Б. А. ГОДУНОВ

СТАТИСТИКА

Часть 1

(Конспект лекций)

МИНИСТЕРСТВО ОБРАЗОВАНИЯ РЕСПУБЛИКИ БЕЛАРУСЬ

УЧРЕЖДЕНИЕ ОБРАЗОВАНИЯ
«БРЕСТСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ»

Кафедра высшей математики

Б. А. ГОДУНОВ

СТАТИСТИКА

Ч а с т ь 1

(Конспект лекций)

Брест 2008

УДК 31
ББК 60. 6я73
Г 59

В данном пособии использованы материалы лекций по курсу «Статистика», читавшихся автором, математиком по основному образованию, на протяжении десяти лет для студентов экономического факультета Брестского государственного технического университета. Поскольку за эти годы менялись количество и соотношение лекционных и семинарских часов, автор не разбивает материал на отдельные лекции и излагает курс в достаточно краткой форме, изредка позволяя себе более подробные комментарии и отступления от конспективного стиля. Кроме этого включены некоторые вопросы, предлагаемые для самостоятельного изучения студентами.

При чтении лекций в основном использовались учебники «Общая теория статистики» под ред. члена-корреспондента РАН И. И. Елисеевой, «Общая теория статистики» под ред. проф. М. Р. Ефимовой и некоторые другие, отмеченные в списке литературы.

Данное учебное пособие поможет студентам самостоятельно освоить некоторые важные описательного плана вопросы и глубже разобраться в тех вопросах, где применяется математический аппарат статистики и статистический анализ данных и результатов их обработки.

Издается в 2-х частях. Часть 1.

Годунов Б. А.
Г 59. **Статистика, часть 1 (конспект лекций)**
Брест, Издательство БрГТУ, 2008,с.85

Рецензенты:
В. Ф. Савчук, доцент, зав. кафедрой алгебры и геометрии УО «Брестский государственный университет им. А. С. Пушкина», к.ф.-м.н.
Кафедра высшей математики УО БрГТУ

Г Л А В А 1

ОСНОВНЫЕ ПОНЯТИЯ СТАТИСТИКИ

1.1. Что такое статистика

Современная статистика – это отрасль знаний, находящая применение практически во всех сферах научной и практической деятельности человека. Мы будем говорить в основном об экономике и производстве.

Слово «статистика» производят от латинского слова *status* – состояние, положение вещей. Как научный термин оно стало применяться в XVIII в. и первоначально употреблялось в значении «государствоведение». Здесь уместно вспомнить итальянские слова *stato* – государство и *statista* – знаток государства.

Корни статистики уходят в глубь веков. Это было связано с зарождением и развитием государств, с потребностями государственного управления. Именно тогда появилась необходимость в сборе сведений о наличии земель, численности населения и его имущественном положении. Известно, что такой учет проводился в Древнем Китае, Греции, Древнем Риме и в Египте. К примеру, греческий философ Аристотель (384 – 322 г. г. до н.э.) составил описание 157 городов и государств своего времени. В средние века предстал свету уникальный памятник – «Книга страшного суда» (1061 г.) – свод материалов всеобщей переписи населения Англии и его имущества, содержащий данные о 240 тыс. дворов.

Со временем благодаря применению математических методов из *описательной* статистики выделяется направление, называемое *выводной* статистикой. Ее задача заключается в том, чтобы в массе событий, связанных со *случайным* проявлением в каждом конкретном случае, найти закономерности и на их основе делать прогнозы на будущее.

В наше время *статистика* может быть определена как *сбор, представление, анализ, интерпретация числовых данных и на основе этого — прогнозирование.*

1.2. Статистические совокупности, статистическая закономерность

Предмет статистики составляют *статистические совокупности* – множества явлений, в каждом из которых *необходимость (некоторый признак),* присущая явлениям данного вида, в каждом отдельном проявлении связана со *случайностью.*

Целью статистики является установление *статистической закономерности* – количественной зависимости, полученной на основе обработки данных по отдельным явлениям всей совокупности. При этом изучаемая закономерность определяет *единицы совокупности* – частный

случай этой закономерности. Действительно, рассмотрим цепочку: *отрасль – завод – цех – рабочий.* Если в качестве статистической совокупности рассматривать отрасль, то единицей совокупности может быть и завод (изучается выполнение плана выпуска продукции), и цех (оснащенность цехов одного профиля современным оборудованием) и рабочий (средняя производительность одного рабочего). В то же время завод можно рассматривать как статистическую совокупность, единицами которой могут быть и цеха и рабочие. Таким образом, одно и то же явление в зависимости от изучаемой статистической закономерности может являться или статистической совокупностью или единицей совокупности. Отсюда еще одна сторона этого понятия: *единицы совокупности* – это предел дробления объекта исследования, при котором сохраняются все свойства изучаемого процесса.

Итак, предметом статистики являются *совокупности – множества однокачественных, варьирующих явлений.* Отметим три стороны этого определения. Первое – *множество* явлений; второе – явления объединены *общим качеством,* отражающим *одну и ту же закономерность*; третье – это множество *варьирующих* явлений, то есть меняющихся по своим характеристикам от явления к явлению.

Цитата из [1]. *Вариация – основа существования мира и источник его развития. Если бы люди не делились на мужчин и женщин, человечество прекратило бы существование; если бы не было различий мнений – истина была бы недостижимой, а жизнь без вариаций – невыносимо скучной!*

1.3. Признаки и их классификация

Признаком называют свойство или качество, присущее всем единицам совокупности. Для одной и той же единицы могут рассматриваться несколько признаков. Например, признаки человека: рост, вес, возраст, форма носа, цвет волос, семейное или социальное положение и т.д. Признаки предприятия: форма собственности, специализация, численность работников, экономическая эффективность деятельности и прочие. *Статистика изучает явления через признаки*.

Признаки отличаются некоторыми особенностями, влияющими на методы и приемы анализа их. Это дает основания для классификации признаков. Таких оснований различают пять: по характеру выражения признаков, по способу измерения, по отношению к изучаемому объекту, по характеру вариации и по отношению ко времени (см. таблицу *1.1*).

Описательные признаки выражаются словесно: пол и национальность человека, разновидности строительных материалов (дерево, цемент, кирпич и т.д.). При этом они подразделяются на *номинальные признаки*, которые нельзя ранжировать (пол, национальность) и *порядковые признаки*, по которым можно провести ранжирование. Например,

пользуясь оценками экспертов, ранжируют фигуристов или прыгунов в воду.

Таблица 1.1

Классификация признаков

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

Количественные признаки выражаются числами. Они преобладают в статистике. Например, надой молока, производительность труда, темпы роста и т.д.

Первичные признаки характеризуют единицу совокупности в целом. Это признаки, которые непосредственно связаны с единицами совокупности и существуют независимо от того, занимается ими статистика или нет. Они могут быть измерены, сосчитаны, взвешены. К ним относятся, например, урожай, произведенная на предприятии продукция, число жителей города.

Вторичные или *расчетные* признаки являются продуктом познания изучаемой статистической совокупности и представляют собой различные соотношения первичных признаков: если разделить собранный урожай на количество гектаров, то получится урожайность, а при делении общих затрат на количество произведенной продукции – себестоимость.

Прямые (непосредственные) признаки относятся ко всей совокупности – продукция, произведенная бригадой рабочих.

Косвенные признаки относятся не ко всей совокупности, а к ее части. Продукция, произведенная каждым рабочим, определяет всю продукцию бригады и является для бригады косвенным признаком. А для рабочего произведенная им продукция – прямой признак.

Альтернативные признаки предполагают два ответа – да, нет. К ним относятся признаки наличия или нет чего-либо. Например, годность изделия для реализации, пол человека, садоводство в сельском предприятии (есть оно или нет его).

Дискретные признаки могут принимать значения, отделенные друг от друга. Как правило, их либо конечное число (они могут повторяться у разных единиц совокупности – число членов семьи, сессионные оценки

студентов), либо это целые числа или числа, отстоящие друг от друга не меньше, чем на некоторую фиксированную величину.

Непрерывные признаки принимают любые значения из некоторого интервала. К ним, в частности, относятся вторичные признаки, так как результат деления может любым числом.

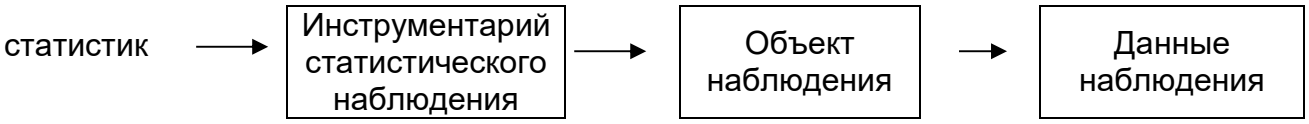
Моментные признаки характеризуют объект исследования на некоторый момент времени – численность населения, состав семьи, техническое оснащение предприятия, парк автобусов или такси.

Интервальные признаки отражают результат процесса за некоторый промежуток времени – число родившихся за год; продукция, произведенная за квартал; число пассажиров железной дороги, перевезенных за год.

Замечание. На практике часто значения непрерывного признака округляют с некоторой степенью точности, и они становятся отделенными друг от друга, то есть *квазидискретными.* И наоборот, есть дискретные, по сути, признаки, принимающие огромное число значений – число работников предприятия, число произведенной продукции в штуках, поголовье скота. В таком случае статистика рассматривает их как *квазинепрерывные.*

1.4. Статистическое наблюдение

Статистическое наблюдение – это научно организованный сбор данных, *определяющий* всю дальнейшую работу статистики. *Достоверность* и *сопоставимость* – основные требования, предъявляемые к сбору данных на всех этапах статистического наблюдения, схему которого для наглядности изобразим ниже:



Достоверность данных зависит от качеств статистика: добросовестность, профессиональная подготовка, коммуникабельность, организаторские навыки.

На втором этапе важную роль играет заранее разработанная программа наблюдений, разработка бланков, анкет и инструкций по их заполнению.

На третьем этапе существенно определить границы объекта, который следует подготовить к наблюдению (предварительное извещение о предстоящем наблюдении, ознакомление с заполняемой документацией, разговор о том, как заполнять бланки и анкеты и т.д.).

На четвертом этапе важно, чтобы собираемые данные соответствовали истинному положению вещей.

Сопоставимость обеспечивается сбором данных с использованием единой методики в одно и то же время или на одну и ту же дату. При этом должны быть сопоставимы единицы наблюдаемой совокупности (нельзя, например, включать в совокупность завод и цех другого завода).

Виды статистических наблюдений

Статистические наблюдения различаются по времени наблюдения и по степени охвата единиц наблюдения.

1) **По времени наблюдения:**
   а) *текущие* или *непрерывные* наблюдения – ведутся систематически по мере возникновения явлений (регистрация рождений, смертей, гражданских браков, пропусков занятий, учет выпуска продукции);
   б) *периодические* наблюдения проводятся через определенные, обычно одинаковые, промежутки времени (учет успеваемости студентов по результатам сессий, техническое оснащение предприятий на начало года, перепись населения);
   в) *разовые* наблюдения проводятся по мере надобности (изучение спроса на товар, школьная перепись, изучение жилого фонда).
2) **По степени охвата:**
   а) *сплошные* наблюдения: регистрации подлежат все единицы совокупности (перепись населения, сбор данных по предприятиям для изучения положения в отрасли);
   б) способ *основного массива* – наблюдению подвергается та часть единиц, которая вносит наибольший вклад в изучаемое явление (при изучении производства куриных яиц учитывают данные птицеферм и не рассматривают данные по подворьям). При этом способе вводится *ценз* – значение признака, ограничивающее единицу наблюдения. Например, рассматривают предприятия с числом работников 100 и более;
   в) *выборочное* наблюдение проводится по части единиц совокупности, отобранной в определенном порядке, а результаты распространяются на всю совокупность (изучение рейтинга кандидатов в депутаты до выборов);
   г) *монографическое* наблюдение заключается в подробном описании *отдельных* единиц совокупности с целью углубленного изучения их. Главную роль играют описательные признаки: качество, поведение, ориентация, перспективы развития и т.д. Например, изучение образа жизни семьи или нескольких семей. Монографическое наблюдение для большой совокупности единиц многозатратно и поэтому не проводится.

Формы наблюдений

Для изучения статистической совокупности используется одна из трех форм наблюдения: непосредственное наблюдение, документальное наблюдение и опрос.

1) *Непосредственное* наблюдение – осмотр, подсчет, взвешивание, снятие показаний приборов. Сюда относятся: учет в магазинах, регистрация цен, инвентаризация, проверка веса расфасованного товара, контроль над качеством продукции, регистрация показаний в метеорологии.
2) Д*окументальный способ (отчет)* – используются отчетные документы различных организаций (результаты описей и инвентаризаций, технические паспорта, данные бухгалтерского учета, переоценка основных фондов предприятий и т.д.).
3) *Опрос* – информацию получают со слов опрашиваемого человека. Возможны три вида опроса. *Экспедиционный* опрос – специально подготовленный регистратор заполняет бланки, одновременно контролирует правильность ответов. Способ надежный, но дорогостоящий. Перепись населения. *Корреспондентский* опрос – рассылка анкет или бланков на предприятия или специально подготовленным людям на них (изучение спроса, социологические обследования). Преимущество – дешевизна, недостаток – не всегда обеспечивает хорошее качество сведений. *Саморегистрация* – работники организации, проводящей опрос, раздают опросные листы или анкеты опрашиваемым людям и инструктируют их. После заполнения ими статистики собирают ответы, контролируя полноту и правильность ответов.

Подготовка статистического наблюдения заключается в составлении организационного плана, включающего следующие пункты:

- цель наблюдения с описанием объекта наблюдения и единиц совокупности;
- органы наблюдения – статистики, сотрудники предприятия, документы данных и документы регистрации и т.д.;
- время наблюдения – период, сезон, критическая дата – время, по состоянию на которое собираются сведения, (критический момент – используется при переписи населения);
- сроки наблюдения;
- территория наблюдения – все места, где расположены единицы совокупности;
- полная программа наблюдений.

Осуществление наблюдения – дорогостоящее предприятие. Поэтому программа должна составляться с учетом выделенных средств и в случае необходимости следует разумно уменьшить число единиц совокупности, не включать никаких лишних вопросов (типа «на всякий случай»),

не включать подозрительные для опрашиваемого вопросы. Программа должна предусматривать контрольные вопросы.

Следует предусмотреть *формы ответов* на вопросы программы – цифровые, альтернативные, многовариантные – меню ответов, изменяющихся от резко отрицательных до положительных. Задача опрашиваемого в этом случае заключается в выборе одного из предложенных ответов.

И, наконец, об **ошибках** наблюдений. Как утверждают авторитеты статистики, *ошибки будут всегда.* Важно организовать над ними контроль, включающий полноту охвата единиц наблюдений, полноту заполнения документов регистрации сведений, проверку ошибок регистрации. Различают два вида ошибок. *Случайные (логические)* ошибки – в совокупности они гасят друг друга и существенно не влияют на результат. *Систематические (преднамеренные)* ошибки могут существенно исказить общую картину. Известно, что многие люди приуменьшают свои доходы, возраст, заявляют о бóльшей осведомленности в различных вопросах, чем это есть на самом деле.

Естественно возникает вопрос о контроле данных, который проводится в двух формах: счетный и логический.

Счетный контроль использует жесткую связь между признаками, которую можно проверить с помощью четырех арифметических операций. Эта связь отражается в заголовках граф отчетности и проверяется, например, по правилу: графа С равняется графе А, деленной на графу В. Счетный контроль используют при проверке итоговых сумм.

Логический контроль учитывает логические связи между признаками. Если, к примеру, при переписи населения дети оказались старше своих родителей, значит, при регистрации данных была допущена ошибка. Можно использовать также максимально и минимально возможные значения признака. Вряд ли стоить верить тому, что в некотором хозяйстве урожайность пшеницы оказалась равной ста пятидесяти центнерам с гектара.

Для проверки полученного материала наблюдения предварительно разрабатывается *схема контроля*. Нельзя произвольно вносить исправления в формуляр. Необходимо либо провести повторное наблюдение, либо согласовать данные с лицами, отвечающими за информацию.

Проверкой данных и внесением в них необходимых изменений завершается начальный этап статистического исследования.

И, справедливости ради, в итоге заметим, что ошибки могут возникать по вине обоих участников наблюдения – и статистика, и наблюдаемого!

Вероятно, эта неизбежность ошибок в статистике побудила английского писателя Дизраэля Б. (1804 – 1901) заявить: „ Имеются три рода лжи – ложь, наглая ложь и статистика“.

1.5. Статистические таблицы

Важную роль на начальном этапе исследований играет вопрос о представлении статистических данных. Выделим три основных способа: включение данных в текст, представление в таблице и графическая интерпретация.

Включение в текст удобно применять в случае малого количества данных. Например, если рассматривается население шести областей республики Беларусь.

Если количество данных достаточно велико, то перечисление в тексте приводит к плохому восприятию их, не говоря уже о невозможности провести предварительный анализ этих данных. В таких случаях наиболее удобной формой представления статистических данных являются *таблицы*.

Статистическая таблица – это одна из форм *наглядного* представления данных. В некоторых случаях таблица предваряется *общим заголовком*, в котором отмечается содержание ее, место и время, к которым относятся данные, и единицы измерения, если они одинаковы для всей совокупности приводимых сведений (см., например, таблицу 2.7, глава 2). Заголовок можно не приводить, если из текста и таблицы ясно, о чем идет речь. Каждая таблица имеет подлежащее и сказуемое. *Подлежащим таблицы* являются единицы или группы единиц статистической совокупности. *Сказуемое таблицы* отражает сведения о подлежащем, чаще всего в количественной форме. Например, в таблице 2.7 подлежащее – число вызовов скорой помощи за один час и соответственно данные первого столбца, а сказуемое – сколько раз наблюдались перечисленные количества, то есть данные второго столбца.

В зависимости от формы подлежащего статистические таблицы подразделяются на *простые*, *групповые* и *комбинационные*.

В подлежащем *простой таблицы* перечисляются все наблюдавшиеся значения, как правило, в порядке возрастания или убывания. Сказуемое при этом должно содержать данные по каждой единице совокупности. По виду подлежащего простая таблица может быть к тому же названа *перечневой* (перечень единиц наблюдения: в таблице 2.6 – это заводы), *хронологической* (подлежащее дано в виде дат) и *территориальными* (в подлежащем перечисляются географические объекты – страны, города, области и т.д.).

В *групповой таблице* подлежащее подразделяется на группы по одному признаку, а в сказуемом указывается число единиц в группах в абсолютных единицах (количество) или в относительных (части или проценты от общего количества). Примерами групповых таблиц являются, например, таблицы 2.7 и 2.9.

В подлежащем *комбинационной таблицы* совокупность подразделяется на группы по двум и более признакам. В качестве примера приве-

дем таблицу, заимствованную из [1], где изучаемая совокупность – постоянное население города Санкт-Петербурга – подразделятся по двум признакам: возрасту и полу.

Таблица 1.2

Численность постоянного населения Санкт-Петербурга по возрастным группам на начало 1992 г., тыс. человек

Группы населения	Количество	В том числе	
		мужчин	женщин
Моложе трудоспособного	*973,1*	*496,8*	*476,3*
Трудоспособные	*2942,2*	*1473,3*	*1468,9*
Старше трудоспособного	*1055,7*	*280,5*	*775,2*
Всего	*4971,0*	*2250,6*	*2720,4*

Приведем основные правила оформления таблиц.

В таблице не должно быть *ни одной лишней линии*. *Необходимо* отделять линией заголовок таблицы от заголовков ее граф, заголовки граф – от числовых данных и в некоторых случаях – итоговую строку. Вертикальные линии могут быть, а могут частично отсутствовать.

Заголовки граф названия показателей и единиц измерения пишутся *без сокращений*, кроме общепринятых (км, чел., млн. руб. и т.д.). Если все данные имеют одну и ту же единицу измерения, то ее можно указать только в заголовке таблицы.

Итоговая строка – нижняя строка. Но иногда она ставится первой, когда хотят подчеркнуть приоритетную значимость итогов. В этом случае во второй строке делается запись «в том числе» и последующие строки содержат составляющие итоговой строки – или все или основные (наиболее значимые для данного исследования).

Цифровые данные в пределах каждой графы записываются с одинаковой степенью точности, при этом одинаковые разряды чисел располагаются на одной вертикали. Если при округлении в пределах принятой точности последняя оставляемая цифра окажется нулем, то его обязательно пишут. Например, при округлении до двух знаков после запятой число 23,197 записывают в виде 23,20, а не 23,2. Если же число равно 0,0025…, то записывают 0,00 и это понимают так, что данное значение признака отлично от нуля, но с более высокой точностью. В таблице не должно быть ни одной пустой клетки. Если показатель равен нулю или не определен, то ставится знак «—» (прочерк); если данные неизвестны, пишут «сведений нет» или знак «…».

Г Л А В А 2

СТАТИСТИЧЕСКИЕ ПОКАЗАТЕЛИ

2.1 Статистические показатели и их классификация

Для характеристики признака каждой единицы совокупности применяют *показатели*, вернее сказать *статистические показатели.* И это не просто абстрактные числа, характеризующие только величину. *Статистический показатель* – это обобщающая характеристика какого-либо свойства изучаемой статистической совокупности. Рассмотрим, например, такой показатель: в России в 1993 году введено в эксплуатацию *41 880, 2 тыс. м*^{2} жилой площади. Отметим на этом примере **атрибуты**, обязательные для всякого показателя (см. таблицу 2.1).

Таблица 2.1

Качественная сторона, его свойства, категория	Количественная сторона: число и единицы измерения	Территориальные, отраслевые и другие границы объекта	Время: интервал или момент времени
Ввод жилой площади	*41 880,2 тыс. м*^{2}	Россия	*1993* год (с *1.01.93* по *31.12.93*)

Как видим, статистический показатель содержит информацию о территориальных (отраслевых, ведомственных...) границах и о привязке ко времени (момент или интервал времени). При отсутствии хотя бы одного из четырех атрибутов делают показатель бессмысленным. Действительно, в чем смысл «укороченных» показателей: «в 1993 году введено в эксплуатацию *41 880,2 тыс. м*^{2} жилой площади» (а где?) или «в России введено в эксплуатацию *41 880,2 тыс. м*^{2} жилой площади» (а когда? за какое время?).

Признаки существуют, даже если их не изучает статистика, а вот статистические показатели – это детище статистики. Отметим, что некоторые показатели могут быть неименованными, например, когда рассматривают отношение части к целому, темпы роста или прироста, различного рода индексы.

На одном признаке можно построить несколько показателей. Пусть признак – возраст человека, а вот возможные показатели этого признака:

- средний возраст,
- возраст по интервалам (от *15* до *20* лет),
- количество людей старше *50* лет,
- максимальный или минимальный возраст.

Классификация статистических показателей

Приводимая ниже таблица отражает основные моменты классификации признаков.

Таблица 2.2

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

Рассмотрим подробнее.

1. Показатели свойств конкретных объектов:
   а) *экономические* – себестоимость, объем продукции, …
   б) *демографические* – сведения по ЗАГСу, по народонаселению, …
   в) *макроэкономические* – характеризуют экономику страны в целом.
2. Показатели статистических свойств массовых явлений и процессов – эти показатели получают в результате статистической обработки данных:
   а) средние величины,
   б) показатели вариации,
   в) структурные показатели …
3. Абсолютные показатели – отражают суммарные свойства объекта (общая выручка, весь урожай, фонд зарплаты и т.д.).
4. Относительные показатели – получаются в результате сравнения абсолютных показателей (средний урожай, средняя производительность труда …) или относительных показателей. В первом случае их называют относительными показателями *первого порядка,* во втором – показателями *высших порядков*.

При построении относительных показателей следует соблюдать некоторые правила:

- в относительном показателе абсолютные показатели должны быть логически связаны;
- исходные показатели при сравнении их должны отличаться только одним атрибутом;
- следует учитывать возможные пределы относительных показателей;
- нельзя сравнивать показатели с разными знаками;

- знаменатель не должен быть близок к нулю.

Системы показателей

Показатели в статистике практически не рассматриваются в одиночку. Обычно учитываются несколько логически связанных между собой показателей. Например, для любого предприятия важными являются абсолютные показатели:

- основные производственные фонды- *300,1 млрд. руб*.;
- среднегодовая численность рабочих *–15,6 млн. чел*.;
- объем произведенной продукции – *559,3 млрд. руб*.
-

На них можно построить следующие относительные показатели:

- фондовооруженность рабочих = $\frac{\text{основные фонды}}{\text{количество рабочих}}$ *=19,24 тыс. руб. на 1чел.*;

- фондоотдача = $\frac{\text{объем продукции}}{\text{основные фонды}}$ *= 1,864 руб. на 1 руб.* основных фондов;

- производительность труда = $\frac{\text{объем продукции}}{\text{численность рабочих}}$ *= 35,85 тыс. руб. на 1 чел.*

Наконец, отметим основные *функции показателей*: *познавательная* – принятие решений происходит на основе анализа статистических показателей, *пропагандистская* – пропаганда достижений.

2.2. Средние величины

При изучении статистикой массовых явлений и процессов первым результатом является большое число значений x_i признака X. Различие этих значений называют ***вариацией***, а сами x_i - ***вариантами***. Множество всех вариант характеризует не только отдельные явления, но и всю статистическую совокупность. Вполне понятно, что при изучении признака X просматривать большое количество значений x_i и трудно и неудобно. Поэтому в качестве обобщающей характеристики рассматривают среднее значение признака, обозначаемое через $\bar{x}$. *Средняя величина $\bar{x}$, обобщая качественно однородные значения признака, является типической характеристикой признака в данной совокупности.* Она относится к относительным показателям.

В зависимости от признака или от набора связанных с ним других признаков в статистике рассматривается несколько видов средних вели-

чин. *Выбор вида средней величины каждый раз должен быть строго обусловлен.* Один из принципов такого выбора связан с тем, что ***при замене всех вариант*** x_i ***на среднее значение*** $\bar{x}$ ***не должен меняться объем некоторого абсолютного признака****, связанного с X, или объем самого признака X, если он является абсолютным признаком.*

Средняя арифметическая величина

Пусть при изучении статистической совокупности *объема n,* то есть числа единиц в ней, относительно абсолютного признака X получены варианты x_i.

Средней арифметической величиной называется такое значение $\bar{x}_{ap}$*, при замене которым всех значений* x_i*, сумма совокупного признака не меняется.* То есть, выполняется равенство $\sum_{i=1}^{n} x_i = n \cdot \bar{x}_{ap}$. Отсюда

$$\bar{x}_{ap} = \frac{\sum x_i}{n}.$$

В дальнейшем пределы суммирования для удобства записи будут опускаться и будут обозначаться только в случаях, когда в этом могут возникнуть сомнения.

Если варианты x_i повторяются n_i раз ($i = 1 \div k, \sum n_i = n$), то общая сумма вариант будет равна $\sum n_i x_i$, а средняя арифметическая запишется в виде *средней взвешенной*

$$\bar{x}_{ap.\,взв.} = \frac{\sum x_i n_i}{\sum n_i}, \text{ или } \quad \bar{x}_{ap.\,взв.} = \frac{\sum x_i n_i}{n}.$$

Числа n_i называются *частотами* вариант x_i или в этой формуле их *весами.*

В качестве весов могут выступать и значения f_i некоторого признака, логически связанного с рассматриваемым признаком. В этом случае применяют формулу

$$\bar{x}_{ap.\,взв.} = \frac{\sum x_i f_i}{\sum f_i}.$$

Рассмотрим пример отыскания средней урожайности в трех колхозах, если для каждого из них известна урожайность и площадь, с которой был собран урожай:

Таблица 2.3

Колхозы	Урожайность, *ц/га* x_i	Площадь, *га* f_i	Расчетная графа $x_i f_i$
1	40	30	1200
2	45	20	900
3	60	10	600
Всего	–	60	2700

В качестве весов рассмотрим площади, с помощью которых найдем общий урожай. Тогда, согласно последней формуле, получим $\overline{x}_{ар.взв} = \frac{\sum x_i f_i}{\sum f_i} = \frac{2700}{60} = 45\ (ц/га)$. Для проверки умножим среднюю урожайность на общую площадь $45\,ц/га \cdot 60\,га = 2700\,ц$ и, как видим, общее количество абсолютного признака (общий урожай) сохранилось.

Свойства средней арифметической

1. *Сумма отклонений вариант от средней арифметической равна нулю.*

Действительно, учитывая, что постоянную величину можно выносить за знак суммы, а $\sum_{i=1}^{n} 1 = n$, получим $\sum (x_i - \overline{x}_{ар}) = \sum x_i - \sum \overline{x}_{ар} =$ $= n\,\overline{x}_{ар} - \overline{x}_{ар} \sum 1 = n\,\overline{x}_{ар} - n\,\overline{x}_{ар} = 0$.

2. *Сумма квадратов отклонений вариант от средней арифметической величины является наименьшей по сравнению с суммой квадратов отклонений их от любого другого числа.*

Действительно, пусть a – любое действительное число. Рассмотрим функцию $\varphi(a) = \sum (x - a)^2$ и найдем точку экстремума, для чего производную по a приравняем к нулю

$\varphi'(a) = \sum 2 \cdot (x_i - a) \cdot (-1) = 0 \Rightarrow \quad \sum (x_i - a) = 0 \Rightarrow \quad \sum x_i - \sum a = 0 \Rightarrow$

$\sum x_i - n\,a = 0$, откуда $a = \frac{\sum x_i}{n} = \overline{x}_{ар}$ – точка минимума.

3. *Если все варианты умножить на одно и то же число, то средняя арифметическая умножится на это число.*

Пусть c – постоянное число. Тогда

$$\overline{cx} = \frac{\sum c\ x_i}{n} = \frac{c \sum x_i}{n} = c\,\overline{x}_{ар}.$$

4. *Если каждую варианту изменить на одно и то же число, то и средняя арифметическая изменится на это число* $\overline{x+d} = \overline{x} + d$.

Обобщая третье и четвертое свойства, получим $\overline{cx+d} = c\,\overline{x} + d$.

5. *Если веса* f_i *умножить или разделить на одно и то же число, то средняя арифметическая взвешенная не изменится.*

Действительно,

$$\frac{\sum x_i (f_i \ c)}{\sum f_i \ c} = \frac{c\sum x_i \ f_i}{c\sum f_i} = \overline{x}.$$

Используя это свойство, можно не указывать абсолютные величины площадей в рассмотренном ранее примере, а рассматривать их части в общей площади, например, в процентах. При этом в качестве объема n принимается 100 %.

Средняя геометрическая величина

Рассмотрим пример: цена товара увеличилась в первый раз в *3* раза, а во второй раз в *2* раза, то есть всего за два раза цена увеличилась в *6* раз. Во сколько раз в среднем увеличивалась цена товара? Если в качестве среднего увеличения рассмотреть $\overline{x}_{ap} = (2+3) : 2 = 2.5$, то увеличение за два раза будет равно $2.5 \cdot 2.5 = 6.25$, то есть общее увеличение не сохранилось. Как видим, средняя арифметическая величина в этом случае не годится.

Поступим по-другому. Пусть a – первоначальная цена товара. Тогда после двух увеличений она станет равной $2 \cdot 3\,a$. И пусть $\overline{x}$ - среднее увеличение, тогда после двух увеличений она станет равной $\overline{x} \cdot \overline{x}\,a = \overline{x}^{\,2}\,a$. Исходя из сформулированного ранее принципа для сохранения *общего* увеличения, должно выполняться равенство $\overline{x}^{\,2}\,a = 2 \cdot 3\,a$, откуда $\overline{x} = \sqrt{2 \cdot 3} \approx 2.45$. Заметим, что в этом примере сохраняется произведение $\overline{x}^{\,2} = 2 \cdot 3$.

В общем случае, если требуется сохранить произведение *положительных* вариант, приходят к *средней геометрической* величине: $\overline{x}^{\,n} = x_1 \cdot x_2 \cdot \ldots \cdot x_n$, откуда

$$\overline{x}_{геом} = \sqrt[n]{x_1 \cdot x_2 \cdot \ldots \cdot x_n}.$$

Средняя *геометрическая взвешенная* имеет вид

$$\overline{x}_{геом\ взвеш} = \sqrt[n]{x_1^{n_1} \cdot x_2^{n_2} \cdot \ldots \cdot x_к^{n_к}},$$

где $n = n_1 + n_2 + ... + n_k$. Так было бы, например, в случае, если увеличение в x_i раз повторялось n_i раз ($i = 1 \div k$).

Средняя квадратическая и средняя кубическая величины

Средняя квадратическая величина сохраняет сумму квадратов вариант $x_1^2 + x_2^2 + ... + x_n^2 = n\,\bar{x}^2$. Отсюда

$$\bar{x}_{кв} = \sqrt{\frac{x_1^2 + x_2^2 + ... + x_n^2}{n}} = \sqrt{\frac{\sum x_i^2}{n}}$$

или

$$\bar{x}_{кв\ взвеш} = \sqrt{\frac{x_1^2 f_1 + x_2^2 f_2 + ... + x_n^2 f_n}{f_1 + f_2 + ... + f_n}} = \sqrt{\frac{\sum x_i^2\ f_i}{\sum f_i}},$$

где f_i – веса вариант x_i (i = 1, 2,…, n).

Рассмотрим пример. Пусть имеются три квадратных участка со сторонами *100, 200* и *300 м*. Чему равна средняя длина участка? И опять средняя арифметическая $\bar{x}_{ар} = (100 + 200 + 300) : 3 = 200\,м$ не подходит, так как общая площадь тогда будет равна $3 \cdot \bar{x}_{ар.}^2 = 120\,000\ м^2$, а реально она равна *10000+40000+90000 = 140000 м²*. Поэтому нужно рассмотреть равенство $3 \cdot \bar{x}^2 = 100^2 + 200^2 + 300^2$, откуда получим $\bar{x}_{кв} = \sqrt{\frac{100^2 + 200^2 + 300^2}{3}} \approx 216\ м$.

Аналогично, при условии сохранения общей суммы кубов вариант возникает *средняя кубическая величина*

$$\bar{x}_{куб} = \sqrt[3]{\frac{x_1^3 + x_2^3 + ... + x_n^3}{n}} = \sqrt[3]{\frac{\sum x_i^3}{n}}$$

и

$$\bar{x}_{куб\ взвеш} = \sqrt[3]{\frac{x_1^3 f_1 + x_2^3 f_2 + ... + x_n^3 f_n}{f_1 + f_2 + ... + f_n}} = \sqrt[3]{\frac{\sum x_i^3\ f_i}{\sum f_i}}.$$

Средняя гармоническая величина

Эта средняя сохраняет сумму величин, обратных вариантам:

$\sum \frac{1}{x_i} = n \cdot \frac{1}{\bar{x}}$ $(i = 1 \div n)$. Отсюда получают *среднюю гармоническую величину*

$$\bar{x}_{гарм.} = \frac{n}{\sum \frac{1}{x_i}} \quad \text{и} \quad \bar{x}_{гарм.\ взвеш.} = \frac{\sum f_i}{\sum \frac{f_i}{x_i}} ,$$

где f_i - веса вариант x_i ($i = 1, 2, ..., n$).

Приведем пример. Автомашина перевозила груз из пункта А в пункт В со скоростью *40 км/час*, а обратно порожняком ехала со скоростью *60 км/час*. Какова средняя скорость машины на всем пути следования? Пусть расстояние между пунктами равно *S*, а средняя скорость – $\bar{x}$. Тогда, исходя из того, что при езде со средней скоростью должно сохраниться общее время движения, получим равенство $\frac{2S}{\bar{x}} = \frac{S}{40} + \frac{S}{60}$, откуда $\bar{x} = \frac{2}{\frac{1}{40} + \frac{1}{60}} = 48\ км/час$. Как видим, средняя скорость посчитана как средняя гармоническая.

А подошла бы средняя арифметическая скорость, равная (*40 + 60*) : *2 = 50 км / час* ? Нет, так как, например при *S = 96 км* реальное время движения равно *96:40 + 96:60 = 4 час*, а при средней арифметической скорости – $\frac{2 \cdot 96}{50} = 3{,}84\ час$, то есть, нарушен принцип сохранения общего времени. Кстати, для средней гармонической скорости получим общее время $\frac{2 \cdot 96}{48} = 4\ час$. Общее время сохранилось.

Обратим внимание еще на один важный факт. Применение вида средней величины существенно зависит от сопутствующих абсолютных величин. Действительно, если бы в примере о средней урожайности из пункта «средняя арифметическая величина» были бы даны не площади, а урожаи колхозов *1200 ц, 900 ц* и *600 ц*, то нам пришлось сохранять общую площадь и исходить из равенства $\frac{1200}{40} + \frac{900}{45} + \frac{600}{60} = \frac{2700}{\bar{x}}$, откуда средняя урожайность, равная $\bar{x} = \frac{2700}{\frac{1200}{40} + \frac{900}{45} + \frac{600}{60}} = 45\ ц/га$, ищется уже как средняя гармоническая.

Отметим общий факт: между средними величинами, вычисленными для одного и того же множества вариант, выполняется *правило мажорантности*:

$$\bar{x}_{гарм} \le \bar{x}_{геом} \le \bar{x}_{ариифм} \le \bar{x}_{квадр} \le \bar{x}_{куб}.$$

При этом все они будут равны между собой только в случае равенства всех вариант и будут различны, если индивидуальные значения признака варьируют. Понятно, что это правило можно использовать в корыстных целях.

Пример из Елисеевой [1]. Статистика может в зависимости от настроения и желания «знатока» либо «утопить», либо «выручить» студента, получившего на сессии оценки *2* и *5* (при пятибальной системе). Каков его средний балл?

Если судить по средней арифметической, равной $3,5$, то он вполне успевающий студент. Но если декан желает «утопить» несчастного студента и вычислит среднюю гармоническую $\bar{x}_{гарм} = \dfrac{2}{\frac{1}{2}+\frac{1}{5}} = \dfrac{20}{7} \approx 2.86$, то и в среднем студент остается двоечником, не дотянувшим до тройки. Однако студенческий комитет может возразить декану и представить среднюю кубическую величину $\bar{x}_{куб} = \sqrt[3]{\dfrac{2^3+5^3}{2}} = \sqrt[3]{66.5} \approx 4.05$. Студент уже выглядит «хорошистом» и даже может претендовать на стипендию!

И только в случае, если студент «провалил бы» оба экзамена, статистика была бы не в силах помочь: увы, все средние из двух двоек равны все той же двойке! Как тут не вспомнить Дизраэля.

И, наконец, отметим одну важную сторону средней величины. Суть в том, что значение статистического признака представляется в виде суммы $x_i = c + \Delta_i$, где c – постоянное значение, обусловленное общей закономерностью для всех единиц совокупности (ожидаемое значение по технологии процесса), а Δ_i – отклонение от c, индивидуальное для каждой единицы совокупности и возникающее за счет влияния на нее случайных факторов. В таком случае

$$\bar{x} = \frac{\sum x_i}{n} = \frac{\sum (c + \Delta_i)}{n} = \frac{n\,c + \sum \Delta_i}{n} = c + \bar{\Delta}.$$

Как видим, в средней сохраняется слагаемое c, отражающее закономерность, а усредняются случайные «помехи». Числа Δ_i могут быть положительными и отрицательными, поэтому при осреднении они взаимно погашаются. Согласно закону больших чисел, чем больше объем совокупности, тем в большей мере проявляется это взаимопогашение и тем в большей мере средняя величина отражает закономерность статистической совокупности.

2.3. Вариация массовых явлений

В каком-то смысле, рассматриваемые в дальнейшем показатели являются также средними величинами и с разных сторон отражают то необходимое, что присуще статистической совокупности. Для изучения их нам придется познакомиться с таким понятием как *вариация массовых явлений*.

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

Не следует путать вариацию с изменением признака у одной и той же единицы совокупности с течением времени (в этом случае говорят о *динамике*, о чем речь будет позже). Например, в 1995 году заводы произвели продукцию на суммы:

Таблица 2.4

1 завод	*2* завод	*3* завод
40 млрд. руб.	*33 млрд. руб.*	*45 млрд. руб.*

Эти данные говорят о *вариации*, а данные по производству продукции на первом заводе (смотрите таблицу 2.5) отражают *динамику* процесса.

Таблица 2.5

1 завод	*1993* год	*1994* год	*1995* год
	36 млрд. руб.	*39 млрд. руб.*	*40 млрд. руб.*

Статистическим рядом распределения называется совокупность вариант или совокупность вариант с их частотами, расположенных в порядке возрастания или убывания. Во втором случае говорят о *сгруппированных данных*. Вариационный ряд еще называют *рядом распределения.*

В статистике рассматривают три формы вариационных рядов: *ранжированный*, *дискретный* и *интервальный* ряды.

Ранжированный ряд (табл. *2.6*) представляет собой перечень единиц совокупности, расположенных в порядке возрастания (убывания) значений изучаемого признака. Рассмотренные ранее данные по трем заводам, ранжированные по количеству произведенной продукции, становятся ранжированным рядом. Обычно такие ряды рассматриваются

Таблица 2.6

Предприятие	Произведенная продукция, *млрд. руб.*
2 завод	*33*
1 завод	*40*
3 завод	*45*

для совокупностей малого объема *n*.

Дискретный вариационный ряд – это перечень наблюдавшихся вариант x_i и их количеств n_i (f_i), называемых *частотой* появления варианты x_i (табл.2.7). Обозначение f_i происходит от английского слова *frequency* – частота. Такие ряды возникают, когда часто наблюдаемый признак принимает небольшое количество повторяющихся значений. В таком случае имеет смысл каждое значение связать с частотой.

Таблица 2.7

Распределение вызовов скорой помощи за один час в течение *5* суток

Число вызовов за один час, x_i	Количество наблюдений числа вызовов, f_i
0	*36*
1	*39*
2	*25*
3	*12*
4	*4*
5	*3*
6	*1*
Всего	$n = 120$

Для *непрерывного признака* в совокупностях больших объемов, как правило, наблюдаются неповторяющиеся или редко повторяющиеся значения вариант. В этом случае строят *интервальное распределение*. Для этого находят x_{min} и x_{max} – наименьшее и наибольшее значения вариант и промежуток между ними разбивают на k интервалов и подсчитывают *интервальные частоты* n_i $(i = 1 \div k)$, то есть число вариант, попавших в каждый интервал разбиения. При этом $\sum n_i = n$ – объем совокупности. *Совокупность интервалов разбиения, расположенных в порядке возрастания, и их интервальных частот называется интервальным распределением*. Отметим, что интервалы называют еще *интервалами группировки*.

Интервалы разбиения принято обозначать $(a_{i-1},\ a_i),\ i = 1 \div k$, причем $a_0 = x_{min}$, $a_k = x_{max}$. Как правило, рассматривают интервалы равной длины. Если при подсчете частот некоторые варианты совпадают с границей двух интервалов, то по ранее установленной договоренности их относят либо к предшествующему интервалу, либо к последующему интервалу. В интервальном распределении отмечают середину каждого интервала x_i $(i = 1 \div k)$ и заменяют все варианты i – того интервала на x_i с частотой n_i. Т. е. получают дискретное распределение $(x_i;\ n_i)$, $i = 1 \div k$, с которым в дальнейшем работают для расчета различного ро-

Таблица 2.8

Общая схема интервального распределения

i	Интервалы $(a_{i-1};\ a_i)$	Середины интервалов x_i	Частота n_i
1	*2*	*3*	*4*
1	$(a_0;\ a_1)$	x_1	n_1
2	$(a_1;\ a_2)$	x_2	n_2
3	$(a_2;\ a_3)$	x_3	n_3
…	…	…	…
k	$(a_{k-1};\ a_k)$	x_k	n_k
Всего			n

да показателей. Оказывается, что такой переход при достаточно большом объеме статистической совокупности и оптимальном числе интервалов группировки практически не влияет на показатели. В таблице *2.8* интервальное распределение – совокупность столбцов *2* и *4*, а столбцы *3* и *4* – соответствующее ему дискретное распределение.
И, наконец, сколько интервалов следует выбрать? Оказывается, что при достаточно больших объемах совокупности число их вычисляют по *формуле Стерджеса*

$$k \approx 1 + 3{,}322 \cdot lg\ n,$$

где n – объем совокупности, и после этого определяется *длина интервала* по формуле:

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

Можно применить и другую схему построения интервалов, в которой сначала по формуле Стерджеса определяется длина интервала:

$$h = \frac{x_{max} - x_{min}}{1 + 3{,}322 \cdot lg\ n},$$

а после этого полагают $a_0 = x_{min} - \frac{h}{2}$, $a_1 = a_0 + h$, $a_2 = a_1 + h$, и так далее, пока некоторое a_k не превзойдет x_{max}.

Замечание. В статистике совокупности, объем которых меньшие *30*, считаются малыми, при объеме между *30* и *50* - средними, а при объеме больше *100* – большими. Анализ совокупностей больших и малых объемов существенно различается, а для средних совокупностей возможно применение обеих методик.

Рассмотрим пример. Пусть среди *143* хозяйств области наименьшая урожайность зерновой культуры оказалась равной *10,6 ц/га*, а наибольшая – 53,2 *ц/га*. Тогда $k \approx 1 - 3{,}322 \cdot lg143 = 8{,}16$. Так как число групп целое, то можно построить *8* или *9* интервалов. Выберем ближайшее значение *8*. Тогда *h = (53,2 – 10, 6) : 8 = 5,325*. При округлении следует прибавлять единицу к последнему сохраняемому знаку. В нашем случае положим *h = 5,4 ц/га*. При этом максимальное значение войдет в восьмой интервал (табл. *2.9*).

А если округлять по правилам математики, то при *h = 5,3 ц/га* общая длина восьми интервалов ($5{,}3 \cdot 8 = 42{,}4$) будет меньше длины интервала вариации признака *53,2 – 10,6 = 42,6* и максимальное значение не

Таблица 2.9

Распределение хозяйств области по урожайности зерновой культуры в 198_ году

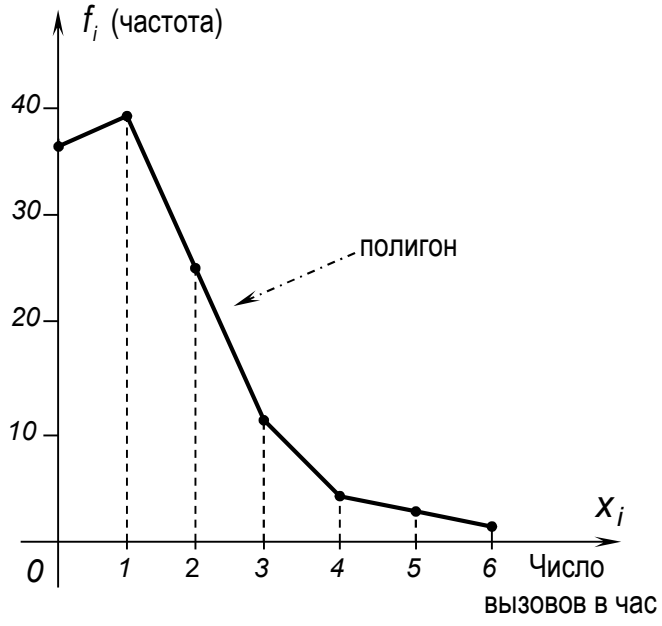
№ п/п	Группы хозяйств по урожайности, *ц/га* $a_{i-1} - a_i$	Число хозяйств n_i	Середины интервалов, *ц/га* x_i	Накопленные частоты n'_i	Условные варианты u_i	Расчетные графы	
						$u_i\ n_i$	$u_i^2\ n_i$
1	***2***	***3***	***4***	***5***	***6***	***7***	***8***
1	10,6 – 16,0	5	13,3	5	– 7	– 35	105
2	16,0 – 21,4	9	18,7	14	– 5	– 45	225
3	21,4 – 26,8	21	24,1	35	– 3	– 63	189
4	26.8 – 32,2	40	29,5	75	– 1	– 40	40
5	32,2 – 37,6	27	34,9	102	1	27	27
6	37,6 – 43,0	21	40,3	123	3	63	189
7	43,0 – 48,4	15	45,7	138	5	75	375
8	48,4 – 53,8	5	51,1	143	7	35	245
И т о г о		n = 143			0	17	1290

войдет в восьмой интервал. Придется добавлять еще и девятый интервал, быть может, только из-за одной варианты.

Результаты группировки представлены в таблице *2.9* (столбцы *2* и *3*).

2.4. Геометрическое изображение вариационных рядов

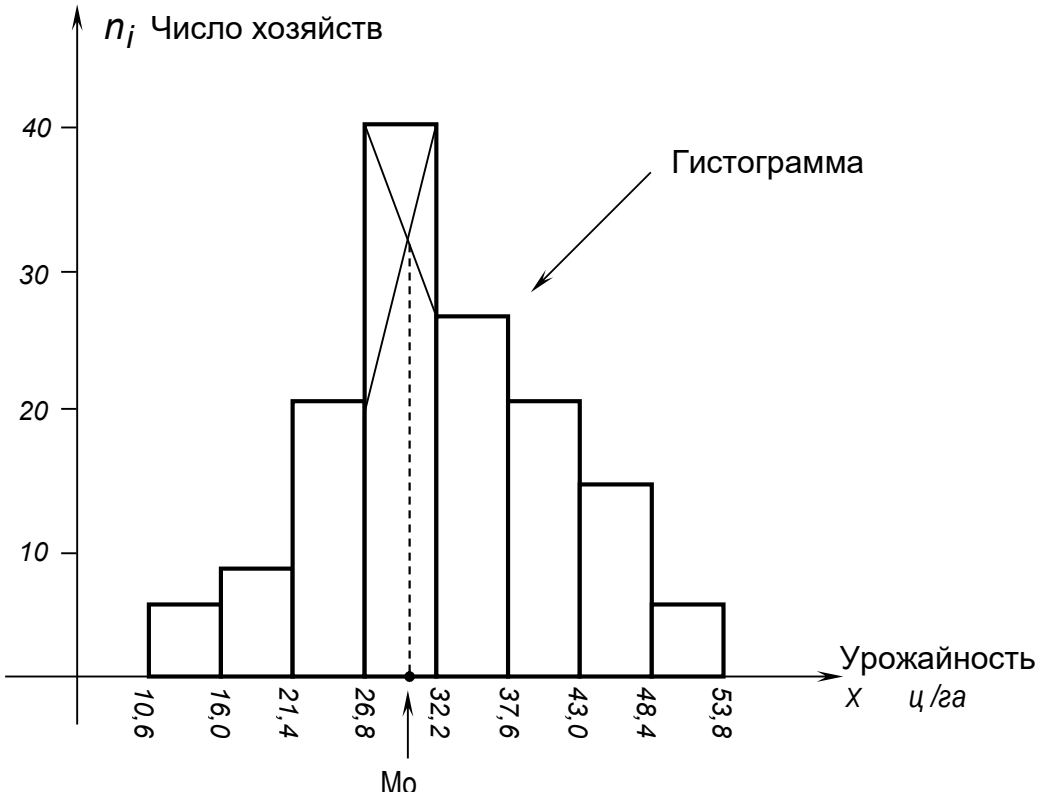
Для большей наглядности ряды распределения интерпретируют геометрически в двумерной системе координат. С этой целью для дискретного ряда на оси абсцисс откладывают значения вариант x_i, а по оси ординат соответствующие частоты f_i или n_i. Затем точки (x_i, f_i) последовательно соединяют отрезками прямых. Полученная ломаная линия называется *полигоном* (от греч. «поли» – много, «гон» – угол). На чертеже построен полигон распределения вызовов скорой помощи (табл. *2.7*). Форма полигона позволяет выдвигать ги-



потезу о виде распределения рассматриваемого признака.
В нашем случае он напоминает многоугольник распределения Пуассона, изучаемого в теории вероятностей и математической статистике.

При изображении интервального вариационного ряда на оси абсцисс откладывают интервалы группировки и на каждом из них строят прямоугольник с высотой, равной интервальной частоте. Получают столбиковую диаграмму, называемую *гистограммой* (от греч. слова «*гистос*» - ткань, строение).

На чертеже изображена гистограмма распределения хозяйств по урожайности зерновой культуры (табл. *2.9*). Ее форма показывает характерные для многих признаков особенности: чаще встречаются средние по величине значения урожайности и реже крайние, то есть большие и малые урожайности. Полученная гистограмма близка по форме к функции плотности нормального распределения. Доказано, что нормальное распределение возникает тогда, когда на вариацию признака

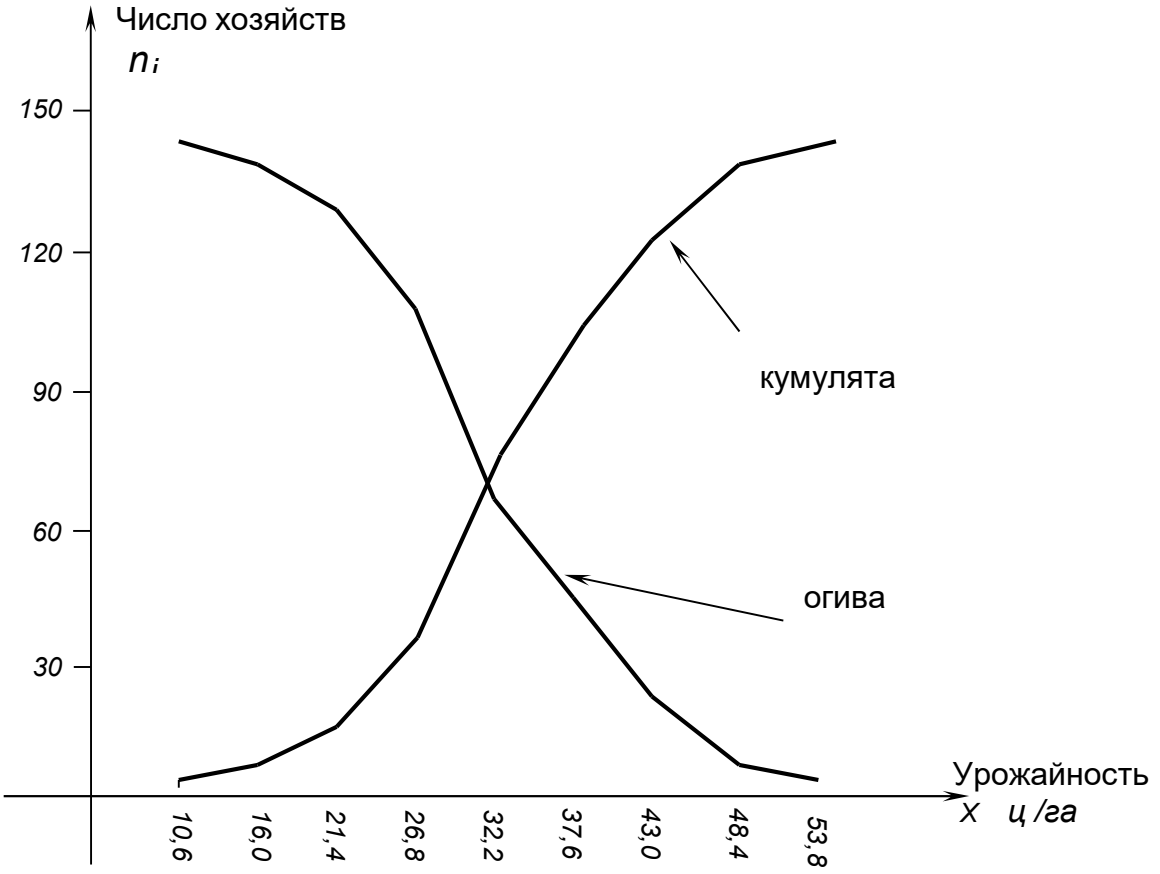


влияет большое число случайных сравнительно равнозначных факторов и ни один из них не имеет преобладающего значения.

Кумулятивное распределение

Накопленной частотой n_x называется количество вариант, не превосходящих числа x. Если после группировки мы знаем только интервальное распределение, и нет информации о наблюдавшихся значениях признака, то точно определяются значения n_x на концах интерва-

лов. По определению считают $n_{x_0} = 0$, так как нет вариант, меньших x_0. Если $x = x_1$, накопленная частота совпадает с частотой первого интервала, то есть $n_{x_1} = n_1$. Если $x = x_2$, то следует посчитать все варианты первого и второго интервалов и $n_{x_2} = n_1 + n_2$. После этого определяем $n_{x_3} = n_1 + n_2 + n_3$ и так далее. В конце последнего интервала получим $n_{x_k} = \sum x_k = n$. Для внутренних точек интервалов группировки принято условно считать, что n_x изменяется линейно от его значения на левом конце интервала до значения на правом конце. Графически для интервала (x_{i-1}, x_1) это соответствует рассмотрению отрезка прямой, соединяющей точки $(x_{i-1}, n_{x_{i-1}})$ и (x_i, n_{x_i}). График ломаной линии кумулятивного распределения называется *кумулятой*. Заметим, что иногда требуется рассмотреть «обратное» накопление частот, аналогичное рассмотренному выше, но начинающееся с максимального значения признака и увеличивающееся при движении влево, то есть накапливаются варианты « не меньшие, чем x ». График такого распределения называется *огивой*.



Наряду с изученными выше вариационными рядами в статистике используют также ряды, в которых вместо частот рассматривают *относи-*

тельные частоты или *частости* $w_i = \frac{n_i}{n}$, выражающие доли частот в общем объеме совокупности, или они же, выраженные в процентах умножением частостей на сто процентов. В силу свойства 5, средняя арифметическая величина в этом случае вычисляется по формуле $\bar{x} = \sum x_i \cdot w_i$, где $\sum w_i = 1$ (или *100* % при процентном выражении частостей). Заметим, что относительные частоты можно применять для вычисления многих других статистических показателей.

Ряд частостей используют в ситуации, когда общий объем статистической совокупности очень велик, а также когда требуется провести сравнение по одному и тому же признаку для разных совокупностей.

Применение частостей позволяет не раскрывать абсолютных величин. Как правило, частости выражают в процентах.

И, наконец, в силу разных причин приходится иметь дело с вариационными рядами с неравными интервалами. Тогда для сопоставимости частот или частостей их приводят к единице интервала, переходя к *плотности распределения:*

$$f_i' = \frac{f_i}{h_i} \quad \text{или} \quad w_i' = \frac{w_i}{h_i},$$

где f_i' называется *абсолютной плотностью* в *i*-ом интервале длиною h_i, а w_i' - *относительная плотность* распределения в *i*-ом интервале.

2.5. Структурные характеристики вариационных рядов

Основной структурной характеристикой вариационного ряда является средняя арифметическая, при вычислении которой используется вся информация об изучаемой статистической совокупности. Для интервального распределения, когда неизвестны конкретные значения признака, средняя арифметическая величина вычисляется как взвешенная с использованием середин интервалов и приписанных им интервальных частот. В ряде «Распределение хозяйств…» (табл. 2.9) получим

$$\bar{x} = \frac{13{,}3 \cdot 5 + 18{,}7 \cdot 9 + 24{,}1 \cdot 21 + \ldots + 51{,}1 \cdot 5}{143} = 32{,}52 \ \text{ц / га.}$$

Кроме того, порой возникает необходимость рассматривать такие показатели, как *мода* и *медиана*.

Мода Mo для дискретного вариационного ряда определяется как варианта, имеющая наибольшую частоту. Для ряда «Вызовы скорой помощи» *Mo = 1* (табл. *2.7*). Для интервального ряда с равными интервалами этот показатель вычисляется по формуле:

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

где x_0 – начало *модального интервала*, имеющего максимальную частоту f_{Mo};

f_{Mo-1} – частота предмодального интервала (предшествующего модальному);

f_{Mo+1} – частота постмодального интервала (следующего за модальным);

h - длина интервала.

В ряде (табл. *2.9*): $x_0 = 26{,}84;\quad h = 5{,}4;\quad f_{Mo} = 40;\quad f_{Mo-1} = 21;$ $f_{Mo+1} = 27$. Тогда

$$Mo = 26{,}84 + 5{,}4 \cdot \frac{40 - 21}{(40 - 21) + (40 - 27)} \approx 30{,}05 \quad ц / га.$$

Медиана Me делит совокупность на две равные по количеству вариант части. В ранжированном ряде с нечетным числом вариант медианой является средняя по порядку варианта. Если x_i: *1,3; 2,1; 2,7; 2,9; 4,1; 5,7; 6,1*, то *Me = 2,9*. Если же ряд имеет четное число членов, то в качестве медианы берут среднее арифметическое двух средних вариант. Для ряда x_i: *1,3; 2,1; 2,7; 2,9; 4,1; 5,7* имеем *Me = (2,7+2,9) : 2 = 2,8.* Для дискретного вариационного ряда удобно использовать накопленные частоты. Например, в распределении

Таблица 2.10

x_i	*2.1*	*3,1*	*4,5*	*4,9*
n_i	*5*	*6*	*8*	*1*
n_x	*5*	*11*	*19*	*20*

объем совокупности *n = 20, n : 2 = 10*, то есть средними являются десятая и одиннадцатая варианта и обе они равны *3,1*, следовательно, *Me = 3,1*. Для отыскания медианы интервального вариационного ряда сначала определяют медианный интервал, содержащий среднюю варианту, а затем применяют формулу:

$$Me = x_{Me} + h \cdot \frac{\frac{\sum f_i}{2} - f'_{Me-1}}{f_{Me}},$$

где x_{Me} – начало медианного интервала с частотой f_{Me};

f'_{Me-1} – накопленная частота до начала медианного интервала;

h – длина интервала группировки;

$\sum f_i = n$ – объем статистической совокупности.

В ряде «Распределение хозяйств …» (табл. *2.9*) средняя варианта имеет номер *72* и принадлежит интервалу с началом *26,8*, имеющем частоту, равную *40*, а накопленная до него частота равна *35*. Получим

$$Me = 26{,}8 + 5{,}4 \cdot \frac{\frac{143}{2} - 35}{40} \approx 31{,}7 .$$

Отметим, что рассмотренные здесь показатели называют *показателями центра распределения.* Основной характеристикой центра распределения считается средняя арифметическая величина, однако, в некоторых случаях в качестве центра существеннее использовать моду или медиану. Например, в статистическом контроле качества продукции лучше пользоваться медианой, которая не чувствительна к крайним значениям контрольной пробы. А мода применяется при изучении спроса населения на товары, когда важно заранее знать, что пользуется наибольшим спросом.

В симметричных рядах мода, медиана и средняя арифметическая равноправны, так как $Me = Mo = \bar{x}$. Это равенство нарушается в асимметрических рядах, но при этом медиана находится между модой и средней арифметической величиной:

$$Mo \le Me \le \bar{x} \quad \text{или} \quad Mo \le Me \le \bar{x} .$$

Именно поэтому медиана часто является предпочтительным показателем центра распределения.

Для более подробного изучения структуры совокупности рассматривают *квартили* Q_i ($i = 1, 2, 3$), разбивающие ее на четыре равные по количеству вариант части. То есть, четверть вариант не превосходят Q_1, четверть вариант лежит между Q_1 и Q_2, еще четверть между Q_2 и Q_3, и варианты последней четверти не меньше, чем Q_3. Для интервального распределения их находят по формулам, аналогичным формуле медианы непрерывного распределения

$$Q_1 = x_{Q_1} + h \cdot \frac{\frac{\sum f_i}{4} - f'_{Q_1-1}}{f_{Q_1}}, \qquad Q_2 = x_{Q_2} + h \cdot \frac{\frac{2\sum f_i}{4} - f'_{Q_2-1}}{f_{Q_2}},$$

$$Q_3 = x_{Q_3} + h \cdot \frac{\frac{3\sum f_i}{4} - f'_{Q_3-1}}{f_{Q_3}}.$$

Очевидно, второй квартиль равен медиане. Поэтому рассмотрим нахождение первого и третьего квартилей для ряда «Распределение хозяйств…». Имеем $n : 4 = 35,75$, следовательно (см. накопленные частоты в табл. *2.9*), начало интервала, содержащего первый квартиль, есть *26,8* и $f_{Q_1} = 40$, $f'_{Q_1-1} = 35$ а *(3п : 4) =107,25*, поэтому *37,6* – начало интервала, содержащего третий квартиль, и $f_{Q_3} = 21$, $f'_{Q_3-1} = 102$. Таким образом, получим

$$Q_1 = 26,8 + 5,4 \cdot \frac{\frac{143}{4} - 35}{40} = 26,90 \ \text{ц/га},$$

$$Q_3 = 37,6 + 5,4 \cdot \frac{\frac{3 \cdot 143}{4} - 102}{21} = 38,95 \ \text{ц/га}.$$

Тот факт, что расстояние между первым квартилем и медианой меньше расстояния между медианой и третьим квартилем, говорит о том, что в центральной части распределения наблюдается асимметрия. Это заметно и на гистограмме (стр. *25*).

Можно рассматривать разбиение на пять равных частей *квинтилями,* на десять – *децилями,* на сто – *перцентилями*. Формулы, по которым вычисляются эти показатели, аналогичны формулам квартилей.

2.6. Показатели размера и интенсивности вариации

Абсолютные показатели

Простейшим показателем является *размах вариации*

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

равный длине интервала вариации. Однако этот показатель не отражает взаимного расположения вариант внутри отрезка вариации. Для характеристики силы вариации естественно учитывать все значения. Можно было бы рассмотреть всевозможные отклонения $x_i - x_j$ вариант друг от друга, но их количество равно числу сочетаний по два из объема совокупности, и при сравнительно небольшом объеме в пятьдесят единиц

пришлось бы учесть $C_{50}^2 = 1225$ таких отклонений. Поэтому в статистике рассматривают среднее отклонение вариант от средней арифметической величины. Но сумма таких отклонений, в силу первого свойства средней арифметической величины, равна нулю. Поэтому предлагается рассмотреть *средний модуль отклонений* или *среднее линейное отклонение*

$$a = \frac{\sum |x_i - \bar{x}|}{n},$$

а для распределения с весами – *среднее взвешенное линейное отклонение*

$$a = \frac{\sum |x_i - \bar{x}| \cdot f_i}{\sum f_i} = \frac{1124,36}{143} = 7,86.$$

Здесь вычислено значение среднего модуля для распределения, рассмотренного в табл. *2.9*.

Этот показатель довольно прост в вычислениях, но неудобен в дальнейших применениях в силу «плохих» свойств модуля. В качестве другой характеристики рассматривают *среднее квадратическое отклонение* (СКО), обозначаемое малой греческой буквой сигма σ или малой латинской s (от «the standard deviaition» – *стандартное отклонение*, сокращенно «*s.d*» или просто «*s*» – обозначение, применяемое в англоязычных компьютерных программах):

$$\sigma = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n}}$$ для ранжированного ряда

или

$$\sigma = \sqrt{\frac{\sum (x_i - \bar{x})^2 \cdot f_i}{\sum f_i}}$$ для дискретного или интервального ряда.

СКО по величине всегда больше среднего модуля отклонений. За эталон принимается величина $\sigma : a \approx 1,2$, наблюдаемая для рядов, достаточно близких к нормальному распределению. Если отношение $\sigma : a$ оказывается больше, то это свидетельствует о «засоренности» совокупности элементами, неоднородными с основной массой. И чем больше это отношение, тем больше такая «засоренность».

Квадрат СКО называется *дисперсией*

$$\sigma^2 = \frac{\sum (x_i - \bar{x})^2}{n} \quad \text{или} \quad \sigma^2 = \frac{\sum (x_i - \bar{x})^2 \cdot f_i}{\sum f_i}.$$

Наряду с СКО дисперсия (*D(x)* – еще одно обозначение) является важным показателем статистического распределения и, кроме того, широко применяется во многих других вопросах статистических исследований, о которых мы будем говорить ниже. А сейчас рассмотрим некоторые свойства ее.

1. $D(x + c) = D(x)$ (c – константа) $\Rightarrow \quad \sigma(x + c) = \sigma(x)$.

2. $D(c \cdot x) = c^2 \cdot D(x) \quad \Rightarrow \sigma(cx) = c \cdot \sigma(x)$.

Другими словами, добавление ко всем вариантам одного и того же числа не меняет дисперсию и СКО, а умножение на одно и то же число увеличивает дисперсию в c^2 раз, а СКО – в c раз.

3. Полезная формула дисперсии (особенно для «ручных» вычислений) получается простыми преобразованиями:

$$\sigma_x^2 = \frac{\sum (x_i - \bar{x})^2}{n} = \frac{\sum \left(x_i^2 - 2x_i \cdot \bar{x} + \bar{x}^2\right)}{n} = \frac{\sum x_i^2}{n} - 2\,\bar{x} \cdot \frac{\sum x_i}{n} + \frac{\sum \bar{x}^2}{n}.$$

Первое слагаемое представляет собой среднее значение квадратов вариант (говорят – *средний квадрат признака*) и обозначается $\overline{x^2} = \frac{\sum x_i^2}{n}$, во втором слагаемом дробь есть средняя арифметическая величина, а $\sum \bar{x}^2 = n \cdot \bar{x}^2$, так как все слагаемые одинаковы (квадрат средней), а суммирование ведется от *1* до *n*. Поэтому $\sigma_x^2 = \overline{x^2} - 2\bar{x} \cdot \bar{x} + \bar{x}^2$. Окончательно получаем так называемую ***вычислительную формулу*** дисперсии

$$\sigma_x^2 = \overline{x^2} - \bar{x}^2 \quad \text{или} \quad \sigma_x^2 = \frac{\sum x_i^2}{n} - \left(\frac{\sum x_i}{n}\right)^2,$$

то есть дисперсия — это *средняя квадрата признака* без *квадрата средней.*

Рассмотренные свойства позволяют достаточно легко вычислять дисперсию с помощью *условных вариант* $u_i = \frac{x_i - c}{h}$, удобных в применении для рядов с равными интервалами или с равноотстоящими вариан-

тами. Отсюда $x_i = h \cdot u_i + c$. Тогда, в силу первого и второго свойств дисперсии и четвертого свойства средней арифметической величины, получим $\bar{x} = h \cdot \bar{u} + c$, а $\sigma_x^2 = h^2 \cdot D(u)$. Если число вариант в распределении нечетно, то в качестве h выбирают длину интервала в интервальном распределении или расстояние между соседними вариантами в дискретном распределении, а в качестве c – среднюю варианту x_i. Тогда условные варианты будут принимать весьма удобные значения $...,-2,-1, 0, 1, 2,...$, где 0 соответствует выбранной средней варианте. Если же число вариант четно, то в качестве c выбирают среднее арифметическое двух средних вариант, которое в случае интервального распределения, очевидно, совпадает с общей границей двух средних интервалов, а в качестве h – половину длины интервала или половину расстояния между соседними вариантами. Тогда условные варианты примут значения $...,-5,-3,-1, 1, 3, 5,...$. В обоих случаях это небольшие целые значения, удобные даже для устного счета.

В примере «Распределение хозяйств…» (табл. *2.9*) число интервалов четное. Половина длины интервала равна $5,4 : 2 = 2,7$. Положим $u_i = \dfrac{x_i - 32,2}{2,7}$. Тогда по данным таблицы $\bar{u} = 17/143 = 0,119$ и $\sigma_u^2 = \dfrac{1290}{143} - (0,119)^2 = 9,007$. И из того, что $x_i = 2,7 \cdot u_i + 32,2$, получим $\bar{x} = 2,7 \cdot 0,119 + 32,2 = 32,521$, $\sigma_x^2 = 2,7^2 \cdot 9.007 = 65,661$ и $\sigma_x = \sqrt{65,661} = 8,103$.

Отметим, что использование условных вариант весьма удобно, когда изучаемый признак принимает большие значения, так как при этом существенно уменьшаются трудоемкость расчетов и погрешности вычислений.

Конечно, кто-то возразит, зачем нужны условные варианты, если достаточно набрать данные в Excel и через секунды получить требуемые величины. Но ведь человек, научившийся ездить на машине и даже что-то знающий об ее устройстве, в случае поломки вынужден идти к специалисту – человеку, который знает машину не только умозрительно, но и «прощупал» ее руками. Так и здесь. Если вы хотите познать статистику, полезно считать «вручную», и условные варианты сыграют свою роль.

Характеристикой силы вариации в центральной части совокупности является *среднее квартильное расстояние*

$$q = \frac{(Q_3 - Me) + (Me - Q_1)}{2} = \frac{Q_3 - Q_1}{2},$$

равное среднему арифметическому длин центральных квартильных

интервалов. Для «Распределения хозяйств…» (табл. *2.9*) $q = (38{,}95 - 26{,}90) : 2 = 6{,}02$ *ц / га*. Сила вариации в центральной части совокупности в основном меньше вариации во всей совокупности. С этим фактом связывают показатель отношения среднего модуля к среднему квартильному расстоянию $a : q = 7{,}86 : 6{,}02 = 1{,}27$. Близость этого показателя к единице говорит о небольшом различии силы вариации в центральной части совокупности и на ее периферии. Большие значения отношения $a : q$ говорят о слабо варьирующем «ядре» (малое рассеяние) и сильном рассеянии окружения «ядра».

Относительные показатели вариации

Начнем с простого примера. Две совокупности *9, 10, 11* и *999, 1000, 1001*. Для обеих из них размах вариации равен *2*, средний модуль отклонений равен *2/3* и СКО равен *1*. Можно ли говорить об одинаковой силе вариации в этих совокупностях? Представим себе, что это доходы представителей двух групп населения. Совершенно очевидно, что изменение дохода на одну единицу в первой группе ощущается значительней, чем во второй. Происходит это из-за существенной разницы значений признака и, в частности, средних величин в этих совокупностях. Это одна из ситуаций, когда возникает необходимость сравнения абсолютных показателей со средней величиной.

Относительные показатели вариации служат не только для оценки интенсивности вариации, но и для сравнения ее в различных совокупностях, а также для сравнения силы вариации различных признаков. Они представляют собой отношение рассмотренных ранее абсолютных показателей к средней арифметической величине признака:

1. $\rho = R : \bar{x}$ — *относительный размах вариации*;
2. $m = a : \bar{x}$ — *относительное отклонение по модулю*;
3. $\nu = \sigma : \bar{x}$ — *коэффициент вариации* или *относительное квадратическое отклонение*;
4. $d = q : \bar{x}$ — *относительное квартильное расстояние*.

В примере из табл. *2.9* эти показатели будут иметь значения:

$\rho = 43{,}2 : 32{,}52 = 1{,}328$ или *132,8* %; $m = 7{,}86 : 32{,}52 = 0{,}242$ или *24,2* %;

$\nu = 8{,}103 : 32{,}52 = 0{,}249$ или *24,9* %; $d = 6{,}02 : 32{,}52 = 0{,}185$ или *18,5* %.

О чем говорят эти значения? Если, как в нашем случае, речь идет об отдельно взятом признаке, то для характеристики степени интенсивности вариации можно применить наработанные практикой эталоны. К

примеру, если совокупность хозяйств относится к одному природному региону, то *вариация* оценивается как *слабая*, если $\nu < 10\ \%$, как *умеренная* при $10\ \% < \nu < 25\ \%$ и как *сильная* при $\nu > 25\ \%$. В нашем примере, исходя из этого, вариацию можно считать умеренной.

Кроме этого, рассматривают еще один показатель – *однородность* совокупности. При коэффициенте вариации меньше 33% она считается однородной для распределений, близких к нормальному. В нашем примере совокупность однородна.

При изучении двух совокупностей или двух признаков сравнение соответствующих относительных показателей позволяет делать выводы о более слабой интенсивности вариации, если показатель меньше, или более сильной, если показатель больше.

Отметим, что можно рассчитать предельно возможные значения рассмотренных относительных показателей вариации в зависимости от объема совокупности n. Вот они: $\rho_{max} = n$, $m_{max} = 2 - \frac{2}{n}$, $\nu_{max} = \sqrt{n-1}$. В нашем примере при $n = 143$ получим $m_{max} = 2 - \frac{2}{143} = 1{,}986$, $\rho_{max} = 143$, $\nu_{max} = \sqrt{143-1} = 11{,}916$. Представляет интерес доля фактического показателя от предельно возможного показателя. Например, можно сравнивать такие доли при изучении совокупностей разных объемов. К тому же, сравнение реальных и предельных показателей дает возможность обнаружить ошибки расчетов.

2.7. Моменты. Показатели формы распределения

Центральным моментом к-го порядка называется величина

$$\mu_k = \frac{\sum (x_i - \bar{x})^k}{n} \quad \text{или взвешенная} \quad \mu_k = \frac{\sum (x_i - \bar{x})^k \cdot f_i}{\sum f_i}.$$

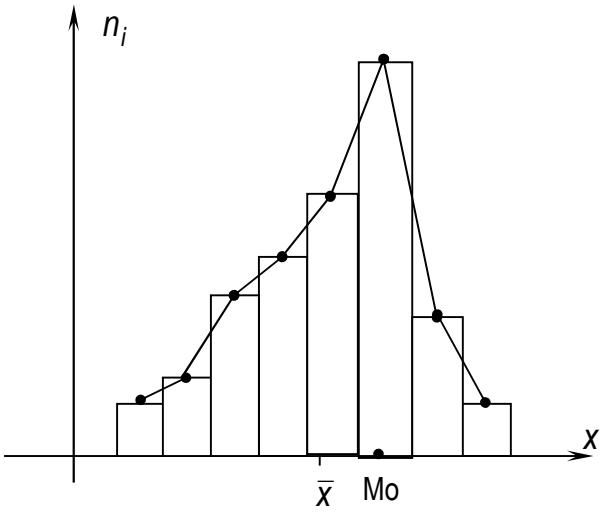
Момент первого порядка $\mu_1 = \frac{\sum (x_i - \bar{x})}{n} = 0$ в силу первого свойства средней арифметической величины. Момент второго порядка, очевидно, совпадает с дисперсией $\mu_2 = \frac{\sum (x_i - \bar{x})^2}{n} = \sigma^2$. Следующий момент третьего порядка $\mu_3 = \frac{\sum (x_i - \bar{x})^3}{n}$ в отличие от дисперсии сохраняет знаки отклонений, и в отличие от момента первого порядка в несимметричных распределениях кубы отрицательных отклонений не погашаются кубами положительных отклонений. На этом построен *показатель асим-*

метрии, характеризующий степень несимметричности рассматриваемого распределения:

$$As = \frac{\mu_3}{\sigma^3}.$$

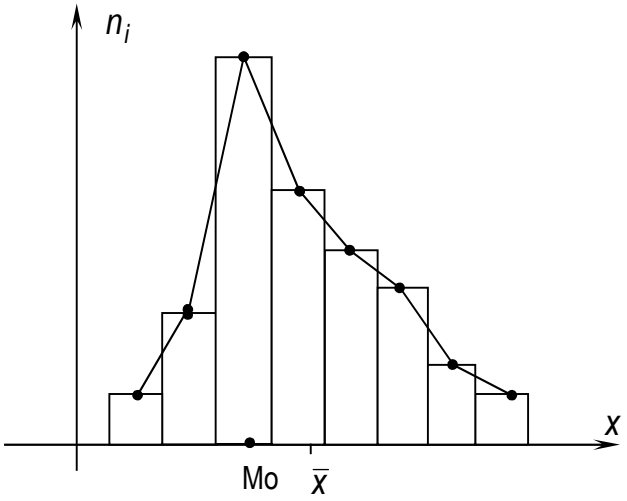
Деление на СКО в кубе не только нормирует μ_3 по величине, но и делает *показатель асимметрии* величиной, не имеющей размерности, так как и σ^3 и μ_3 имеют размерность, равную кубу размерности рассматриваемого признака.

Если $As < 0$, то асимметрия называется *левосторонней*, при этом средняя арифметическая величина лежит левее моды, а если $As > 0$, то *правосторонней*, и средняя арифметическая лежит правее моды. При левосторонней асимметрии левая от моды часть гистограммы или полигона более вытянута, чем правая, а при правосторонней – наоборот.



Левосторонняя асимметрия

As < 0



Правосторонняя асимметрия

As > 0

Английский ученый К. Пирсон ввел более простой показатель асимметрии, используя взаимное расположение моды и средней:

$$As_П = \frac{\bar{x} - Mo}{\sigma}.$$

Этот показатель отражает степень асимметрии в центре распределения в отличие от показателя, рассмотренного выше, который учитывает всю совокупность. Левосторонняя и правосторонняя асимметрии и в

этом случае имеют те же знаки. Заметим, что для симметричных распределений выполняется равенство $Mo = Me = \bar{x}$, и оба показателя асимметрии равны нулю.

По данным «Распределения хозяйств…» (табл. 2.9) получим

$$As = \frac{3773,23}{143 \cdot 8,103^3} = 0,050 \quad , \quad \text{а} \quad As_{П} = \frac{32,52 - 30,05}{8,103} = 0,305 .$$

Как видим, правосторонняя асимметрия более выражена в центре, чем во всей совокупности, что видно и на гистограмме и на полигоне.

2.8. Эксцесс

В теории вероятностей моментом четвертого порядка непрерывной случайной величины называют

$$\mu_4 = \int_{-\infty}^{+\infty} (x + M(x))^4 \cdot f(x)\, dx ,$$

где $M(x)$ – математическое ожидание, $f(x)$ – ее функция плотности. Напомним, что функция плотности наиболее распространенного нормального распределения имеет вид

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{(x-a)^2}{2\sigma^2}} ,$$

при этом a – математическое ожидание, σ – СКО, а σ^2 – дисперсия этого распределения. Рассмотрим четвертый момент для нормального распределения.

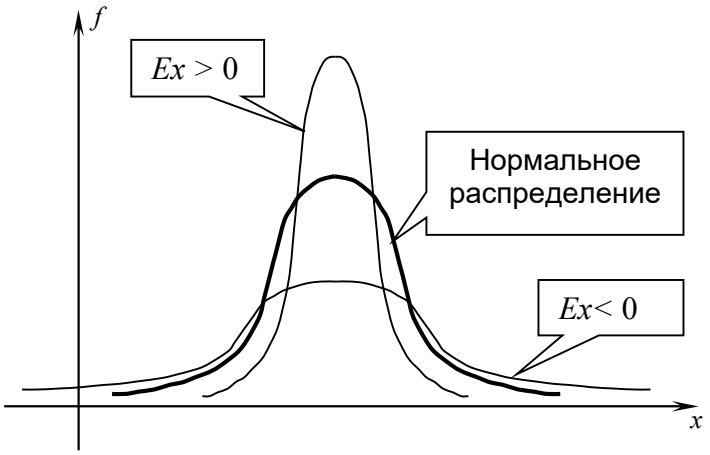
$\mu_4 = \frac{1}{\sigma\sqrt{2\pi}} \cdot \int_{-\infty}^{+\infty} (x-a)^4 e^{-\frac{(x-a)^2}{2\sigma^2}} dx$. Сделаем замену переменной $\frac{x-a}{\sigma} = t$, отсюда $x - a = \sigma t$, $x = \sigma t + a$ и $dx = \sigma \cdot dt$. Пределы интегрирования не изменятся в силу положительности сигма. Используя в последующем интегрирование по частям, получим

$$\mu_4 = \frac{1}{\sigma\sqrt{2\pi}} \cdot \int_{-\infty}^{+\infty} \sigma^4 t^4 e^{-\frac{t^2}{2}} \sigma dt = \left| \begin{array}{l} u = t^3, \; du = 3t^2 dt, dv = t e^{-\frac{t^2}{2}} dt, \\ v = \int t e^{-\frac{t^2}{2}} dt = -\int e^{-\frac{t^2}{2}} d\left(-\frac{t^2}{2}\right) = -e^{-\frac{t^2}{2}} \end{array} \right| =$$

$= \frac{\sigma^4}{\sqrt{2\pi}}\left(-t^3 \cdot e^{-\frac{t^2}{2}}\Big|_{-\infty}^{+\infty} + 3\int_{-\infty}^{+\infty} t^2 \cdot e^{-\frac{t^2}{2}} dt \right)$. Первое слагаемое в скобке равно нулю, как интеграл от нечетной функции по симметричному промежутку. Во втором слагаемом опять применим интегрирование по частям, полагая $u = t, du = dt$, а dv и v те же, что и на предыдущем шаге. Тогда $\mu_4 = \frac{3\sigma^4}{\sqrt{2\pi}}\left(-t \cdot e^{-\frac{t^2}{2}}\Big|_{-\infty}^{+\infty} + \int_{-\infty}^{+\infty} e^{-\frac{t^2}{2}} dt \right)$. Первое слагаемое в скобке равно нулю (опять нечетная функция), а второе – интеграл, «не берущийся» с помощью первообразных функций, называется *интегралом Пуассона* и вычисляется другим методом. Доказано, что $\int_{-\infty}^{+\infty} e^{-\frac{t^2}{2}} dt = \sqrt{2\pi}$. Учитывая это, окончательно получим $\mu_4 = 3\sigma^4$. Отсюда следует $\frac{\mu_4}{\sigma^4} = 3$, то есть, каковы бы ни были параметры a и σ нормального распределения, отношение момента четвертого порядка к четвертой степени СКО всегда равно трем. В статистике отклонение этого отношения от трех связывают с термином *эксцесс* (от латинского *excessus* – отступление). В данном случае – отступление от формы графика функции плотности нормального распределения. Однако, для удобства в качестве *эксцесса* применяют формулу:

$$Ex = \frac{\mu_4}{\sigma^4} - 3,$$

то есть разделяют эту характеристику на положительную или отрицательную. Если эксцесс положительный, то график распределения имеет более острую вершину, чем у нормального распределения, а если отри-



цательный, то форма более плоская. В первом случае у распределения

есть слабо варьирующее «ядро», окруженное рассеянным «гало» (от французского *halo* – белые или радужные светлые круги или *пятна* вокруг Солнца или Луны). При значительном отрицательном эксцессе такого «ядра» вовсе нет.

И, наконец, заметим, что эксцесс имеет смысл рассматривать для симметричных или «почти» симметричных распределений.

Г Л А В А 3

ГРУППИРОВКА

3. 1. Задачи и значение группировки

После сбора данных и первичной обработки результатов наблюдений возникает необходимость анализа их с целью выявления закономерностей, присущих изучаемой статистической совокупности. Трудно и практически невозможно установить эти закономерности по множеству отдельных значений признака или нескольких признаков, особенно если объем совокупности велик. А именно, большие объемы – это та база, на которой можно с большой долей уверенности выявить особенности статистического объекта. Одним из методов, применяемых с этой целью, является *группировка*.

Группировка – это распределение единиц статистической совокупности по группам, при котором различия между единицами, отнесенными к одной группе, должны быть меньше, чем между единицами из разных групп.

Например, что можно сказать, имея только данные о возрасте населения? Практически ничего, кроме того, что мы, мол, в курсе дела. А вот, выделив интервалы по возрасту и подсчитав количества людей, отнесенных к каждому из них, мы получаем возможность планирования мест в дошкольных учреждениях, в школах и других учебных заведениях, в производстве и т. д.

Значение группировки заключается в том, что она позволяет обобщать данные, представлять в компактном, а, главное, в достаточно обозримом виде. И что важно, группировка создает основу для дальнейшего анализа данных.

В статистике разработано несколько видов группировок, но при всем их различии, в них должны соблюдаться *общие правила*. Сводные показатели, рассчитанные по сгруппированным данным, будут *типичными* и *устойчивыми*, если: 1) группировка проведена правильно, 2) число единиц в группе достаточно большое. Первое требование связано с тем, что выбор количества групп или выделение групп по типам единиц совокупности, как правило, не являются очевидными. Выполнение второго правила обеспечивает погашение влияния случайных факторов и про-

явление закономерного, типичного для совокупности. Если говорить о выполнении второго правила, то понятно, что сама совокупность должна иметь довольно большой объем. А об остальном разговор будет ниже.

Для правильности группировки следует соблюдать правила отнесения единицы к той или иной группе. С этой целью определяются *группировочные признаки,* по которым будет осуществляться группировка, и их значения, отделяющие одну группу от другой и определяющие так называемые *интервалы группировки.*

Если группировка проводится по одному признаку, то она называется *простой* или *монотетической,* а если по двум и более, то *сложной* или *политетической*. Обычно сложная группировка проводится в *комбинированном* виде, когда группы, выделенные по одному признаку, разбиваются на подгруппы по другому признаку, которые в свою очередь затем разбиваются на группы по следующему признаку, и т.д. Использование нескольких группировочных признаков, конечно же, богаче по возможностям анализа, чем простая группировка. Однако такой способ сопровождается существенным увеличением числа групп (в геометрической прогрессии, если количество группировочных интервалов по каждому признаку одинаково), что приводит к измельчению групп. Так, для совокупности из ста единиц уже после группировки по второму признаку при рекомендуемых семи группировочных интервалах получим $7 \cdot 7 = 49$ групп, то есть по две единицы в каждой группе и так будет при равномерном распределении значений признаков. А если совокупность имеет «ядро», то многие группы окажутся и вовсе «пустыми». Из-за малого числа единиц совокупности в группах данные становятся плохо обозримыми, а групповые и тем самым межгрупповые и совокупные показатели становятся ненадежными (нетипичными и неустойчивыми). Как видим, комбинированный способ надежен для существенно больших совокупностей.

Другим способом является *многомерная группировка*, которая после предварительной обработки данных ведется одновременно по всем признакам.

Интервалы группировки бывают *закрытые*, когда указывается верхняя и нижняя границы их, и *открытые*, имеющие только верхнюю или только нижнюю границу. Например, интервалы возрастных групп рабочих предприятия *25 – 35*, *35 – 45*, *45 – 55* лет являются закрытыми, интервалы, определяемые неравенствами «до *25* лет» и «*55* и более лет» являются открытыми. Важно еще установить, к какой группе относить единицу, если значение признака ее совпало с границей двух интервалов. В этом случае заранее устанавливают, что будут каждый раз включать ее в предыдущий интервал (или в последующий).

Верхняя и нижняя границы соседних интервалов могут браться несовпадающими, если признак принимает целочисленные значения или значения, расположенные с некоторым необязательно целым шагом,

или в случае, когда при группировке используются только целые части значений признака. Например, при группировке по возрасту часто рассматривают число полных лет. В таком случае в приведенном выше примере интервалы будут иметь вид: «до *25* лет», *26 – 35*, *36 – 45*, *46 – 55*, «*56* и более лет». В этом случае не возникает разногласий по вопросу отнесения единиц совокупности к той или иной группе, однако возникают разногласия при выборе середины интервала.

Об определении равных по длине закрытых интервалов речь шла во второй главе. Кроме этого, группировка может проводиться с *равнонаполненными* интервалами, когда разбиение производится на группы равного объема, равного $m = n : k$, где n – объем всей совокупности, а k – число групп. В этом случае данные предварительно ранжируют. Затем в первую группу включают первые m единиц, во вторую – следующие m единиц и т.д. В качестве границ интервалов берутся фактически наблюдавшиеся значения признака – первое и последнее в каждой группе. В результате получается распределение с неравными интервалами.

3.2. Виды группировок

Целью группировки является изучение структуры совокупности, выявление закономерностей и статистических связей в изучаемой совокупности. В зависимости от поставленной цели применяют следующие виды группировок: *типологическая, структурная* и *аналитическая*.

Типологическая группировка служит для выделения социально-

Таблица 3.1

Группировка акционерных компаний города N по уровню выплаты дивидендов за 200 __ год .

Отрасль промышленности	Показатель выплаты дивидендов, %	Тип компании	Число компаний
А	Б	В	Г
Производство детских игрушек	до *30*	низкий	—
	30 – 50	средний	*1*
	50 и выше	высокий	*4*
Производство животного масла	до *20*	низкий	*1*
	20 – 40	средний	*2*
	40 и выше	высокий	—
Производство х / б тканей	до *10*	низкий	*2*
	10 – 30	средний	*4*
	30 и выше	высокий	*1*
И т о г о			*15*

экономических типов в совокупности, включающей сугубо различные (несравнимые) области социально-экономической деятельности. В основе такой группировки лежит мнение эксперта, определяющего типы явлений в области, в которой он является специалистом. При этом во всех областях привлекаются свои специалисты. Дальнейшее изучение этого вопроса проведем, рассматривая пример, заимствованный из [1].

Требуется выделить типы акционерных компаний некоторого региона с высокими, средними и низкими дивидендами и установить долю каждого типа в этом регионе.

Напомним, что показатель выплаты дивидендов показывает долю чистого дохода, выплачиваемую как дивиденды:

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

Этот коэффициент зависит от структуры акционерного капитала фирмы, возраста и перспектив ее роста. Обычно молодые, набирающие силы компании выплачивают низкие дивиденды или вообще не выплачивают, а зрелые стремятся дать более высокие дивиденды. И структура капитала и показатель выплаты дивидендов зависят также от отраслевой принадлежности фирмы. Именно поэтому, проводя группировку, кроме показателя выплаты дивидендов, мы должны учитывать еще один группировочный признак — отрасль (подотрасль).

Как видно из таблицы *3.1*, границы интервалов, определяющих типы компаний, различны. Их устанавливают упоминавшиеся выше специалисты каждой из отраслей. Эта разница называется *специализацией интервалов группировочного признака.* Специализация интервалов уравнивает оценки компаний в разных отраслях, что позволяет объединить выделенные группы в три типа независимо от отрасли. Суммируя числа компаний по типам, получим итоговую таблицу 3.2 – результат типологической группировки.

Таблица 3.2

Распределение акционерных компаний города N по уровню выплаты дивидендов за 200 __ год .

Тип компании	Число компаний	
	абсолютное	доля в %
Низкий	*3*	*20.0*
Средний	*7*	*46,7*
Высокий	*5*	*33,3*
И т о г о	*15*	*100,0*

Возвращаясь к общим вопросам, заметим, что иногда условия формирования типов приводят к различиям в их описании. Например, выделение крупных, средних и мелких предприятий в разных отраслях должно проводиться по разным характеристикам.

В энергоемких отраслях за основу берут потребление электроэнергии; в трудоемких – численность рабочих; в капиталоемких – стоимость оборудования и т.д. Это изменение круга группировочных признаков при выделении одних и тех же типов в разных условиях называется *спецификацией группировочных признаков*.

Как видим, рассмотренный метод группировки позволяет избежать дробления совокупности, но он слишком субъективен: *эксперт определяет*, какие типы должны быть выделены, по каким признакам и какими должны быть границы интервалов. Да и число группировочных признаков ограничено двумя-тремя. Но, если объект исследования хорошо изучен, если есть развитая теория, то типологическая группировка может дать хорошо интерпретируемые результаты.

Структурная группировка осуществляется при изучении структуры совокупности относительно *одного* признака. Правила такой группировки описаны в параграфе «Вариация массовых явлений», и примеры приведены в таблицах *2.7* и *2.9*. Также можно рассматривать относительные частоты или доли $w_i = \frac{n_i}{n}$ вместо частот. Структурная группировка позволяет решать многие вопросы изучения признака, оперируя небольшим числом числовых данных по сравнению с исходными данными. Кроме этого она дает возможность следить за динамикой структуры совокупности, используя распределение долей в нескольких последовательных периодах.

По известным долям w_{i0} и w_{i1} двух периодов можно рассчитать показатель среднего абсолютного изменения структуры

$$d_{w_1-w_2} = \frac{\sum_{i=1}^{k} |w_{i1} - w_{i0}|}{k},$$

где k – число групп, или сводный показатель структурных сдвигов по виду среднего квадратического отклонения

$$s_{w_1-w_0} = \sqrt{\frac{\sum_{i=1}^{k} (w_{i1} - w_{i0})^2}{k}}.$$

При отсутствии структурных сдвигов оба показателя равны нулю. Их величины тем больше, чем значительнее структурные сдвиги, причем квадратичный коэффициент более чутко реагирует на изменения структуры. Предполагается, что число групп в обоих периодах одинаково.

Аналитическая группировка позволяет установить и изучить связь между несколькими признаками, среди которых один выделяется как результат (*признак-следствие* *Y*), а остальные как факторы (*признаки-факторы X*), от которых зависит результат. Признак-следствие и признаки-факторы в статистике являются аналогами соответственно функции и аргументов в математике.

Если признак-фактор один, то говорят об *однофакторной* группировке, а если факторов два или более, то о *многофакторной* группировке.

Однофакторная аналитическая группировка начинается с ранжирования пар $(x_i; y_i)$ по признаку-фактору с последующим определением группировочных интервалов и подсчета количества единиц n_j ($j = 1 \div k$, $\sum n_j = n$ – объем совокупности). Для каждой группы указываются середины интервалов x'_j признака-фактора и вычисляются *групповые средние* признака-следствия по формуле $\bar{y}_j = \dfrac{\sum_{i=1}^{n_j} y_{ij}}{n_j}$, где y_{ij} – i –ое значение признака-следствия в j-ой группе. Результаты представляются в таблице.

Таблица 3.3

Зависимость прибыли предприятий от оборачиваемости оборотных средств в 199_ году

№ п/п	Продолжительность оборота средств, *дни*	Число предприятий	Середина интервала, *дни*	Средняя прибыль, *млн. р.*	Изменение средней прибыли, *млн. р.*
	x	n_j	x'_j	$\bar{y}_j$	$y_j - y_{j-1}$
1	40 –50	6	45	14,57	—
2	50 – 70	8	60	12,95	– 1,62
3	70 - 100	6	85	7,40	– 5,55
Итого		20		$\bar{y} = 11{,}77$	

По данным таблицы *3.3* легче обнаружить связь между признаками. И если бы интервалы были равными по длине, то можно было бы сказать, что имеется обратная связь, то есть с возрастанием продолжительности оборота средств уменьшается средняя прибыль (столбцы второй и пятый). Но если интервалы разные по длине, то рассматривают изменения результата на единицу изменений фактора – показатели, называемые *силой связи*:

$$b_{yx1} = \frac{\bar{y}_2 - \bar{y}_3}{x'_2 - x'_1} = \frac{-1{,}62}{15} = -0{,}108 \text{ млн. р./день},$$

$$b_{y\,x2} = \frac{\bar{y}_3 - \bar{y}_2}{x'_3 - x'_2} = \frac{-5{,}55}{25} = -0{,}222 \quad \textit{млн. р./день}.$$

Отрицательность показателей силы связи говорит об обратной связи между признаками, а их значения показывают величину снижения прибыли при замедлении оборачиваемости на *1* день. В нашем случае – на *0,108 млн. р*. при замедлении оборачиваемости от *40* до *60* дней и на *0,222 млн. р*. при замедлении от *60* до *100* дней. Для линейной связи показатели силы связи практически равны, а при существенно различных значениях, как в нашем случае, следует говорить о нелинейной связи. И отметим, что $b_{y\,x} > 0$ соответствует прямой связи между признаками, то есть при возрастании признака-фактора возрастает и признак-следствие, а $b_{y\,x} < 0$ — обратной связи, при которой с возрастанием признака-фактора убывает признак-следствие.

При рассмотрении зависимости от двух и более факторов каждая группа по первому из них разбивается на группы по второму фактору, которые в свою очередь делятся на группы по третьему фактору и т.д. Недостатком такого метода является сильное дробление групп с малым числом вариант. Тем не менее, и в этом случае для выявления влияния некоторого фактора на признак-следствие рассматривают частные показатели силы связи, вычисляемые при фиксированных остальных факторах. Более подробно этот вопрос можно посмотреть в [1].

Сила связи является простейшим показателем зависимости между признаками и удобным на первом этапе изучения этого вопроса. Более качественный анализ проводится с привлечением дисперсии признака-следствия, которая по результатам группировки по фактору разлагается на так называемые *внутригрупповую* и *межгрупповую дисперсии*.

Рассмотрим подробнее. В обозначениях аналитической группировки общая дисперсия будет иметь вид

$$S_y^2 = \frac{\sum_{j=1}^{k} \sum_{i=1}^{n_j} \left(y_{i\,j} - \bar{y}\right)^2}{\sum_{j=1}^{k} n_j}.$$

Двойное суммирование означает, что сначала находятся суммы по i в каждой группе, и полученные результаты суммируются по группам, то есть по j. Преобразуем дисперсию, добавив и отняв в каждой скобке групповые средние и раскрывая квадрат суммы с последующим почленным делением ее на знаменатель:

$$S_y^2 = \frac{\sum_j \sum_i \left(y_{ij} - \bar{y}_j + \bar{y}_j - \bar{y}\right)^2}{\sum n_j} =$$

$$= \frac{\sum_j \sum_i \left(y_{ij} - \bar{y}_j\right)^2}{\sum n_j} + 2\frac{\sum_j \sum_i \left(y_{ij} - \bar{y}_j\right)\cdot\left(\bar{y}_j - \bar{y}\right)}{\sum n_j} + \frac{\sum_j \sum_i \left(\bar{y}_j - \bar{y}\right)^2}{\sum n_j}.$$

Во втором слагаемом вторая скобка не зависит от *j*, поэтому ее можно вынести за знак внутренней суммы, после чего оно примет вид

$$2\frac{\sum_j \left(\left(\bar{y}_j - \bar{y}\right)\cdot\sum_i \left(y_{ij} - \bar{y}_j\right)\right)}{\sum n_j}.$$

Теперь внутренняя сумма представляет собой сумму отклонений вариант *j*-ой группы от средней этой группы и в силу первого свойства средней арифметической величины равна нулю для каждого *j*. Поэтому второе слагаемое равно нулю. Теперь введем в рассмотрение *групповые дисперсии* $S_{yxj}^2 = \frac{\sum_i \left(y_{ij} - \bar{y}_j\right)^2}{n_j}$, $(j = 1 \div k)$, или $\sum_i \left(y_{ij} - \bar{y}_j\right)^2 = S_{yxj}^2 \cdot n_j$.

Тогда первое слагаемое примет вид

$$\overline{S_{yx}^2} = \frac{\sum_j S_{yxj}^2 \cdot n_j}{\sum n_j}.$$

Полученная величина есть средняя взвешенная групповых дисперсий с весами, равными объемам групп, и называется *внутригрупповой дисперсией.*

В последнем слагаемом преобразуется внутренняя сумма $\sum_{i=1}^{n_j} \left(\bar{y}_j - \bar{y}\right)^2 = \left(\bar{y}_j - \bar{y}\right)^2 \cdot \sum_{i=1}^{n_j} 1 = \left(\bar{y}_j - \bar{y}\right)^2 \cdot n_j$. В силу этого третье слагаемое

примет вид

$$S_{\bar{y}x}^2 = \frac{\sum_j \left(\bar{y}_j - \bar{y}\right)^2 \cdot n_j}{\sum n_j}.$$

Оно представляет собой дисперсию средних групповых величин относительно общей средней и называется *межгрупповой* или *факторной дисперсией*.

Таким образом, общая дисперсия разлагается на внутригрупповую и межгрупповую дисперсии по формуле:

$$S_y^2 = \overline{S_{y\,x}^2} + S_{\bar{y}\,x}^2 .$$

Межгрупповая дисперсия отражает влияние фактора на признак-следствие, поэтому в качестве количественной характеристики связи рассматривают ее долю в общей дисперсии — *коэффициент детерминации* η^2 (греческая буква «эта»)

$$\eta^2 = \frac{S_{\bar{y}\,x}^2}{S_y^2} .$$

Отметим, что $0 \le \eta^2 \le 1$, и, чем ближе η^2 к 1, тем теснее связь между признаками (ближе к функциональной), а если оно близко к *0*, то можно предполагать отсутствие связи.

Наряду с этим показателем рассматривают *эмпирическое корреляционное отношение*

$$\eta = \pm\sqrt{\frac{S_{\bar{y}\,x}^2}{S_y^2}} .$$

Знак выбирается в зависимости от вида связи: (+) для прямой связи, (–) – для обратной. По данным таблицы 3.3 получим $\eta = 0{,}88$, то есть связь тесная.

В случае многофакторной группировки аналогично вычисляется показатель множественной тесноты связи — совокупное эмпирическое отношение (см. [1]).

Многомерные группировки в некотором смысле являются альтернативой многофакторным группировкам. Во-первых, они применимы в случае анализа совокупности относительно большого числа факторов, во-вторых, они не требуют дробления ее на группы. К тому же, многомерная группировка позволяет выделить, например, типы предприятий по финансовому положению или по эффективности их деятельности, используя множество признаков.

Мы рассмотрим один из простейших методов такой группировки, основанный на сравнении многомерных средних величин. Проблема заключается в том, что нельзя рассчитывать среднюю величину разнока-

чественных признаков. Поэтому в предлагаемом ниже *методе многомерной группировки* для каждой единицы совокупности (*i* – номер единицы) по каждому признаку (*j* – номер признака) вычисляются относительные показатели признака $p_{ij} = \frac{x_{ij}}{\bar{x}_j}$, где x_{ij} – абсолютное значение *j*–го признака *i*–ой единицы совокупности, $\bar{x}_j$ – среднее значение *j*–го признака. На основании этих показателей и вычисляется *многомерная средняя* для каждой единицы совокупности

$$\bar{p}_i = \frac{\sum_{j=1}^{k} p_{ij}}{k},$$

где *k* – число признаков.

При малом числе единиц по полученным средним можно провести «визуальную» группировку по типам, а если число единиц достаточно большое, то по значениям $\bar{p}_i$ можно провести структурную группировку.

В качестве примера рассмотрим группировку семи хозяйств, для удобства обозначенных числами от *1* до *7*, по четырем признакам [1].

Таблица 3.4

Показатели интенсивности производства в семи хозяйствах в 198_ году

i	Затраты труда, *час/ га*		Мощность двигателей, *л. с. / га*		Оборотные фонды, *руб. / га*		Основные фонды, *руб. / га*		Многомерный средний показатель, %
	x_{i1}	$\frac{x_{i1}}{\bar{x}_1}100\%$	x_{i2}	$\frac{x_{i2}}{\bar{x}_2}100\%$	x_{i3}	$\frac{x_{i3}}{\bar{x}_3}100\%$	x_{i4}	$\frac{x_{i4}}{\bar{x}_4}100\%$	
1	*292*	*79,3*	*5,55*	*76,5*	*1 042*	*81,5*	*4 063*	*88,8*	*81,5*
2	*189*	*51,3*	*3,58*	*49,3*	*683*	*53,4*	*1 990*	*43,5*	*49,4*
3	*423*	*114,9*	*8,62*	*118,8*	*1 766*	*138,2*	*4 508*	*98,5*	*117,6*
4	*353*	*95,8*	*6,47*	*89,1*	*1 339*	*104,7*	*5 061*	*110,6*	*100,05*
5	*290*	*78,7*	*6,08*	*83,8*	*1 187*	*92,9*	*3 533*	*77,2*	*83,15*
6	*693*	*188,2*	*13,94*	*192,0*	*2 186*	*171,0*	*8 190*	*179,0*	*182,0*
7	*338*	*91,8*	*6,57*	*90,5*	*745*	*58,3*	*4 682*	*102,3*	*85,7*
В среднем	*368,3*	*100*	*7,26*	*100*	*1 278*	*100*	*4 575*	*100*	*100*

Поясним вычисления. В графе «Затраты труда» для данных абсолютных величин вычисляем среднее арифметическое значение *368,3* и на-

ходим отношение показателя каждого хозяйства к полученной средней в процентах – второй столбец этой графы. Аналогично проделываем для остальных признаков. Затем для каждого хозяйства по данным вторых столбцов всех признаков вычисляем среднее значение и записываем в последний столбец.

Рассчитанные таким образом многомерные средние позволяют сравнивать хозяйства по совокупности четырех признаков и определить типы. Выделяется резко отстающее хозяйство *2*, далее идут хозяйства *1*, *5* и *7*, имеющие более высокие приблизительно равные показатели уровня интенсивности, типа «ниже среднего». В группу «немного выше среднего» входят хозяйства *3* и *4*, а хозяйство *6* имеет «очень высокий» уровень интенсивности производства. В итоге, как говорится, «на глаз» мы провели классификацию хозяйств, используя многомерную оценку их.

Заметим только, что в рассмотренном методе все признаки приходится считать равнозначными, что экономически, конечно же, не верно.

ГЛАВА 4

ВЫБОРОЧНЫЕ МЕТОДЫ

4.1. Причины и виды выборочного наблюдения

Изучаемая статистическая совокупность называется *генеральной совокупностью*. Однако часто возникает необходимость рассмотрения только некоторой части ее, называемой *выборочной совокупностью* или *выборкой*. Перенесение результатов, полученных по выборке, на генеральную совокупность и называют *выборочным методом*.

Рассмотрим причины использования выборочного метода. Их несколько.

1. Как ни странно, повышается точность: уменьшение числа единиц при переходе к выборке значительно уменьшает *ошибки регистрации*. И несмотря на то, что при этом возникают так называемые *ошибки репрезентации*, все ж общая ошибка оказывается значительно меньшей, чем при *сплошном* наблюдении.

2. Экономия материальных, трудовых и финансовых ресурсов и времени. К примеру, наблюдение за бюджетами семей с целью изучения дифференциации населения по уровню жизни, денежного обращения, определения черты бедности и т.д. при ежедневной регистрации доходов и расходов в условиях Республики Беларусь потребует порядка пятидесяти-ста тысяч статистиков. И это не считая расходов на канцелярские принадлежности и последующую обработку данных.

3. Если наблюдение связано с порчей наблюдаемых объектов, выборочный метод не имеет альтернативы. Такая ситуация возникает при изучении таких показателей качества продукции, как показатели пре-

дельной нагрузки, получение которых связано с разрушением испытуемых образцов, содержание полезных и вредных веществ в продуктах питания или в предметах гигиены и т.п.

4. В случае, когда изучаемый признак гипотетически может принимать бесконечное число значений (например, в физике или химии). В этом случае проводят отдельные эксперименты и тем самым получают выборочную совокупность.

5. Если требуется установить закон распределения случайной величины, то организуют выборочное наблюдение и на основании полученных данных выдвигают и проверяют гипотезы о виде распределения и его параметрах.

На практике выборочный метод реализуется в разнообразных формах. Например, проведя сплошное наблюдение, можно сделать выборку при разработке данных: отбирают часть данных для более подробной разработки по расширенной программе. Иногда в процессе сбора данных проводят параллельно сплошное и несплошное наблюдение. При переписях населения в СССР (*1959, 1970, 1979* гг.) собирались сведения о каждом лице по *11* признакам, а *25*% населения давали более подробную информацию (*18* вопросов).

Начало применения и формирования выборочного метода относится к XVII – XVIII векам. Важную роль сыграли работы Якоба Бернулли (1654 – 1705). Заметный вклад в разработку теоретических основ выборочного метода внесли русские математики П. Л. Чебышев, А. М. Ляпунов, А. А. Марков. Позже теория этого метода получила развитие в трудах известного русского статистика А. А. Чупрова и в работе А.Г. Ковалевского «Основы теории выборочного метода», вышедшей в 1924году. В 1930 году известные советские статистики А. Я Боярский и Б. С. Ястремский классифицировали формы выборочного наблюдения. В последние годы выборочные методы стали применяться практически во всех областях человеческой деятельности.

Возникновение выборочного метода определило разделение статистики на *описательную (дискриптивную)* и *выводную.* Целью первой является сбор данных по всем единицам генеральной совокупности, их обработка и получение сводных показателей, характеризующих только эту совокупность. Если же изучаемая совокупность является частью некоторой совокупности, и полученные результаты переносят на генеральную совокупность, то говорят о выводной статистике. Например, можно изучить успеваемость в студенческой группе по данным на всех ее студентов, то это описательная статистика. А если по этим данным судить обо всем курсе, то это уже выводная статистика.

В выводной статистике устанавливается соответствие показателей генеральной и выборочной совокупностей. Это соответствие позволяет использовать показатели выборки как приближенные значения (*оценки*) соответствующих неизвестных параметров генеральной совокупности.

В связи с этим возникает вопрос о *репрезентативности* (представительности) выборки, т. е. наиболее полном отражении свойств генеральной совокупности данными, полученными по выборочной совокупности.

В дальнейшем будут использоваться следующие обозначения:

Таблица 4.1

Генеральная совокупность	Параметры	Выборочная совокупность
μ, $\bar{x}_г$	Математическое ожидание, средняя величина	$\bar{x}$, $\bar{x}_в$
σ^2	Дисперсия	s^2, $s_в^2$
p	Доля признака (относительная частота)	w
ρ	Коэффициент корреляции	r
N	Объем совокупности	n
θ	Общее обозначение параметра	$\hat{\theta}$

4.2. Требования, предъявляемые к выборке. Способы отбора

Репрезентативность выборки может быть обеспечена исключительно за счет *объективности отбора* данных из генеральной совокупности, что обеспечивается выполнением двух требований:

- *случайность* включения единиц генеральной совокупности в выборку;
- *достаточный объем* выборочной совокупности.

Различают три вида отбора.

1. Простой случайный отбор, производящийся из всей генеральной совокупности без предварительного расчленения ее на части, и единица отбора совпадает с единицей наблюдения. При этом возможно применение двух схем отбора:

а) повторная выборка, осуществляемая по принципу возвращенного шара, когда одна и та же единица генеральной совокупности может быть учтена в выборке несколько раз;

б) бесповторная выборка – случайный отбор, при котором отобранная единица не возвращается обратно, и, следовательно, в выборке все единицы совокупности различны.

В экономике и социологии применяется, как правило, бесповторный отбор. При расчетах следует учитывать тот факт, что различие этих схем сказывается на вероятности попадания единиц в выборку. В слу-

чае *а)* она одинакова для всех единиц и равна *1 / N,* а в случае *б)* она увеличивается по мере отбора и равна *1 / (N – к + 1)* для *к*-ой отбираемой единицы. Естественно, это сказывается на величинах выборочных оценок и, следовательно, должно учитываться при их расчете.

Случайность отбора обеспечивается за счет применения принципа лототрона. Для этого каждой единице генеральной совокупности предварительно приписывается код (цифровой номер или код). Затем производится жеребьевка с помощью шаров или карточек с тем же набором кодов. Выпавшие коды, число которых заранее планируется, определяют единицы совокупности, попавшие в выборку. Однако более надежным является отбор с помощью таблиц случайных чисел, который можно осуществить на компьютере, что существенно упрощает всю процедуру.

2. Сложный отбор – вся совокупность предварительно «разбивается» на части:

а) механический отбор, при котором генеральная совокупность делится на *п* равных групп, и случайным образом выбирается по одной единице из каждой группы. Если в совокупности установлен порядок (нумерация), то, например, при *10*%-ой выборке можно, выбрав случайно третью из первых десяти единиц, затем отбирать единицы с номерами *13*, *23*, *33* и т.д.

в) серийный отбор – генеральная совокупность делится на серии, и в выборку попадают *все* единицы случайным образом отобранных нескольких серий. Например, изучается вся продукция двух-трех станков из всех, производящих одно и то же изделие.

с) *типический отбор*, при котором в генеральной совокупности выделяются *типы* единиц *(страты* или *районы)*, и проводится случайный отбор из каждого типа, при этом число единиц каждого типа должно быть пропорционально объемам районов (страт), а их общее число равно *п*. Выборка при этом виде отбора достаточно хорошо отражает структуру генеральной совокупности.

В заключение отметим, что отбор, проведенный без какой либо схемы, может привести к нерепрезентативной выборке.

4.3. Ошибки выборки (репрезентации)

Ошибкой выборки или *ошибкой репрезентативности* называют разницу между генеральным параметром и значением выборочного показателя, оценивающего этот показатель. Для рассмотренных выше показателей вводятся следующие обозначения:

$\varepsilon_{\bar{x}} = \bar{x} - \mu$ или $\varepsilon_{\bar{x}} = \bar{x}_в - \bar{x}_г$ – ошибка выборочной средней,

$\varepsilon_{s^2} = s^2 - \sigma^2$ – ошибка выборочной дисперсии,

$\varepsilon_w = w - p$ – ошибка выборочной доли.

Естественно возникает вопрос: как, не зная значения параметра генеральной совокупности (в дальнейшем «теоретический параметр»), оценить ошибку выборки?

Пусть θ – некоторый (неизвестный) теоретический параметр, а θ^* – его оценка по выборке. Представим себе, что было проведено бесконечное число выборок, для которых получены оценки $\theta_1^*, \theta_2^*, \theta_3^*, \ldots$. Рассматривая $\xi_i^2 = \varepsilon_{\theta_i^*}^2 = (\theta_i^* - \theta)^2$ как случайную величину, найдем среднеквадратическую ошибку $s_{\theta^*} = \sqrt{\frac{\sum \xi_i^2 f_i}{\sum f_i}} = \sqrt{\frac{\sum (\theta_i^* - \theta)^2 f_i}{\sum f_i}}$ (здесь θ_i^* повторяется f_i раз). Параметр θ неизвестен, но если оценка θ^* *несмещенная*, то есть $M(\theta^*) = \theta$, тогда

$$s_{\theta^*} = \sqrt{\frac{\sum (\theta_i^* - M(\theta^*))^2 f_i}{\sum f_i}} = \sqrt{D(\theta^*)}.$$

Как видим, отыскание *средней ошибки* s_{θ^*} сводится к нахождению дисперсии выборочного показателя θ^*.

Возникает вопрос, насколько можно доверять средней ошибке? Для этого следует оценить вероятность отклонения показателя θ^*, полученного по выборке, от оцениваемого теоретического параметра θ. Однако на практике важнее уверенность в ошибке. Ее связывают с рассмотрением равенства

$$P\left(\left|\theta - \theta^*\right| < t \cdot \sigma_\theta\right) = \gamma.$$

При этом *доверительную вероятность* γ задают заранее, исходя из поставленной задачи, и находят соответствующее ей значение параметра *t.* Величина $t \cdot \sigma_\theta$ называется *предельной ошибкой* выборочной оценки θ^*, а интервал

$$\theta^* - t \cdot \sigma_\theta < \theta < \theta^* + t \cdot \sigma_\theta,$$

полученный из неравенства, стоящего под знаком вероятности, *доверительным интервалом*. Если, к примеру, $\gamma = 0{,}95$, то говорят, что в 95%

случаев доверительный интервал накрывает параметр θ, или о *95*-процентной гарантии полученной оценки.

4.4. Средняя и предельная ошибки выборочной средней

Напомним, что формулы средней и предельной ошибок получены в предположении *несмещенности* оценки. Рассмотрим этот вопрос для выборочной средней величины.

Пусть X – некоторый признак, математическое ожидание которого равно μ, а дисперсия - σ^2, и $x_1, x_2, ..., x_n$ - выборочные данные, по которым вычислена средняя $\bar{x}_в = \dfrac{x_1 + x_2 + ... + x_n}{n}$. Каждое из выборочных значений x_i можно рассматривать как случайную величину X_i с теми же параметрами распределения, что и породивший их признак X, то есть $M(X_i) = \mu$, а $\sigma^2(X_i) = \sigma^2$. Тогда $\bar{x}$ можно рассматривать как среднюю случайных величин $\bar{x}_в = \dfrac{X_1 + X_2 + ... + X_n}{n}$, и учитывая свойства математического ожидания и дисперсии, получим

$$M(\bar{x}_в) = M\left(\frac{X_1 + X_2 + ... + X_n}{n}\right) = \frac{M(X_1) + M(X_2) + ... + M(X_n)}{n} = \frac{n\,M(X)}{n} = \mu,$$

Из этого равенства вытекает несмещенность $\bar{x}$ как оценки математического ожидания или генеральной средней. Тогда

$$D(\bar{x}_в) = D\left(\frac{X_1 + X_2 + ... + X_n}{n}\right) = \frac{D(X_1) + D(X_2) + ... + D(X_n)}{n} = \frac{nD(X)}{n^2} = \frac{\sigma^2}{n}$$

и получаем *стандартное* или среднеквадратическое отклонение выборочной средней величины

$$\sigma_{\bar{x}_в} = \frac{\sigma}{\sqrt{n}}.$$

Полученные результаты позволяют утверждать, что *средняя ошибка выборочной средней величины* как оценки математического ожидания или генеральной средней величины имеет вид:

$$s_{\bar{x}в} = \frac{\sigma}{\sqrt{n}}.$$

Отсюда видно, что средняя ошибка уменьшается с увеличением объема выборки *n*. А учитывая, что для нормального распределения верно $\sigma \cong \frac{R}{6}$, то, чем больше размах вариации *R* признака *X*, тем больше средняя ошибка выборочной средней.

Рассматривая выборочную среднюю как случайную величину (в силу случайности выборки), отметим, что при достаточно большом объеме выборки ее распределение является нормальным или близко к таковому, независимо от того, какое распределение имел признак в генеральной совокупности. С увеличением числа выборок их общая средняя величина будет приближаться к генеральной средней величине или математическому ожиданию. Следовательно, возникает возможность с определенной вероятностью оценить надежность ошибки.

Напомним, что для нормально распределенной случайной величины с параметрами распределения μ и σ доверительная вероятность вычисляется по формуле $P\left(\left|x-\mu\right|<t\sigma\right)=2\Phi\left(t\right)$, где $\Phi(t)=\frac{1}{\sqrt{2\pi}}\int_0^t e^{-\frac{x^2}{2}}dx$.

Для величины $\bar{x}_в$ следует заменить μ на $\bar{x}_г$, а σ на $s_{\bar{x}_в}=\frac{\sigma}{\sqrt{n}}$. Тогда получим $P\left(\left|\bar{x}_в-\bar{x}_г\right|<t\frac{\sigma}{\sqrt{n}}\right)=2\Phi(t)$. Величина $\gamma=2\Phi(t)$ называется *доверительной вероятностью*, а величина

$$\Delta_{\bar{x}}=t\frac{\sigma}{\sqrt{n}}$$

называется *доверительной ошибкой выборки* или *предельной ошибкой выборочной средней.* На практике задаются доверительной вероятностью γ, определяющей степень уверенности; из равенства $\Phi(t)=\frac{\gamma}{2}$ по таблицам этой функции (приложение *2*) находят значение аргумента t, а затем находят предельную ошибку и доверительный интервал $\left(\bar{x}_в-t\frac{\sigma}{\sqrt{n}};\ \bar{x}_в+t\frac{\sigma}{\sqrt{n}}\right)$. Таким образом, средняя величина генеральной совокупности с доверительной вероятностью γ или с уверенностью $\gamma\cdot100\%$ заключена в пределах

$$\bar{x}_в-t\frac{\sigma}{\sqrt{n}}<\bar{x}_г<\bar{x}_в+t\frac{\sigma}{\sqrt{n}}.$$

Отметим, что на практике доверительную вероятность принимают равной одному из следующих значений *0,95*, *0,954*, *0,997*, или *0,999*. Это означает, что при уровне доверия *0,95* только в *5* случаях из *100* генеральная средняя величина выйдет за указанные пределы. При вероятности *0,954* это может произойти в *46* случаях из *1000*, при *0,997* – в *3* случаях из *1000*, при *0,999* – в 1 случае из *1000*. При всех рассмотренных вероятностях можно говорить о практически полной уверенности в определении предельной ошибки.

Итак, для вычисления предельной ошибки при заданной доверительной вероятности потребуется дисперсия признака генеральной совокупности, которая на практике, как правило, не известна. Но можно определить выборочную дисперсию s^2, которая является *смещенной* оценкой дисперсии σ^2 с коэффициентом $(n-1)/n$, то есть имеет место равенство

$$M\left(s_в^2\right) = \frac{n-1}{n} \cdot \sigma^2 .$$

Поэтому рассмотрим *исправленную дисперсию*

$$s^2 = \frac{n}{n-1} \cdot s_в^2 ,$$

которая уже является несмещенной оценкой, так как

$$M\left(s^2\right) = M\left(\frac{n}{n-1} \cdot s_в^2\right) = \frac{n}{n-1} \cdot M\left(s_в^2\right) = \frac{n}{n-1} \cdot \frac{n-1}{n} M\left(\sigma^2\right) = M\left(\sigma^2\right).$$

Подставляя в формулу средней ошибки выборочной средней исправленную дисперсию, получим среднюю и предельную ошибки

$$s_{\bar{x}_в} = \frac{s}{\sqrt{n-1}} , \qquad \Delta_{\bar{x}_в} = t\frac{s}{\sqrt{n-1}}$$

При больших n ($n > 100$) величина $(n-1)/n$ практически равна единице. Поэтому выборочную дисперсию можно принимать за несмещенную оценку генеральной дисперсии и пользоваться формулами:

$$s_{\bar{x}_в} = \frac{s}{\sqrt{n}}, \quad \Delta_{\bar{x}_в} = t\frac{s}{\sqrt{n}}$$

Заметим, что индекс $в$ в $\bar{x}_в$ для упрощения записи можно опускать.

Рассмотрим *п р и м е р*. Для изучения вопроса об обеспеченности населения одного населенного пункта был проведен случайный опрос 200 человек. В результате выяснилось, что в среднем на одного человека приходится *9,61 м²* жилой площади со стандартным отклонением *3,33 м²*. С доверительной вероятностью *0,95* определить пределы для средней жилой площади на одного человека в этом населенном пункте.

По таблице (приложение *2*) из равенства $\Phi(t) = \frac{0,95}{2} = 0,475$ находим $t = 1,96$. Так как объем выборки *200* достаточно большой, то нет необходимости исправлять стандартное отклонение. Тогда средняя ошибка будет равна

$$s_{\bar{x}_в} = \frac{3.33}{\sqrt{200}} = \pm 0.24 \ м^2,$$

а предельная ошибка при доверительной вероятности *0,95* составит

$$\Delta_{\bar{x}_в} = 1.96 \cdot 0.24 = 0.47 \ м^2.$$

Найдем доверительный интервал $9,61 - 0,47 < \bar{x}_г < 9.61 + 0.47$ или $9,14 м^2 < \bar{x}_г < 10,08 м^2$. Следовательно, с *95*-процентной уверенностью можно утверждать, что средняя жилая площадь, приходящаяся на одного человека в данном населенном пункте, не менее *9,14 м²*, но не более *10,08 м²*.

4.5. Средняя и предельная ошибки выборочной доли

Долей или *выборочной относительной величиной* (*p*) называется отношение объема (числа) единиц, отнесенных к некоторой категории, к объему (числу единиц) всей совокупности.

Доля порождает *альтернативную* или *дихотомическую* переменную, которая принимает значение *1*, если единица совокупности принадлежит к выделенной категории, и *0*, если нет. Пусть n – *объем* выборочной совокупности, f_1 – количество единиц, относящихся к доле, f_2 – не относящихся к ней: $f_1 + f_2 = n$. Тогда среднее значение этой переменной будет

$$\frac{1 \cdot f_1 + 0 \cdot f_2}{f_1 + f_2} = \frac{f_1}{n} = w,$$

где w – выборочная доля. Найдем ее дисперсию

$$s^2 = \frac{(1-w)^2 f_1 + (0-w)^2 f_2}{f_1 + f_2} = \frac{f_1 - 2wf_1 + w^2 f_1 + w^2 f_2}{n} =$$

$$= \frac{w^2 (f_1 + f_2) + (1-2w) f_1}{n} = w^2 \cdot \frac{n}{n} + \frac{f_1}{n} \cdot (1-2w) = w^2 + w\ (1-2w) = w(1-w).$$

Отсюда $s_w = \sqrt{w\ (1-w)}$. Следовательно, *средняя ошибка выборочной доли* будет иметь вид:

$$s_w = \sqrt{\frac{w\ (1-w)}{n}},$$

а *предельная ошибка выборочной доли* с заданной доверительной вероятностью будет иметь вид:

$$\Delta_w = t \cdot s_w = t \cdot \sqrt{\frac{w \cdot (1-w)}{n}}.$$

Рассмотрим *п р и м е р.* В условиях предыдущего примера была выделена доля: из *200* опрошенных *46* жителей имеют жилплощадь в пределах от *11,1 м²* до *17,0 м²*. При доверительной вероятности *0,924* найти пределы изменения доли этой категории для всего населенного пункта.

Решение. Найдем выборочную долю *w = 46 : 200 = 0.23* и из равенства $\Phi(t) = \frac{0{,}924}{2} = 0{,}462$ по *таблице 2* найдем *t = 1,77*. Тогда $s_w = \sqrt{\frac{0.23 \cdot 0.77}{200}} = \pm 0.0297$ и $\Delta_w = \pm 1.77 \cdot 0.0297 = 0.053$. Таким образом, с вероятностью *0,924* генеральная доля заключена в пределах *0,177 < p < 0,283* или в процентах *17,7% < p < 28,3%.*

4.6. Предельная ошибка выборки при неизвестном σ. Распределение Стьюдента

Выше было показано, что для вычисления предельной ошибки требуется знать дисперсию признака в генеральной совокупности, которая на практике, как правило, неизвестна. В таком случае можно было бы рассмотреть выборочную дисперсию $s_в^2$, но она является смещенной оценкой дисперсии σ^2 и использование ее приведет к регулярным (не случайным) ошибкам. Поэтому рассматривают *исправленную дисперсию*

$$s^2 = \frac{n}{n-1} \cdot s_в^2,$$

которая является уже несмещенной оценкой дисперсии генеральной совокупности. Тогда рассмотренная ранее вероятность отклонения выборочной средней от генеральной средней запишется в виде $P\left(|\bar{x}_в - \bar{x}_г| < t \cdot \frac{s}{\sqrt{n-1}}\right)$. Так как s является случайной величиной (в силу случайности выборки), то перепишем его в виде $P\left(\frac{|\bar{x}_в - \bar{x}_г|}{s/\sqrt{n-1}} < t\right)$. Случайная величина в левой части неравенства имеет уже распределение, отличное от нормального, и следовательно, отыскание параметра t по формуле $\Phi(t) = \frac{\gamma}{2}$ должно быть поставлено под сомнение, так как это будет приводить к ошибкам, особенно заметным при малых значениях *n*.

Английский статистик В. Госсет (псевдоним «Стьюдент») изучил случайную величину (статистику)

$$T = \frac{\bar{x} - a}{s/\sqrt{n-1}},$$

распределение которой называют *распределением Стьюдента* или *t-распределением* [2]. Оказалось, что дифференциальная функция ее не зависит от неизвестных параметров *a* и *σ*, а зависит только от объема выборки *n* (или, что то же, от числа степеней свободы *d.f* : *k* = *n* – *1* в обозначениях некоторых источников).

Тогда предельная ошибка выборки будет вычисляться в виде

$$\Delta_{\bar{x}_в} = t_\gamma \frac{s}{\sqrt{n}} \quad \text{или} \quad \Delta_{\bar{x}_в} = t_\gamma \frac{s_в}{\sqrt{n-1}},$$

где $t_\gamma = t(\gamma; n)$ определяется по таблице распределения Стьюдента по объему выборки n (или по числу степеней $n - 1$) и доверительной вероятности γ.

Замечание. При $n > 100$ считается, что $(n-1)/n \approx 1$, и поэтому в первой из формул предельной ошибки применяют выборочную дисперсию $s_в^2$ вместо исправленной дисперсии s^2.

Отметим важное для практики свойство распределения Стьюдента. При $n \to \infty$ оно стремится к нормальному распределению с функцией

плотности $\phi(t) = \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{t^2}{2}}$. Поэтому для больших n при отыскании параметра t можно исходить из нормального распределения, используя равенство $\Phi(t) = \frac{\gamma}{2}$. Причем, при доверительной вероятности $0,95$ относительная погрешность предельной ошибки при такой перестановке не превысит одного процента (малость с точки зрения статистики). И она будет уменьшаться по мере увеличения n.

Пример. Для изучения вопроса об обеспеченности жителей одного населенного пункта был проведен случайный опрос 200 человек. В результате выяснилось, что в среднем на одного человека приходится $9,61\ м^2$ жилой площади со стандартным отклонением $3,33\ м^2$, найденными по выборке. С доверительной вероятностью $0,95$ определить пределы для средней жилой площади на одного человека в этом населенном пункте.

У нас $\gamma = 0,95$, а число степеней свободы $k = 200 - 1 = 199$. Тогда по таблице распределения Стьюдента определяем $t_\gamma = t(0,95;199) = 1,96$. Так как объем 200 выборки достаточно большой, то нет необходимости исправлять стандартное отклонение. После этого предельная ошибка составит

$$\Delta_{\bar{x}} = 1,96 \cdot \frac{3,33}{\sqrt{200}} = \pm 0,47 \quad м^2.$$

Найдем доверительный интервал $9,61 - 0,47 < x_г < 9,61 + 0,47$ или $9,14\ м^2 < x_г < 10,08\ м^2$. Следовательно, с 95–процентной уверенностью можно утверждать, что средняя жилая площадь на одного человека в данном населенном пункте больше, чем $9,14\ м^2$, но меньше, чем $10,08\ м^2$.

Заметим, что и из равенства $\Phi(t) = \frac{0,95}{2} = 0,475$ по *таблице 2* находим также $t = 1,96$.

4.7. Влияние вида выборки на величину ошибки

Как мы выяснили ранее, основным и самым трудоемким показателем, влияющим на ошибки выборки, является дисперсия. Естественно, ее величина зависит как от способа отбора, так и от вида выборки.

Рассмотрим сначала случай, когда при любом виде выборки осуществляется *повторный отбор*. Тогда, как отмечалось ранее, вероятность попадания любой единицы в выборочную совокупность остается неизменной на протяжении всей процедуры отбора. При этом под единицей

понимается собственно единица совокупности или серия, или страт. Поэтому дисперсия будет зависеть только выбора формулы.

В *серийной выборке* дисперсия определяется как колеблемость между сериями и по виду является аналогом межгрупповой дисперсии:

$$s_{\breve{x}}^2 = \frac{\sum_{j=1}^{r} (\breve{x}_j - \bar{x})^2 f_j}{\sum f_j},$$

где $\breve{x}_j$ – среднее значение признака в *j*-ой серии; $\bar{x}$ – среднее значение признака во всей выборке и f_j – объем *j*-ой серии. Если объемы серий одинаковы, то формула преобразуется к виду

$$s_{\breve{x}}^2 = \frac{\sum_{j=1}^{r} (\breve{x}_j - \bar{x})^2}{r},$$

где r – число отобранных серий.

При *типической выборке* дисперсия вычисляется как средняя (взвешенная) внутрирайонных дисперсий, то есть по схеме внутригрупповой дисперсии:

$$\overline{s^2} = \frac{\sum_{j=1}^{m} s_{x_j}^2 n_j}{\sum_{j=1}^{m} n_j},$$

где $s_{x_j}^2 = \frac{\sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j)^2}{n_j}$ – выборочная дисперсия признака *x* в *j*-ом страте (районе) объема n_j; m – число районов.

Исходя из правила сложения дисперсий, заключаем, что дисперсии, полученные по этим формулам, будут меньше своих общих дисперсий. Поэтому использование их уменьшит величину ошибки, и в этих двух случаях она будет меньше ошибки простой выборки.

Часто используемое сочетание районированного (стратифицированного) отбора с отбором сериями в каждом районе обеспечивает преимущества в организации выборки и уменьшение ошибки выборки. Дисперсия такой выборки $\overline{s_{\breve{x}}^2}$ вычисляется как средняя межсерийных дисперсий $s_{\breve{x}_j}^2$ для каждого из *m* районов:

$$\overline{s^2_{\breve{x}}} = \frac{\sum_{j=1}^{m} s^2_{\breve{x}_j} r_j}{\sum_{j=1}^{m} r_j},$$

где $s^2_{\breve{x}_j}$ – дисперсия для *j*-го района, вычисляемая по формуле:

$$s^2_{\breve{x}_j} = \frac{\sum_{i=1}^{r_j} (\breve{x}_{ij} - \bar{x}_j)^2}{r_j}.$$

Здесь $\breve{x}_{ij}$ – средняя в *i*-й серии *j*-го района; $\bar{x}_j$ – средняя выборочная в *j*-м районе; r_j – число серий, отобранных в *j*-м районе; *m* – число районов. Последняя формула применяется в случае, когда объемы серий одинаковы. В противном случае следует вычислять величину взвешенной дисперсии.

Для доли дисперсия вычисляется аналогично:
при *серийной выборке*

$$s^2_w = \frac{\sum_{1}^{r} (w_j - w)^2}{r},$$

где w_j – доля единиц определенной категории в *j*-й серии, *w* – доля единиц этой категории в выборке;
при *районированной серийной* выборке

$$\overline{s^2_w} = \frac{\sum_{1}^{m} s^2_{w_j} r_j}{\sum_{1}^{m} r_j},$$

где $s^2_{w_j}$ – межсерийная дисперсия доли в *j*-м районе, r_j – число серий, отобранных в *j*-м районе, *m* – число районов.

Если в рассмотренных выше случаях производится *бесповторный отбор*, то по причине, упоминавшейся выше, дисперсия корректируется множителем *(N – n) / (N – 1)* при отборе единицами или *(R – r) / (r – 1)* при отборе сериями.

Таким образом, средняя ошибка выборочной средней и выборочной доли для разных видов выборки будет иметь вид, представленный в приводимой ниже таблице.

Таблица 4.2

Вид выборки	Средняя ошибка	
	Выборочной средней	Выборочной относительной величины (доли)
1	*2*	*3*
Повторная, отбор единицами	$\sqrt{\frac{s^2}{n}}$	$\sqrt{\frac{w(1-w)}{n}}$
Бесповторная, отбор единицами	$\sqrt{\frac{s^2}{n}\cdot\frac{N-n}{N-1}}$	$\sqrt{\frac{w(w-1)}{n}\cdot\frac{N-n}{N-1}}$
Серийная (нерайонированная) бесповторная	$\sqrt{\frac{s_{\bar{x}}^2}{r}\cdot\frac{R-r}{R-1}}$	$\sqrt{\frac{s_w^2}{r}\cdot\frac{R-r}{R-1}}$
Районированная, отбор единицами, бесповторная	$\sqrt{\frac{\overline{s^2}}{n}\cdot\frac{N-n}{N-1}}$	$\sqrt{\frac{\overline{s_w^2}}{n}\cdot\frac{N-n}{N-1}}$
Районированная, отбор сериями, бесповторная	$\sqrt{\frac{\overline{s_{\bar{x}}^2}}{r}\cdot\frac{R-r}{R-1}}$	$\sqrt{\frac{\overline{s_w^2}}{r}\cdot\frac{R-r}{R-1}}$

4.8. Определение объема выборки

Одним из важных вопросов планирования выборочного наблюдения является вопрос об объеме выборки. Как мы уже убедились, для получения выборочных оценок, наиболее адекватно отражающих показатели генеральной совокупности, желательно получить гарантированно малую (с большой вероятностью) среднюю ошибку $s = \frac{\sigma}{\sqrt{n}}$. Отсюда видно, что любую степень точности можно получить, увеличивая *n*, что, однако, приводит к нежелательному увеличению затрат, зато малый объем порождает большие ошибки. К тому же эта формула не учитывает веро-

ятностной оценки. Поэтому при планировании объема выборки исходят из предельной ошибки *повторной* выборки

$$\Delta = t\frac{\sigma}{\sqrt{n}}.$$

Выражая отсюда n, получим

$$n = \frac{t^2\sigma^2}{\Delta^2}.$$

Напомним, что здесь σ^2 – дисперсия признака в генеральной совокупности, а Δ – предельная ошибка его. Выше в качестве признака мы рассматривали среднюю величину или долю.

Исходя из целей выборочного наблюдения, при планировании его *задают величину предельной ошибки* Δ и *доверительную вероятность*, которая определяет параметр t. Что же касается дисперсии, которая, как правило, заранее не известна, то на практике используют некоторые методы оценки ее величины. Вот некоторые из них.

1. Если оценивается средняя признака, имеющего нормальное распределение или близкое к нему, то можно воспользоваться «правилом трех сигм», согласно которому практически все значения признака лежат в интервале длиною 6σ. Отсюда $\sigma = R/6$, где $R = x_{max} - x_{min}$. А наибольшее и наименьшее значения признака оцениваются экспертами.

2. Если асимметрия распределения признака существенна, то полагают

$$\sigma = \frac{1}{5}(x_{max} - x_{min}).$$

3. Можно организовать «пробную» выборку небольшого объема k, по которой определяется дисперсия

$$\sigma^2_{проб.} = \frac{\sum_{i=1}^{k}(x_i - \bar{x}_{проб.})^2}{k-1},$$

используемая затем в качестве оценки генеральной дисперсии.

4. Если ранее проводились выборочные исследования аналогичной статистической совокупности, то полученную тогда дисперсию берут в качестве оценки дисперсии изучаемой совокупности.

5. Можно воспользоваться тем, что максимум дисперсии доли достигается при $w = 0{,}5$, и $max\ \sigma_w^2 = 0{,}5\cdot(1-0{,}5) = 0{,}25$.

В любом случае, в связи с тем, что генеральная дисперсия оценивается приближенно, объем выборки n рекомендуется округлять в сторону увеличения. При этом не стоит гнаться за малыми значениями Δ и большими значениями t, что, с одной стороны, приводит к большей точности и большей уверенности, но с другой, к увеличению затрат средств, труда и времени, на что можно не обращать внимание только в исключительных случаях.

Если планируется *бесповторный отбор,* то следует исходить из формулы:

$$\Delta = t\frac{s}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}},$$

где s^2 – оценка дисперсии генеральной совокупности. Отсюда получим скорректированное выражение объема выборки

$$n = \frac{n_0 N}{n_0 + (N-1)},$$

где $n_0 = t^2\frac{\sigma^2}{\Delta^2}$ – объем выборки при повторном отборе. Легко видеть, что при большом объеме генеральной совокупности скорректированный объем выборки незначительно отличается от n_0.

Если планируется районированная выборка, то объем выборки распределяют пропорционально объемам генеральной совокупности, приходящимся на районы:

$$n_i = n\cdot\frac{N_i}{N},$$

где n_i – объем выборки для i–го района в генеральной совокупности, n – общий объем выборки, N – общий объем генеральной совокупности.

При любом виде проектируемой выборки расчет объема начинают по формуле повторного отбора. Если в результате расчета доля отбора n_0/N превысит 5% генеральной совокупности, то проводят второй вариант расчета по формуле для бесповторного отбора, согласно которой объем выборки n будет меньше, чем n_0. Если же доля отбора меньше *пяти процентов*, то в таком переходе нет смысла, так как это несущественно скажется на величине объема выборки.

Выборка должна *обеспечить репрезентативность* основных показателей генеральной совокупности. Поэтому ее объем рассчитывают многократно, исходя из допустимых ошибок разных показателей. При

этом к практическому использованию выбирают *наибольшее из получившихся значений n*.

Пример. На электроламповом заводе для проверки взято *100* ламп. Средняя продолжительность горения их оказалась раной *1420 ч*. со среднеквадратическим отклонением *61, 03 ч*. Поскольку представляет интерес качество всей партии (*50 тыс*. ламп), то оценивают точность полученной средней $s_{\bar{x}} = \frac{61,03}{\sqrt{100}} = \pm 6,1$ *ч*. При доверительной вероятности *0,954* из равенства $\Phi(t) = \frac{0,954}{2} = 0,477$ определяем $t = 2$. После этого получим предел возможной ошибки

$$\Delta_{\bar{x}} = 2 \cdot (\pm 6,1) = \pm 12,2 \text{ ч}.$$

Следовательно, с вероятностью *0,954* можно утверждать, что средняя продолжительность горения одной электролампы во всей партии будет находиться в пределах от *1408 ч.* до *1432 ч.*, при этом только *46* ламп из каждой тысячи (*1000 (1 – 0,954)=46*) могут иметь срок горения, выходящий за эти пределы. Но поскольку представляют интерес отклонения в сторону меньшей продолжительности горения, то некачественными признают половину – *23* лампы из тысячи. На оновании этого решается вопрос о годности всей партии электроламп.

Решение вопроса можно уточнить. Определим, у какой доли ламп срок службы окажется меньше установленного лимита – *1410 ч.* Продукция с меньшим сроком горения считается некачественной.

При контрольной проверке *100* ламп меньше *1410 ч*. горели *10* ламп, доля которых в выборке составила *0,1* или *10*%. Тогда средняя ошибка этой доли будет

$$s_w = \sqrt{\frac{0,1 \cdot (1 - 0,1)}{100}} = \pm 0,03 \text{ или } \pm 3\%.$$

С вероятностью *0,954* определяем предельную ошибку доли $\Delta_w = 2 \cdot (\pm 0,03) = \pm 0,06$ или *6%.* Следовательно, с уверенностью *95,4%* можно ожидать, что во всей партии окажется от четырех до шестнадцати процентов некачественных электроламп (10% ± 6%),

4.9. Малые выборки

Напомним, что результаты предыдущего пункта были получены в предположении большого объема выборки *n*, что позволяет пользоваться нормальным распределением при отыскании параметра *t* и предельных ошибок. Безусловно большими являются *n > 100*, но уже при

$n < 100$ распределения выборочных средней и доли ощутимо отличаются от нормального и его использование приводит к заметным погрешностям. При объеме выборки от *30* до *100* эти погрешности считаются незначительными, и поэтому параметр t можно находить, используя как нормальное распределение, так и распределение Стьюдента.

При объемах выборки, меньших *30*, разница становится существенной. Такие выборки называют *малыми*. Для них *обязательно* исполнять два правила:

1. параметр t находить по таблице *t*-распределения Стьюдента;

2. использовать *только* «исправленную» дисперсию.

Использование малых выборок зачастую продиктовано необходимостью. Например, стендовые испытания различных конструкций на предельные нагрузки ведутся до необратимого разрушения их. А это приводит к большим затратам. В других случаях причиной являются «технологические» условия. Так, в селекционной работе «чистый» опыт легче поставить на небольшом числе делянок.

П р и м е р. Для изучения использования рабочего времени проведено наблюдение за *10* отобранными рабочими, из которых *4* человека работали все время. Доля их равна *0,4*, дисперсия $0{,}4 \cdot (1 - 0.4) = 0{,}24$. При доверительной вероятности *0,95* и числе степеней свободы $k = 10 - 1 = 9$ по таблице распределения Стьюдента находим $t = 2{,}66$. Тогда предельная ошибка будет равна $\Delta w = 2{,}26 \cdot \sqrt{\frac{0{,}24}{9}} = \pm 0{,}36$.

Таким образом, с вероятностью *0,95* доля чистого рабочего времени в фонде времени рабочих данного цеха находится в пределах от *39,64*% до *40,36*%.

Если при той же вероятности использовать $t = 1{,}96$, найденное из нормального распределения, то предельная ошибка получится равной *0,31*. С одной стороны, меньшая ошибка – хорошо, а с другой – это неправда.

Обобщая сказанное, приведем схему применимости нормального распределения и распределения Стьюдента для отыскания параметра t:

Таблица 4.3

Объем выборки n	Из нормального распределения $\Phi(t) = \frac{\gamma}{2}$	Из распределения Стьюдента $t_\gamma = t(\gamma, k)$
Менее *30*	Нет	Обязательно
От *30* до *100*	Можно	Предпочтительнее
Более *100*	Да	Да

И в завершение напомним, что *распределение Стьюдента рассматривается только для нормально распределенного исходного признака X .*

Глава 5

Основы дисперсионного анализа

5.1. Теоретические основы

Основной областью применения статистического метода дисперсионного анализа является ситуация, когда требуется решить вопрос о влиянии одного или нескольких качественных факторов на количественный признак *Y*, оценить их влияние и выбрать наиболее важный фактор. Если исследуется влияние одного фактора, то говорят об *однофакторном дисперсионном анализе* или об *однофакторной классификации*, если факторов два, то речь идет о *двухфакторном дисперсионном анализе* или о *двухфакторной классификации* и т. д.

Мы рассмотрим однофакторный комплекс. Вначале приведем некоторые примеры. Если речь идет о влиянии различных видов удобрений на урожайность некоторой сельскохозяйственной культуры, то фактором *F* является удобрение, а количественным признаком *Y* – урожайность. Другая ситуация: одна и та же деталь изготавливается на разных по типу станках или разными рабочими. Тогда в качестве факторов можно рассматривать типы станков или рабочих, а количественным признаком может быть число бракованных деталей, затраты времени на изготовление одной детали или отклонение от заданных размеров и т. д. Во всех случаях для фактора *F* выделяют *уровни фактора* F_1 , F_2 , … ,F_m. В приведенных выше примерах уровнями фактора являются различные виды удобрений, различные станки или разные рабочие.

На каждом уровне F_j $\left(j = 1 \div m\right)$ организуют выборочное наблюдение признака *Y* объемов n_j и для каждой из *m* получившихся групп данных y_{ij} $\left(i = 1 \div n_j ; j = 1 \div m\right)$, то есть на каждом уровне, вычисляют средние величины $\overline{y}_j$, так называемые групповые средние.

Основной постулат. Если рассматриваемый фактор не влияет на признак Y, то он не влияет и на средние величины, то есть они должны быть равны или незначительно отличаться друг от друга. Если же фактор влияет на Y, то различие между средними величинами должно быть существенным.

Поскольку выборка случайна, то, как правило, средние значения отличаются друг от друга. Возникает необходимость выяснить, случайно или закономерно расхождение между ними. Для этого проверяют гипо-

тезу о равенстве средних величин. В качестве нулевой гипотезы рассматривают

$$H_0 : \bar{y}_1 = \bar{y}_2 = ... = \bar{y}_m,$$

а в качестве альтернативной

$$H_1 : \bar{y}_1 \neq \bar{y}_2 \neq ... \neq \bar{y}_m.$$

Гипотеза H_0 проверяется сравнением межгрупповой и внутригрупповой дисперсий по F- критерию, предложенному английским статистиком Рональдом Фишером (1890 – 1968). Если эти дисперсии отличаются незначительно или межгрупповая дисперсия меньше внутригрупповой, то нулевая гипотеза принимается, и влияние фактора исключается. Если же межгрупповая дисперсия существенно больше внутригрупповой, то заключают, что различие в средних обусловлено не только случайностями выборок, но и воздействием исследуемого фактора. Кстати, использование дисперсий для проверки гипотезы о равенстве средних определило название рассматриваемого метода.

Аналогично тому, как это делалось в главе 3, общую сумму квадратов отклонений вариант y_{ij} от общей средней $\bar{y}$ можно представить в виде суммы двух слагаемых

$$\sum_{j=1}^{m}\sum_{i=1}^{n_j}\left(y_{ij} - \bar{y}\right)^2 = \sum_{j=1}^{m}\left(\bar{y}_j - \bar{y}\right)^2 n_j + \sum_{j=1}^{m}\sum_{i=1}^{n_j}\left(y_{ij} - \bar{y}_j\right)^2$$

или при соответствующих обозначениях

$$D_{общ} = D_{факт} + D_{ост}.$$

Величина $D_{общ}$ измеряет общую вариацию признака Y. Слагаемое $D_{факт}$ отражает ту часть общей вариации, которая обусловлена влиянием фактора F, а $D_{ост}$ представляет собой вариацию за счет влияния прочих факторов. Указанные выше суммы квадратов отклонений позволяют получить оценки трех дисперсий, используя, согласно Р. Фишеру, соответствующие числа степеней свободы, которые равны:

$d.f._{общ} = n - 1$, – для общей вариации $(n = \sum_{j=1}^{m} n_j)$;

$d.f._{факт} = m - 1$ – для межгрупповой вариации ;

$$d.f._{ост} = \sum_{j=1}^{m}(n_j - 1) = \sum_{j=1}^{m} n_j - \sum_{j=1}^{m} 1 = n - m \ -$$ для внутригрупповой.

Заметим, как и суммы квадратов отклонений, числа степеней свободы связаны аналогичным равенством:

$$d.f._{общ} = d.f._{факт} + d.f._{ост}.$$

Выборочные оценки трех дисперсий находят в виде:

$$s^2_{общ} = \frac{D_{общ}}{n-1} = \frac{1}{n-1}\sum_{j=1}^{m}\sum_{i=1}^{n_j}(y_{ij} - \bar{y})^2 ,$$

$$s^2_{факт} = \frac{D_{факт}}{m-1} = \frac{1}{m-1}\sum_{j=1}^{m}(\bar{y}_j - \bar{y})^2 n_j ,$$

$$s^2_{ост} = \frac{D_{ост}}{n-m} = \frac{1}{n-m}\sum_{j=1}^{m}\sum_{i=1}^{n_j}(y_{ij} - \bar{y}_j)^2 .$$

По ним вычисляется наблюдаемое значение статистики, называемой *F*-критерием Фишера:

$$F_{набл} = \frac{s^2_{факт}}{s^2_{ост}},$$

при этом предполагается, что $s^2_{факт} > s^2_{ост}$. В противном случае, как мы уже говорили, влияние изучаемого фактора исключается. Имеются две статистические таблицы критических значений *F*-критерия Фишера $F_{крит} = F(\alpha, d.f._1, d.f._2)$ при уровнях значимости $\alpha = 0{,}05$ и $\alpha = 0{,}01$. Здесь $d.f._1$ и $d.f._2$ – числа степеней свободы факторной и остаточной дисперсий соответственно.

Если окажется, что $F_{набл} > F_{крит}$, влияние фактора является существенным или статистически значимым. Если же $F_{набл} \le F_{крит}$, следует поставить под сомнение вопрос о влиянии фактора на количественный признак.

Остается заметить, что рассмотренный метод верен для случая, когда признак *Y* на каждом уровне распределен нормально с одинаковыми дисперсиями.

5.2. Пример применения дисперсионного анализа

В трех соседних колхозах (уровни фактора F_1 , F_2 , F_3) были собраны данные об урожайности пшеницы за четыре последних года . Результаты представлены в таблице *5.1* на следующей странице. Методом дисперсионного анализа при уровне значимости *0,05* требуется проверить нулевую гипотезу о равенстве средних урожайностей и тем самым проверить влияние фактора колхоза на урожайность. Предполагается, что урожайность в каждом колхозе имеет нормальное распределение с одинаковыми дисперсиями.

Р е ш е н и е. В нашем случае число уровней $m = 3$, число наблюдений n_j на каждом уровне одинаково и равно *4*. Отсюда $n = 4 \cdot 3 = 12$. Групповые средние $\bar{y}_j$ – в последней строке таблицы.

Так как число наблюдений на каждом уровне одинаково, то общую среднюю можно вычислить как простую среднюю арифметическую групповых средних $\bar{y} = \dfrac{54 + 55 + 47}{3} = 52$. Теперь можно вычислить суммы

Таблица 5.1

Годы i	Уровни фактора F_j		
	F_1	F_2	F_3
1	51	52	42
2	52	54	44
3	56	56	50
4	57	58	52
$\bar{y}_j$	54	55	47

квадратов отклонений:

$D_{факт} = 4\left((54 - 52)^2 + (55 - 52)^2 + (47 - 52)^2\right) = 152$.

Предварительно найдем по уровням суммы квадратов отклонений от групповых средних:

$D_{1ост} = (51 - 54)^2 + (52 - 54)^2 + (56 - 54)^2 + (57 - 54)^2 = 26$,

$D_{2ост} = (52 - 55)^2 + (54 - 55)^2 + (56 - 55)^2 + (58 - 55)^2 = 20$,

$D_{3ост} = (42 - 47)^2 + (44 - 47)^2 + (50 - 47)^2 + (52 - 47)^2 = 68$.

Тогда $D_{ост} = 26 + 20 + 68 = 114$. Найдем теперь факторную и остаточную дисперсии

$$s^2_{факт} = \frac{D_{факт}}{m - 1} = \frac{152}{3 - 1} = 76,$$

$$s^2_{ост} = \frac{D_{ост}}{n-m} = \frac{114}{12-3} = 12{,}67 .$$

Отсюда получим наблюдаемое значение *F*-критерия:

$$F_{набл} = \frac{s^2_{факт}}{s^2_{ост}} = \frac{76}{12{,}67} = 6{,}00 .$$

Учитывая, что уровень значимости $\alpha = 0{,}05$ и числа степеней свободы $d.f._1 = 3-1 = 2$, $d.f._2 = 12-3 = 9$, по таблице находим критическое значение $F = F(0{,}05; 2; 9) = 4{,}26$.

Так как $F_{набл} > F_{крит}$, то нулевую гипотезу о равенстве средних отвергаем, то есть они различаются значимо. Таким образом, урожайность пшеницы в группе рассмотренных колхозов существенным образом зависит от фактора – колхоз.

5.3. Случай количественного фактора

До сих пор мы говорили о качественном (описательном) факторе и о его влиянии на количественный признак *Y*. Однако дисперсионный анализ можно применить и в случае количественного фактора – признака *X*. В качестве уровней фактора в этом случае можно принять группировочные интервалы, рассмотренные нами в главе 3 в пункте «Аналитическая группировка».

Далее, используя сгруппированные данные y_{ij} признака-следствия *Y*, как и в предыдущем пункте находят наблюдаемое значение *F*-критерия и, сравнивая его с критическим значением, делают соответствующие выводы.

Мы же покажем, как использовать результаты аналитической группировки для получения того же результата. С этой целью заметим, что из рассмотренных там дисперсий вытекают равенства: $D_{общ} = n \cdot S^2_y$, $D_{факт} = n \cdot S^2_{\overline{yx}}$ и $D_{ост} = n \cdot \overline{S^2_{yx}}$. Отсюда очевидно вытекает равенство $\eta^2 = \frac{D_{факт}}{D_{общ}} = 0{,}88^2 = 0{,}774$. Тогда, учитывая результаты пункта *5.1*, с помощью несложных выкладок получим: $\frac{D_{факт} + D_{ост}}{D_{факт}} = \frac{1}{0{,}774} = 1{,}292$ или $1 + \frac{D_{ост}}{D_{факт}} = 1{,}292$, откуда $\frac{D_{факт}}{D_{ост}} = \frac{1}{0{,}292} = 3{,}425$.

Учитывая, что объем всей совокупности $n = 20$, а число групп (уровней фактора) $m = 3$, вычислим числа степеней свободы. Получим $d.f._1 = 3 - 1 = 2$ и $d.f._2 = 20 - 3 = 17$ для межгрупповой и остаточной дисперсий соответственно. Тогда наблюдаемое значение F-критерия:

$$F_{набл} = \frac{D_{факт}}{2} : \frac{D_{ост}}{17} = \frac{17}{2} \cdot \frac{D_{факт}}{D_{ост}} = 8{,}5 \cdot 3.425 = 29{,}11 .$$

По таблице определяем $F_{крит} = F(0{,}05; 2; 17) = 3{,}59$.

Так как $F_{набл} > F_{крит}$, делаем вывод о том, что влияние признака-фактора X на признак-следствие Y статистически существенно, так гипотеза о равенстве средних отвергается. То есть скорость оборота средств является важным фактором формирования прибыли. На это же указывало достаточно близкое к единице значение эмпирического корреляционного отношения $\eta = 0.88$.

Кстати заметим, что в статистике принято считать, что подтверждение вывода различными методами делает этот вывод более достоверным.

ПРИЛОЖЕНИЯ

Статистическо-математические таблицы

Приложение1.

Значения функции $\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$

x	Сотые доли									
	0	1	2	3	4	5	6	7	8	9
0,0	0,3989	0,3989	0.3989	0,3988	0,3986	0,3984	0,3982	0,3980	0,3977	0,3973
0,1	3970	3965	3961	3956	3951	3945	3939	3932	3925	3918
0,2	3910	3902	3894	3885	3876	3867	3857	3847	3836	3825
0,3	3814	3802	3790	3778	3765	3752	3739	3726	3712	3697
0,4	3683	3668	3653	3637	3621	3605	3589	3572	3555	3538
0,5	3521	3503	3485	3467	3448	3429	3410	3391	3372	3352
0,6	3332	3312	3292	3271	3251	3230	3209	3187	3166	3144
0,7	3123	3101	3079	3056	3034	3011	2989	2966	2943	2920
0,8	2897	2874	2850	2827	2803	2780	2756	2732	2709	2685
0,9	2661	2637	2613	2589	2565	2541	2516	2492	2468	2444
1,0	0,2420	0,2396	0,2371	0,2347	0,2323	0,2299	0,2275	0,2251	0,2227	0,2203
1,1	2179	2155	2331	2107	2083	2059	2036	2012	1989	1965
1,2	1942	1919	1895	1872	1849	1826	1804	1781	1758	1736
1,3	1714	1691	1669	1647	1626	1604	1582	1561	1539	1518
1,4	1497	1476	1456	1435	1415	1394	1374	1354	1334	1315
1,5	1295	1276	1257	1238	1219	1200	1182	1163	1145	1127
1,6	1109	1092	1074	1057	1040	1023	1006	0989	0973	0957
1,7	0940	0925	0909	0893	0878	0863	0848	0833	0818	0804
1,8	0790	0775	0761	0748	0734	0721	0707	0694	0681	0669
1,9	0656	0644	0632	0620	0608	0596	0584	0573	0562	0551
2,0	0,0540	0,0529	0,0519	0,0508	0,0498	0,0488	0,0478	0,0468	0,0459	0,0449
2,1	0440	0431	0422	0413	0404	0396	0387	0379	0371	0363
2,2	0355	0347	0339	0332	0325	0317	0310	0303	0297	0290
2,3	0283	0277	0270	0264	0258	0252	0246	0241	0235	0229
2,4	0224	0219	0213	0208	0203	0198	0194	0189	0184	0180
2,5	0175	0171	0167	0163	0158	0154	0151	0147	0143	0139
2,6	0136	0132	0129	0126	0122	0119	0116	0113	0110	0107
2,7	0104	0101	0099	0096	0093	0091	0088	0086	0084	0081
2,8	0079	0077	0075	0073	0071	0069	0067	0065	0063	0061
2,9	0060	0058	0056	0055	0053	0051	0050	0048	0047	0046
3,0	0,0044	0,0043	0,0042	0,0040	0,0039	0,0038	0,0037	0,0036	0,0035	0034
3,1	0033	0032	0031	0030	0029	0028	0027	0026	0025	0025
3,2	0024	0023	0022	0022	0021	0020	0020	0019	0018	0018
3,3	0017	0017	0016	0016	0015	0015	0014	0014	0013	0013
3,4	0012	0012	0012	0011	0011	0010	0010	0010	0009	0009
3,5	0009	0008	0008	0008	0008	0007	0007	0007	0007	0006
3,6	0006	0006	0006	0005	0005	0005	0005	0005	0005	0004
3,7	0004	0004	0004	0004	0004	0004	0003	0003	0003	0003
3,8	0003	0003	0003	0003	0003	0002	0002	0002	0002	0002
3,9	0002	0002	0002	0002	0002	0002	0002	0001	0001	0001

При $x \geq 4$ принимают $\varphi(x) = 0$.

Приложение 2.

Значения функции $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_0^x e^{-\frac{t^2}{2}} dt$

x	Φ (x)	x	Φ (x)	x	Φ (x)	**x**	Φ (x)	x	Φ(x)	x	Φ(x)
0.00	0.0000	0.45	0.1736	0.90	0.3159	1.35	0.4115	1.80	0.4641	2.50	0.4938
0.01	0.0040	0.46	0.1772	0.91	0.3186	1.36	0.4131	1.81	0.4649	2.52	0.4941
0.02	0.0080	0.47	0.1808	0.92	0.3212	1.37	0.4147	1.82	0.4656	2.54	0.4945
0.03	0.0120	0.48	0.1844	0.93	0.3238	1.38	0.4162	1.83	0.4664	2.56	0.4948
0.04	0.0160	0.49	0.1879	0.94	0.3264	1.39	0.4177	1.84	0.4671	2.58	0.4951
0.05	0.0199	0.50	0.1915	0.95	0.3289	1.40	0.4192	1.85	0.4678	2.60	0.4953
0.06	0.0239	0.51	0.1950	0.96	0.3315	1.41	0.4207	1.86	0.4686	2.62	0.4956
0.07	0.0279	0.52	0.1985	0.97	0.3340	1.42	0.4222	1.87	0.4693	2.64	0.4959
0.08	0.0319	0.53	0.2019	0.98	0.3365	1.43	0.4236	1.88	0.4699	2.66	0.4961
0.09	0.0359	0.54	0.2054	0.99	0.3389	1.44	0.4251	1.89	0.4706	2.68	0.4963
0.10	0.0398	0.55	0.2088	1.00	0.3413	1.45	0.4265	1.90	0.4713	2.70	0.4965
0.11	0.0438	0.56	0.2123	1.01	0.3438	1.46	0.4279	1.91	0.4719	2.72	0.4967
0.12	0.0478	0.57	0.2157	1.02	0.3461	1.47	0.4292	1.92	0.4726	2.74	0.4969
0.13	0.0517	0.58	0.2190	1.03	0.3485	1.48	0.4306	1.93	0.4732	2.76	0.4971
0.14	0.0557	0.59	0.2224	1.04	0.3508	1.49	0.4319	1.94	0.4738	2.78	0.4973
0.15	0.0596	0.60	0.2257	1.05	0.3531	1.50	0.4332	1.95	0.4744	2.80	0.4974
0.16	0.0636	0.61	0.2291	1.06	0.3554	1.51	0.4345	1.96	0.4750	2.82	0.4976
0.17	0.0675	0.62	0.2324	1.07	0.3577	1.52	0.4357	1.97	0.4756	2.84	0.4977
0.18	0.0714	0.63	0.2357	1.08	0.3599	1.53	0.4370	1.98	0.4761	2.86	0.4979
0.19	0.0753	0.64	0.2389	1.09	0.3621	1.54	0.4382	1.99	0.4767	2.88	0.4980
0.20	0.0793	0.65	0.2422	1.10	0.3643	1.55	0.4394	2.00	0.4772	2.90	0.4981
0.21	0.0832	0.66	0.2454	1.11	0.3665	1.56	0.4406	2.02	0.4783	2.92	0.4982
0.22	0.0871	0.67	0.2486	1.12	0.3686	1.57	0.4418	2.04	0.4793	2.94	0.4984
0.23	0.0910	0.68	0.2517	1.13	0.3708	1.58	0.4429	2.06	0.4803	2.96	0.4985
0.24	0.0948	0.69	0.2549	1.14	0.3729	1.59	0.4441	2.08	0.4812	2.98	0.4986
0.25	0.0987	0.70	0.2580	1.15	0.3749	1.60	0.4452	2.10	0.4821	3.00	04987
0.26	0.1026	0.71	0.2611	1.16	0.3770	1.61	0.4463	2.12	0.4830	3.20	0.4993
0.27	0.1064	0.72	0.2642	1.17	0.3790	1.62	0.4474	2.14	0.4838	3.40	0.4997
0.28	0.1103	0.73	0.2673	1.18	0.3810	1.63	0.4484	2.16	0.4846	3.60	0.4998
0.29	0.1141	0.74	0.2703	1.19	0.3830	1.64	0.4495	2.18	0.4854	3.80	0.4999
0.30	0.1179	0.75	0.2734	1.20	0.3849	1.65	0.4505	2.20	0.4861	4.00	0.4999
0.31	0.1217	0.76	0.2764	1.21	0.3869	1.66	0.4515	2.22	0.4868	4.50	0.5000
0.32	0.1255	0.77	0.2794	1.22	0.3883	1.67	0.4525	2.24	0.4875	5.00	0.5000
0.33	0.1293	0.78	0.2823	1.23	0.3907	1.68	0.4535	2.26	0.4881		
0.34	0.1331	0.79	0.2852	1.24	0.3925	1.69	0.4545	2.28	0.4887	↓	↓
0.35	0.1368	0.80	0.2881	1.25	0.3944	1.70	0.4554	2.30	0.4893	$+\infty$	0.5
0.36	0.1406	0.81	0.2910	1.26	0.3962	1.71	0.4564	2.32	0.4898		
0.37	0.1443	0.82	0.2939	1.27	0.3980	1.72	0.4573	2.34	0.4904		
0.38	0.1480	0.83	0.2967	1.28	0.3997	1.73	0.4582	2.36	0.4909		
0.39	0.1517	0.84	0.2995	1.29	0.4015	1.74	0.4591	2.38	0.4913		
0.40	0.1554	0.85	0.3023	1.30	0.4032	1.75	0.4599	2.40	0.4918		
0.41	0.1591	0.86	0.3051	1.31	0.4049	1.76	0.4608	2.42	0.4922		
0.42	0.1628	0.87	0.3078	1.32	0.4066	1.77	0.4616	2.44	0.4927		
0.43	0.1654	0.88	0.3106	1.33	0.4082	1.78	0.4625	2.46	0.4931		
0.44	0.1700	0.89	0.3133	1.34	0.4099	1.79	0.4633	2.48	0.4934		

Приложение 3.

Распределение Стьюдента (двусторонняя критическая область),

α – уровень значимости, $\gamma = 1 - \alpha$ – доверительная вероятность, v – число степеней свободы, $n = v + 1$ – объем выборки

$v \downarrow$ \ α / γ	0,10 0,90	0,05 0,95	0,02 0,98	0,01 0,99	0,002 0,998	0,001 0,999
1	6,314	12,71	31,82	63,66	318,3	636,6
2	2,920	4,303	6,965	9,925	22,33	31,60
3	2,353	3,182	4,541	5,841	10,22	12,94
4	2,132	2,776	3,747	4,604	7,173	8,610
5	2,015	2,571	3,365	5,032	5,893	6,859
6	1,943	2,447	3,143	3,707	5,208	5,959
7	1,895	2,365	2,998	3,499	4,785	5,405
8	1,860	2,306	2,896	3,355	4,501	5,041
9	1,833	2,262	2,821	3,250	4,297	4,781
10	1,812	2,228	2,764	3,169	4,144	4,587
11	1,796	2,201	2,718	3,106	4,025	4,437
12	1,782	2,179	2,681	3,055	3,930	4,318
13	1,771	2,160	2,650	3,012	3,852	4,221
14	1,761	2,145	2,624	2,977	3,787	4,140
15	1,753	2,131	2,602	2,947	3,733	4,073
16	1,746	2,120	2,583	2,921	3,686	4,015
17	1,740	2,110	2,567	2,898	3,646	3,965
18	1,734	2,101	2,552	2,878	3,611	3,922
19	1,729	2,093	2,539	2,861	3,579	3,883
20	1,725	2,086	2,528	2,845	3,562	3,850
21	1,721	2,080	2,518	2,831	3,527	3,819
22	1,717	2,074	2,508	2,819	3,505	3,792
23	1,714	2,069	2,500	2,807	3,485	3,767
24	1,711	2,064	2,492	2,797	3,467	3,745
25	1,708	2,060	2,485	2,787	3,450	3,725
26	1,706	2,056	2,479	2,779	3,435	3,707
27	1,703	2,052	2,473	2,771	3,421	3,690
28	1,701	2,048	2,467	2,763	3,408	3,674
29	1,699	2,045	2,462	2,756	3,396	3,659
30	1,697	2,042	2,457	2,750	3,385	3,646
40	1,684	2,021	2,423	2,704	3,307	3,551
50	1,676	2,009	2,403	2,678	3,262	3,495
60	1,671	2,000	2,390	2,660	3,232	3,460
80	1,664	1,990	2,374	2,639	3,195	3,415
100	1,660	1,984	2,365	2,626	3,174	3,389
200	1,653	1,972	2,345	2,601	3,131	3,339
300	1,648	1,965	2,334	2,586	3,106	3,310
∞	1,645	1,960	2,326	2,576	3,090	3,291

Приложение 4.

F **-** распределение (Фишера – Снедекора) при уровне значимости 0,01

k_2 – степени свободы для меньшей дисперсии	k_2 ↓	k_1 – степени свободы для большей дисперсии											
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
	1	4 052	4 999	5 403	5 625	5 764	5 889	5 928	5 981	6 022	6 056	6 082	6 106
	2	98,49	99,01	99,17	99,25	99,30	99,33	99,34	99,36	99,38	99,40	99,41	99,42
	3	34,12	30,81	29,46	28,71	28,24	27,91	27,67	27,49	27,34	27,23	27,13	27,05
	4	21,20	18,00	16,69	15,98	15,52	15,21	14,98	14,80	14,66	14,54	14,45	14,37
	5	16,26	13,27	12,06	11,39	10,97	10,67	10,45	10,27	10,15	10,05	9,96	9,89
	6	13,74	10,92	9,78	9,15	8,75	8,47	8,26	8,10	7,98	7,87	7,79	7,72
	7	12,25	9,55	8,45	7,85	7,46	7,19	7,00	6,84	6,71	6,62	6,54	6,47
	8	11,26	8,65	7,59	7,01	6,63	6,37	6,19	6,03	5,91	5,82	5,74	5,67
	9	10,56	8,02	6,99	6,42	6,06	5,80	5,62	5,47	5,35	5,26	5,18	5,11
	10	10,04	7,56	6,55	5,99	5,64	5,39	5,21	5,06	4,95	4,85	4,78	4,71
	11	9,85	7,20	6,22	5,67	5,32	5,07	4,88	4,74	4,63	4,54	4,46	4,40
	12	9,33	6,93	5,95	5,41	5,06	4,82	4,65	4,50	4,39	4,30	4,22	4,16
	13	9,07	6,70	5,74	5,20	4,86	4,62	4,44	4,30	4,19	4,10	4,02	3,96
	14	8,86	6,51	5,56	5,03	4,69	4,46	4,28	4,14	4,03	3,94	3,86	3,80
	15	8,68	6,36	5,42	4,89	4,56	4,32	4,14	4,00	3,89	3,80	3,73	3,67
	16	8,53	6,23	5,29	4,77	4,44	4,20	4,03	3,89	3,78	3,69	3,61	3,55
	17	8,40	6,11	5,18	4,67	4,34	4,10	3,93	3,79	3,68	3,59	3,52	3,45
	20	8,10	5,85	4,94	4,43	4,10	3,87	3,71	3,56	3,45	3,37	3,30	3,23
	24	7,82	5,61	4,72	4,22	3,90	3,67	3,50	3,36	3,25	3,17	3,09	3,03
	30	7,56	5,39	4,51	4,02	3,70	3,47	3,30	3,17	3,06	2,98	2,90	2,84
		14	**16**	**20**	**24**	**30**	**40**	**50**	**75**	**100**	**200**	**500**	**∞**
	1	6 142	6 169	6 208	6 234	6 258	6 286	6 302	6 323	6 334	6 352	6 361	6 366
	2	99,43	99,44	99,45	99,46	99,47	99,48	99,48	99,49	99,49	99,49	99,50	99,50
	3	26,92	26,83	26,69	26,60	26,50	26,41	26,35	26,27	26,23	26,18	26,14	26,12
	4	14,24	14,15	14,02	13,93	1383	13,74	13,69	13,61	13,57	13,52	13,48	13,46
	5	9,77	9,68	9,55	9,47	9,38	9,29	9,24	9,17	9,13	9,07	9,04	9,02
	6	7,60	7,52	7,39	7,31	7,23	7,14	7,09	7,02	6,99	6,94	6,90	6,88
	7	6,35	6,27	6,15	6,07	5,98	5,90	5,85	5,78	5,75	5,70	5,67	5,65
	8	5,56	5,48	5,36	5,28	5,20	5,11	5,06	5,00	4,96	4,91	4,88	4,86
	9	5,00	4,92	4,80	4,73	4,64	4,56	4,51	4,45	4,41	4,36	4,33	4,31
	10	4,60	4,52	4,41	4,33	4,25	4,17	4,12	4,05	4,01	3,96	3,93	3,91
	11	4,29	4,21	4,10	4,02	3,94	3,86	3,80	3,74	3,70	3,66	3,62	3,60
	12	4,05	3,98	3,86	3,78	3,70	3,61	3,56	3,49	3,46	3,41	3,38	3,36
	13	3,85	3,78	3,67	3,59	3,51	3,42	3,37	3,30	3,27	3,21	3,18	3,16
	14	3,70	3,62	3,51	3,43	3,34	3,26	3,21	3,14	3,11	3,06	3,02	3,00
	15	3,56	3,48	3,36	3,29	3,20	3,12	3,07	3,00	2,97	2,92	2,89	2,87
	16	3,45	3,37	3,25	3,18	3,10	3,01	2,96	2,89	2,86	2,80	2,77	2,75
	17	3,35	3,27	3,16	3,08	3,00	2,92	2,86	2,79	2,76	2,70	2,67	2,65
	20	3,13	3,05	2,94	2,86	2,77	2,69	2,63	2,56	2,53	2,47	2,44	2,42
	24	2,93	2,85	2,74	2,66	2,58	2,49	2,44	2,36	2,33	2,27	2,23	2,21
	30	2,74	2,66	2,55	2,47	2,38	2,29	2,24	2,16	2,13	2,07	2,03	2,01

Приложение 5.

F - распределение (Фишера – Снедекора) при уровне значимости 0,05.

k_2 ↓	k_1 – степени свободы для большей дисперсии											
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
1	161	200	216	225	230	234	237	239	241	242	243	244
2	18,51	19,00	19,16	19,25	19,30	19,33	19,36	19,37	19,38	19,39	19,40	19,41
3	10,13	9,55	9,28	9,12	9,01	8,94	8,88	8,84	8,81	8,78	8,76	8,74
4	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6,00	5,96	5,93	5,91
5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,78	4,74	4,70	4,68
6	5,99	5,14	4,76	4,35	4,39	4,28	4,21	4,15	4,10	4,06	4,03	4,00
7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,63	3,60	3,57
8	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39	3,34	3,31	3,28
9	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,13	3,10	3,07
10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,97	2,94	2,91
11	4,84	3,98	3,59	3,36	3,20	3,09	3,01	2,95	2,90	2,86	2,82	2,79
12	4,75	3,88	3,49	3,26	3,11	3,00	2,92	2,85	2,80	2,76	2,72	2,69
13	4,67	3,80	3,41	3,18	3,02	2,92	2,84	2,77	2,72	2,67	2,63	2,60
14	4,60	3,74	3,34	3,11	2,96	2,85	2,77	2,70	2,65	2,60	2,56	2,53
15	4,54	3,68	3,29	3,06	2,90	2,79	2,70	2,64	2,59	2,55	2,51	2,48
16	4,49	3,63	3,24	3,01	2,85	2,74	2,66	2,59	2,54	2,49	2,45	2,42
17	4,45	3,59	3,20	2,96	2,81	2,70	2,62	2,55	2,50	2,45	2,41	2,38
20	4,35	3,49	3,10	2,87	2,71	2,60	2,52	2,45	2,40	2,35	2,31	2,28
24	4,26	3,40	3,01	2,78	2,62	2,51	2,43	2,36	2,30	2,26	2,22	2,18
30	4,17	3,32	2,92	2,69	2,53	2,42	2,34	2,27	2,21	2,16	2,12	2,09
	14	**16**	**20**	**24**	**30**	**40**	**50**	**75**	**100**	**200**	**500**	**∞**
1	245	246	248	249	250	251	252	253	253	254	254	254
2	19,42	19,43	19,44	19,45	19,46	19,47	19,47	19,48	19,49	19,49	19,50	19,50
3	8,71	8,69	8,66	8,64	8,62	8,60	8,58	8,57	8,56	8,54	8,54	8,53
4	5,87	5,84	5,80	5,77	5,74	5,71	5,70	5,68	5,66	5,65	5,64	5,63
5	4,64	4,60	4,56	4,53	4,50	4,46	4,44	4,42	4,40	4,38	4,37	4,36
6	3,96	3,92	3,87	3,84	3,81	3,77	3,75	3,72	3,71	3,69	3,68	3,67
7	3,52	3,49	3,44	3,41	3,38	3,34	3,32	3,29	3,28	3,25	3,24	3,23
8	3,23	3,20	3,15	3,12	3,08	3,05	3,03	3,00	2,98	2,96	2,94	2,93
9	3,02	2,98	2,93	2,90	2,86	2,82	2,80	2,77	2,76	2,73	2,72	2,71
10	2,86	2,82	2,77	2,74	2,70	2,67	2,64	2,61	2,59	2,56	2,55	2,54
11	2,74	2,70	2,65	2,61	2,57	2,53	2,50	2,47	2,45	2,42	2,41	2,40
12	2,64	2,60	2,54	2,50	2,46	2,42	2,40	2,36	2,35	2,32	2,31	2,30
13	2,55	2,51	2,46	2,42	2,38	2,34	2,32	2,28	2,26	2,24	2,22	2,21
14	2,48	2,44	2,39	2,35	2,31	2,27	2,24	2,21	2,19	2,16	2,14	2,13
15	2,43	2,39	2,33	2,29	2,25	2,21	2,18	2,15	2,12	2,10	2,08	2,07
16	2,37	2,33	2,28	2,24	2,20	2,16	2,13	2,09	2,07	2,04	2,02	2,01
17	2,33	2,29	2,23	2,19	2,15	2,11	2,08	2,04	2,02	1,99	1,97	1,96
20	2,23	2,18	2,12	2,08	2,04	1,99	1,96	1,92	1,90	1,87	1,85	1,84
24	2,13	2,09	2,02	1,98	1,94	1,89	1,86	1,82	1,80	1,76	1,74	1,73
30	2,04	1,99	1,93	1,89	1,84	1,79	1,76	1,72	1,69	1,66	1,64	1,62

k_2 – степени свободы для меньшей дисперсии

Приложение 6.

χ^2 – распределение

ν - число степеней свободы, α - уровень значимости

ν \ α	0,20	0,10	0,05	0,02	0,01	0,001
1	1,642	2,706	3,841	5,412	6,635	10,827
2	3,219	4,605	5,991	7,824	9,210	13,815
3	4,642	6,251	7,815	9,837	11,345	16,266
4	5,989	7,779	9,488	11,668	13,237	18,467
5	7,289	9,236	11,070	13,388	15,086	20,515
6	8,558	10,645	12,592	15,033	16,812	22,457
7	9,803	12,017	14,067	16,622	18,475	24,322
8	11,030	13,362	15,507	18,168	20,090	26,125
9	12,242	14,684	16,919	19,679	21,666	27,877
10	13,442	15,987	18,307	21,161	23,209	29,588
11	14,631	17,275	19,675	22,618	24,795	31,264
12	15,812	18,549	21,026	24,054	24,217	32,909
13	16,985	19,812	22,362	25,472	27,688	34,528
14	18,151	21,064	23,685	26,783	29,141	36,123
15	19,311	22,307	24,996	28,259	30,578	37,697
16	20,465	23,542	26,296	29,633	32,000	39,252
17	21,615	24,769	27,587	30,995	32,409	40,790
18	22,760	25,989	28,869	32,346	34,805	42,312
19	23,900	27,204	30,144	33,678	36,191	43,820
20	25,038	28,412	31,410	35,020	37,566	45,315
21	26,171	29,615	32,671	36,343	38,932	46,797
22	27,301	30,813	33,924	37,659	40,289	48,268
23	28,429	32,007	35,172	38,968	41,638	49,728
24	29,553	33,196	36,415	40,270	42,980	51,179
25	30,675	34,382	37,652	41,566	42,314	52,620
26	31,795	35,563	38,885	42,856	45,642	54,052
27	32,912	36,741	40,113	44,140	46,963	55,476
28	34,027	37,916	41,337	45,419	48,278	56,893
29	35,139	39,087	42,557	46,693	49,588	58,302
30	36,250	40,256	43,773	47,962	50,892	59,703

Приложение 7.

Таблица значений $q = q(\gamma, n)$.

(1 – q) s < σ < (1+ q) s, если q < 1, 0 < σ < (1+ q) s , если q > 1.

$$\left(s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2} \right).$$

n \ γ	0,95	0,99	0,999	n \ γ	0,95	0,99	0,999
5	1,37	2,67	5,64	20	0,37	0,58	0,88
6	1,09	2,01	3,88	25	0,32	0,49	0,73
7	0,92	1,62	2,98	30	0,28	0,43	0,63
8	0,80	1,38	2,42	35	0,26	0,38	0,56
9	0,71	1,20	2,06	40	0,24	0,35	0,50
10	0,65	1,08	1,80	45	0, 22	0,32	0,46
11	0,59	0,98	1,60	50	0,21	0,30	0,43
12	0,55	0,90	1,45	60	0,188	0,269	0,38
13	0,52	0,83	1,33	70	0,174	0,245	0,34
14	0,48	0,78	1,23	80	0,161	0,226	0,31
15	0,46	0,73	1,15	90	0,151	0,211	0,29
16	0,44	0,70	1,07	100	0,143	0,198	0,27
17	0,42	0,66	1,01	150	0,115	0,160	0,211
18	0,40	0,63	0,96	200	0,099	0,136	0,185
19	0,39	0,60	0,92	250	0,089	0,120	0,162

Приложение 8.

Критические значения λ_α -распределения Колмогорова:

$P(\lambda \geq \lambda_\alpha) = \alpha$

Уровень значимости α	0,20	0,10	0,05	0,02	0,01	0,001
λ_α	1,073	1,224	1,358	1,520	1,627	1,950

Греческий алфавит

Прописные		*Строчные*		Название	Прописные		Строчные		Название
1	***2***	***1***	***2***	***3***	***1***	***2***	***1***	***2***	***3***
A	A	α	$\mathit{\alpha}$	а́льфа	N	N	ν	$\mathit{\nu}$	ни (ню)
B	B	β	$\mathit{\beta}$	бе́та	Ξ	$\mathit{\Xi}$	ξ	$\mathit{\xi}$	кси
Γ	$\mathit{\Gamma}$	γ	$\mathit{\gamma}$	га́мма	O	O	o	o	о микро́н
Δ	$\mathit{\Delta}$	δ	$\mathit{\delta}$	де́льта	Π	$\mathit{\Pi}$	π	$\mathit{\pi}$	пи
E	E	ε	$\mathit{\varepsilon}$	э псило́н	P	P	ρ	$\mathit{\rho}$	ро
Z	Z	ζ	$\mathit{\zeta}$	дзе́та	Σ	$\mathit{\Sigma}$	$\sigma\ \varsigma$	$\mathit{\sigma\ \varsigma}$	си□гма
H	H	η	$\mathit{\eta}$	э□та	T	T	τ	$\mathit{\tau}$	та́у
Θ	$\mathit{\Theta}$	θ	$\mathit{\theta}$	те́та	Y	Y	υ	$\mathit{\upsilon}$	и псило́н
I	I	ι	$\mathit{\iota}$	ио́та	Φ	$\mathit{\Phi}$	φ	$\mathit{\varphi}$	фи
K	K	κ	$\mathit{\kappa}$	ка́ппа	X	X	χ	$\mathit{\chi}$	хи
Λ	$\mathit{\Lambda}$	λ	$\mathit{\lambda}$	ла́мбда	Ψ	$\mathit{\Psi}$	ψ	$\mathit{\psi}$	пси
M	M	μ	$\mathit{\mu}$	ми (мю)	Ω	$\mathit{\Omega}$	ω	$\mathit{\omega}$	о ме́га

В таблице приводится написание греческих букв курсивом двумя шрифтами: первые столбцы – «Times New Roman», вторые – «Arial».

Рекомендуемая литература

1. Елисеева И.И., Юзбашев М.М. Общая теория статистики: Учебник. -М.: Финансы и статистика, 1996.
2. Ефимова М.Р. и др. Общая теория статистики: Учебник. -М.: ИНФРА –М, 1997.
3. Общая теория статистики./Под ред. Спирина А.А., Башиной О.Э.-М.: Финансы и статистика, 1995.
4. Статистика: национальные счета, показатели и методы анализа: Справочное пособие/ под общей ред. И.Э. Теслюка.-БГЭУ, 1995.
5. Общая теория статистики: Практикум / Под общей ред. М.М. Новикова.-Мн.: БГЭУ, 1996.
6. Теория статистики./Под ред. Р.А.Шмойловой.: Финансы и статистика, 1998.
7. Практикум по теории статистики: Учебное пособие / Под ред. проф. Р.А.Шмойловой.-М.: Финансы и статистика, 1999.
8. Громыко Г.М., Общая теория статистики: Практикум: ИНФРА-М, 1999.
9. Елисеева И.И. Моя профессия – статистик.- М.:Финансы и статистика, 1992.
10. Статистический словарь/Под ред. М.А. Королева. 2-е изд.-М.: Финансы и статистика, 1989.

Оглавление

УЧЕБНОЕ ИЗДАНИЕ

Годунов Борис Алексеевич

СТАТИСТИКА

Часть 1

(Конспект лекций)

Ответственный за выпуск Годунов Б. А.
Редактор Строкач Т. В.
Компьютерная верстка Боровикова Е. А.
Корректор Никитчик Е. В.

Лицензия № 02330/0133017 от 30.04.2004 г.
Подписано к печати 16.06.2008 г.
Формат 60×84 $^{1}/_{16}$. Бумага «Снегурочка».
Усл.п.л. 4,9.. Уч.изд.л. 5,25.
Тираж 200 экз. Заказ № 635. Отпечатано
на ризографе УО «Брестский
государственный технический университет».
224017, Брест, ул. Московская, 267.
Лицензия № 02330/0148711 от 30.04. 2004.

ISBN 978-985-493-086-2

9 789854 930862