявить важные моменты в состоянии хозяйства барщинных крестьян. С конца XVIII до середины XIX в. размеры надела пашни у барщинных крестьян обоих районов существенно сократились и равнялись минимальным их размерам, т. е. 2,0—2,5 дес. на душу. В середине XIX в. во многих губерниях нечерноземной полосы и в некоторых черноземных губерниях имелась значительная прослойка малоземельных крестьян. Большинству таких крестьян надел не давал даже необходимых им средств к существованию. Резкое сокращение крестьянских наделов к концу крепостнической эпохи свидетельствовало об отсутствии сколько-нибудь широких возможностей для успешного развития крестьянского хозяйства на основе феодально-крепостнических отношений, поскольку условием такого развития является наделение крестьян землей в размерах, допускающих расширение крестьянского хозяйства и повышение его уровня [1].

Как видим, количественные методы позволяют решать важные и трудные задачи по установлению достоверности и репрезентативности массовых количественных данных, содержащихся в источниках как по социально-экономической истории (о чем шла речь в данном разделе), так и по другим явлениям общественной жизни.

[1] См.: *Ковальченко И. Д.* Русское крепостное крестьянство в первой половине XIX века, с. 273—274.

В следующей главе будет показана эффективность количественных методов и при источниковедческом анализе нарративных источников.

§ 2. Анализ структуры социально-экономических явлений

Применение количественных методов при изучении структур социально-экономических явлений имеет определенную историографическую традицию. Отсутствие в источниках массовых систематических данных, характеризующих непосредственно социально-экономическое содержание изучаемых явлений, делает неизбежным изучение их путем структурного анализа, опирающегося на имеющиеся в источниках показатели. Конкретно задача сводится к выявлению взаимосвязей этих признаков и их содержательной интерпретации. Рассмотрим некоторые примеры.

Важное значение для раскрытия уровня развития аграрного капитализма в России второй половины XIX — начала XX в. имеет анализ внутреннего строя помещичьего хозяйства.

В пореформенную эпоху существовали две экономические формы организации помещичьего хозяйства: капиталистическая, основанная на собственной производственно-технической базе помещика (орудия, инвентарь, рабочий скот и др.) и эксплуатации свободной наемной рабочей силы, и отработочная, при которой помещик эксплуатировал закабаленных окрестных крестьян и использовал их рабочий скот и инвентарь (так называемые отработки первого рода). Эволюция помещичьего хозяйства шла по линии перехода от отработков к собственно капиталистическим способам ведения хозяйства. Однако в реальности в помещичьем хозяйстве в тех или иных сочетаниях переплетались и отработочная и капиталистическая системы[1]. Это делает практически невозможным выявление непосредственного соотношения отработочной и капиталистической систем в помещичьем хозяйстве. Для этого, как указывал В. И. Ленин, «...потребовался бы учет не только всех имений, но и всех хозяйственных операций во всех имениях»[2]. Такого учета не существовало. Более того, источники не содержат и систематических данных о соотношении отработочных и

---

[1] См.: *Ленин В. И.* Полн. собр. соч., т. 3, с. 187.
[2] Там же, с. 188.

капиталистических форм ведения хозяйства. Такие сведения имеются лишь в отношении немногих отдельных вотчин. В итоге сложилась весьма своеобразная ситуация, когда при обилии массовых данных о помещичьем хозяйстве периода капитализма (сведения земской статистики, земельных банков, сельскохозяйственных переписей 1916, 1917 гг.) историки не могли решить традиционными методами вопрос о характере социального строя помещичьего хозяйства России.

Решение этого вопроса потребовало совершенствования методов исследования — применения системного подхода и структурно-количественных методов анализа, т. е. перехода к целостному рассмотрению внутренней структуры помещичьего хозяйства на основе привлечения имеющихся в источниках сведений о различных сторонах организации помещичьего хозяйства, к созданию моделей его структуры.

Первый этап решения задачи состоял в четком теоретическом описании двух крайних вариантов организации пореформенного помещичьего хозяйства — капиталистического и отработочного (точнее такого, где преобладали бы отработки первого рода). Основой построения таких сущностно-содержательных моделей послужили общая марксистская характеристика капиталистического производства и ленинская характеристика пореформенного помещичьего хозяйства. Структура капиталистически организованного помещичьего хозяйства должна была характеризоваться взаимозависимостью и тесной сбалансированностью его компонентов, пропорциональностью в обеспеченности собственными средствами производства, потреблением рабочей силы, расходами и доходами. Структура же отработочного хозяйства в силу того, что оно преимущественно основывалось на производственно-технической базе крестьянского хозяйства, такой сбалансированности, пропорциональности не имела. Исходя из этого проводился отбор показателей, обработка и анализ которых математическими методами давали бы информацию о внутреннем строе помещичьего хозяйства.

Наиболее существенными показателями социально-экономической структуры помещичьего хозяйства являются сведения о применении в хозяйстве помещика наемной рабочей силы, наличии владельческого скота, инвентаря, доходах и расходах помещика[1]. Что касается метода анализа данных,

[1] Эти сведения должны учитываться не в абсолютном, а в относительном выражении, т. е. в расчете на десятину помещичьей пашни или посевов.

Таблица 4. Корреляционные модели капиталистической и отработочной систем помещичьего хозяйства

Признаки	Коэффициент корреляции					
	1	2	3	4	5	6
	Отработочная					
Стоимость инвентаря	×	0,62	—0,35	0,52	—0,57	—0,06
Стоимость скота	0,75	×	—0,21	0,26	—0,44	—0,22
Расход на рабочих	0,91	0,68	×	0,10	0,67	—0,66
Расход на тягловый скот	0,61	0,49	0,58	×	0,07	0,35
Общий расход	0,79	0,58	0,91	0,56	×	—0,51
Доход от сельского хозяйства	0,57	0,37	0,69	0,39	0,87	×

Капиталистическая

то он должен раскрывать тесноту взаимосвязи между выделенными показателями. Наиболее надежным математическим показателем тесноты взаимосвязи количественных признаков является коэффициент корреляции. Проверить адекватность предложенного метода сути изучаемого явления возможно на основе обработки и анализа такого источника о помещичьем хозяйстве, в котором содержались бы и прямые указания на способ ведения хозяйства помещиком и количественные характеристики состояния помещичьего хозяйства (данные о применении рабочей силы, наличии скота, инвентаря и др.). Основой для такого анализа могут служить оценочные описания имений, заложенных в Государственном дворянском земельном банке. В некоторых из них содержатся указания на способ ведения хозяйства либо на основе испольщины, издольщины, отработков, либо на основе свободного найма.

Далее, были вычислены коэффициенты корреляции между показателями, характеризующими состояние помещичьего хозяйства (применение рабочей силы, наличие скота, инвентаря и др.) соответственно двух групп имений с отработочным и капиталистическим типом хозяйства. Тем самым открылась возможность как бы в чистом виде выявить структуру двух типов помещичьих хозяйств и построить их корреляционные модели. Эти модели (см. табл. 4) иллюстрируют адекватность выбранного математического метода и являются определенными эталонами структуры двух типов помещичьего хозяйства.

Как видно из таблицы, в группе имений с капиталистической организацией производства имела место положительная и в большинстве случаев очень тесная взаимосвязь между основными производственными компонентами. Особенно важно, что обеспеченность инвентарем и скотом находилась в очень тесной взаимосвязи с затратами на рабочую силу. Это значит, что расширение масштаба применения наемной рабочей силы сопровождалось и расширением производственной базы, т. е. хозяйство велось с помощью свободного найма сельскохозяйственных рабочих на основе использования помещичьего тягла и инвентаря, т. е. по-капиталистически. Обратная картина наблюдается в группе отработочных имений. Обеспеченность инвентарем и скотом здесь естественно была пропорциональной. Но эта обеспеченность даже в малой степени не соотносилась с затратами на рабочую силу (коэффициенты корреляции низкие и отрицательные). Она была проявлением того, что хозяйство в этой группе имений велось прежде всего на основе использования инвентаря и скота окрестных крестьян. Поэтому и увеличение (или сокращение) используемой в хозяйстве рабочей силы не требовало от помещиков этих имений соответственного увеличения (или сокращения) наличного тягла или инвентаря. Полученная корреляционная модель характеризует отработочный тип помещичьего хозяйства.

Основной итог построения корреляционных моделей состоял в том, что была показана возможность выявления социально-экономического строя помещичьего хозяйства без непосредственно выражающих его характер конкретно-исторических данных, т. е. без сведений о соотношении вольного и кабально-отработочного найма, либо удельном весе отработочной и капиталистической систем в различных отраслях производства и др. Корреляционная модель дает информацию, которая и без таких данных позволяет и притом значительно более точно раскрыть суть внутреннего социально-экономического строя помещичьего хозяйства. Тем самым открывается путь к широкому анализу характера этого строя на основе массовых исторических источников, которые до этого оставались фактически вне поля зрения историков и анализ которых позволяет решить вопрос о социально-экономическом строе помещичьего хозяйства в масштабах всей страны.

Так был проведен анализ социально-экономического строя помещичьего хозяйства европейской России и его отдельных регионов на основе обработки материалов сельскохозяйственной и поземельной переписи 1917 г. Корреляционная модель социально-экономического строя помещичь-

Таблица 5. Корреляционная модель социально-экономического строя помещичьего хозяйства европейской России

Признаки в расчете на десятину посева	Коэффициенты корреляции				
	1	2	3	4	5
Рабочий скот	×				
Продуктивный скот	0,70	×			
Наемные рабочие	0,80	0,67	×		
Плуги	0,76	0,69	0,84	×	
Усовершенствованные орудия	0,83	0,63	0,82	0,86	×

его хозяйства, полученная на основе обработки данных о помещичьем хозяйстве 34 губерний европейской России, содержащихся в материалах переписи 1917 г., свидетельствует об общей высокой сбалансированности основных компонентов помещичьего хозяйства России (см. табл. 5). Взаимосвязь между основными признаками, характеризующими социальный тип организации хозяйства, была, как видим, очень тесной. Тем самым доказывается, что к исходу капиталистической эпохи, накануне Великой Октябрьской социалистической революции, в помещичьем хозяйстве безусловно господствующей была капиталистическая система.

Предложенный метод позволяет изучить динамику, выявить этапы буржуазной эволюции помещичьего хозяйства путем сравнения корреляционных моделей помещичьего хозяйства за различные периоды времени на основе обработки источников, содержащих сведения о помещичьем хозяйстве за значительные исторические периоды (например, описания имений, заложенных в дворянском земельном банке), а также изучать региональные особенности социально-экономического строя помещичьего хозяйства.

Высокоэффективными оказались результаты корреляционного анализа и при изучении внутреннего строя крестьянского хозяйства как в эпоху феодализма и капитализма, так и в советский период. Покажем это на примере изучения крестьянского хозяйства Урала 20-х годов по данным бюджетных обследований[1]. Для анализа внутренней структуры крестьянских хозяйств разных типов по

---

[1] См.: *Обожда В. А.* К вопросу о социально-экономической группировке крестьянских хозяйств доколхозной деревни (по материалам Урала). — В кн.: Математические методы в историко-экономических и историко-культурных исследованиях.

публикации данных бюджетов крестьянских хозяйств Урала за 1925/26 год было отобрано для анализа 43 показателя, характеризующих различные аспекты крестьянского хозяйства.

Группировка хозяйств по величине валового сельскохозяйственного дохода позволила выделить пять групп дворов, существенно отличавшихся не только по размеру хозяйств, но и по характеру социально-экономических отношений. На основе традиционного анализа средних показателей по группам хозяйства I группы с доходом до 250 руб. на двор являлись полупролетарскими, бедняцкими (42 хозяйства); хозяйства II группы с доходом от 250 до 500 руб.— маломощными, середняцкими (119 хозяйств); хозяйства III группы с доходом от 500 до 900 руб.— типично середняцкими (116 хозяйств); хозяйства IV группы с доходом от 900 до 1500 руб.— зажиточными (89 хозяйств) и, наконец, хозяйства с доходом свыше 1500 руб. были кулацкими (32 хозяйства).

Вместе с тем традиционный анализ средних показателей не может раскрыть многих особенностей внутреннего строя крестьянских хозяйств разных социально-экономических типов. Сделать это можно путем выявления взаимосвязей отдельных его элементов, т. е. на основе корреляционного анализа.

Посмотрим прежде всего, какова была взаимосвязь обеспеченности крестьянских хозяйств рабочей силой и средствами производства во всех группах хозяйств (см. табл. 6).

Из таблицы четко видна разница в уровне сбалансированности элементов крестьянского хозяйства высших и низших групп. В беднейшей группе заметна взаимосвязь рабочей силы хозяйства и размера посева (0,46). В то же время она неуклонно уменьшается от группы к группе и в высшей, кулацкой, группе исчезает совсем, так как здесь большую роль играли наемные работники. Наиболее резкие различия низших и высшей групп хозяйств видны по характеру взаимосвязей таких пар признаков, как посев — аренда, аренда — рабочий скот. В первых двух группах взаимосвязи посев — аренда нет совсем. В третьей группе заметна лишь взаимосвязь посев — аренда, т. е. налицо у незначительной части хозяйств этой группы признаки предпринимательства (чем больше посев, тем больше аренда). В двух же высших группах эта связь уже очень существенна, т. е. товарный характер аренды имеется у большинства хозяйств. Взаимосвязь аренда — рабочий скот реально

Таблица 6. Взаимосвязь обеспеченности рабочей силой и средствами производства (коэффициенты корреляции)

Пары показателей	Группы				
	I	II	III	IV	V
Работники — всего посева	0,46	0,22	0,20	0,09	0,04
Работники — стоимость средств производства	0,24	0,01	0	0,18	—0,02
Всего пашни — весь посев	0,48	0,56	0,75	0,84	0,98
Посев — рабочий скот	0,62	0,41	0,39	0,41	0,70
Посев — стоимость средств производства	0,31	0,20	0,22	0,29	0,54
Посев — аренда	0	0,23	0,51	0,66	0,79
Аренда — рабочий скот	0	0,02	0,21	0,28	0,53
Аренда — стоимость средств производства	0	0,07	0,09	0,20	0,24
Рабочий скот — инвентарь	0,51	0,39	0,20	0,41	0,59
Инвентарь — стоимость всего скота	0,51	0,54	0,35	0,61	0,73
Рабочий скот — хозяйственные постройки	—0,11	0,02	0,02	0,08	0,42

значима лишь в V группе (0,53). Взаимосвязь аренда — стоимость средств производства едва заметна даже у кулацкой группы (0,24), что свидетельствует о том, что кулацкие хозяйства еще некрепки и практика аренды земли еще не сбалансирована со стоимостью средств производства. Вместе с тем у них с инвентарем тесно взаимосвязаны такие показатели, как рабочий скот (0,59) и особенно стоимость всего скота (0,73). В бедняцкой группе эта связь слабее, а в наиболее натуральной середняцкой III группе она слабее всего (0,35). В кулацкой группе ощутима связь: рабочий скот — хозпостройки (0,42) и т. д. В целом только в высшей кулацкой группе можно видеть устойчивую тесную взаимосвязь признаков, включающих показатели размера посева, количества рабочего скота, стоимость всего скота, стоимость инвентаря и всех средств производства. Недостаточная сбалансированность этих признаков лежала в основе неустойчивости групп середняцкого типа.

Перейдем к анализу товарно-денежных отношений в крестьянском хозяйстве (см. табл. 7).

В первую очередь здесь следует отметить резкие различия в характере взаимосвязей почти всех рассматриваемых пар признаков, с одной стороны, в хозяйствах кулацкой группы и, с другой,— в хозяйствах остальных четырех групп.

Таблица 7. Характер товарно-денежных отношений (коэффициенты корреляции)

Пары показателей	Группы				
	I	II	III	IV	V
Продано и отдано в обмен продуктов полеводства — доход от полеводства	0,50	0,44	0,36	0,55	0,90
Продано и сдано продуктов сельского хозяйства — доход от полеводства	0,16	0,09	0,18	0,29	0,87
Куплено и получено всего — доход от промыслов	0,36	0,48	0,49	0,51	0,04
Весь денежный доход — доход от сельского хозяйства	—0,21	—0,12	0,09	0,39	0,87
Весь денежный доход — куплено продуктов сельского хозяйства	0,89	0,72	0,50	0,57	0,32
Весь денежный доход — потребительский расход	0,75	0,71	0,69	0,47	0,57

Наличие тесных взаимосвязей между показателями объема сельскохозяйственного производства и размерами продукции, реализованной на рынке, должно указывать на целенаправленное товарное производство, предпринимательский характер производственной деятельности кулацкого хозяйства (0,90). Чем больше хозяйство производит, тем больше оно продает, т. е. основная цель хозяйства не удовлетворение внутренних потребностей, а товарное производство.

С другой стороны, отсутствие или наличие относительно слабой связи данных признаков указывает на нерегулярный, неустойчивый характер связи с рынком. В этом случае размер товарной части продукции определялся в значительной степени не общим объемом производства, а другими неучтенными факторами — срочная потребность в деньгах и прочее.

На преобладающую роль сельскохозяйственного производства и его товарности в денежном бюджете кулацких хозяйств указывает коэффициент корреляции между размером всего денежного дохода и дохода от продажи продуктов сельского хозяйства (0,87). В маломощных и типично середняцких хозяйствах различия в денежном доходе определялись в первую очередь доходами от неземледельческих занятий. (Коэффициенты от 0,36 до 0,51). Кулаки промыслами не занимались (0,04).

Таблица 8. Корреляционная модель кулацких хозяйств Урала

Показатели	Коэффициенты корреляции 0, . . .								
	1	2	3	4	5	6	7	8	9
Стоимость инвентаря	×	93	72	77	77	81	74	81	77
Стоимость средств производства		×	78	86	80	83	82	83	86
Доход от полеводства			×	90	90	87	82	87	80
Доход от сельского хозяйства				×	87	89	86	90	87
Продано и отдано продуктов полеводства					×	95	87	97	85
Продано и отдано продуктов сельского хозяйства						×	87	97	87
Куплено и получено в обмен всех продуктов							×	88	96
Денежный доход от сельского хозяйства								×	87
Весь денежный доход									×

Очень важное значение имеет выявление с помощью корреляционного анализа устойчивого и существенного ослабления тесноты взаимосвязей потребительских расходов на продукты сельского хозяйства с величиной денежного дохода от низшей группы к высшей (строки 5 и 6). Данная тенденция указывает на строгое ограничение расходов доходами в хозяйствах менее состоятельных групп (0,75; 0,71), на независимость потребительских расходов от доходов в крупных, более зажиточных хозяйствах (0,47; 0,57).

Сильные различия не только в тесноте, но и в самом характере взаимосвязей свидетельствуют о том, что внутренний строй крестьянского хозяйства полярных групп имел существенные различия. Это отчетливо выявляется в корреляционных моделях этого строя, которые отражают наиболее тесные, устойчивые взаимосвязи элементов, характеризующих их внутреннюю структуру. Внутренний строй кулацкого хозяйства характеризуется тесной взаимосвязью производственной базы, доходов и расходов крестьян (см. табл. 8).

Кроме того, в этой группе хозяйств были тесно взаимосвязаны размеры всей удобной земли, пашни, аренды и посевов.

В отличие от кулацкой группы в бедняцкой и середняцкой группах нельзя выделить столь разветвленной системы взаимосвязанных компонентов.

Таблица 9. Взаимосвязь признаков бедняцких хозяйств Урала

Показатели	Коэффициенты корреляции 0, . . .			
	1	2	3	4
Куплено и получено в обмен всех продуктов	×	76	74	83
Куплено продуктов продовольствия		×	74	89
Расход на личное потребление			×	79
Весь денежный доход				×

Но и здесь были тесные взаимосвязи, характерные для бедняцких хозяйств, например наличие тесной связи между денежным доходом и потребительскими расходами (см. табл. 9).

Как видно из матрицы, весь расход бедняцкого хозяйства обусловлен главным образом величиной денежного дохода, размеры которого, как было показано, во многом определялись неземледельческими занятиями. Следует подчеркнуть и наименее тесную из всех групп связь размера посева с: 1) размером пашни, 2) всей удобной и 3) арендованной земли. Следовательно, и у бедняцких хозяйств внутренняя структура имеет свою специфику.

Что касается середняцких хозяйств, то здесь важно подчеркнуть отсутствие каких-либо специфических взаимосвязей, свойственных именно середняцким группам хозяйств. От группы маломощных середняков к группе зажиточных четко прослеживается ослабление специфических взаимосвязей, характерных для бедняцких хозяйств, и усиление тесноты взаимосвязей элементов, характерных для кулацких дворов.

Таким образом, на основе корреляционного анализа с новой стороны подтверждается определение классовой сущности крестьянского хозяйства, его двойственной природы и тенденций развития, неоднократно отмечавшихся В. И. Лениным. Вместе с тем обнаруженные с помощью корреляционного анализа особенности взаимосвязей элементов крестьянских хозяйств кардинально углубляют наши знания внутренней структуры хозяйств разных социально-экономических типов.

В целом, как видно из рассмотренных примеров[1], системный подход и структурный анализ являются весьма эффективным и мощным средством (даже при использовании несложного математического аппарата) для углубленного изучения как социально-экономических, так и других явлений исторического прошлого.

§ 3. Структурно-типологическое изучение социально-экономических явлений

В практике исторических исследований одной из важнейших задач является типизация изучаемых явлений и объектов. На простейшем уровне, когда предметом типизации служат объекты с одним ярко выраженным однородным признаком, задача эта решается сравнительно легко. Историк классифицирует, например, помещичьих крестьян в период позднего феодализма на оброчных и барщинных, поскольку уже со второй половины XVIII в. чистые формы ренты являются преобладающими. Однако задача резко усложняется, если поставить вопрос о типологии самого барщинного или оброчного крестьянского хозяйства.

Здесь уже генерализующий признак (форма ренты) исчезает и исследователь имеет дело с целым рядом разнородных признаков (характеристик) этого хозяйства (рабочий и продуктивный скот, размер усадьбы, пашни, сенокосов и т. д.). Учитывая все эти признаки, традиционными методами невозможно корректно выделить основные типы крестьянского хозяйства. Но эта задача может быть решена с помощью современных методов многомерного анализа.

В нашей литературе уже накоплен известный опыт типологической многомерной классификации различных сложных социально-экономических объектов и явлений. В области аграрной истории одним из наиболее трудных объектов классификации является крестьянское хозяйство феодальной эпохи, т. е. того периода, когда в силу слабого развития товарно-денежных отношений еще не выделились типы хозяйств, резко отличающиеся своим внутренним экономическим строем.

---

[1] Существенные результаты дал структурный анализ крестьянского хозяйства и при изучении характера крестьянской аренды в России в конце XIX в. — См.: *Моисеенко Т. Л.* Методы изучения крестьянской аренды в России по данным земской статистики конца XIX в. — История СССР, 1979, № 4.

Примером одной из попыток классификации объектов этого класса является типология крестьянских хозяйств Эстонии начала XIX в., проведенная Ю. Ю. Кахком [1] с целью сравнения их экономического состояния и тяжести лежащих на них феодальных повинностей. В условиях господства барщинной системы существовало деление крестьянских хозяйств на разряды в зависимости от тяжести барщины (гаковые хозяйства, полугаковые и т. д.). Исследование преследовало цель выявить адекватность этой классификации экономическим потенциям каждого хозяйства. В основу работы был взят материал 357 хозяйств Северной и 502 хозяйств Южной Эстонии по 10 показателям. Программа обработки данных состояла в вычислении статистических «расстояний» между всеми индивидуальными хозяйствами крестьян и конструировании на их основе матрицы показателей сходства. Итогом анализа было получение «ступенчатой социальной лестницы» крестьянских хозяйств, на одном конце которой были самые слабые, а на другом — самые состоятельные хозяйства. Многомерная типологическая классификация внесла весьма существенные коррективы в деление на разряды крестьянских хозяйств, которое в свое время проводилось лишь на основе различий в размере земли на хозяйство.

Другой опыт многомерной классификации крестьянских хозяйств был основан на факторном анализе. Этот метод был впервые использован В. П. Пушковым и И. М. Промахиной для группировки крестьянских хозяйств советской доколхозной деревни 20-х годов на материалах бюджетных обследований Украины [2].

Из всей совокупности факторов (а их авторы выявили пять) было взято два, обладающие наибольшим весом и вобравшие в себя наибольшее количество признаков. Затем хозяйства были ориентированы в «пространстве» двух факторов и из образовавшихся на графике более или менее обособленных друг от друга совокупностей точек — объектов формировались группы хозяйств.

Необходимость применения многомерной типологической классификации еще острее ощущается при обобщенном

---

[1] См.: *Кахк Ю. Ю.* К вопросу о типологии крестьянских хозяйств в Эстонии в начале XIX века. Таллин, 1975.

[2] См.: *Пушков В. П., Промахина И. М.* Опыт применения факторного анализа для классификации, изучения структуры и моделирования социальных типов крестьянских хозяйств. (По материалам крестьянских бюджетов Украины 20-х годов XX в.). — В кн.: Математические методы в социально-экономических и археологических исследованиях.

исследовании сложных систем макрообъектов. В частности, сложнейшее переплетение компонентов представляет собой экономика европейской России эпохи империализма в региональном аспекте. На обширном пространстве размещались многочисленные сравнительно однородные районы, отличавшиеся вместе с тем друг от друга важными чертами в силу исторических, природно-географических, демографических, экономических, политических и других причин. Выделение этих районов чаще всего предстает как процедура экономико-географического районирования, а основные конкретно-научные приемы могут быть определены как эмпирически-интуитивные. В качестве критериев для выделения районов используются от силы несколько разнородных признаков: природно-климатические условия как важнейший признак, далее степень развития сельского хозяйства, степень развития промышленности. Однако при этом практически невозможно учесть совокупность основных производственно-экономических и социально-экономических характеристик. Решить эту сложную задачу позволяют методы многомерного анализа. Они и были использованы, во-первых, при выявлении аграрной типологии губерний европейской России на рубеже XIX—XX вв. и, во-вторых, при анализе их структуры и определении общего уровня аграрного развития [1].

Набор признаков при изучении аграрной типологии европейской России включал 19 показателей, характеризующих основные производственно-экономические и социально-экономические условия сельскохозяйственного производства, общий уровень развития его ведущих отраслей и степень их интенсификации.

Цель методов многомерной классификации — выделить в многомерном пространстве признаков компактные группы, скопления объектов, близких друг к другу. В данном случае классификация проводилась с помощью алгоритмов кластер-анализа.

Обработка исходных данных на ЭВМ позволила наглядно раскрыть пространственную структуру системы аграрных отношений на рубеже XIX—XX вв. Эта структура характеризовалась наличием совокупности групп губерний,

---

[1] См.: *Ковальченко И. Д., Бородкин Л. И.* Аграрная типология губерний европейской России на рубеже XIX—XX вв. (Опыт многомерного анализа).— История СССР, 1979, № 1; *их же.* Структура и уровень аграрного развития районов европейской России на рубеже XIX—XX вв. (Опыт многомерного анализа). — История СССР, 1981, № 1.

Типы аграрной структуры и образующие их районы и губернии

Нечерноземный тип

1. Владимирская, Костромская, Ярославская
2. Тверская, Псковская, Новгородская, Вологодская
3. Гродненская, Виленская, Минская, Витебская
4. Смоленская, Калужская, Черниговская, Могилевская, Волынская
5. Вятская, Пермская, Уфимская

Среднечерноземный тип

6. Рязанская, Тульская, Орловская, Курская, Тамбовская, Пензенская
7. Нижнегородская, Казанская, Симбирская
8. Воронежская, Саратовская
9. Харьковская, Полтавская
10. Киевская, Подольская
11. Бессарабская

Степной тип

12. Херсонская, Таврическая, Екатеринославская, Донская
13. Самарская, Оренбургская, Астраханская

Прибалтийский тип

14. Эстляндская, Лифляндская, Курляндская, Ковенская

Столичный тип

15. Московская
16. Петербургская

Северный тип

17. Архангельская, Олонецкая

в каждую из которых входили сходные по своему аграрному облику губернии. Районирование, основанное на результатах машинной классификации, представляет собой следующую картину.

Классификация выявила 17 групп губерний (кластеров). Укрупняясь, они образуют шесть типов аграрного развития. Составляющие эти группы губернии образуют определенные регионы, включающие территориально смежные губернии.

Анализ средних значений и коэффициентов вариации по 19 показателям показал, с одной стороны, высокую степень однородности губерний, включенных в каждый из кластеров и типов аграрного развития, с другой стороны, была выявлена и специфика каждого кластера и типа.

Задача второго этапа работы заключалась в том, чтобы перейти от 19 исходных показателей к меньшему числу характеристик, но таких, которые позволили бы углубить анализ и получить интегральные, обобщенные показатели, характеризующие аграрную структуру губерний и районов и уровень их развития. Эта задача решалась с помощью нового метода факторного анализа — метода экстремальной группировки параметров. Основная идея этого метода состоит в том, что совокупность исходных признаков распадается на группы, каждая из которых отражает действие определенных факторов — причин. Поскольку признаки внутри каждой из групп связаны между собой более тесно, чем признаки разных групп, то задача сводится к выделению сильно закоррелированных групп и признаков, что позволяет выявить соответствующие факторы.

Число факторов в данной работе было взято равным 5, исходя как из формальных, так и из содержательных критериев. Суть построенных на ЭВМ факторов раскрывается содержательным смыслом характеризующих их признаков.

Выявленные факторы приводятся ниже в той последовательности, которая соответствует их роли в общей аграрной структуре. Указанные в скобках коэффициенты корреляции показывают тесноту взаимосвязи признаков с соответствующими факторами.

I. Характер разложения крестьян и уровень развития земледелия.

1. Доля (%) безлошадных и однолошадных дворов	(—0,884)
2. Лошадей на душу населения	(0,887)
3. Доля (%) дворов с 4 и более лошадьми	(0,856)
4. Сбор хлебов и картофеля на душу населения	(0,820)
5. Посев на душу населения	(0,849)

II. Интенсификация животноводства

1. Продуктивного скота на десятину посева	(0,860)
2. Лошадей на десятину посева	(0,758)
3. Осенние цены ржи	(0,754)
4. Цена десятины земли	(—0,651)
5. Арендная плата за десятину пашни	(—0,596)

III. Капитализация и интенсификация земледелия

1. Наемных сельскохозяйственных рабочих на десятину посева	(0,978)
2. Наемных рабочих в % к местным работникам	(0,925)
3. Урожайность зерновых в пудах с десятины	(0,643)

IV. Характер земельных отношений

1. Земельный надел на душу	(0,743)
2. Доля дворянской пашни в удобной земле	(—0,840)
3. Арендованная земля в % к надельной	(—0,660)

V. Положение крестьян и сельскохозяйственных рабочих

1. Продуктивного скота на душу населения	(0,809)
2. Поденная плата сельскохозяйственным рабочим в уборку урожая	(0,651)
3. Проданных частновладельческих земель в % к общей площади	(0,567)

Важно отметить, что все 19 признаков тесно связаны с соответствующими факторами, т. е. все они являются существенными в плане анализа аграрной структуры губерний. При этом все пять факторов предполагают четкую интерпретацию, что дает возможность перейти к обобщенной характеристике аграрного строя районов европейской России.

Методы многомерного анализа позволяют не только выделить факторы и определить их сравнительную роль, но и установить вес каждого объекта, в данном случае каждой из 50 губерний по данному фактору, т. е. провести ранжирование губерний по степени их развитости в аспекте, характеризуемом каждым из факторов. При этом средний уровень развития принимается равным нулю. Губернии с уровнем развития выше среднего получают положительный показатель, а ниже среднего — отрицательный. Эти показатели называются факторными весами объектов. На их основе можно получить факторные веса для соответствующих районов путем выведения среднего показателя из показателей губерний, входящих в эти районы. Далее, на основе факторных весов можно вывести совокупный показатель для губерний и районов по всем рассматриваемым факторам, т. е. получить интегральный, обобщенный индекс степени развитости аграрной структуры.

Рассмотрим для примера вопрос о характере разложения крестьян и уровне развития земледелия (1-й фактор). Лидирующее положение по степени развития этого фактора занимали два района, образующие степной тип аграрной структуры. В целом степная полоса (индекс 0,745) отличалась наиболее широким слоем зажиточных и наименьшей долей беднейших крестьян и самым высоким уровнем развития земледелия. Вслед за ней с несущественным отрывом (0,271) шел Прибалтийский район. Этот отрыв был обусловлен более низким уровнем развития земледелия. Во всех других районах этот уровень был ниже среднего. Наиболее низким было значение первого фактора в Столичном (—0,859) и Северном (—0,689) районах.

Аналогичный анализ, проведенный по всем пяти факторам, показал, что по одним аспектам аграрной структуры районы могли быть «лидирующими», а по другим — «отстающими». Неоднотипность аграрной структуры отдельных районов и разная направленность действия определяющих ее факторов требуют для оценки общего уровня аграрного развития районов введения интегрального показателя. Это и было целью последнего этапа работы. В результате было

Таблица 10. Общий уровень аграрного развития районов европейской России на рубеже XIX—XX вв.

Районы	Индекс	Районы	Индекс
I. Высокий уровень		**III. Низкий уровень**	
1. Южный Степной	0,276	9. Центрально-Южно-Черноземный	—0,056
2. Прибалтийский	0,264	10. Центрально-Черноземный	—0,094
3. Юго-Восточный	0,169	11. Петербургский	—0,102
4. Северный	0,164	12. Западный	—0,120
II. Средний уровень		13. Центрально-Промышленный	—0,131
5. Центрально-Западный	0,024	14. Средневолжский	—0,154
6. Северо-Западный	0,008	15. Левобережный	—0,204
7. Приуральский	0,001	16. Юго-Западный	—0,211
8. Бессарабский	—0,011	17. Московский	—0,239

получено ранжирование всех 17 районов европейской России в соответствии с интегральным показателем, отражающим совокупный вес пяти факторов с учетом их сравнительной роли (табл. 10).

Главное, что обращает на себя внимание в общем уровне аграрного развития отдельных районов европейской России,— это то, что из четырех районов с наиболее высоким уровнем аграрного развития три (Южный Степной, Юго-Восточный и Северный) являлись районами, в которых господствующим был буржуазно-демократический, крестьянский («американский») тип буржуазной аграрной эволюции. И это несмотря на то, что господствующие классы и выражавшее их интересы самодержавное государство прилагали все усилия к тому, чтобы обеспечить господство буржуазно-консервативного, помещичьего («прусского») пути буржуазной аграрной эволюции.

Проведенный анализ убедительно подтверждает, что объективно-исторически именно развитие капитализма на основе крестьянского хозяйства обеспечивало наиболее быстрый и благоприятный для широких масс общественный прогресс.

Подводя итог применению методов многомерного анализа, следует особо подчеркнуть, что они могут использоваться при изучении любых других совокупностей объектов, характеризуемых тем или иным набором количественных показателей, которые целесообразно использовать не выборочно, а во всей их совокупности как при классификации изучаемых объектов, так и при анализе образуемых ими структур.

§ 4. Изучение динамики социально-экономических процессов

Изучение динамики социально-экономического развития является одной из важнейших задач исторических исследований. Проблемы анализа динамики возникают перед историками в самых различных аспектах и на разных (микро- и макро-) уровнях. На микроуровне изучение динамики обычно связано с выявлением тех изменений во времени, которые были присущи различным социально-экономическим объектам (промышленным предприятиям, помещичьим имениям, крестьянским хозяйствам, отдельным семьям, лицам и т. д.). Задача заключается в том, чтобы выявить состояние этих объектов в различные временные моменты. Решение ее сводится к розыску каждого из этих объектов в их совокупности в соответствующие, зафиксированные источниками моменты. При решении подобных поисковых задач весьма эффективным может быть применение ЭВМ. Это хорошо видно на примере изучения расслоения крестьянства путем выявления тенденций в изменении хозяйственной состоятельности отдельных крестьянских дворов[1].

Для решения этой задачи в распоряжении исследователя имелись подворные описания крестьянских хозяйств за различные годы первой половины XIX в. На основе их была предпринята попытка наряду с традиционной обработкой данных подворных описей выявить изменения в хозяйственной самостоятельности отдельных крестьянских дворов, а именно проследить их перемещения по различным хозяйственно-экономическим группам за более или менее длительные периоды времени. Для этого было необходимо осуществить розыск каждого двора, зафиксированного в начальной подворной описи, во всех последующих описях, разделенных периодами примерно в 15—20 лет. Для реализации этой трудоемкой процедуры была использована ЭВМ. Дворы разыскивались по именному составу семей, т. е. именам и отчествам крестьян, входящих в состав двора. Для обработки данных подвор-

---

[1] См.: *Ковальченко И. Д.* Русское крепостное крестьянство в первой половине XIX века, с. 356—366.

ных переписей на ЭВМ была проведена кодировка материала. Все имена получили цифровые коды и материал был введен в ЭВМ. Отождествление дворов происходило на основе совпадения не менее двух имен и отчеств. Дворы, по которым совпадения двух имен и отчеств не оказалось, рассматривались как прекратившие свое существование. Розыск дворов осуществлялся по 11 имениям (всего обработаны данные по 2000 дворов). Итоги розыска позволили собрать следующие сведения: какое число раз дворы каждой группы (по состоятельности было выделено три группы), зафиксированные в начальной описи, встретились в последующих описях, какое число раз дворы каждой группы встретились в тех или иных группах, сколько дворов все время оставалось в первоначальной группе, сколько дворов, зафиксированных в первой переписи, оказались в последней и как они распределены по группам. Эти данные позволили выявить процесс мобильности крестьянских хозяйств. Так было установлено, что наиболее существенные изменения в этом процессе в первой половине XIX в. произошли в земледельческой деревне.

Если в первой трети XIX в. движение средних дворов, которые составляли 60—70% крестьянских хозяйств, в зажиточную группу преобладало над переходом их в беднейшую группу, то в 30—50-е годы переход в беднейшую становится господствующим. В зажиточной группе дворов преобладание стабильности группы в первой трети века сменяется перевесом перехода этих дворов в среднюю и беднейшую группы в 30—50-х годах XIX в. Качественный сдвиг заключается, таким образом, в установлении четкой тенденции движения «вниз» (из зажиточной в низшие группы, из средней соответственно в беднейшую) при исчезновении или резком ослаблении движения «вверх». В целом использование ЭВМ позволило изучить процесс мобильности крестьянских хозяйств не в общем плане, когда о нем судят по изменению соотношения различных по своей хозяйственной состоятельности групп крестьян, а с учетом изменения статуса каждого крестьянского двора, т. е. значительно более глубоко и конкретно.

Использование ЭВМ для решения подобного рода задач получает все более широкое распространение в исследовательской практике историков. Оно расширяет возможности изучения динамики явлений на их микроуровне, т. е. в ее наиболее конкретном выражении.

Другим аспектом изучения динамики в историческом развитии является анализ обширных в пространственном, длительных во временном и весьма сложных в содержательном плане процессов, в которых проявляются наиболее общие закономерности тех или иных сфер общественной жизни. Успешный и конкретный анализ этих процессов фактически невозможен без применения количественных методов и ЭВМ. Ярким примером эффективного применения этих методов на макроуровне является изучение процесса формирования единого аграрного рынка в России в XVIII — начале XX в.[1]

Изучение динамики такого сверх сложного процесса, как формирование единого рынка, решается путем последовательного анализа структуры рынка на разных этапах исторического развития.

Для решения этой задачи необходимо содержательное изучение самого понятия «рынок» и создание качественной модели «рынка».

Суть вопроса вкратце состоит в следующем.

В условиях господства натурального хозяйства и ограниченного развития товарно-денежных отношений существует сеть мелких местных рынков. Каждый из таких локальных рынков имеет свои рыночные цены, отражающие местные условия производства и сбыта тех или иных товаров.

С ростом производительных сил происходит постепенное углубление общественного разделения труда, что приводит к укрупнению местных рынков и в конечном счете к формированию единого рынка.

По мере своего развития рынок становится не просто ареной обмена товарных изделий, товарных стоимостей, а и регулятором общественного производства. Экономическим рычагом регулирования общественного производства является закон стоимости, прокладывающий себе дорогу через механизм взаимодействия спроса и предложения товаров. Следовательно, изучать процесс развития рынка можно путём выяснения степени проявления закона стоимости как с точки зрения глубины его воздействия на производство, так и с точки зрения выявления территориальной протяженности, сферы его действия. Выяснив это, мы выясним уровень развития рынка как регулятора общественного производства. Сделать это можно, изучая цены на основные товары, ибо в степени их сопряженности, т. е. согла-

---

[1] См.: *Ковальченко И. Д., Милов Л. В.* Всероссийский аграрный рынок, XVIII — начало XX в. (Опыт количественного анализа).

сованности и пропорциональности их изменений в различных пунктах, отражается действие закона стоимости.

Внутренний механизм процесса был следующим. На ранней стадии развития рынка, когда существует сеть мелких местных рынков, движение цен на какой-то товар в целом по стране должно выглядеть как хаотическое и лишь в рамках небольших местных рынков это движение должно быть согласованным. В самом деле, ведь если пункты А и Б лежат в пределах единого рынка, то механизм соотношения спроса и предложения в них отличаются общностью в своих основных моментах. Следовательно, увеличение цен в А вызывает и увеличение цен в Б, и наоборот. Если же пункты А и Б расположены в разных местных рынках и механизмы соотношения спроса и предложения цен у них разные, то увеличение цен в А не вызовет соответствующего увеличения цен в Б. В основе последнего явления лежит экономическая изоляция пунктов А и Б друг от друга. Следовательно, синхронное изменение цен в А и Б есть свидетельство действия каких-то общих экономических факторов, которые в конечном счете отражаются в более или менее синхронном действии механизма спроса и предложения. Общность экономических факторов может создать синхронность движения цен даже при условии отсутствия между ними движения .товаров. Следовательно, этапы интеграции мелких рынков в более крупные соответственно отразятся на синхронности и пропорциональности изменения в них цен. В едином рынке в этом случае должен быть единый механизм движения цен на ту или иную продукцию.

Выявить сопряженность колебания цен традиционными методами не представляется возможным. Здесь нужны методы количественного анализа. Наиболее эффективным в данном случае является корреляционный анализ динамических рядов цен. Поскольку изучался процесс формирования аграрного рынка, анализу подверглись цены на рожь и овес, являвшиеся основными земледельческими культурами, а также на землю, тягловый скот и рабочую силу. Анализ первых охватывал следующие рубежи: а) середина XVIII в., так как в 50-х годах XVIII в. была проведена отмена внутренних таможенных пошлин; б) начало XIX в., относящееся к периоду разложения феодально-крепостнической системы; в) середина XIX в.— период кризиса этой системы; г) 80—90-е годы XIX в.— период интенсивного развития капитализма в сельском хозяйстве. Для каждого из этих рубежей были взяты цены на рожь и овес за десятилетние периоды. Это обусловлено тем, что десять лет являются тем минимальным отрезком времени, в пределах которого укладывается определенное сочетание низких, средних и высокоурожайных лет. Выяснение степени сопряженности цен в каждом из периодов и расширение масштабов и степени этой сопряженности от периода к периоду позволяет проследить процесс формирования единого товарного рынка. Корреляционный же анализ цен на средства производства и рабочую силу, охвативший период конца XIX — начала XX в., показывает степень развития аграрного капиталистического рынка.

Каждый из выделенных рубежей характеризовался динамическими рядами цен на указанные товары по каждой из 50 губерний европейской России. Эти ряды были подвергнуты математико-статистической обработке, которая состояла, во-первых, в их аналитическом выравнивании по прямой линии, в результате чего были получены значения трендов, уровней этих рядов и случайных отклонений от них фактических данных о ценах. Во-вторых, были вычислены коэффициенты корреляции фактических значений цен и их случайных отклонений от уровней для каждой губернии со всеми другими. В-третьих, вычислялись коэффициенты автокорреляции натуральных рядов цен, которые учитывались при вычислении доверительных интервалов к коэффициентам корреляции этих рядов.

В итоге была получена огромная новая информация о движении и взаимосвязи цен, анализ которой раскрывает процесс формирования единого всероссийского аграрного рынка.

Общие итоги изучения процесса складывания единого рынка на товарную продукцию выглядит следующим образом.

В середине XVIII в. рынок на основные сельскохозяйственные продукты (рожь, овес) достиг очень высокого уровня интеграции. В это время уже нет системы мелких местных рынков. Сформировался обширный Волжский региональный рынок ржи, куда целиком или частично вошли территории примерно десяти будущих губерний европейской России (Псковской, Новгородской, Вологодской, Тверской, Ярославской, Владимирской, Костромской, Нижегородской, Казанской и Симбирской). Уровень взаимосвязи случайных колебаний цен был очень высоким (средний коэффициент $0{,}74 \pm 0{,}03$). Второй региональный рынок ржи — Центрально-Черноземный — объединял территории запада и юга Московской провинции, а также Тульскую, Переяславль-Рязанскую, Орловскую, Елецкую, Белгородскую, Севскую, Воронежскую провинции, Слобожанщину и Киев. Уровень тесноты случайных колебаний здесь еще выше ($0,82 \pm 0,02$). Следовательно, на долю общих факторов, влияющих на механизм случайных колебаний, приходится 64—71% от общего веса общих и местных факторов (в процентах выражен коэффициент детерминации). Характерно, что территория Московской провинции входила в регион, экономическая конъюнктура которого обнимала гораздо меньшую территорию (только Рязанскую, Тульскую и частично Орловскую провинции). Этот Московско-Черноземный регион (уровень сопряженности $0,73 \pm 0,03$), видимо, наиболее древний, так же как и Волжский регион. Что касается Центрально-Черноземного, то он, вероятно, оформился к последней трети XVII в., так как его территория не выходила за рамки засечных черт XVII в.

В высшей степени важно, что корреляционный анализ случайных колебаний цен на овес дает очень сходные итоги. В середине XVIII в. сформировался Волжский региональный рынок овса, контуры которого в целом совпадают с региональным рынком ржи. Есть и второй региональный рынок овса — Черноземный.

В начале XIX в. в европейской России в принципе продолжали существовать те же региональные рынки ржи и овса (Волжский и Центрально-Черноземный). Вместе с тем Волжский регион заметно вырос и включил Прибалтику на северо-западе и Уфу и Симбирск на юго-востоке. Это — обширнейший Прибалтийский-Волжский регион ржи. Степень его единства, однако, не столь высока ($0,71 \pm 0,07$). Вполне оформился процесс интеграции черноземных районов с промышленными. Рынок ржи здесь формируется на основе территории Курской, Орловской, Воронежской, Тамбовской, Тульской, Рязанской, Калужской, Московской и Владимирской губерний ($0,80 \pm 0,04$). Эта тенденция принципиально новая. Вместе с тем итоги бурного хозяйственного освоения юга и юго-восточных районов европейской России в течение второй половины XVIII в. в конечном счете сказались на формировании здесь особого регионального рынка ржи — Черноморско-Уральского. Он охватывал территории Херсонщины, Екатеринослава, Крыма, через Дон к Саратову, Симбирску, Пензе и далее к Уфе. Перми и даже Тобольску. Уровень взаимосвязи случайных колебаний цен ржи высок ($0,75 \pm 0,04$). Весьма важно, что этот региональный рынок имеет тенденцию к расширению на север вдоль Волги (в регион включены районы Казани, Нижнего Новгорода и даже Костромы и Ярославля).

Динамика рынка овса в 10-х годах XIX в. дала итог гораздо более значительный, чем по ржи. Объясняется это тем, что овес в крестьянском хозяйстве стал быстрее товарной культурой, чем рожь. Поэтому в 1809—1819 гг. мы видим уже сформировавшуюся тенденцию не к образованию регионов, а образованию основы единого рынка. Корреляция случайных колебаний цен на овес дала в итоге гигантский регион, охватывающий территории 26 губерний. Средний уровень тесноты взаимосвязи случайных колебаний довольно высок ($0{,}74 \pm 0{,}03$), т. е. от 50 до 60% от суммы всех факторов, воздействующих на случайные колебания цен на овес, составляют факторы, общие для всего региона.

Середина XIX в. наиболее интересный этап в эволюции рынка ржи. Здесь переплетаются две тенденции. С одной стороны, увеличивается число региональных рынков, что отражает резкую активность процессов вовлечения в товарное производство гигантских резервов необъятной массы крестьянских хозяйств при одновременном развитии товарного зернового производства помещичьим классом. С другой стороны, на первый план выходит процесс интеграции региональных рынков. В итоге в середине XIX в. возможно выделение обширных пространств, рыночная конъюнктура которых принадлежит одновременно двум и даже трем региональным рынкам.

Рыночная конъюнктура движения цен на рожь образует в это время на уровне случайных колебаний цен пять региональных рынков. Прежде всего это Центрально-Юго-Западный регион, включавший 6 губерний Промышленного центра, 9 губерний Черноземного центра и Среднего Поволжья и 4 губернии Украины и Бессарабии. В отдельный Западный регион слились Прибалтика, Северо-Запад, Белоруссия, Смоленская и Волынская губернии. Третий региональный рынок Юго-Западный (Черниговская, Полтавская, Харьковская, Киевская, Волынская, Подольская, Херсонская, Таврическая и Бессарабская губернии). Четвертый регион — Центрально-Северо-Западный. В него слились Север, Северо-Запад, Промышленный центр и Черноземный центр (13 губерний). Наконец, в очень деформированном виде сохранялся Черноморско-Уральский регион (Пермская, Казанская, Пензенская, Саратовская, Астраханская и Екатеринославская губернии). Наиболее древний Волжский регион распался. «География» региональных рынков ржи причудлива из-за «наплыва» одного региона на другой. Так, Владимирская, Нижегородская, Воронежская, Казанская, Симбирская и Пензенская губернии одновременно входят в первый и пятый регионы. А Московская, Ярославская, Костромская, Калужская, Тульская, Орловская и Курская губернии одновременно принадлежат первому и четвертому регионам и т. д.

В середине XIX в. фактически сформировался единый товарный рынок на овес. Он охватил 40 губерний европейской России с уровнем тесноты взаимосвязей, равной 0,74+0,02. Сюда не вошли лишь районы Прибалтики и часть Белоруссии. Однако и они были теснейшим образом связаны общностью механизма движения цен с Севером, Промышленным центром и Юго-Западом.

Наконец, к 80-м годам XIX в. единый рынок ржи и овса стал реальным фактом. Теснота взаимосвязи очень высокой степени ($K=0{,}89+0{,}01$) характерна для всей обширнейшей территории европейской России, включавшей 49 губерний. Доля общих факторов, определявших механизм случайных колебаний цен на рожь, составляла 79—83%. Единый механизм случайных колебаний цен на овес включал в 80-е годы XIX в. территорию 46 губерний ($K=0{,}78+0{,}02$). Формирование товарного аграрного всероссийского рынка завершилось.

Остановимся вкратце на проблемах формирования аграрного капиталистического рынка.

Широкая и глубокая товаризация сельского хозяйства (и прежде всего крестьянского производства), отразившаяся в становлении единого товарного рынка, создавала решающие предпосылки для утверждения в сельском хозяйстве капитализма, так как ускоряла формирование рынка рабочей силы и средств производства. В этом смысле единый товарный рынок есть предпосылка для единого аграрного капиталистического рынка. Вместе с тем, если простой товарный рынок может формироваться в течение длительного периода на некапиталистической основе и только завершающая стадия формирования его происходит на базе капиталистического способа производства, то базисом складывания капиталистического аграрного рынка мог быть только капиталистический способ производства, ибо здесь товары обмениваются не просто как товары, но как продукты капиталов. Развитие рынков на землю, рабочий скот и рабочую силу приводит в конечном счете к тому, что стоимость каждого из них определяется общими (в рамках страны) факторами образования и движения цен, а именно общественно необходимыми затратами на их воспроизводство. Это приводит к складыванию средней нормы прибыли и цены производства (общественно необходимые затраты производства плюс средняя прибыль), которая выступает регулятором цен на сельскохозяйственную продукцию. Тем самым степень развития товарного аграрного рынка на средства производства непосредственно отражает степень развития капитализма в земледелии. Основные итоги анализа развития капиталистического аграрного рынка сводятся к следующему.

Цена земли не стала капиталистической рентой и не изменялась в полном соответствии с движением арендной платы и банковского процента. В виду полуфеодального характера землевладения, преобладания на земельном рынке дворянских земель земельный рынок оставался чрезвычайно узким и в начале XX в. не мог достичь высокого уровня развития даже в пределах определенных регионов.

Громадные изменения претерпевает рынок на рабочий скот. Уже в 80-е годы XIX в. имела место тесная взаимосвязь общих изменений цен и сформировались гигантские региональные рынки. В начале XX в. рынок на рабочий скот достиг чрезвычайно высокой степени интеграции и представлял собой уже единый всероссийский рынок в своей высшей фазе (на уровне случайных колебаний цен).

Наиболее интересен и важен итог по исследованию рынка рабочей силы. На низшем уровне, на уровне взаимосвязи фактических изменений цен, рынок рабочей силы в конце XIX — начале XX в. испытывает стремительные темпы развития. Если в 80-е годы XIX в. были лишь региональные рынки, то в начале XX в. формирование единого рынка рабочей силы по существу завершается. Движение цен на рабочую силу определялось на 55—64% общими социально-экономическими факторами, т. е. единой стоимостью и единой рыночной ценой рабочей силы и их изменениями. В своей высшей фазе, на уровне случайных колебаний цен, рынок рабочей силы претерпел стремительное превращение от совокупности региональных рынков, каждый из которых охватывал территорию 5—8 губерний, до выделения в начале XX в. гигантского региона в 18—20 губерний европейской России, на который «наплывали» четыре региональных рынка Центра, Поволжья, Украины, степного Юга и Юго-Запада. Это был канун их слияния в единый рынок на его высшей стадии, на уровне случайных колебаний цен.

Полученные итоги имеют большое значение для раскрытия уровня буржуазного аграрного развития России в конце XIX — начале XX в.

* *
*

Таким образом, применение количественных методов в изучении социально-экономических явлений и процессов имеет широкий диапазон и весьма значительные итоги в исследовательской практике.

Во-первых, применение этих методов расширяет возможности историков в установлении достоверности и обеспечении репрезентативности фактических данных, привлекаемых в исследовании. Во-вторых, системный подход и моделирование социально-экономических явлений и процессов открывают принципиально иные сравнительно с традиционными возможности в получении новой, непосредственно невыраженной информации об этих явлениях и процессах. Это позволяет проверять на эмпирических данных гипотезы, выдвигаемые для их объяснения, а следовательно, на качественно ином уровне раскрыть внутреннюю суть, закономерности и особенности функционирования и развития изучаемых явлений и процессов.

В-третьих, методы многомерного анализа открывают перед исследователями широкие возможности в решении такой сложнейшей задачи, как классификация объектов и явлений и анализ их внутренней структуры на основе использования всей исходной информации и оперирования полученными на ее основе обобщающими, интегральными показателями.

Короче говоря, количественные методы при глубоком теоретико-методологическом, содержательном подходе, при репрезентативности конкретно-исторических данных и корректном применении математических средств действительно дают возможность поднять изучение социально-экономических явлений и процессов, особенно массовых, на качественно новый уровень.

ГЛАВА 13

КОЛИЧЕСТВЕННЫЕ МЕТОДЫ АНАЛИЗА СОЦИАЛЬНО-ПОЛИТИЧЕСКИХ И ИСТОРИКО-КУЛЬТУРНЫХ ЯВЛЕНИЙ

Область социально-политическая, пожалуй, наиболее сложная и труднодоступная для применения количественных методов. Однако уровень и масштаб задач, поднимаемых современными гуманитарными науками (социологией, историей и т. п.) делает неизбежным применение современной методики и техники исследования и в этой области.

Социально-политические явления, как известно, отличаются чрезвычайной сложностью и многообразием взаимосвязей огромного количества элементов, в свою очередь образующих в рамках общей системы целый ряд подсистем. В качестве элементов таких систем могут выступать социальные группы, классы, государство и его учреждения, общественные организации и т. д. Среди этого буквально необъятного множества признаков и свойств сложно выделить самые существенные, абсолютно необходимые для решения той или иной поставленной исследователями задачи. С этим непосредственно связана задача разработки методологических и теоретических проблем политической истории, структуры политических систем и т. д. Кроме того, теоретические концепции политических отношений и процессов должны иметь четко разработанный категориальный и понятийный аппарат, что является условием успеха последующей формализации и количественного анализа, выбора адекватной теории (т. е. качественной содержательной модели) и математической модели или алгоритма для обработки на ЭВМ. Процесс подготовки к формализации и непосредственно процедура измерения весьма сложны и ввиду неразработанности операций измерения качественных признаков, которые часто трудно формализовать даже в простейших шкалах. Однако на макроуровне, т. е. в рамках крупномасштабного обобщающего историко-культурного исследования индивидуальные проявления культуры обнаруживают определенное сходство в части каких-то своих характеристик. Следовательно, огромное число объектов повторяемостью части своих свойств дает в итоге совокупность свойств, которая может быть определена как массовое явление или массовый процесс. Так, среди огромного числа сугубо индивидуальных явлений (таких, как, например, литературное произведение и др.), взятых в совокупности, мы можем выявить ряд общих черт (тематика, степень популярности тех или иных жанров, типы сюжетов и т. д.), которые образуют ряд «параметров», поддающихся счету. Таким образом, изучение культурных явлений на макроуровне, когда задачей исследователя служит выявление тех или иных общих характеристик или закономерностей, вполне органично связано с количественными методами.

Количественные методы могут с успехом углублять наши знания и в тех случаях, когда исследователь имеет дело с сугубо качественным явлением во всей его неповторимости и уникальности. Возможность применения количественных методов теоретически и практически реальна здесь потому, что, как уже говорилось, во всех общественных явлениях наблюдается единство количества и качества. Каждое качество имеет свои количественные характеристики, которые в принципе могут быть измерены. Для этого индивидуальное явление должно быть рассмотрено как определенная система с присущими ей элементами. Вся задача, следовательно, состоит в том, чтобы суметь найти и выделить эти элементы в соответствии с поставленной исследовательской задачей.

Таким образом, любое уникальное явление можно изучать «количественно», рассмотрев его как микросистему, элементы которой поддаются измерению.

В советской литературе уже накоплен определенный опыт применения количественных методов при изучении как массовых, так и индивидуальных социально-политических и историко-культурных явлений. Эти методы используются не только для раскрытия сути исследуемых явлений, но и при источниковедческой критике нарративных, прежде всего древних, источников. Центральной проблемой здесь является выяснение происхождения и установление авторства текстов.

§ 1. Формализованные методы в изучении происхождения и в атрибуции древних текстов

Примером применения количественных методов и ЭВМ в изучении происхождения нарративных источников может служить работа по построению «генеалогического древа» (стеммы) древнейшего юридического памятника славянского права IX в.— Закона Судного Людем (ЗСЛ). ЗСЛ — своеобразный раннехристианский, по своему типу юридический памятник, созданный одним из славянских просветителей Кириллом-Константином в пределах Велико-Моравского княжества, нашедший практическое применение в Болгарии конца IX — начала X в. Вместе с тем тексты этого памятника сохранились только на Руси, в составе древнерусских юридических сборников XIII—XVII вв. В связи с этим задача выявления наиболее древнего текста, т. е. текста, близкого к архетипу, является особенно актуальной.

ЗСЛ краткой, т. е. древнейшей, редакции сохранился в подавляющем большинстве списков в одних и тех же юридических сборниках с Русской Правдой. Сборники эти называются «Кормчими книгами». Соседство в книгах является не просто механическим. Как доказал в свое время М. Н. Тихомиров, разнообразные тексты каждой из кормчих книг носят следы единой редакции, единой правки, отражая общий подход редактора-юриста к содержащимся в кормчей юридическим памятникам. Вследствие этого существует возможность суждения о потенциальном сходстве генеалогии списков как Русской Правды, так и ЗСЛ.

Историками проведено фундаментальное текстологическое исследование ста с лишним списков Русской Правды Пространной редакции. Вся совокупность списков типологически (т. е. по степени общности текста) была разделена на три вида: 1) Синодально-Троицкий вид, списки которого сохранились в составе Кормчих книг и Мерил Праведных, представляющих разные варианты юридических сборников церковных и гражданских законов; 2) Пушкинско-Археографический вид, списки которого сохранились в юридических сборниках особого состава; 3) Карамзинский вид, списки которого сохранились в позднейших сборниках XV—XVI вв. В свою очередь, виды подразделяются на изводы. Так, Синодально-Троицкий вид (или группа) делится на Чудовский извод (или Розенкампфовский), Софийский извод (или Новгородско-Софийский), Ферапонтовский извод, извод Мерила Праведного и т. д. Одни из них считаются поздними по своему происхождению (например, Ферапонтовский извод), другие — ранними (извод Мерила Праведного). Вместе с тем исследователи, пользуясь традиционными методами «ручного» сличения разночтений (т. е. различных вариантов чтения одного и того же текста), не могли создать для Русской Правды Пространной редакции единой системы хронологического и генеалогического расположения всех списков, т. е. создать стемму, разветвленное древо генеалогии, где есть один исходный список, от которого в коиечном счете образовались все другие.

Трудностью в создании схемы генеалогии является прежде всего сложность в таких соотношениях списков, когда один список ведет свое происхождение не от какого-то определенного списка, а сразу от двух списков, т. е. является сводным списком, объединившим особенности текста обоих списков. Кроме того, важнейшим препятствием в создании «древа» является неспособность человека удерживать и, сопоставляя, определять генетическую последовательность нескольких тысяч разночтений из десятков списков. Это можно сделать только пользуясь ЭВМ.

Отсюда возникает задача создания программы автоматической классификации разночтений большого числа списков какого-либо одного древнего памятника. Принципиально существует реальная возможность перевода разночтений текста на основе теории множеств на формальный язык символов с дальнейшим построением генетического соотношения списков на основе этих разночтений. Из всей совокупности текстов выбирается условно один из них (желательно наиболее старший и исправный) в качестве так называемого «экземпляра ссылок», т. е. эталона, по отношению к которому все разночтения остальных списков трактуются либо как ошибки, либо как позднейшие изменения. Поскольку каждое слово текста изучаемого памятника закодировано, как закодированы его варианты и разночтения других списков, то в рамках программ ЭВМ возможно системное сравнение гигантской массы разночтений и выявление их «очередности» с точки зрения выделения ряда последовательностей в изменении каждого слова основного текста.

Этот метод называется «методом групп» и впервые был предложен французским текстологом Ж. Фроже. Метод применим не ко всем типам текстов, поскольку имеет ряд довольно жестких условий. Предлагаемая модель процесса генерации списков предполагает, что: а) у каждого списка есть только один «прямой предок», т. е. протограф; б) что в каждом списке имеются все ошибки или варианты протографа и в итоге происходит накопление их в отдаляющихся от протографа списках; в) одинаковые ошибки не содержатся в списках, имеющих в качестве своих протографов независимые друг от друга списки. Нарушение этих условий делает этот метод менее корректным.

Исходя из названных условий, в качестве материала для эксперимента, проведенного Л. И. Бородкиным и Л. В. Миловым [1], был взят текст Закона Судного Людем. Первое знакомство со списками ЗСЛ Краткой редакции показывает, что именно этот текст, как правило, бережно переписывавшийся русскими книжниками, максимально отвечает данным ограничениям. В то время как Русская Правда, имея принципиальное сходство в характере разночтений, отличается гораздо большим количеством сложных случаев, что и делает ее текст менее пригодным к данной процедуре. Для обработки на ЭВМ по созданной Л. И. Бородкиным программе было взято 54 списка этого памятника и закодировано более 15 тыс. вариантов разночтений. Все разночтения устанавливались на основе «экземпляра ссылок», в качестве которого был взят список ЗСЛ из Новгородской Кормчей 1280 г., т. е. самый древний список.

Работа программы ЭВМ рассчитана при необходимости на периодический контакт с экспертом, так как при работе могли встречаться так называемые конфликтные группы, когда эксперту нужно было выбрать из двух разночтений с примерно одинаковой частотой одно. Правда, таких случаев на практике было очень мало и касались они лишь одной ветви создаваемой стеммы — Софийского извода. В итоге работы была получена следующая стемма (рис. 31).

Как видно из стеммы, в итоге получена новая, огромная и принципиально важная информация. Важнейшая часть ее относится к условной «реконструкции» большого числа не сохранившихся до наших дней списков ЗСЛ (они на стемме обозначены кружками без номеров). По сравнению с известными списками (54) они увеличивают число их примерно на 60% (31 реконструкция). Больше того, каждая реконструкция поставлена на определенное место в генеалогии списков. Тем самым мы можем вполне реально представить себе то, что исчезло казалось бы навсегда [2].

Корректность полученной стеммы видна из того факта, что уцелевшие списки выстроились в цепях «древа» точно по фактической хронологии, хотя сведения о дате списков в машину не вводились.

---

[1] См.: *Бородкин Л. И., Милов Л. В.* О некоторых аспектах автоматизации текстологического исследования (Закон Судный Людем). — В кн.: Математические методы в историко-экономических и историко-культурных исследованиях.

[2] Подчеркнем, что на стемме реконструированы лишь списки, несущие часть анализируемых разночтений, а ведь огромное количество исчезнувших списков могло быть точной копией каждого из известных списков.

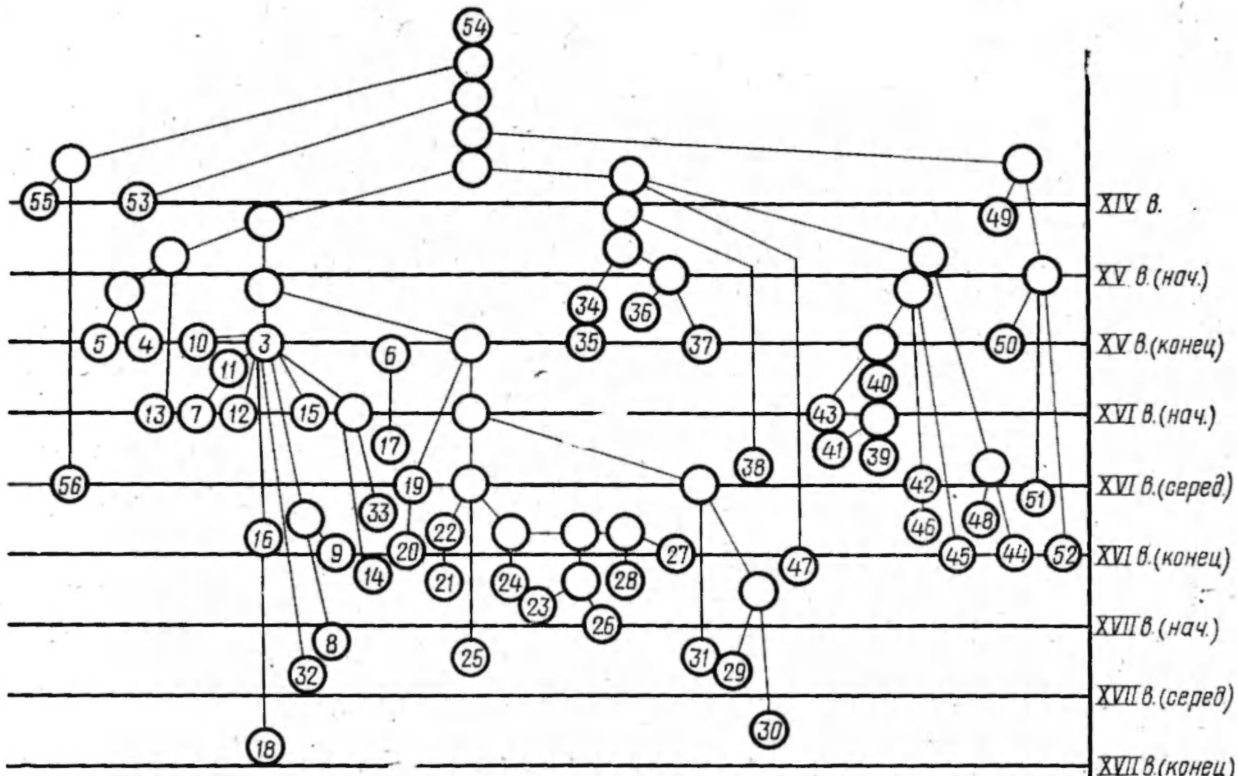

Рис. 31. Стемма «Закона Судного Людем»

Наконец, корректность стеммы доказывается тем обстоятельством, что стемма, созданная на основе разночтений ЗСЛ, одновременно является по существу и стеммой списков Русской Правды Синодально-Троицкой группы. Из всей совокупности списков только в Устюжской Кормчей, Иосифовской Кормчей и Варсонофьевской Кормчей нет текста Русской Правды. Отсюда следует, что оценка соотношения списков Русской Правды этой группы, данная историками-текстологами на основе традиционных методов, должна либо подтвердиться стеммой, созданной с помощью ЭВМ (и тогда ее корректность безукоризненна), либо не подтвердиться (и тогда наша стемма, видимо, негодна).

Анализ стеммы и сопоставления ее с классификацией списков Пространной Русской Правды, данной в академическом издании текстов Русской Правды, показывают их полное совпадение[1]. Основные группы ветвей древа по составу списков идентичны группам, составленным учеными «вручную», за исключением того, что исследователи формировали группы просто «списком», без определения их генеалогического соотношения. Больше того, машина «учла» даже оценки степени древности и чистоты текста, данные учеными тому или иному отдельному списку (например, Софийскому, Вязниковскому и т. д.). Стемма дает и принципиально новые материалы. Вопреки утвердившемуся в литературе, но слабо аргументированному суждению о том, что группа списков так называемого Ферапонтовского извода является лишь позднейшей чисто литературной традицией переписки текстов Русской Правды (а значит, и ЗСЛ), итоги автоматической классификации показывают весьма древнее происхождение этой ветви списков ЗСЛ, (и Русской Правды). Таким образом, списки Ферапонтовского извода были действующим источником права.

Как уже говорилось, стемма создавалась при условии, что «корнем» генеалогического древа a'priori был взят древнейший список Новгородской Кормчей 1280 г. После того как стемма создана, она нуждается в переориентации «корня» древа на основе экспертной работы историка, чтобы наиболее правильно избрать исходный список и к нему «подвесить» всю систему с учетом всех соотношений списков друг к другу, которые при этом не меняются.

В итоге, после перебора вариантов и изучения вопроса об исходном списке или исходном протографе, стемма при-

---

[1] См.: Правда Русская/Под ред. акад. Б. Д. Грекова. М.— Л., 1940 т. 1 (Тексты).

обретает вид древа, где исходной точкой является протограф Троицкого списка юридического сборника «Мерило Праведное». Больше того, проведенная классификация создала принципиальную возможность построения и стеммы для списков Синодально-Троицкой группы текстов Русской Правды, что дает принципиально новые возможности для исследования времени и места происхождения Пространной Русской Правды.

* *
*

К типу задач, в которых реализуется подход к индивидуальному явлению как определенной микросистеме с соответствующими элементами, поддающимися счету, относится и задача атрибуции древних памятников на основе изучения авторских особенностей текстов литературных и исторических произведений.

Необходимо напомнить, что область языкознания в широком смысле этого слова уже давно является объектом применения математических методов исследования. Выделилась особая наука: структурная лингвистика, синтезирующая методы лингвистики и математики, интенсивно применяются количественные методы в текстологии и т. д. Есть уже большое число работ по количественному изучению стилей языка и стилей речи, разрабатываются вопросы статистики речи, статистики лексикографии, составляются частотные словари писателей и т. д.

Названная выше задача изучения авторских особенностей текстов нарративных источников количественными методами была поставлена Л. И. Бородкиным, Л. В. Миловым, Л. Е. Морозовой применительно к памятникам Древней Руси [1]. Это одна из самых сложных проблем, имеющих свою специфику, вытекающую из специфики объекта исследования. Дело в том, что авторское начало довольно трудноуловимо в древнерусской литературе, а авторские произведения до XVII в. включительно почти никогда не имели авторской подписи, хотя авторство некоторых крупных дея-

[1] См.: *Бородкин Л. И., Милов Л. В., Морозова Л. Е.* К вопросу о формальном анализе авторских особенностей стиля в произведениях Древней Руси. — В кн.: Математические методы в историко-экономических и историко-культурных исследованиях.

телей средневековой культуры письменная традиция бережно хранила. Слабость авторского начала в древнерусских нарративных текстах обусловлена прямо или косвенно многими обстоятельствами. Среди них можно упомянуть и специфику эрудиции древнерусского «интеллигента» (а это были почти исключительно деятели церкви), которая была сопряжена с механическим запоминанием гигантского материала различного рода авторитетных текстов, религиозно-философского, литературно-религиозного содержания. Это и огромная роль цитирований, составляющих органическую часть средневекового текста. Это и большая распространенность компилятивного метода в творчестве того или иного автора. Наконец, это и бытование так называемого литературного этикета (определение Д. С. Лихачева), проявлявшегося, в частности, в своего рода канонизации определенных жанровых и стилистических приемов. До наших дней дошли, например, так называемые «формулярники» — целые сборники образцов текстов произведений различных средневековых жанров («Посланий», «Слов», «Похвальных слов» и т. п.).

Сложность проблемы выявления авторства связана и с вопросом чистоты авторского текста. Древнерусские авторские тексты были очень изменчивы в силу традиционного в ту эпоху вмешательства в текст со стороны переписчиков и редакторов, многие из которых в той или иной мере становились по существу соавторами текста. Этим объясняется, в частности, что наиболее популярные произведения Древней Руси, имевшие сотни (а может, и тысячи) списков (рукописных копий) сохранились во множестве редакций (кратких, пространных и т. д.). Так, «Житие» Александра Невского дошло до нас во многих сотнях списков пятнадцати редакций.

Вследствие названных, а также ряда иных обстоятельств вопрос об авторстве того или иного произведения является в литературе, как правило, либо спорным, либо до конца нерешенным. В дополнительной аргументации в пользу того или иного авторства нуждается, пожалуй, большинство из принятых в историографии атрибуций. В перспективе здесь еще огромное поле для работы. Не менее важной проблемой является установление авторства анонимных произведений, а если персонификация анонима невозможна, то не менее важна разработка классификации анонимов с присвоением каждому из них комплекса его индивидуальных признаков. В советской историографии эта тема совершенно не затронута исследованием.

Излагаемый здесь принцип фиксации авторских особенностей нарративного текста основан на следующих моментах.

Выявление авторских особенностей в нашей историографии осуществляется преимущественно путем обнаружения деталей стиля, присущих тому или иному автору, и, в частности, излюбленных слов, терминов, оборотов и выражений, т. е. субъективно-осознанных или субъективно признанных существенными для стиля автора. Этот путь имеет свои слабости, поскольку лексика автора диктуется жанром, а автор мог работать в разных жанрах. Кроме того, если у автора в языке нет явно выраженных «симпатий» и штампов, в распоряжении исследователя почти не остается никаких аргументов, кроме концепционных. Излишне напоминать, что автор может и подражать авторитету, и быть просто компилятором и т. д. Исходя из этого в названной разработке проблемы атрибуции нарративных текстов Древней Руси было решено сделать индикаторы стиля (а точнее, языка некоего автора), его подсознательные особенности. В качестве материала для выявления подсознательных особенностей письменного языка была взята структура предложения, вернее, специфика употребления и чередования каким-либо автором различных грамматических (морфологических) форм (существительных, прилагательных, глаголов и т. п.). При этом исследование не носит филологического характера, поскольку цель его в выявлении специфики чередования слов в предложении не как языкового явления, а как чисто индивидуальной авторской манеры. Введение в исследовательский материал грамматических форм есть лишь способ обобщения и упорядочения бесчисленного разнообразия лексики автора, есть способ **формализации** конкретного языкового материала, перевода его из конкретно-исторической формы в **абстрактную** форму, удобную для количественного анализа. По существу передача конкретной лексики какого-либо произведения есть процедура измерения по шкале наименований или классификационной шкале. Классы этой шкалы (которые, в свою очередь, при вводе в ЭВМ нумеруются) в данном случае получили наименования: «глагол», «наречие», «причастие» и т. д. (всего их 132).

В формализованном таким образом тексте затем выявлялись частоты парных встречаемостей (т. е. соседства) тех или иных грамматических классов (форм). Частота парных встречаемостей у различных авторов неизбежно различается. Исключением могут быть откровенные подражания

	1	2	3	4	5	6
1	5	4	3	8	2	1
2	3	2	1	2	3	7
3	8	1	3	9	2	3
4	3	4	5	6	4	6
5	1	3	2	3	1	4
6	5	6	7	2	1	3

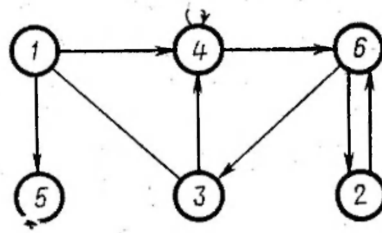

Рис. 32. Пример матрицы и графа

и «плагиаты», но и в этих случаях минимум индивидуальных особенностей должен обнаруживаться. В задаче учитывались лишь связи «слева-направо», т. е. в направлении развертывания текста, а не наоборот. В итоге получался огромный набор частот, из которых отбирали для исследования лишь наиболее весомые частоты парной встречаемости, достигающие известной величины («порога»), устанавливаемой экспериментально.

Для работы над текстом произведения создавалась выборка в 1000 значимых слов (размер ее установлен экспериментально). Этот массив информации подвергался формализации путем преобразования последовательности слов в последовательность грамматических классов. Далее материал обрабатывался по специальной программе на ЭВМ, после чего строилась матрица частот парной встречаемости (статистических связей) на уровне грамматических классов данного текста. Порог частоты (статистической связи) выбирался не ниже 5. Это означает, что в выборке на тысячу слов данное конкретное сочетание встретилось не менее пяти раз. Это частота, при столь ограниченном объеме текста, весьма высокая.

Полученная система статистических связей грамматических классов может служить основой для построения графа, т. е. формализации разных аспектов понятия совокупности статистических связей посредством языка теории графов. Граф состоит из «вершин» и «дуг», где вершина есть грамматический класс, а дуга — сильная связь (не ниже заданного порога). Слабые связи отсекаются при построении графа. Таким образом, чем выше задаваемый порог частоты встречаемости, тем меньше число дуг в графе.

Приведем пример построения графа сильных связей по условной матрице парной встречаемости (рис. 32). На матрице отражены частоты встречаемости в тексте шести грам-

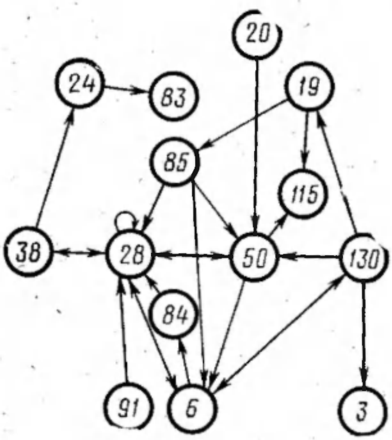

Рис. 33. Граф «Жития» Феодосия Печерского

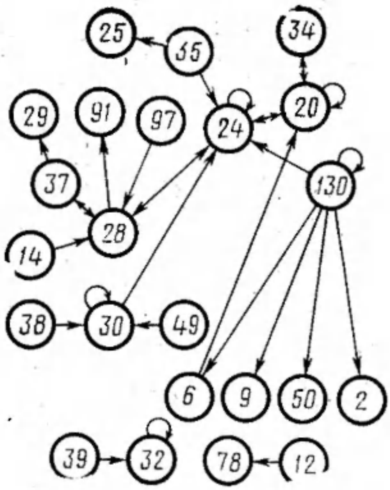

Рис. 34. Граф «Жития» Стефана Пермского

матических классов (1 — существительное, 2 — глагол, 3 — наречие и т. д.). Задаем порог встречаемости не ниже 6.

На рис. 32 видно, что частота выше порога в первой строке матрицы у пар 1 : 4 (8) и 1 : 5 (12). В соответствии с этим строятся вершины первого, четвертого и пятого классов, а дуги обозначают направление сильной связи. Соответственно фиксируются связи второй строки матрицы 2 : 6 (7), третьей строки — 3 : 1 (8), 3 : 4 (9), четвертой строки — 4 : 4 (6) и 4 : 6 (6). В пятой строке сильных связей нет, а в шестой — 6 : 2 (6), 6 : 3 (7). В образовавшемся графе вершины несут разную «нагрузку» по числу дуг. Вершины с большим числом дуг (не менее 4) именуются «узлами».

Приведем теперь примеры графов, отражающих частоты парных встречаемостей нескольких авторов Древней Руси: 1) Нестора-летописца — граф «Жития» Феодосия Печерского (конца XI — начала XII в.), 2) Епифания Премудрого — граф «Жития» Стефана Пермского (конец XIV в.), 3) Пахомия Серба (Логофета) — граф «Жития» Кирилла Белозерского (XV в.), 4) Иосифа Волоцкого — граф «Послания» И. И. Третьякову (первая четверть XVI в.), 5) Зиновия Отенского — граф выборки из полемического сочинения «Истины показание» (середина XVI в.) (рис. 33—37).

Внимательное знакомство с характером графов пяти крупнейших писателей Древней Руси XI—XVI вв. позволяет прежде всего установить наличие во всех графах так называемого «общеязыкового ядра», т. е. фундаментальных черт языка, отражающихся в инвариантных «узлах» всех графов парных встречаемостей. Такие дистрибутивные «узлы», как 28 (сущ. в винит. падеже), 24 (сущ. в родит. падеже), 20 (сущ. в имен. падеже), 130 (наречие), присутствуют во всех графах независимо от жанра. Вычет «общеязыковых» элементов графа дает в итоге индивидуальные черты того или иного автора либо в виде целых «узлов», либо в виде наиболее характерных связей. Кроме того, в рамках «общеязыковых» черт характер дистрибуции инвариантного узла у разных авторов может быть резко отличным. Так, в графе Нестора наряду с «общеязыковым ядром» обнаруживаются индивидуальные узлы (85 — указат. местоимение в винит. падеже; 50 — краткое причастие прошедш. времени). «Узел» 6 (глагол в аористе 3 лица) имеет весьма развитую систему связей, что является индивидуальной чертой. Примечательно сопоставление графов Епифания Премудрого — выдающегося писателя конца XIV — начала XV в. и Пахомия Серба, жившего позднее (ум. в 1485 г.) Пахомий известен не только как писатель, но и как редактор многих сочинений Епифания, а в некоторых случаях он кардинально их переделывал. Ученые предполагают , что «Житие» Стефана Пермского, изобретателя азбуки для народа коми, «уцелело» от пера Пахомия. Это заметно и по графу, имеющему богатую структуру детализации «общеязыковых узлов», чего совсем нет у Пахомия. Ярче у Епифания чередование однородных членов предложения (24 с 24, 20 с 20 и т. д.). Есть у него и кардинальные индивидуальные особенности (дистрибуция «узла» 28, «узел» 30 — сущ. в творит. падеже). Граф Пахомия беднее, хотя видна и его склонность к однородным членам предложения. Яркой индивидуальной чертой графа Пахомия выступает сильно развитая система связей «узла» 130. В графе Иосифа Волоцкого резко заметными особенностями выступают «узлы» 30 и 19, а также пристрастие к глаголу в аористе (6). Эта архаическая черта стиля Иосифа выступает, пожалуй, сильнее, чем у Нестора. Наконец, в графе писателя-публициста, правоверного борца с ересью Феодосия Косого, ка-

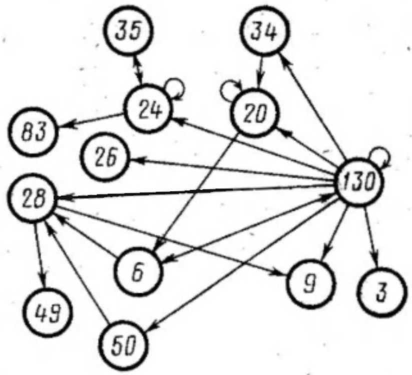

Рис. 35. Граф «Жития» Кирилла Белозерского

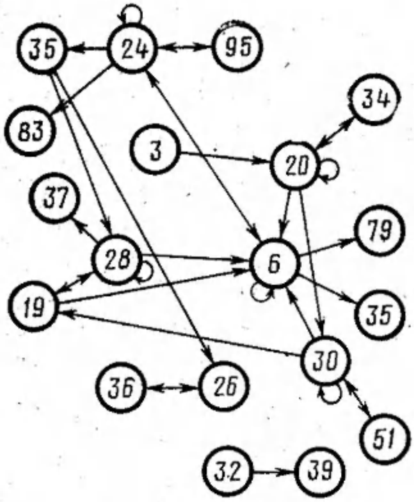

Рис. 36. Граф «Послания» И. И. Третьякову

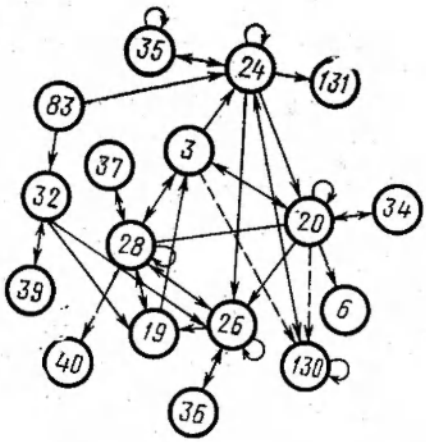

Рис. 37. Граф выборки из «Истины показание»

ким был Зиновий Отенский, сильно развиты специфические дистрибутивные «узлы» (26 — сущ. в дат. падеже, 19 — глагол в инфинитиве, 3 — глагол наст. времени 3 лица, 32 — сущ. в местном падеже).

Разумеется, графы разных произведений какого-либо одного автора не имеют буквального сходства (особенно, когда сопоставляются тексты разных жанров) и степень их общности неизбежно колеблется. Вместе с тем и в этих случаях получены позитивные результаты.

Данная методика способна выявить даже компиляцию и указать ее источник. На рис. 38 и 39 приведены графы «Жития» Леонтия Ростовского в третьей редакции, созданной, как предполагают, в XV в., и «Жития» Макария Калязинского, написанного в период подготовки в середине XVI в. «Великих Минеи — Четьи», этого грандиозного собрания житий, похвальных слов, посвященных канонизированным персонажам и деятелям христианства. Оба произведения компилятивны и источник вдохновения их авторов в конечном счете один — «Житие» Феодосия Печерского. Причем граф «Жития» Макария показывает, что автор его пользовался в первую очередь текстом жития Леонтия, а влияние Нестора здесь косвенное.

Как видим, количественные методы и ЭВМ дают возможность эффективно решать важнейшие проблемы источниковедческого анализа древних текстов, а именно выявить их происхождение и устанавливать авторство.

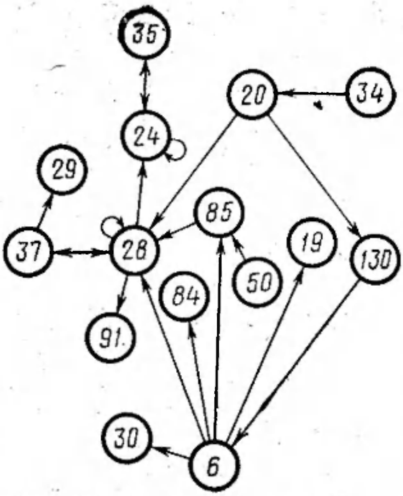

Рис. 38. Граф «Жития» Леонтия Ростовского

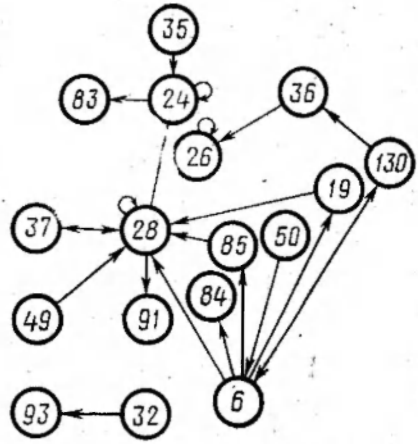

Рис. 39. Граф «Жития» Макария Калязинского

§ 2. Количественный анализ политических и культурных явлений

Количественные методы применяются при анализе как массовых, так и индивидуальных явлений социально-политического и историко-культурного характера. Остановимся на наиболее ярких примерах такого применения в советской историографии.

Интересной и перспективной является работа, выполненная Ю. Ю. Кахком и Х. М. Лиги[1]. Они выясняли зависимость размаха крестьянских волнений от экономического состояния крестьянского хозяйства и тяжести лежавших на них феодальных повинностей в начале XIX в.

Решение этой задачи потребовало от историков измерения положения крестьян. Для этого учитывалось количество дней, которые отрабатывал крестьянин у помещика на барщине до и после реформы 1804 г. Кроме того, вводились показатели относительной тяжести барщинных повинностей и обеспеченности крестьян землей и хлебом[2].

---

[1] См.: *Кахк Ю. Ю., Лиги Х. М.* О связи между антифеодальными выступлениями крестьян и их положением.— История СССР, 1976, № 2.

[2] В соответствии с задачей исследования было отобрано 49 имений, охваченных волнениями, которые равномерно распределялись по всем районам материковой части Южной Эстонии, и данные по 43 имениям, в которых волнения крестьян не наблюдались и которые относились к тем же районам. Учитывалось также распределение крестьян-

Методика расчета показателей относительной тяжести барщинных повинностей состояла в вычислении:

а) отношения количества дней, обрабатываемых на барщине, к общему числу дней, находящихся в течение лета в распоряжении крестьянских хозяйств, умноженному на количество упряжек, которое рассчитывалось, в свою очередь, на основе данных о рабочем скоте и числе трудоспособных мужчин в хозяйстве;

б) так называемого индекса рабочей силы, означавшего отношение суммы дней, которые необходимы крестьянину для обработки своих полей и барщинных дней с упряжкой, к общему числу имеющихся в распоряжении крестьянства дней с упряжкой. Если индекс меньше 1,0, то, следовательно, общее число дней, нужное для выполнения собственных работ и барщины, меньше числа имеющихся в их распоряжении дней и положение крестьян относительно лучшее; если же оно больше 1,0, то крестьянину требуется больше времени, чем есть в его распоряжении, т. е. он не сумеет исполнить барщину и качественно обработать свои поля;

в) объема остатка или недостатка хлеба, рассчитываемых на основе данных о размере пашни, урожайности и потребности в зерне на хозяйственные и продовольственные нужды [1].

Изучение зависимости между положением крестьянства и участием его в волнениях велось как на основе сравнения основных расчетных показателей по группам крестьянских хозяйств, охваченных волнениями, и «спокойных», так и путем корреляционного анализа показателей положения крестьян с показателем участия крестьян в волнениях. В данном случае качественный показатель участия или неучастия крестьян в волнениях был формализован путем придания охваченным волнениями имениям дополнительно параметра «1», а неохваченным —«0». Полученные коэффициенты корреляции были невысоки (от 0,35 до 0,56). Это свидетельствует об отсутствии прямой тесной связи между участием или неучастием крестьян в антифеодальных выступлениях и состоянием хозяйства и положением крестьян.

Исследователи пришли к выводу об отсутствии прямой связи между положением крестьян в каком-то имении и фактом участия в антифеодальных волнениях. Этот вывод

ских хозяйств по основным группам, которые выделялись по числу дней работы на барщине. Эта группа «двухдневных» и «трехдневных» хозяйств и т. д.

[1] Всего взято из источников и рассчитано исследователями 14 показателей положения крестьянства.

является, по мнению авторов, отражением такой характерной черты массовых антифеодальных волнений, как стихийность. Борьба крестьян несомненно была причинно связана с их положением и вырастала из беспросветной нужды народных масс. Но занимаясь историей классовой борьбы крестьянства, необходимо особое внимание уделять и поводам выступлений. Именно тот комплекс явлений, которые составляют сложную структуру повода (не причины), определяет, когда, в каких имениях и почему вспыхивают крестьянские волнения и из каких структурных элементов состоит мозаика крестьянских волнений. Видимо, речь должна идти о том, что положение всех крестьян к началу XIX в. достигло той степени угнетения, за которой их общая готовность к волнению могла быть реализована лишь с помощью того или иного повода, т. е. моментами уже не принципиальными. В 1802—1804 гг. крестьяне поднялись на борьбу вследствие убеждения, что они должны отказаться от выбора присяжных в волостные суды, ибо они якобы в противном случае лишаются льгот, обещанных царем, но скрытых от них дворянами. В 1820—1823 гг. они отказались дать согласие на навязанные им помещиками арендные контракты, чтобы не лишиться опять же царской милости — «обещанных земли и свободы». Таким образом, обращение к количественным методам помогло более глубоко вникнуть в механизм классовой борьбы крестьянства. Подобный подход применим и при изучении других массовых социально-политических явлений.

* *
*

Количественные методы могут успешно применяться при изучении не только массовых, но и индивидуальных социально-политических и историко-культурных явлений, зафиксированных в нарративных источниках. Эффективность такого применения демонстрирует опыт количественного анализа древней восточной летописи «Чуньцю»— источника, содержащего огромное количество отрывочных конкретных сведений по политической истории древнекитайских государств и их соседей в VIII— начале V в. до н. э. [1]. Несмотря на его большое значение, основные принципиальные

---

[1] См.: *Деопик Д. В.* Опыт количественного анализа древней восточной летописи «Чуньцю». — В кн.: Математические методы в историко-экономических и историко-культурных исследованиях.

вопросы источниковедческого характера не были решены. Это прежде всего вопрос о типе памятника: является ли он историко-философским произведением, содержащим изложение Конфуцием его философской концепции общества на материале истории государств бассейна среднего и нижнего течения Хуанхэ в VIII—начале V в. до н. э., или летописью, возникшей в одном из государств региона — царстве Лу. Значимость этого вопроса очевидна, ибо от его решения зависит установление характера изображения в нем реальной картины исторической действительности тех веков, определение достоверности и представительности содержащихся в нем сведений. Вот почему прежде всего перед исследователем стояла задача выработки методики анализа текста, позволяющей доказательно ответить на вопрос о типе памятника.

Придерживаясь точки зрения, что «Чуньцю» является летописью царства Лу, автор в своем доказательстве исходил из гипотезы о возрастании подробности изложения на протяжении всей летописи, как бы демонстрирующей естественное возникновение летописи по обычным законам этого исторического жанра. Каким же образом можно выявить эту закономерность? Очевидно, это возможно только с помощью структурного анализа текста памятника.

Количественный анализ структуры текста состоял в выделении основных элементов текста (иероглифы) и их изучения с помощью объективно измеряемых величин (число иероглифов, приходящихся на основные структурные единицы текста: год правления, сезон и др.). Выяснилось, что объем информации на год растет, увеличиваясь в поздних шести правлениях (6718 иероглифов на 114 лет в шести ранних правлениях и 9539 иероглифов на 130 лет в поздних).

Итоги оценки подробности изложения позволили сделать вывод о том, что текст памятника по данному критерию является летописью. Расчет объемов текста и его частей позволяет оценить представительность фактического материала в различные периоды для различных государств.

Решение источниковедческих задач было лишь предварительным этапом исследования «Чуньцю» как источника по политической истории древнекитайских государств. Далее перед историком встала задача — на основе разрозненных многочисленных мелких фактов (при отсутствии в источнике их обобщений, оценок, объяснений) воссоздать картину социально-политической истории. Исходя из специфики отражения социально-политических процессов в памятнике в форме фиксирования простейших действий и осуществляя подход к тексту «Чуньцю» как источнику, содержащему массовые данные о политической истории Дальнего Востока VIII—V вв. до н. э., автор выделил основные качественные элементы, характеризующие исследуемую политическую систему отношений. Ими являются так называемые «простейшие события». Например, «Ци убил вельможу» или «У воевало с Лу» и т. д. Эти события поддаются счету и может быть установлено их распределение во времени и пространстве, а также совместная встречаемость. Эта новая информация дает основу для выявления многих сторон и черт в жизни древнего дальневосточного общества.

Рассмотрим характер возможного исследования на примере анализа одного из аспектов социально-политической истории [1], а именно военной истории.

Содержание этой разновидности исторических явлений устанавливается на основе анализа «простейших событий», зафиксированных в тексте летописи. Сначала из текста были извлечены все простейшие события, содержательно относящиеся к аспекту военной истории. Они были сгруппированы в три основные подгруппы: 1) непосредственные военные действия (война, осада, постройка крепости, поход, военная помощь); 2) военные территориальные приобретения (вступление на территорию противника, захват страны); 3) битвы (сражения, взятие города). Так были систематизированы мелкие, бессистемные факты.

Следующий этап исследования состоял в **пространственно-временном анализе связей по простейшим событиям.** Рассмотрим его на примере анализа простейшего события —«война». Оно насчитывает 184 упоминания. Анализ его состоял в извлечении из текста указаний о субъектах, объектах войн и обстоятельствах времени и места и представлении этого материала в виде таблиц, на поле которых с помощью различных обозначений отображены отношения между 14 «основными» государствами эпохи и их племенами, отдельно для раннего (722—600 гг. до н. э.) и позднего (599—479 гг. до н. э.) периодов. Эти данные позволили увидеть ряд закономерностей в системе межгосударственных отношений азиатского дальневосточного общества, их развитие в различные исторические периоды.

---

[1] Анализ «Чуньцю» проведен Д. В. Деопиком по целому ряду аспектов: 1) внешняя политика, 2) военная история, 3) быт монархов, 4) внутренняя политика, 5) экономика, 6) сакральные сюжеты, 7) природные явления.

Так выявилось плотное ядро государств по среднему течению р. Хуанхэ, которые между собой либо совсем не воевали, либо сражались вместе в рядах одной каолиции. Им противостояли государства, не образующие ядра, а враждующие и с названной группой и между собой. Это сильные периферийные государства на юге и западе.

Основная тенденция развития межгосударственных отношений состояла в поляризации, обострении противостояния центрального «блока» его врагам и размывании промежуточных групп. Так, в поздний период военные отношения становятся более определенными: пар, воюющих «всегда вместе» или «всегда против», в поздний период было 38 против 27 в ранний, а пар, у которых число совместных выступлений равно числу конфликтов, меньше в поздний период в 11 раз.

Весьма эффективным в работе оказалось применение графов при описании конкретных областей политической истории, например, для анализа устойчивых групп государств, характерных для раннего и позднего периодов. На них хорошо видны группы государств, их постоянные и временные союзники, соотношения между группами. Текстовая запись таких закономерностей для всей эпохи почти невозможна и даже в частичном виде очень громоздка, да и сложна, тем более что в тексте ни разу не сказано, кто кому друг, кто враг.

Таким образом, в результате количественного анализа текста «Чуньцю» по многочисленным «мелким» фактам можно проследить определенные общие и частные тенденции в истории общества VIII—V вв. до н. э. Анализ показал возникновение обширной системы изолированных единиц — государств, отношения между которыми все в меньшей степени нормировались традицией и соглашениями. Со всей очевидностью проявлялся процесс слома старой системы групп средних и мелких государств, связанных через отношения монархов, и формирование сравнительно единого социально-политического комплекса за счет роста различных новых связей между государствами.

Как видим, применение количественных методов позволило значительно углубить анализ различного рода социально-политических процессов, зафиксированных в сугубо индивидуальном описательном историческом источнике.

* *
*

Примером успешного применения количественных методов при изучении историко-культурных явлений на основе нарративных источников является анализ философско-этических трактатов итальянских гуманистов XV в. Кристофоро Ландино («Об истинном благородстве») и Джованни Нези («О нравах»), проделанный Л. М. Брагиной[1]. Именно в этой работе был реализован подход к каждому из индивидуальных трактатов как к сложному объекту с выделением в качестве его основных характеристик совокупности его терминов. Причем, характер и направленность отбираемых терминов как единицы измерения должны строго соответствовать задаче, которую исследователь ставит перед собой.

При изучении трактата «О благородстве» ставилась задача возможно более полного выяснения всей палитры оттенков значения термина nobilitas, nobilis — благородство. Толкование этого термина не должно привносить современные нам аспекты интерпретации. А поскольку термин благородство выступает в трактате в огромном многообразии сочетаний с другими терминами, то синтезировать этот сложный и разнообразный материал, идя традиционными методами логико-содержательного анализа, не представляется возможным. Такая же задача была поставлена и при изучении трактата «О нравах», где основная идея автора заключена в мысли о том, что добродетель — главный путь к счастью.

Первый этап исследования состоял в составлении частотного словаря терминов. При этом делалась опора на существительные, хотя привлекались и наречия, и прилагательные, и некоторые глагольные формы. Критерием отбора здесь служила логическая нагрузка термина. Частотный словарь выявляет главные термины обоих трактатов. Для трактата «О благородстве» — это термин благородство (nobilitas, nobilis) и добродетель (virtus), частота которых соответственно 190 и 89. В трактате «О нравах» это термины добродетель (virtus) и счастье (foelicitas ets), частота которых 429 и 305. Выделение главных терминов трактата возможно потому, что круг понятий, несущих основную концепционную нагрузку, сравнительно невелик.

Поскольку оба трактата отличаются сюжетной строгостью и одноплановостью и авторы их придерживаются чет-

---

[1] См.: *Брагина Л. М.* Методика количественного анализа философских трактатов эпохи Возрождения. — В кн.: Математические методы в историко-экономических и историко-культурных исследованиях.

кой терминологии, тем самым создается основа для статистической обработки частотного словаря.

Первым этапом его обработки является процедура «уплотнения словаря», которая состояла в группировке терминов, близких или адекватных по своей смысловой, семантической нагрузке. В итоге этой процедуры, например, термины civitas, coetus hominum, respublica, societas выступают как один термин со значением государство, общество. Так же объединяются и другие термины. Таким образом в основу группировки или «уплотнения» кладется так называемое «лексико-семантическое поле».

Следующим этапом работы был поиск показателя взаимосвязи терминов, их сопряженности. Сопряженность их выступает как частота совместного употребления пары каких-то терминов. Критерий совместного употребления лежит в границах контекста, очерченных тремя фразами, если в центральной из них фигурирует главный изучаемый термин. Таким образом, частота совместного употребления каких-либо пар терминов ограничена «полем контекста» из трех фраз.

Частоты сопряженности каждого из основных терминов образуют целую систему количественных характеристик, которая служит основой для математико-статистического анализа текстов и, в частности, корреляционного анализа. Для того чтобы стал возможным подсчет коэффициентов корреляции, необходимо для какой-либо пары главных терминов (например, а) nobilitas, nobilis и б) virtus) выделить такую комбинацию сопряженных терминов, которая была бы общей для обоих терминов. Таким образом, изучается не просто частота встречаемости двух терминов **непосредственно**, а изучается частота встречаемости этих терминов **опосредованно**, через сопряженные термины, отражающие всю совокупность оттенков, в которых тем самым выступает основной изучаемой термин (nobilitas, nobilis) и сопряженный термин (в данном случае virtus). Набор сопряженных терминов для обоих указанных главных терминов выглядит следующим образом (см. табл. на с. 367).

Из таблицы видно, что число сопряженных терминов, общих для пары указанных главных терминов, равно 14. Эта систематизация ясно убеждает в том, что как термин nobilitas, так и термин virtus совместно употребляются с 14 общими терминами, но частота их употребления неодинакова. Сказать же, в какой мере вариация сопряжена с другим термином без подсчета коэффициента корреляции, невозможно. Подсчет же коэффициента корреляции дает

Сопряженные термины	nobilitas	virtus	Сопряженные термины	nobilitas	virtus
1. Actiones	5	6	7 Dignitas	20	10
2 Animus	30	25	8 Genus, origo	45	17
3. Ars; industria, labor, opera	20	8	9 Gloria, fama	15	6
			10 Honor, honestas	9	11
4. Beatitudo, foelicitas	17	10	11 Ingenium	14	12
5 Bonum	7	4	12 Mens	16	12
6 Civitas, coetus hominum recpublica societas	63	26	13 Ratio	23	15
			14 Sapientia, scietia	21	8

точную меру тесноты этой взаимосвязи (0,83). Взаимосвязь, как видим, очень сильная.

Точно так же подсчитывается теснота взаимосвязи всех главных терминов трактата. Во всех случаях, с одной стороны, фигурирует термин nobilitas, nobilis, а с другой — поочередно все остальные термины. В итоге все сопряженные термины ранжируются по степени взаимосвязи с главным и изучаемым термином (nobilitas, nobiils). Самая сильная по уровню коэффициента взаимосвязь термина nobilitas с термином virtus (0,83), далее идет термин ars (0,81), затем термин genus (0,77) и т. д.

Таким образом, в рассуждениях Ландино в содержании понятия «благородство» главную роль играет добродетель (virtus), потом (в порядке убывания силы связи) творчество (ars ets), происхождение (genus, origo), мудрость, знание (Sapientia ets), общество, государство (civitas ets), душа (anima), достоинство (dignitas), разум (ratio) и т. д.

В обоих трактатах исследовано таким образом внушительное количество терминов. Ранжированные по тесноте взаимосвязи ряды, естественно, не могут быть сами по себе непосредственным итогом работы. Они лишь дают информацию для содержательного анализа, который и раскрывает суть воззрений авторов трактатов. Однако охарактеризованную выше стадию работы успешно выполнить путем традиционной методики не представляется возможным.

Приведенные примеры свидетельствуют об эффективности, перспективности применения количественных методов при изучении социально-политических и историко-культурных явлений, отраженных в нарративных исторических источниках.

* *
*

Краткий обзор основных направлений и некоторых конкретных итогов применения количественных методов в исследованиях советских историков позволяет сделать ряд выводов, которые подтверждают основные теоретико-методологические и методические положения, изложенные в предыдущих частях работы.

Прежде всего очевидна возможность и эффективность применения количественных методов в изучении самых различных объектов, явлений и процессов исторического прошлого, зафиксированных как в массовых, так и индивидуальных исторических источниках и выраженных как количественно, так и описательно.

Во-первых, применение количественных методов открывает возможность более надежно решать ряд источниковедческих проблем, а именно устанавливать достоверность и репрезентативность массовых количественных данных, а также выявлять происхождение и авторство нарративных памятников.

Во-вторых, используя систему количественных показателей и подвергая их математической обработке, историки получают такую информацию, которая не может быть выявлена описательными методами. Все это позволяет строить количественные модели этих явлений и процессов, наиболее адекватно и глубоко характеризующие их внутреннюю суть, функционирование и развитие.

Весь опыт применения количественных методов советскими историками подтверждает то теоретическое несомненное положение, что количественные методы, как и методы описательные, могут дать должный эффект, т. е. продвинуть историков по пути углубления знаний о прошлом лишь в том случае, когда формально-количественный анализ сочетается с качественным, сущностно-содержательным, а этот последний опирается на такую теорию и методологию исторического познания, которые допускают возможность и требуют получения в полной мере объективных знаний об этом прошлом. Именно превосходством марксистской теории и методологии исторического познания обусловлено то, что применение количественных методов в советской историографии имеет более органичный и основательный характер, приносит большие научные результаты и не влечет за собой тех существенных издержек и негативных явлений, которые имеют место в буржуазной историографии.

Наконец, практика показывает, что наряду с обоснованностью и глубиной сущностно-содержательного подхода успех и эффективность количественных методов определяются адекватностью избранных математических средств и корректностью их применения. Те издержки, которые имеют место в применении количественных методов в советской историографии, как правило, и обусловлены некорректностью решения указанных конкретно-методологических и методических проблем.

В заключение еще раз отметим, что настоящая работа имеет своей целью охарактеризовать основной круг теоретико-методологических, научно-методических и конкретно-исторических проблем, возникающих в связи с применением количественных методов в исторических исследованиях. Более глубокое уяснение этих проблем, так же как и овладение навыками применения наиболее сложных из охарактеризованных методов, естественно требует ознакомления со специальной литературой и совместной работы историков с математиками.

ПРИЛОЖЕНИЕ

Таблица 1. Значения функции $\Phi(t)$

t	0	2	4	6	8
0,0	00000	01596	03191	04784	06376
0,1	07966	09552	11134	12712	14285
0,2	15852	17413	18967	20514	22052
0,3	23582	25103	26614	28115	29605
0,4	31084	32551	34006	35448	36877
0,5	38292	39694	41080	42452	43809
0,6	45149	46474	47783	49075	50350
0,7	51607	52848	54070	65275	56461
0,8	57629	58778	59909	61021	62114
0,9	63188	64243	65278	66294	67291
1,0	68269	69227	70166	71086	71986
1,1	72867	73729	74571	75395	76200
1,2	76986	77754	78502	79233	79945
1,3	80640	81316	81975	82617	83241
1,4	83849	84439	85013	85571	86113
1,5	86639	87149	87644	88124	88589
1,6	89040	89477	89899	90309	90704
1,7	91087	91457	91814	92159	92492
1,8	92814	93124	93423	93711	93989
1,9	94257	94514	94762	95000	95230
2,0	95450	95662	95865	96060	96247
2,1	96427	96599	96765	96923	97074
2,2	97219	97358	97491	97618	97739
2,3	97855	97966	98072	98172	98269
2,4	98360	98448	98531	98611	98686
2,5	98758	98826	98891	98953	99012
2,6	99068	99121	99171	99219	99264
2,7	99307	99347	99386	99422	99456
2,8	99489	99520	99549	99576	99602
2,9	99627	99650	99672	99692	99712
3,0	99730	99747	99763	99779	99793
3,1	99806	99819	99831	99842	99853
3,2	99863	99872	99880	99889	99896
3,3	99903	99910	99916	99922	99928
3,4	99933	99937	99942	99946	99950
3,5	99953	99957	99960	99963	99966
3,6	99968	99971	99973	99975	99977
3,7	99978	99980	99982	99983	99984
3,8	99986	99987	99988	99989	99990
3,9	99990	99991	99992	99992	99993
4	99994	99997	99999	—	—

Примечание. В таблицах, там, где это не вызывает сомнений, приведены лишь дробные части значений функций.

Таблица 2. Распределение нормированных отклонений в малой выборке t (k)

k	$1-\alpha=P$					k	$1-\alpha=P$				
	0,9	0,95	0,98	0,99	0,999		0,9	0,95	0,98	0,99	0,999
1	6,31	12,71	31,82	63,66	63,62	17	1,74	2,11	2,57	2,90	3,97
2	2,92	4,30	6,97	9,93	31,60	18	1,73	2,10	2,55	2,88	3,92
3	2,35	3,18	4,54	5,84	12,94	19	1,73	2,09	2,54	2,86	3,88
4	2,13	2,78	3,75	4,60	8,61	20	1,72	2,09	2,53	2,85	3,85
5	2,02	2,57	3,37	4,03	6,86	21	1,72	2,08	2,52	2,83	3,82
6	1,94	2,45	3,14	3,70	5,96	22	1,72	2,07	2,51	2,82	3,79
7	1,90	2,37	3,00	3,50	5,40	23	1,71	2,07	2,50	2,81	3,77
8	1,86	2,30	2,90	3,36	5,04	24	1,71	2,06	2,49	2,80	3,75
9	1,83	2,26	2,82	3,25	4,78	25	1,71	2,06	2,49	2,79	3,72
10	1,81	2,23	2,76	3,17	4,59	26	1,71	2,06	2,48	2,78	3,71
11	1,80	2,20	2,72	3,11	4,49	27	1,70	2,05	2,47	2,77	3,69
12	1,78	2,18	2,68	3,06	4,32	28	1,70	2,05	2,47	2,76	3,67
13	1,77	2,18	2,65	3,06	4,12	29	1,70	2,05	2,46	2,76	3,66
14	1,76	2,14	2,62	2,98	4,14	30	1,70	2,04	2,46	2,75	3,65
15	1,75	2,13	2,60	2,95	4,07	40	1,68	2,02	2,42	2,70	3,55
16	1,75	2,12	2,58	2,92	4,02	60	1,67	2,00	2,39	2,66	3,36
						120	1,66	1,98	2,36	2,62	3,37
						∞	1,65	1,96	2,33	2,58	3,29

Таблица 3. Значения χ^2, соответствующие значениям α и числам степеней свободы k

k	α							
	0,99	0,95	0,90	0,25	0,10	0,05	0,025	0,01
1	0,032	0,024	0,02	1,32	2,71	3,84	5,02	6,63
2	0,02	0,10	0,21	2,77	4,61	5,99	7,38	9,21
3	0,12	0,35	0,58	4,11	6,25	7,81	9,35	11,3
4	0,30	0,71	1,06	5,39	7,78	9,49	11,1	13,3
5	0,55	1,15	1,61	6,63	9,24	11,1	12,8	15,1
6	0,87	1,64	2,20	7,84	10,6	12,6	14,4	16,8
7	1,24	2,17	2,83	9,04	12,0	14,1	16,0	18,5
8	1,65	2,73	3,49	10,2	13,4	15,5	17,5	20,1
9	2,09	3,33	4,17	11,4	14,7	16,9	19,0	21,7
10	2,56	3,94	4,87	12,5	16,0	18,3	20,5	23,2
11	3,05	4,57	5,58	13,7	17,3	19,7	21,9	24,7
12	3,57	5,23	6,30	14,8	18,5	21,0	23,3	26,2
13	4,11	5,89	7,04	16,0	19,8	22,4	24,7	27,7

Продолжение табл. 3

k	α							
	0,99	0,95	0,90	0,25	0,10	0,05	0,025	0,01
14	4,66	6,57	7,79	17,1	21,1	23,7	26,1	29,1
15	5,23	7,26	8,55	18,2	22,3	25,0	27,5	30,6
16	5,81	6,96	9,31	19,4	23,5	26,3	28,8	32,0
17	6,41	8,67	10,1	20,5	24,8	27,6	30,2	33,4
18	7,01	9,39	10,9	21,6	26,0	28,9	31,5	34,8
19	7,63	10,1	11,7	22,7	27,2	30,1	32,9	36,2
20	8,26	10,9	12,4	23,8	28,4	31,4	34,2	37,6
21	8,90	11,6	13,2	24,9	29,6	32,7	35,5	38,9
22	9,54	12,3	14,0	26,0	30,8	33,9	36,8	40,3
23	10,2	13,1	14,8	27,1	32,0	35,2	38,1	41,6
24	10,9	13,8	15,7	28,2	33,2	36,4	39,4	43,0
25	11,5	14,6	16,5	29,3	34,4	37,7	40,6	44,3
26	12,2	15,4	17,3	30,4	35,6	38,9	41,9	45,6
27	12,9	16,2	18,1	31,5	36,7	40,1	43,2	47,0
28	13,6	16,9	18,9	32,6	37,9	41,3	44,5	48,3
29	14,3	17,7	19,8	33,7	39,1	42,6	45,7	49,6
30	15,0	18,5	20,6	34,8	40,3	43,8	47,0	50,9

Таблица 4А. Доверительные границы для $F[k_1;\ k_2]$ с уровнем значимости 5%; k_1 — число степеней свободы числителя, k_2 — число степеней свободы знаменателя

k_2	k_1 (число степеней свободы числителя)							
	1	2	3	4	5	8	11	15
1	161	200	216	225	230	239	243	246
2	18,5	19,0	19,2	19,2	19,3	19,4	19,4	19,4
3	10,1	9,55	9,28	9,12	9,01	8,85	8,76	8,70
4	7,71	6,94	6,59	6,39	6,26	6,04	5,94	5,86
5	6,61	5,79	5,41	5,19	5,05	4,82	4,70	4,62
6	5,99	5,14	4,76	4,53	4,39	4,15	4,03	3,94
7	5,59	4,74	4,35	4,12	3,97	3,73	3,60	3,51
8	5,32	4,46	4,07	3,84	3,69	3,44	3,31	3,22
9	5,12	4,26	3,86	3,63	3,48	3,23	3,10	3,01
10	4,96	4,10	3,71	3,48	3,33	3,07	2,94	2,85
11	4,84	3,98	3,59	3,36	3,20	2,95	2,82	2,72
12	4,75	3,89	3,49	3,26	3,11	2,85	2,72	2,62
13	4,67	3,81	3,41	3,18	3,03	2,77	2,63	2,53
14	4,60	3,74	3,34	3,11	2,96	2,70	2,57	2,46
15	4,54	3,68	3,29	3,06	2,90	2,64	2,51	2,40
16	4,49	3,63	3,24	3,01	2,85	2,59	2,46	2,35
17	4,45	3,59	3,20	2,96	2,81	2,55	2,41	2,31

k_2	k_1 (число степеней свободы числителя)							
	1	2	3	4	5	8	11	15
18	4,41	3,55	3,16	2,93	2,77	2,51	2,37	2,27
19	4,38	3,52	3,13	2,90	2,74	2,48	2,34	2,23
20	4,35	3,49	3,10	2,87	2,71	2,45	2,31	2,20
21	4,32	3,47	3,07	2,84	2,68	2,42	2,28	2,18
22	4,30	3,44	3,05	2,82	2,66	2,40	2,26	2,15
23	4,28	3,42	3,03	2,80	2,64	2,37	2,23	2,13
24	4,26	3,40	3,01	2,78	2,62	2,36	2,21	2,11
25	4,24	3,39	2,99	2,76	2,60	2,34	2,20	2,09
26	4,23	3,37	2,98	2,74	2,59	2,32	2,18	2,07
27	4,21	3,35	2,96	2,73	2,57	2,31	2,17	2,06
28	4,20	3,34	2,95	2,71	2,56	2,29	2,15	2,04
29	4,18	3,33	2,93	2,70	2,55	2,28	2,14	2,03
30	4,17	3,32	2,92	2,69	2,53	2,27	2,13	2,01
32	4,15	3,29	2,90	2,67	2,51	2,24	2,10	1,99
34	4,13	3,28	2,88	2,65	2,49	2,23	2,08	1,97
36	4,11	3,26	2,87	2,63	2,48	2,21	2,07	1,95
38	4,10	3,24	2,85	2,62	2,46	2,19	2,05	1,94
40	4,08	3,23	2,84	2,61	2,45	2,18	2,04	1,92
42	4,07	3,22	2,83	2,59	2,44	2,17	2,03	1,91
44	4,06	3,21	2,82	2,58	2,43	2,16	2,01	1,90
46	4,05	3,20	2,81	2,57	2,42	2,15	2,00	1,89
48	4,04	3,19	2,80	2,57	2,41	2,14	1,99	1,88
50	4,03	3,18	2,79	2,56	2,40	2,13	1,99	1,87
60	4,00	3,15	2,76	2,53	2,37	2,10	1,95	1,84
70	3,98	3,13	2,74	2,50	2,35	2,07	1,93	1,81
80	3,96	3,11	2,72	2,49	2,33	2,06	1,91	1,79
90	3,95	3,10	2,71	2,47	2,32	2,04	1,90	1,78
100	3,94	3,09	2,70	2,46	2,31	2,03	1,89	1,77
125	3,92	3,07	2,68	2,44	2,29	2,01	1,87	1,75
150	3,90	3,06	2,66	2,43	2,27	2,00	1,85	1,73
200	3,89	3,04	2,65	2,42	2,26	1,98	1,84	1,72
300	3,87	3,03	2,63	2,40	2,24	1,97	1,82	1,70
500	3,86	3,01	2,62	2,39	2,23	1,96	1,81	1,69
1000	3,85	3,00	2,61	2,38	2,22	1,95	1,80	1,68

Таблица 4Б. Доверительные границы для $F[k_1; k_2]$ с уровнем значимости 1%; k_1 — число степеней свободы числителя, k_2 — число степеней свободы знаменателя

k_2	k_1 (число степеней свободы числителя)							
	1	2	3	4	5	8	11	15
2	98,5	99,0	99,2	99,2	99,3	99,4	99,4	99,4
3	34,1	30,8	29,5	28,7	28,2	27,5	27,1	26,9
4	21,2	18,0	16,7	16,0	15,5	14,8	24,4	14,2

k_2	k_1 (число степеней свободы числителя)							
	1	2	3	4	5	8	11	15
5	16,3	13,3	12,1	11,4	11,0	10,3	9,96	9,72
6	13,7	10,9	9,78	9,15	8,75	8,10	7,79	7,56
7	12,2	9,55	8,45	7,85	7,46	6,84	6,54	6,31
8	11,3	8,65	7,59	7,01	6,63	6,03	5,73	5,52
9	10,6	8,02	6,99	6,42	6,06	5,47	5,18	4,96
10	10,0	7,56	6,55	5,99	5,64	5,06	4,77	4,56
11	9,65	7,21	6,22	5,67	5,32	4,74	4,46	4,25
12	9,33	6,93	5,95	5,41	5,06	4,50	4,22	4,01
13	9,07	6,70	5,74	5,21	4,86	4,30	4,02	3,82
14	8,86	6,51	5,56	5,04	4,69	4,14	3,86	3,66
15	8,63	6,36	5,42	4,89	4,56	4,00	3,73	3,52
16	8,53	6,23	5,29	4,77	4,44	3,89	4,62	3,41
17	8,40	6,11	5,18	4,67	4,34	3,79	3,52	3,31
18	8,29	6,01	5,09	4,58	4,25	3,71	3,43	3,23
19	8,18	5,93	5,01	4,50	4,17	3,63	3,36	3,15
20	8,10	5,85	4,94	4,43	4,10	3,56	3,29	3,09
21	8,02	5,78	4,87	4,37	4,04	3,51	3,24	3,03
22	7,95	5,72	4,82	4,31	3,99	3,45	3,18	2,98
23	7,88	5,66	4,76	4,26	3,94	3,41	3,14	2,93
24	7,82	5,61	4,72	4,22	3,90	3,36	3,09	2,89
25	7,77	5,57	4,68	4,18	3,86	3,32	3,06	2,85
26	7,72	5,53	4,64	4,14	3,82	3,29	3,02	2,82
27	7,68	5,49	4,60	4,11	3,78	3,26	2,99	2,78
28	7,64	5,45	4,57	4,07	3,75	3,23	2,96	2,75
29	7,60	5,42	4,54	4,04	3,73	3,20	2,93	2,73
30	7,56	5,39	4,51	4,02	3,70	3,27	2,91	2,70
32	7,50	5,34	4,46	3,97	3,65	3,13	2,86	2,66
34	7,44	5,29	4,42	3,93	3,61	3,09	2,82	2,62
36	7,40	5,25	4,38	3,89	3,57	3,05	2,79	2,58
38	7,35	5,21	4,34	3,86	3,54	3,02	2,75	2,55
40	7,31	5,18	4,31	3,83	3,51	2,99	2,73	2,52
42	7,28	5,15	4,29	3,80	2,49	2,97	2,70	2,50
44	7,25	5,12	4,26	3,78	3,47	2,95	2,68	2,47
46	7,22	5,10	4,24	3,76	3,44	2,93	2,66	2,45
48	7,19	5,08	4,22	3,74	3,43	2,91	2,64	2,44
50	7,17	5,06	4,20	3,72	3,41	2,89	2,63	2,42
55	7,12	5,01	4,16	3,68	3,37	2,85	2,59	2,38
60	7,08	4,98	4,13	3,65	3,34	2,82	2,56	2,35
70	7,01	4,92	4,08	3,60	3,29	2,78	2,51	2,31
80	6,96	4,88	4,04	3,56	3,26	2,74	2,48	2,27
90	6,93	4,85	4,01	3,54	3,23	2,72	2,45	2,24
100	6,90	4,82	3,98	3,51	3,21	2,69	2,43	2,22
125	6,84	4,78	3,94	3,47	3,17	2,66	2,39	2,19
150	6,81	4,75	3,92	3,45	3,14	2,63	2,37	2,16
200	6,76	4,71	3,88	3,41	3,11	2,60	2,34	2,13
300	6,72	4,68	3,85	3,38	3,08	2,57	2,31	2,10
500	6,69	4,65	3,82	3,36	3,05	2,55	2,28	2,07
1000	6,66	4,63	3,80	3,34	3,04	2,53	2,27	2,06

Таблица 5. Границы критической области для критерия знаков

n	α						n	α					
	5%		2%		1%			5%		2%		1%	
5	0	5	0	5	0	5	45	16	29	15	30	14	31
6	1	5	0	6	0	6	46	16	30	15	31	14	32
7	1	6	1	6	0	7	47	17	30	16	31	15	32
8	1	7	1	7	1	7	48	17	31	16	32	15	33
9	2	7	1	8	1	8	49	18	31	16	33	16	33
10	2	8	1	9	1	9	50	18	32	17	33	16	34
11	2	9	2	9	1	10	51	19	32	17	34	16	35
12	3	9	2	10	2	10	52	19	33	18	34	17	35
13	3	10	2	11	2	11	53	19	34	18	35	17	36
14	3	11	3	11	2	12	54	20	34	19	35	18	36
15	4	11	3	12	3	12	55	20	35	19	36	18	37
16	4	12	3	13	3	13	56	21	35	19	37	18	38
17	5	12	4	13	3	14	57	21	36	20	37	19	38
18	5	13	4	14	4	14	58	22	36	20	38	19	39
19	5	14	5	14	4	15	59	22	37	21	38	20	39
20	6	14	5	15	4	16	60	22	38	21	39	20	40
21	6	15	5	16	5	16	61	23	38	21	40	21	40
22	6	16	6	16	5	17	62	23	39	22	40	21	41
23	7	16	6	17	5	18	63	24	39	22	41	21	42
24	7	17	6	18	6	18	64	24	40	23	41	22	42
25	8	17	7	18	6	19	65	25	40	23	42	22	43
26	8	18	7	19	7	19	66	25	41	24	42	23	43
27	8	19	8	19	7	20	67	26	41	24	43	23	44
28	9	19	8	20	7	21	68	26	42	24	44	23	45
29	9	20	8	21	8	21	69	26	43	25	44	24	45
30	10	20	9	21	8	22	70	27	43	25	45	24	46
31	10	21	9	22	8	23	71	27	44	26	45	25	46
32	10	22	9	23	9	23	72	28	44	26	46	25	47
33	11	22	10	23	9	24	73	28	45	27	46	26	47
34	11	23	10	24	10	24	74	29	45	27	47	26	48
35	12	23	11	24	10	25	75	29	46	27	48	26	49
36	12	24	11	25	10	26	76	29	47	28	48	27	49
37	13	24	11	26	11	26	77	30	27	28	49	27	50
38	13	25	12	26	11	27	78	30	48	29	49	28	50
39	13	26	12	27	12	27	79	31	48	29	50	28	51
40	14	26	13	27	12	28	80	31	49	30	50	29	51
41	14	27	13	28	12	29	81	32	49	30	51	29	52
42	15	27	14	28	13	29	82	32	50	31	51	29	53
43	15	28	14	29	13	30	83	33	50	31	52	30	53
44	16	28	14	30	14	30	84	33	51	31	53	30	54
	5%		2%		1%			5%		2%		1%	
	Двусторонние границы							Двусторонние границы					

Таблица 6. Значения величины z для значений r от 0,00 до 0,99

r	0	2	4	6	8
0,0	0,0000	0,0200	0,0400	0,0601	0,0802
1	0,1003	0,1206	0,1409	0,1614	0,1820
2	0,2027	0,2237	0,2448	0,2661	0,2877
3	0,3095	0,3317	0,3541	0,3769	0,4001
4	0,4236	0,4477	0,4722	0,4973	0,5230
5	0,5493	0,5763	0,6042	0,6328	0,6625
6	0,6931	0,7250	0,7582	0,7928	0,8291
7	0,8673	0,9076	0,9505	0,9962	1,0454
8	1,0986	1,1568	1,2212	1,2933	1,3758
9	1,4722	1,5890	1,7380	1,9459	2,2976

Таблица 7. Значения функции $F(t)$

t	0	2	4	6	8
0,0	50000	50798	51595	52392	53188
0,1	53983	54776	55567	56356	57142
0,2	57926	58706	59483	60257	61026
0,3	61791	62552	63307	64058	64803
0,4	65542	66276	67003	67724	68439
0,5	69146	69847	70540	71226	71904
0,6	72575	73237	73891	74537	75175
0,7	75804	76424	77035	77637	78230
0,8	78814	79389	79955	80511	81057
0,9	81594	82121	82639	83147	83646
1,0	84134	84614	85083	85543	85993
1,1	86433	86864	87286	87698	88100
1,2	88493	88877	89251	89617	89973
1,3	90320	90658	90988	91308	91621
1,4	91924	92220	92507	92786	93056
1,5	93319	93574	93822	94062	94295
1,6	94520	94738	94960	95154	95352
1,7	95543	95728	95907	96080	96246
1,8	96407	96562	96712	96856	96995
1,9	97128	97257	97381	97500	97615
2,0	97725	97831	97932	98030	98124
2,1	98214	98300	98382	98461	98537
2,2	98610	98679	98745	98809	98870
2,3	98928	98983	99036	99086	99134
2,4	99180	99224	99266	99305	99343

t	0	2	4	6	8
2,5	99379	99413	99446	99477	99506
2,6	99534	99560	99585	99609	99632
2,7	99653	99674	99693	99711	99728
2,8	99744	99760	99774	99788	99801
2,9	99813	99825	99836	99846	99856
3,0	99865	99874	99882	99889	99896
3,1	99903	99910	99916	99921	99926
3,2	99931	99936	99940	99944	99948
3,3	99952	99955	99958	99961	99964
3,4	99966	99969	99971	99973	99975
3,5	99977	99978	99980	99981	99983
3,6	99984	99985	99986	99987	99988
3,7	99989	99990	99991	99992	99992
3,8	99993	99993	99994	99994	99995
3,9	99995	99996	99996	99996	99997
4	99997	99999	99999	—	—

Таблица 8

Объем выборки N	Критические значения коэффициента					
	эксцесса E при $1-\alpha$				асимметрия A при $1-\alpha$	
	0,99	0,95	0,05	0,01	0,95	0,99
50	4,92	4,01	2,13	1,95	0,533	0,787
100	40	3,77	35	2,18	389	567
150	14	66	45	30	321	464
200	3,98	57	51	37	280	403
250	87	51	55	42	251	360
300	79	47	59	46	230	329
350	72	44	62	50	213	305
400	67	41	64	52	200	285
450	63	39	66	55	188	269
500	60	37	67	57	179	255
550	57	35	69	58	171	243
600	54	34	70	60	163	233
650	52	33	71	61	157	224
700	50	31	72	62	151	215
750	48	30	73	64	146	208
800	46	29	74	65	142	202
850	45	28	74	66	138	196
900	43	28	75	66	134	190

Объем выборки N	Критические значения коэффициента					
	эксцесса E при $1-\alpha$				асимметрия A при $1-\alpha$	
	0,99	0,95	0,05	0,01	0,95	0,99
950	42	27	76	67	130	185
1000	41	26	76	68	127	180
1200	37	24	78	71	116	165
1400	34	22	80	72	107	152
1600	32	21	81	74	100	142
1800	30	20	82	76	095	134
2000	28	18	83	77	090	127
2500	25	16	85	79	080	114
3000	22	15	86	81	973	104
3500	21	14	87	82	068	096
4000	19	13	88	83	064	090
4500	18	12	88	84	060	085
5000	17	12	89	85	057	081

Таблица 9. Случайные числа

(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)
5489	5583	3156	0835	1988	3912	0938	7460
3522	0935	7877	5665	7020	9255	7379	7124
7555	7579	2550	2487	9477	0864	2349	1012
5759	3554	5080	9074	7001	6249	3224	6368
6303	6895	3371	3196	7231	2918	7380	0438
7351	5634	5323	2623	7803	8374	2191	0464
7068	7803	8832	5119	6350	0120	5026	3684
3613	1428	1796	8447	0503	5654	3254	7336
5143	4534	2105	0368	7890	2473	4240	8652
9815	5144	7649	8638	6137	8070	5345	4865
5780	1277	6316	1013	2867	9938	3930	3203
1187	0951	5991	5245	5700	5564	7352	0891
4184	2179	4554	9083	2254	2436	2965	5154
2916	2972	9885	0275	0144	8034	8122	3213
5524	1341	9860	6565	6981	9842	0171	2284
0146	6291	2354	5694	0377	5336	6460	9585
4920	2826	5238	5402	7937	1993	4332	2327
7978	1947	6380	3425	7267	7285	1130	7722
7453	0653	3645	7497	5969	8682	4191	2976
1473	6938	4899	5348	1641	3652	0852	5296
8162	8797	8000	4700	1880	9660	8446	1883
5642	4219	0807	3301	4279	4168	4305	9937
2042	1192	1175	8851	6432	4635	5757	6656
5470	7702	6958	9080	5925	8519	0127	9233

(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)
4045	1730	6005	1704	0345	3275	4738	4862
5880	1257	6163	4439	7276	6353	6912	0731
9083	4260	5277	4998	4298	5204	3965	4028
1762	8713	1189	1090	8989	7273	3213	1935
2023	2589	1740	0424	8924	0005	1969	1636
7965	3855	4765	0703	1678	0841	7543	0308
7690	0480	8098	9629	4819	7219	7241	5128
9292	0426	9573	4903	5916	6576	8368	3270
0867	1656	7016	4220	2533	6345	8227	1904
0505	2127	8255	5276	2233	3956	4118	8199
6295	9795	1112	5761	2575	6837	3336	9322
6323	2615	3410	3365	1117	2417	3176	2434
8672	8536	2966	5773	6412	8114	0030	4697
1422	5507	7596	0670	3013	1351	3886	3268
2653	1472	5113	5735	1469	9545	9331	5303
0438	4376	3328	8649	8327	0110	4549	7955
2851	2157	0047	7085	1129	0460	6821	8373
7962	2753	3077	8718	7418	8004	3425	3706
3837	4098	0220	1217	4732	0150	1637	1097
8542	4126	9274	2251	0607	4301	8730	7690
0139	0765	8039	9484	2577	7859	1976	0623
6687	1943	4307	0579	8171	8224	8641	7034
6242	5582	5872	3197	4919	2792	5991	4058
6859	9606	0522	4993	0345	8958	1289	8825
6590	1932	6043	3623	1973	4112	1795	8465
3482	0478	0221	6738	7313	5643	4767	0106

Таблица 10. Критические значения корреляционного отношения η^2 и коэффициента детерминации R^2 ($\alpha=0{,}01$)

k_2	k_1									
	1	2	3	4	5	6	7	10	20	40
3	919	954	967	975	979	982	985	989	994	997
4	841	900	926	941	951	958	963	973	986	993
5	765	842	879	901	916	928	936	953	974	987
6	969	785	830	859	979	894	906	929	961	979
7	636	732	784	818	842	860	875	904	946	971
8	585	684	740	778	806	827	844	879	931	962
9	540	641	700	741	771	795	814	854	914	953
10	501	602	663	706	738	764	784	829	898	943
11	467	567	629	673	707	734	757	805	882	933
12	437	536	598	643	678	707	730	782	865	923

k_2	k_1									
	1	2	3	4	5	6	7	10	20	40
13	411	508	570	616	652	681	705	759	849	913
14	388	482	544	590	626	656	681	738	834	903
15	367	459	520	566	603	633	659	717	818	893
16	342	438	498	544	581	612	638	698	803	883
17	331	418	478	523	560	591	618	679	788	873
18	315	401	459	504	541	572	599	669	774	863
19	201	284	442	486	523	554	581	644	760	853
20	288	369	426	470	506	537	564	627	746	843
21	276	355	410	454	490	521	548	612	733	834
22	265	342	396	440	475	506	533	597	720	824
23	255	330	383	426	461	492	519	583	707	815
24	246	319	371	413	448	478	505	569	695	806
25	237	308	359	401	435	465	492	556	683	797
26	229	298	349	389	423	453	479	543	672	788
27	221	289	338	378	412	442	468	531	661	779
28	214	280	329	368	401	430	456	520	650	771
29	208	272	319	358	391	420	446	509	640	762
30	201	264	311	349	381	410	435	498	630	754
32	190	250	295	332	364	391	416	479	611	738
34	179	237	280	316	347	374	398	459	592	722
36	170	226	267	302	332	358	382	443	574	707
38	162	215	255	289	318	344	367	426	558	693
40	155	206	244	277	305	310	353	412	542	679
50	125	168	201	229	254	276	297	351	475	615
60	106	142	171	196	218	238	256	305	423	563
80	080	109	132	151	169	186	201	242	345	479
100	065	088	107	123	138	152	165	201	292	417
120	054	074	090	104	117	129	140	171	253	370
200	033	045	055	064	072	080	087	106	165	253
400	016	023	028	033	037	011	045	056	086	141

ЛИТЕРАТУРА

Ленин В. И. Новые хозяйственные движения в крестьянской жизни. — Полн. собр. соч., т. 1.

Ленин В. И. Кустарная перепись 1894/95 года в Пермской губернии и общие вопросы «кустарной» промышленности. — Полн. собр. соч., т. 2.

Ленин В. И. Развитие капитализма в России. — Полн. собр. соч., т. 3.

Ленин В. И. К вопросу о нашей фабрично-заводской статистике. — Полн. собр. соч., т. 4.

Ленин В. И. О статистике стачек в России. — Полн. собр. соч., т. 19.

Ленин В. И. Заработки рабочих и прибыль капиталистов в России. — Полн. собр. соч., т. 22.

Ленин В. И. Рабочий день и рабочий год в Московской губернии. — Полн. собр. соч., т. 22.

Ленин В. И. Роль сословий и классов в освободительном движении. — Полн. собр. соч., т. 23.

Ленин В. И. Язык цифр. — Полн. собр. соч., т. 23.

Ленин В. И. К вопросу о задачах земской статистики. — Полн. собр. соч., т. 24.

Ленин В. И. Статистика и социология. — Полн. собр. соч., т. 30.

Ленин В. И. Удержат ли большевики государственную власть? — Полн. собр. соч., т. 34.

Айвазян С. А., Бежаева З. И., Староверов О. В. Классификация многомерных наблюдений. М., 1974.

Бешелев С. Д., Гурвич Ф. Г. Экспертные оценки. М., 1973.

Бирюков В. В., Геллер Е. С. Кибернетика в гуманитарных науках. М., 1973.

Бокарев Ю. П. Бюджетные обследования крестьянских хозяйств 20-х годов как исторический источник. М., 1981.

Бородкин Л. И., Селунская Н. Б. Методы изучения социальной истории в американской историографии. (По поводу «Проекта социальной истории Филадельфии», США). — История СССР, 1978, № 2.

Бородкин Л. И., Соколов А. К. Историк и изучение социальных процессов. (Об использовании массовых источников и количественных методов их анализа в новейшей зарубежной историографии). — История СССР, 1983, № 1.

Венецкий И. Г. Вариационные ряды и их характеристики. М., 1970.

Венецкий И. Г., Кильдишев Г. С. Основы теории вероятностей и математической статистики. М., 1968.

Гельман К. Б. Машиночитаемые документы в СССР. М., 1980, вып. 1

Герчук Я. П. Графики в математико-статистическом анализе. М., 1972.

Головач А. В., Ерина А. М., Трофимов В. И. Критерии математической статистики в экономических исследованиях. М., 1973.

Гренджер К., Хатанака М. Спектральный анализ временных рядов в экономике. М., 1972.

Громыко Г. Л. Статистика, М., 1976.

Джини К. Средние величины. М., 1970.

Дрейпер Н., Смит Г. Прикладной регрессионный анализ. М., 1973.
Дробижев В. З., Соколов А. К., Устинов В. А. Рабочий класс советской России в первый год диктатуры пролетариата. (Опыт структурного анализа). М., 1975.
Дружинин Н. К. Выборочный метод и его применение в социально-экономических исследованиях. М., 1970.
Дружинин Н. К. Логика оценки статистических гипотез. М., 1973.
Дружинин Н. К. Математическая статистика в экономике. Введение в математико-статистическую методологию. М., 1971.
Езекиэл М., Фокс К. А. Методы анализа корреляций и регрессий. М., 1966.
Елисеева И. И. Статистические методы измерения связей. Л., 1982.
Елисеева И. И., Рукавишников В. О. Группировка, корреляция, распознавание образов. М., 1977.
Жуковская В. М., Мучник И. Б. Факторный анализ в социально-экономических исследованиях. М., 1976.
Йейтс Ф. Выборочный метод в переписях и обследованиях. М., 1965.
Кендэл М. Временные ряды. М., 1981.
Кендэл М. Ранговые корреляции. М., 1975.
Кильдишев Г. С., Френкель А. А. Анализ временных рядов и прогнозирование. М., 1973.
Кахк Ю., Ковальченко И. Д. Методологические проблемы применения количественных методов в исторических исследованиях. — История СССР, 1974, № 5.
Ковальченко И. Д., Милов Л. В. Всероссийский аграрный рынок XVIII — начала XX в. (Опыт количественного анализа). М., 1974.
Ковальченко И. Д., Селунская Н. Б., Литваков Б. М. Социально-экономический строй помещичьего хозяйства европейской России эпохи капитализма. (Источники и методы изучения). М., 1982.
Ковальченко И. Д. О моделировании исторических явлений и процессов. — Вопросы истории, 1978, № 8.
Ковальченко И. Д., Бородкин Л. И. Аграрная типология губерний европейской России на рубеже XIX—XX вв. (Опыт многомерного количественного анализа). — История СССР, 1979, № 1.
Ковальченко И. Д., Бородкин Л. И. Структура и уровень аграрного развития районов европейской России на рубеже XIX—XX вв. (Опыт многомерного анализа). — История СССР, 1981, № 1.
Козлов В. А., Обожда В. А., Пушков В. П. Опыт изучения особенностей культурного развития советского доколхозного крестьянства. (По данным бюджетных обследований крестьянских хозяйств 20-х годов). — История СССР, 1978, № 5.
Количественные методы в гуманитарных науках. М., 1981.
В. И. Ленин и современная статистика. М., 1970, т. 1; М., 1971, т. 2; М., 1973, т. 3.
Массовые источники по социально-экономической истории России периода капитализма. М., 1979.
Массовые источники по социально-экономической истории советского общества. М., 1979.
Математические методы в исторических исследованиях. М., 1972.
Математические методы в исследованиях по социально-экономической истории. М., 1975.
Математические методы в историко-экономических и историко-культурных исследованиях. М., 1977.
Математические методы в социально-экономических и археологических исследованиях. М., 1981.

Мелихов С. В. Количественные методы в американской политологии. М., 1979.

Миллс Ф. Статистические методы. М., 1958.

Милов Л. В., Булгаков М. Б., Гарскова И. М. Об экономической структуре поместья и вотчины начала XVII в. (Опыт количественного анализа). Россия на путях централизации. М., 1982.

Миркин Б. Г. Анализ качественных признаков и структур. М., 1980.

Миронов Б. Н., Степанов З. В. Историк и математика. М., 1975.

Многомерный статистический анализ в социально-экономических исследованиях. М., 1974.

Моисенко Т. Л. Методы изучения крестьянской аренды в России по данным земской статистики конца XIX в. — История СССР, 1979, № 4.

Мошков Ю. А. Опыт применения корреляционного анализа в изучении отчетов о распределении колхозных доходов в начале массовой коллективизации (по материалам Северокавказского края). Источниковедение отечественной истории. М., 1973.

Новикова Н. Н., Клосс Б. М. Чернышевский — автор прокламации-газеты «Великорусс». Чернышевский и его эпоха. Революционная ситуация в России в 1854—1861 гг. М., 1979.

Осипов Г. В., Андреев Э. И. Методы измерения в социологии. М., 1977.

Палли Э. Х. ЭВМ и новые возможности хранения исторической информации. Источниковедение отечественной истории. М., 1977.

Пфанцагль И. Теория измерения. М., 1976.

Рабочая книга социолога. М., 1977.

Растригин Л. А. Вычислительные машины, системы, сети. М., 1982.

Рунион Р. Справочник по непараметрической статистике. М., 1982.

Рябушкин М. В. Ленинское наследие и статистика. М., 1978.

Славко Т. И. Математико-статистические методы в исторических исследованиях. М., 1981.

Социологические исследования: построение и сравнение показателей. М., 1981.

Статистические методы анализа информации в социологических исследованиях. М., 1979.

Суслов И. П. Теория статистических показателей. М., 1975.

Суслов И. П. Основы теории достоверности статистических показателей. Новосибирск, 1979.

Типология и классификация в социологических исследованиях. М., 1982.

Урланис У. Б. Общая теория статистики. М., 1962.

Устинов В. А., Фелингер А. Ф. Историко-социальные исследования, ЭВМ и математика. М., 1973.

Хвостова К. В. Количественный подход в средневековой социально-экономической истории. М., 1980.

Хвостова К. В. Количественные методы в историческом познании. — Вопросы истории, 1983, № 4.

Юл Дж., Кендэл М. Теория статистики. М., 1960.

ОГЛАВЛЕНИЕ