МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ НАУКИ

Федеральное государственное бюджетное образовательное учреждение
высшего профессионального образования
«Пензенский государственный университет
архитектуры и строительства»
(ПГУАС)

И.Н. Максимова

МЕТОДЫ ОБРАБОТКИ ЭКСПЕРИМЕНТАЛЬНЫХ ДАННЫХ

Рекомендовано Редсоветом университета
в качестве учебного пособия для студентов,
обучающихся по направлению 27.03.01
«Стандартизация и метрология»

Пенза 2014

УДК 519.2(075.8)
ББК 22.17я73
М17

Рецензенты: доктор технических наук, профессор И.А. Гарькина (ПГУАС);
кандидат технических наук, начальник Управления по контролю и надзору в сфере образования Министерства образования Пензенской области А.А. Федосеев

Максимова И.Н.

М17 Методы обработки экспериментальных данных: учеб. пособие / И.Н. Максимова. – Пенза: ПГУАС. – 2014. – 116 с.

Учебное пособие знакомит с основными теоретическими положениями дисциплины «Методы обработки экспериментальных данных» и их практическим применением. Наибольшее внимание уделено методам обработки экспериментальных данных, таким как проверка статистических гипотез, методы оценки параметров распределения, аппроксимация закона распределения экспериментальных данных, однофакторный дисперсионный анализ, корреляционный анализ.

Разделы учебного пособия имеют вопросы для самооценки и задачи для самостоятельного решения, позволяющие качественно и эффективно освоить базовый уровень знаний.

Учебное пособие подготовлено на кафедре «Управление качеством и технология строительного производства» согласно федеральному государственному образовательному стандарту по направлению 27.03.01 «Стандартизация и метрология».

© Пензенский государственный университет архитектуры и строительства, 2014
© Максимова И.Н., 2014

ПРЕДИСЛОВИЕ

Объекты исследований, как правило, сложны и связаны со значительным количеством как управляемых, так и неуправляемых (независимых) факторов. На параметры их состояния могут существенно влиять элементы случайностей, имеющих сложную природу происхождения. Для установления закономерностей функционирования этих объектов в реальных условиях одних теоретических исследований недостаточно, так как аналитически описать изучаемый объект с достаточной точностью не всегда представляется возможным. Такие объекты характерны практически для всех направлений прикладных исследований: как в технологии, так и в технике, и в области естественных наук.

Экспериментальные данные могут быть использованы для проверки и уточнения рабочих гипотез, а также обоснования направления исследований в соответствующей области. Эффективность исследований в целом повышается, если теоретические предпосылки уточняются опытным путем, а экспериментальные данные анализируются и обобщаются на базе теоретических положений соответствующих отраслей наук.

Учебное пособие посвящено изложению исходных понятий по обработке результатов наблюдений, раскрытию сущности задач и методов обработки экспериментальных данных, относящихся к одному простому свойству исследуемого объекта. Рассмотрены три основные группы вопросов: оценка параметров распределения; проверка статистических гипотез; подбор (аппроксимация) закона распределения для описания данных. В приложении приведены фрагменты статистических таблиц, широко применяемых в задачах обработки экспериментальных данных.

Учебное пособие подготовлено на кафедре «Управление качеством и технологии строительного производства» согласно федеральному государственному образовательному стандарту по направлению 27.03.01 – Стандартизация и метрология (квалификация (степень) выпускника – бакалавр) и рабочему учебному плану 221700_62–09–1234–3467 по дисциплине вариативной части математического и естественнонаучного цикла дисциплин Б2.В.ДВ.1.1 «Методы обработки экспериментальных данных».

ВВЕДЕНИЕ

Эмпирические исследования являются основным источником объективной информации о характеристиках процессов, протекающих в реальных объектах. Целью обработки экспериментальных данных является выявление закономерностей в характеристиках исследуемых объектов и процессов. Результаты обработки экспериментальных данных позволяют оценить качество объекта, они необходимы для оперативного управления процессами, решения задач адаптации объекта к изменившимся условиям или формирования требований ко вновь создаваемым системам.

Получение экспериментальной информации связано с решением ряда проблем по организации регистрации первичных параметров, их сбора и обработки. Те данные, которые можно непосредственно зарегистрировать, обычно лишь косвенно отражают существенные свойства изучаемого процесса или объекта. Многие показатели качества автоматизированных систем носят случайный характер и по этой причине не могут быть непосредственно измерены. Ряд событий в системах происходит крайне редко, и получить для них достаточный объем эмпирических данных (в частности, получить данные по отказам систем с высокой надежностью) невозможно.

Методы обработки экспериментальных данных начали разрабатываться более двух веков тому назад в связи с необходимостью решения практических задач по агробиологии, медицине, экономике, социологии. Полученные при этом результаты составили фундамент такой научной дисциплины как математическая статистика. В последние десятилетия математический аппарат обработки экспериментальных данных получил значительное развитие в связи с необходимостью решения принципиально новых задач. И к настоящему времени он включает множество различных направлений, которые выходят за пределы классической математической статистики. Многие методы нашли применение при исследовании технических и человеко-машинных систем, а также при обработке результатов имитационного (статистического) моделирования.

В пределах одного учебного пособия изложить все многообразие основных методов обработки экспериментальных данных невозможно. Материал ограничен раскрытием основ обработки экспериментальных данных применительно к стационарному режиму функционирования объекта, вопросы оценки характеристик случайных функций в пособии не затрагиваются. Из всех форм представления экспериментальных данных рассматривается только одна наиболее универсальная – числовая. Предполагается, что экспериментальные данные получены в результате проведения пассивного эксперимента, а объем данных фиксирован к началу обработки.

В основу учебного пособия положен одноименный курс лекций, читаемый студентам направления 27.03.01 – Стандартизация и метрология.

1. ОБЩАЯ ХАРАКТЕРИСТИКА ЭКСПЕРИМЕНТАЛЬНЫХ ДАННЫХ

1.1. Понятие экспериментального исследования

Эксперимент – это метод исследования, состоящий в целенаправленном воздействии на объект в заданных контролируемых условиях, позволяющих следить за ходом его проведения с точным фиксированием значений заранее намеченных параметров исследуемого объекта с требуемой надежностью и точностью и воссоздать его каждый раз по мере необходимости при повторении тех же условий его проведения. При этом как условия, так и параметры исследуемого объекта (параметры рабочих органов машин и оборудования, отдельных операций технологических процессов, характеристики явлений и т.д.) могут меняться в заранее заданных интервалах варьирования.

При проведении эксперимента исследователь прибегает к другим (более простым) методам эмпирического исследования:

- *наблюдению*, основанному на целенаправленном восприятии явлений (причем исследователю на основе имеющихся у него знаний известно, что и как наблюдать);
- *описанию*, основанному на фиксации сведений, полученных на основе наблюдения;
- *измерению,* заключающемуся в сравнении объектов по каким-либо сходным свойствам, признакам с эталоном и установлении количественных характеристик.

Основной задачей любого эксперимента является не только получение неизвестных ранее сведений об объекте исследования, но и достоверное установление закономерностей его поведения в изменяющихся условиях, совпадающих с условиями его функционирования в природе, технике, общественной жизни и т.д. С помощью эксперимента могут быть получены данные, обработка которых позволит получить математические модели, достоверно описывающие изучаемый объект, вскрыть закономерности его поведения в изменяющихся условиях, т.е. решить задачу идентификации.

В зависимости от особенностей объекта и поставленных целей экспериментальные исследования могут проводиться в различных условиях. При этом различают лабораторные, лабораторно-полевые, заводские и т.п. исследования.

Для получения надежных и достоверных результатов экспериментальных исследований необходимо осуществить:

- анализ характеристик исследуемого объекта во всем многообразии свойств, предусмотренных целью проведения исследования, на основе имеющихся сведений, полученных другими исследователями и опубликованных в источниках информации;

- разработку программы экспериментальных исследований;
- обоснование выбора количественных параметров (критериев или измеряемых величин) оценки свойств объекта, выбор их размерности и способов измерения в ходе эксперимента;
- определение всех факторов, влияющих на выбранные на основе поисковых исследований (если в этом есть необходимость) для наблюдения параметры рассматриваемого объекта в условиях протекания изучаемых явлений и процессов;
- рассмотрение причинно-следственных связей между параметрами оценки свойств объекта и выявленными факторами;
- ранжирование факторов по степени их влияния на параметры (критерии) оценки свойств объекта и выделение из них основных (доминирующих);
- определение рациональных интервалов варьирования выделенных факторов для установления соответствующих закономерностей, предусмотренных программой исследований;
- фиксирование остальных факторов на определенных (возможно, лучших) уровнях варьирования;
- разработку конструктивно-технологических схем опытно-экспериментальных установок или стендов, обеспечивающих реализацию намеченной программы исследований;
- изучение возможностей моделирования объекта;
- подбор соответствующей существующей или разработка новой измерительной аппаратуры, фиксирующей во время проведения экспериментов измеряемые величины через исполнительные органы (датчики, усилители, компьютеры и т.д.);
- разработку методики тарировки выбранных средств измерения, их установки для надежного измерения или регистрации контролируемых величин;
- разработку методики обработки первичной документации, в том числе журналов наблюдений, протоколов или актов исследований, с целью обеспечения требований надежности, точности и достоверности результатов эксперимента.

В ходе обработки результатов эксперимента устанавливаются закономерности исследуемых явлений и процессов, происходящих с участием изучаемого объекта, которые позволяют получить достоверный ответ на интересующие исследователя задачи и вопросы. Для иллюстрации установленных закономерностей принято использовать таблицы, диаграммы, графики, математические модели и др.

При исследовании сложных систем часто возникают ситуации, когда однозначно нельзя выделить и изолированно изучить отдельные явления или процессы. В этом случае в экспериментальном исследовании объекта используют построение математических моделей, которые с соответст-

вующей степенью достоверности описывают реальный изучаемый объект. При этом точное понятие «закон» или «закономерность» заменяется более приблизительным и абстрактным понятием «модель», которое носит элемент многозначности и какой-то неопределенности, однако практически более понятный и удобный в применении. Безусловно, здесь нет противоречий, если эта модель описывает объект исследований с требуемой надежностью и точностью оценочных параметров. Как и в случае теоретических исследований, при построении моделей в эксперименте одни и те же системы и процессы могут быть описаны разными моделями и с различной точностью – в зависимости от конкретных условий.

Иногда при обработке данных эксперимента ставится задача оптимизации параметров объекта по каким-то количественным или качественным критериям. С этой целью применяются известные методики, соответствующая программа и технические средства обработки данных. Достоверность того, что параметры оптимизации действительно позволяют оптимизировать свойство объекта, должна быть подтверждена прямым экспериментом, условия проведения которого соответствуют оптимизированным параметрам. Лишь в этом случае можно говорить о достоверности полученных практических выводов и рекомендаций.

В прикладных исследованиях, особенно технического профиля, заключительным этапом является проведение испытаний исследуемого объекта в условиях производства. Испытание – это разновидность научных экспериментальных исследований, при которых исследуемый объект подвергается оценке в производственных условиях, для работы в которых он, собственно, и предназначен. При испытаниях не изменяют параметров его эксплуатации, кроме тех, которые предусмотрены соответствующими требованиями инструкций по эксплуатации и техническому обслуживанию в виде отдельных регулировок механизмов. Цель таких испытаний состоит в определении соответствия данного объекта исследования тем производственным требованиям, которые были первоначально поставлены перед исследователями (разработчиками).

Государственными нормативными документами сегодня предусматривается проведение порядка сорока различных видов испытаний. Основными из них являются:

- предварительные заводские или полевые испытания опытного образца;
- приемочные испытания доработанных образцов или опытных партий (установочной серии);
- контрольные испытания при массовом производстве машин;
- испытания образцов после капитального ремонта.

Первые два вида испытаний применяются на стадии проектирования, научных исследований и доработки новых конструкций машин и оборудования до их работоспособного состояния. С их помощью оценивается эф-

фективность идей, технологических и технических решений, обоснованность выбора величины отдельных параметров, конструктивно-технологических и компоновочных схем, заложенных в такие машины и оборудование, степень обоснованности и оптимальности базовых (основных) величин параметров. При этом выявляются ошибки, допущенные при проектировании, уточняются параметры основных элементов исследуемого объекта, возможные отклонения, надежность работы в производственных условиях и дается вывод о перспективности дальнейшего использования его по основному назначению.

1.2. Источники и вид представления экспериментальных данных

Экспериментальные исследования событий и процессов основаны на наблюдениях, в ходе которых регистрируются различные факты искусственного и естественного происхождения.

Источниками экспериментальных данных являются:

– результаты наблюдения за реальными объектами и протекающими в них процессами. Наблюдения могут проводиться в ходе испытаний или в ходе обычной эксплуатации;

– результаты моделирования объектов. В первую очередь к ним следует отнести результаты имитационного моделирования;

– технические, экономические, научные отчеты и обозрения, публикуемые в различных изданиях, например, сведения о результатах испытаний или о характеристиках однотипных устройств различных производителей;

– результаты опросов специалистов и другие источники.

Обработка экспериментальных данных, получаемых от различных источников, имеет много общего. Однако организация сбора и интерпретации экспериментальных данных специфична для конкретной предметной области. В дальнейшем обработка экспериментальных данных будет рассматриваться применительно к результатам наблюдения за функционированием физических величин и средств измерений, их элементов или их моделей.

Вид экспериментальных данных определяет форму представления, степень зависимости от времени, характер данных.

Одной из основных форм представления является *символьная,* которая включает представление данных в виде чисел, двоичных величин или текста. Для задания значений соответствующих величин применяются различные шкалы измерений. Описательные (качественные) признаки измеряются на основе номинальных и порядковых шкал. Номинальные шкалы обеспечивают только группирование объектов по признаку наличия у них

некоторых общих свойств, но не позволяют проводить ранжирование объектов. Порядковые шкалы обеспечивают возможность упорядочивания данных по признакам «больше», «меньше», «равно», но при этом не указывается, насколько одно значение признака больше или меньше другого.

Количественные свойства отображаются числами в относительных или абсолютных шкалах измерений. В относительных шкалах точки начала отсчета и масштаб измерений имеют условный характер. Например, температуру можно измерять в относительных шкалах по Цельсию, Реомюру, Фаренгейту. Исходя из этого, результаты количественного сравнения величин зависят от используемой шкалы, а некоторые операции над количественными признаками недопустимы. Например, температура одного объекта выше температуры другого на три градуса Цельсия, но эти три градуса не равны трем градусам шкалы Фаренгейта. Бессмысленно говорить, во сколько раз температура одного объекта выше температуры другого, в частности, нельзя сказать, что температура +10 ºС в два раза выше, чем +5 ºС. Абсолютная шкала обеспечивает однозначное представление точки отсчета и масштаба. Примерами абсолютных шкал является шкала температур по Кельвину, шкала вероятностей. Эти шкалы позволяют дать однозначные ответы на вопросы о том, насколько или во сколько раз одна величина больше (меньше) другой. Именно применение относительных и абсолютных шкал дает возможность проводить количественную обработку экспериментальных данных. Но при обработке следует применять только те операции, которые допускаются применяемой шкалой измерений.

Количественные характеристики (параметры) представимы дискретными или непрерывными величинами. Дискретные параметры принимают только отдельные значения, без промежуточных значений между ними. Часто дискретные параметры имеют такое большое количество допустимых значений, что их на практике считают квазинепрерывными. Непрерывные параметры способны принимать любые значения из некоторого допустимого диапазона. В процессе обработки непрерывные величины всегда округляют и представляют ограниченным числом разрядов, т.е. они становятся квазинепрерывными. На практике экспериментальные данные отображают с разрядностью, обеспечивающей относительную погрешность не более единиц или десятых долей процента.

Экспериментальные данные могут быть представлены не только в символьной, но и в других формах: *графической* (графики, осциллограммы, штриховые рисунки, цветные изображения и полутоновые рисунки); *аудиоданных*. Такие данные обрабатывают непосредственно или предварительно преобразуют в числовую форму.

Мы будем рассматривать только числовую форму представления данных как универсальную и широко распространенную форма представления информации, количественно характеризующую параметры объектов и процессов. А сами параметры будем считать непрерывными величинами, если

особо не оговорено иное. Обработка данных, представленных в других формах, обладает существенной спецификой и требует отдельного рассмотрения.

Параметры, вообще говоря, зависят или не зависят от времени. В учебном пособии мы будем изучать оценки параметров, не зависящие от времени.

Результаты наблюдений носят детерминированный или случайный характер. Большинство событий и процессов в технологических системах можно считать случайными. Именно на обработке таких данных сосредоточено внимание в пособии.

Существенной особенностью задач оценивания параметров является наличие цензурирования, под которым понимается отсутствие в результатах наблюдений каких-либо элементов. Цензурирование возникает по ряду причин, например, как следствие разновременности начала и окончания работы различных устройств или из-за того, что время свершения некоторого события выходит за пределы периода наблюдения. Обработка таких экспериментальных данных требует применения специального математического аппарата.

Таким образом, в данном учебном пособии *в качестве объекта обработки рассматриваются совокупности числовых данных, характеризующих не зависящие от времени случайные значения непрерывных параметров*.

1.3. Цели обработки экспериментальных данных

Основные показатели средств измерений, например, показатели производительности и надежности, носят вероятностный характер и не могут быть непосредственно измерены. Для их оценки следует применять косвенные способы на основе регистрации соответствующих первичных параметров и последующей обработки накопленных данных с привлечением специальных математических методов. Иначе говоря, экспериментальные данные представляют собой лишь наборы возможных случайных значений показателей, зарегистрированных в некоторые моменты времени. Например, продолжительность наработок до отказа некоторой совокупности однотипных устройств можно рассматривать как множество возможных случайных значений показателя «наработка до отказа». Именно по наработкам и необходимо оценить значение этого показателя. Сам показатель как случайная величина характеризуется законом распределения, моментами распределения или другими параметрами, которые и следует определить.

Основными целями обработки экспериментальных данных [17] являются следующие:

– оценка значений показателей качества измерительных средств, комплексов или системы в целом. На стадиях создания такая оценка проводится в интересах обоснования принимаемых решений по построению объектов, проверки показателей на соответствие требованиям, выявления существенных факторов, влияющих на функционирование объектов, выявления причин несоответствия требованиям. На стадии эксплуатации обработка экспериментальных данных проводится также для решения задач управления объектом: изменения режимов работы объекта; изменения порядка обработки информации; обоснования данных для модернизации объекта (изменения конфигурации технических и программных средств); адаптации объекта к условиям функционирования;

– сжатие информации о функционировании объекта, ее обобщение для последующего применения в интересах исследования подобных объектов, обоснования данных для создания новых систем;

– выявление закономерностей функционирования объекта в конкретных условиях эксплуатации, т.е. установление зависимостей между параметрами объекта, внешней среды и показателями качества объекта. Выявленные закономерности применяют для поиска оптимальных значений параметров при синтезе новых систем, для упрощенного описания объекта в модели суперсистемы;

– выявление существенных параметров системы и внешней среды;

– изучение типологии объектов (распознавание образов, классификация объектов);

– прогнозирование развития объектов в интересах организационного и технологического управления.

Следует помнить, что результаты обработки экспериментальных данных не гарантируют достоверного описания неизвестных показателей или закономерностей, их необходимо рассматривать только лишь как более-менее удачную аппроксимацию соответствующих характеристик.

Необходимость сбора и обработки экспериментальных данных обусловлена объективными обстоятельствами, так как действительные значения показателей качества сложных объектов обычно существенно отличаются от рассчитанных на стадиях проектирования. Эти различия являются следствием ряда причин:

– на стадиях проектирования нет достаточно полных и точных представлений о характеристиках процессов, протекающих в объекте и во внешней среде, поэтому при проектировании приходится вводить существенные допущения и ограничения. На практике часть принятых допущений оказывается не вполне справедливой;

– с вводом в действие объекта внешняя среда постепенно меняет свои характеристики, например, меняются стиль и методы работы пользователей, появляются дополнительные потребности в решении задач. Такие изменения заранее предусмотреть невозможно;

– условия эксплуатации, квалификация обслуживающего персонала в разных организациях имеют свою специфику;

– в ходе эксплуатации изменяются характеристики технических, программных и информационных средств, меняется их взаимное отображение. Эти обстоятельства приводят к изменению характеристик потоков запросов на решение задач и параметров их обслуживания.

Таким образом, фактические значения показателей качества не только отличаются от расчетных, но и меняются с течением времени, имеют свои особенности для одних и тех же типов объектов, эксплуатируемых в различных организациях, претерпевают колебания, зависящие от времени, характера выполняемых на объекте работ.

В зависимости от стадии жизненного цикла объекта задачи обработки экспериментальных данных имеют ряд особенностей. На стадии создания имеется принципиальная возможность проведения *активных и пассивных экспериментов*. Понятие "*активный*" подразумевает возможность выбора объема экспериментов, последовательности и значений характеристик воздействий на объект по желанию исследователя. Проведение активных экспериментов позволяет расширить диапазон условий, при которых проводится оценивание качества. А специальным образом подобранные условия проведения исследования и порядок задания внешних воздействий, т.е. рационально обоснованные планы экспериментов, обеспечивают взаимную статистическую независимость результатов испытаний. Эти обстоятельства значительно облегчают обработку экспериментальных данных, повышают качество получаемых оценок, позволяют разделить влияние различных факторов при построении модели функционирования объекта. Активные эксперименты в основном проводятся на завершающих стадиях создания в виде испытаний опытных образцов, фрагментов систем и т.п. Вопросы постановки активных экспериментов, методов обработки их результатов изучаются в рамках специальной теории планирования экспериментов.

Реальные условия эксплуатации объекта в конкретной организации могут отличаться от предполагаемых при проведении испытаний. В связи с этим, показатели, установленные в ходе испытаний, несут в себе известную долю абстракции. На стадии создания время испытаний весьма ограничено, что не дает возможности сформировать большой объем экспериментальных данных для полноценной оценки искомых показателей.

В *пассивных* экспериментах количество наблюдений, последовательность и значения воздействий определяются реальной обстановкой использования объектов. Иначе говоря, исследователь практически лишен возможности управления качеством и количеством экспериментальных данных.

На стадии эксплуатации возможности проведения активных экспериментов значительно ограничены или вообще отсутствуют, что обычно приводит к взаимной зависимости результатов наблюдений. Наличие такой за-

висимости затрудняет обработку полученных данных, а игнорирование данного обстоятельства приводит к смещению значений получаемых оценок, невозможности разделения влияния различных факторов на показатели функционирования и к другим нежелательным последствиям. Нестационарные условия эксплуатации, влияние на объекты периодических или нерегулярно изменяющихся воздействий (трендов) обусловливают необходимость рассмотрения характеристик не как случайных величин, а как случайных функций. Однако из-за слабой разработанности методов и средств оценки параметров случайных процессов по результатам наблюдения обычно предполагается, что процесс функционирования объекта носит стационарный или кусочно-стационарный характер.

1.4. Основные задачи математической статистики

Математические законы теории вероятностей не являются беспредметными абстракциями, лишенными физического содержания; они представляют собой математическое выражение реальных закономерностей, фактически существующих в массовых случайных явлениях природы [4].

Говоря о законах распределения случайных величин, необходимо поставить вопрос о том, откуда берутся, на каком основании устанавливаются эти законы распределения. Ответ на вопрос вполне определенен – в основе всех этих характеристик лежит опыт; каждое исследование случайных явлений, выполняемое методами теории вероятностей, прямо или косвенно опирается на экспериментальные данные. Оперируя такими понятиями, как события и их вероятности, случайные величины, их законы распределения и числовые характеристики, теория вероятностей дает возможность теоретическим путем определять вероятности одних событий через вероятности других, законы распределения и числовые характеристики одних случайных величин через законы распределения и числовые характеристики других. Такие косвенные методы позволяют значительно экономить время и средства, затрачиваемые на эксперимент, но отнюдь не исключают самого эксперимента. Каждое исследование в области случайных явлений, как бы отвлеченно оно ни было, корнями своими всегда уходит в эксперимент, в опытные данные, в систему наблюдений.

Разработка методов регистрации, описания и анализа статистических экспериментальных данных, получаемых в результате наблюдения массовых случайных явлений, составляет предмет специальной науки – математической статистики.

Все задачи математической статистики касаются вопросов обработки наблюдений над массовыми случайными явлениями, но в зависимости от характера решаемого практического вопроса и от объема имеющегося экс-

периментального материала эти задачи могут принимать ту или иную форму.

Охарактеризуем вкратце некоторые типичные задачи математической статистики, часто встречаемые на практике.

1. Задача определения закона распределения случайной величины (или системы случайных величин) по статистическим данным

Закономерности, наблюдаемые в массовых случайных явлениях, проявляются тем точнее и отчетливее, чем больше объем статистического материала. При обработке обширных по своему объему статистических данных часто возникает вопрос об определении законов распределения тех или иных случайных величин. Теоретически при достаточном количестве опытов свойственные этим случайным величинам закономерности будут осуществляться сколь угодно точно. На практике всегда приходится иметь дело с ограниченным количеством экспериментальных данных; в связи с этим результаты наших наблюдений и их обработки всегда содержат больший или меньший элемент случайности. Возникает вопрос о том, какие черты наблюдаемого явления относятся к постоянным, устойчивым и действительно присущи ему, а какие являются случайными и проявляются в данной серии наблюдений только за счет ограниченного объема экспериментальных данных. Естественно, к методике обработки экспериментальных данных следует предъявить такие требования, чтобы она, по возможности, сохраняла типичные, характерные черты наблюдаемого явления и отбрасывала все несущественное, второстепенное, связанное с недостаточным объемом опытного материала. В связи с этим возникает характерная для математической статистики задача сглаживания или выравнивания статистических данных, представления их в наиболее компактном виде с помощью простых аналитических зависимостей.

2. Задача проверки правдоподобия гипотез

Эта задача тесно связана с предыдущей; при решении такого рода задач мы обычно не располагаем настолько обширным статистическим материалом, чтобы выявляющиеся в нем статистические закономерности были в достаточной мере свободны от элементов случайности. Статистический материал может с большим или меньшим правдоподобием подтверждать или не подтверждать справедливость той или иной гипотезы. Например, может возникнуть такой вопрос: согласуются ли результаты эксперимента с гипотезой о том, что данная случайная величина подчинена закону распределения $F(x)$? Другой подобный вопрос: указывает ли обнаруженная в опыте тенденция на наличие действительной объективной зависимости между двумя случайными величинами или же она объясняется случайными причинами, связанными с недостаточным объемом наблюдений? Для решения подобных вопросов математическая статистика выработала ряд специальных приемов.

3. Задача нахождения неизвестных параметров распределения

Часто при обработке статистического материала вовсе не возникает вопрос об определении законов распределения исследуемых случайных величин. Обыкновенно это бывает связано с крайне недостаточным объемом экспериментального материала. Иногда же характер закона распределения качественно известен до опыта, из теоретических соображений; например, часто можно утверждать заранее, что случайная величина подчинена нормальному закону. Тогда возникает более узкая задача обработки наблюдений – определить только некоторые параметры (числовые характеристики) случайной величины или системы случайных величин. При небольшом числе опытов задача более или менее точного определения этих параметров е может быть решена; в этих случаях экспериментальный материал содержит в себе неизбежно значительный элемент случайности; поэтому случайными оказываются и все параметры, вычисленные на основе этих данных. В таких условиях может быть поставлена только задача об определении так называемых «оценок» или «подходящих значений» для искомых параметров, т.е. таких приближенных значений, которые при массовом применении приводили бы в среднем к меньшим ошибкам, чем всякие другие. С задачей отыскания «подходящих значений» числовых характеристик тесно связана задача оценки их точности и надежности.

Итак, современная математическая статистика разрабатывает способы определения числа необходимых испытаний до начала исследования (планирование эксперимента), в ходе исследования (последовательный анализ) и является наукой о принятии решений в условиях неопределенности; ее задача состоит в создании методов сбора и обработки статистических данных для получения научных и практических выводов. В последующих главах мы вкратце познакомимся с некоторыми, наиболее элементарными задачами математической статистики и с методами их решения.

Вопросы для самооценки

1. Что такое эксперимент?
2. С какими методами эмпирического исследования Вы знакомы?
3. Что является задачей эксперимента?
4. Каковы могут быть условия проведения эксперимента?
5. Какова постановка задачи эксперимента?
6. Как обрабатывают результаты эксперимента?
7. Что является источником экспериментальных данных?
8. Какой может быть форма представления экспериментальных данных?
9. Каковы цели обработки экспериментальных данных?
10. Каковы основные задачи математической статистики?

2. БАЗОВЫЕ ПОНЯТИЯ И ОПЕРАЦИИ ОБРАБОТКИ ЭКСПЕРИМЕНТАЛЬНЫХ ДАННЫХ

2.1. Эмпирическая функция распределения

Методы обработки экспериментальных данных опираются на базовые понятия теории вероятностей и математической статистики. К их числу относятся понятия генеральной и выборочной совокупности, выборки, эмпирической функции распределения [5, 17].

Объемом совокупности (выборочной или генеральной) называют число объектов этой совокупности.

Выборочной совокупностью или просто *выборкой* называют совокупность случайно отобранных объектов.

Под *генеральной совокупностью* понимают все возможные значения параметра, которые могут быть зарегистрированы в ходе неограниченного по времени наблюдения за объектом. Такая совокупность состоит из бесконечного множества элементов. В результате наблюдения за объектом формируется ограниченная по объему совокупность значений параметра $x_1, x_2, ...x_n$. С формальной точки зрения такие данные представляют собой *выборку* из генеральной совокупности. Будем считать, что выборка содержит полные наработки до системных событий (цензурирование отсутствует). Наблюдаемые значения x_i называют *вариантами*, их количество – объемом выборки n.

Для того чтобы по результатам наблюдения можно было достаточно уверенно судить об интересующем признаке генеральной совокупности, необходимо, чтобы объекты выборки его правильно представляли, т.е. выборка должна быть *репрезентативной* (*представительной*). Это требование выполняется, если объем выборки достаточно велик, а каждый элемент генеральной совокупности имеет одинаковую вероятность попасть в выборку.

Пусть из генеральной совокупности извлечена выборка, при этом значение параметра x_1 наблюдалось n_1 раз, значение x_2 – n_2 раз, значение x_k – n_k раз, причем $n_1 + n_2 + ... + n_k = n$. Последовательность вариант, записанных в порядке их возрастания, называют *вариационным рядом*, величины n_i – *частотами*, а их отношения к объему выборки $W_i = n_i / n$ – *относительными частотами* (или *частостями*). Очевидно, что сумма относительных частот равна единице.

Статистическим распределением выборки называют перечень вариант и соответствующих им частот (или относительных частот). Статистическое распределение можно задать также в виде последовательности интервалов и соответствующих им частот (в качестве частоты, соответствующей интервалу, принимают сумму частот, попавших в этот интервал).

Вообще, в теории вероятностей под *распределением* понимают соответствие между возможными значениями случайной величины и их вероятностями, а в математической статистике – соответствие между наблюдаемыми вариантами и их частотами, или относительными частотами.

При большом числе наблюдений простая статистическая совокупность перестает быть удобной формой записи статистического материала – она становится слишком громоздкой и мало наглядной. Для придания ему большей компактности и наглядности статистический материал должен быть подвергнут дополнительной обработке – строится так называемый *статистический ряд*.

Пусть известно статистическое распределение частот количественного признака X. Введем обозначения: n_x – число наблюдений, при которых наблюдалось значение признака, меньшее x; n – общее число наблюдений (объем выборки). Понятно, что относительная частота события $X < x$ равна n_x/n. Если x изменяется, то изменяется и относительная частота, таким образом, относительная частота n_x/n есть функция от x. Так как эта функция находится эмпирическим (опытным) путем, то ее называют эмпирической.

Эмпирической функцией распределения (функцией распределения выборки) называют функцию $F^*(x)$, определяющую для каждого значения x относительную частоту события $X < x$.

Согласно определению: $F^*(x) = n_x/n$, где n_x – число вариант, меньших x; n – объем выборки.

Тогда для того чтобы найти, например, $F^*(x_2)$, надо число вариант, меньших x_2, разделить на объем выборки: $F^*(x_2) = n_{x_2}/n$.

В отличие от эмпирической функции распределения выборки функцию распределения $F(x)$ генеральной совокупности называют *теоретической функцией распределения*.

Различие между эмпирической и теоретической функциями распределения состоит в том, что теоретическая функция $F(x)$ определяет вероятность события $X < x$, а эмпирическая функция $F^*(x)$ определяет относительную частоту этого же события.

Из теоремы Бернулли следует, что относительная частота события $X < x$, т.е. $F^*(x)$ стремится по вероятности к вероятности $F(x)$ этого события. Другими словами, при больших n числа $F^*(x)$ и $F(x)$ мало отличаются одно от другого в том смысле, что

$$\lim_{n\to\infty} P\left[\left|F(x) - F^*(x)\right| < \varepsilon\right] = 1(\varepsilon > 0).$$

Уже отсюда следует целесообразность использования эмпирической функции распределения выборки для приближенного представления теоретической (интегральной) функции распределения генеральной совокупности.

Данное заключение подтверждается и тем, что $F^*(x)$ обладает всеми свойствами $F(x)$. Действительно, из определения функции $F^*(x)$ вытекают следующие ее свойства:

1) значения эмпирической функции принадлежат отрезку [0, 1];

2) $F^*(x)$ – неубывающая функция;

3) если x_1 – наименьшая варианта, то $F^*(x)=0$ при $x \le x_1$; если x_k – наибольшая варианта, то $F^*(x)=1$ при $x > x_k$.

Таким образом, эмпирическая функция распределения выборки служит для оценки теоретической функции распределения генеральной совокупности.

График эмпирической функции $F_n(x)$ представляет собой ломаную линию. В промежутках между соседними членами вариационного ряда $F_n(x)$ сохраняет постоянное значение. При переходе через точки оси x, равные членам выборки, $F_n(x)$ претерпевает разрыв, скачком возрастая на величину $1/n$, а при совпадении l наблюдений – на l/n.

Пример 2.1. Построить вариационный ряд и график эмпирической функции распределения по результатам наблюдений табл. 2.1.

Таблица 2.1

i	1	2	3	4	5	6
x_i	51	43	56	60	64	56

Решение. Построим вариационный ряд, упорядочив по возрастанию значения варианты (табл. 2.2).

Таблица 2.2

i	1	2	3	4	5	6
X_i	43	51	56	56	60	64

Искомая эмпирическая функция (рис. 2.1):

$$F_6(x)=\begin{cases}0, & \text{при } x\le 43,\\ 0{,}16, & \text{при } 43<x\le 51,\\ 0{,}33, & \text{при } 51<x\le 56,\\ 0{,}67, & \text{при } 56<x\le 60,\\ 0{,}84, & \text{при } 60<x\le 64,\\ 1, & \text{при } x>64.\end{cases}$$

При большом объеме выборки (понятие «большой объем» зависит от целей и методов обработки, в данном случае будем считать n большим, если $n>40$) в целях удобства обработки и хранения сведений прибегают к группированию экспериментальных данных в интервалы. Количество интервалов следует выбрать так, чтобы в необходимой мере отразилось разнообразие значений параметра в совокупности, и в то же время, – закономерность распределения не искажалась случайными колебаниями частот по отдельным разрядам.

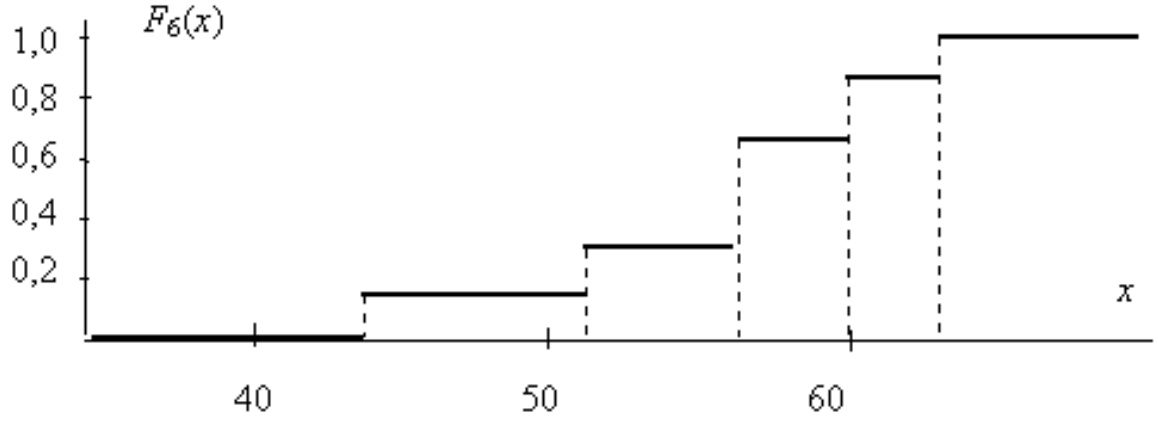

Рис. 2.1. Эмпирическая функция распределения

Существуют нестрогие рекомендации по выбору количества n и размера (длины) h таких интервалов, в частности:

– в каждом интервале должно находиться не менее 5-7 элементов. В крайних разрядах допустимо всего два элемента;

– количество интервалов не должно быть очень большим или очень маленьким. Минимальное значение n должно быть не менее 6-7. При объеме выборки, не превышающем несколько сотен элементов, величину n задают в пределах от 10 до 20. Для очень большого объема выборки ($n>1000$) количество интервалов может превышать указанные значения. Некоторые исследователи рекомендуют пользоваться соотношением $k=1{,}44\ln n+1$; длины интервалов удобно выбирать одинаковыми и равными величине $h=\left(x_{\max}-x_{\min}\right)/k$, где $x_{\max}$ – максимальное и $x_{\min}$ – минимальное значения вариант. При значительной неравномерности закона

распределения длины интервалов можно задавать меньшего размера в области быстрого изменения плотности распределения.

Группирование результатов наблюдений по интервалам предусматривает: определение размаха изменений параметра x; выбор количества интервалов и их величины; подсчет для каждого i-го интервала $[x_i - x_{i+1}]$ частоты n_i или относительной частоты (частости W_i) попадания варианты в интервал. В результате формируется представление экспериментальных данных в виде *интервального или статистического ряда*.

Полигоном частот называют ломаную линию, отрезки которой соединяют точки (x_1, n_1), (x_2, n_2), …, (x_k, n_k). Для построения полигона частот на оси абсцисс откладывают варианты x_i, а на оси ординат – соответствующие им частоты n_i. Точки (x_i, n_i) соединяют отрезками прямых и получают полигон частот.

Полигоном относительных частот называют ломаную линию, отрезки которой соединяют точки (x_1, W_1), (x_2, W_2), …, (x_k, W_k). Для построения полигона относительных частот на оси абсцисс откладывают варианты x_i, а на оси ординат – соответствующие им относительные частоты W_i. Точки (x_i, W_i) соединяют отрезками прямых и получают полигон относительных частот.

В случае непрерывного признака целесообразно строить гистограмму, для чего интервал, в котором заключены все наблюдаемые значения признака, разбивают на несколько частичных интервалов длиной h и находят для каждого частичного интервала n_i – сумму частот вариант, попавших в i-й интервал.

Гистограммой частот называют ступенчатую фигуру, состоящую из прямоугольников, основаниями которых служат интервалы длиной h, а высоты равны отношению n_i/h – плотность частоты. Для построения гистограммы частот на оси абсцисс откладывают частичные интервалы, а над ними проводят отрезки, параллельные оси абсцисс, на расстоянии n_i/h. Площадь i-го частичного прямоугольника равна $h \cdot \frac{n_i}{h} = n_i$ – сумме частот вариант i-го интервала. Следовательно, *площадь гистограммы частот равна сумме всех частот, т.е. объему выборки.*

Гистограммой относительных частот называют ступенчатую фигуру, состоящую из прямоугольников, основаниями которых служат интервалы длиной h, а высоты равны отношению W_i/h – плотности относительной частоты. Для построения гистограммы относительных частот на оси абсцисс откладывают частичные интервалы, а над ними проводят отрезки, параллельные оси абсцисс, на расстоянии W_i/h. Площадь i-го час-

тичного прямоугольника равна $h \cdot \frac{W_i}{h} = W_i$ – сумме относительных частот вариант, попавших в i-й интервал. Следовательно, *площадь гистограммы относительных частот равна сумме всех относительных частот, т.е. единице.*

Пример 2.2. Имеются результаты регистрации значений затухания сигнала x_i на частоте 1000 Гц коммутируемого канала телефонной сети. Эти значения, измеренные в дБ, в виде вариационного ряда представлены в табл. 2.3. Необходимо построить статистический ряд.

Таблица 2.3

i	1	2	3	4	5	6	7	8	9	10	11
x_i	25,79	25,98	25,98	26,12	26,13	26,49	26,52	26,60	26,66	26,69	26,74
i	12	13	14	15	16	17	18	19	20	21	22
x_i	26,85	26,90	26,91	26,96	27,02	27,11	27,19	27,21	27,28	27,30	27,38
i	23	24	25	26	27	28	29	30	31	32	33
x_i	27,40	27,49	27,64	27,66	27,71	27,78	27,89	27,89	28,01	28,10	28,11
i	34	35	36	37	38	39	40	41	42	43	44
x_i	28,37	28,38	28,50	28,63	28,67	28,90	28,99	28,99	29,03	29,12	29,28

Решение. Количество разрядов статистического ряда следует выбрать минимальным, чтобы обеспечить достаточное количество попаданий в каждый из них, возьмем $k = 6$. Определим размер разряда (длину интервала) $h = \left(x_{\max} - x_{\min}\right)/k = (29{,}28 - 25{,}79)/6 = 0{,}58$.

Сгруппируем наблюдения по разрядам (табл. 2.4).

Таблица 2.4

i	1	2	3	4	5	6
x_i	25,79	26,37	26,95	27,53	28,12	28,70
n_i	5	9	10	9	5	6
$W_i = n_i/n$	0,114	0,205	0,227	0,205	0,114	0,136
W_i/h	0,196	0,353	0,392	0,353	0,196	0,235

На основе статистического ряда построим гистограмму (рис. 2.2) и график эмпирической функции распределения (рис. 2.3).

График эмпирической функции распределения (см. рис. 2.3) отличается от графика, представленного на рис. 2.1, равенством шага изменения варианты и величиной шага приращения функции (при построении по вариационному ряду шаг приращения кратен $1/n$, а по статистическому ряду – зависит от частости в конкретном разряде).

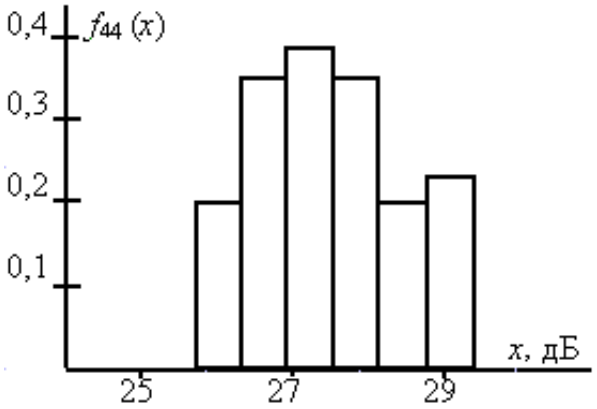

Рис. 2.2. Гистограмма распределения

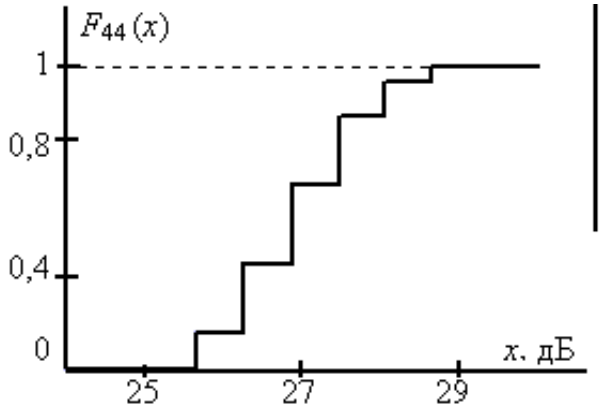

Рис. 2.3. Эмпирическая функция распределения

Рассмотренные представления экспериментальных данных являются исходными для последующей обработки и вычисления различных параметров.

2.2. Оценки параметров распределения и их свойства

Значение параметра, вычисленное по ограниченному объему экспериментальных данных, является случайной величиной, т.е. значение такой величины от выборки к выборке может меняться заранее непредвиденным образом. Следовательно, в результате обработки экспериментальных данных определяется не значение параметра X, а только лишь его приближенное значение – статистическая *оценка* параметра $\overline{X}$. Итак, *статистической оценкой* параметра теоретического распределения называют функцию от наблюдаемых случайных результатов наблюдения, которая и даст приближенное значение искомого параметра:

$$\overline{X} = \left(X_1 + X_2 + \ldots + X_n\right)/n .$$

Различают два вида оценок – точечные и интервальные. *Точечными* называют такие оценки, которые определяются одним числом. При малых

объемах выборки точечные оценки могут значительно отличаться от истинных значений параметров, поэтому их применяют при большом объеме выборки. *Интервальные* оценки задаются двумя числами, определяющими вероятный интервал (диапазон) возможного значения параметра. Эти оценки применяются для малых и для больших выборок, позволяют установить точность и надежность оценок. Рассмотрим вначале точечные оценки.

Применительно к каждому оцениваемому параметру закона распределения генеральной совокупности существует множество функций, позволяющих вычислить искомые значения. Например, оценку математического ожидания можно вычислить, взяв среднее арифметическое выборочных значений, половину суммы крайних членов вариационного ряда, средний член выборки и т.д. Указанные функции отличаются качеством оценок и трудоемкостью реализации.

Качество оценок характеризуется такими свойствами, как состоятельность, несмещенность, эффективность и достаточность [5, 7, 11].

Состоятельность характеризует сходимость по вероятности оценки $\overline{X}$ к истинному значению параметра X при неограниченном увеличении объема выборки *n*. Для состоятельности оценки достаточно, но не обязательно, чтобы математическое ожидание квадрата отклонения оценки от параметра $M\left(X-\overline{X}\right)^2$ стремилось к нулю с увеличением объема выборки (здесь и далее символ *M* означает математическое ожидание). Свойство состоятельности проявляется при неограниченном увеличении *n*, а при небольших объемах экспериментальных данных наличие этого свойства еще недостаточно для применения оценки.

Несмещенность характеризует отсутствие систематических (в среднем) отклонений оценки от параметра при любом конечном, в том числе, и малом, объеме выборки, т.е. $M\left(\overline{X}\right)=X$. Использование статистической оценки, математическое ожидание которой не равно оцениваемому параметру, приводит к систематическим ошибкам. Не всегда наличие смещения плохо. Оно может быть существенно меньше погрешности регистрации значений параметра или давать дополнительную гарантию выполнения требований к значению параметра (если даже при положительном смещении оценка $\overline{X}$ меньше предельно допустимого значения, то несмещенное значение тем более будет отвечать этому условию). В таких ситуациях допустимо применение смещенных оценок, если они вычисляются проще, чем несмещенные. Но даже несмещенная оценка может быть удалена от истинного значения.

Эффективность характеризует разброс случайных значений оценки около истинного значения параметра. Среди всех оценок следует выбрать ту, значения которой теснее сконцентрированы около оцениваемого пара-

метра. Для многих применяемых способов оценивания выборочные распределения параметров асимптотически нормальны, поэтому часто мерой эффективности служит дисперсия оценки. В таком понимании эффективная оценка – это оценка с минимальной дисперсией.

Достаточность характеризует полноту использования информации, содержащейся в выборке. Другими словами, оценка $\overline{X}$ будет достаточной, если все другие независимые оценки на основе данной выборки не дают дополнительной информации об оцениваемом параметре. Эффективная оценка обязательно является и достаточной.

Рассмотренные свойства применимы также и к экспериментальным данным, которые характеризуются многомерными распределениями вероятностей.

Подходы к формированию оценок разработаны в теории несмещенных оценок, предложенной А.Н. Колмогоровым и С. Рао. В данной теории предполагается известным с точностью до параметра T вид функции плотности распределения наблюдаемой величины $f(x, T)$. Вид распределения устанавливается исходя из априорных соображений, например, на основе общепринятых суждений о характере безотказной работы технических средств. Тогда задача сводится к нахождению такой функции от результатов наблюдений, которая дает несмещенную и эффективную оценку.

2.3. Оценки моментов и квантилей распределения

Для характеристики эмпирического распределения можно использовать оценки центральных и начальных моментов. Применение находят моменты до четвертого порядка включительно, так как точность выборочных моментов резко падает с увеличением их порядка, в частности, дисперсия начальных моментов порядка r зависит от моментов порядка $2r$. Она становится значительной для моментов высокого порядка даже при больших объемах выборки. Выборочные значения моментов определяют непосредственно по выборке или по сгруппированным данным [5, 7, 11].

Выборочные значения центральных моментов случайной величины x вычисляются по выборке с применением с формул

$$\mu_1 = \frac{1}{n}\sum_{i=1}^{n} x_i,$$
$$\tilde{\mu}_k = \frac{1}{n}\sum_{i-1}^{n}\left(x_1 - \mu_1\right)^k, \text{ где } k = 2, 3, 4. \qquad (2.1)$$

Указанные величины являются оценками соответствующих теоретических моментов $\mu_1 - \mu_4$ и должны рассматриваться как случайные. Вычис-

ления по формулам (2.1) дают состоятельные, но смещенные оценки моментов старше первого. Смещение удается устранить введением поправочных коэффициентов, зависящих от объема выборки. Несмещенными и состоятельными будут оценки.

$$\mu_2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \mu_1\right)^2,$$

$$\mu_3 = \frac{n^2}{(n-1)(n-2)}\tilde{\mu}_3, \qquad (2.2)$$

$$\mu_4 = \frac{n\left(n^2-2n+3\right)\tilde{\mu}_4 - 3n(2n-3)\tilde{\mu}_2^2}{(n-1)(n-2)(n-3)}.$$

Оценки моментов по сгруппированным экспериментальным данным

$$\tilde{\mu}_{1,g} = \frac{1}{n}\sum_{j=1}^{\psi} n_j X_{\text{ц},j},$$

$$\tilde{\mu}_{k,g} = \frac{1}{n}\sum_{j=1}^{\psi} n_j \left(X_{\text{ц},j} - \tilde{\mu}_{1,g}\right)^k, \text{ где } k = 2, 3, 4, ..., \qquad (2.3)$$

где $X_{\text{ц},j}$ – центр j-го интервала; ψ – количество интервалов.

Группирование и приписывание соответствующей частости значения варианты в середине интервала группирования вносят некоторые искажения. Если распределение непрерывно и имеет достаточно высокий порядок соприкосновения с осью абсцисс (значения функции плотности распределения быстро убывают при удалении от центра распределения), то для снижения ошибок группирования используют поправки Шеппарда. Уточненные значения выборочных моментов для случая равной длины всех интервалов определяются через оценки моментов, вычисленные по сгруппированным данным:

$$\mu_1 = \tilde{\mu}_{1,g};$$

$$\mu_2 = \tilde{\mu}_{2,g} - h^2/12;$$

$$\mu_3 = \tilde{\mu}_{3,g}; \qquad (2.4)$$

$$\mu_4 = \tilde{\mu}_{4,g} - \tilde{\mu}_{2,g}\, h^2/2 + 7h^4/240,$$

где h – длина интервала группирования. Указанные поправки ведут к уточнению только при соблюдении указанного условия, в противном случае они могут привести к еще большей ошибке.

Начальный эмпирический момент порядка r по несгруппированным данным определяется соотношением

$$\eta_r = \frac{1}{n}\sum_{i=1}^{n} x_i^r, \quad r = 1,\ 2,\ 3, \ldots . \qquad (2.5)$$

Центральные и начальные оценки моментов связаны между собой следующими зависимостями:

$$\begin{aligned}
&\mu_1 = \eta_1;\\
&\tilde{\mu}_2 = \eta_2 - \eta_1^2;\\
&\tilde{\mu}_3 = \eta_3 - 3\eta_1\eta_2 + 2\eta_1^3;\\
&\tilde{\mu}_4 = \eta_4 - 4\eta_1\eta_3 + 6\eta_1^2\eta_2 - 3\eta_1^4.
\end{aligned} \qquad (2.6)$$

В процессе обработки экспериментальных данных проще сначала определить оценки начальных моментов, потом перейти к смещенным оценкам центральных моментов и затем вычислить несмещенные оценки.

Квантилью, отвечающей уровню вероятности γ, называют такое значение варианты $x\gamma$, при котором функция распределения случайной величины принимает значение γ, т.е. квантиль – это значение аргумента $x\gamma$ функции распределения, при котором $F(x\gamma) = \gamma$. Эмпирическую квантиль находят по заданному значению вероятности γ, используя вариационный ряд или ступенчатую ломаную линию.

Наряду с указанными параметрами для описания распределений применяются и другие характеристики:

– среднеквадратическое отклонение $\sigma = \sqrt{\mu_2}$;

– коэффициент асимметрии $\beta_1 = \mu_3 / \sigma^3$;

– коэффициент эксцесса $\beta_2 = \mu_4 / \mu_2^2$;

– стандартизованные переменные $u = (x - \mu_1)/\sigma$.

Коэффициент асимметрии характеризует "скошенность" распределения относительно симметричного нормального распределения (у любого симметричного распределения $\beta_1 = 0$) (рис. 2.4). Этот показатель в основном зависит от крайних значений выборки.

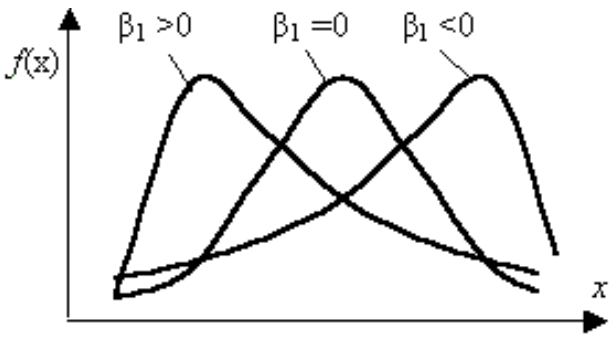

Рис. 2.4. Асимметрия распределения

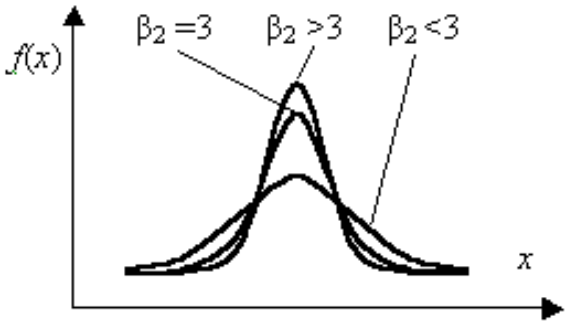

Рис. 2.5. Эксцесс распределения

Коэффициент эксцесса характеризует островершинность распределения относительно нормального распределения (этот коэффициент у нормального распределения равен трем) (рис. 2.5). Термин "эксцесс" (превышение) целесообразно применять не к величине β_2, а к сравнению этой ве-

личины изучаемого распределения с величиной данного коэффициента нормального распределения, т.е. с величиной, равной трем. Исходя из этого, часто вместо β_2 используют величину $\beta_2 - 3$.

Стандартизация переменной позволяет упростить расчеты, кроме того, в литературе многие справочные статистические таблицы приводятся именно для стандартизованных переменных. Нетрудно показать, что математическое ожидание стандартизованной переменной равно нулю, а дисперсия равна единице, т.е. после такого преобразования экспериментальных данных справедливы следующие соотношения:

$$M(u) = \frac{1}{n}\sum_{i=1}^{n} u_i = 0; \quad D(u) = \frac{1}{n}\sum_{i=1}^{n}(u_i - 0)^2 = 1.$$

Величина u называется *центрированной и нормированной*. Переход от центрированной и нормированной величины к исходной осуществляется простым преобразованием $x = u\sigma + \mu_1$. Потери информации при стандартизации и обратном преобразовании не происходит.

Пример 2.3. Необходимо определить числовые характеристики распределения по данным, представленным в виде вариационного и статистического ряда (табл. 2.3 и 2.4), соответственно.

Решение. Вычислим значения центральных моментов по вариационному ряду, пользуясь формулами (2.1):

$$\mu_1 = 27{,}508;\ \tilde{\mu}_2 = 0{,}893;\ \tilde{\mu}_3 = 0{,}123;\ \tilde{\mu}_4 = 1{,}656.$$

Эти оценки, кроме математического ожидания, являются смещенными. Несмещенные оценки получим на основе (2.2):

$$\tilde{\mu}_2 = 0{,}913;\ \tilde{\mu}_3 = 0{,}132;\ \tilde{\mu}_4 = 1{,}819;\ \sigma = 0{,}956.$$

Вычисление оценок моментов на основе статистического ряда по (2.3) дает следующие результаты:

$$\mu_1 = 27{,}482;\ \tilde{\mu}_2 = 0{,}805;\ \tilde{\mu}_3 = 0{,}137;\ \tilde{\mu}_4 = 1{,}656.$$

Судя по гистограмме, по крайней мере, левый край распределения не имеет гладкого соприкосновения с осью x, поэтому поправки Шеппарда нецелесообразны.

Значения оценок моментов различаются при их вычислении по вариационному ряду и по сгруппированным данным. Можно предполагать, что оценки, вычисленные по вариационному ряду, будут точнее оценок, рассчитанных по статистическому ряду.

Оценка коэффициента асимметрии $\beta_1 = 0{,}132/0{,}913^{1{,}5} = 0{,}15$ говорит о небольшой положительной асимметрии распределения (мода функции плотности распределения находится левее математического ожидания), а оценка коэффициента эксцесса $\beta_2 = 1{,}819/0{,}913^2 = 2{,}18$ – о пологости распре-

деления («островершинность» выражена слабее, чем у нормального распределения).

Анализируя назначение рассмотренных параметров, необходимо отметить следующее. Одни параметры характеризует средние величины, а другие – вариацию. *Главное назначение средних величин (оценок начальных моментов и, в первую очередь, первого момента распределения) состоит в их обобщающей функции.* Это обобщение позволяет заменить множество различных индивидуальных значений показателя средней величиной, характеризующей всю однородную совокупность. Иначе говоря, средняя величина является типической характеристикой варианты в конкретной выборке. Иногда средняя величина обобщает и неоднородные показатели однотипных объектов.

Каждый элемент экспериментальных данных формируется под влиянием как общих закономерностей, так и особых условий и случайных событий. Следовательно, в обработке экспериментальных данных большой интерес представляют вопросы оценки величин, характеризующих вариацию значений параметра у разных объектов или у одного и того же объекта в разные моменты времени. *Вариацией какого-либо параметра (показателя) в совокупности наблюдений называется различие его значений у разных элементов этой совокупности.* Именно это свойство является объектом исследования большинства методов обработки экспериментальных данных. Для характеристики вариации нет единого показателя, в этих целях применяются моменты распределения выше первого, производные от них величины, размах выборки, квантили и др.

Задачи.

2.1. Составить вариационный ряд, график эмпирической функции распределения, построить полигон частот для следующих значений длины случайным образом отобранных заготовок: 39, 41, 40, 43, 41, 44, 43, 41, 41, 42, 43, 39, 40, 42, 44, 41, 42, 38, 42, 41, 41, 42, 40, 40, 43, 41, 38, 39, 41, 42.

2.2. Составить вариационный ряд, график эмпирической функции распределения, построить полигон частот для следующих значений отклонений диаметра ролика от номинального при измерении микрометром с ценой деления 0,01 мм:

0,08;	0,09;	0,03;	0,11;	0,05;	0,09;	0,06;	0,11;	0,09;	0,04;
0,03;	0,08;	0,07;	0,01;	0,12;	0,05;	0,09;	0,06;	0,06;	0,06;
0,04;	0,06;	0,08;	0,13;	0,07;	0,11;	0,05;	0,07;	0,07;	0,06.

2.3. Для задач 1 и 2 определить числовые характеристики распределения: значения центральных моментов по вариационному ряду по формулам (2.1); несмещенные и состоятельные оценки по формулам (2.2); оценки моментов по сгруппированным экспериментальным данным по (2.3); коэффициент асимметрии и коэффициент эксцесса.

3. СТАТИСТИЧЕСКАЯ ПРОВЕРКА СТАТИСТИЧЕСКИХ ГИПОТЕЗ

3.1. Сущность задачи проверки статистических гипотез

Часто необходимо знать закон распределения генеральной совокупности. Если закон распределения неизвестен, но имеются основания предположить, что он имеет определенный вид (назовем его A): выдвигают гипотезу: генеральная совокупность распределена по закону A. Таким образом, в этой гипотезе речь идет о *виде предполагаемого распределения*.

Возможен случай, когда закон распределения известен, а его параметры неизвестны. Если есть основания предположить, что неизвестный параметр θ равен определенному значению θ_0, выдвигают гипотезу: $\theta = \theta_0$. Таким образом, в этой гипотезе речь идет о *предполагаемой величине параметра* известного распределения.

Возможны и другие гипотезы: о равенстве параметров двух или нескольких распределений, о независимости выборок и многие другие.

Статистической называют гипотезу о виде неизвестного распределения, или о параметрах известных распределений [5, 7, 14].

Например, статистическими гипотезами являются гипотезы:

1) генеральная совокупность распределена по экспоненциальному закону;

2) математические ожидания двух экспоненциально распределенных выборок равны друг другу.

В первой из них высказано предположение о виде закона распределения, а во второй – о параметрах двух известных распределений.

Гипотезы, в основе которых нет никаких допущений о конкретном виде закона распределения, называют *непараметрическими*, в противном случае – *параметрическими*.

Наряду с выдвинутой гипотезой рассматривают и противоречащую ей гипотезу. Если выдвинутая гипотеза будет отвергнута, то имеет место противоречащая гипотеза. По этой причине эти гипотезы целесообразно различать.

Нулевой (основной) называют выдвинутую гипотезу H_0. Наряду с основной гипотезой рассматривают и *альтернативную* (*конкурирующую*, *противоречащую*) ей гипотезу H_1. И если нулевая гипотеза будет отвергнута, то будет иметь место альтернативная гипотеза.

Например, если нулевая гипотеза состоит в предположении, что математическое ожидание m нормального распределения равно 10, то конкурирующая гипотеза, в частности, может состоять в предположении, что $m \neq 10$. Запись гипотез выглядит так: $H_0: m = 10$; $H_1: m \neq 10$.

Различают гипотезы, содержащие одно либо более одного предположений. Гипотезу называют *простой*, если содержит только одно предположение. Например, если λ является параметром экспоненциального распределения, то гипотеза H_0: $\lambda = 5$ – простая гипотеза. Гипотеза H_0: математическое ожидание нормального распределения равно 3 (σ известно) – простая.

Сложной называют гипотезу, которая состоит из конечного или бесконечного множества простых гипотез. Сложная гипотеза H_0: $\lambda > 5$ состоит из бесконечного множества простых гипотез H_i: $\lambda = b_i$, где b_i – любое число, большее 5. Гипотеза H_0: математическое ожидание нормального распределения равно 3 (σ неизвестно) – сложная. Сложной гипотезой будет предположение о распределении случайной величины X по нормальному закону, если не фиксируются конкретные значения математического ожидания и дисперсии.

Проверка нулевой гипотезы основывается на вычислении некоторой случайной величины – критерия, точное или приближенное распределение которого известно. Обозначают эту величину через U или Z, если она распределена нормально; F или v^2 – если она распределена по закону Фишера – Снедекора; T – по закону Стьюдента; χ^2 – по закону хи-квадрат Пирсона и т.д.

В общем случае, когда не принимается во внимание вид распределения, *статистическим критерием* (или просто *критерием*) назовем случайную величину K, которая служит для проверки нулевой гипотезы.

Например, если проверяют гипотезу о равенстве дисперсий двух нормальных генеральных совокупностей, то в качестве критерия K принимают отношение исправленных выборочных дисперсий:

$$K = F = s_1^2 / s_2^2 .$$

Это величина случайная, потому что в различных опытах дисперсии принимают различные, наперед неизвестные значения, и распределена по закону Фишера – Снедекора.

Для проверки гипотезы по данным выборок вычисляют частные значения входящих в критерий величин и таким образом получают частное (наблюдаемое) значение критерия.

Наблюдаемым значением $K_{\text{набл}}$ называют значение критерия, вычисленное по выборкам. Например, если по двум выборкам найдены исправленные выборочные дисперсии $s_1^2 = 20$ и $s_2^2 = 5$, то наблюдаемое значение критерия F будет равно: $F_{\text{набл}} = s_1^2 / s_2^2 = 20/5 = 4$.

После выбора определенного критерия множество всех его возможных значений разбивают на два непересекающихся подмножества: одно из них

содержит значение критерия, при которых нулевая гипотеза отвергается, а другая – при которых она принимается.

Критической областью называют совокупность значений критерия, при которых нулевую гипотезу отвергают.

Областью принятия гипотезы (областью допустимых значений) называют совокупность значений критерия, при которых гипотезу принимают.

Основной принцип проверки статистических гипотез сформулирован так: если наблюдаемое значение критерия принадлежит критической области – гипотезу отвергают, если наблюдаемое значение критерия принадлежит области принятия гипотезы – гипотезу принимают.

Выдвинутая гипотеза может быть правильной, либо неправильной, поэтому ее проверяют, а так как проверку производят статистическими методами, ее называют *статистической.* В итоге статистической проверки гипотезы в двух случаях может быть принято неправильное решение, т.е. могут быть допущены ошибки двух родов.

Ошибка первого рода возникает с вероятностью α тогда, когда отвергается верная гипотеза H_0 и принимается конкурирующая гипотеза H_1.

Ошибка второго рода возникает с вероятностью β в том случае, когда принимается неверная гипотеза H_0, в то время как справедлива конкурирующая гипотеза H_1.

Правильное решение может быть принято также в двух случаях: 1) гипотеза принимается, причем и в действительности она правильная; 2) гипотеза отвергается, причем и в действительности она неверная.

Вероятность совершить ошибку первого рода принято обозначать через α; ее называют также *уровнем значимости*. Как правило, уровень значимости принимают равным 0,05 или 0,01, что означает, соответственно, что пяти (или одном) случаях из ста имеется риск допустить ошибку первого рода (отвергнуть правильную гипотезу).

Доверительная вероятность, $1-\alpha$, – это вероятность не совершить ошибку первого рода и принять верную гипотезу H_0. Вероятность отвергнуть ложную гипотезу H_0 называется *мощностью критерия*, $1-\beta$. Следовательно, при проверке гипотезы возможны четыре варианта исходов (табл. 3.1).

Таблица 3.1

Гипотеза H_0	Решение	Вероятность	Примечание
Верна	Принимается	$1-\alpha$	Доверительная вероятность
	Отвергается	α	Вероятность ошибки первого рода
Неверна	Принимается	β	Вероятность ошибки второго рода
	Отвергается	$1-\beta$	Мощность критерия

Например, рассмотрим случай, когда некоторая несмещенная оценка параметра θ вычислена по выборке объема n, и эта оценка имеет плотность распределения $f(\theta)$ (рис. 3.1).

Рис. 3.1. Области принятия и отклонения гипотезы

Предположим, что истинное значение оцениваемого параметра равно T. Если рассматривать гипотезу H_0: $\theta = T$, то насколько велико должно быть различие между θ и T, чтобы эту гипотезу отвергнуть? Ответить на данный вопрос можно в статистическом смысле, рассматривая вероятность достижения некоторой заданной разности между θ и T на основе выборочного распределения параметра θ.

Целесообразно полагать одинаковыми значения вероятности выхода параметра θ за нижний и верхний пределы интервала. Такое допущение во многих случаях позволяет минимизировать доверительный интервал, т.е. повысить мощность критерия проверки. Суммарная вероятность того, что параметр θ выйдет за пределы интервала с границами $\theta_{1-\alpha/2}$ и $\theta_{\alpha/2}$, составляет величину α. Эту величину следует выбрать настолько малой, чтобы выход за пределы интервала был маловероятен. Если оценка параметра попала в заданный интервал, то в таком случае нет оснований подвергать сомнению проверяемую гипотезу, следовательно, гипотезу равенства $\theta = T$ можно принять. Но если после получения выборки окажется, что оценка выходит за установленные пределы, то в этом случае есть серьезные основания отвергнуть гипотезу H_0. Отсюда следует, что вероятность допустить ошибку первого рода равна α (равна уровню значимости критерия).

Если предположить, например, что истинное значение параметра в действительности равно $T + d$, то согласно гипотезе H_0: $\theta = T$ – вероятность того, что оценка параметра θ попадет в область принятия гипотезы, составит β (рис. 3.2).

При заданном объеме выборки вероятность совершения ошибки первого рода можно уменьшить, снижая уровень значимости α. Однако при этом увеличивается вероятность ошибки второго рода β (снижается мощность критерия). Аналогичные рассуждения можно провести для случая, когда истинное значение параметра равно $T - d$.

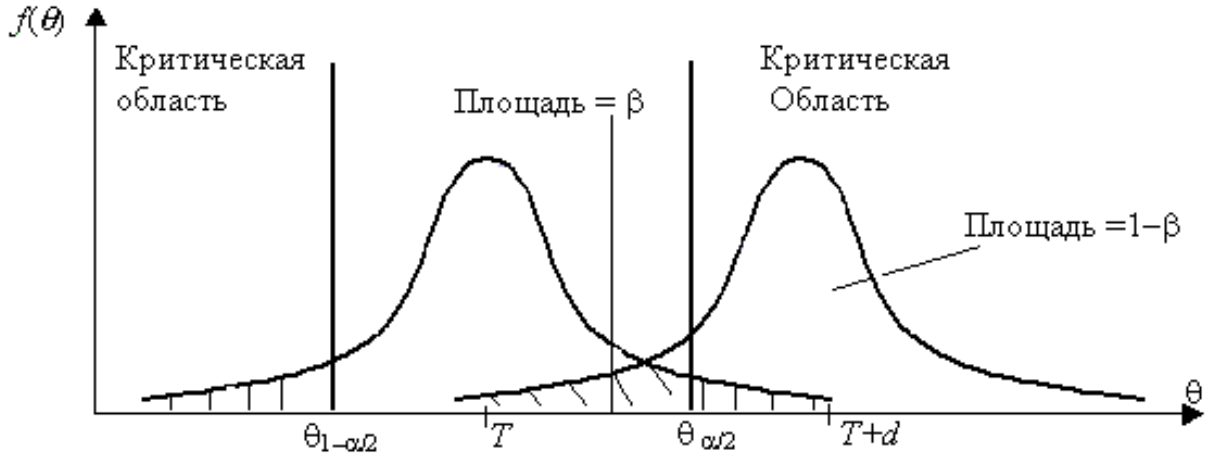

Рис. 3.2

Единственный способ уменьшить обе вероятности – состоит в увеличении объема выборки (плотность распределения оценки параметра при этом становится более «узкой»). При выборе критической области руководствуются правилом Неймана – Пирсона: следует так выбирать критическую область, чтобы вероятность α была мала, если гипотеза верна, и велика – в противном случае. Однако выбор конкретного значения α относительно произволен. Употребительные значения лежат в пределах от 0,001 до 0,2. В целях упрощения ручных расчетов составлены таблицы интервалов с границами $\theta_{1-\alpha/2}$ и $\theta_{\alpha/2}$ для типовых значений α и различных способов построения критерия.

При выборе уровня значимости необходимо учитывать мощность критерия при альтернативной гипотезе. Иногда большая мощность критерия оказывается существеннее малого уровня значимости, и его значение выбирают относительно большим, например 0,2. Такой выбор оправдан, если последствия ошибок второго рода более существенны, чем ошибок первого рода. Например, если отвергнуто правильное решение «продолжать строительство жилого дома», то ошибка первого рода повлечет материальный ущерб; если же принято неправильное решение «продолжать строительство», несмотря на опасность обвала грунта, то такая ошибка второго рода может повлечь за собой гибель людей.

Кстати, при контроле качества продукции вероятность признать негодной партию годных изделий называют «риском производителя», а вероятность принять негодную партию – «риском потребителя».

В зависимости от сущности проверяемой гипотезы и используемых мер расхождения оценки характеристики от ее теоретического значения применяют различные критерии.

К числу наиболее часто применяемых критериев для проверки гипотез о законах распределения относят критерии хи-квадрат Пирсона, Колмогорова, Мизеса, Вилкоксона, о значениях параметров – критерии Фишера, Стьюдента.

3.2. Типовые распределения

При проверке гипотез широкое применение находит ряд теоретических законов распределения. Наиболее важным из них является нормальное распределение. С ним связаны распределения хи-квадрат, Стьюдента, Фишера, а также интеграл вероятностей. Для указанных законов функции распределения аналитически не представимы. Значения функций определяются по таблицам или с использованием стандартных процедур пакетов прикладных программ. Указанные таблицы обычно построены в целях удобства проверки статистических гипотез в ущерб теории распределений – они содержат не значения функций распределения, а критические значения аргумента $z(\alpha)$.

Для односторонней критической области $z(\alpha)=z_{1-\alpha}$, т.е. критическое значение аргумента $z(\alpha)$ соответствует квантили $z_{1-\alpha}$ уровня $1-\alpha$, так как $\int\limits_{z(\alpha)}^{\infty} f(z)dz=\alpha=1-\int\limits_{-\infty}^{z(\alpha)} f(z)dz$, рис. 3.3.

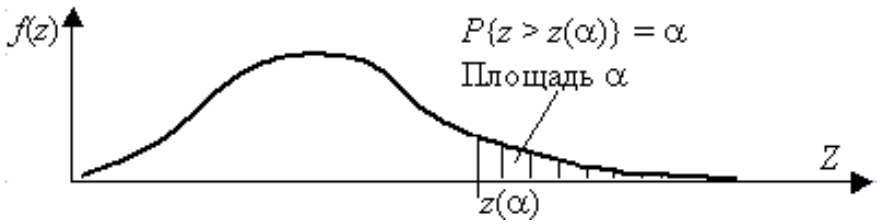

Рис. 3.3. Односторонняя критическая область

Для двусторонней критической области, с уровнем значимости α, размер левой области $\alpha/2$, правой $\alpha/2$, ($\alpha/2+\alpha/2=\alpha$), рис. 3.4.

Значения $-z(\alpha/2)$ и $z(\alpha/2)$ связаны с квантилями распределения соотношениями $-z(\alpha/2)=z_{1-\alpha/2}$, $z(\alpha/2)=z_{\alpha/2}$, так как

$$\int\limits_{-\infty}^{z(\alpha)} f(z)dz=1-\alpha=1-\int\limits_{z(\alpha)}^{\infty} f(z)dz.$$

Для симметричной функции плотности распределения $f(z)$ критическую область выбирают из условия $\alpha_1 = \alpha_2 = \alpha/2$ (обеспечивается наибольшая мощность критерия). В таком случае левая и правая границы будут равны $|z(\alpha/2)|$.

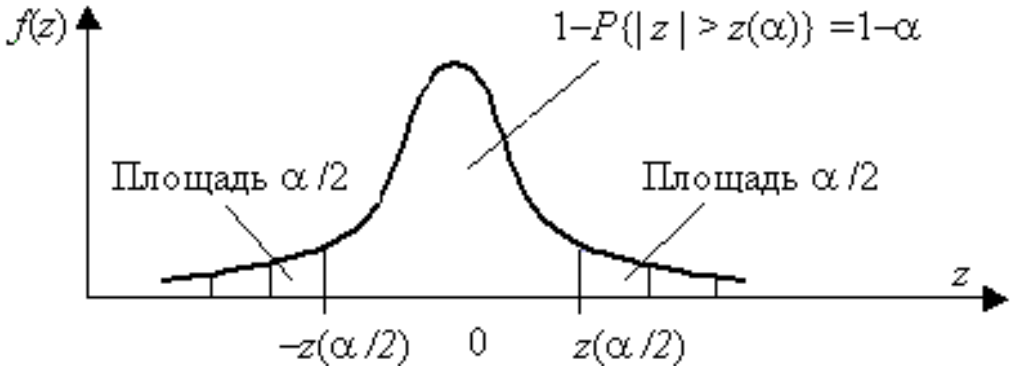

Рис. 3.4. Двусторонняя критическая область

Нормальное распределение

Этот вид распределения является наиболее важным в связи с центральной предельной теоремой теории вероятностей: распределение суммы независимых случайных величин стремится к нормальному с увеличением их количества при произвольном законе распределения отдельных слагаемых, если слагаемые обладают конечной дисперсией. Так как реальные физические явления часто представляют собой результат суммарного воздействия многих факторов, то в таких случаях нормальное распределение является хорошим приближением наблюдаемых значений. Функция плотности нормального распределения

$$f(x) = \frac{1}{\sqrt{2\pi m_2}} \exp\left[\frac{-(x-m_1)^2}{2m_2}\right] \qquad (3.1)$$

– унимодальная, симметричная, аргумент x может принимать любые действительные значения (рис. 3.5):

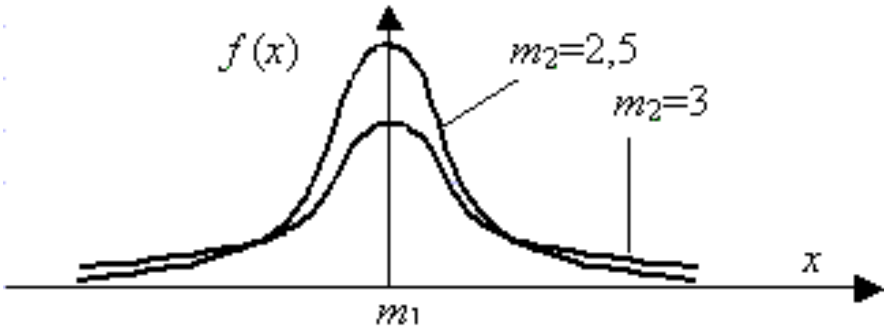

Рис. 3.5. Плотность нормального распределения

Функция плотности нормального распределения стандартизованной величины u имеет вид

$$f(u)=\frac{1}{\sqrt{2\pi}}\exp\left[\frac{-u^2}{2}\right].$$

Вычисление значений функции распределения $\Phi(u)$ для стандартизованного неотрицательного аргумента u $(u\geq 0)$ можно произвести с помощью полинома наилучшего приближения:

$$\Phi(u)=1-0,5\left(1+0,196854u+0,115194u^2+0,000344u^3+0,019527u^4\right)^{-4}. \quad (3.2)$$

Такая аппроксимация обеспечивает абсолютную ошибку не более 0,00025. Для вычисления $\Phi(u)$ в области отрицательных значений стандартизованного аргумента u $(u<0)$ следует воспользоваться свойством симметрии нормального распределения $\Phi(u)=1-\Phi(-u)$.

Иногда в справочниках вместо значений функции $\Phi(u)$ приводят значения *интеграла вероятностей*

$$F(u)=\frac{1}{\sqrt{2\pi}}\int_0^u \exp\left(-x^2/2\right)dx,\ u>0. \quad (3.3)$$

Интеграл вероятностей связан с функцией нормального распределения соотношением $\Phi(u)=0,5+F(u)$.

Распределение хи-квадрат

Распределению хи-квадрат (χ^2-распределению) с k степенями свободы соответствует распределение суммы $\chi^2=\sum_{i=1}^{n}u_i^2$ квадратов n стандартизованных случайных величин u_i, каждая из которых распределена по нормальному закону, причем k из них независимы, $n\geq k$. Функция плотности распределения хи-квадрат с k степенями свободы:

$$f(x)=\left[2^{k/2}\Gamma(k/2)\right]^{-1}(x)^{k/2-1}e^{-x/2},\ x\geq 0, \quad (3.4)$$

где $x=\chi^2$, $\Gamma(k/2)$ – гамма-функция.

Число степеней свободы k определяет количество независимых слагаемых в выражении для χ^2. Функция плотности при k, равном одному или двум, – монотонная, а при $k>2$ – унимодальная, несимметричная (рис. 3.6).

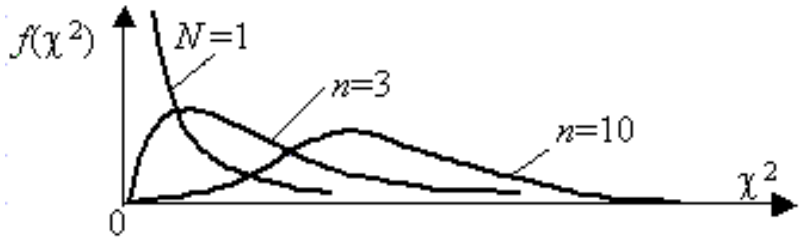

Рис. 3.6. Плотность распределения хи-квадрат

Математическое ожидание и дисперсия величины χ^2 равны соответственно k и $2k$. Распределение хи-квадрат является частным случаем более общего гамма-распределения, а величина, равная корню квадратному из хи-квадрат с двумя степенями свободы, подчиняется распределению Рэлея.

С увеличением числа степеней свободы ($k > 30$) распределение хи-квадрат приближается к нормальному распределению с математическим ожиданием k и дисперсией $2k$. В таких случаях критическое значение $\chi^2(k;\ \alpha) \approx u_{1-\alpha}(k,\ 2k)$, где $u_{1-\alpha}(k,\ 2k)$ – квантиль нормального распределения. Погрешность аппроксимации не превышает нескольких процентов.

Распределение Стьюдента

Распределение Стьюдента (t-распределение, предложено в 1908 г. английским статистиком В. Госсетом, публиковавшим научные труды под псевдонимом Student), – характеризует распределение случайной величины $t = \dfrac{u_0}{\sqrt{\left(u_1^2 + u_2^2 + \ldots + u_k^2\right)/k}}$, где $u_0,\ u_1,\ \ldots,\ u_k$ – взаимно независимые нормально распределенные случайные величины с нулевым средним и конечной дисперсией. Аргумент t не зависит от дисперсии слагаемых. Функция плотности распределения Стьюдента:

$$f(t) = \frac{\Gamma\left[(k+1)/2\right]}{\sqrt{\pi k}\,\Gamma(k/2)}\left[1 + \frac{t^2}{k}\right]^{-(k+1)/2}. \qquad (3.5)$$

Величина k характеризует количество степеней свободы. Плотность распределения – унимодальная и симметричная функция, похожая на нормальное распределение (рис. 3.7):

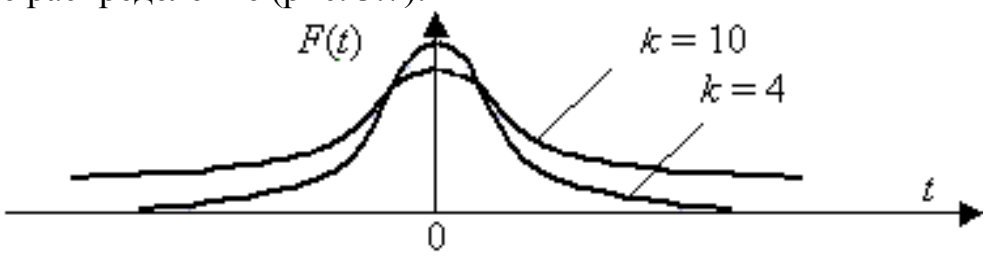

Рис. 3.7. Плотность распределения Стьюдента

Область изменения аргумента t от $-\infty$ до ∞. Математическое ожидание и дисперсия равны 0 и $k/(k-2)$, соответственно, при $k>2$. По сравнению с нормальным, распределение Стьюдента более пологое, оно имеет меньшую дисперсию. Это отличие заметно при небольших значениях k, что следует учитывать при проверке статистических гипотез (критические значения аргумента распределения Стьюдента превышают аналогичные показатели нормального распределения). Таблицы распределения содержат значения для односторонней $\int_{t(k;\,\alpha)}^{\infty} f(t)\,dt=\alpha$ или двусторонней $\int_{-t(k;\,\alpha)}^{t(k;\,\alpha)} f(t)\,dt=\alpha$ критической области.

Распределение Стьюдента применяется для описания ошибок выборки при $k\le 30$. При $k>100$ данное распределение практически соответствует нормальному, для $30<k<100$ различия между распределением Стьюдента и нормальным распределением составляют несколько процентов. Поэтому, относительно оценки ошибок, малыми считаются выборки объемом не более 30 единиц, большими – объемом более 100 единиц. При аппроксимации распределения Стьюдента нормальным распределением для односторонней критической области вероятность

$$P\{t>t(k;\,\alpha)\}=u_{1-\alpha}\left(0,\ k/(k-2)\right),$$

где $u_{1-\alpha}\left(0,\ k/(k-2)\right)$ – квантиль нормального распределения. Аналогичное соотношение можно составить и для двусторонней критической области.

Распределение Фишера

Распределению Р.А. Фишера (F-распределению Фишера – Снедекора) подчиняется случайная величина $x=\left[(y_1/k_1)/(y_2/k_2)\right]$, равная отношению двух случайных величин y_1 и y_2, имеющих хи-квадрат распределение с k_1 и k_2 степенями свободы. Область изменения аргумента x от 0 до ∞. Плотность распределения

$$f(x)=\left[\frac{k_1}{k_2}\right]^{k_1/2}\frac{\Gamma\left(\dfrac{k_1+k_2}{2}\right)}{\Gamma(k_1/2)\,\Gamma(k_2/2)}x^{(k_1-2)/2}\left(1+\frac{k_1}{k_2}x\right)^{-(k_1+k_2)/2}. \qquad (3.6)$$

В этом выражении k_1 обозначает число степеней свободы величины y_1 с большей дисперсией, k_2 – число степеней свободы величины y_2 с

меньшей дисперсией. Плотность распределения – унимодальная, несимметричная (рис. 3.8):

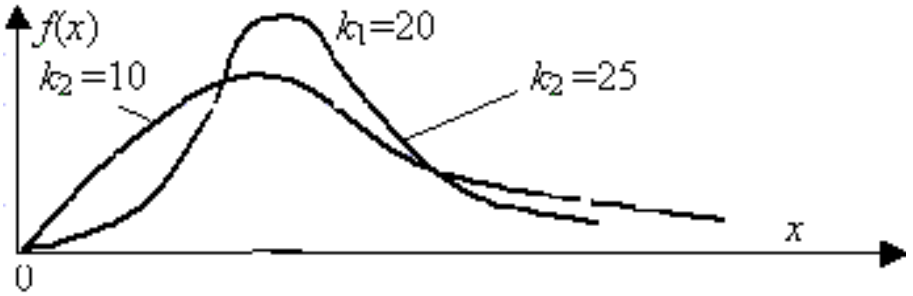

Рис. 3.8. Плотность распределения Фишера

Математическое ожидание случайной величины x равно $k_2/(k_2-2)$ при $k_2>2$, дисперсия $m_2=\left[2k_2^2(k_1+k_2-2)\right]/\left[k_1(k_2-2)^2(k_2-4)\right]$ при $k_2>4$. При $k_1>30$ и $k_2>30$ величина x распределена приближенно нормально с центром $(k_1-k_2)/(2k_1k_2)$ и дисперсией $(k_1+k_2)/(2k_1k_2)$.

3.3. Проверка гипотез о законе распределения

Обычно сущность проверки гипотезы о законе распределения экспериментальных данных заключается в следующем. Имеется выборка экспериментальных данных фиксированного объема, выбран или известен вид закона распределения генеральной совокупности. Необходимо оценить по этой выборке параметры закона, определить степень согласованности экспериментальных данных и выбранного закона распределения, в котором параметры заменены их оценками. Пока не будем касаться способов нахождения оценок параметров распределения, а рассмотрим только вопрос проверки согласованности распределений с использованием наиболее употребительных критериев.

Критерий хи-квадрат К. Пирсона

Использование этого критерия основано на применении такой меры (статистики) расхождения между теоретическим $F(x)$ и эмпирическим распределением $F_n(x)$, которая приближенно подчиняется закону распределения χ^2. Гипотеза H_0 о согласованности распределений проверяется путем анализа распределения этой статистики. Применение критерия требует построения статистического ряда.

Итак, пусть выборка представлена статистическим рядом с количеством разрядов ψ. Наблюдаемая частота попаданий в i-й разряд равна n_i. В соответствии с теоретическим законом распределения ожидаемая частота попаданий в i-й разряд составляет F_i. Разность между наблюдаемой и

ожидаемой частотой составит величину $(n_i - F_i)$. Для нахождения общей степени расхождения между $F(x)$ и $F_n(x)$ необходимо подсчитать взвешенную сумму квадратов разностей по всем разрядам статистического ряда

$$\chi^2 = \sum_{i=1}^{\psi} \frac{(n_i - F_i)^2}{F_i}. \qquad (3.7)$$

Величина χ^2 при неограниченном увеличении n имеет распределение хи-квадрат (асимптотически распределена как хи-квадрат). Это распределение зависит от числа степеней свободы k, т.е. количества независимых значений слагаемых в выражении (3.7). Число степеней свободы равно числу ψ минус число линейных связей, наложенных на выборку. Одна связь существует в силу того, что любая частота может быть вычислена по совокупности частот в оставшихся $\psi - 1$ разрядах. Кроме того, если параметры распределения неизвестны заранее, то имеется еще одно ограничение, обусловленное подгонкой распределения к выборке. Если по выборке определяются ϕ параметров распределения, то число степеней свободы составит $k = \psi - \phi - 1$.

Область принятия гипотезы H_0 определяется условием $\chi^2 \le \chi^2(k; \alpha)$, где $\chi^2(k; \alpha)$ – критическая точка распределения хи-квадрат с уровнем значимости α. Вероятность ошибки первого рода равна α, вероятность ошибки второго рода четко определить нельзя, потому что существует бесконечно большое множество различных способов несовпадения распределений. Мощность критерия зависит от количества разрядов и объема выборки. Критерий рекомендуется применять при $n > 200$, допускается применение при $n > 40$, именно при таких условиях критерий состоятелен (как правило, отвергает неверную нулевую гипотезу).

Пример 3.1. Проверить с помощью критерия хи-квадрат гипотезу о нормальности распределения случайной величины, представленной статистическим рядом в табл. 2.4 при уровне значимости $\alpha = 0,05$.

Решение. В примере 2.3 были вычислены значения оценок моментов: $\mu_1 = 27,51$, $\mu_2 = 0,91$, $\sigma = 0,96$. На основе табл. 2.4 построим табл. 3.2, иллюстрирующую расчеты.

Таблица 3.2

Номер интервала, i	1	2	3	4	5	6
n_i	5	9	10	9	5	6
x_i	26,37	26,95	27,53	28,12	28,70	∞
$F(x_i)$	0,117	0,280	0,508	0,737	0,892	1
ΔF_i	0,117	0,166	0,228	0,228	0,155	0,108
F_i	5,148	7,304	10,032	10,032	6,820	4,752
$(n_i - F_i)^2/F_i$	0,004	0,394	0,0001	0,1062	0,486	0,328

В этой таблице:

n_i – частота попаданий элементов выборки в i-й интервал;

x_i – верхняя граница i-го интервала;

$F(x_i)$ – значение функции нормального распределения;

ΔF_i – теоретическое значение вероятности попадания случайной величины в i-й интервал:

$$\Delta F_i = \frac{1}{\sigma\sqrt{2\pi}} \int_{x_{i-1}}^{x_i} \exp\left(-\frac{1}{2\sigma^2}(x-\mu_1)^2 \right) = F_x(x_i) - F_x(x_{i-1}) =$$

$$= \frac{1}{\sigma\sqrt{2\pi}} \int_{\infty}^{x_i} \exp\left(-\frac{1}{2\sigma^2}(x-\mu_1)^2 \right) - \frac{1}{\sigma\sqrt{2\pi}} \int_{\infty}^{x_{i-1}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu_1)^2 \right);$$

$F_i = \Delta F_i \cdot n$ – теоретическая частота попадания случайной величины в i-й интервал;

$(n_i - F_i)^2 / F_i$ – взвешенный квадрат отклонения.

Для нормального закона возможные значения случайной величины лежат в диапазоне от –∞ до ∞, поэтому при расчетах оценок вероятностей крайний левый и крайний правый интервалы расширяются до –∞ и ∞ соответственно. Вычислить значения функции нормального распределения можно, воспользовавшись стандартными функциями табличного процессора или полиномом наилучшего приближения.

Сумма взвешенных квадратов отклонения $\chi^2 = 1{,}32$. Число степеней свободы $k = 6 - 1 - 2 = 3$ (уклонения связаны линейным соотношением $\sum_{i=1}^{6}(n_i - F_i) = 0$, кроме того, на уклонения наложены еще две связи, так как по выборке были определены два параметра распределения). Критическое значение $\chi^2(3;\ 0{,}05) = 7{,}815$ определяется по табл. П.3 приложения. Поскольку соблюдается условие $\chi^2 \le \chi^2\ (3;\ 0{,}05)$, то полученный результат нельзя считать значимым, и гипотеза о нормальном распределении генеральной совокупности не противоречит экспериментальным данным.

Критерий А.Н. Колмогорова

Для применения критерия А.Н. Колмогорова экспериментальные данные требуется представить в виде вариационного ряда (экспериментальные данные недопустимо объединять в разряды). В качестве меры расхождения между теоретической $F(x)$ и эмпирической $F_n(x)$ функциями распреде-

ления непрерывной случайной величины X используется модуль максимальной разности

$$d_n = \max\left|F(x) - F_n(x)\right|. \qquad (3.8)$$

Колмогоров А.Н. доказал, что какова бы ни была функция распределения $F(x)$ величины X при неограниченном увеличении количества наблюдений n функция распределения случайной величины $d_n\sqrt{n}$ асимптотически приближается к функции распределения $K(\lambda) = P\left(d\sqrt{n} < \lambda\right) = \sum_{k=-\infty}^{\infty} (-1)^k \exp\left(-2k^2\lambda^2\right)$. Иначе говоря, критерий А.Н. Колмогорова характеризует вероятность того, что величина $d_n\sqrt{n}$ не будет превосходить параметр λ для любой теоретической функции распределения. Уровень значимости α выбирается из условия $P\left(d_m\sqrt{n} > \lambda\right) = \alpha = 1 - K(\lambda)$, в силу предположения, что почти невозможно получить это равенство, когда существует соответствие между функциями $F(x)$ и $F_n(x)$. Критерий А.Н. Колмогорова позволяет проверить согласованность распределений по малым выборкам, он проще критерия хи-квадрат, поэтому его часто применяют на практике. Но требуется учитывать два обстоятельства.

Во-первых, в точном соответствии с условиями его применения необходимо пользоваться следующим соотношением

$$d_n = \max\left(d_n^+,\ d_n^-\right),$$

где $d_n^+ = \underbrace{\max}_{1\le i\le n}\left|\frac{i}{n} - F(x_i)\right|$; $\quad d_n^- = \underbrace{\max}_{1\le i\le n}\left|F(x_i) - \frac{i-1}{n}\right|$.

Во-вторых, условия применения критерия предусматривают, что теоретическая функция распределения известна полностью (известны вид функции и ее параметры). Но на практике параметры обычно неизвестны и оцениваются по экспериментальным данным. Это приводит к завышению значения вероятности соблюдения нулевой гипотезы, т.е. повышается риск принять в качестве правдоподобной гипотезу, которая плохо согласуется с экспериментальными данными (повышается вероятность совершить ошибку второго рода). В качестве меры противодействия такому выводу следует увеличить уровень значимости α, приняв его равным 0,1-0,2, что приведет к уменьшению зоны допустимых отклонений.

Пример 3.2. Проверить с помощью критерия А.Н. Колмогорова гипотезу о том, что экспериментальные данные, представленные в табл. 2.3, подчиняются нормальному распределению при уровне значимости $\alpha = 0{,}1$.

Решение. Исходные данные и результаты вычислений сведены в табл. 3.3. Необходимые вычисления можно провести с использованием табличного процессора: значение эмпирической функции распределения $F_n(x_i) = i/44$; значения теоретической функции $F(x_i)$ – это значение функции нормального распределения в точке x_i.

Таблица 3.3

i	1	2	3	4	5	6	7	8	9	10	11
x_i	25,79	25,98	25,98	26,12	26,13	26,49	26,52	26,60	26,66	26,69	26,74
$F_n(x_i)$	0,023	0,046	0,068	0,091	0,114	0,136	0,159	0,182	0,204	0,227	0,250
$F(x_i)$	0,036	0,055	0,055	0,073	0,075	0,144	0,151	0,170	0,188	0,196	0,211
d_n^+	0,014	0,009	0,013	0,018	0,038	0,008	0,008	0,012	0,016	0,032	0,039
d_n^-	0,036	0,032	0,010	0,005	0,016	0,031	0,014	0,011	0,006	0,009	0,016
i	12	13	14	15	16	17	18	19	20	21	22
x_i	26,85	26,90	26,91	26,96	27,02	27,11	27,19	27,21	27,28	27,30	27,38
$F_n(x_i)$	0,273	0,296	0,318	0,341	0,364	0,386	0,409	0,432	0,455	0,477	0,500
$F(x_i)$	0,246	0,263	0,267	0,284	0,305	0,337	0,371	0,378	0,406	0,412	0,447
d_n^+	0,027	0,032	0,051	0,057	0,059	0,050	0,038	0,054	0,049	0,065	0,053
d_n^-	0,004	0,010	0,028	0,034	0,036	0,027	0,015	0,031	0,026	0,042	0,031
i	23	24	25	26	27	28	29	30	31	32	33
x_i	27,40	27,49	27,64	27,66	27,71	27,78	27,89	27,89	28,01	28,10	28,11
$F_n(x_i)$	0,523	0,546	0,568	0,591	0,614	0,636	0,659	0,682	0,705	0,727	0,750
$F(x_i)$	0,456	0,492	0,555	0,561	0,583	0,610	0,656	0,656	0,701	0,731	0,735
d_n^+	**0,067**	0,053	0,013	0,030	0,031	0,026	0,003	0,026	0,003	0,004	0,015
d_n^-	0,044	0,031	0,010	0,007	0,008	0,003	0,019	0,003	0,020	0,027	0,008
i	34	35	36	37	38	39	40	41	42	43	44
x_i	28,37	28,38	28,50	28,63	28,67	28,90	28,99	28,99	29,03	29,12	29,28
$F_n(x_i)$	0,773	0,795	0,818	0,841	0,864	0,886	0,909	0,932	0,955	0,977	1,000
$F(x_i)$	0,817	0,819	0,851	0,879	0,888	0,928	0,939	0,940	0,944	0,954	0,968
d_n^+	0,044	0,024	0,032	0,038	0,024	0,042	0,030	0,008	0,010	0,024	0,032
d_n^-	**0,067**	0,046	0,055	0,061	0,047	0,064	0,053	0,031	0,013	0,001	0,009

В данном примере максимальные значения d_n^+ и d_n^- одинаковы и равны 0,067. Из табл. П.1 при $\alpha = 0{,}1$ найдем $\lambda = 1{,}22$. Для $n = 44$ критическое значение $d_m(0{,}1) = 1{,}22/\sqrt{44} = 0{,}184$. Поскольку max величина $d_n = 0{,}067$ меньше критического значения, гипотеза о принадлежности выборки нормальному закону не отвергается.

Критерий Мизеса

В качестве меры различия теоретической функции распределения $F(x)$ и эмпирической $F_n(x)$, по критерию Мизеса (критерию ω^2), выступает средний квадрат отклонений по всем значениям аргумента x:

$$\omega_n^2 = \int_{-\infty}^{\infty} \left[F_n(x) - F(x) \right]^2 dF(x). \qquad (3.9)$$

Статистика критерия

$$n\omega_n^2 = \frac{1}{12n} + \sum_{i=1}^{n} \left[F(x_i) - \frac{i - 0{,}5}{n} \right]^2. \qquad (3.10)$$

При неограниченном увеличении n существует предельное распределение статистики $n\omega_n^2$. Задав значение вероятности α, можно определить критические значения $n\omega_n^2(\alpha)$. Проверка гипотезы о законе распределения осуществляется обычным образом: если фактическое значение $n\omega_n^2$ окажется больше критического или равно ему, то, согласно критерию Мизеса с уровнем значимости α, гипотеза H_0 о том, что закон распределения генеральной совокупности соответствует $F(x)$, должна быть отвергнута.

Пример 3.3. Проверить с помощью критерия Мизеса гипотезу о том, что экспериментальные данные, представленные вариационным рядом, табл. 2.3, подчиняются нормальному распределению при уровне значимости $\alpha = 0{,}1$.

Решение. Исходные данные и результаты вычислений представлены в табл. 3.4.

Таблица 3.4

i	1	2	3	4	5	6	7	8	9	10	11
x_i	25,79	25,98	25,98	26,12	26,13	26,49	26,52	26,60	26,66	26,69	26,74
$F_n(x_i)$	0,011	0,034	0,057	0,080	0,102	0,125	0,148	0,171	0,193	0,216	0,237
$F(x_i)$	0,036	0,055	0,055	0,073	0,075	0,144	0,151	0,170	0,188	0,196	0,211
Δ_i	0,618	0,429	0,003	0,047	0,726	0,378	0,009	0,000	0,025	0,409	0,742
i	12	13	14	15	16	17	18	19	20	21	22
x_i	26,85	26,90	26,91	26,96	27,02	27,11	27,19	27,21	27,28	27,30	27,38
$F_n(x_i)$	0,261	0,284	0,307	0,330	0,352	0,375	0,398	0,421	0,443	0,466	0,489
$F(x_i)$	0,246	0,263	0,267	0,284	0,305	0,337	0,371	0,378	0,406	0,412	0,447
Δ_i	0,231	0,439	1,572	2,071	2,243	1,467	0,717	1,790	1,391	2,866	1,755

Окончание табл. 3.4

i	23	24	25	26	27	28	29	30	31	32	33
x_i	27,40	27,49	27,64	27,66	27,71	27,78	27,89	27,89	28,01	28,10	28,11
$F_n(x_i)$	0,511	0,534	0,557	0,580	0,602	0,625	0,648	0,671	0,693	0,716	0,739
$F(x_i)$	0,456	0,492	0,555	0,561	0,583	0,610	0,656	0,656	0,701	0,731	0,735
Δ_i	3,103	1,765	0,003	0,332	0,374	0,216	0,063	0,213	0,067	0,238	0,013
i	34	35	36	37	38	39	40	41	42	43	44
x_i	28,37	28,38	28,50	28,63	28,67	28,90	28,99	28,99	29,03	29,12	29,28
$F_n(x_i)$	0,761	0,784	0,807	0,830	0,852	0,875	0,898	0,921	0,943	0,966	0,989
$F(x_i)$	0,817	0,819	0,851	0,879	0,888	0,928	0,939	0,940	0,944	0,954	0,968
Δ_i	3,090	1,230	1,908	2,461	1,271	2,791	1,737	0,381	0,001	0,149	0,432

В этой таблице:

$F_n(x_i) = (1-0,5)/44$ – значение эмпирической функции распределения;

$F(x_i)$ – значение теоретической функции распределения, соответствует значению функции нормального распределения в точке x_i;

$\Delta_i = 1000\left[F_n(x_i) - F(x_i)\right]^2$. Здесь масштабный множитель 1000 введен для удобства отображения данных в таблице, при расчетах он не используется.

Критическое значение статистики критерия Мизеса при заданном уровне значимости равно 0,347, табл. П.2. Фактическое значение статистики $n\omega_n^2 = \dfrac{1}{12 \cdot 44} + \sum_{i=1}^{44} \Delta_i / 1000 = 0,044$, что меньше критического значения. Следовательно, гипотеза H_0 не противоречит имеющимся данным.

Достоинством критерия Мизеса является быстрая сходимость к предельному закону, для этого достаточно не менее 40 наблюдений в области часто используемых на практике больших значений $n\omega_n^2$ (а не несколько сот, как для критерия хи-квадрат).

Сопоставляя возможности различных критериев, необходимо отметить следующие особенности.

Критерий Пирсона устойчив к отдельным случайным ошибкам в экспериментальных данных. Однако его применение требует группирования данных по интервалам, выбор которых относительно произволен и подвержен противоречивым рекомендациям.

Критерий Колмогорова слабо чувствителен к виду закона распределения и подвержен влиянию помех в исходной выборке, но прост в применении.

Критерий Мизеса имеет ряд общих свойств с критерием Колмогорова: оба основаны непосредственно на результатах наблюдения и не требуют построения статистического ряда, что повышает объективность выводов;

оба не учитывают уменьшение числа степеней свободы при определении параметров распределения по выборке, а это ведет к риску принятия ошибочной гипотезы. Их предпочтительно применять в тех случаях, когда параметры закона распределения известны априори, например, при проверке датчиков случайных чисел.

При проверке гипотез о законе распределения следует помнить, что слишком хорошее совпадение с выбранным законом распределения может быть обусловлено некачественным экспериментом («подчистка» экспериментальных данных) или предвзятой предварительной обработкой результатов (некоторые результаты отбрасываются или округляются).

Выбор критерия проверки гипотезы относительно произволен. Разные критерии могут давать различные выводы о справедливости гипотезы, окончательное заключение в таком случае принимается на основе неформальных соображений. Точно также нет однозначных рекомендаций по выбору уровня значимости.

Рассмотренный подход к проверке гипотез, основанный на применении специальных таблиц критических точек распределения, сложился в эпоху «ручной» обработки экспериментальных данных, когда наличие таких таблиц существенно снижало трудоемкость вычислений. В настоящее время математические пакеты включают процедуры вычисления стандартных функций распределений, что позволяет отказаться от использования таблиц, но может потребовать изменения правил проверки. Например, соблюдению гипотезы H_0 соответствует такое значение функции распределения критерия, которое не превышает значение доверительной вероятности $1-\alpha$ (оценка статистики критерия соответствует доверительному интервалу). В частности, для примера 3.1 значение статистики критерия хи-квадрат равно 1,318. А значение функции распределения хи-квадрат для этого значения аргумента при трех степенях свободы составляет 0,275, что меньше доверительной вероятности 0,95. Следовательно, нет оснований отвергать нулевую гипотезу.

Задачи.

3.1. Проверить с помощью критерия хи-квадрат гипотезу о нормальности распределения случайной величины, представленной статистическим рядом в задаче 2.1 при уровне значимости $\alpha = 0,05$.

3.2. Проверить с помощью критерия хи-квадрат гипотезу о нормальности распределения случайной величины, представленной статистическим рядом в задаче 2.2 при уровне значимости $\alpha = 0,01$.

3.3. Проверить с помощью критерия Мизеса гипотезу о том, что экспериментальные данные, представленные вариационным рядом, задача 2.1, подчиняются нормальному распределению при уровне значимости $\alpha = 0,1$.

3.4. Проверить с помощью критерия Мизеса гипотезу о том, что экспериментальные данные, представленные вариационным рядом, задача 2.2, подчиняются нормальному распределению при уровне значимости $\alpha = 0,1$.

4. МЕТОДЫ ОЦЕНКИ ПАРАМЕТРОВ РАСПРЕДЕЛЕНИЯ

4.1. Точечная оценка параметров распределения

Сущность задачи точечного оценивания параметров

Точечная оценка предполагает нахождение единственной числовой величины, которая и принимается за значение параметра. Такую оценку целесообразно определять в тех случаях, когда объем экспериментальных данных достаточно велик. Причем не существует единого понятия о достаточном объеме экспериментальных данных, его значение зависит от вида оцениваемого параметра (к этому вопросу предстоит вернуться при изучении методов интервальной оценки параметров, а предварительно будем считать достаточной выборку, содержащую не менее чем десять значений). При малом объеме экспериментальных данных точечные оценки могут значительно отличаться от истинных значений параметров, что делает их непригодными для использования.

Задача точечной оценки параметров в типовом варианте постановки состоит в следующем [5].

Имеется: выборка наблюдений $(x_1, x_2, ..., x_n)$ за случайной величиной X. Объем выборки n фиксирован.

Известен вид закона распределения величины X, например, в форме плотности распределения $f(T, x)$, где T – неизвестный (в общем случае векторный) параметр распределения. Параметр является неслучайной величиной.

Требуется найти оценку θ параметра T закона распределения.

Ограничения: выборка представительная.

Существует несколько методов решения задачи точечной оценки параметров, наиболее употребительными из них являются методы максимального (наибольшего) правдоподобия, моментов и квантилей.

Метод максимального правдоподобия

Метод предложен Р. Фишером в 1912 году. Метод основан на исследовании вероятности получения выборки наблюдений $(x_1, x_2, ..., x_n)$. Эта вероятность равна $f(x_1, T) f(x_2, T) ... f(x_n, T) dx_1 dx_2 ... dx_n$.

Совместная плотность вероятности

$$L(x_1, x_2, ..., x_n; T) = f(x_1, T) f(x_2, T) ... f(x_n, T), \qquad (4.1)$$

рассматриваемая как функция параметра *T*, называется *функцией правдоподобия*.

В качестве оценки θ параметра *T* следует взять то значение, которое обращает функцию правдоподобия в максимум. Для нахождения оценки необходимо заменить в функции правдоподобия *T* на θ и решить уравне-

ние $\partial L/\partial\theta = 0$. В целях упрощения вычислений переходят от функции правдоподобия к ее логарифму $\ln L$. Такое преобразование допустимо, так как функция правдоподобия – положительная функция, и она достигает максимума в той же точке, что и ее логарифм. Если параметр распределения векторная величина $\theta = (\theta_1, \theta_2, ..., \theta_n)$, то оценки максимального правдоподобия находят из системы уравнений:

$$\begin{aligned} &\partial \ln L(\theta_1, \theta_2, ..., \theta_n)/\partial_{\theta_1} = 0; \\ &\partial \ln L(\theta_1, \theta_2, ..., \theta_n)/\partial_{\theta_2} = 0; \\ &..... \\ &\partial \ln L(\theta_1, \theta_2, ..., \theta_n)/\partial_{\theta_n} = 0. \end{aligned} \qquad (4.2)$$

Для проверки того, что точка оптимума соответствует максимуму функции правдоподобия, необходимо найти вторую производную от этой функции. И если вторая производная в точке оптимума отрицательна, то найденные значения параметров максимизируют функцию.

Итак, нахождение оценок максимального правдоподобия включает следующие этапы: построение функции правдоподобия (ее натурального логарифма); дифференцирование функции по искомым параметрам и составление системы уравнений; решение системы уравнений для нахождения оценок; определение второй производной функции, проверку ее знака в точке оптимума первой производной и формирование выводов.

Пример 4.1. Будем считать, что случайная величина X, выборка значений которой представлена в табл. 2.3, имеет нормальное распределение. Необходимо найти оценки максимального правдоподобия параметров μ и σ этого распределения.

Решение. Функция правдоподобия для выборки экспериментальных данных объемом n

$$L(\mu, \sigma) = \frac{1}{\left(\sigma\sqrt{2\pi}\right)^n} \exp\left[-\sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2\sigma^2} \right].$$

Логарифм функции правдоподобия

$$\ln L(\mu, \sigma) = -n \ln\left(\sqrt{2\pi}\right) - n \ln\sigma - \left\{ \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2\sigma^2} \right\}.$$

Система уравнений для нахождения оценок параметров:

$$\partial \ln L(\mu, \sigma)/\partial\mu = \left\{ \sum_{i=1}^{n} \frac{(x_i - \mu)}{\sigma^2} \right\} = 0;$$

$$\partial \ln L(\mu, \sigma)/\partial\sigma = -n/\sigma + \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{\sigma^3} = 0.$$

Из первого уравнения следует: $\sum_{i=1}^{n}(x_i-\mu)=0;\ \ \mu=\sum_{i=1}^{n}x_i/n,$ т.е. среднее арифметическое является оценкой максимального правдоподобия для математического ожидания. Из второго уравнения можно найти $\sigma^2=\sum_{i=1}^{n}(x_i-\mu)^2/n$. Эмпирическая дисперсия является смещенной. После устранения смещения $\sigma^2=\sum_{i=1}^{n}(x_i-\mu)^2/(n-1)$.

Фактические значения оценок параметров: $\mu=27{,}51,\ \ \sigma^2=0{,}91.$

Для проверки того, что полученные оценки максимизируют значение функции правдоподобия, возьмем вторые производные:

$$\partial^2\ln(\mu,\ \sigma)/\partial\mu^2=-n/\sigma^2;$$

$$\partial^2\ln(\mu,\ \sigma)/\partial\sigma^2=n/\sigma^2-3\sum_{i=1}^{n}(x_i-\mu)^2\Big/\sigma^4=n/\sigma^2-3n/\sigma^2=-2n/\sigma^2.$$

Вторые производные от функции $\ln L(\mu,\ \sigma)$, независимо от значений параметров, – меньше нуля, следовательно, найденные значения параметров являются оценками максимального правдоподобия.

Метод максимального правдоподобия позволяет получить состоятельные, эффективные (если таковые существуют, то полученное решение даст эффективные оценки), достаточные, асимптотически нормально распределенные оценки. Этот метод может давать как смещенные, так и несмещенные оценки. Смещение удается устранить введением поправок. Метод особенно полезен при малых выборках. Оценка инвариантна относительно преобразования параметра, т.е. оценка некоторой функции $\varphi(T)$ от параметра T является эта же функция от оценки $\varphi(\theta)$. Если функция максимального правдоподобия имеет несколько максимумов, то из них выбирают глобальный.

Метод моментов

Метод предложен К. Пирсоном в 1894 г.

Сущность метода:

– выбирается столько эмпирических моментов, сколько требуется оценить неизвестных параметров распределения. Желательно применять моменты младших порядков, так как погрешности вычисления оценок резко возрастают с увеличением порядка момента;

– вычисленные по экспериментальным данным оценки моментов приравниваются к теоретическим моментам;

– параметры распределения определяются через моменты, и составляются уравнения, выражающие зависимость параметров от моментов, в результате получается система уравнений. Решение этой системы дает оценки параметров распределения генеральной совокупности.

Пример 4.2. Предположим, что случайная величина X, выборка значений которой представлена в табл. 2.3, имеет гамма-распределение. Необходимо найти оценки параметров этого распределения (можно отметить, что нормальное распределение является частным случаем гамма-распределения).

Решение. Функция плотности гамма-распределения имеет вид

$$f(x,\ \lambda) = \frac{\lambda^{\nu}}{\Gamma(\nu)} x^{\nu-1} \exp(-\lambda x),\ \ x \geq 0,\ \ \lambda > 0,\ \ \nu \geq 0.$$

Распределение характеризуется двумя параметрами ν и λ, поэтому следует выразить один параметр через оценку математического ожидания, а другой – через оценку дисперсии. Математическое ожидание и дисперсия этого распределения равны ν/λ и ν/λ^2 соответственно. Их оценки определены в примере 2.3: $\mu_1 = 27{,}51,\ \ \mu_2 = 0{,}91.$ Тогда получим систему уравнений для оцениваемых параметров

$$\frac{\nu}{\lambda} = \frac{1}{n}\sum_{i=1}^{n} x_i = \mu_1,$$

$$\frac{\nu}{\lambda^2} = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \mu)^2 = \mu_2.$$

Разделив оценку математического ожидания на оценку дисперсии, получим $\lambda = \mu_1/\mu_2 = 30{,}12$, следовательно, $\nu = \lambda \cdot \mu_1 = 828{,}61.$

Метод моментов позволяет получить состоятельные, достаточные оценки, они при довольно общих условиях распределены асимптотически нормально. Смещение удается устранить введением поправок. Эффективность оценок невысокая, т.е. даже при больших объемах выборок дисперсия оценок относительно велика (за исключением нормального распределения, для которого метод моментов дает эффективные оценки). В реализации метод моментов проще метода максимального правдоподобия. Напомним, что метод целесообразно применять для оценки не более чем четырех параметров, так как точность выборочных моментов резко падает с увеличением их порядка.

Метод квантилей

Сущность метода квантилей схожа с методом моментов: выбирается столько квантилей, сколько требуется оценить параметров; неизвестные теоретические квантили, выраженные через параметры распределения,

приравниваются к эмпирическим квантилям. Решение полученной системы уравнений дает искомые оценки параметров.

Дисперсия $D(x_\alpha)$ выборочной квантили обратно пропорциональна квадрату плотности распределения $D(x_\alpha)=\left[\alpha(1-\alpha)\right]/\left[nf^2(x_\alpha)\right]$ в окрестностях точки x_α. Поэтому следует выбирать квантили вблизи тех значений x, в которых плотность вероятности максимальна.

Пример 4.3. Оценить методом квантилей параметры нормального распределения случайной величины, выборочные значения которой представлены в табл. 2.3.

Решение. Так как требуется определить два параметра распределения μ и σ, то выберем из вариационного ряда две эмпирические квантили.

Например, можно взять

α_1 =5/44 =0,114; $x_{\alpha 1}$ = 26,13;

α_2 =31/44=0,705; $x_{\alpha 2}$ = 28,01

Используя стандартные функции математических пакетов, для выбранных значений α_1 и α_2 определим значения аргументов теоретической функции распределения для стандартизованной переменной $u\alpha_1$ = – 1, 207; $u\alpha_2$ = 0,538.

Составим систему из двух уравнений

$u\alpha_1 = (x\alpha_1 - \mu)/\sigma$;

$u\alpha_1 = (x\alpha_2 - \mu)/\sigma$.

Решение системы позволит найти искомые оценки параметров

$$\mu = (u\alpha_2 x\alpha_1 - u\alpha_1 x\alpha_2)/(u\alpha_2 - u\alpha_1) = 27{,}42;\ \sigma = (x\alpha_1 - \mu)/u\alpha_1 = 1{,}07.$$

Метод квантилей позволяет получить асимптотически нормальные оценки, однако они несут в себе некоторый субъективизм, связанный с относительно произвольным выбором квантилей. Эффективность оценок не выше метода моментов. Определение оценок может приводить к необходимости численного решения достаточно сложных систем уравнений.

Оценки, вычисленные на основе различных методов, различаются. Универсального ответа на вопрос, какой из рассмотренных методов лучше или следует ли положиться на данный метод при решении любой задачи, нет. Значение оценки в каждом конкретном случае (для разных выборок) отличается от истинного значения параметра на неизвестную величину, иначе говоря, существует некоторая доля неопределенности в знании действительного значения параметра. Качество оценок можно определить косвенно путем проверки согласованности эмпирических данных и теоретического закона распределения.

4.2. Интервальная оценка параметров распределения

Сущность задачи интервального оценивания параметров

Интервальный метод оценивания параметров распределения случайных величин заключается в определении интервала (а не единичного значения), в котором с заданной степенью достоверности будет заключено значение оцениваемого параметра. *Интервальная оценка* характеризуется двумя числами – концами интервала, внутри которого предположительно находится истинное значение параметра. Иначе говоря, вместо отдельной точки для оцениваемого параметра можно установить интервал значений, одна из точек которого является своего рода "лучшей" оценкой. Интервальные оценки являются более полными и надежными по сравнению с точечными, они применяются как для больших, так и для малых выборок. Совокупность методов определения промежутка, в котором лежит значение параметра T, получила название методов интервального оценивания. К их числу принадлежит метод Неймана.

Постановка задачи интервальной оценки параметров заключается в следующем [5, 14].

Имеется: выборка наблюдений $(x_1, x_2, ..., x_n)$ за случайной величиной X. Объем выборки n фиксирован.

Необходимо с доверительной вероятностью $\gamma = 1 - \alpha$ определить интервал $t_0 - t_1$ $(t_0 < t_1)$, который накрывает истинное значение неизвестного скалярного параметра T (здесь, как и ранее, величина T является постоянной, поэтому некорректно говорить, что значение T попадает в заданный интервал).

Ограничения: выборка представительная, ее объем достаточен для оценки границ интервала.

Эта задача решается путем построения доверительного утверждения, которое состоит в том, что интервал от t_0 до t_1 накрывает истинное значение параметра T с доверительной вероятностью не менее γ. Величины t_0 и t_1 называются нижней и верхней доверительными границами (НДГ и ВДГ, соответственно). Доверительные границы интервала выбирают так, чтобы выполнялось условие $P(t_0 \le \theta \le t_1) = \gamma$. В инженерных задачах доверительную вероятность γ назначают в пределах от 0,95 до 0,99. В доверительном утверждении считается, что статистики t_0 и t_1 являются случайными величинами и изменяются от выборки к выборке. Это означает, что доверительные границы определяются неоднозначно, существует бесконечное количество вариантов их установления.

На практике применяют два варианта задания доверительных границ:

– доверительные границы устанавливают симметрично относительно оценки параметра, т.е. $t_0 = \theta - E\gamma$, $t_1 = \theta + E\gamma$, где $E\gamma$ выбирают так, чтобы

выполнялось доверительное утверждение. Следовательно, величина абсолютной погрешности оценивания $E\gamma$ равна половине доверительного интервала;

– доверительные границы устанавливают из условия равенства вероятностей выхода за верхнюю и нижнюю границу $P\left(T > \theta + E_{1,\gamma}\right) = P\left(T < \theta - E_{2,\gamma}\right) = \alpha/2$. В общем случае величина $E_{1,\gamma}$ не равна $E_{2,\gamma}$. Для симметричных распределений случайного параметра θ в целях минимизации величины интервала значения $E_{1,\gamma}$ и $E_{2,\gamma}$ выбирают одинаковыми, следовательно, в таких случаях оба варианта эквивалентны.

Нахождение доверительных интервалов требует знания вида и параметров закона распределения случайной величины θ. Для ряда практически важных случаев этот закон можно определить из теоретических соображений.

Общий метод построения доверительных интервалов

Метод позволяет по имеющейся случайной выборке построить функцию $u\left(T, \theta\right)$, распределенную асимптотически нормально с нулевым математическим ожиданием и единичной дисперсией. В основе метода лежат следующие положения. Пусть:

$f\left(x, \theta\right)$ – плотность распределения случайной величины X;

$\ln\left[L\left(x, \theta\right)\right]$ – логарифм функции правдоподобия;

$y = \dfrac{\partial}{\partial\theta}\ln f\left(x, \theta\right)$;

$A^2 = M\left(y\right)^2$ – дисперсия y.

Если математическое ожидание $M\left(y\right) = 0$ и дисперсия y конечна, то распределение случайной величины $w = \dfrac{1}{A\sqrt{n}}\dfrac{\partial}{\partial\theta}\ln\left(x, \theta\right)$ асимптотически нормально с параметрами 0 и 1 при $n \to \infty$.

Пример 4.4. Построить доверительный интервал с надежностью $\gamma = 1 - \alpha$ для оценки μ_1 математического ожидания m_1 случайной величины x, имеющей экспоненциальное распределение с функцией плотности $f\left(x, l\right) = l\exp\left(-lx\right)$.

Решение. Известно, что для экспоненциального закона распределения математическое ожидание $m_1 = 1/l$, а дисперсия $m_2 = 1/l^2$. Обозначим через λ оценку параметра l. Определим оценку математического ожидания μ_1, вспомогательную переменную y, производную от логарифма функции правдоподобия:

$$\mu_1 = \left(x_1 + x_2 + \ \ldots\ + x_n\right)/n = 1/\lambda;\ M\left(\lambda\right) = l;\ m_1 = M\left(1/\lambda\right);$$

$$y = \frac{\partial}{\partial\lambda}\ln\left(\lambda\exp\left(-\lambda x\right)\right) = \frac{\partial}{\partial\lambda}\left[\ln\lambda - \lambda x\right] = \lambda^{-1} - x;$$

$$z = \frac{\partial}{\partial\lambda}\ln L\left(x,\,\lambda\right) = \frac{\partial}{\partial\lambda}\ln\left[\lambda^{n}\exp\left(-\lambda\sum_{i=1}^{n}x_i\right)\right] = \frac{\partial}{\partial\lambda}\left[n\ln\lambda - \lambda\sum_{i=1}^{n}x_i\right] = \frac{n}{\lambda} - \sum_{i=1}^{n}x_i\,.$$

Оценка μ_1 параметра m_1 является состоятельной и несмещенной, следовательно: $M\left(y\right) = M\left(\lambda^{-1} - x\right) = 0$; значение $A^2 = M\left(\lambda^{-1} - x\right)^2 = \lambda^{-2}$ – конечно. Тогда случайная величина

$$w = \frac{1}{A\sqrt{n}}z = \frac{\lambda}{\sqrt{n}}\left[\frac{n}{\lambda} - \sum_{i=1}^{n}x_i\right] = \frac{\lambda n}{\sqrt{n}}\left[\frac{1}{\lambda} - \frac{1}{n}\sum_{i=1}^{n}x_i\right] = \lambda\sqrt{n}\left[m_1 - \mu_1\right]$$

распределена нормально с параметрами 0 и 1.

Нормальное распределение симметрично, поэтому границы интервала следует выбрать симметрично относительно нулевой точки. Вероятность $\gamma = 1 - \alpha$ того, что модуль величины w не превысит некоторого заданного значения δ, составит

$$P\left(\left|\lambda\sqrt{n}\left[m_1 - \mu_1\right]\right| \le \delta\right) = 1 - \alpha = \Phi\left(\delta\right) - \Phi\left(-\delta\right) = -1 + 2\Phi\left(\delta\right),$$

где $\Phi\left(\delta\right)$ – значение функции нормального распределения в точке δ.

Величина δ равна квантили $u_{1-\alpha/2}$ стандартного нормального распределения уровня $1 - \alpha/2$. Значение абсолютной погрешности оценивания равно $E = \left|m_1 - \mu_1\right| = \delta\Big/\left(\lambda n^{0,5}\right) = u_{1-\alpha/2}\Big/\left(\lambda n^{0,5}\right)$. Итак, имея достаточный объем выборки экспериментальных данных и задаваясь определенным уровнем надежности γ, можно определить доверительный интервал $t_0 = \mu_1 - E$, $t_1 = \mu_1 + E$, который с заданной вероятностью содержит неизвестный параметр m_1.

Аналогичные результаты для некоторых параметров распределения можно получить, используя более простые рассуждения.

Доверительный интервал для математического ожидания

Пусть по выборке достаточно большого объема, $n > 30$, и при заданной доверительной вероятности $1 - \alpha$ необходимо определить доверительный интервал для математического ожидания m_1, в качестве оценки которого используется среднее арифметическое $\mu_1 = \sum_{i=1}^{n}x_i/n$.

Закон распределения оценки математического ожидания близок к нормальному (распределение суммы независимых случайных величин с ко-

нечной дисперсией асимптотически нормально). Если потребовать абсолютную надежность оценки математического ожидания, то границы доверительного интервала будут бесконечными $[-\infty,\ \infty]$. Выбор любых более узких границ связан с риском ошибки, вероятность которой определяется уровнем значимости α. Интерес представляет максимальная точность оценки, т.е. наименьшее значение интервала. Для симметричных функций минимальный интервал тоже будет симметричным относительно оценки μ_1. В этом случае выражение для доверительной вероятности имеет вид $P\left(\left|\mu_1 - m_1\right| \le E\right) = 1 - \alpha$, где E – абсолютная погрешность оценивания.

Нормальный закон полностью определяется двумя параметрами – математическим ожиданием и дисперсией. Величина μ_1 является несмещенной, состоятельной и эффективной оценкой математического ожидания, поэтому ее значение принимаем за значение математического ожидания. Определим оценку дисперсии случайного параметра μ_1, учитывая, что этот параметр равен среднему арифметическому одинаково распределенных случайных величин x_i (следовательно, их дисперсии $D\left(x_i\right)$ одинаковы и равны μ_2)

$$D\left(\mu_i\right) = D\left[\frac{1}{n}\sum_{i=1}^{n} x_i\right] = \frac{1}{n^2}\left[\sum_{i=1}^{n} D\left(x_i\right)\right] = \frac{n\mu_2}{n^2} = \frac{\mu_2}{n}.$$

Итак, случайная величина μ_1 распределена по нормальному закону с параметрами μ_1 и μ_2/n. Для установления необходимых соотношений целесообразно перейти к центрированным и нормированным величинам. Выражение $\mu_1 - m_1$ можно трактовать как центрирование случайной величины μ_1. Нормирование осуществляется делением на величину среднеквадратического отклонения оценки μ_1:

$$P\left(\frac{\left|\mu_1 - m_1\right|}{\sqrt{\mu_2/n}} \le \frac{E}{\sqrt{\mu_2/n}}\right) = \gamma.$$

Для стандартизованной величины $z = \left(\mu_1 - m_1\right)\Big/\sqrt{\mu_2/n}$ вероятность соблюдения неравенства определяется по функции нормального распределения

$$P\left(\left|z\right| \le \beta\right) = \frac{1}{\sqrt{2\pi}}\int_{-\beta}^{\beta} \exp\left(-t^2/2\right)dt = \Phi\left(\beta\right) - \Phi\left(-\beta\right) = -1 + 2\Phi\left(\beta\right) = 1 - \alpha,$$

где $\beta = E\Big/\sqrt{\mu_2/n}$. Значение β равно квантили $u_{1-\alpha/2}$ стандартного нормального распределения уровня $1 - \alpha/2$. В частности, уровням надежности

0,9, 0,95 и 0,99 соответствуют значения допустимого отклонения $u_{1-\alpha/2}$ величины z, равные 1,64, 1,96 и 2,58 соответственно. Окончательно можно записать

$$u_{1-\alpha/2}^2 = nE^2 \big/ \mu_2 . \qquad (4.3)$$

Нетрудно заметить, что это выражение аналогично по своему содержанию формуле, полученной с использованием общего метода построения доверительного интервала.

При фиксированном объеме выборки из (4.3) следует, что чем больше доверительная вероятность $1-\alpha$, тем шире границы доверительного интервала (тем больше ошибка в оценке математического ожидания). Это равенство позволяет определить необходимый объем выборки для получения оценки математического ожидания с заданной надежностью и требуемой точностью (погрешностью): $n = \mu_2 u_{1-\alpha/2}^2 \big/ E^2$. Если перейти к относительной погрешности $\varepsilon = E/\mu_1$, то

$$n = \mu_2 u_{1-\alpha/2}^2 \Big/ \left(\varepsilon^{2\mu} 1^2\right). \qquad (4.4)$$

Таким образом, чтобы снизить относительную погрешность на порядок, необходимо увеличить объем выборки на два порядка. Приведенная формула часто используется в статистическом моделировании для определения необходимого количества испытаний модели.

Во многих случаях предположение о нормальном распределении случайной величины μ_1 становится приемлемым при $n > 4$ и вполне хорошо оправдывается при $n > 10$. Оценка μ_1 вполне пригодна для применения вместо m_1. Но не так обстоит дело с дисперсией, правомочность ее замены на μ_2 не обоснована даже в указанных случаях. При небольшом объеме выборки, $n > 30$, закон распределения оценки дисперсии μ_2 принимать за нормальный неоправданно. Ее распределение следует аппроксимировать распределением хи-квадрат как суммы квадратов центрированных величин (хи-квадрат распределение сходится к нормальному при количестве слагаемых, превышающем 30). Но это утверждение обосновано только для случая, когда случайная величина X распределена нормально.

С учетом сделанных допущений величина z будет подчиняться закону распределения Стьюдента с $(n-1)$ степенями свободы (одна степень свободы использована для определения оценки дисперсии). Распределение Стьюдента симметричное, поэтому полученное соотношение между точностью, надежностью оценки и объемом выборки сохраняется, меняются только значения квантилей. Вместо квантили нормального распределения $u_{1-\alpha/2}$ следует взять квантиль $t_{1-\alpha/2}(n-1)$ распределения Стьюдента с

$(n-1)$ степенями свободы. Значения критических точек распределения Стьюдента для некоторых вероятностей и различных степеней свободы представлены в табл. П.4. Как можно видеть из табл. П.4, квантили распределения Стьюдента больше квантилей нормального распределения того же уровня надежности при малом n.

Иначе говоря, применение нормального распределения при небольшом объеме выборки экспериментальных данных приводит к неоправданному завышению точности оценки.

Пример 4.5. Определить с надежностью $\gamma = 0,9$ доверительный интервал для математического ожидания случайной величины, выборка которой представлена вариационным рядом, табл. 2.3.

Решение. Ранее были определены оценки $\mu_1 = 27,51$ и $\mu_2 = 0,91$. Интервал двусторонний, симметричный, $\alpha = 0,1$. Объем выборки можно считать большим, поэтому воспользуемся нормальным распределением, тогда $u_{0,95} = 1,96$. Допустимое отклонение $E = u_{0,95}\left(\mu_2 / n\right)^{0,5} = 1,96\left(0,91/44\right)^{0,5} = 0,28$.

С вероятностью 0,9 нижняя допустимая граница (НДГ) интервала составит $t_0 = 27,51 - 0,28 = 27,23$, верхняя допустимая граница (ВДГ) интервала $t_1 = 27,51 + 0,28 = 27,79$. Другими словами, с вероятностью 0,9 значение математического ожидания лежит в пределах от 27,23 до 27,79.

Пример 4.6. Определить с надежностью $\gamma = 0,9$ $(\alpha = 0,1)$ доверительный интервал для математического ожидания случайной величины, представленной выборкой (см. табл. 2.1).

Решение. Определим оценки $\mu_1 = 55$ и $\mu_2 = 658,6$. Объем выборки $n = 6$ нельзя считать большим, поэтому воспользуемся распределением Стьюдента при числе степеней свободы, равном 5. Тогда для двусторонней критической области, в соответствии с табл. П.4, допустимое отклонение стандартизованной случайной величины составит $t_{0,9}\left(4\right) = t\left(5;\ 0,1\right) = 2,015$. Допустимое отклонение исходной величины составит

$$E = t_{0,9}\left(5\right)\left(\mu_2 / n\right)^{0,5} = 2,015\left(658,6/6\right)^{0,5} = 21,11.$$

Границы интервала: $t_0 = 55 - 21,11 = 33,89$ и $t_1 = 55 + 21,11 = 76,11$.

В данном примере использование нормального распределения вместо распределения Стьюдента приведет к неоправданному сужению интервала в $2,015/1,64 = 1,3$ раза.

Доверительный интервал для дисперсии

По выборке достаточно большого объема ($n > 30$) и при заданной надежности $(1-\alpha)$ необходимо определить доверительный интервал для дисперсии m_2, оценка которой $\mu_2 = \sigma^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}(x_i - \mu_i)^2$.

Если стандартизовать оценку дисперсии, то величина $(n-1)\sigma^2/m_2$ имеет распределение хи-квадрат с $(n-1)$ степенями свободы. Из этого вытекает вероятностное утверждение относительно выборочной дисперсии

$$P\left[(n-1)\sigma^2/m_2 > \chi^{2\alpha}(n-1)\right] = \alpha. \qquad (4.5)$$

Функция хи-квадрат несимметричная, поэтому границы интервала $\chi_1^2(n-1)$ и $\chi_2^2(n-1)$ выбирают из условия равной вероятности выхода за их пределы $P\left[(n-1)\sigma^2/m_2 < \chi_1^2(n-1)\right] = P\left[(n-1)\sigma^2/m_2 > \chi_2^2(n-1)\right] = \alpha/2$ или

$$P\left[(n-1)\sigma^2/\chi_1^2(n-1) < m_2\right] = P\left[(n-1)\sigma^2/\chi_2^2(n-1) > m_2\right] = \alpha/2. \qquad (4.6)$$

Значения границ соответствуют квантилям распределения хи-квадрат уровня $\alpha/2$ и $1-\alpha/2$ с количеством степеней свободы $(n-1)$. Нижняя граница $\chi_1^2(n-1)$ равна квантили $\chi_{\alpha/2}^2(n-1)$, а верхняя – квантили $\chi_{1-\alpha/2}^2(n-1)$. Если воспользоваться критическими точками распределения, то следует записать $\chi_1^2(n-1) = \chi^2(1-\alpha/2;\ n-1)$ и $\chi_2^2(n-1) = \chi^2(\alpha/2;\ n-1)$.

Пример 4.7. Определить с надежностью 0,9 доверительный интервал для дисперсии случайной величины, представленной выборкой, табл. 2.3.

Решение. Количество степеней свободы $44-1=43$. Вероятности выхода за нижнюю и верхнюю границы $(1-0,9)/2 = 0,05$. По распределению хи-квадрат находим квантили $\chi_{0,05}^2(43) = 28,96$, $\chi_{0,95}^2(43) = 59,30$. Следовательно, НДГ для дисперсии равна $t_0 = (n-1)\sigma^2/\chi_{0,95}^2(43) = 43\cdot 0,91/59,30 = 0,66$, ВДГ – $t_1 = (n-1)\sigma^2/\chi_{0,05}^2(43) = 43\cdot 0,91/28,96 = 1,36$.

Доверительный интервал для вероятности

Пусть случайная величина X имеет только два возможных значения: 0 и 1. В результате проведения достаточно большого количества наблюдений эта случайная величина приняла единичное значение m раз. Необхо-

димо при заданной надежности $(\alpha - 1)$ определить доверительный интервал для вероятности p, оценка которой соответствует частоте $h = m/n$.

Оценка h вероятности p является состоятельной, эффективной и несмещенной. Если оцениваемая вероятность не слишком мала и не слишком велика $(0{,}05 < p < 0{,}95)$, то можно считать, что распределение случайной величины h близко к нормальному. Этим допущением можно пользоваться, если np и $n(1-p)$ больше четырех. Параметры нормального распределения частоты $m_1 = p$, $m_2 = p(1-p)/n$ (дисперсия $\mu_2(m)$ количества успехов m составляет величину $np(1-p)$, а дисперсия частоты $\mu_2(m)/n^2$). Тогда по аналогии с определением доверительного интервала для математического ожидания нормально распределенной величины h можно записать

$$E = |h - p| = u_{1-\alpha/2}\left(\mu_2(m)\right)^{0,5} = u_{1-\alpha/2}\left(p(1-p)/n\right)^{0,5},$$

где $u_{1-\alpha/2}$ – квантиль стандартизованного нормального распределения.

Чтобы связать доверительный интервал с исходными параметрами n, h и $u_{1-\alpha/2}$, возведем выражение для E в квадрат, т.е. преобразуем равенство к виду $(h-p)^2 = u^2_{1-\alpha/2}(1-p)\,p/n$. Доверительные границы можно получить, решив это уравнение второй степени

$$p_{2,\,1} = \left\{nh + 0{,}5u^2_{1-\alpha/2} \pm u_{1-\alpha/2}\left[nh(1-h) + 0{,}25u^2_{1-\alpha/2}\right]^{0,5}\right\} \Big/ \left(n + u^2_{1-\alpha/2}\right). \quad (4.7)$$

С увеличением объема выборки, $(nh > 200,\ nh(1-h) > 200)$, такими слагаемыми как $u^2_{1-\alpha/2}$, $0{,}5u^2_{1-\alpha/2}$ и $0{,}25u^2_{1-\alpha/2}$ можно пренебречь, тогда приближенно имеем:

$$\begin{aligned} p_1 &= h - u_{1-\alpha/2}\left[h(1-h)/n\right]^{0,5}, \\ p_2 &= h + u_{1-\alpha/2}\left[h(1-h)/n\right]^{0,5}. \end{aligned} \quad (4.8)$$

Более общие результаты получены с учетом того, что случайная величина h распределена по биномиальному закону [14]

$$F(h) = \sum_{k=0}^{n} C_n^k p^k (1-p)^{n-k}, \quad (4.9)$$

где C_n^k – число сочетаний из n по k.

Исходя из этого положения, для практического применения получены значения нижней p_1 и верхней p_2 доверительных границ

$$p_1 = \frac{\chi^2_{\alpha/2}(2m)}{2n - m + 1 + 0{,}5\chi^2_{\alpha/2}(2m)}; \quad (4.10)$$

$$p_2 = \frac{\chi^2_{1-\alpha/2}\left(2(m+1)\right)}{2n - m + 0{,}5\chi^2_{1-\alpha/2}\left(2(m+1)\right)}, \qquad (4.11)$$

где $\chi^2_{\xi}(k)$ – квантиль распределения хи-квадрат уровня ξ с числом степеней свободы k.

Формулы (4.10) и (4.11) можно применять и в тех случаях, когда частость h события близка (равна) нулю или близка (равна) количеству экспериментов n соответственно. В первом случае НДГ p_1 принимается равной нулю и рассчитывается только ВДГ p_2. Во втором случае рассчитывается НДГ p_1, а верхняя граница $p_2 = 1$.

Пример 4.8. В результате наблюдения за 58 изделиями не было зафиксировано ни одного отказа. Определить доверительный интервал для вероятности отказа с надежностью 0,9.

Решение. Нижнюю доверительную границу p_1 следует принять равной нулю, ВДГ $p_2 = \dfrac{\chi^2_{0,95}(2)}{116 - 0 + 0{,}5\chi^2_{0,95}(2)} = \dfrac{6{,}0}{119} = 0{,}05$.

Таким образом, доверительный интервал с нижней границей 0 и верхней границей 0,05 с вероятностью 0,9 накрывает истинное значение вероятности отказа изделий.

Задачи.

4.1. Необходимо найти оценки максимального правдоподобия параметров μ и σ распределения, представленного в задаче 2.1, считая данное распределение случайной величины нормальным.

4.2. Предположим, что случайная величина X, выборка значений которой представлена в задаче 2.2, имеет гамма-распределение. Необходимо найти оценки параметров этого распределения (можно отметить, что нормальное распределение является частным случаем гамма-распределения).

4.3. Определить с надежностью $\gamma = 0{,}9$ доверительный интервал для математического ожидания случайной величины, выборка которой представлена вариационным рядом (задача 2.1).

4.4. Определить с надежностью $\gamma = 0{,}9$ $(\alpha = 0{,}1)$ доверительный интервал для математического ожидания случайной величины, представленной выборкой (задача 2.2).

4.5. Определить с надежностью $\gamma = 0{,}9$ доверительный интервал для дисперсии случайной величины, выборка которой представлена вариационным рядом (задача 2.1).

4.6. Определить с надежностью $\gamma = 0{,}9$ доверительный интервал для дисперсии случайной величины, выборка которой представлена вариационным рядом (задача 2.2).

5. АППРОКСИМАЦИЯ ЗАКОНА РАСПРЕДЕЛЕНИЯ ЭКСПЕРИМЕНТАЛЬНЫХ ДАННЫХ

5.1. Задачи аппроксимации

Конкретное содержание обработки одномерных экспериментальных данных зависит от поставленных целей исследования. В простейшем случае достаточно определить первый момент распределения. В других случаях требуется установить вероятностно-временные характеристики распределения, например, оценить вероятность своевременной обработки запросов или вероятность безотказной работы системы в течение заданного периода времени. Для нахождения таких значений требуется знание закона распределения как наиболее полной характеристики соответствующей случайной величины.

В классической математической статистике вид закона распределения предполагается известным и производится оценка значений его параметров по результатам наблюдений. Но обычно заранее вид закона распределения неизвестен, а теоретические предположения не позволяют его однозначно установить. Обработка экспериментальных данных также не позволит точно вычислить истинный закон распределения показателя. В таком случае следует говорить только об аппроксимации (приближенном описании) реального закона некоторым другим, который не противоречит экспериментальным данным и в каком-то смысле похож на этот неизвестный истинный закон.

В соответствии с этими положениями *постановка задачи аппроксимации закона распределения* экспериментальных данных формулируется следующим образом.

Имеется выборка наблюдений $(x_1, x_2, ..., x_n)$ за случайной величиной X. Объем выборки n фиксирован.

Необходимо подобрать закон распределения (вид и параметры), который бы в статистическом смысле соответствовал имеющимся наблюдениям.

Ограничения: выборка представительная, ее объем достаточен для оценки параметров и проверки согласованности выбранного закона распределения и экспериментальных данных; плотность распределения унимодальная.

Наличие в функции плотности распределения нескольких мод может быть следствием различных причин, например, существования сезонной разницы остроты пара при тепловлажностной обработке железобетонных изделий. Выборку с несколькими модами разделяют на составные части так, чтобы каждая из них имела одну моду. В последнем случае функция распределения исходной выборки представляет собой взвешенную сумму

соответствующих функций отдельных выборок: $F(x)=\sum_{i=1}^{s} p_i F_i(x)$, где s – количество выборок, выбранное исходя из требований унимодальности распределения; p_i – вероятность принадлежности элемента выборки к выборке i; $F_i(x)$ – функция распределения выборки i.

Решение поставленной задачи аппроксимации осуществляется на основе применения "типовых" распределений, специальных рядов или семейств универсальных распределений [5, 9, 10, 11, 15].

5.2. Аппроксимация на основе типовых распределений

Задача аппроксимации на основе типовых распределений решается итерационно и включает выполнение трех основных шагов:

– предварительного выбора вида закона распределения;

– определения оценок параметров закона распределения;

– оценки согласованности закона распределения и экспериментальных данных.

Если заданный уровень согласованности достигнут, то задача считается решенной, а если нет, то шаги повторяются снова, начиная с первого шага, на котором выбирается другой вид закона, или начиная со второго – путем некоторого уточнения параметров распределения.

Выбор вида закона распределения осуществляется посредством анализа гистограммы распределения, оценок коэффициентов асимметрии и эксцесса. По степени «похожести» гистограммы и графиков плотностей распределения типовых законов или по «близости» значений оценок коэффициентов и диапазонов их теоретических значений выбираются распределения – кандидаты для последующей оценки параметров. На рис. 3.5, 5.1–5.4 представлены графики типовых функций плотностей распределения, часто применяемых в задачах аппроксимации экспериментальных данных, а в табл. 5.1 приведены функции плотности и теоретические параметры этих распределений.

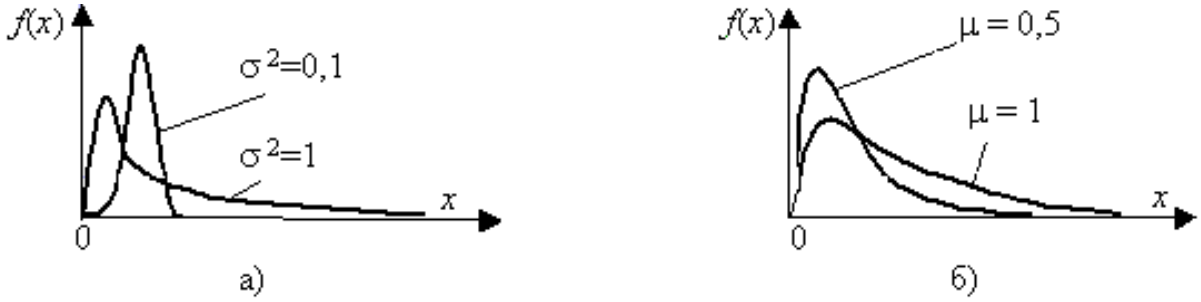

Рис. 5.1. Логарифмически нормальное распределение

а) $\mu = 0$;

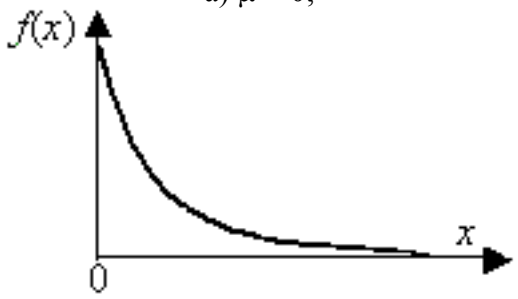

Рис. 5.2. Экспоненциальное распределение

б) $\sigma = 1$

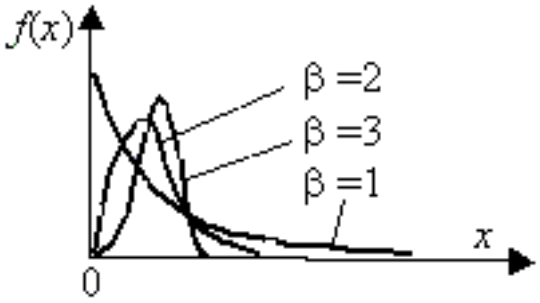

Рис. 5.3. Распределение Вейбулла

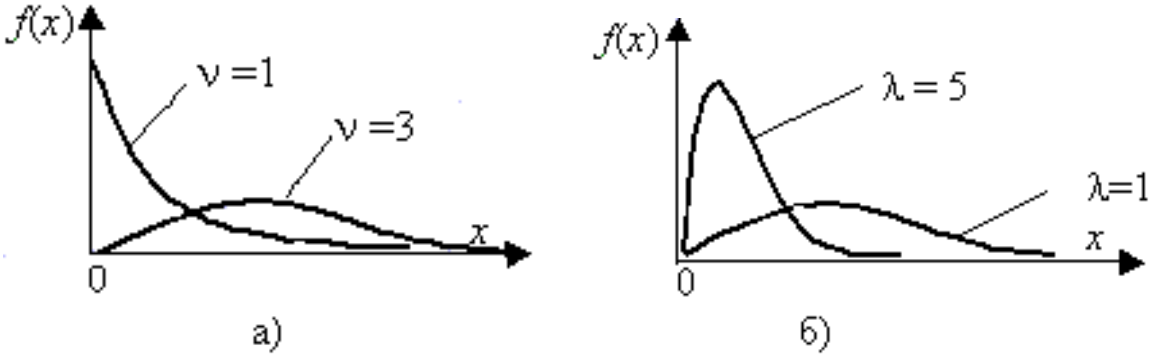

Рис. 5.4. Гамма-распределение:
а) $\lambda = 1$; б) $\nu = 3$

Следует отметить, что гамма-распределение соответствует распределению Эрланга, если λ – целое, и экспоненциальному распределению при $\nu = 1$.

После выбора подходящего вида распределения производится оценка его параметров, используя методы максимального правдоподобия, моментов или квантилей. В целях упрощения решения задачи в табл. 5.2 приведены расчетные формулы для вычисления оценок параметров типовых распределений.

Применительно к выбранному закону распределения производится проверка гипотезы о том, что имеющаяся выборка может принадлежать этому закону. Если гипотеза не отвергается, то можно считать, что задача аппроксимации решена. Если гипотеза отвергается, то возможны следующие действия: изменения значений оценок параметров распределения; выбор другого вида закона распределения; продолжение наблюдений и пополнение выборки. Конечно, такой подход не гарантирует нахождение «истинного» или даже подбора подходящего закона распределения.

Таблица 5.1

Тип и функция плотности распределения	Математическое ожидание m_1, дисперсия m_2, асимметрия $b_1 = m_3 / m_2^{3/2}$, эксцесс $b_2 = m_4 / m_2^2$
Нормальное $\frac{1}{\sigma\sqrt{2\pi}} \exp\left(-(x-\mu_1)^2 / \left(2\sigma^2\right)\right)$, $-\infty < x < \infty$	$m_1 = \mu_1;\ m_2 = \sigma^2 = 2;$ $b_1 = 0;\ b_2 = 3$
Логарифмически нормальное $\frac{1}{\sigma x\sqrt{2\pi}} \exp\left(-\frac{(\ln x - \mu_1)^2}{2\sigma^2}\right)$, $x > 0,\ 0,\ x \le 0$	$m_1 = \exp(\mu_1 + 0{,}5\mu_2);$ $m_2 = \exp(2\mu_1 + \mu_2)(\exp(\mu_2) - 1);$ $b_1 = (\exp(\mu_2) + 2)\sqrt{\exp(\mu_2) - 1};$ $b_2 = \exp(4\mu_4) + 2\exp(3\mu_2) + 3\exp(2\mu_2) - 3$
Экспоненциальное $\lambda \exp(-\lambda x)$, $x \ge 0,\ 0,\ x < 0$	$m_1 = 1/\lambda;\ m_2 = 1/\lambda^2;$ $b_1 = 2;\ b_2 = 9$
Вейбулла $\frac{\beta}{\delta}\left(\frac{x}{\delta}\right)^{\beta-1} \exp\left(-\left(\frac{x}{\delta}\right)^{\beta}\right)$, $x \ge 0,\ 0,\ x < 0,\ \delta > 0,\ \beta > 0$	$m_1 = \delta g_1;\ m_2 = \delta^2\left(g_2 - g_1^2\right);$ $b_1 = \left(g_3 - 3g_1g_2 + 2g_1^3\right) / \left(g_2 - g_1^2\right)^{3/2};$ $a = \left(g_4 - 4g_1g_3 + 6g_2g_1^2 - 3g_1^4\right);$ $b_2 = a / \left(g_2 - g_1^2\right)^2;$ $g_1 = \Gamma(1 + i/\beta)$
Гамма $\frac{\lambda^{\nu}}{\Gamma(\nu)} x^{\nu-1} \exp(-\lambda x)$, $x \ge 0,\ 0,\ x < 0,\ \nu > 0,\ \lambda > 0$	$m_1 = \nu/\lambda;\ m_2 = \nu/\lambda^2;$ $b_1 = 2/\sqrt{\nu};\ b_2 = 3(\nu + 2)/\nu$

Преимущество применения типовых законов распределения состоит в их хорошей изученности и возможности получения состоятельных, несмещенных и относительно высокоэффективных оценок параметров. Однако рассмотренные выше типовые законы распределения не обладают необходимым разнообразием форм, поэтому их применение не дает необходимой общности представления случайных величин, которые встречаются при обработке экспериментальных данных.

Таблица 5.2

Тип распределения	Оценка параметров распределения по выборочным данным
Нормальное	$\mu_1 = \frac{1}{n}\sum_{i=1}^{n} x_i\,;\ \mu_2 = \sigma^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \mu\right)^2$
Логарифмически нормальное	$\mu_1 = \frac{1}{n}\sum_{i=1}^{n} \ln x_i\,;\ \mu_2 = \sigma^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(\ln x_i - \mu\right)^2$
Экспоненциальное	$\lambda = \frac{1}{\frac{1}{n}\sum_{i=1}^{n} x_i}$
Вейбулла	$\delta = \exp\left(\frac{\ln a \ln x_q - \ln b \ln x_p}{\ln a - \ln b}\right);\ \beta = \frac{\ln a - \ln b}{\ln x_q - \ln x_p};$ $0 < q < p < 1,\ a = -\ln\left(1-p\right),\ b = -\ln\left(1-q\right);$ $x_q,\ x_p$ – выборочные квантили
Гамма	$\alpha = \begin{cases} \frac{\left(0{,}5001 - 0{,}1649q - 0{,}0544q^2\right)}{q} - 1, & 0 < q \le 0{,}577, \\ \frac{\left(8{,}899 + 9{,}060q + 0{,}9775q^2\right)}{\left(17{,}80 + 11{,}97q + q^2\right)q} - 1, & 0{,}577 < q \le 17, \end{cases}$ где $q = \ln\left(\mu_1/6\right),\ \beta = \mu_1/\left(1+\alpha\right),\ \mu_1 = \frac{1}{n}\sum_{i=1}^{n} x_i$

5.3. Аппроксимация на основе специальных рядов

Типовые ряды, известные из математического анализа (ряды Тейлора, Фурье), не подходят для описания функций распределений, так как не обладают свойствами, присущими этому виду функций. Для подобного описания предложены специальные функции, например, основанные на полиномах Чебышева – Эрмита. К числу таких функций относится *ряд Грама – Шарлье*

$$F\left(x\right) = \Phi\left(u\right) - \frac{1}{6}\sqrt{\frac{\mu_3^2}{\mu_2^3}}\Phi^{(3)}\left(u\right) + \frac{1}{24}\left(\frac{\mu_4}{\mu_2^2} - 3\right)\Phi^{(4)}\left(u\right) + \frac{1}{72}\frac{\mu_3^2}{\mu_2^3}\Phi^{(6)}\left(u\right) + \ldots, \quad (5.1)$$

где $\Phi(u)$ – функция нормального распределения центрированной и нормированной случайной величины $u=(x-\mu_1)/\mu_2^{0,5}$, $\Phi^{(k)}(u)$ – k-я производная от функции нормального распределения.

Вычисление $\Phi(u)$ не требует численного интегрирования, так как имеются ее приближения на основе полиномов, а производные представимы элементарными функциями:

$$\begin{aligned}
\Phi^{(3)}(u) &= \left(u^2-1\right) f_н(u),\\
\Phi^{(4)}(u) &= \left(-u^3+3u\right) f_н(u),\\
\Phi^{(6)}(u) &= \left(-u^5+10u^3-15u\right) f_н(u),\\
f_н(u) &= (2\pi)^{-0,5}\exp\left(-u^2/2\right).
\end{aligned} \qquad (5.2)$$

Ряд Грама – Шарлье целесообразно использовать для описания распределений, близких к нормальному. В других случаях начинают проявляться серьезные недостатки: ряд может вести себя нерегулярно (увеличение количества членов ряда иногда снижает точность аппроксимации); ошибки аппроксимации возрастают с удалением от центра распределения; сумма конечного числа членов ряда при большой асимметрии распределения приводит к отрицательным значениям функций, особенно на краях распределений. Этот ряд применяют только при весьма умеренном коэффициенте асимметрии, не превышающем 0,7. Следовательно, применение рядов тоже не обеспечивает необходимой общности решения задач аппроксимации.

Пример 5.1. Оценить качество аппроксимации экспериментальных данных, табл. 2.4, на основе ряда Грама – Шарлье. Проверку согласованности провести с использованием критерия хи-квадрат при уровне значимости $\alpha=0,5$.

Решение. В примере 2.3 были вычислены значения оценок моментов: $\mu_1=27,508$, $\mu_2=0,913$, $\mu_3=0,132$, $\mu_4=1,819$.

На основе табл. 2.4 построим табл. 5.3.

Таблица 5.3

i	1	2	3	4	5	6
n_i	5	9	10	9	5	6
Верхняя граница, x_i	26,37	26,95	27,53	28,11	28,69	∞
$F(x_i)$	0,127	0,303	0,517	0,721	0,877	1
ΔF_i	0,127	0,176	0,214	0,204	0,156	0,123
F_i	5,588	7,744	9,416	9,976	6,864	5,412
$(n_i-F_i)^2/F_i$	0,062	0,204	0,036	0,000	0,506	0,063

В таблице значения функции распределения $F(x_i)$ для верхней границы интервала и теоретическое значение оценки вероятности ΔF_i попадания случайной величины в i-й интервал вычислены на основе ряда Грама-Шарлье. Обозначения оценки частоты попадания $F_i = \Delta F_i \cdot n$ случайной величины в i-й интервал, вероятности ΔF_i попадания случайной величины в интервал $x_i - x_{i-1}$, взвешенного квадрата отклонения $(n_i - F_i)^2 / F_i$ аналогичны табл. 3.2. Сумма взвешенных квадратов отклонения $\chi^2 = 0{,}872$ (критическое значение составляет 7,815).

Выборка имеет слабо выраженную асимметрию. По сравнению с аналогичным значением $\chi^2 = 1{,}318$ при аппроксимации экспериментальных данных нормальным распределением, ряд Грама – Шарлье дает более «точное» описание данных.

5.4. Аппроксимация на основе универсальных семейств распределений

Существуют различные подходы к построению универсальных семейств распределений. Рассмотрим два наиболее типичных. Первый подход является дальнейшим развитием метода моментов, а второй основан на замене исходной выборки другой, распределение которой является стандартным.

Аппроксимация на основе семейства распределений К. Пирсона

В рамках первого подхода одно из универсальных семейств распределений предложил К. Пирсон. Моменты распределения случайной величины, даже если все они существуют, не характеризуют полностью этого распределения, но они определяют его однозначно при некоторых условиях, которые выполняются почти для всех используемых на практике распределений. Иначе говоря, при решении задач обработки экспериментальных данных знание моментов эквивалентно знанию функции распределения, и совпадение значений первых r моментов двух распределений говорит о приблизительной одинаковости распределений. Не зная точно вид функции распределения, но найдя r первых моментов, можно подобрать другое распределение с теми же первыми моментами. Практически такая аппроксимация оказывается хорошей при совпадении первых трех-четырех моментов.

Анализ характерных черт функций плотности унимодальных распределений показывает, что эти распределения начинаются с нуля, поднимаются до максимума, а затем уменьшаются снова до нуля. Это означает, что

для описания подобных функций плотности распределений $f(x)$ необходимо выбрать такие уравнения, для которых $df(x)/dx = 0$ при следующих условиях:

1) $f(x) = 0$, тогда, по крайней мере, на одном краю распределения будет соприкосновение с осью абсцисс высшего порядка;

2) $x = a$, где величина a соответствует моде распределения.

Этим условиям для *центрированной* переменной x удовлетворяет дифференциальное уравнение $df/dx = (x-a)f\Big/\left(b_0 + b_1 x + b_2 x^2\right)$, решение которого приводит к семейству *распределений Пирсона.* Действительно, в этом уравнении $df(x)/dx$ равно нулю, если $f(x) = 0$ или $x = a$. Семейство распределений Пирсона включает не только унимодальные распределения, но и распределения, имеющие U-образную форму (две моды).

Уравнение содержит четыре неизвестных параметра. Их вычисление основано на методе моментов – четыре выборочных момента приравниваются к соответствующим моментам теоретического распределения, являющимся функциями от неизвестных параметров. Решая полученную систему уравнений относительно неизвестных параметров, получают искомые оценки параметров в виде функций выборочных моментов

$$\begin{aligned} a &= \mu_3\left(\mu_4 + 3\mu_2^2\right)\Big/A, \\ B_0 &= -\mu_2\left(4\mu_2\mu_4 - 3\mu_3^2\right)\Big/A, \\ B_1 &= -\mu_3\left(\mu_4 + 3\mu_2^2\right)\Big/A, \\ B_2 &= -\left(2\mu_2\mu_4 - 3\mu_3^2 - 6\mu_2^3\right)\Big/A, \\ A &= 10\mu_2\mu_4 - 18\mu_2^3 - 12\mu_3^2. \end{aligned} \qquad (5.3)$$

Выражения для плотности $f(x)$ выводятся путем интегрирования дифференциального уравнения. Интегрирование позволяет получить одиннадцать типов функций плотности распределения, три из которых являются основными, а восемь – их частными случаями, в том числе, и такие общеизвестные, как нормальное, экспоненциальное, гамма-распределение. Распределение $f(x)$ сосредоточено:

– на конечном интервале, если корни уравнения $B_0 + B_1 x + B_2 x^2 = 0$ представляют собой действительные числа различных знаков;

– на положительной полупрямой, если корни – действительные числа одного знака и $a > 0$, или на отрицательной полупрямой при $a < 0$;

– на всей оси абсцисс, если уравнение не имеет действительных корней.

Принимая моду за начало отсчета исходной центрированной величины, т.е. полагая $t = x - a$, исходное уравнение представим в виде

$$\frac{d}{dt}\left(\ln f(t)\right) = t \Big/ \left(B_0 + B_1 t + B_2 t^2\right).$$

Первый основной тип распределения получается в случае, когда корни уравнения $B_0 + B_1 t + B_2 t^2 = 0$ являются действительными числами с различными знаками. Обозначим корни уравнения через $-c_1$ и c_2 соответственно, причем, величины c_1 и c_2 – положительные числа. Тогда по известной теореме $B_0 + B_1 t + B_2 t^2 = B_2\left(t + c_1\right)\left(t - c_2\right)$.

Исходное уравнение преобразуем к виду

$$\frac{d}{dt}\left(\ln f_1(t)\right) = \frac{t}{B_2\left(t + c_1\right)\left(t - c_2\right)} = \frac{c_1}{B_2\left(c_1 + c_2\right)}\frac{1}{\left(t + c_1\right)} + \frac{c_2}{B_2\left(c_1 + c_2\right)}\frac{1}{\left(t - c_2\right)}.$$

Обозначим $\gamma = c_1 \Big/ \left(B_2\left(c_1 + c_2\right)\right)$ и $\eta = c_2 \Big/ \left(B_2\left(c_1 + c_2\right)\right)$. Тогда можно записать $d\left(\ln f_1(t)\right) = d\left[\ln\left(t + c_1\right)^{\gamma} + \ln\left(c_2 - t\right)^{\eta}\right]$. Решение дифференциального уравнения с точностью до некоторого коэффициента k_1 можно представить в виде $f_1(t) = k_1\left(c_1 + t\right)^{\gamma}\left(c_2 - t\right)^{\eta}$. Размах данного распределения сосредоточен на интервале $\left(-c_1,\ c_2\right)$. Проведем замену переменной $t = \left(c_1 + c_2\right)y - c_1$, учитывая, что $dt = \left(c_1 + c_2\right)dy$, включим постоянный сомножитель $\left(c_1 + c_2\right)^{\gamma+\eta+1}$ в состав коэффициента k_1. В итоге получим $f_1(y) = k_1 y\gamma(1 - y)\eta$, где y изменяется в пределах от 0 до 1. Интегрируя в этих пределах функцию $f_1(t)$, можно найти значение k_1 из условия $\int_0^1 k_1 y^{\gamma}\left(1 - y\right)^{\eta} dy = 1$. Интеграл в данном выражении по определению соответствует бета-функции $B\left(\gamma + 1,\ \eta + 1\right)$, которая определяется через гамма-функцию $B\left(\gamma + 1,\ \eta + 1\right) = \Gamma\left(\gamma + 1\right)\Gamma\left(\eta + 1\right) \Big/ \Gamma\left(\gamma + \eta + 2\right)$. Итак, $k_1 = 1 \Big/ B\left(\gamma + 1,\ \eta + 1\right)$. Окончательно плотность распределения

$$f_1(y) = \left(1 \Big/ B\left(\gamma + 1,\ \eta + 1\right)\right) y^{\gamma}\left(1 - y\right)^{\eta}, \qquad (5.4)$$

где $0 \le y \le 1$.

Переменная y определяется через исходный (не центрированный и несмещенный) аргумент x в соответствии с ранее введенными подстановками: $y = \left(c_1 + x - \mu_1 - a\right) \Big/ \left(c_1 + c_2\right)$.

Функция плотности распределения первого типа соответствует бета-распределению (рис. 5.5). Функция распределения

$$F_1(y)=\frac{\Gamma(\gamma+\eta+2)}{\Gamma(\gamma+1)\Gamma(\eta+1)}\int_0^y y^{\gamma}(1-y)^{\eta}\,dy. \qquad (5.5)$$

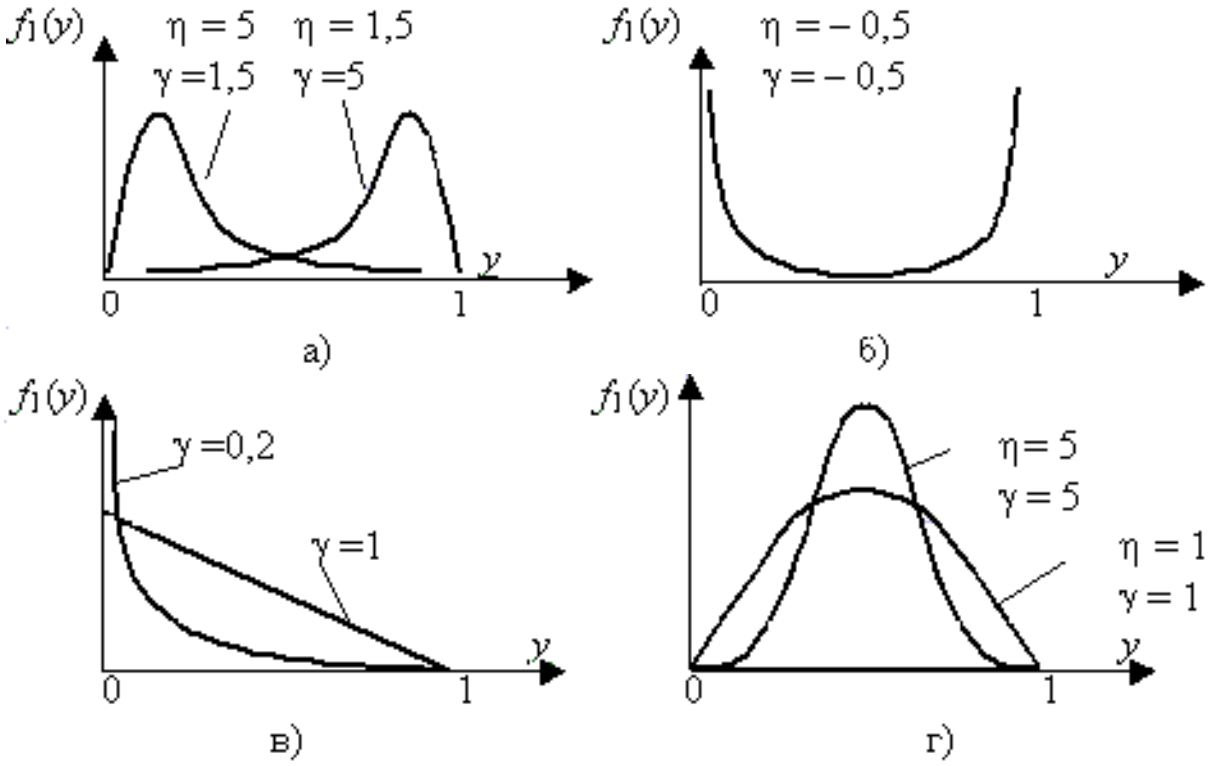

Рис. 5.5. Распределение Пирсона первого типа (бета-распределение):
а – $\eta>1$, $\gamma>1$; б – $\eta<1$, $\gamma<1$; в – $\eta=2$, $\gamma\le 1$; г – $\eta=\gamma\ge 1$

При наличии действительных корней одного знака получается распределение Пирсона шестого типа. Пусть корни $-c_1$ и $-c_2$ меньше нуля, т.е. B_2, c_1 и c_2 положительны $(c_1<c_2)$, тогда можно записать

$$\frac{d}{dt}\left(\ln f_6(t)\right)=\frac{-c_1}{B_2(-c_1+c_2)}\frac{1}{(t+c_1)}+\frac{c_2}{B_2(-c_1+c_2)}\frac{1}{(t+c_2)},$$

где $-c_1<t<\infty$. Обозначим $\alpha=-c_1/B_2(c_1-c_2)$ и $\beta=c_2/B_2(c_1-c_2)$.

После преобразований получим $d\left(\ln f_6(t)\right)=d\left[\ln\left[(c_1+t)^{\alpha}(c_2+t)^{\beta}\right]\right]$ или $f_6(t)=k_6(c_1+t)\alpha(c_2+t)\beta$. Здесь, как и для распределения первого типа, $t=x-\mu_1-a$.

Используем подстановку $(c_1-c_2)/(c_2+t)$, тогда $dt=-(c_2-c_1)z^{-2}dz$. Функция плотности распределения шестого типа примет вид $f_6(t)=k_6(1-z)\alpha\cdot z^{-(\alpha+\beta+2)}$. Нормировочный коэффициент определяется

аналогично ранее рассмотренному варианту. Нормирующее условие имеет вид $1 = k_6 \int_0^1 z^{-(\alpha+\beta+2)} (1-z)^{\alpha}\, dz$. Следовательно, коэффициент определяется через бета-функцию: $k_6 = 1/B(-\alpha-\beta-1,\ \alpha+1)$.

Окончательно функция плотности распределения шестого типа

$$f_6(z) = 1/\left[B(-\alpha-\beta-1,\ \alpha+1)\right] z^{-(\alpha+\beta+2)} (1-z)\alpha. \qquad (5.6)$$

Функция распределения шестого типа

$$F_6(z) = 1 - \frac{\Gamma(-\beta)}{\Gamma(-\alpha-\beta-1)\Gamma(\alpha+1)} \int_0^z z^{-(\alpha+\beta+2)} (1-z)^{\alpha}\, dz. \qquad (5.7)$$

Для положительных корней уравнения $B_0 + B_1 t + B_2 t^2 = 0$ диапазон изменения аргумента $-\infty < t < c_1$, а выражения для плотности и функции распределения получаются такие же, только при выводе используется другая подстановка $z = (c_2 - c_1)/(c_2 - t)$. Таким образом, шестой тип распределения является разновидностью первого типа.

Функции распределения (5.5) и (5.7) представляют собой неполные бета-функции $B_y(p,\ q)$. Когда оба показателя степени в формулах (5.4) и (5.6) больше нуля, плотность имеет единственную моду и обращается в нуль на краях интервала. Если один из показателей отрицателен, то значение плотности на одном краю интервала стремится к бесконечности и распределение имеет L- или J-образную форму. При двух отрицательных показателях распределения принимают U-образную форму, значения функций плотности стремятся к бесконечности на обоих краях. В указанных случаях применение численного интегрирования для вычисления значений функций распределения невозможно.

Вычисления значений функций распределения первого и шестого типов целесообразно осуществлять разложением интеграла (неполной бета-функции) в гипергеометрический ряд. Гипергеометрический ряд

$$F(a,b,c,w) = 1 + \frac{ab}{1\cdot c} w + \frac{a(a+1)b(b+1)}{1\cdot 2\cdot c(c+1)} w^2 + \ldots \qquad (5.8)$$

сходится абсолютно и равномерно при $|w| < 1$. Для ускорения сходимости ряда неполную бета-функцию вычисляют по различным формулам в зависимости от значения предела интегрирования

$$B_z(p,\ q) = \begin{cases} \dfrac{1}{p} z^p (1-z)^q F(1,\ p+q,\ p+1,\ z), & \text{при } z \le 0{,}5, \\ 1 - B_{1-z}(p,q), & \text{при } z > 0{,}5. \end{cases} \qquad (5.9)$$

В формуле (5.9) для распределения первого типа $p=\gamma+1$ и $q=\eta+1$, а для распределения шестого типа $p=-\alpha-\beta-1$, $q=\alpha+1$.

Если корни уравнения $B_0+B_1t+B_2t^2=0$ комплексные числа, то получается распределение Пирсона четвертого типа с диапазоном изменения переменной по всей оси абсцисс и единственной модой. Путем тождественных преобразований, вводя соответствующие обозначения, исходное дифференциальное уравнение представим в виде

$$\frac{d}{dt}\left(\ln f_4(t)\right)=\frac{t}{B_2\left\{\left(t+\frac{B_1}{2B_2}\right)^2+\frac{B_0}{B_2}-\frac{B_1^2}{4B_2^2}\right\}}=\frac{t}{B_2\left(\left(t+\varphi\right)^2+\delta^2\right)},$$

где $\varphi=B_1/\left(2B_2\right)$, $\delta^2=B_0/B_2-B_1^2/\left(4B_2^2\right)$.

Используя правила интегрирования элементарных дробей, уравнение преобразуем к виду

$$\ln f_4(t)=\ln R+\frac{1}{2B_2}\ln\left(\left(t+\varphi\right)^2+\delta^2\right)-\frac{\varphi}{B_2\delta}\operatorname{arctg}\frac{t+\varphi}{\delta}.$$

Следовательно, функция плотности четвертого типа

$$f_4(t)=R\left\{\left(t+\varphi\right)^2+\delta^2\right\}^{1/(2B_2)}\cdot\exp\left\{-\varphi/\left(B_2\delta\right)\cdot\operatorname{arctg}\left(\left(t+\varphi\right)/\delta\right)\right\}. \quad (5.10)$$

Коэффициент R находится из нормирующего условия (интеграл от плотности распределения в пределах изменения переменной равен единице). Для вычисления коэффициента приходится проводить численное интегрирование, так как первообразная функция через элементарные функции не представима. Чтобы перейти к конечным пределам при численном интегрировании, воспользуемся заменой переменной $\nu=\operatorname{arctg}\left(\frac{t+\varphi}{\delta}\right)$, тогда интегрирование следует провести в пределах от $-\pi/2$ до $\pi/2$ (здесь, как и ранее, $t=x-\mu_1-a$). Окончательно получим

$$F_4(\nu)=\frac{\int\limits_{-\pi/2}^{\nu}\cos^{-2\left(1/(2B_2)+1\right)}(\nu)\exp\left(-\varphi\nu/\left(B_2\delta\right)\right)d\nu}{\int\limits_{-\pi/2}^{\pi/2}\cos^{-2\left(1/(2B_2)+1\right)}(\nu)\exp\left(-\varphi\nu/\left(B_2\delta\right)\right)d\nu}. \quad (5.11)$$

Последовательность подгонки описания эмпирических данных распределениями Пирсона включает следующие этапы:

– вычисление значений оценок первых четырех моментов эмпирического распределения путем обработки экспериментальных данных;

– вычисление параметров B_0, B_1, B_2, a семейства распределений, переход от исходной переменной x к центрированной и смещенной переменной t;

– анализ корней квадратного уравнения B_0, B_1, B_2 и определение типа распределения. При этом реальная область значений случайной величины играет второстепенную роль. Например, четвертое распределение Пирсона может служить хорошей аппроксимацией распределения ограниченной случайной величины, или, наоборот, первое распределение – для случайной величины с бесконечными пределами изменения;

– вычисление параметров выбранного типа распределения;

– проверку гипотезы о возможности применения выбранного распределения для описания экспериментальных данных.

Распределения Пирсона вполне удовлетворительно обобщают результаты наблюдений. Но эти оценки не являются наилучшими, так как имеют неминимальные дисперсии, а, следовательно, не являются наилучшими оценками параметров генеральной совокупности.

Области в плоскости квадрата коэффициента асимметрии b_1^2 и коэффициента эксцесса b_2, соответствующие различным распределениям семейства Пирсона, показаны на рис. 5.6. Из рисунка видно, что распределения Пирсона охватывают широкую область возможных видов распределений и включают в себя как частные случаи, – нормальное (н.р.), экспоненциальное (э.р.), гамма (г.р.) и другие типовые распределения. Нормальное и экспоненциальное распределения не имеют параметров формы, поэтому на рисунке отображаются точками, гамма-распределение имеет только один параметр формы и ему соответствует линия. Иначе говоря, типовые распределения обладают скромными возможностями по аппроксимации экспериментальных данных.

Рис. 5.6. Области аппроксимации экспериментальных данных семейством распределений Пирсона

Недостаток рассмотренного метода состоит в большой трудоемкости расчетов значений функции распределения.

Пример 5.2. Необходимо подобрать распределение Пирсона для описания экспериментальных данных, табл. 2.4, и оценить качество аппроксимации. Проверку согласованности провести с использованием критерия хи-квадрат при уровне значимости $\alpha = 0,05$.

Решение. Значения оценок моментов были вычислены ранее:

$$\mu_1 = 27,508,\ \mu_2 = 0,913,\ \mu_3 = 0,132,\ \mu_4 = 1,819.$$

По формулам (5.3) вычислим параметры распределения:

$$A = 2,6995;\ a = 0,2112;\ B_0 = -2,2290;\ B_1 = -0,2112;\ B_2 = 0,4804.$$

Корни уравнения $b_0 + b_1 x + b_2 x^2 = 0$ – действительные числа различных знаков: $-c_1 = -1,945;\ c_2 = 2,385$. Значит, распределение относится к первому типу и сосредоточено на ограниченном интервале. Построим табл. 5.4, иллюстрирующую расчеты.

Таблица 5.4

i	1	2	3	4	5	6
n_i	5	9	10	9	5	6
Верхняя граница, x_i	26,37	26,95	27,53	28,11	28,69	∞
$F(x_i)$	0,165	0,348	0,550	0,740	0,892	1
ΔF_i	0,165	0,183	0,202	0,190	0,152	0,108
F_i	7,260	8,052	8,888	8,360	6,688	4,752
$(n_i - F_i)^2 / F_i$	0,703	0,112	0,139	0,049	0,426	0,327

В таблице значения функции распределения $F(x_i)$ для верхней границы интервала и теоретическое значение оценки вероятности ΔF_i попадания случайной величины в i-й интервал вычислены на основе распределения Пирсона первого типа. Расчет оценки частоты $F_i = \Delta F_i \cdot n$, вероятности ΔF_i попадания случайной величины в интервал $x_i - x_{i-1}$, взвешенного квадрата отклонения $(n_i - F_i)^2 / F_i$ проводится аналогично примеру 5.1. Значение критерия составляет $\chi^2 = 1,757$.

По сравнению с критическим значением хи-квадрат, равным 7,815, аппроксимация с помощью распределения Пирсона дает вполне допустимый результат, хотя в данном случае и уступает по «точности» аппроксимации с помощью ряда Грама – Шарлье ($\chi^2 = 0,872$). Повысить точность аппроксимации можно, если проанализировать плотность аппроксимирующего распределения. Полученная функция плотности имеет небольшой коэффициент эксцесса, поэтому наблюдаются относительно большие отклонения

функции распределения от экспериментальных данных. Такая ситуация является следствием значительной погрешности в оценке четвертого момента из-за ограниченного объема выборки. Следовательно, для повышения качества аппроксимации необходимо увеличить значение четвертого момента. Увеличим значение четвертого момента до 2,2 (ошибки в 20-25 % при оценке четвертого момента по выборке малого объема вполне реальны) и пересчитаем все параметры. В результате получится значение $\chi^2 = 0{,}864$, что практически одинаково с аппроксимацией рядом Грама – Шарлье.

Потенциально аппроксимация по Пирсону является более универсальной по сравнению с рядами Грама – Шарлье. Семейство Пирсона охватывает широкий класс законов распределений, а не только близкие к нормальному, как это имеет место при применении рядов.

Задачи.

5.1. Оценить качество аппроксимации сгруппированных экспериментальных данных, задача 2.1, на основе ряда Грама – Шарлье. Проверку согласованности провести с использованием критерия хи-квадрат при уровне значимости $\alpha = 0{,}5$.

5.2. Оценить качество аппроксимации сгруппированных экспериментальных данных, задача 2.2, на основе ряда Грама – Шарлье. Проверку согласованности провести с использованием критерия хи-квадрат при уровне значимости $\alpha = 0{,}1$.

5.3. Необходимо подобрать распределение Пирсона для описания сгруппированных экспериментальных данных, задача 2.1, и оценить качество аппроксимации. Проверку согласованности провести с использованием критерия хи-квадрат при уровне значимости $\alpha = 0{,}05$.

5.4. Необходимо подобрать распределение Пирсона для описания сгруппированных экспериментальных данных, задача 2.2, и оценить качество аппроксимации. Проверку согласованности провести с использованием критерия хи-квадрат при уровне значимости $\alpha = 0{,}01$.

6. ОДНОФАКТОРНЫЙ ДИСПЕРСИОННЫЙ АНАЛИЗ

Дисперсионный анализ (от латинского Dispersio – рассеивание) – статистический метод, позволяющий анализировать влияние различных факторов на исследуемую переменную. Метод был разработан биологом Р. Фишером в 1925 году и применялся первоначально для оценки экспериментов в растениеводстве. В дальнейшем выяснилась общенаучная значимость дисперсионного анализа для экспериментов в психологии, педагогике, медицине и др. В литературе также встречается обозначение ANOVA (от англ. *ANalysis Of VAriance*).

Целью дисперсионного анализа является проверка значимости различия между средними с помощью сравнения дисперсий. Дисперсию измеряемого признака разлагают на независимые слагаемые, каждое из которых характеризует влияние того или иного фактора или их взаимодействия. Последующее сравнение таких слагаемых позволяет оценить значимость каждого изучаемого фактора, а также их комбинации.

При истинности нулевой гипотезы (о равенстве средних в нескольких группах наблюдений, выбранных из генеральной совокупности), оценка дисперсии, связанной с внутригрупповой изменчивостью, должна быть близкой к оценке межгрупповой дисперсии.

При проведении исследования рынка часто встает вопрос о сопоставимости результатов. Например, проводя опросы по поводу потребления какого-либо товара в различных регионах страны, необходимо сделать выводы, насколько данные опроса отличаются или не отличаются друг от друга. Сопоставлять отдельные показатели не имеет смысла и поэтому процедура сравнения и последующей оценки производится по некоторым усредненным значениям и отклонениям от этой усредненной оценки. Изучается вариация признака. За меру вариации может быть принята дисперсия. Дисперсия σ^2 – мера вариации, определяемая как средняя из отклонений признака, возведенных в квадрат.

На практике часто возникают задачи более общего характера – задачи проверки существенности различий средних выборочных нескольких совокупностей. Например, требуется оценить влияние различного сырья на качество производимой продукции, решить задачу о влиянии количества удобрений на урожайность с/х продукции.

Иногда дисперсионный анализ применяется, чтобы установить однородность нескольких совокупностей (дисперсии этих совокупностей одинаковы по предположению; если дисперсионный анализ покажет, что и математические ожидания одинаковы, то в этом смысле совокупности однородны). Однородные же совокупности можно объединить в одну и, тем самым, получить о ней более полную информацию, следовательно, и более надежные выводы.

Основные понятия дисперсионного анализа

В процессе наблюдения за исследуемым объектом качественные факторы произвольно или заданным образом изменяются. Конкретная реализация фактора (например, определенный температурный режим, выбранное оборудование или материал) называется уровнем фактора или способом обработки. Модель дисперсионного анализа с фиксированными уровнями факторов называют моделью I, модель со случайными факторами – моделью II. Благодаря варьированию фактора можно исследовать его влияние на величину отклика. В настоящее время общая теория дисперсионного анализа разработана для моделей I.

В зависимости от количества факторов, определяющих вариацию результативного признака, дисперсионный анализ подразделяют на однофакторный и многофакторный.

Основными схемами организации исходных данных с двумя и более факторами являются:

– перекрестная классификация, характерная для моделей I, в которых каждый уровень одного фактора сочетается при планировании эксперимента с каждой градацией другого фактора;

– иерархическая (гнездовая) классификация, характерная для модели II, в которой каждому случайному, наудачу выбранному значению одного фактора соответствует свое подмножество значений второго фактора.

Если одновременно исследуется зависимость отклика от качественных и количественных факторов, т.е. факторов смешанной природы, то используется ковариационный анализ.

При обработке данных эксперимента наиболее разработанными и поэтому распространенными считаются две модели. Их различие обусловлено спецификой планирования самого эксперимента. В модели дисперсионного анализа с фиксированными эффектами исследователь намеренно устанавливает строго определенные уровни изучаемого фактора. Термин «фиксированный эффект» в данном контексте имеет тот смысл, что самим исследователем фиксируется количество уровней фактора и различия между ними. При повторении эксперимента он или другой исследователь выберет те же самые уровни фактора. В модели со случайными эффектами уровни значения фактора выбираются исследователем случайно из широкого диапазона значений фактора, и при повторных экспериментах, естественно, этот диапазон будет другим.

Таким образом, данные модели отличаются между собой способом выбора уровней фактора, что, очевидно, в первую очередь влияет на возможность обобщения полученных экспериментальных данных. Для дисперсионного анализа однофакторных экспериментов различие этих двух моделей не столь существенно, однако в многофакторном дисперсионном анализе оно может оказаться весьма важным.

При проведении дисперсионного анализа должны выполняться следующие статистические допущения: независимо от уровня фактора величины отклика имеют нормальный (Гауссовский) закон распределения и одинаковую дисперсию. Такое равенство дисперсий называется гомогенностью. Таким образом, изменение способа обработки сказывается лишь на положении случайной величины отклика, которое характеризуется средним значением или медианой. Поэтому все наблюдения отклика принадлежат сдвиговому семейству нормальных распределений.

Говорят, что техника дисперсионного анализа является «робастной». Этот термин, используемый статистиками, означает, что данные допущения могут быть в некоторой степени нарушены, но, несмотря на это, технику можно использовать.

При неизвестном законе распределения величин отклика используют непараметрические (чаще всего ранговые) методы анализа.

Типы дисперсионного анализа

Суть дисперсионного анализа сводится к изучению влияния одной или нескольких независимых переменных, обычно именуемых факторами, на зависимую переменную. Зависимые переменные представлены в виде шкал. Независимые переменные являются номинативными, то есть отражают групповую принадлежность, и могут иметь две или более градации (или уровня). Градации, соответствующие независимым выборкам объектов, называются межгрупповыми, а градации, соответствующие зависимым выборкам, — внутригрупповыми.

В зависимости от типа и количества переменных различают:

– однофакторный и многофакторный дисперсионный анализ (одна или несколько независимых переменных);

– одномерный и многомерный дисперсионный анализ (одна или несколько зависимых переменных);

– дисперсионный анализ с повторными измерениями (для зависимых выборок);

– дисперсионный анализ с постоянными факторами, случайными факторами, и смешанные модели с факторами обоих типов.

Принципы и применение

Исходными положениями дисперсионного анализа являются:

– нормальное распределение зависимой переменной;

– равенство дисперсий в сравниваемых генеральных совокупностях;

– случайный и независимый характер выборки.

Нулевой гипотезой в дисперсионном анализе является утверждение о равенстве средних значений:

$$H_0: \ \mu_1 = \mu_2 = ... = \mu_j .$$

При отклонении нулевой гипотезы принимается альтернативная гипотеза о том, что не все средние равны, то есть имеются, по крайней мере, две группы, отличающиеся средними значениями:

$$H_0: \ \mu_1 \neq \mu_2 \neq ... \neq \mu_j .$$

Однофакторный дисперсионный анализ

Простейшим случаем дисперсионного анализа является одномерный однофакторный анализ для двух или нескольких независимых групп, когда все группы объединены по одному признаку. В ходе анализа проверяется нулевая гипотеза о равенстве средних [5]. При анализе двух групп дисперсионный анализ тождественен двухвыборочному t-критерию Стьюдента для независимых выборок, и величина F-статистики равна квадрату соответствующей t-статистики.

Пусть генеральные совокупности $X_1, X_2, ..., X_p$ распределены нормально и имеют одинаковую, хотя и неизвестную дисперсию; математические ожидания также неизвестны, но могут быть различными. Требуется при заданном уровне значимости по выборочным средним проверить нулевую гипотезу $H_0: \ M(X_1) = M(X_2) = ... = M(X_p)$ о равенстве всех математических ожиданий. Другими словами, требуется установить, значимо или незначимо различаются выборочные средние. Казалось бы, для сравнения нескольких средних $(p > 2)$ можно сравнить их попарно. Однако, с возрастанием числа средних возрастает и наибольшее различие между ними: среднее новой выборки может оказаться больше наибольшего или меньше наименьшего из средних, полученных до нового опыта. По этой причине для сравнения нескольких средних пользуются другим методом, который основан на сравнении дисперсий и поэтому назван дисперсионным анализом (в основном развит английским статистиком Р.Фишером).

Основная идея дисперсионного анализа состоит в сравнении межгрупповой (факторной) дисперсии, порождаемой воздействием фактора, и внутригрупповой (остаточной) дисперсии, обусловленной случайными причинами. Если различие между этими дисперсиями значимо, то фактор оказывает существенное влияние на X; в этом случае средние наблюдаемых значений на каждом уровне (групповые средние) различаются также значимо.

Если уже установлено, что фактор существенно влияет на X, а требуется выяснить, какой из уровней оказывает наибольшее воздействие, то дополнительно производят попарное сравнение средних.

Иногда дисперсионный анализ применяется, чтобы установить однородность нескольких совокупностей (дисперсии этих совокупностей одинаковы по предположению; если дисперсионный анализ покажет, что и математические ожидания одинаковы, то в этом смысле совокупности од-

нородны). Однородные же совокупности можно объединить в одну, и тем самым получить о ней более полную информацию, следовательно, и более надежные выводы.

В более сложных случаях исследуют воздействие нескольких факторов на нескольких постоянных или случайных уровнях и выясняют влияние отдельных уровней и их комбинаций (*многофакторный анализ*).

Изучим простейший случай однофакторного анализа, когда на X воздействует фактор F, который имеет p постоянных уровней. При этом предположим, что число испытаний (экспериментальных данных) на каждом уровне одинаково и равно q.

Пусть наблюдалось $n = pq$ значений x_{ij} признака X, где i – номер испытания $(i = 1,\ 2,\ ...,\ p)$, j – номер уровня фактора (группы) $(j = 1,\ 2,\ ...,\ q)$. Результаты наблюдений приведены в табл. 6.1.

Таблица 6.1

Номер испытания	Уровни фактора F_j			
	F_1	F_2	…	F_p
1	x_{11}	x_{12}	…	x_{1p}
2	x_{21}	x_{22}	…	x_{2p}
…	…	…	…	…
q	x_{q1}	x_{q2}	…	x_{qp}
Групповая средняя	$\overline{x}_{\text{гр}1}$	$\overline{x}_{\text{гр}2}$	…	$\overline{x}_{\text{гр}p}$

По определению, *общая сумма* квадратов отклонений наблюдаемых значений от общей средней $\overline{x}$:

$$S_{\text{общ}} = \sum_{j=1}^{p}\sum_{i=1}^{q}\left(x_{ij} - \overline{x}\right)^2,$$

факторная (межгрупповая) сумма квадратов отклонений групповых средних от общей средней, характеризующей рассеяние «между группами»:

$$S_{\text{факт}} = q\sum_{j=1}^{p}\left(\overline{x_{\text{гр}\,j}} - \overline{x}\right)^2,$$

остаточная (внутригрупповая) сумма квадратов отклонений наблюдаемых значений группы от своей групповой средней, характеризующей рассеяние «внутри группы»:

$$S_{\text{ост}} = \sum_{i=1}^{q}\left(x_{i1} - \overline{x}_{\text{гр}1}\right)^2 + \sum_{i=1}^{q}\left(x_{i2} - \overline{x}_{\text{гр}2}\right)^2 + ... + \sum_{i=1}^{q}\left(x_{ip} - \overline{x}_{\text{гр}p}\right)^2.$$

На практике остаточную сумму находят по равенству:

$$S_{\text{ост}} = S_{\text{общ}} - S_{\text{факт}}.$$

Элементарными преобразованиями можно получить формулы, более удобные для расчетов:

$$S_{\text{общ}} = \sum_{j=1}^{p} P_j - \frac{\left(\sum_{j=1}^{p} R_j\right)^2}{pq}; \tag{6.1}$$

$$S_{\text{факт}} = \frac{\sum_{j=1}^{p} R_j^2}{q} - \frac{\left(\sum_{j=1}^{p} R_j\right)^2}{pq}, \tag{6.2}$$

где $P_j = \sum_{i=1}^{q} x_{ij}^2$ – сумма квадратов значений признака на уровне F_j; $R_j = \sum_{i=1}^{q} x_{ij}$ – сумма значений признака на уровне F_j.

Для упрощения вычислений вычитают из каждого наблюдаемого значения одно и то же число C, примерно равное общей средней. Если принять $y_{ij} = x_{ij} - C$, то

$$S_{\text{общ}} = \sum_{j=1}^{p} Q_j - \frac{\left(\sum_{j=1}^{p} T_j\right)^2}{pq}; \tag{6.3}$$

$$S_{\text{факт}} = \frac{\sum_{j=1}^{p} T_j^2}{q} - \frac{\left(\sum_{j=1}^{p} T_j\right)^2}{pq}, \tag{6.4}$$

где $Q_j = \sum_{i=1}^{q} y_{ij}^2$ – сумма квадратов уменьшенных значений признака на уровне F_j; $T_j = \sum_{i=1}^{q} y_{ij}$ – сумма уменьшенных значений признака на уровне F_j.

Для вывода формул (6.3) и (6.4) достаточно подставить $x_{ij} = y_{ij} + C$ в отношение (6.1) и $R_j = \sum_{i=1}^{q} x_{ij} = \sum_{i=1}^{q} \left(y_{ij} + C \right) = \sum_{i=1}^{q} y_{ij} + qC = T_j + qC$ в отношение (6.2).

Разделив суммы квадратов отклонений на соответствующее число степеней свободы, можно получить общую, факторную и остаточную дисперсии:

$$s^2_{\text{общ}} = \frac{S_{\text{общ}}}{pq-1}, \quad s^2_{\text{факт}} = \frac{S_{\text{факт}}}{p-1}, \quad s^2_{\text{ост}} = \frac{S_{\text{ост}}}{p\left(q-1\right)},$$

где p – число уровней фактора; q – число наблюдений на каждом уровне; $pq-1$ – число степеней свободы общей дисперсии; $p-1$ – число степеней свободы факторной дисперсии; $p\left(q-1\right)$ – число степеней свободы остаточной дисперсии.

Если нулевая гипотеза о равенстве средних справедлива, то все эти дисперсии являются несмещенными оценками генеральной дисперсии. Например, учитывая, что объем выборки $n = pq$, получаем, что

$$s^2_{\text{общ}} = \frac{S_{\text{общ}}}{pq-1} = \frac{S_{\text{общ}}}{n-1} \; -$$

исправленная выборочная дисперсия, являющаяся несмещенной оценкой генеральной дисперсии.

При этом число степеней свободы $p\left(q-1\right)$ остаточной дисперсии равно разности между числами степеней свободы общей и факторной дисперсий:

$$\left(pq-1\right) - \left(p-1\right) = pq - p = p\left(q-1\right).$$

Вернемся к нашей задаче: проверить при заданном уровне значимости нулевую гипотезу о равенстве нескольких ($p > 2$) средних нормальных совокупностей с неизвестными, но одинаковыми дисперсиями. Посмотрим, что решение этой задачи сводится к сравнению факторной и остаточной дисперсий по критерию Фишера – Снедекора:

1. Пусть нулевая гипотеза о равенстве нескольких средних (будем называть их групповыми) правильна. В этом случае факторная и остаточная дисперсии являются несмещенными оценками неизвестной генеральной дисперсии и, следовательно, различаются незначимо. Если сравнить эти оценки по критерию F, то, очевидно, критерий укажет, что нулевую гипотезу о равенстве факторной и остаточной дисперсий надо принять. Таким образом, если гипотеза о равенстве групповых средних правильна, то и верна гипотеза о равенстве факторной и остаточной дисперсий.

2. Пусть нулевая гипотеза о равенстве групповых средних ложна. В этом случае с возрастанием расхождения между групповыми средними увеличивается факторная дисперсия, а вместе с ней и отношение $F_{\text{набл}} = s^2_{\text{факт}} \big/ s^2_{\text{ост}}$. В итоге $F_{\text{набл}}$ окажется больше $F_{\text{кр}}$ и, следовательно, гипотеза о равенстве дисперсий будет отвергнута. Таким образом, если гипотеза о равенстве групповых средних ложна, то ложна и гипотеза о равенстве факторной и остаточной дисперсий.

Легко доказать от противного справедливость обратных утверждений: из правильности (ложности) гипотезы о дисперсиях следует правильность (ложность) гипотезы о средних.

Таким образом, *для того чтобы проверить нулевую гипотезу о равенстве групповых средних нормальных совокупностей с одинаковыми дисперсиями, достаточно проверить по критерию F нулевую гипотезу о равенстве факторной и остаточной дисперсий.* В этом и состоит метод дисперсионного анализа.

При этом если факторная дисперсия окажется меньше остаточной, то уже отсюда следует справедливость гипотезы о равенстве групповых средних и, следовательно, нет необходимости прибегать к критерию F.

Если нет уверенности в справедливости предположения о равенстве дисперсий рассматриваемых p совокупностей, то это предположение следует проверить по критерию Кохрена.

Пример 6.1. Произведено по 4 испытания на каждом из трех уровней. Результаты испытаний приведены в табл. 6.2. Методом дисперсионного анализа при уровне значимости 0,05 проверить нулевую гипотезу о равенстве групповых средних. Предполагается, что выборки извлечены из нормальных совокупностей с одинаковыми дисперсиями.

Таблица 6.2

Номер испытания	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
$\overline{x}_{\text{гр}j}$	54	55	47

Решение. Для упрощения расчета вычтем $C = 52$ из каждого наблюдаемого значения: $y_{ij} = x_{ij} - 52$. Получим расчетную табл. 6.3.

Таблица 6.3

Номер испытания	Уровни фактора F_j						Итоговый столбец
	F_1		F_2		F_3		
	y_{i1}	y_{i1}^2	y_{i2}	y_{i2}^2	y_{i3}	y_{i3}^2	
1	-1	1	0	0	-10	100	
2	0	0	2	4	-8	64	
3	4	16	4	16	-2	4	
4	5	25	6	36	0	0	
$Q_j = \sum_{i=1}^{4} y_{ij}^2$		42		56		168	$\sum Q_j = 266$
$T_j = \sum_{i=1}^{4} y_{ij}$	8		12		-20		$\sum T_j = 0$
T_j^2	64		144		400		$\sum T_j^2 = 608$

Пользуясь таблицей и учитывая, что число уровней фактора $p = 3$, число испытаний на каждом уровне $q = 4$, найдем общую и факторную суммы квадратов отклонений (согласно (6.3) и (6.4)):

$$S_{\text{общ}} = \sum_{j=1}^{p} Q_j - \frac{\left(\sum_{j=1}^{p} T_j\right)^2}{pq} = 266 - 0 = 266;$$

$$S_{\text{факт}} = \frac{\sum_{j=1}^{p} T_j^2}{q} - \frac{\left(\sum_{j=1}^{p} T_j\right)^2}{pq} = \left(608/4\right) - 0 = 152.$$

Найдем остаточную сумму квадратов отклонений:

$$S_{\text{ост}} = S_{\text{общ}} - S_{\text{факт}} = 266 - 152 = 114.$$

Найдем факторную и остаточную дисперсии:

$$s_{\text{факт}}^2 = \frac{S_{\text{факт}}}{p-1} = \frac{152}{3-1} = 76;$$

$$s_{\text{ост}}^2 = \frac{S_{\text{ост}}}{p\left(q-1\right)} = \frac{114}{3\left(4-1\right)} = \frac{114}{9} = 12,67.$$

Сравним факторную и остаточную дисперсии по критерию F, для чего найдем наблюдаемое значение критерия:

$$F_{\text{набл}} = s_{\text{факт}}^2 \big/ s_{\text{ост}}^2 = 76/12,67 = 6.$$

Учитывая, что число степеней свободы числителя $k_1 = p-1 = 2$, а знаменателя $k_2 = p(q-1) = pq - p = n - p = 9$, и уровень значимости $\alpha = 0{,}05$, по таблице приложения П.6 находим критическую точку:

$$F_{\text{кр}}(0{,}05;\ 2;\ 9) = 4{,}26.$$

Так как $F_{\text{набл}} > F_{\text{кр}}$, – нулевую гипотезу о равенстве нулевых средних отвергаем. Другими словами, групповые средние «в целом» различаются значимо.

Если требуется сравнить средние попарно, следует воспользоваться критерием Стьюдента.

Следует обратить внимание, что если наблюдаемые значения x_{ij} – десятичные дроби с одним знаком после запятой, то целесообразно перейти к числам $y_{ij} = 10x_{ij} - C$, где C – примерно среднее значение чисел $10x_{ij}$. В итоге получим сравнительно небольшие целые числа. При этом следует помнить, что факторная и остаточная дисперсии увеличиваются в 10^2 раз, но их отношение не изменится.

Аналогично поступают, если после десятичной запятой имеется k знаков:

$$y_{ij} = 10^k x_{ij} - C.$$

Мы решили задачу, когда число испытаний на различных уровнях предполагалось одинаковым. Теперь обсудим *неодинаковое число испытаний на различных уровнях*.

Пусть произведено q_1 испытаний на уровне F_1, q_2 испытаний – на уровне F_2, …, q_p испытаний – на уровне F_p. В этом случае общую сумму квадратов отклонений находят по формуле:

$$S_{\text{общ}} = \left[P_1 + P_2 + \ldots + P_p\right] - \left[\left(R_1 + R_2 + \ldots + R_p\right)^2 \Big/ n\right],$$

где $P_1 = \sum_{i=1}^{q_1} x_{i1}^2$ – сумма квадратов наблюдавшихся значений признака на уровне F_1;

$P_2 = \sum_{i=1}^{q_2} x_{i2}^2$ – сумма квадратов наблюдавшихся значений признака на уровне F_2;

. . . . . .

$P_p = \sum_{i=1}^{q_p} x_{ip}^2$ – сумма квадратов наблюдавшихся значений признака на уровне F_p;

$R_1 = \sum_{i=1}^{q_1} x_{i1}$, $R_2 = \sum_{i=1}^{q_2} x_{i2}$, …, $R_p = \sum_{i=1}^{q_p} x_{ip}$ – суммы наблюдавшихся значений признака соответственно на уровнях F_1, F_2, …, F_p;

$n = q_1 + q_2 + ... + q_p$ – общее число испытаний (объем выборки).

Если для упрощения вычислений из каждого наблюдавшегося значения x_{ij} вычитать одно и то же значение C и принять $y_{ij} = x_{ij} - C$, то получим:

$$S_{\text{общ}} = \left[Q_1 + Q_2 + ... + Q_p \right] - \left[\left(T_1 + T_2 + ... + T_p \right)^2 / n \right],$$

где $Q_1 = \sum_{i=1}^{q_1} y_{i1}^2$, $Q_2 = \sum_{i=1}^{q_2} y_{i2}^2$, …, $Q_p = \sum_{i=1}^{q_p} y_{ip}^2$;

$T_1 = \sum_{i=1}^{q_1} y_{i1}$, $T_2 = \sum_{i=1}^{q_2} y_{i2}$, …, $T_p = \sum_{i=1}^{q_p} y_{ip}$.

Факторную сумму квадратов отклонений находим по формуле:

$$S_{\text{факт}} = \left[\left(R_1^2 / q_1 \right) + \left(R_2^2 / q_2 \right) + ... + \left(R_p^2 / q_p \right) \right] - \left[\left(R_1 + R_2 + ... + R_p \right)^2 / n \right];$$

при уменьшении значения признака ($y_{ij} = x_{ij} - C$) получим:

$$S_{\text{факт}} = \left[\left(T_1^2 / q_1 \right) + \left(T_2^2 / q_2 \right) + ... + \left(T_p^2 / q_p \right) \right] - \left[\left(T_1 + T_2 + ... + T_p \right)^2 / n \right].$$

Остальные вычисления производим, как и в случае одинакового числа испытаний:

$$S_{\text{ост}} = S_{\text{общ}} - S_{\text{факт}},$$

$$s_{\text{факт}}^2 = \frac{S_{\text{факт}}}{p-1}, \quad s_{\text{ост}}^2 = \frac{S_{\text{ост}}}{n-p}.$$

Пример 6.2. Произведено 10 испытаний, из них 4 – на первом уровне фактора, 4 – на втором и 2 – на третьем. Результаты испытаний приведены в табл. 6.4. Методом дисперсионного анализа при уровне значимости 0,01 проверить нулевую гипотезу о равенстве групповых средних. Предполагается, что выборки извлечены из нормальных совокупностей с одинаковыми дисперсиями.

Таблица 6.4

Номер испытания	Уровни фактора F_j		
	F_1	F_2	F_3
1	40	62	92
2	44	80	76
3	48	71	
4	36	91	
$\overline{x}_{\text{гр}j}$	42	76	84

Решение. Для упрощения расчета вычтем $C = 67$ из каждого наблюдаемого значения: $y_{ij} = x_{ij} - 67$. Составим расчетную табл. 6.5:

Таблица 6.5

Номер испытания	Уровни фактора F_j						Итоговый столбец
	F_1		F_2		F_3		
i	y_{i1}	y_{i1}^2	y_{i2}	y_{i2}^2	y_{i3}	y_{i3}^2	
1	-27	729	-5	25	25	625	
2	-23	529	13	169	9	81	
3	-19	361	4	16			
4	-31	961	24	576			
$Q_j = \sum_{i=1}^{4} y_{ij}^2$		2580		786		706	$\sum Q_j = 4072$
$T_j = \sum_{i=1}^{4} y_{ij}$	-100		36		34		$\sum T_j = -30$
T_j^2	10000		1296		1156		$\sum T_j^2 = 12452$

Используя табл. 6.5, найдем общую и факторную суммы квадратов отклонений:

$$S_{\text{общ}} = \sum_{j=1}^{p} Q_j - \frac{\left(\sum_{j=1}^{p} T_j\right)^2}{n} = 4072 - \left((-30)^2/10\right) = 4072 - 90 = 3982;$$

$$S_{\text{факт}} = \frac{\sum_{j=1}^{p} T_j^2}{q} - \frac{\left(\sum_{j=1}^{p} T_j\right)^2}{n} = \left(10000/4 + 1296/4 + 1156/2\right) - \left((-30)^2/10\right) =$$

$$= 2500 + 324 + 578 - 90 = 3312.$$

Найдем остаточную сумму квадратов отклонений:

$$S_{\text{ост}} = S_{\text{общ}} - S_{\text{факт}} = 3982 - 3312 = 670.$$

Найдем факторную и остаточную дисперсии:

$$s^2_{\text{факт}} = \frac{S_{\text{факт}}}{p-1} = \frac{3312}{3-1} = 1656;$$

$$s^2_{\text{ост}} = \frac{S_{\text{ост}}}{n-p} = \frac{670}{10-3} = 95,71.$$

Сравним факторную и остаточную дисперсии по критерию F, для чего найдем наблюдаемое значение критерия:

$$F_{\text{набл}} = s^2_{\text{факт}} \big/ s^2_{\text{ост}} = 1656/95,71 = 17,3.$$

Учитывая, что число степеней свободы числителя $k_1 = p-1 = 2$, а знаменателя $k_2 = p(q-1) = n-p = 10-3 = 7$, и уровень значимости $\alpha = 0,01$, по таблице приложения П.6 находим критическую точку:

$$F_{\text{кр}}(0,01;\ 2;\ 7) = 9,55.$$

Так как $F_{\text{набл}} > F_{\text{кр}}$, – нулевую гипотезу о равенстве нулевых средних отвергаем. Другими словами, групповые средние различаются значимо.

Задачи.

В задачах 6.1-6.3 требуется при уровне значимости 0,05 проверить нулевую гипотезу о равенстве групповых средних. Предполагается, что выборки извлечены из нормальных совокупностей с одинаковыми генеральными дисперсиями.

6.1.

Номер испытания	Уровни фактора F_j				
i	F_1	F_2	F_3	F_4	F_5
1	42	66	35	64	70
2	55	91	50	70	79
3	67	96	60	79	88
4	67	98	69	81	90
$\overline{x}_{\text{гр}j}$					

6.2.

Номер испытания	Уровни фактора F_j			
i	F_1	F_2	F_3	F_4
1	6	6	9	7
2	7	7	12	9
3	8	11	13	10
4	11	12	14	10
$\overline{x}_{\text{гр}j}$				

6.3.

Номер испытания	Уровни фактора F_j		
i	F_1	F_2	F_3
1	37	60	69
2	47	86	100
3	40	67	98
4	60	92	
5		95	
6		98	
$\bar{x}_{\text{гр}j}$			

7. КОРРЕЛЯЦИОННЫЙ АНАЛИЗ

Термин «корреляция» был введен в науку выдающимся английским естествоиспытателем Френсисом Гальтоном в 1886 году. Однако точную формулу для подсчета коэффициента корреляции разработал его ученик Карл Пирсон.

Задачи с одним выходным параметром имеют очевидные преимущества. Но на практике чаще всего приходится учитывать несколько выходных параметров. Иногда их число довольно велико. Так, например, при производстве лакокрасочной продукции приходится учитывать физико-механические, технологические, экономические, художественно-эстетические и другие параметры (прочность, долговечность, эластичность, водоотталкивающие свойства, химическая стойкость, антисептическая стойкость, огнестойкость и т.д.). Математические модели можно построить для каждого из параметров, но одновременно оптимизировать несколько функций невозможно.

Обычно оптимизируется одна функция, наиболее важная с точки зрения цели исследования, при ограничениях, налагаемых другими функциями. Поэтому из многих выходных параметров выбирается один в качестве параметра оптимизации, а остальные служат ограничениями. Всегда полезно исследовать возможность уменьшения числа выходных параметров. Для этого и используется корреляционный анализ.

С использованием результатов корреляционного анализа исследователь может делать определённые выводы о наличии и характере взаимозависимости, что уже само по себе может представлять существенную информацию об исследуемом объекте. Результаты могут подсказать и направление дальнейших исследований, и совокупность требуемых методов, в том числе статистических, необходимых для более полного изучения объекта [8].

Особенно реальную пользу применение аппарата корреляционного анализа может принести на стадии ранних исследований в областях, где характеры причин определённых явлений ещё недостаточно понятны. Это может касаться изучения очень сложных систем различного характера: как технических, так и социальных.

Понятие корреляционной зависимости

Во многих задачах требуется установить и оценить зависимость изучаемой случайной величины Y от одной случайной (или неслучайной) величины X, а затем от нескольких величин.

Две случайные величины могут быть связаны либо функциональной зависимостью, либо зависимостью другого рода, называемой статистической, либо быть независимыми.

Функциональной называется зависимость, если каждому возможному значению величины X соответствует одно возможное значение случайной величины Y: $Y = f(X)$.

Но строгая функциональная зависимость реализуется редко, так как обе величины, или одна из них подвержены еще действию случайных факторов, причем, среди них могут быть и общие для обеих величин (под «общими» подразумевают факторы, воздействующие и на Y, и на X).

Статистической называют зависимость, при которой изменение одной из величин влечет изменение распределения другой. В частности, статистическая зависимость может проявляться в том, что при изменении одной из величин изменяется среднее значение другой; в этом случае статистическую зависимость называют *корреляционной*.

Корреляционная зависимость – это изменения, которые вносят значения одного признака в вероятность появления разных значений другого признака.

Задача корреляционного анализа сводится к установлению направления (положительное или отрицательное) и формы (линейная, нелинейная) связи между варьирующими признаками, измерению ее тесноты, и, наконец, к проверке уровня значимости полученных коэффициентов корреляции.

Корреляционные связи различаются по форме, направлению и степени (силе).

По форме корреляционная связь может быть прямолинейной или криволинейной.

По направлению корреляционная связь может быть положительной («прямой») и отрицательной («обратной»). При положительной прямолинейной корреляции более высоким значениям одного признака соответствуют более высокие значения другого, а более низким значениям одного признака – низкие значения другого (рис. 7.1). При отрицательной корреляции соотношения обратные (рис. 7.2). При положительной корреляции коэффициент корреляции имеет положительный знак, при отрицательной корреляции – отрицательный знак [4].

Степень, сила или теснота корреляционной связи определяется по величине коэффициента корреляции. Сила связи не зависит от ее направленности и определяется по абсолютному значению коэффициента корреляции.

В зависимости от коэффициента корреляции различают следующие корреляционные связи:

1) сильная, или тесная – при значениях модуля коэффициента корреляции $|r| \geq 0,70$;

2) средняя – при $0,50 \leq |r| \leq 0,69$;

3) умеренная – при $0,30 \leq |r| \leq 0,49$;

4) слабая – при $0,20 \leq |r| \leq 0,29$;

5) очень слабая – при $|r| \leq 0,19$.

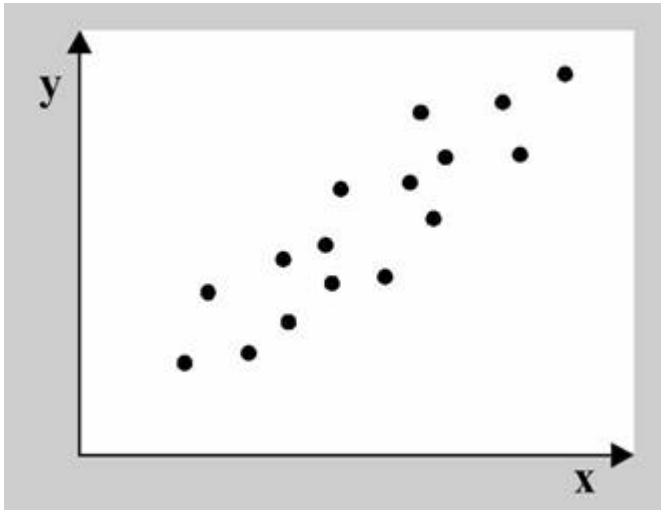

Рис. 7.1. Прямая корреляция

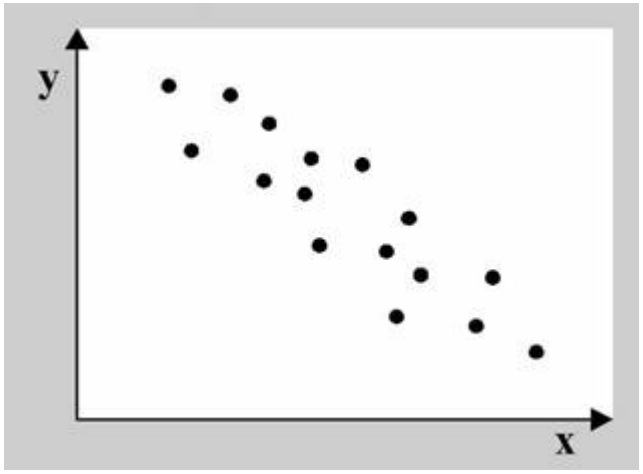

Рис. 7.2. Обратная корреляция

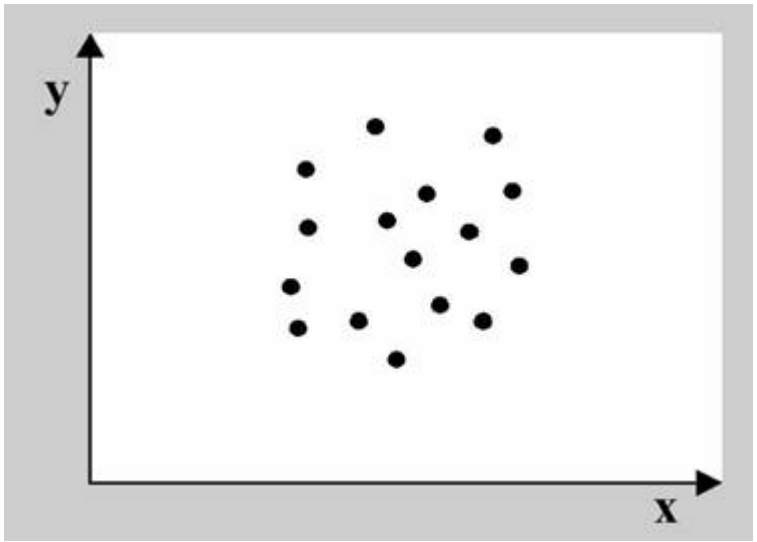

Рис. 7.3. Отсутствие корреляции

Коэффициенты корреляции являются общепринятой в математической статистике характеристикой связи между двумя случайными величинами. Коэффициент корреляции – показатель степени взаимозависимости, статистической связи двух переменных; изменяется в пределах от -1 до +1. Значение коэффициента корреляции 0 указывает на возможное отсутствие зависимости, значение +1 свидетельствует о согласованности переменных.

Различают следующие коэффициенты корреляции:

- дихотомический – показатель связи признаков (переменных), измеряемых по дихотомическим шкалам наименований;
- Пирсона (Pearson product-moment correlation) – коэффициент корреляции, используемый для континуальных переменных;
- ранговой корреляции Спирмена (Spearmens rank-order correlation) – коэффициент корреляции для переменных, измеренных в порядковых (ранговых) шкалах;
- точечно-бисериальной корреляции (point-biserial correlation) – коэффициент корреляции, применяемый в случае анализа отношения переменных, одна из которых измерена в континуальной шкале, а другая – в строго дихотомической шкале наименований;
- j – коэффициент корреляции, используемый в случае, если обе переменные измерены в дихотомической шкале наименований.
- тетрахорический (четырехпольный) (tetrachoric) – коэффициент корреляции, используемый в случае, если обе переменные измерены в континуальных шкалах.

Корреляционная таблица

При большом числе наблюдений одно и то же значение x может встретиться n_x раз, одно и то же значение y – n_y раз, одна и та же пара чисел (x, y) может наблюдаться n_{xy} раз. Поэтому данные наблюдений группируют, т.е. подсчитывают частоты n_x, n_y, n_{xy}. Все сгруппированные данные заносят в таблицу, которую называют *корреляционной* [5].

Пример корреляционной таблицы смотри в табл. 7.1.

Таблица 7.1

Y	X				
	10	20	30	40	n_y
0,4	5	-	7	14	26
0,6	-	2	6	4	12
0,8	3	19	-	-	22
n_x	8	21	13	18	n=60

В первой строке таблицы указаны наблюдаемые значения (10; 20; 30; 40) признака X, а в первом столбце – наблюдаемые значения (0,4; 0,6; 0,8)

признака Y. На пересечении строк и столбцов находятся частоты n_{xy} наблюдаемых пар значений признаков. Например, частота 5 указывает, что пара чисел (10; 0,4) наблюдалась 5 раз. Все частоты помешаны в прямоугольнике, стороны которого проведены жирными отрезками. Черточка означает, что соответствующая пара чисел, например, (20; 0,4), не наблюдалась.

В последнем столбце записаны суммы частот строк. Например, сумма частот первой строки «жирного» прямоугольника равна $n_y = 5 + 7 + 14 = 26$; это число указывает, что значение признака Y, равное 0,4, в сочетании с различными значениями признака X наблюдалось 26 раз.

В последней строке записаны суммы частот столбцов. Например, число 8 указывает, что значение признака X, равное 10, в сочетании с различными значениями признака Y наблюдалось 8 раз.

В клетке, расположенной в нижнем правом углу таблицы, помещена сумма всех частот – общее число всех наблюдений n. Очевидно, что $\sum n_x = \sum n_y = n$. В рассматриваемом примере: $\sum n_x = 8 + 21 + 13 + 18 = 60$ и $\sum n_y = 26 + 12 + 22 = 60$.

Выборочный коэффициент корреляции. Вычисление

Выборочный коэффициент корреляции определяется равенством:

$$r_{\text{в}} = \frac{\sum n_{xy} xy - n\overline{x}\,\overline{y}}{n\overline{\sigma}_x \overline{\sigma}_y}, \qquad (7.1)$$

где x, y – варианты (наблюдавшиеся значения) признаков X и Y; n_{xy} – частота наблюдавшейся пары вариант (x, y); n – объем выборки (сумма всех частот); $\overline{\sigma}_x$, $\overline{\sigma}_y$ – выборочные средние квадратические отклонения; $\overline{x}$, $\overline{y}$ – выборочные средние.

Известно, что если величины Y и X независимы, то коэффициент корреляции $r = 0$; если $r = \pm 1$, то Y и X связаны *линейной* функциональной зависимостью. Отсюда следует, что коэффициент корреляции r измеряет силу (тесноту) линейной связи между Y и X.

Выборочный коэффициент корреляции $r_{\text{в}}$ является оценкой коэффициента корреляции r генеральной совокупности и поэтому также служит для измерения линейной связи между величинами Y и X. Допустим, что выборочный коэффициент корреляции, найденный по выборке, оказался отличным от нуля. Так как выборка отобрана случайно, то отсюда еще нельзя заключить, что коэффициент корреляции генеральной совокупности также отличен от нуля. Возникает необходимость проверить гипотезу о значимости (существенности) выборочного коэффициента корреляции (или, что то же, о равенстве нулю коэффициента корреляции генеральной совокупно-

сти). Если гипотеза о равенстве нулю генерального коэффициента корреляции будет отвергнута, то выборочный коэффициент корреляции значим, а величины X и Y коррелированны; если гипотеза принята, то выборочный коэффициент корреляции незначим, а величины X и Y не коррелированны.

Пусть требуется по данным корреляционной таблицы вычислить выборочный коэффициент корреляции. Можно значительно упростить расчет, если перейти к условным вариантам, при этом величина $r_в$ не изменится:

$$u_i = \left(x_i - C_1\right)/h_1 \text{ и } \upsilon_j = \left(y_j - C_2\right)/h_2 .$$

В этом случае выборочный коэффициент корреляции вычисляют по формуле

$$r_в = \left(\sum n_{u\upsilon} u\upsilon - n\bar{u}\bar{\upsilon}\right)/\left(n\bar{\sigma}_u\bar{\sigma}_\upsilon\right). \qquad (7.2)$$

Величины $\bar{u}$, $\bar{\upsilon}$, $\bar{\sigma}_u$, $\bar{\sigma}_\upsilon$ можно найти методом произведений, а при малом числе данных – непосредственно исходя из определений этих величин.

Метод произведений дает удобный способ вычисления условных моментов различных порядков вариационного ряда с равноотстоящими вариантами. Зная же условные моменты, нетрудно найти начальные и центральные эмпирические момента. В частности методом произведений удобно вычислять выборочную среднюю и выборочную дисперсию. Целесообразно пользоваться расчетной таблицей, которая составляется так:

4) в первый столбец таблицы записывают выборочные (первоначальные) варианты, располагая их в возрастающем порядке;

5) во второй столбец записывают частоты вариант; складывают все частоты, и их сумму (объем выборки n) помещают в нижнюю клетку столбца;

6) в третий столбец записывают условные варианты $u_i = \left(x_i - C\right)/h$, причем в качестве ложного нуля C выбирают варианту, которая расположена примерно в середине вариационного ряда, и полагают h равным разности между любыми двумя соседними вариантами. Практически же третий столбец заполняется так: в клетке строки, содержащий выбранный ложный нуль, пишут 0; в клетках над нулем пишут последовательно -1, -2, -3 и т.д., в клетках под нулем – 1, 2, 3 и т.д.;

7) умножают частоты на условные варианты и записывают их произведения $n_i u_i$ в четвертый столбец; сложив все полученные по столбцу числа, их сумму $\sum n_i u_i$ помещают в нижнюю клетку столбца;

8) умножают частоты на квадраты условных вариант и записывают их произведения $n_i u_i^2$ в пятый столбец; сложив все полученные числа, их сумму $\sum n_i u_i^2$ помещают в нижнюю клетку столбца;

9) умножают частоты на квадраты условных вариант, увеличенных каждый на единицу, и записывают произведения $n_i(u_1+1)^2$ в шестой контрольный столбец; сложив все полученные числа, их сумму $\sum n_i(u_1+1)^2$ помещают в нижнюю клетку столбца.

При заполнении таблицы целесообразно отдельно складывать отрицательные и положительные числа четвертого столбца, при этом сумму отрицательных чисел A_1 записывают в клетку строки, содержащей ложный нуль, а сумму положительных чисел A_2 записывают в предпоследнюю снизу клетку столбца, тогда $\sum n_i u_i = A_1 + A_2$.

Далее, при вычислении произведений $n_i u_i^2$ пятого столбца целесообразно числа $n_i u_i$ четвертого столбца умножать на u_1.

Шестой столбец служит для контроля вычислений: если сумма $\sum n_i(u_1+1)^2$ окажется равной сумме $\sum n_i u_i^2 + 2\sum n_i u_i + n$ (как и должно быть в соответствии с тождеством $\sum n_i(u_i+1)^2 = \sum n_i u_i^2 + 2\sum n_i u_i + n$), то вычисления проведены правильно.

После того, как расчетная таблица заполнена и проверена правильность вычислений, вычисляют условные моменты:

$$M_1^* = \left(\sum n_i u_i\right)/n, \quad M_2^* = \left(\sum n_i u_i^2\right)/n.$$

Наконец, вычисляют выборочные среднюю и дисперсию по формулам:

$$\overline{x}_в = M_1^* h + C, \quad D_в = \left[M_2^* - (M_1^*)^2\right]h^2.$$

Пример 7.1. Найти методом произведений выборочные среднюю и дисперсию следующего статистического распределения:

Варианты	10,2	10,4	10,6	10,8	11,0	11,2	11,4	11,6	11,8	12,0
Частоты	2	3	8	13	25	20	12	10	6	1

Решение. Составим расчетную табл. 7.2, для чего:

1) запишем варианты в первый столбец;

2) запишем частоты во второй столбец; сумму частот (100) поместим в нижнюю клетку столбца;

3) в качестве ложного нуля выберем варианту 11,0 (эта варианта расположена примерно в середине вариационного ряда); в клетке третьего столбца, которая принадлежит строке, содержащей выбранный ложный нуль, пишем 0; над нулем записываем последовательно -1, -2, -3, -4, а под нулем – 1, 2, 3, 4, 5;

4) произведения частот на условные варианты записываем в четвертый столбец; отдельно находим сумму (-46) отрицательных и отдельной

сумму (103) положительных чисел; сложив эти числа, получаем сумму (57), записываем ее в нижнюю клетку столбца;

5) произведения частот на квадраты условных вариант запишем в пятый столбец; сумму чисел столбца (383) помещаем в нижнюю клетку столбца;

6) произведения частот на квадраты условных вариант, увеличенных на единицу, запишем в шестой контрольный столбец; сумму (597) чисел контрольного столбца помещаем в нижнюю клетку столбца.

Таблица 7.2

1	2	3	4	5	6
x_i	n_i	u_i	n_iu_i	$n_iu_i^2$	$n_i(u_i+1)^2$
10,2	2	-4	-8	32	18
10,4	3	-3	-9	27	12
10,6	8	-2	-16	32	8
10,8	13	-1	-13	13	0
11,0	25	0	A_1=–46		25
11,2	20	1	20	20	80
11,4	12	2	24	48	108
11,6	10	3	30	90	160
11,8	6	4	24	96	150
12,0	1	5	5	25	36
			A_2=103		
	n=100		$\sum n_iu_i = 57$	$\sum n_iu_i^2 = 383$	$\sum n_i(u_i+1)^2 = 597$

Контроль: $\sum n_i{u_i}^2 + 2\sum n_iu_i + n = 383 + 2\cdot 57 + 100 = 597$;

$\sum n_i(u_i+1)^2 = 597$.

Вычисления произведены правильно.

Вычислим условные моменты первого и второго порядков:

$$M_1^* = \left(\sum n_iu_i\right)\Big/ n = 57/100 = 0,57;$$

$$M_2^* = \left(\sum n_i{u_i}^2\right)\Big/ n = 383/100 = 3,83.$$

Найдем шаг: $h = 10,4 - 10,2 = 0,2$.

Вычислим искомые выборочные среднюю и дисперсию:

$$\bar{x}_в = M_1^* h + C = 0,57\cdot 0,2 + 11,0 = 11,1;$$

$$D_в = \left[M_2^* - \left(M_1^*\right)^2\right]h^2 = \left[3,83 - \left(0,57\right)^2\right]\cdot 0,2^2 = 0,14.$$

Вернемся к выражению (7.2).

Следующим шагом необходимо указать способ вычисления $\sum n_{u\upsilon}u\upsilon$, где $n_{u\upsilon}$ – частота пары условных вариант $(u,\ \upsilon)$.

Доказано [5], что справедливы формулы:

$$\sum n_{u\upsilon} u\upsilon = \sum \upsilon U, \text{ где } U = \sum n_{u\upsilon} u,$$

$$\sum n_{u\upsilon} u\upsilon = \sum u V, \text{ где } V = \sum n_{u\upsilon} \upsilon. \qquad (7.3)$$

Для контроля целесообразно выполнить расчеты по обеим формулам и сравнить результаты; их совпадение свидетельствует о правильности вычислений.

В примере разберем, как пользоваться приведенными формулами (7.3).

Пример 7.2. Вычислить $\sum n_{u\upsilon} u\upsilon$ по данным корреляционной табл. 7.3.

Таблица 7.3

Y	X						n_y
	10	20	30	40	50	60	
15	5	7	–	–	–	–	12
25	–	20	23	–	–	–	43
35	–	–	30	47	2	–	79
45	–	–	10	11	20	6	47
55	–	–	–	9	7	3	19
n_x	5	27	63	67	29	9	n=200

Решение. Перейдем к условным вариантам: $u_i = (x_i - C_1)/h_1 = (x_i - 40)/10$ (в качестве ложного нуля C_1 взята варианта $x = 40$, расположенная примерно в середине вариационного ряда; шаг h_1 равен разности между двумя соседними вариантами: 20-10=10) и $\upsilon_j = (y_j - C_2)/h_2 = (y_j - 35)/10$ (в качестве ложного нуля C_2 взята варианта $y = 35$, расположенная примерно в середине вариационного ряда; шаг h_2 равен разности между двумя соседними вариантами: 25-15=10).

Составим корреляционную таблицу в условных вариантах. Практически это делают так: в первом столбце вместо ложного нуля C_2 (варианты 35) пишут 0; над нулем последовательно записывают -1, -2, под нулем пишут 1, 2. В первой строке вместо ложного нуля C_1 (варианты 40) пишут 0; слева от нуля последовательно записывают -1, -2, -3, справа от нуля пишут 1, 2. Все остальные данные переписывают из первоначальной корреляционной таблицы. В итого получаем корреляционную табл. 7.4 в условных вариантах.

Таблица 7.4

υ	u						n_υ
	-3	-2	-1	0	1	2	
-2	5	7	–	–	–	–	12
-1	–	20	23	–	–	–	43
0	–	–	30	47	2	–	79
1	–	–	10	11	20	6	47
2	–	–	–	9	7	3	19
n_u	5	27	63	67	29	9	n=200

Теперь для вычисления искомой суммы $\sum n_{u\upsilon} u\upsilon$ составим расчетную табл. 7.5.

Пояснения к составлению табл. 7.5:

1) В каждой клетке, в которой частота $n_{u\upsilon} \neq 0$, записывают в правом верхнем углу произведение частоты $n_{u\upsilon}$ на варианту u. Например, в правых верхних углах клеток первой строки записаны произведения: $5 \cdot (-3) = -15$; $7 \cdot (-2) = 14$.

2) Складывают все числа, помещенные в правых верхних углах одной строки, и их сумму записывают в клетку этой же строки столбца U. Например, для первой строки: $U = -15 + (-14) = -29$.

3) Умножают варианту υ на U и полученное произведение записывают в клетку этой же строки, т.е. в клетку столбца υU. Например, в первой строке таблицы $\upsilon = -2$, $U = -29$, следовательно, $\upsilon U = (-2) \cdot (-29) = 58$.

4) Наконец, сложив все числа столбца υU, получают сумму $\sum_{\upsilon} \upsilon U$, которая равна искомой сумме $\sum n_{u\upsilon} u\upsilon$. Например, для табл. 7.5 имеем $\sum_{\upsilon} \upsilon U = 169$, следовательно, искомая сумма $\sum n_{u\upsilon} u\upsilon = 169$.

Для контроля аналогичные вычисления производят по столбцам: произведения $n_{u\upsilon}\upsilon$ записывают в левый нижний угол клетки, содержащей частоту $n_{u\upsilon} \neq 0$; все числа, помещенные в левых нижних углах клеток одного столбца, складывают, и их сумму записывают в строку V; далее умножают каждую варианту u на V и результат записывают в клетках последней строки.

Наконец, сложив все числа последней строки, получают сумму $\sum_{u} uV$, которая также равна искомой сумме $\sum n_{u\upsilon} u\upsilon$. Например, для табл. 7.5 имеем $\sum_{u} uV = 169$, следовательно, $\sum n_{u\upsilon} u\upsilon = 169$.

Теперь, когда мы научились вычислять $\sum n_{u\upsilon} u\upsilon$, вычислим, наконец, выборочный коэффициент корреляции согласно выражению (7.2). Рассмотрим пример на отыскание выборочного коэффициента корреляции.

Таблица 7.5

υ	u						$U=\sum n_{u\upsilon}u$	υU
	-3	-2	-1	0	1	2		
-2	-15 5 -10	-14 7 -14	-	-	-	-	-29	58
-1	-	-40 20 -20	-23 23 -23	-	-	-	-63	63
0	-	-	-30 30 0	0 47 0	2 2 0	-	-28	0
1	-	-	-10 10 10	0 11 11	20 20 20	12 6 6	22	22
2	-	-	-	0 9 18	7 7 14	6 3 6	13	26
$V=\sum n_{u\upsilon}\upsilon$	-10	-34	-13	29	34	12		$\sum_{\upsilon}\upsilon U=169$
uV	30	68	13	0	34	24	$\sum_{u}uV=169$	контроль

Пример 7.3. Вычислить выборочный коэффициент корреляции $r_{\text{в}} = \dfrac{\sum n_{u\upsilon} u\upsilon - n\overline{u}\,\overline{\upsilon}}{n\overline{\sigma}_u \overline{\sigma}_\upsilon}$ по данным корреляционной табл. 7.3.

Решение. Перейдя к условным вариантам, получим корреляционную табл. 7.4. Величины $\overline{u}$, $\overline{\upsilon}$, $\overline{\sigma}_u$, $\overline{\sigma}_\upsilon$ можно найти методом произведений, однако, поскольку числа u_i и υ_j малы, вычислим $\overline{u}$ и $\overline{\upsilon}$, исходя из определения средней, а $\overline{\sigma}_u$ и $\overline{\sigma}_\upsilon$ по формулам:

$$\overline{\sigma}_u = \sqrt{\overline{u^2} - \left(\overline{u}\right)^2}, \quad \overline{\sigma}_\upsilon = \sqrt{\overline{\upsilon^2} - \left(\overline{\upsilon}\right)^2}.$$

Найдем $\overline{u}$ и $\overline{\upsilon}$:

$$\overline{u} = \left(\sum n_u u\right)\Big/ n = \left[5\cdot(-3) + 27\cdot(-2) + 63\cdot(-1) + 29\cdot 1 + 9\cdot 2\right]\Big/ 200 = -0{,}425;$$

$$\overline{\upsilon} = \left(\sum n_\upsilon \upsilon\right)\Big/ n = \left[12\cdot(-2) + 43\cdot(-1) + 47\cdot 1 + 19\cdot 2\right]\Big/ 200 = 0{,}09.$$

Вычислим вспомогательную величину $\overline{u^2}$, а затем $\overline{\sigma}_u$:

$$\overline{u^2} = \left(\sum n_u u^2\right)\Big/ n = \left(5\cdot 9 + 27\cdot 4 + 63\cdot 1 + 29\cdot 1 + 9\cdot 4\right)\Big/ 200 = 1{,}405;$$

$$\overline{\sigma}_u = \sqrt{\overline{u^2} - \left(\overline{u}\right)^2} = \sqrt{1{,}405 - \left(0{,}425\right)^2} = 1{,}106.$$

Аналогично вычисляем $\overline{\upsilon^2}$, а затем $\overline{\sigma}_\upsilon$:

$$\overline{\upsilon^2} = \left(\sum n_\upsilon \upsilon^2\right)\Big/ n = \left(12\cdot 4 + 43\cdot 1 + 47\cdot 1 + 19\cdot 4\right)\Big/ 200 = 1{,}07;$$

$$\overline{\sigma}_\upsilon = \sqrt{\overline{\upsilon^2} - \left(\overline{\upsilon}\right)^2} = \sqrt{1{,}07 - \left(0{,}09\right)^2} = 1{,}03.$$

Найдем искомый выборочный коэффициент корреляции, учитывая, что ранее уже вычислена сумма $\sum n_{u\upsilon} u\upsilon = 169$:

$$r_{\text{в}} = \frac{\sum n_{u\upsilon} u\upsilon - n\overline{u}\,\overline{\upsilon}}{n\overline{\sigma}_u \overline{\sigma}_\upsilon} = \frac{169 - 200\cdot\left(-0{,}425\right)\cdot 0{,}09}{200\cdot 1{,}106\cdot 1{,}03} = 0{,}775.$$

Итак, выборочный коэффициент корреляции равен $r_{\text{в}} = 0{,}775$. Таким образом, в рассматриваемом примере мы имеем сильную (тесную) корреляционную связь.

Выборочное уравнение прямой линии регрессии

После того, как мы научились вычислять выборочный коэффициент корреляции, разберем пример на отыскание уравнения прямой линии регрессии.

Поскольку при нахождении $r_в$ мы уже вычислили $\overline{u}$, $\overline{\upsilon}$, $\overline{\sigma}_u$, $\overline{\sigma}_\upsilon$, целесообразно воспользоваться формулами:

$$\overline{\sigma}_x = h_1\overline{\sigma}_u; \quad \overline{\sigma}_y = h_2\overline{\sigma}_\upsilon; \quad \overline{x} = \overline{u}h_1 + c_1; \quad y = \overline{\upsilon}h_2 + c_2.$$

Студенту рекомендуется вывести данные формулы самостоятельно.

Пример 7.4. Найти выборочное уравнение прямой линии регрессии Y на X по данным корреляционной табл. 7.3.

Решение. Напишем искомое уравнение в общем виде:

$$\overline{y}_x - \overline{y} = r_в\frac{\overline{\sigma}_y}{\overline{\sigma}_x}\left(x - \overline{x}\right). \qquad (7.4)$$

Коэффициент корреляции вычислен в примере 7.3. Остается найти $\overline{x}$, $\overline{y}$, $\overline{\sigma}_x$, $\overline{\sigma}_y$:

$$\overline{x} = \overline{u}h_1 + c_1 = -0,424\cdot 10 + 40 = 35,75;$$

$$y = \overline{\upsilon}h_2 + c_2 = 0,09\cdot 10 + 35 = 35,9;$$

$$\overline{\sigma}_x = h_1\overline{\sigma}_u = 1,106\cdot 10 = 11,06; \qquad \overline{\sigma}_y = h_2\overline{\sigma}_\upsilon = 1,03\cdot 10 = 10,3.$$

Подставив полученные величины в (7.4), получим искомое уравнение:

$$\overline{y}_x - 35,9 = 0,775\frac{10,3}{11,06}\left(x - 35,75\right),$$

что окончательно составляет

$$\overline{y}_x = 0,72x + 10,09.$$

Сравним условные средние, вычисленные по: а) этому уравнению; б) данным корреляционной табл. 7.3.

Например, при $x = 30$:

а) $\overline{y}_{30} = 0,72\cdot 30 + 10,09 = 31,69$;

б) $\overline{y}_{30} = \left(23\cdot 25 + 30\cdot 35 + 10\cdot 45\right)/63 = 32,94$.

Как видим, согласование расчетного и наблюдаемого условных средних – удовлетворительное.

Мера любой корреляционной связи

Мы рассмотрели оценку тесноты линейной корреляционной связи. Как же возможно оценить тесноту любой корреляционной связи?

Пусть данные наблюдений над количественными признаками X и Y сведены в корреляционную таблицу. Можно считать, что тем самым наблюдаемые значения Y разбиты на группы; каждая группа содержит те

значения Y, которые соответствуют определенному значению X. Например, данные наблюдений сведены в корреляционную табл. 7.6.

Таблица 7.6

Y	X	
	8	9
3	4	13
5	6	7
n_x	10	20
$\overline{y}_x$	4,2	3,7

К первой группе относятся те десять значений Y (четыре раза наблюдалось $y_1 = 3$ и шесть раз $y_2 = 5$), которые соответствуют $x_1 = 8$.

Ко второй группе относятся те двадцать значений Y (тринадцать раз наблюдалось $y_1 = 3$ и семь раз $y_2 = 5$), которые соответствуют $x_2 = 9$.

Условные средние теперь можно назвать групповыми средними:

– групповая средняя первой группы $\overline{y}_8 = (4 \cdot 3 + 6 \cdot 5)/10 = 4,2$;

– групповая средняя второй группы $\overline{y}_9 = (13 \cdot 3 + 7 \cdot 5)/20 = 3,7$.

Поскольку все значения признака Y разбиты на группы, можно представить общую дисперсию признака в виде суммы внутригрупповой и межгрупповой дисперсий:

$$D_{\text{общ}} = D_{\text{внгр}} + D_{\text{межгр}}. \qquad (7.5)$$

При этом согласно определениям функциональной и корреляционной зависимостей Гмурмана В.Е., справедливы следующие утверждения:

1) если Y связан с X функциональной зависимостью, то $D_{\text{межгр}}/D_{\text{общ}} = 1$;

2) если Y связан с X корреляционной зависимостью, то $D_{\text{межгр}}/D_{\text{общ}} < 1$.

Таким образом, видно, что чем связь между признаками ближе к функциональной, тем меньше $D_{\text{внгр}}$, и, следовательно, чем больше приближается $D_{\text{межгр}}$ к $D_{\text{общ}}$, а, значит, отношение $D_{\text{межгр}}/D_{\text{общ}}$ – к единице. Отсюда ясно, что целесообразно рассматривать в качестве меры тесноты корреляционной зависимости отношение межгрупповой дисперсии к общей, или, что то же, отношение межгруппового среднего квадратического отклонения к общему среднему квадратическому отклонению.

Выборочное корреляционное отношение. Свойства

Для оценки тесноты линейной корреляционной связи между признаками в выборке служит выборочный коэффициент корреляции.

Для оценки связи *нелинейной* корреляционной связи вводят новые сводные характеристики:

η_{yx} – выборочное корреляционное отношение Y к X;

η_{xy} – выборочное корреляционное отношение X к Y.

Выборочным корреляционным отношением Y к X называют отношение межгруппового среднего квадратического отклонения к общему среднему квадратическому отклонению признака Y:

$$\eta_{yx} = \sigma_{\text{межгр}} \big/ \sigma_{\text{общ}},$$

или в других обозначениях:

$$\eta_{yx} = \sigma_{\bar{y}_x} \big/ \bar{\sigma}_y .$$

Здесь

$$\sigma_{\bar{y}_x} = \sqrt{D_{\text{межгр}}} = \sqrt{\left(\sum n_x \left(\bar{y}_x - \bar{y}\right)^2\right) \Big/ n};$$

$$\bar{\sigma}_y = \sqrt{D_{\text{общ}}} = \sqrt{\left(\sum n_y \left(y - \bar{y}\right)^2\right) \Big/ n},$$

где n – объем выборки (сумма всех частот); n_x – частота значения x признака X; n_y – частота значения y признака Y; $\bar{y}$ – общая средняя признака Y; $\bar{y}_x$ – условная средняя признака Y.

Аналогично определяется выборочное корреляционное отношение X к Y:

$$\eta_{xy} = \sigma_{\bar{x}_y} \big/ \bar{\sigma}_x .$$

Пример 7.5. Найти η_{yx} по данным корреляционной табл. 7.7.

Таблица 7.7

Y	X			n_y
	10	20	30	
15	4	26	6	38
25	6	–	6	12
n_x	10	26	12	$n = 50$
$\bar{y}_x$	21	15	20	

Решение. Найдем общую среднюю:

$$\bar{y} = \left(\sum n_y y\right) \big/ n = \left(38 \cdot 15 + 12 \cdot 25\right) \big/ 50 = 17{,}4 .$$

Найдем общее среднее квадратическое отклонение:

$$\overline{\sigma}_y = \sqrt{\left(\sum n_y \left(y - \overline{y}\right)^2\right) \Big/ n} = \sqrt{\left(38 \cdot (15 - 17{,}4)^2 + 12 \cdot \left(25 - 17{,}4\right)^2\right) \Big/ 50} = 4{,}27 .$$

Найдем межгрупповое среднее квадратическое отклонение:

$$\sigma_{\overline{y}_x} = \sqrt{\left(\sum n_x \left(\overline{y}_x - \overline{y}\right)^2\right) \Big/ n} =$$

$$= \sqrt{\left[10\left(21 - 17{,}4\right)^2 + 28\left(15 - 15{,}4\right)^2 + 12\left(20 - 17{,}4\right)^2\right] \Big/ 50} = 2{,}73 .$$

Искомое корреляционное отношение равно

$$\eta_{yx} = \sigma_{\overline{y}_x} \Big/ \overline{\sigma}_y = 2{,}73/4{,}27 = 0{,}64 .$$

Поскольку η_{xy} обладает теми же свойствами, что и η_{yx}, перечислим свойства только одного из них – η_{yx}, которое для упрощения обозначим через η и для простоты назовем «корреляционным отношением»:

– свойство 1: *корреляционное отношение удовлетворяет двойному неравенству* $0 \le \eta \le 1$*;*

– свойство 2: *если* $\eta = 0$*, то признак* Y *с признаком* X *корреляционной зависимостью не связан;*

– свойство 3: *если* $\eta = 1$*, то признак* Y *связан с признаком* X *функциональной зависимостью;*

– свойство 4: *выборочное корреляционное отношение не меньше абсолютной величины выборочного коэффициента корреляции:* $\eta \ge \left|r_{\text{в}}\right|$*;*

– свойство 5: *если выборочное корреляционное отношение равно абсолютной величине выборочного коэффициента корреляции, то имеет место точная линейная корреляционная зависимость.*

Другими словами, если $\eta = \left|r_{\text{в}}\right|$, то точки $\left(x_1;\ y_1\right)$, $\left(x_2;\ y_2\right)$, …, $\left(x_n;\ y_n\right)$ лежат на прямой линии регрессии, найденной способом наименьших квадратов.

И, говоря о достоинствах и недостатках корреляционного отношения как меры корреляционной связи, следует отметить, что, поскольку в рассуждениях не делалось никаких допущений о форме корреляционной связи, – η служит *мерой тесноты связи любой, в том числе и линейной, формы.* В этом состоит преимущество корреляционного отношения перед коэффициентом корреляции, который оценивает тесноту лишь линейной зависимости.

Вместе с тем, корреляционное отношение обладает недостатком: оно не позволяет судить, насколько близко расположены точки, найденные по данным наблюдений, к кривой определенного вида, например, к параболе,

гиперболе и т.п. Это как раз и объясняется тем, что при определении корреляционного отношения форма связи во внимание не принималась.

Задачи.

В задачах 7.1-7.2 даны корреляционные табл. 7.8 и 7.9. Найти: а) $r_{в}$; б) выборочные уравнения прямых регрессии; в) η_{yx} и η_{xy}.

7.1.

Таблица 7.8

Y	X					
	5	10	15	20	n_y	$\overline{x}_y$
10	2	–	–	–	2	5
20	5	4	1	–	10	8
30	3	8	6	3	20	12,25
40	–	3	6	6	15	16
50	–	–	2	1	3	16,67
n_x	10	15	15	10	$n=50$	
$\overline{y}_x$	21	29,33	36	38		

7.2.

Таблица 7.9

Y	X							
	65	95	125	155	185	215	n_y	$\overline{x}_y$
30	5	–	–	–	–	–	5	65
40	4	12	–	–	–	–	16	87,5
50	–	8	5	4	–	–	17	101,18
60	–	1	5	7	2	–	15	145
70	–	–	–	–	1	1	2	200
n_x	9	21	10	11	3	1	$n=55$	
$\overline{y}_x$	34,44	44,76	55	56,36	63,33	70		

ЗАКЛЮЧЕНИЕ

В учебном пособии изложены некоторые основные сведения о методах обработки экспериментальных данных. Учебное пособие, составленное по одноименному курсу лекций, читаемому студентам направления 27.03.01 – Стандартизация и метрология, также может быть полезно аспирантам и научным сотрудникам, обрабатывающим экспериментальные данные.

Наибольшее внимание в учебном пособии уделено таким методам обработки экспериментальных данных, как проверка статистических гипотез, методы оценки параметров распределения, аппроксимация закона распределения экспериментальных данных, однофакторный дисперсионный анализ и корреляционный анализ.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. Адлер, Ю.П. Планирование эксперимента при поиске оптимальных условий [Текст] / Ю.П. Адлер, Ю.В. Грановский, Е.В. Маркова. – М.: Наука, 1976.
2. Боровиков, В.П. STATISTICA – Статистический анализ и обработка данных в среде Windows [Текст] / В.П. Боровиков, И.П. Боровиков. – М.: Информационно-издательский дом "Филинъ", 1998.
3. Браверман, Э.М. Структурные методы обработки эмпирических данных [Текст] / Э.М. Браверман, И.Б. Мучник. – М.: Наука, 1983.
4. Вентцель, Е.С. Теория вероятностей [Текст]: учеб. для вузов / Е.С. Вентцель. – 6-е изд. стер. – М.: Высшая школа, 1999.
5. Гмурман, В.Е. Теория вероятностей и математическая статистика [Текст] / В.Е. Гмурман. – М.: Высшая школа, 2003.
6. Дубров, А.М. Многомерные статистические методы [Текст] / А.М. Дубров, В.С. Мхитарян, Л.И. Трошин. – М.: Финансы и статистика, 1998.
7. Елисеева, И.И. Общая теория статистики [Текст] / И.И. Елисеева, М.М. Юзбашев. – М.: Финансы и статистика, 1999.
8. Иванов, А.Ю. Военно-технические основы построения и математическое моделирование перспективных средств и комплексов автоматизации [Текст] / А.Ю. Иванов, С.П. Полковников, Г.Б. Ходасевич. – СПб.: ВАС, 1997.
9. Кендал, М. Статистические выводы и связи [Текст] / М. Кендал, А. Стьюарт. – М.: Наука, 1973.
10. Кендал, М. Теория распределений [Текст] / М. Кендал, А. Стьюарт. – М.: Наука, 1966.
11. Корн, Г. Справочник по математике (для научных работников и инженеров) [Текст] / Г. Корн, Т. Корн. – М.: Наука, 1974.
12. Кремер, Н.Ш. Теория вероятности и математическая статистика [Текст] / Н.Ш. Кремер. – М.: Юнити–Дана, 2002.
13. Надежность и эффективность в технике: Справочник в десяти томах. Т.6. Экспериментальная отработка и испытания [Текст]. – М.: Машиностроение, 1989.
14. Рябинин, И.А. Надежность и безопасность структурно-сложных систем [Текст] / И.А. Рябинин. – СПб.: Политехника, 2000.
15. Северцев, Н.А. Надежность сложных систем в эксплуатации и обработке [Текст] / Н.А. Северцев. – М.: Высшая школа, 1989.
16. Скрипник, В.М. Оценка надежности технических систем по цензурированным выборкам [Текст] / В.М. Скрипник, А.Е. Назин. – Минск: Наука и техника, 1981.
17. Ходасевич, Г.Б. Обработка экспериментальных данных на ЭВМ: обработка одномерных массивов [Текст]: учеб. пособие / Г.Б. Ходасевич. – СПб.: ГУ Телекоммуникаций, 2008 – 60 с. (http://dvo.sut.ru/libr/opds/i130hod2/index.htm

ПРИЛОЖЕНИЕ

Начертания букв латинского и греческого алфавитов, используемых в научных текстах

Латинский алфавит

Начертание		Название	Начертание		Название
прямое	*курсивное*		прямое	*курсивное*	
Aa	*Aa*	а	Nn	*Nn*	эн
Bb	*Bb*	бэ	Oo	*Oo*	о
Cc	*Cc*	це	Pp	*Pp*	пе
Dd	*Dd*	дэ	Qq	*Qq*	ку
Ee	*Ee*	э	Rr	*Rr*	эр
Ff	*Ff*	эф	Ss	*Ss*	эс
Gg	*Gg*	гэ	Tt	*Tt*	тэ
Hh	*Hh*	аш	Uu	*Uu*	у
Ii	*Ii*	и	Vv	*Vv*	вэ
Jj	*Jj*	йот	Ww	*Ww*	дубль-вэ
Kk	*Kk*	ка	Xx	*Xx*	икс
Ll	*Ll*	эль	Yy	*Yy*	игрек
Mm	*Mm*	эм	Zz	*Zz*	зет

Греческий алфавит

Начертание		Название	Начертание		Название
прямое	*курсивное*		прямое	*курсивное*	
Αα	*Αα*	альфа	Νν	*Νν*	ни
Ββ	*Ββ*	бета	Ξξ	*Ξξ*	кси
Γγ	*Γγ*	гамма	Οο	*Οο*	омикрон
Δδ	*Δδ*	дельта	Ππ	*Ππ*	пи
Εε	*Εε*	эпсилон	Ρρ	*Ρρ*	ро
Ζζ	*Ζζ*	дзета	Σσ	*Σσ*	сигма
Ηη	*Ηη*	эта	Ττ	*Ττ*	тау
Θθ	*Θθ*	тета	Υυ	*Υυ*	ипсилон
Ιι	*Ιι*	йота	Φφ	*Φφ*	фи
Κκ	*Κκ*	каппа	Χχ	*Χχ*	хи
Λλ	*Λλ*	ламбда	Ψψ	*Ψψ*	пси
Μμ	*Μμ*	ми	Ωω	*Ωω*	омега

Таблица П.1

Распределение А.Н. Колмогорова

$P\{\lambda > \lambda\,\alpha\} = \alpha$			
α	0,10	0,05	0,01
$\lambda\,\alpha$	1,22	1,36	1,63

Таблица П.2

Распределение Мизеса

$P\{n\omega^2_n > n\omega^2_\alpha\} = \alpha$			
α	0,10	0,05	0,01
$\lambda\,\alpha$	0,347	0,461	0,744

Таблица П.3

Распределение хи-квадрат

Вероятность $P\{\chi^2 > \chi^2(\alpha\,;\,k)\} = \alpha$, где k – число степеней свободы							
k	α			k	α		
	0,10	0,05	0,01		0,10	0,05	0,01
1	2,706	3,841	6,635	17	24,769	27,587	33,409
2	4,605	5,991	9,210	18	25,989	28,869	34,805
3	6,251	7,815	11,341	19	27,204	30,144	36,191
4	7,779	9,488	13,277	20	28,412	31,410	37,566
5	9,236	11,070	15,086	21	29,615	32,671	38,932
6	10,645	12,592	16,812	22	30,813	33,924	40,289
7	12,017	14,067	18,475	23	32,007	35,172	41,638
8	13,362	15,507	20,090	24	33,196	36,415	42,980
9	14,684	16,919	21,666	25	34,382	37,652	44,314
10	15987	18,307	23,209	26	35,563	38,885	45,642
11	17,275	19,675	24,725	27	36,741	40,113	46,963
12	18,549	21,026	26,217	28	37,916	41,337	48,278
13	19,812	22,362	27,688	29	39,087	42,557	49,588
14	21,064	23,685	29,141	30	40,256	43,773	50,892
15	22,307	24,996	30,578	40	51,805	55,758	63,691
16	23,542	26,296	32,000	60	74,397	79,082	88,3379

Таблица П.4

Распределение Стьюдента

Вероятность $P\{t > t(k; \alpha)\} = \alpha$, где k – число степеней свободы							
k	α, односторонняя область			k	α, односторонняя область		
	0,10	0,05	0,01		0,10	0,05	0,01
	α, двусторонняя область				α, двусторонняя область		
	0,20	0,10	0,02		0,20	0,10	0,02
1	3,078	6,314	31,821	17	1,333	1,740	2,567
2	1,886	2,920	6,965	18	1,330	1,734	2,552
3	1,638	2,353	4,541	19	1,328	1,729	2,539
4	1,533	2,132	3,747	20	1,325	1,725	2,528
5	1,476	2,015	3,365	21	1,323	1,721	2,518
6	1,440	1,943	3,143	22	1,321	1,717	2,508
7	1,415	1,895	2,998	23	1,319	1,714	2,500
8	1,397	1,860	2,896	24	1,318	1,711	2,492
9	1,383	1,833	2,821	25	1,316	1,708	2,485
10	1,372	1,812	2,764	26	1,315	1,706	2,479
11	1,363	1,796	2,718	27	1,314	1,703	2,473
12	1,356	1,782	2,681	28	1,313	1,701	2,467
13	1,350	1,771	2,650	29	1,311	1,699	2,462
14	1,345	1,761	2,624	30	1,310	1,697	2,457
15	1,341	1,753	2,602	40	1,303	1,684	2,423
16	1,337	1,746	2,583	60	1,296	1,671	2,390

Продолжение приложения

Таблица П.5

Распределение Вилкоксона

Объемы выборок		α/2			Объемы выборок		α/2		
n_1	n_2	0,05	0,025	0,005	n_1	n_2	0,05	0,025	0,005
6	6	28	26	23	7	7	39	36	32
	7	30	27	24		8	41	38	34
	8	31	29	25		9	43	40	35
	9	33	31	26		10	45	42	37
	10	35	32	27		11	47	44	38
	11	37	34	28		12	49	46	40
	12	38	35	30		13	52	48	41
8	8	51	49	43	9	9	66	62	56
	9	54	51	45		10	69	65	58
	10	56	53	47		11	72	68	61
	11	59	55	49		12	75	71	63
	12	62	58	51		13	78	73	65
	13	64	60	53		14	81	76	67
	14	67	62	54		15	84	79	69
10	10	82	78	71	11	11	100	96	87
	11	86	81	73		12	104	99	90
	12	89	84	76		13	108	103	93
	13	92	88	79		14	112	106	96
	14	96	91	81		15	116	110	99
	15	99	94	84		16	120	113	102
	16	103	97	86		17	123	117	105

Таблица П.6

Распределение Р. Фишера (*F*-распределение)

Уровень значимости α = 0,10											
k_2	k_1										
	2	3	4	5	6	7	8	9	10	11	12
2	9,00	9,16	9,24	9,29	9,33	9,35	9,37	9,38	9,39	9,40	9,41
3	5,46	5,39	5,34	5,31	5,28	5,27	5,25	5,24	5,23	5,22	5,22
4	4,32	4,19	4,11	4,05	4,01	3,98	3,95	3,94	3,92	3,91	3,90
5	3,78	3,62	3,52	3,45	3,40	3,37	3,34	3,32	3,30	3,28	3,27
6	3,46	3,29	3,18	3,11	3,05	3,01	2,98	2,96	2,94	2,92	2,90
7	3,26	3,07	2,96	2,88	2,83	2,78	2,75	2,72	2,70	2,68	2,67
8	3,11	2,92	2,81	2,73	2,67	2,62	2,59	2,56	2,54	2,52	2,50
9	3,01	2,81	2,69	2,61	2,55	2,51	2,47	2,44	2,42	2,40	2,38
10	2,92	2,73	2,61	2,52	2,46	2,41	2,38	2,35	2,32	2,30	2,28
11	2,86	2,66	2,54	2,45	2,39	2,34	2,30	2,27	2,25	2,23	2,21
12	2,81	2,61	2,48	2,39	2,33	2,28	2,24	2,21	2,19	2,17	2,15
13	2,76	2,56	2,43	2,35	2,28	2,23	2,20	2,16	2,14	2,12	2,10
14	2,73	2,52	2,39	2,31	2,24	2,19	2,15	2,12	2,10	2,07	2,05

Окончание табл. П.6

Уровень значимости $\alpha = 0,05$											
k_2	k_1										
	2	3	4	5	6	7	8	9	10	11	12
2	19,0	19,2	19,3	19,3	19,3	19,3	19,4	19,4	19,4	19,4	19,4
3	9,55	9,28	9,12	9,01	8,94	8,89	8,85	8,81	8,79	8,76	8,74
4	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6,00	5,96	5,94	5,91
5	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,77	4,74	4,70	4,68
6	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10	4,06	4,03	4,00
7	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,64	3,60	3,57
8	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39	3,35	3,31	3,28
9	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,14	3,10	3,07
10	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,98	2,94	2,91
11	3,98	3,59	3,36	3,20	3,09	3,01	2,95	2,90	2,85	2,82	2,79
12	3,89	3,49	3,26	3,11	3,00	2,91	2,85	2,80	2,75	2,72	2,69
13	3,81	3,41	3,18	3,03	2,92	2,83	2,77	2,71	2,67	2,63	2,60
14	3,74	3,34	3,11	2,96	2,85	2,76	2,70	2,65	2,60	2,57	2,53
Уровень значимости $\alpha = 0,01$											
k_2	k_1										
	2	3	4	5	6	7	8	9	10	11	12
2	99,0	99,2	99,3	99,3	99,3	99,4	99,4	99,4	99,4	99,4	99,4
3	30,8	29,5	28,7	28,2	27,9	27,7	27,5	27,3	27,2	27,1	27,1
4	18,0	16,7	16,0	15,5	15,2	15,0	14,8	14,7	14,6	14,5	14,4
5	13,3	12,1	11,4	11,0	10,7	10,5	10,3	10,2	10,1	10,0	9,9
6	10,9	9,78	9,15	8,75	8,47	8,26	8,10	7,98	7,87	7,79	7,72
7	9,55	8,45	7,85	7,46	7,19	6,99	6,84	6,72	6,62	6,54	6,47
8	8,65	7,59	7,01	6,63	6,37	6,18	6,03	5,91	5,81	5,73	5,67
9	8,02	6,99	6,42	6,06	5,80	5,61	5,47	5,35	5,26	5,18	5,11
10	7,56	6,55	5,99	5,64	5,39	5,20	5,06	4,94	4,85	4,77	4,71
11	7,21	6,22	5,67	5,32	5,07	4,89	4,74	4,63	4,54	4,46	4,40
12	6,93	5,95	5,41	5,06	4,82	4,64	4,50	4,39	4,30	4,22	4,16
13	6,70	5,74	5,21	4,86	4,62	4,44	4,30	4,19	4,10	4,02	3,96
14	6,51	5,56	5,04	4,69	4,46	4,28	4,14	4,03	3,94	3,86	3,80

Окончание приложения

Таблица П.7

Распределение Кохрена

Уровень значимости $\alpha = 0{,}05$										
k_1	m									
	4	5	6	7	8	9	10	16	36	144
2	0,906	0,872	0,853	0,833	0,816	0,801	0,788	0,734	0,660	0,581
3	0,746	0,707	0,677	0,653	0,633	0,617	0,603	0,547	0,475	0,403
4	0,629	0,590	0,560	0,536	0,518	0,502	0,488	0,437	0,372	0,309
5	0,544	0,507	0,478	0,456	0,439	0,424	0,411	0,365	0,307	0,251
6	0,480	0,445	0,418	0,398	0,382	0,368	0,357	0,314	0,261	0,212
7	0,431	0,397	0,373	0,354	0,338	0,326	0,315	0,276	0,228	0,183
8	0,391	0,360	0,336	0,319	0,304	0,293	0,283	0,246	0,202	0,162
9	0,358	0,329	0,307	0,290	0,277	0,266	0,257	0,223	0,182	0,145
10	0,331	0,303	0,282	0,267	0,254	0,244	0,235	0,203	0,166	0,131
12	0,288	0,262	0,244	0,230	0,219	0,210	0,202	0,174	0,140	0,110
15	0,242	0,220	0,203	0,191	0,182	0,174	0,167	0,143	0,144	0,089
20	0,192	0,174	0,160	0,150	0,142	0,136	0,130	0,111	0,088	0,068
40	0,108	0,097	0,089	0,082	0,078	0,075	0,071	0,060	0,046	0,035
60	0,077	0,068	0,062	0,058	0,055	0,052	0,050	0,041	0,031	0,023
Уровень значимости $\alpha = 0{,}01$										
k_1	m									
	4	5	6	7	8	9	10	16	36	144
2	0,959	0,937	0,917	0,899	0,882	0,867	0,854	0,795	0,707	0,606
3	0,834	0,793	0,761	0,734	0,711	0,691	0,674	0,606	0,515	0,423
4	0,721	0,676	0,641	0,613	0,590	0,570	0,554	0,488	0,406	0,325
5	0,633	0,588	0,553	0,526	0,504	0,485	0,470	0,409	0,335	0,264
6	0,564	0,520	0,487	0,461	0,440	0,433	0,408	0,353	0,286	0,223
7	0,508	0,466	0,435	0,411	0,391	0,375	0,362	0,311	0,249	0,193
8	0,463	0,423	0,393	0,370	0,352	0,337	0,323	0,278	0,221	0,170
9	0,425	0,387	0,359	0,338	0,321	0,307	0,295	0,251	0,199	0,152
10	0,393	0,357	0,331	0,311	0,295	0,281	0,270	0,230	0,181	0,138
12	0,343	0,310	0,286	0,268	0,254	0,242	0,232	0,196	0,154	0,116
15	0,288	0,260	0,239	0,223	0,210	0,200	0,192	0,161	0,125	0,093
20	0,229	0,205	0,188	0,175	0,165	0,157	0,150	0,125	0,096	0,071
40	0,128	0,114	0,103	0,096	0,090	0,085	0,082	0,067	0,050	0,036
60	0,090	0,080	0,072	0,067	0,063	0,059	0,057	0,046	0,034	0,025

О Г Л А В Л Е Н И Е

Учебное издание

Максимова Ирина Николаевна

МЕТОДЫ ОБРАБОТКИ
ЭКСПЕРИМЕНТАЛЬНЫХ ДАННЫХ
Учебное пособие

В а в т о р с к о й р е д а к ц и и
В е р с т к а Н.В. Кучина

Подписано в печать 16.05.14. Формат 60×84/16.
Бумага офисная «Снегурочка». Печать на ризографе.
Усл.печ.л. 6,74. Уч.-изд.л. 7,25. Тираж 80 экз.
Заказ № 154.

Издательство ПГУАС.
440028, г. Пенза, ул. Германа Титова, 28